MIDAS MICHIGAN INSTITUTE FOR DATA & AI IN SOCIETY
UNIVERSITY OF MICHIGAN

MIDAS AT 10

BREAKTHROUGHS, PEOPLE, AND IMPACT

TABLE OF CONTENTS

LETTER FROM OUR EXECUTIVE DIRECTOR

A Personal Perspective on Ten Years of MIDAS

When I started my initial position at MIDAS almost nine years ago as its Senior Scientist and Industry Partnership Leader, I did not realize that I'd be part of such an exciting journey of MIDAS: going from a new institute with a bold idea to successfully defining its identity, building a great community, and emerging as a national leader. Throughout the years, the institute's leaders, from the inaugural Directors Drs. Alfred Hero and Brian Athey to the most recent Director Dr. H. V. Jagadish, have built a community that supports collective success with a culture of generosity and ambition.

I also did not expect to witness how data science and AI take the center stage in scientific discovery by not only being fields of study in their own rights, but also becoming essential research methodologies across departments and disciplines. U-M's leadership, especially the former and current leaders in the Office of the Vice President for Research (Drs. Hu, Cunningham, Lupia, Michielssen, and Orr), guided us with strong and sustained vision. The university was one of the nation's first to establish a data science institute. The university was also among the first in the nation to formally recognize AI as central to the mission of the institute, and the importance of data science and AI not only to research but also to society. Hence the change of MIDAS from the Michigan Institute for Data Science to the Michigan Institute for Data and AI in Society.

Looking back on ten years of MIDAS, what stands out to me is not just the collection of amazing research achievements that we helped happen, but the character of our research community. Our researchers and trainees think boldly and bring endless dedication and creativity to discovery and impact. The numerous departments, institutes and programs on campus who collaborate with us have together created a culture where researchers can explore ideas collaboratively, where research trustworthiness and true impact matter, and where everyone's success becomes the community's shared success.

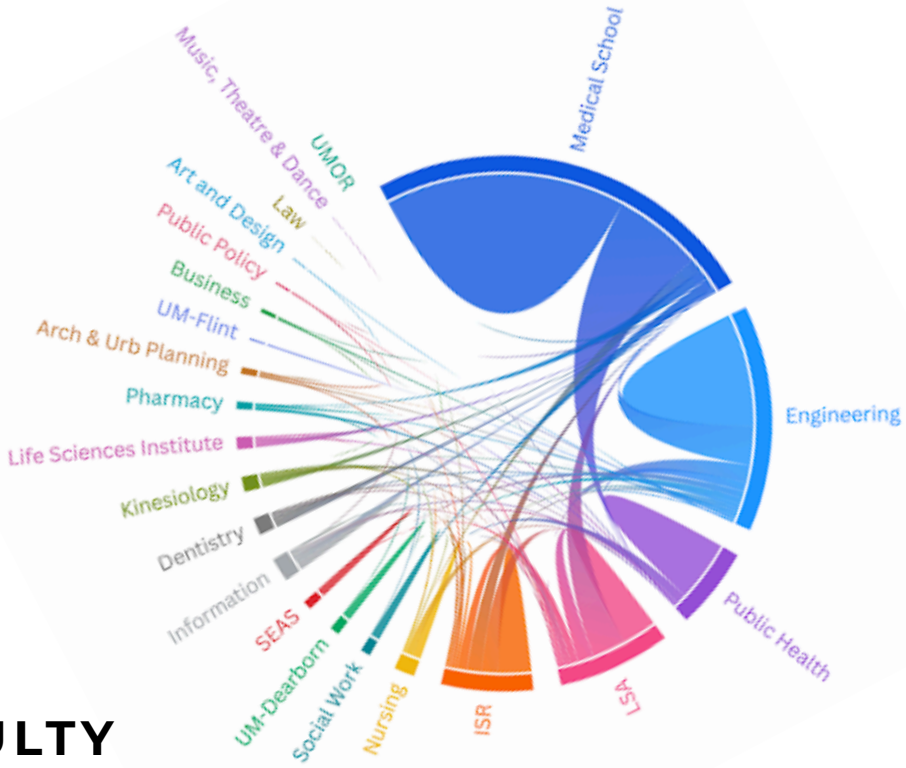
As we enter our second decade, our work will only become more exciting yet more challenging. AI is completely transforming how science is done; it requires us to not just gain new technical expertise but also define a new research ecosystem, the role of human scientists, the meaning of discovery, and the impact science should have in society. MIDAS is here to help our researchers adapt with innovation and agility, and stay true to our mission of discovery and service.

What you see in the following pages is our initial effort to provide a small sample of the incredible achievements of U-M data science and AI in the past 10 years. If you are inspired to write a story about achievements in a significant area of research to add to our collection, please do get in touch with us.

Jing Liu
Executive Director

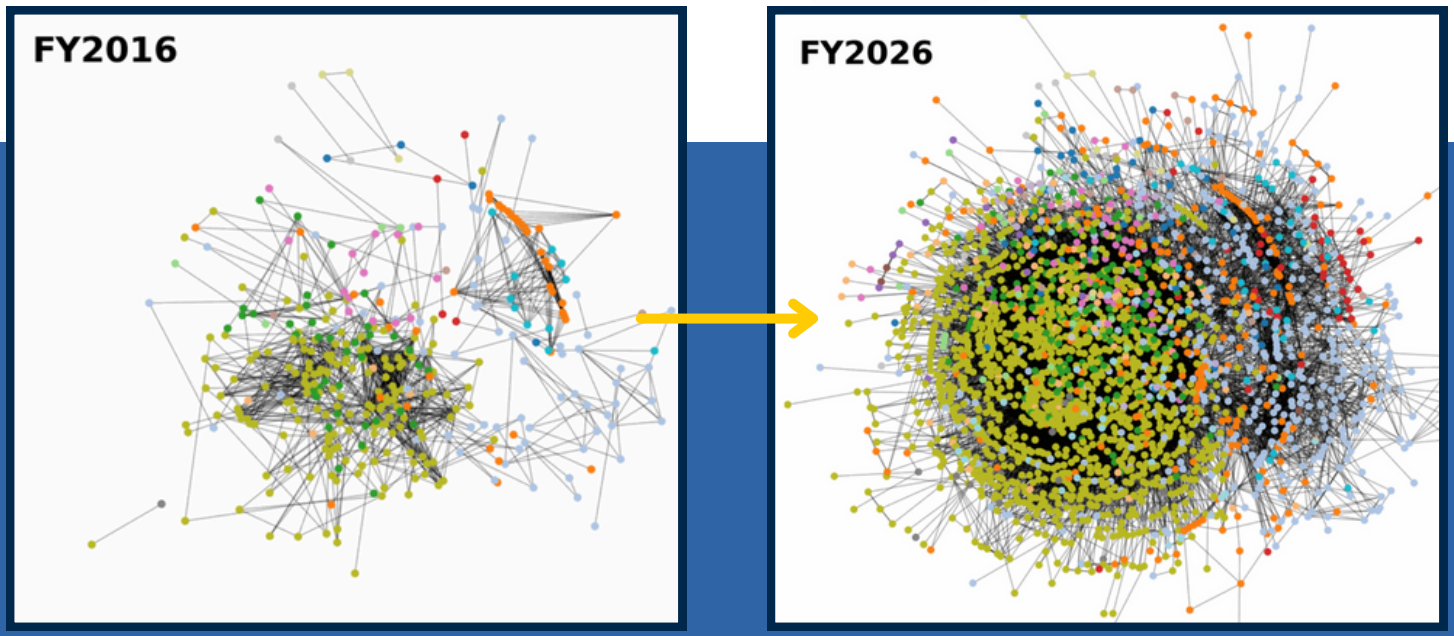
MIDAS AFFILIATED FACULTY

From 2016 to 2026, the MIDAS faculty community has grown from 200 to nearly 700 members across all schools, colleges, and campuses. By providing research resources, fostering groundbreaking ideas and interdisciplinary collaboration, and supporting rigorous and ethical research practices, MIDAS has helped drive breakthroughs in data science and AI.



A DECADE OF FACULTY COLLABORATION

This decade of growth reflects a community-wide commitment to advancing data- and AI-intensive intensive research for lasting scientific and societal impact.



Dots represent **individual faculty affiliates**
Lines represent **collaborations on grant proposals**
Colors represent **different schools and colleges (complete information not included)**

TRAINEES

MIDAS supports a broad community of postdoctoral fellows, graduate students, and student organizations, and hosts one of the largest postdoctoral training programs in the world on data science and AI for scientific discovery. Our postdoctoral programs advance research in AI and data science. Graduate students enrolled in the Data Science Certificate enhance their disciplinary research with the latest in data science and AI instruction and implementation. The MIDAS Student Organizations Council connects student groups across campus to collaborate, share resources, and create inclusive learning and research opportunities.

The four MIDAS postdoc programs (detailed on page 7) has expanded from 7 postdocs in the first cohort in 2019 to a large community that currently is home to almost 50 postdocs who collaborate with colleagues across campus.

The Graduate Data Science Certificate has been a strong component of data science training at MIDAS since the very beginning of MIDAS. Since we have been busy making changes to the curriculum to bring the certificate to leading edge of data science and AI developments on campus.

In addition, we support student-led data science and AI organizations at the undergraduate and graduate levels through the MIDAS Student Organization Council (SOC).

MIDAS postdoc research meeting

MIDAS postdocs

Student Organizations Council

Current and former members of the Council liaise and coordinate with MIDAS, connect with the MIDAS research community, share resources and engage in real-world research.

AI in Medicine (AIM)
BLUElab Data
Leaders in Ethical AI Development Now (LEAD Now)
Master of Applied Data Science (MADS)
Michigan AI Safety Initiative (MAISI)
Michigan Data Science Team
Michigan EcoData
Michigan Student AI Lab
Student Actuaries at Michigan (SAM)
Statistics in the Community (STATCOM)
Tech 4 Social Good

MIDAS SUPPORTED PROJECTS

Since 2016, MIDAS has funded **113** high-risk, high-reward research projects with nearly **$14** million in seed funding, and supported many other projects in various ways. We enabled many research breakthroughs and helped U-M teams secure more than **$239** million in external support for **235** research projects.

SUPPORTING BREAKTHROUGHS ACROSS DOMAINS

Over the past decade, MIDAS has funded research that advances data science and AI across disciplines, supporting innovative projects, enabling transformative discoveries, and strengthening the research infrastructure that drives scientific and societal impact.

AI & ML Foundations

Core advances in data science and AI methodology (see pages 12-14)

Data Infrastructure & Research Workflow

Digitization, privacy, and AI-ready research data pipelines (see pages 15-18, 34-36)

Data, AI, Society & Responsible Technology

Trust, governance, and socially responsible data and AI applications (see pages 12-14, 15-18, 22-24)

Social Science & Policy Analytics

Large-scale social, behavioral, and policy data analysis (see pages 15-18, 19-21, 25-27)

Engineering & Infrastructure

Data- and AI-driven mobility, energy, and engineering research (see pages 28-30)

Health & Biomedical Data Science

AI-enhanced diagnostics, prediction, and data-intensive discovery (see pages 19-21, 31-33, 34-36)

Biodiversity & Digital Biology

AI for biological collections, organism detection, and molecular design (see pages 37-39)

Environment & Climate Analytics

Data-driven modeling of environmental systems (see pages 37-39)

POSTDOCTORAL PROGRAMS

In 2019, the first cohort of seven Michigan Data Science (MDS) Fellows started at MIDAS, marking the beginning of our postdoctoral training program. With the MDS program as the foundation, we received funding from Schmidt Sciences to start the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship program, and the associated AI in Science African Faculty Fellows program. The three programs together train outstanding postdoctoral fellows to apply AI and data science methods across research domains, and have made MIDAS the site of one of the largest postdoctoral programs in data science and AI. Our postdoc community is at the core of MIDAS effort to support a much larger campus community that leverages AI and data science for research breakthroughs. In addition to conducting their own research, the postdocs develop broad research collaboration at U-M and beyond, serve as instructors in our training programs, and develop their own initiatives such as AI for Social Good. To further increase the impact of our postdoc training programs, MIDAS started the Postdoc Affiliates program in 2025.

Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship

Eric and Wendy Schmidt AI in Science African Faculty Fellowship

Michigan Data Science Fellowship

MIDAS Postdoctoral Affiliates Program

MIDAS postdoc fellows participating in research meetings and university discussions

Arya Farahi
Assistant Professor, Department of Statistics and Data Sciences, University of Texas at Austin

Elyse Thulin, PhD
Research Assistant Professor, Institute for Firearm Injury Prevention, University of Michigan

POSTDOC HIGHLIGHTS

MIDAS postdoc alumnus Arya Farahi is advancing the frontiers of trustworthy AI, decision-making and data science. Farahi directs the Data, Discovery, Decision Lab at UT Austin and co-leads the @CosmicAI Institute, an NSF Simons AI Institute based at UT Austin. While at MIDAS, Farahi carried out research on dark matter and dark energy, and developed a project with the City of Detroit to transform urban mobility systems which was featured by the World Economic Forum.

MIDAS postdoc alumna Elyse Thulin is a leading voice in the country on firearm injury prevention. Having developed her data science and AI skills during her tenure at MIDAS, Thulin has since secured multiple grants and awards, and is focusing on examining the risks that technology and online space pose on youth well-being and school safety, and developing interventions.

TRAINING FOR FACULTY AND STAFF

Data science and AI methods can enable research breakthroughs only if researchers are equipped with sufficient technical expertise. But most researchers across research fields are not trained in data science and AI. MIDAS started its first intensive summer academy in 2020, focusing on building data science and AI proficiency for biomedical faculty and staff. Since then, MIDAS has expanded the summer academies significantly. From the initial biomedical summer academy, we experimented with summer academies for social science and for environmental science. We expanded the curriculum from introductory levels to also covering advanced topics. In the summer of 2025, we offered a much more structured three-week series of AI for Scientists and Engineers. Our training effort now goes far beyond U-M. In 2025, we also offered our first AI for Scientists and Engineers short course to researchers in African universities.

When generative AI emerged as a powerful tool for research, MIDAS started the generative AI tutorial series. We modify the curricula of this series and the summer academies constantly in order to help our researchers stay updated with the newest research methodologies.

In 2025, we have also started the AI Sandbox and AI consultations, to support customized learning and active projects.

SUMMER ACADEMIES

MIDAS strengthens U-M's research community through year-round training programs, including three signature Summer Academies that equip researchers and faculty with essential data science and AI skills.

- **Biomedical Summer Academy**
- **AI for Scientists and Engineers Summer Academy**
- **Data and AI Intensive Research with Rigor and Reproducibility Academy (DAIR3)**

GENERATIVE AI TUTORIALS

MIDAS equips U-M researchers with practical, responsible, and discipline-spanning skills through our Generative AI Tutorial series. Over the past year, participants explored topics such as **coding with GenAI**, **critical evaluation of AI output**, and **AI-assisted literature discovery and synthesis**.

Looking ahead, the 2026 series will expand these foundations with hands-on sessions, including:

- **AI-Powered Literature Reviews**
- **Generative AI for Visualization**
- **Emerging Generative AI Agents**

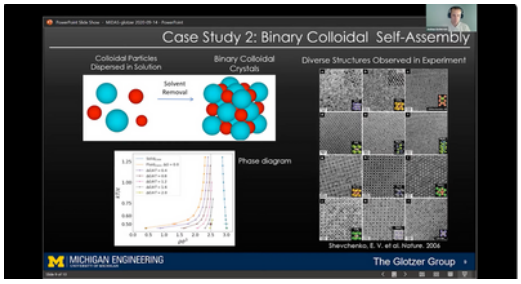
Participants engage in hands-on learning and discussion during the MIDAS Summer Academy, bringing together researchers from across disciplines to build data science and AI skills and strengthen the MIDAS research community.

RESPONSIBLE RESEARCH

The widespread and rapid adoption of data science and AI research methods comes with the danger that such methods are not used rigorously and reproducibly. MIDAS is dedicated to promoting and enabling the rigor and reproducibility in data- and AI-intensive research. In 2020, we started building a collection of technical solutions for research reproducibility and rewarding best practices. Now, we run a national training program and expand our effort into ensuring that the rapid adoption of AI is accompanied with ethical and rigorous practices.

REPRODUCIBILITY CHALLENGE

Through the Reproducibility Challenges in 2020 and 2021, MIDAS highlighted U-M researchers' exemplary work to demonstrate how complex data science and AI projects can be rigorously designed, documented, and shared to ensure reproducibility.



The image above features a case study on binary colloidal self-assembly showing how computational models and experimental data combine to predict complex crystal formation.

A NATIONAL TRAINING PROGRAM

Our initial work to promote research rigor and reproducibility led to an NIH-funded national training program: Data and AI Intensive Research with Rigor and Reproducibility (DAIR^3). This multi-university collaboration offers immersive weeklong summer bootcamps and a one-year mentoring program for biomedical science faculty and research staff across the country. This is one of the few academic research training programs entirely focused on research rigor and reproducibility. We update the program curriculum each year to include rigor and reproducibility practices with the most cutting-edge AI tools.

FUTURE LEADERS SUMMIT AND ETHICAL AI SYMPOSIA

From 2019 through 2024, the MIDAS Future Leaders Summit, often in conjunction with our Ethical AI Symposia, brought together outstanding graduate students, postdocs, and early-career faculty from across the U.S. The major theme is responsible research, including topics such as fairness, explainability, reproducibility, and the societal impact of data science and AI.

The Future Leaders Summit left a lasting legacy by supporting the community of emerging leaders in data science and AI and fostering collaborations that advance both innovation and social responsibility.

Participants collaborate around responsible data science and AI, building a national community of emerging leaders focused on fairness, transparency, and societal impact.

CROSS-SECTOR COLLABORATION

The rapid advancement of AI and data science is transforming the academia-industry research collaboration ecosystem. Cross-sector collaboration is of paramount importance now that data, AI models and even experimentation are increasing vastly in scale and becoming non-local, combined with the unprecedented speed of discovery. In the past 10 years, MIDAS has collaborated with more than 50 companies, initially focusing on industry-defined research initiatives, faculty-led projects, and talent recruitment, but evolving to joint discovery and collaborative visioning to shape the future of the research and innovation ecosystem.

Below are some examples:

Drug Discovery

In 2021, MIDAS launched the only NSF-funded Industry/University Cooperative Research Center on AI for drug discovery: the Center for Data-Driven Drug Development and Treatment Assessment (DATA). A dozen industry partners - big pharma, startups, AI tech companies - work with U-M researchers to advance precompetitive research in drug design, treatment monitoring, and pharmacovigilance using cutting-edge AI techniques.

Collaboration by Design

With funding from the National Science Foundation, our Collaborative Design of Data and AI Systems for Science and Society project built on our vast experience of industry collaboration and developed a structured model to enable collaboration across academia, industry, and government. With Microsoft and Detroit as our initial partners, we brought together AI developers, database experts, and data and AI end users to workshops and hackathons.

Ethical AI

The Rocket Companies has been a close collaborator with MIDAS ever since MIDAS was established. The company established its Ethical AI team, in part because of its interactions with MIDAS. Bernardo Modenesi, a Michigan Data Science Fellow, was sponsored by Rocket to focus on ethical AI research.

Understanding AI policy and Impact

Microsoft has been a close collaborators with MIDAS since 2019 and the focus of collaboration has always been ethical and responsible technology for social good. With support from Microsoft, U-M researchers examine AI policy, governance and societal impact, advance solutions for regulatory compliance, assess community impacts and inform emerging governance frameworks.

The Research and Development Strategic Visioning team host the inaugural workshop on the Collaborative Design of Data and AI Systems for Science and Society

DATA AND AI FOR SOCIAL GOOD

Data science and AI are changing every aspect of how the government works and makes policies, how communities support their members, and how we live our lives. Academic researchers have an important role in leveraging data science and AI methods to address significant societal challenges and ensuring their positive impact on lives and society in equitable ways. The MIDAS research community has supported many local, regional and national organizations, including the City of Detroit, the Native American tribal nations in Michigan, and the US Environmental Protection Agency. Below are some highlights.

Advancing Digital Inclusion in Detroit

In collaboration with Microsoft, Dr. Jing Liu led a team of students to better understand and address digital inequity across the city of Detroit. As Microsoft's first academic partner in its Airband Initiative for metropolitan areas, the team applied advanced statistical and AI methods to identify neighborhoods, communities and population most in need of broadband access and how assistance should be allocated.

Supporting Tribal Education

Native American Tribal Nations serve resource-restricted populations and face unique challenges for adopting technology to support policy. A team of U-M undergraduate students, led by Dr. Tayo Fabusuyi, developed a secure, customized database to consolidate previously scattered education data and support the Tribal Nations' effort to improve their student outcomes.

Improving Youth Programs in Detroit

The Detroit Police Athletic League (PAL) offers many academic, leadership, and athletic programs to support the youth in Detroit. Drs. Brady West and Paul Schulz analyzed a large volume of survey data and helped PAL assess how its programs improved achievements, life skills, community engagement, and perceptions of police among youth participants and their families.

Fair Representation in Arts and Data

Drs. Sophia Brueckner, Kerby Shedden, Jing Liu, and collaborators from the U-M Museum of Art (UMMA) used AI to examine the entire collection (24,000 pieces) at UMMA to determine how US populations were represented across 150 years of artwork, and the changes in the practice of museum collections curation over time. The project culminated in UMMA's White Cube Black Box exhibition, which translated the research into an accessible public installation featuring composite imagery, digital displays and narrative storytelling. By bringing artists, data scientists and curators together, MIDAS helped spark community dialogue about fair representation in data, algorithmic bias and the broader impact of Big Data on society.

BUILDING A BLUEPRINT FOR BETTER LLMS

Academic research is quietly shaping the next wave of trustworthy, useful, and equitable AI, right here at U-M.

In just a few years, large language models (LLMs) such as ChatGPT, Gemini, and Claude have gone from research curiosity to household names. They have the power to chat, translate, and generate content, and are becoming increasingly embedded in domains spanning science, medicine, business, and education. Tech companies are pouring billions of dollars into expanding the size and capability of LLMs, rapidly transforming these models from niche tools into essential infrastructure that is reshaping how we work, learn, and connect.

Despite this revolutionary growth, today's LLMs remain, in many respects, virtual black boxes. Beneath the surface of rapid innovation is a parallel race: U-M researchers, such as the affiliate members of the Michigan Institute for Data and AI in Society (MIDAS) featured in this story, are working to expose these models' inner workings and hidden pitfalls. Through their explorations, these researchers are getting to the heart of what LLMs actually "know" and, in the process, ensuring safer, smarter, and more responsible AI tools for real-world use.

Language, vision, and what makes us human

Understanding LLMs means understanding the roots of language development itself. "Language didn't emerge for humans until about 70,000 years ago, long after we'd already developed vision and physical navigation of the world," said Joyce Chai, professor of Computer Science of Engineering.

This linguistic origin story continues to shape how people, and the machines we create, communicate and learn. "We learn to speak after we see, move, and act first," Chai noted. "Language builds on our embodied, shared experiences in the world."

Her research aims to close the gap between words and the world, building AIs that aren't just text-only chatbots but agents that can both "see" and "say", linking images, video, and physical context to language for a more grounded, robust understanding.

For instance, in a recent COLM paper, Chai's team developed a novel framework to bootstrap the development of visual dialogue agents that can guide humans through complex tasks in physical environments. Other projects explore how robots learn tasks through trial, feedback, and spatial reasoning.

Across these projects, Chai's team is devising new architectures and training paradigms to bring language, perception, and experience closer together. From examining how well AI reasons based on embodied context to exploring whether today's models can represent spatial and physical relationships, her work is bringing us closer to AI tools that communicate, collaborate, and learn in ways that mirror human experience.

Chai points out that while LLMs are being widely deployed, they are trained based on people's perspectives of the world through text; they are not grounded to the true physical world. Next-generation multimodal models - ones that can watch and act as well as chat - will require fresh architectures and a deeper understanding of both human and machine learning.

"Human learning is incredibly data efficient, incorporating feedback, grounding, and social interaction in a rapid process," emphasized Chai. "If we could give models some of those advantages, the impact could be huge, across everything from healthcare to education."

LLMs' trust problem

Understanding is only half the story. As LLMs become more widespread, new dangers surface: misinformation, overconfidence, and hallucinations have become some of the field's most urgent challenges. Lu Wang, associate professor of Computer Science and Engineering, is tackling these risks head-on, designing techniques that make models smarter and more trustworthy.

