Proposal Title: #1001: AEO/GEO Improving Companies' AI Visibility in Chatbot Responses

Department: Kelley School of Business, Operations and Decision Technologies, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Answer engine optimization (AEO) / Generative search optimization (GEO) is quickly replacing traditional search engine optimization. AEO/GEO is an optimization strategy focused on enhancing the visibility of companies in AI chatbot responses. This project introduces an AEO/GEO conversion tool designed to help businesses transform their digital content into a vector database for Retrieval-Augmented Generation (RAG). The tool requires six core components: an input module, a web crawler, a