"A lot of people have heard stories about AI chatbots making things up: false medical advice, legal citations that don't exist, or flat-out dangerous recommendations," said Wang. "This isn't just annoying. These errors can have very real consequences in domains including healthcare, justice, and education."

The root problem, Wang explains, is that LLMs absorb nearly the whole internet, which itself is riddled with errors, and learn to reproduce patterns rather than truths, sometimes with surprising confidence. Even newer, more powerful models aren't necessarily better at sticking to the facts: "In some cases, hallucinations actually get worse with stronger models, because newer training methods are designed to boost general capabilities, but not factual grounding," said Wang.

Wang's group is addressing these shortcomings by pushing for models that actively retrieve and cite evidence from reliable sources. Such methods improve accuracy and build a transparent, attributable "factual trail" for users. She and her lab are also pioneering the approach of confidence calibration, encouraging LLMs to say "I'm not certain" or provide confidence scores when unsure, making it easier for users to judge and manage risk.

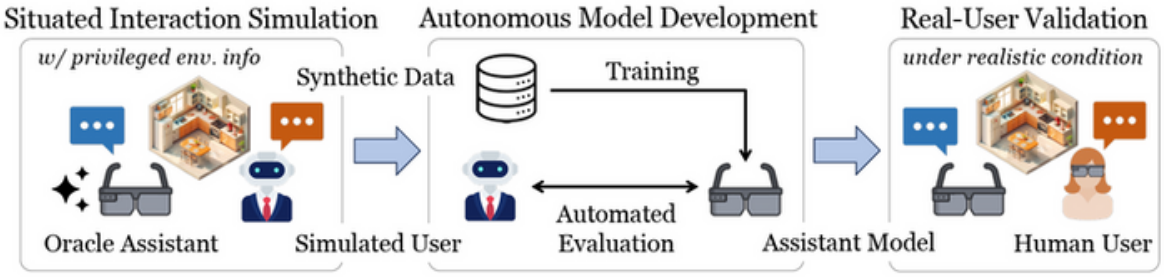


Figure from Zhang et al. (2025) - BASIS presents a three-stage pipeline for visual assistant development

Beyond technical fixes, Wang believes that academic research plays a vital role in holding AI technology accountable and ultimately ensuring it serves society's best interests: "Industry moves fast, but universities have the objectivity and independence to probe, critique, and improve these tools, identifying weaknesses and designing the guardrails that are sorely needed."

Models for social insight

Trust and transparency matter even more when language touches on personal, cultural, or ethical dimensions. David Jurgens, associate professor of Computer Science and Engineering and Information, leads research exploring the complex social and cultural factors that shape language, and how LLMs can (or fail to) reflect them.

His group develops models and datasets that capture how people from different backgrounds and cultures communicate, and how language reflects underlying values and priorities. For instance, a recent project from his group, which won Best Resource Paper Award at ACL 2025, tackled the challenge of moral and cultural alignment in LLMs. Their findings revealed that models often struggle with classic moral dilemmas—such as whether to tell a difficult truth or protect someone's feelings—especially when cultural context shapes possible answers.

"We found that today's models are nowhere near capturing the rich diversity in human responses to these questions," he said. "They're not random, but they miss a lot of what drives moral decision-making." This gap highlights both the remarkable potential and the existing shortcomings of LLMs in truly human-centered understanding.

Jurgens' team also systematically tests LLM "psychology," discovering that demographic differences drive disagreement in annotating subjective language, and highlighting ongoing gaps in model empathy, authenticity, and connection. As Jurgens put it, "The aim isn't to replace human interaction but to meaningfully support it, so people can have deeper and more effective conversations."

In another ongoing project, supported by a MIDAS pilot grant, Jurgens and his collaborators are working to improve patient-physician interactions by using LLMs to better understand how patients experience language used during clinical encounters. This work investigates how Black and White patients perceive physician comments and is part of a broader research effort supported by MIDAS.

Jurgens credits the collaborative, interdisciplinary culture at U-M for making this kind of research possible. "We're constantly reaching across computing, psychology, political science, and more to work on questions that really matter for society," he noted. "It's a huge strength of Michigan."

AI across cultures

As LLMs shape communication and decision-making around the world, Janice M. Jenkins Collegiate Professor of Computer Science and Engineering and director of the Michigan AI Lab Rada Mihalcea works to confront the critical challenge of making these technologies truly inclusive. Much of her current research focuses on cross-cultural understanding, developing methods to identify people's values, worldviews, and behaviors from language, and ensuring that AI systems serve not just English-speaking or Western users, but a diverse, global audience.

"If we want AI to be truly transformative, we need to address not only what models can do, but who they work for, how they handle diverse backgrounds, and whether they respect the values of the communities they support," Mihalcea emphasized.

> **There's so much still to discover. This isn't just about beating benchmarks—it's about building AI that actually helps.**
>
> **-Joyce Chai**

For example, her 2024 EMNLP paper analyzed how LLMs encode age-related values, demonstrating the importance of demographic-aware modeling for truly fair systems. In a 2025 NAACL study, Mihalcea's team showed how prompting models with geographic and socioeconomic context can improve performance on data from low-income or marginalized communities—favoring perspectives too often overlooked by standard AI benchmarks.

Her work in multilingual sentiment and emotion analysis offers vital benchmarks to counteract bias and avoid narrow stereotypes, advancing language models that can understand and respond to emotional cues across different regions and languages.

From building cross-lingual semantic tools to investigating multimodal sensing of human behavior, Mihalcea's work is shaping the next generation of AI to be not just smarter but also more empathetic and equitable. "AI technologies shouldn't just be powerful," she noted. "They should adapt, empathize, and reflect the richness of human diversity."

Connecting the dots: Michigan's unique contribution

Michigan's AI researchers aren't just tackling benchmarks or chasing the latest leaderboard. Their work is deeply collaborative and interdisciplinary, spanning computer science, psychology, information science, medicine, and robotics. At U-M, advancing the science of LLMs means thinking as much about meaning, fairness, and impact as about performance or scale.

By focusing on grounded, transparent, and user-centered models, researchers are charting a path beyond the current AI hype. They're showing how LLMs can be made more reliable, more inclusive, and ultimately more useful for the complex world outside the lab.

PROTECTING PEOPLE ONLINE: AI, HARMS, AND INFORMATION INTEGRITY

From non-consensual intimate images and cyberbullying to white supremacist speech and misinformation, online harms are a daily reality. With support from MIDAS, U-M researchers are using AI and data science to detect abuse, design interventions and inform policy, helping reshape how platforms, organizations and governments protect people in digital spaces.

The message arrived on a Tuesday afternoon. Someone sent her a screenshot: an intimate photo she had shared months earlier with a partner she trusted. Now it was circulating in a private Discord server with hundreds of strangers.

The college student - we'll call her Maya - felt her stomach drop. She searched frantically across platforms. The image had already been reposted. She reported it everywhere she could, but each site had different rules and processes. Some takedowns happened quickly. Others took days. New copies kept appearing.

Her phone became a source of dread. She stopped going to campus events, afraid someone would recognize her. Her grades slipped. She withdrew from friends because explaining what happened meant reliving it.

Maya's experience is far from rare. With support from the Michigan Institute for Data and AI in Society, University of Michigan researchers are using AI and data science to understand online harms, design better interventions, and inform policy, in order to reshape how platforms, schools, and governments protect people in digital spaces.

The scale of online harm

Online harassment has become a pervasive part of digital life. A 2020 Pew Research Center survey found that 41 percent of U.S. adults had experienced online harassment; 25 percent reported more severe forms such as threats or sustained abuse. Nearly half of U.S. teens report being bullied online.

These experiences are not merely unpleasant. Sustained harassment is linked to anxiety, depression, and withdrawal from public life. Journalists, activists, and researchers, especially women, often self-censor or leave platforms entirely.

Non-consensual intimate images are among the most damaging forms of abuse. Survivors describe violations that echo physical sexual assault. The images can resurface years later, affecting careers, relationships, and mental health.

"For survivors, the harm is deeply personal," said Sarita Schoenebeck, a professor at the School of Information and director of the Living Online Lab. "But it's also structural. People cannot participate safely online or offline when their image or likeness can be modified and shared online as nude or sexual."

Beyond detection

For years, platform responses followed a familiar pattern: users reported abuse, human moderators reviewed it, and content was removed - or not. Billions of posts circulate daily; no human workforce can review them all. Machine learning offered a way to scale detection. AI systems can flag hate speech, threats, and harassment, performing reasonably well for overt cases. But they struggle with context and evolving tactics.

White supremacist speech, for example, often avoids explicit slurs. Harassment campaigns may rely on coordinated but individually mild messages that evade filters. More fundamentally, detection-focused systems overlook a deeper question: what actually helps people who've been harmed?

"That's where our work begins," Schoenebeck said. "Instead of only asking how to detect harm, we ask what survivors need to be safer and regain agency."

Listening to survivors

Schoenebeck and collaborators surveyed nearly 4,000 people across 14 regions worldwide about their experiences with online harassment and the platform responses they found most helpful. The study, *Women's Perspectives on Harm and Justice after Online Harassment*, examined reactions to insults, stalking, threats, and non-consensual image sharing.

The findings were consistent. Women perceived greater harm across nearly all scenarios and strongly favored responses such as rapid content removal and permanent bans for repeat offenders. Financial compensation or temporary suspensions were seen as far less effective. The research also showed that context matters. Trust in platforms, cultural norms, and legal recourse varied across regions, meaning no single response works everywhere.

"But the principles are clear," Schoenebeck said. "Victims want agency, accountability, and responsiveness. Delayed or dismissive responses overlook and enable the problem."

These insights challenge moderation approaches that focus solely on punishment or takedowns without addressing safety, dignity, and repair.

Designing deterrence

Another strand of work in Schoenebeck's group focuses on preventing harm before it happens through design.

One project, led by PhD student Qiwei Li, developed a prototype app called Hands-Off, which requires users to make a specific hand gesture above their phone before viewing an image. The gesture makes simultaneous screenshotting nearly impossible, adding friction that deters non-consensual sharing.

Performing interlace and binoculars gestures in Hands-Off to deter non-consensual mobile screenshots.

"Technology isn't neutral," Schoenebeck said. "Design choices shape what's easy and what's hard for a user to do. Thoughtful friction can deter harmful behaviors without compromising usability."

The approach doesn't eliminate abuse, but it raises the barrier for casual violations and shifts responsibility from victims to system design.

What platforms miss

While Schoenebeck centers survivor experience, Libby Hemphill, associate professor at the School of Information and director of the Social Media Archive at ICPSR, studies how platforms detect (and fail to detect) harmful content.

In the project *What Social Media Platforms Miss About White Supremacist Speech*, Hemphill analyzed thousands of posts from forums such as Stormfront and extremist subreddits. The research showed that harmful ideology often spreads through euphemisms, humor, historical references, and claims of victimhood,

strategies that evade keyword-based moderation.

“Platforms are getting better at catching the most obvious content,” Hemphill said. “But that pushes extremists to become more sophisticated.”

The work produced carefully curated datasets that support better detection models and are shared through the Social Media Archive, allowing other researchers to build on validated, ethically managed data.

De-escalation, not just deletion

Hemphill’s research also asks whether platforms can intervene before conversations spiral into abuse.

One project tested bots that post gentle reminders when conversations show signs of escalation. The messages encourage users to pause, reflect, or reconsider tone. The approach reframes AI as a tool for de-escalation rather than punishment. Early results suggest modest but measurable reductions in hostile follow-ups. “We focus so much on deletion,” Hemphill said. “But what if platforms helped people step back before harm happens?”

Generative AI raises the stakes

Generative AI has intensified these challenges. Deepfake tools now allow anyone to create realistic non-consensual sexual images using a single photo. What once required technical expertise now takes seconds.

“Generative AI has changed the scale and accessibility of harm,” Schoenebeck said. “It forces us to rethink consent, privacy, and platform responsibility.”

In 2025, MIDAS provided another pilot grant to Schoenebeck and Li, along with co-advisor Eric Gilbert, to build web-based AI agents that act on behalf of victim-survivors to locate, report, and monitor non-consensual content across the web.

“The law is catching up slowly,” Schoenebeck said. “In the meantime, platforms need better tools; and victims need faster, clearer pathways to help.”

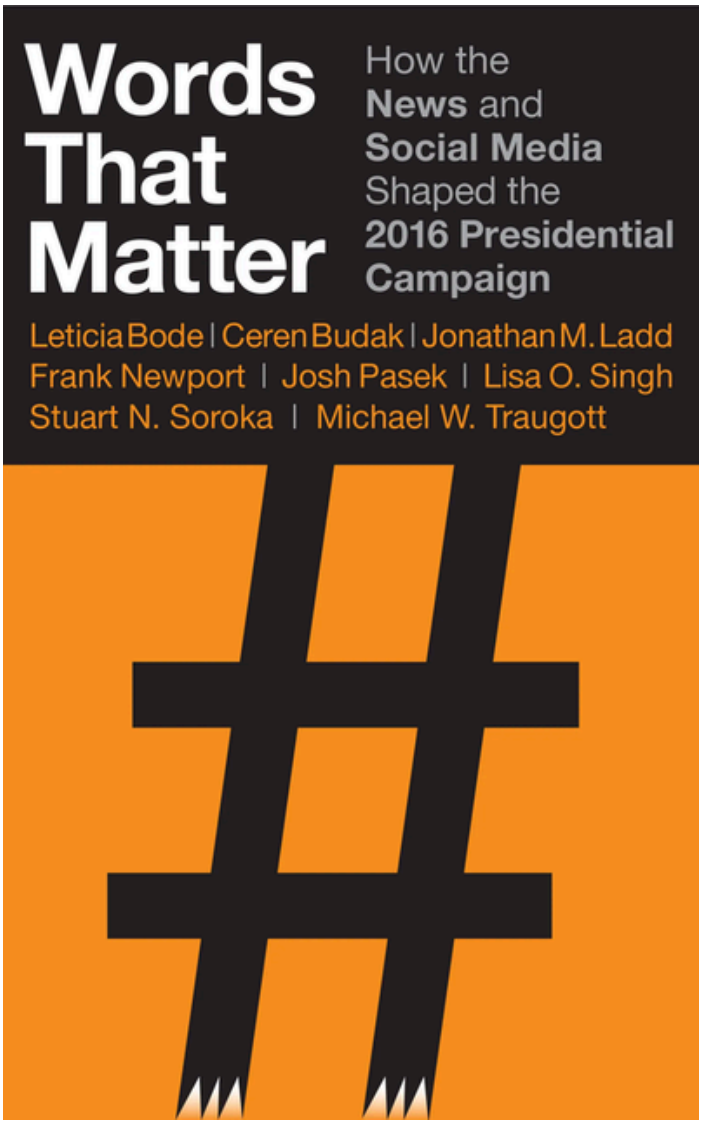
Technology and misinformation

Harassment is not the only major threat in the online world. Another threat is misinformation, and the broader swirl of political information that shapes what people believe and how they act.

With MIDAS support, Michael Traugott and colleagues have studied how political information and misinformation move through online ecosystems, and how exposure to that content connects to people’s political attitudes and perceptions. Their work treats “what spreads” as a measurable phenomenon, tracking how campaign messages travel across news and social media, who encounters them, and what that means for public opinion.

That research thread is captured in *Words That Matter: How the News and Social Media Shaped the 2016 Presidential Campaign*, a book that synthesizes evidence on how campaign information flowed through journalism, social media, and public attention, and the analysis of misinformation and “fake news” as part of that information environment.

> **Social media platforms can enable social support, political dialogue and productive collective action. But the companies behind them have civic responsibilities to combat abuse and prevent hateful users and groups from harming others. We hope these findings and recommendations help platforms fulfill these responsibilities now and in the future.**
>
> **-Libby Hemphill**



What's next

The challenges continue to evolve. New platforms emerge. Moderation policies shift. Generative tools accelerate the creation and spread of misinformation.

"The landscape shifts rapidly," Schoenebeck said. "But the core principles are durable. People deserve safety, agency, and accountability online."

Maya eventually found support. The images never disappeared entirely, but the most widely shared copies were removed. She now advocates for clearer policies and resources for others facing similar harm.

Research such as that supported by MIDAS gives substance to that advocacy. It shows that online harms and misinformation are not inevitable, and that AI, when guided by care and evidence, can be part of the solution instead of the problem.

WHEN THE WORLD CHANGED OVERNIGHT: HOW MIDAS HELPED MICHIGAN RESPOND TO COVID-19

When COVID-19 arrived in Michigan, decisions about shutdowns, reopening and resource allocation couldn't wait for perfect data. MIDAS moved quickly, funding rapid-response projects, convening modelers and public health experts, and building tools that showed where the virus was spreading and who was being left behind.

The Zoom call stretched late into the night in early April 2020. Scattered across the screen, University of Michigan administrators were weighing whether, and if so, how, to bring students back to campus. Amongst them, epidemiologists and data scientists were sharing maps and models that made the decision anything but straightforward.

One map, built by infectious disease modeler Jon Zelner and his colleagues, showed COVID-19 spreading across Michigan counties in alarming waves. Another visualization, created using anonymized WiFi data, traced how students moved through campus building (where they clustered and how long they stayed), trying to predict which spaces posed the highest risk for transmission.

The data were imperfect. Testing was scarce and uneven across regions. As case counts rose, No one knew how much of the rising case counts reflected real spread or increased testing. Hospital capacity projections also swung wildly depending on what assumptions were being used.

“We were making high-stakes decisions with incomplete information,” said H. V. Jagadish, MIDAS Director from 2019 to 2025. “The question wasn’t whether to use data; it was how to use messy, biased, rapidly changing data responsibly, and be honest about what we did and didn’t know.”

The tension between urgency and uncertainty defined Michigan’s (and the world’s) early pandemic response. What helped set the university apart was how quickly MIDAS mobilized its network of data scientists, public health researchers, clinicians, and modelers to tackle problems that could not wait for traditional grant cycles.

Between April and May 2020, MIDAS reviewed 49 proposals and funded seven interdisciplinary COVID-19 projects, launching all of them within 30 days.

Why case counts weren’t enough

In the pandemic’s early months, public attention focused on confirmed cases, hospitalizations, and deaths. But those numbers told an incomplete story.

Biostatistician Bhramar Mukherjee spelled out a “testing paradox.” When testing is limited and selective, raw case counts severely underestimate true infection rates and make trends difficult to interpret. Everyone was flying blind. Some counties reported hundreds of cases and others almost none, but that often reflected access to testing more than actual spread.

With MIDAS funding, Mukherjee developed methods to estimate underlying infection prevalence and optimize how limited tests should be allocated. Her team combined infectious disease models with survey-sampling techniques to infer who was being missed. The team was able to also predict where testing would be most informative, helping public health officials interpret positivity rates and decide where to deploy

mobile testing units. It wasn't just about the math, it was about communicating uncertainty in ways that supported real decisions.

Following students through WiFi signals

As state-level researchers tracked the virus, another team focused on the campus itself. If students returned, where would transmission risk be highest?

Quan Nguyen, Christopher Brooks, and Daniel Romero used anonymized WiFi connection data to map how students moved through campus spaces. Aggregated patterns revealed which buildings became crowded and for how long. It allowed the team to see empirically where risks were highest. not just in terms of occupancy, but also in duration and overlap between groups.

Across universities, classroom schedules were adjusted, study spaces reconfigured to reduce proximity, and high-risk building usage managed to reduce simultaneous occupancy. The insights from this and similar studies fed directly into such reopening plans.

Seeing the virus across the state

Zelner's lab zoomed out to analyze COVID-19 spread across Michigan using high-resolution spatial modeling. Beyond mapping cases, the team integrated demographic data and structural factors such as housing density, occupational exposure, and healthcare access.

Their interactive Michigan COVID-19 Tracker became a widely used public resource for journalists, health officials, and residents.

"Place matters because of what place represents," Zelner said. "Social and economic structures that shape who gets exposed, who gets sick, and who survives."

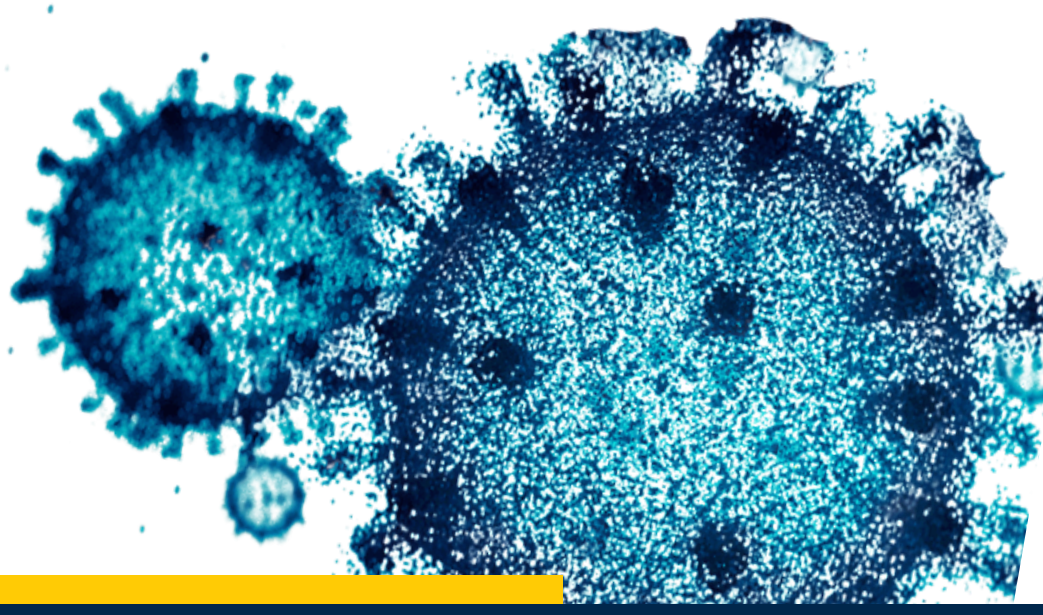
The work showed that predominantly Black neighborhoods in Detroit and Flint faced compounded risk—patterns tied to decades of policy decisions rather than chance. These insights guided targeted testing, vaccination outreach, and resource deployment.

Zelner now contributes to the CDC-funded research collaborative, the Michigan Public Health Integrated Center for Outbreak Analytics and Modeling, which extends this work to future outbreaks.

Mapping inequality

Early mortality data showed that Black Michiganders were dying from COVID-19 at disproportionately high rates, but the numbers alone did not explain why. Epidemiologist Nancy Fleischer, whose work focuses on health disparities, used MIDAS funding to build a probability sample of people who tested positive for COVID-19 and survey them about illness, care, and recovery.

The results revealed inequities at every stage of the disease experience. Black respondents reported more severe symptoms, longer hospital stays, worse treatment experiences, and less follow-up care—even after accounting for underlying health conditions and socioeconomic factors. Disparities also extended to mental health impacts and what would later be recognized as long COVID.



"We knew from mortality data that Black communities were bearing a disproportionate burden," Fleischer said. "But our new research showed how deeply those inequities were embedded."

The findings informed resource allocation decisions and broadened conversations about structural racism in healthcare.

Inside the hospital

At Michigan Medicine, clinicians faced a different challenge: predicting which hospitalized patients would deteriorate.

A MIDAS-funded team led by Andrew Admon and Christopher Gillies developed real-time models using electronic health record data to track patient trajectories. The approach combined transfer learning with ordinal regression to monitor risk hour by hour.

"COVID overwhelmed hospitals in ways we weren't prepared for," Admon said. "Even imperfect early warnings helped us be proactive rather than constantly reacting."

The tools supported staffing and ICU planning and informed later efforts to design clinical decision systems that account for equity and population differences.

The 30-day turnaround

What made this response possible was not just expertise, but institutional agility.

Traditional grant cycles take months. MIDAS compressed the entire process, from proposal submission to review and to funding, into 30 days. "We accelerated everything," Jing Liu, MIDAS Executive Director, said. "Researchers and reviewers understood the urgency and the potential impact. While everyone was trying to also pivot how to teach and do research and adjust to a new way of life, they helped MIDAS make this rapid funding happen."

MIDAS also convened cross-campus working groups that met throughout 2020, connecting public health researchers, clinicians, data scientists, policy advisors, and administrators.

The legacy

The pandemic emergency has passed, but its infrastructure remains. Data pipelines linking clinical, mobility, and demographic data persist, alongside stronger cross-campus relationships and a shared understanding of what crisis-ready analytics require: speed, transparency, equity, and connection to decisions.

Methods developed during COVID-19 now inform broader work: Mukherjee's approaches support ongoing surveillance, Zelner's modeling extends to future outbreak analytics, and Fleischer's research continues through long-COVID studies.

> **"The pandemic was a stress test. What worked didn't happen by accident. It worked because we'd invested for years in relationships, trust, and flexible infrastructure.**
>
> **-Jing Liu**

TRUSTWORTHY BY DESIGN: HOW MIDAS IS SHAPING RESPONSIBLE AI

AI is now embedded in everything from sepsis prediction and credit scoring to public policy and online platforms. MIDAS is building the foundations for responsible AI and enabling responsible practice through research, training, policy partnerships and human-centered AI.

The Epic Sepsis Model was supposed to be a breakthrough. Embedded in electronic health record systems at hundreds of hospitals, the AI tool promised to detect sepsis, a condition that kills more than 250,000 Americans each year, early. This is an urgent task for health care staff at hospitals where every hour of delayed treatment raises mortality risk. Epic Systems reported accuracy rates between 76 and 83 percent. Hospitals adopted the model widely, trusting that real-time alerts would help clinicians intervene sooner.

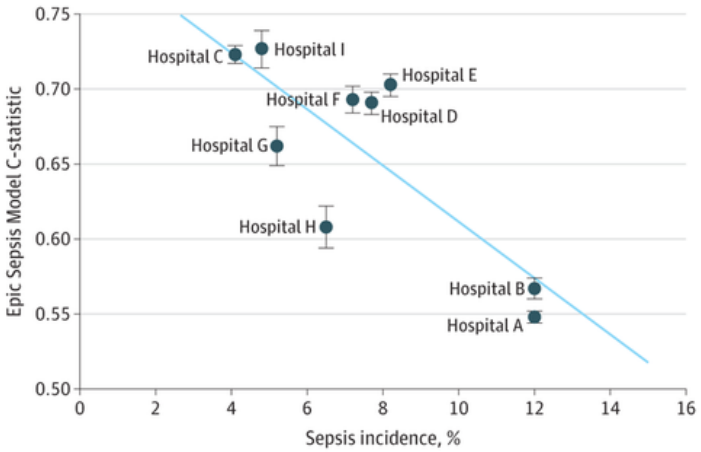


Figure from Patrick G. Lyons et. al (2023) - Association Between Hospital-Level Sepsis Incidence and Epic Sepsis Model C-Statistic Across 9 US Hospitals in a Network

Each hospital is represented by a blue point (A through I), with 95% CIs represented by vertical bars. The diagonal line represents the line of best fit among hospitals A through I.

Then researchers at Michigan Medicine decided to validate it independently.

In a 2021 study of nearly 40,000 hospitalizations, the model missed 67% of sepsis cases. It flagged only 7 percent of patients that clinicians had not already identified as high risk, while generating alerts for nearly one in five hospitalized patients, which overwhelmed staff with alarm fatigue. A follow-up study spanning more than 800,000 patient encounters across multiple hospitals showed wildly uneven performance, with the model working the worst at hospitals that serve sicker, more complex patients.

“Such wake-up calls have been sounding for many years,” said Jing Liu, Executive Director at MIDAS. “And with researchers rushing into data- and AI-intensive research, the issue of research rigor, reproducibility and the trustworthiness of science is never more important.”

“The reproducibility crisis”, the widespread phenomenon that research studies fail to be reproduced or validated which damages the trustworthiness of science, is something MIDAS has been trying to address for years through the development of training programs and supporting research to improve research.

Why trustworthiness matters

AI now shapes decisions across nearly every consequential domain: healthcare, hiring, credit, education, criminal justice, public policy, and scientific discovery itself. A single flaw in a widely deployed model can do serious harm to people and erode trust in institutions that rely on automated systems. In academic research, the rush to publish often leads to hastily done research without proper validation or the proper sharing of data and models. Such issues are exacerbated as increasingly more researchers rely on data that they did not collect themselves and models that they did not train themselves.

“We’re in a moment where AI and data science capabilities are advancing incredibly fast,” said Liu, “but our ability to rigorously use such capabilities hasn’t kept pace. That gap is dangerous.”

The knowing–doing gap

By the late 2010s, there was broad consensus about what responsible data science should involve: fairness, transparency, accountability, robustness, privacy. Professional societies published principles. Companies released ethics statements. Policymakers issued guidelines.

But principles alone can't change practice.

Many research teams lacked concrete tools and workflows to implement best practices consistently. Ethical reviews, when they happened, often came after data were collected, models trained, and decisions locked in. Small changes in preprocessing or modeling choices could produce very different results, yet were often not documented.

"Saying you care about fairness and actually building fair systems are very different things," said H. V. Jagadish, MIDAS Director from 2019 to 2025. "You need methods, tools, training, and institutional structures that make doing the right thing the default, not a heroic extra effort."

MIDAS set out to close that gap by treating responsible AI as culture.

Building the culture

Responsible research is one of MIDAS's core focus areas, running through all of the institute's work rather than sitting on the margins and spanning reproducibility, data equity, ethics, and human-centered design.

In 2020, MIDAS built a Reproducibility Hub, an online repository of tools, templates, and best practices contributed by researchers across the university, including those who won the MIDAS Reproducibility Challenge, to highlight exemplary work and provide models others can follow.

"Reproducibility feels abstract and, honestly, boring, until you're trying to understand why a colleague's analysis doesn't match yours," Liu said. "The Hub gives people concrete solutions instead of abstract guidance."

With such work as the foundation, MIDAS started DAIR³: Data and AI Intensive Research with Rigor and Reproducibility, a national NIH-funded multi-university collaboration. DAIR³ combines intensive bootcamps with year-long follow-up mentoring that help faculty, staff, and advanced trainees nationwide incorporate new skills and mentality into their research and teaching. Participants learn how to document data provenance, design reproducible workflows, develop robust machine learning models and statistical analysis, and rigorously compare results from multiple studies. The goal is not awareness but habit change.

"We're trying to redefine what normal practice looks like," Jagadish said. "If you train enough people and give them tools that make rigor easier, it becomes the standard rather than the exception."

Data equity and FIDES

Trustworthy AI also depends on equitable data practices. Jagadish led the Framework for Integrative Data Equity Systems (FIDES), a $2.5 million NSF-funded, multi-institution effort.

FIDES focused on identifying inequities in data collection, modeling, and deployment to design systems that actively counter those inequities rather than amplifying them. The work spanned domains including health, mobility, and public services.

"Data equity isn't just about representation," Jagadish said. "It's about understanding how data and algorithms shape power, to build systems that work against existing imbalances."

From research to practice

MIDAS's commitment to trustworthy AI extends beyond principles into practice through strategic partnerships and targeted research investment. A cornerstone of this work is its collaboration with Microsoft, launched in 2024 to advance research and policy solutions for responsible AI. Research projects funded by Microsoft's Office of Responsible AI span both technical and socio-policy dimensions of responsible AI. Researchers carry out work on human–AI system design, building equitable predictive modeling, developing community-inclusive innovation frameworks, enabling appropriate reliance on generative AI, creating interventions to address online harms such as non-consensual intimate media, and building data-driven tools to inform public decision-making in areas like infrastructure and climate policy.

"Together, these projects form a pipeline from research to real-world impact by guiding responsible system design, improving governance, and strengthening society's capacity to adopt AI thoughtfully." Said Liu.

By connecting academic research to policy and practice, the MIDAS–Microsoft collaboration demonstrates how university–industry partnerships can help ensure that AI development is grounded not only in technical excellence, but also in accountability, equity, and public trust.

Human-centered AI

Even statistically fair and robust systems can fail if people don't understand when to trust them.

Nikola Banovic, MIDAS Associate Director and Associate Professor of Computer Science and Engineering, designs explanation mechanisms that help users build accurate mental models of AI systems to understand both their strengths and limits. Vera Liao, Associate Professor of Computer Science and Engineering, studies AI transparency from an Human-computer Interaction perspective, asking how explanations can be designed to support real decision-making rather than overwhelm users with technical detail.

Human-centered AI work connects directly to responsible AI goals. A system might be statistically fair and technically robust, but if people can't understand how it works or don't trust it, it won't be used appropriately. Conversely, a system that's presented with misleading explanations or overconfident predictions can lead people to trust it in situations where they shouldn't.

Together, their work underscores a core MIDAS principle: trust is relational, not just statistical.

A different kind of breakthrough

MIDAS is betting on a different narrative about AI progress. Not just models that are faster or more accurate, but also systems that are fairer, more transparent, and more robust.

The Epic Sepsis Model story could have ended badly. Instead, independent validation exposed its limits, sparked changes in how hospitals evaluate clinical AI, and reinforced the need for rigorous testing across settings.

That outcome wasn't inevitable. It happened because Michigan researchers had the tools, training, and institutional support to ask hard questions, and because that expectation was part of the culture.

> **"In a world where AI shapes who gets healthcare, education, and opportunity," Liu said, "trustworthiness isn't optional. It's foundational.**
>
> **-Jing Liu**

DATA FOR DEMOCRACY: BUILDING THE SOCIAL SCIENCE INFRASTRUCTURE BEHIND BETTER POLICY

From digitized GI Bill records that reveal who was (and who wasn't) helped to buy a home after World War II, to social media archives that help researchers study harassment, misinformation and political discourse, U-M is a national powerhouse for social science data. With support from MIDAS, our researchers are building the infrastructure, standards and training that turn messy data into tools for advancing democracy and evidence-based policy.

The index card is yellowed with age, its typed entries faded but still legible. It records a mortgage guarantee issued by the Veterans Administration in 1947, including the loan number, property location, veteran's name and amount borrowed. One card among thousands filling boxes at the National Archives, it is a remnant of the postwar housing boom when millions of returning World War II veterans bought homes with federally backed loans.

For decades, those cards sat underutilized. Historians could not analyze them systematically without investing enormous labor. The records were too voluminous to transcribe by hand, too inconsistent in format for simple scanning and too inaccessible for most researchers.

That changed when a team of University of Michigan researchers received MIDAS funding for a project called "Images to Integrated Data." They developed machine learning tools to digitize and parse those index cards, converting tens of thousands mortgage guarantees from 1946 to 1954 into clean, analyzable data.

What emerged was the first public dataset documenting individual GI Bill mortgage guarantees, and it told a story that supported long-standing anecdotal accounts from veterans. The data showed exactly where loans went, who received them and who did not. When coded onto maps, the pattern became stark. Black veterans were systematically excluded from federally backed mortgages in many regions, perpetuating housing segregation that was widespread at the time.

"These records sat in boxes for decades," said J. Trent Alexander, a researcher at the Inter-university Consortium for Political and Social Research (ICPSR) who led the digitization project. "Now that they're digitized and linkable to other data, policymakers and communities have the potential to explore the role federal programs can play in creating or stifling economic opportunity and equality."

That transformation from isolated index cards to usable evidence illustrates the value of social science data infrastructure. It is not glamorous work, but without it, many critical social science questions cannot be answered.

Why data infrastructure matters

Good policy requires good data. That is true whether the question is how to fund education, reform veterans' benefits, protect voting rights, address housing discrimination or counter online extremism.

But useful data does not simply exist. Someone has to collect them, clean them, document what they mean, secure them against misuse and make them accessible to researchers who can analyze them rigorously. Such work often happens sporadically, driven by individual research projects or agency mandates. Datasets lived in scattered archives with inconsistent documentation and high barriers to access. Researchers often spent years just getting permission to use data, then more years cleaning and organizing them before any analysis could begin.

The result was enormous inefficiency and duplicate effort. Studies were difficult to reproduce because documentation was incomplete. Important research questions went unasked because the costs of assembling the necessary data were too high.

“Social science has always been data-intensive, but the infrastructure for managing that data lagged far behind what researchers actually needed,” said Margaret Levenstein, director of ICPSR. “We had surveys and administrative records, but we didn’t have the systems to make them findable, accessible, interoperable and reusable, what we now call FAIR principles.”

The challenge has grown more complex as new data sources have emerged. Researchers now also work with social media posts, mobile sensor data, digitized historical documents and other forms of digital trace information that reveal how people behave, not just what they report in surveys. These sources create powerful opportunities to understand social phenomena at unprecedented scale and resolution. They also raise difficult questions about privacy, consent, potential misuse and who owns such data.

Building infrastructure that addresses those challenges responsibly, protecting privacy while enabling research, documenting provenance and limitations and establishing ethical guidelines, requires sustained institutional investment.

The Michigan advantage

The University of Michigan has been a leader in social science data for decades, with ICPSR being an international leader that has archived and distributed social science data since 1962. ICPSR data catalog contains studies that are associated with more than 80,000 datasets covering topics from election studies and health surveys to international development indicators, and it provides data to researchers at 800 member institutions worldwide.

Even with its long history and deep expertise, ICPSR needed to modernize for an era of big data, digital trace information and AI-powered analysis methods.

That is where MIDAS support came in. “We recognized that traditional surveys were being strained by societal changes, including declining response rates, higher costs and limited ability to capture fast-moving phenomena like social media dynamics,” said Dr. H. V. Jagadish, MIDAS director from 2019 to 2025. “At the same time, new data sources were emerging that could complement or even replace surveys for some purposes. Using those sources responsibly required new infrastructure, new methods and new norms.”

With MIDAS as a key collaborator, ICPSR secured $38M of funding from the National Science Foundation for its Research Data Ecosystem, to modernize how social science data are curated, discovered, accessed and analyzed.

The Research Data Ecosystem includes tools such as Explore Data, which allows researchers to browse thousands of datasets to identify relevant variables; Researcher Passport, which streamlines secure access to restricted-use data; and TurboCurator, which uses AI to suggest standardized metadata and keywords that make datasets easier to find and use.

> **“Now that they’re digitized and linkable to other data, policymakers and communities have the potential to explore the role federal programs can play in creating or stifling economic opportunity and equality.**
>
> **-J. Trent Alexander**

“These aren’t just technical improvements,” Levenstein said. “They fundamentally change what’s possible. When researchers can quickly find and access the data they need, when datasets are well documented and follow common standards and when the infrastructure handles security and privacy protections automatically, researchers can focus on the questions that matter rather than logistics.”

Archiving social media

Another major infrastructure effort emerged as social media data became essential for understanding public discourse, misinformation, political mobilization and online harms. Most researchers, however, lacked the technical capacity or resources to collect and manage that data themselves.

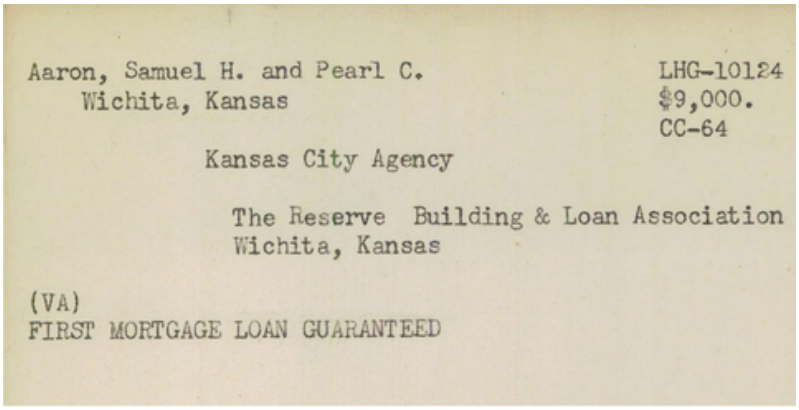
Libby Hemphill, a professor at the School of Information, received a MIDAS pilot grant in 2021 to develop standards and infrastructure for ethical social media research. The project, "Ensuring FAIRness in Social Media Archives," examined how to build archives that provide reliable, equitable access to social media data while protecting privacy and preventing misuse.

That work led to a partnership with Meta and ICPSR to launch SOMAR, the Social Media Archive. With $1.3 million in initial support, SOMAR is building a centralized repository for curated social media research datasets, complete with documentation, ethical guidelines and access controls.

The archive includes datasets such as #MeToo tweet IDs, congressional messaging collections and a white supremacist speech corpus combining Stormfront posts and Reddit comments.

These resources support research on content moderation, platform policies, political communication and the spread of extremist ideology.

"Social media platforms generate enormous amounts of data about public discourse, but that data is hard to access, unstable over time and often disappears when platforms change their APIs or policies," Hemphill said. "SOMAR is working to create a stable archive of these data, democratizing access to powerful but sensitive digital trace data."


Aaron, Samuel H. and Pearl C.
Wichita, Kansas

LHG-10124
$9,000.
CC-64

Kansas City Agency

The Reserve Building & Loan Association
Wichita, Kansas

(VA)
FIRST MORTGAGE LOAN GUARANTEED

One of 25,744 index card records from the administration of G.I. Bill Mortgages from 1946 to 1954, housed at the National Archives in College Park, Maryland.

FROM CRASH DATA TO SAFER STREETS: HOW DATA SCIENCE IS REDEFINING MOBILITY

Every day, hundreds of thousands of people move through Michigan's roads, buses and intersections. MIDAS supports researchers to turn vast crash and mobility datasets into tools that help prevent crashes, redesign dangerous corridors and rethink public transit.

The arterial road runs straight through suburban Detroit, six lanes wide, lined with strip malls and bus stops. For years, it appeared on Michigan crash maps as a dense red band—rear-end collisions at stoplights, left-turn crashes at busy intersections, pedestrians struck while crossing. For years, serious-injury crashes were recorded along this stretch. The numbers showed that crashes were happening, but not why, or how to make the road safer.

Researchers at the University of Michigan Transportation Research Institute (UMTRI) approached the problem from a data science perspective. They combined crash reports with traffic volumes, weather data, road geometry, and signal timing. Patterns emerged: intersections where sun glare blinded drivers at rush hour, left turns without protection, crosswalks that forced pedestrians to navigate multiple conflict points. “We used to wait until locations accumulated enough crashes to justify action,” said Carol Flannagan, who leads UMTRI’s Data Science Group. “Now we can identify risk factors earlier and test interventions with data.”

The research insights, along with other work and policy advice from researchers at UMTRI, helped guide safety technologies, driver safety programs and driving laws, which led to a notable decline in crashes in recent years.

The stakes

In 2023, Michigan recorded nearly 288,000 crashes, killing 1,095 people and injuring more than 71,000. Pedestrians and bicyclists accounted for 207 deaths. Nationally, traffic fatalities remain near 41,000 per year.

These harms are not evenly distributed. Crashes cluster along wide, fast arterials that often cut through low-income neighborhoods and communities of color. These are areas with fewer safe crossings and limited transit alternatives.

“Where you live strongly affects your risk of being injured or killed.” Flannagan said. “Data helps us see those patterns and invest more fairly.”

Why traditional approaches weren’t enough

For decades, safety analysis focused on high-crash locations identified from historical records. While effective, that approach was slow and reactive. Data lived in silos: crash reports, traffic counts, vehicle data, and medical outcomes rarely connected easily.

At the same time, transportation was changing. Ride-sharing altered travel patterns. Delivery vehicles flooded residential streets. Advanced driver assistance systems (ADAS) spread rapidly across vehicle fleets.

“We had rich datasets but limited ability to link them and analyze them at scale,” Flannagan said. “Agencies needed answers faster than traditional methods could deliver.”

MIDAS funds data infrastructure

In 2016, MIDAS launched a Data-Intensive Transportation Research Initiative to address those limitations. Two projects anchored the effort.

The first, led by Flannagan, focused on creating a transportation data ecosystem, integrating decades of crash records, roadway characteristics, vehicle specifications, and medical outcomes into a usable system. The second, Reinventing Public Urban Transportation and Mobility (RITMO), explored how to redesign public transit using machine learning and optimization.

“MIDAS was created to support high-risk, high-potential ideas,” said Alfred Hero, MIDAS’s inaugural co-director. “At the time, integrating transportation data at this scale, or rethinking transit algorithms, seemed ambitious. That’s exactly why we invested.”

The resulting infrastructure underpins **Michigan Traffic Crash Facts**, a public-facing platform that provides accessible crash statistics, maps, and analyses for journalists, planners, advocates, and policymakers. “Making data usable matters,” Flannagan said. “Michigan Traffic Crash Facts opens information that once required specialized expertise.”

Evidence that saves lives

One of UMTRI’s most influential contributions came from evaluating advanced driver assistance systems using real-world data.

In a study funded by the National Highway Traffic Safety Administration, researchers linked optional safety features on about 1.2 million General Motors vehicles to police-reported crashes across multiple states.

The results were striking. Vehicles equipped with front automatic braking and forward collision alerts experienced roughly 45 percent fewer relevant crashes. Forward collision alerts alone reduced crashes by about 16 percent. Other systems showed benefits that varied by design.

“Policy decisions need evidence from the real world,” Flannagan said. “This study showed which technologies actually reduce crashes.”

> **We're at a point where almost every transportation decision involves data and algorithms in some way. The challenge is making sure those tools serve everyone...**
>
> **-Al Hero**

UMTRI is extending this work using connected-vehicle data from Ann Arbor testbeds, where vehicles communicate with roadside infrastructure to warn drivers about hazards such as red-light violations or sudden braking ahead.

Rethinking transit systems

Data science is also reshaping how communities think about mobility, especially in places where transportation systems have long fallen short of meeting residents’ needs. With MIDAS support, engineers led by Jerome Lynch have been building data-driven approaches to help Michigan communities design safer, more connected, and more equitable transportation networks.

One line of work focused on smart infrastructure, using sensor networks and high-resolution data to understand how roads, bridges, and vehicles interact in real time. Lynch’s team has developed methods to fuse information from connected vehicles, roadway sensors, and environmental monitoring systems, allowing engineers to identify crash-prone areas, detect structural vulnerabilities, and model traffic flow more accurately than traditional tools. These insights help transportation agencies prioritize safety investments, mitigate congestion, and prepare for emerging mobility technologies such as vehicle automation.

Another thread centered on community-driven mobility planning. In Benton Harbor, for example, Lynch's group partnered with local leaders to diagnose transportation barriers faced by residents that resulted in limited access to jobs, healthcare, and grocery stores. By combining mobility data with extensive community engagement, the team helped the city envision practical solutions ranging from optimized bus routes to shared-mobility services. The project became a model for how data can support mobility justice in small and mid-sized cities.

Together, these efforts show what rethinking a transportation system can look like: embedding sensing, connectivity, and analytics into the built environment; partnering with communities to address real mobility needs; and designing infrastructure that is ready for the next generation of mobility technologies.

The gig economy and transportation

MIDAS-supported research also examined how digital platforms reshape transportation labor and access.

Qiaozhu Mei and Yan Chen, professors at the School of Information, analyzed large datasets from Didi Chuxing, a major ride-sharing platform, to study how matching algorithms and incentive structures affect driver behavior and earnings. His work and strategies to build incentive structures showed that team-based incentives improved driver retention and income compared with individual bonuses, while algorithmic design influenced service availability in low-density areas.

"These platforms make millions of decisions daily," Mei said. "Data science can design better systems. But we have to ask, better for whom? Improving gig workers' job satisfaction should be part of the design."

New risks in automated mobility

As vehicles become more automated, MIDAS researchers have examined emerging vulnerabilities.

Atul Prakash, a professor of computer science, demonstrated that subtle physical alterations to road signs, such as carefully placed stickers, could mislead machine-learning systems used in autonomous vehicles.

"These systems are sophisticated, but brittle in ways human drivers aren't," Prakash said. "That creates new safety risks as autonomy expands."

The findings have informed broader efforts to improve robustness and testing standards for safety-critical AI.

Who benefits and what comes next

The impacts of such research extend across Michigan and beyond. Cities use crash analytics to justify protected bike lanes and intersection redesigns. Regulators gain clarity about which safety technologies save lives. Transit agencies experiment with more flexible, equitable service models.

As transportation systems evolve, with electric vehicles, connected infrastructure, and algorithm-mediated services, data science will shape nearly every decision.

"Mobility isn't just an engineering problem," Hero said. "It's a data and policy problem. MIDAS brings those perspectives together."

The suburban arterial is safer now than it was five years ago. Fewer crashes. Fewer injuries. Fewer families facing preventable loss. That is what happens when data moves from spreadsheets into action, and when institutions invest in the infrastructure to make safer streets possible.

FROM CELLS TO SYSTEMS: HOW SINGLE-CELL DATA AT U-M IS REWRITING HUMAN BIOLOGY

From treating tissues as black boxes to mapping biology cell by cell, U-M researchers are using single-cell and spatial genomics to transform reproductive health research. With early support from MIDAS, this work has built campus-wide infrastructure, attracted major external funding, and produced landmark cellular atlases revealing how individual cells interact, fail, and regenerate, and opening new paths for understanding infertility, disease, and clinical care.

For most of modern biology, tissues were treated as black boxes. Scientists ground up millions of cells, averaged their molecular signals, and inferred what was happening inside. The approach revealed broad patterns; but it hid rare cell types, erased spatial context, and blurred the earliest signs of disease.

Single-cell and spatial genomics have changed that. By measuring gene activity in individual cells and mapping those signals back to precise locations in tissue, researchers can now watch how individual cells interact, specialize, and fail.

“Single-cell data let us stop treating a tissue as a black box,” said Sue Hammoud, associate professor of human genetics, obstetrics and gynecology, and urology at Michigan Medicine. “We can ask which exact cells are present, how they talk to each other, and which ones fail first when something goes wrong.”

At the University of Michigan, that shift has transformed reproductive biology. A $1.25 million seed investment from the Michigan Institute for Data and AI in Society in 2017 catalyzed a campus-wide single-cell program that has since attracted more than $30 million in external funding and positioned Michigan as a global leader in reproductive atlases.

The work addresses an urgent health challenge. Nearly one in eight U.S. couples experiences infertility, and about half of those cases involve abnormalities in eggs or sperm. Cancer treatments such as chemotherapy can damage reproductive organs, leaving young survivors with limited options to preserve fertility or restore hormone production.

Building infrastructure and a community

The turning point came with a MIDAS Challenge Award awarded to a cross-campus team determined to make single-cell analysis a core capability at Michigan. That funding launched the Michigan Center for Single-Cell Genomic Data Analytics, co-led by geneticist Jun Li and collaborators from the Medical School, College of Engineering, LSA, and School of Public Health.

The center built computational infrastructure and analytical pipelines to handle the massive, sparse datasets generated by single-cell experiments. Just as importantly, it created a community by hosting retreats and symposia that brought biologists, statisticians, and computer scientists into sustained collaboration.

“These experiments are expensive and technically demanding,” said Brian Athey, MIDAS inaugural co-director. “The Challenge Award gave researchers room to take risks and a way to share tools instead of every lab starting from scratch.”

The initial MIDAS funding helped unlock $7 million in internal support through the U-M Biosciences Initiative. The effort has since grown into today's Single-Cell Spatial Analysis Program, which supports dozens of labs across campus.

"The Challenge Award gave researchers room to tackle the hardest data science problems," said Li, now Chair of the Department of Molecular Genetics and Genome Sciences at the University of Oklahoma Health Sciences Center. "And it knitted together a community so breakthroughs could benefit many labs."

Mapping the human ovary

One of the program's most visible achievements is a new cellular and spatial atlas of the human ovary, led by Ariella Shikanov, Li, and Hammoud. Working with an organ procurement organization, the team combined single-cell RNA sequencing with spatial transcriptomics to analyze ovaries from five healthy young donors.

They mapped gene activity across key regions, especially around follicles which are the hormone-producing structures that contain immature eggs. The atlas, published in Science Advances and named one of the top science breakthroughs of 2024, revealed dozens of genes uniquely active in oocytes and their supporting cells.

> **"Most human follicles never make it to a mature egg. By seeing exactly what a healthy follicle looks like in its tissue context, we can start to understand why some succeed and so many fail.**
>
> **-Ariella Shikanov**

The work has immediate clinical relevance. Surgeons can reimplant frozen ovarian tissue to restore hormones and fertility for some cancer survivors, but only a small fraction of follicles survive. The atlas provides a roadmap for designing engineered ovarian tissues that could sustain hormone production and egg maturation for much longer.

Understanding the uterus

The team has also applied single-cell tools to the uterus, an organ that undergoes dramatic remodeling throughout life.

In a recent study, Hammoud, Li, and collaborators profiled more than 50,000 cells from healthy premenopausal uterine tissue and integrated those data with additional datasets. They identified previously unrecognized cell types and progenitor populations in both the endometrium (the lining that supports embryo implantation) and the muscular layer that gives the uterus its strength.

"The uterus is constantly rebuilding itself," Hammoud said. "Single-cell data let us see which cells drive regeneration, which signals make the lining receptive to an embryo, and which changes might predispose someone to implantation failure or disease."

This work feeds into a broader Human Cell Atlas of the female reproductive system, supported by the Chan Zuckerberg Initiative and led by Hammoud, Li, reproductive endocrinologist Erica Marsh, and Shikanov. The project is generating open atlases of the ovaries, fallopian tubes, and uterus for researchers worldwide.

The algorithms behind the atlases

Behind every atlas is a suite of computational tools that translate noisy, high-dimensional sequencing data into biological insight. Xiang Zhou, professor of biostatistics, is one of the methodology researchers on Li's team, together with several other methodologists, to develop widely used approaches suitable for the analysis of sparse signals for single-cell and spatial analysis.

"Our job is to turn millions of sequencing reads into something biologists and clinicians can reason about," Zhou said. "We design methods that respect the structure and uncertainty of the data so the biological story becomes clearer."

Although developed in the context of reproductive biology, these tools are now used to study tumors, immune responses, and brain tissue.

Connecting environment and disease

Single-cell atlases also support research linking environmental exposures to long-term health risks. Justin Colacino, associate professor of environmental health sciences and nutritional sciences, uses single-cell and spatial omics to study how pollutants affect stem and progenitor cells in tissues such as breast and brain.

“Two people can have similar exposure to a chemical, yet only one develops disease,” Colacino said. “Single-cell approaches show which specific cell populations are most vulnerable and how their gene programs change.”

Colacino, a leadership member of the Single-Cell Spatial Analysis Program and a 2024 Rogel Scholar, integrates tissue atlases with epidemiologic cohorts and molecular data to identify more precise targets for prevention and regulation.

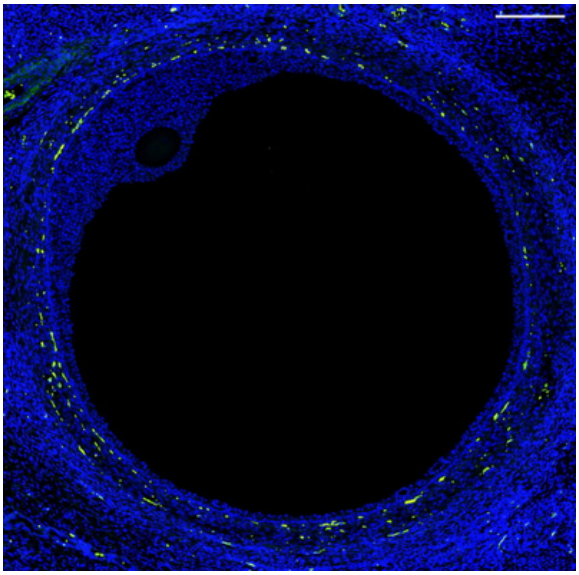
Clinical impact and what’s next

The impact of this work is already visible. Reproductive endocrinologists use atlas data to refine their understanding of uterine receptivity. Oncology teams explore engineered ovarian tissues to extend fertility windows for childhood cancer survivors. Basic scientists probe how germ cells form and fail in ways that bulk tissue studies never revealed.

The atlases are also pushing the frontier of AI in biology. Their massive, multimodal datasets are ideal testbeds for machine learning models that simulate cellular trajectories, annotate cell types, and predict how tissues respond to perturbations.

“I’m very excited about what the next decade will look like,” Li said. “As AI tools mature, we’ll move from describing what cells do to predicting how tissues behave if we tweak one gene or signal.”

In addition to conducting their own research, the MIDAS-funded single-cell research team has also provided consultation to several dozens of research groups on campus to help them develop single-cell research projects. Such generous effort has greatly helped Michigan to become a leader for single-cell science, one where data scientists and biologists work side by side to reveal how individual cells build healthy tissues, and what happens when that process goes off course.



A fluorescent image of a human ovarian follicle U-M researchers collected during spatial analysis—clearly showing all the different compartments including the oocyte (the small oval), surrounding hormone-producing cells, blood vessels, immune cells and compartments. The scale bar is 0.2mm. Credit: University of Michigan.

ALGORITHMS IN THE CLINIC AND THE LAB: CLINICAL AI AND A NEW MODEL FOR DRUG DISCOVERY

Over the past decade, MIDAS has helped turn AI from a buzzword into a practical tool in medicine, from models that track sepsis risk in the emergency department to a national center that uses AI to design and test drugs more efficiently.

In a busy emergency department, monitors stream a familiar rhythm of numbers: heart rate, blood pressure, oxygen saturation. On a nearby tablet, another signal catches a resident physician's eye, a rising curve that suggests a patient with vague abdominal pain and a low-grade fever may be heading toward sepsis.

It is not a diagnosis. But it is enough to prompt a closer look, additional tests, or earlier antibiotics if the clinical picture supports it. In emergency medicine, where minutes matter, that early warning can make the difference between a short hospital stay and weeks in intensive care.

"The hard part about sepsis is that early on, it can look like a lot of other things," said Kayvan Najarian, professor of Computational Medicine and Bioinformatics, Emergency Medicine, and Electrical Engineering and Computer Science. "The question is whether we can see deterioration before it becomes obvious."

Najarian's lab has spent years developing AI models that do exactly that. Using continuous streams of vital signs and waveforms, the models track how a patient's risk evolves minute by minute. They rely on tensor-based mathematics capable of integrating multiple time series, such as heart rhythms, breathing patterns, and blood pressure fluctuations, into a dynamic picture of clinical risk.

As the models matured, Najarian and his collaborators recognized something important: the same mathematical tools that make sense of noisy, complex physiological data could also be applied far beyond the clinic. The methods developed to track a patient's shifting trajectory could also be used to untangle biological complexity at other scales: from gene interactions to molecular behavior.

That insight opened a second frontier: drug discovery.

A national hub for data-driven drug development

The Center for Data-Driven Drug Development and Treatment Assessment (DATA), was established in 2021 as a National Science Foundation Industry/University Cooperative Research Center (I/UCRC). DATA is the only I/UCRC focused on using AI to make drug development faster, cheaper, and more reliable; and it complements other drug discovery research on campus and strengthens U-M's leadership role in this area.

Najarian leads DATA together with H. V. Jagadish, MIDAS Director from 2019 to 2025. The center brings together data scientists, clinicians, pharmaceutical companies, and regulators to work on shared challenges in drug design, toxicity prediction, treatment assessment, and pharmacovigilance.

The stakes are high. Bringing a new drug to market typically takes 10 to 15 years and costs between $1 billion and $2.6 billion. Only about 10 percent of drug candidates that enter clinical trials are ever approved, with most failures occurring late, after enormous investments have already been made.

"The idea is precompetitive research," Jagadish said. "The research problems are of interest to many pharmaceutical companies, so it makes sense to pool expertise rather than solve them independently behind closed doors."

Industry partners including Amgen, AbbVie, and Sanofi co-fund research through DATA. Company scientists help select projects and gain access to validated methods, trained talent, and favorable intellectual property terms. DATA fills a national niche for data- and AI-driven drug development, and has already produced new AI technologies for drug discovery and assessment and continues to expand rapidly.

From silos to shared infrastructure

A decade ago, both clinical AI and drug development research looked very different. Clinical AI efforts were often small pilots confined to single hospital units, with limited access to integrated data or validation across sites. Electronic health record data largely sat unused.

“There was a lot of talk about using AI in medicine,” Najarian said, “but very little infrastructure to do it at scale.”

Drug development faced similar fragmentation. Academic labs developed promising algorithms or compounds, but rarely had pathways to test them on industrial datasets. Pharmaceutical companies maintained in-house analytics teams but had little incentive to collaborate on foundational methods.

MIDAS helped change that landscape. “MIDAS gave researchers room to take risks,” Jagadish said. “It helped build a culture where clinicians, engineers, statisticians, and computer scientists could work together.”

Those early investments aligned with broader efforts at Michigan Medicine to operationalize AI responsibly and complemented other central efforts. The AI & Digital Health Innovation group, for example, helped ensure that models developed in research settings could integrate into clinical workflows and align with health system priorities.

On the drug development side, MIDAS played a key role in securing the NSF designation for DATA, positioning the university as a national hub for precompetitive collaboration. DATA developed a Secure Data Hub that allows companies to share encrypted datasets and created federated learning frameworks so models can be trained across multiple databases without data leaving their home institutions.

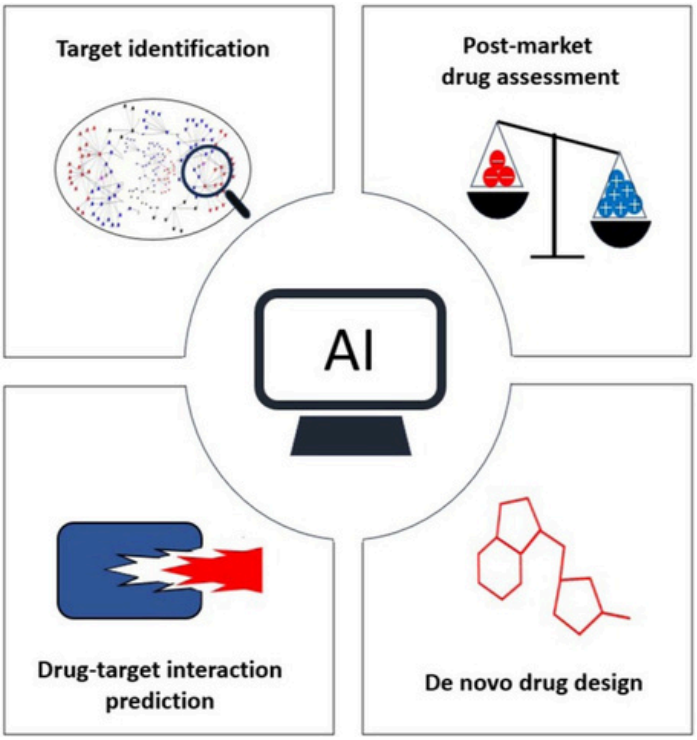


Figure from Minocerri et al. (2025) - AI applications along four critical steps of the DDD process.

Saving time and saving lives

Back in the emergency department, Najarian’s sepsis trajectory models are designed to act as a “second set of eyes.” They do not replace clinical judgment, but provide continuous risk updates that help clinicians prioritize attention in fast-moving situations. In some cases, earlier identification can mean starting antibiotics an hour sooner, which, in some cases, could make a difference in significantly reducing mortality. AI, when used properly and rigorously, can save lives.

Meanwhile, projects funded by DATA are illustrating how AI can reshape drug development upstream.

One project, led by Duxin Sun in the College of Pharmacy, uses machine learning to analyze preclinical and molecular data and predict safety issues before drugs enter human trials. “If you can predict safety failures early, you save years of work and avoid exposing patients to risky compounds,” Sun said.

Another project, led by Daniel Beard in the Department of Molecular and Integrative Physiology,

uses quantitative systems pharmacology models, which means detailed simulations of how drugs affect biological pathways, to predict toxicity arising from emergent system-level effects "We're moving toward 'digital twins,'" Beard said, "models that let you test therapies in silico before running human trials."

Another project, co-led by Peter Tessier in the Department of Pharmaceutical Sciences and Najarian, applies generative AI to design new therapeutic antibodies with improved binding, lower toxicity, and better manufacturability.

Privacy, validation, and governance

One of DATA's most ambitious efforts focuses on privacy-preserving machine learning. In a project led by Najarian, researchers use fully homomorphic encryption, enabling models to train on encrypted data without ever decrypting it. "That unlocks enormous collaborative potential," Najarian said. "Organizations can work together without exposing sensitive data."

Pharmaceutical partners gain access to validated methods and talent at lower cost, while regulators and payers benefit from more transparent, rigorously evaluated models for safety and real-world evidence.

Researchers are also acutely aware of equity concerns. Studies at Michigan Medicine, including evaluations of commercial sepsis tools, show that models trained in one setting may perform poorly in others, particularly across different patient populations. "That's why validation and ongoing monitoring matter, and why DATA regards it as a central focus." Jagadish said.

Looking ahead

From digital twins to adaptive clinical trials, AI-driven approaches are beginning to shift medicine from describing what is happening to predicting what might happen next. DATA's national role offers a model for how universities, industry, and government can collaborate on problems too large for any one institution to solve alone.

From emergency department monitors to encrypted drug discovery pipelines, the work reflects a shared belief that AI can make medicine faster and safer, when built, validated, and governed with care.

> **AI and data science aren't magic bullets, but used responsibly and collaboratively, they're powerful tools. That's what MIDAS and DATA are trying to demonstrate at scale.**
>
> **-H.V. Jagadish**

DATA BENEATH THE WAVES: PROTECTING THE NORTH AMERICAN GREAT LAKES AND A CHANGING CLIMATE

The North American Great Lakes hold about 21 percent of Earth's surface freshwater and supply drinking water, habitat, transportation and recreation for more than 40 million people in the U.S. and Canada. As climate change, pollution and invasive species stress this system, MIDAS is helping researchers use AI and advanced data analytics to forecast harmful algal blooms, model ice and water levels, and plan for a hotter, more volatile future.

From space, western Lake Erie looks almost beautiful in mid-July, swirled with bands of bright green. Up close, the color tells a different story. The green comes from cyanobacteria, organisms that can produce toxins dangerous to people, pets, and wildlife.

Water plant managers monitor forecasts that blend satellite imagery, buoy data, and computational models to predict how large a bloom will grow and whether it will produce toxins like microcystin. Those forecasts can arrive weeks in advance, time to adjust treatment processes, move intake depths, or issue advisories before contaminated water reaches the tap.

Those bulletins increasingly rely on machine learning and AI, part of a broader effort, supported by the Michigan Institute for Data and AI in Society, to bring data science into every aspect of Great Lakes research and management.

What's at stake

The Great Lakes hold about 21 percent of Earth's surface freshwater and support more than 40 million people across the United States and Canada. They provide drinking water, transportation, fisheries, recreation, and habitat, and they are changing rapidly.

Over recent decades, the lakes have experienced record-low winter ice cover, extreme water temperature events, dramatic swings between high and low water levels, and longer growing seasons for harmful algal blooms. During some winters, ice cover during peak season drops to just a few percent, altering lake-effect snow, shipping operations, and aquatic ecosystems.

"The Great Lakes and their coastal ecosystems are changing from a wide range of human impacts," said Bill Currie, a professor at the School for Environment and Sustainability. "We need tools that help us anticipate those changes and understand how to protect these systems."

The data challenge

The problem is not a lack of data. Federal agencies such as NOAA, the EPA, and the USGS collect enormous streams of information from satellites, buoys, ships, and shore stations. Universities and state agencies add decades of additional measurements.

A big challenge is integration. Climate data, watershed runoff models, lake temperature and chemistry measurements, and biological monitoring interact in complex ways, but traditional approaches often examine them in isolation.

Machine learning offers ways to handle that complexity by identifying patterns across massive, multi-source datasets and assessing uncertainty across systems. But using these tools well requires close collaboration: environmental scientists need to understand what AI can and cannot do, and data scientists need to understand the physical and biological processes they are modeling.

MIDAS makes connections

MIDAS identified environmental and climate science as areas where data science could have a disproportionate impact and began supporting cross-disciplinary projects through its seed funding programs. These efforts brought together environmental scientists, computer scientists, statisticians, and policy researchers to tackle real-world problems.

One early project focused on forecasting ice conditions in the St. Marys River, the narrow channel connecting Lake Superior and Lake Huron and home to the Soo Locks, critical infrastructure through which roughly 80 million tons of cargo pass each year.

Ice conditions in the river are notoriously difficult to predict, shaped by weather, currents, lake conditions, and lock operations. With support from a MIDAS pilot grant, Ayumi Fujisaki-Manome, a researcher at the School for Environment and Sustainability and the Cooperative Institute for Great Lakes Research, and co-investigator Christiane Jablonowski developed machine learning models to forecast ice cover seven to thirty days in advance.

“This isn’t just about convenience,” Fujisaki-Manome said. “It’s about safety. With climate change, we’re seeing much more variability in ice cover, and forecasting that variability helps everyone adapt.”

That work continues through larger projects supported by the National Oceanic and Atmospheric Administration to develop Great Lakes Earth system models that couple ice dynamics with waves, currents, and climate processes.

Preparing researchers and building community

MIDAS also recognized that lasting impact required training. Many environmental scientists have strong quantitative skills but limited exposure to modern machine learning methods, while data scientists often lack deep knowledge of environmental systems.

To bridge that gap, MIDAS launched the Environmental Data Science Summer Academy, a three-day intensive that taught faculty and staff researchers practical skills in machine learning, spatial statistics, time-series analysis, and high-performance computing. “The goal isn’t to turn ecologists into machine learning experts,” said Jing Liu, Executive Director of MIDAS. “It’s to help them collaborate effectively and understand when AI methods are appropriate.”

> **“AI isn’t a magic solution. It has to be combined with domain expertise, rigorous validation, and honest communication about uncertainty.**
>
> **-Dani Jones**

In 2024, MIDAS and the Cooperative Institute for Great Lakes Research co-hosted an AI Horizons summit focused on the future of AI in Great Lakes science. Researchers from across the country outlined priorities, including better data integration, development of “digital twin” lake models, optimized sensor networks, and expanded AI-based forecasting.

Shifting seasons, shifting risks

Climate change is not only warming the Great Lakes region, it is reshaping the timing of the biological events that unfold across it. MIDAS postdoctoral fellow Yiluan Song and faculty mentor Kai Zhu study these shifts by combining large ecological datasets with machine learning to understand how seasonal cycles are changing and what that means for ecosystems and human health.

Their recent work focuses on fungal spores, an important but underexamined driver of allergy seasons and ecological processes. By analyzing decades of atmospheric, climate, and land-use data, Zhu and Song found that spore seasons now begin more than three weeks earlier than they did in the 1970s, a direct consequence of warming temperatures and changing humidity patterns. The lengthening of these seasons influences respiratory health, forest dynamics, soil processes, and the timing of numerous other species interactions.

To anticipate these changes, Zhu and Song are developing AI models that integrate weather patterns, environmental conditions, and long-term biological records. These tools provide early warnings for shifts in spore activity and other seasonal biological markers, helping public health agencies and environmental managers prepare for climate-driven risks.

Karen Alofs points to a map that shows how fish ranges are shifting as Great Lakes waters warm. (Still image via Aaron Martin)

Life beneath the surface

MIDAS also supports work that looks past the water's surface to understand how climate change is reshaping Great Lakes ecosystems. A major line of research is led by Karen Alofs, whose group examines how fish communities respond to warming waters, invasive species, and habitat loss.

Alofs' team uses long-term fisheries datasets, digitized historical records, angler reports, and environmental monitoring data to trace how species such as walleye, yellow perch, smallmouth bass, and lake trout have shifted their ranges and behavior over the past century. Integrating archival data with modern analytical tools, the group uncovers trends that would otherwise remain hidden, such as how warming favors warm-water species, alters predator–prey dynamics, and changes which lakes can support which fisheries. These shifts directly influence commercial harvests, recreational fishing, and the cultural identities of coastal communities.

The work extends deeper as well. Katie Skinner develops autonomous underwater vehicles and AI methods to map lake bottoms and identify submerged habitats. Originally designed to locate shipwrecks, these technologies now help researchers detect spawning grounds, monitor habitat change, and identify areas vulnerable to invasive species.

Together, these projects reveal an underwater world in rapid transition. By combining ecological data, AI, and advanced sensing technologies, Alofs, Skinner, and collaborators are generating the knowledge needed to protect fisheries, support communities, and anticipate ecological surprises as the Great Lakes warm.

Who benefits and what comes next

The impacts of such research extend across the Great Lakes region. Water utilities use bloom forecasts to protect drinking water. Shipping companies rely on ice forecasts to manage risk. Coastal planners use water level and temperature data to guide infrastructure investments. The data also inform binational governance and state-level decision-making.

Looking ahead, researchers envision integrated digital twins of the Great Lakes models that allow for testing of scenarios before making costly or irreversible decisions.

As climate change reshapes freshwater systems worldwide, the Great Lakes serve as both a local lifeline and a global testbed. Through early investments, training, and collaboration, MIDAS is helping turn vast environmental data into tools communities can use to navigate an uncertain future.

MIDAS MICHIGAN INSTITUTE FOR DATA & AI IN SOCIETY
UNIVERSITY OF MICHIGAN

CREDITS

Graphic Design:
Catherine Leonhard
Justin Varney

Content Contributors:
Christina Certo
Nathan Fox
Jing Liu
Do-Hee Morsman
Eric Shaw
Justin Varney

All images without a caption are stock or AI generated

Michigan Institute for Data & AI in Society
Weiser Hall, Suite 600
500 Church Street Ann Arbor, MI 48109

midas.umich.edu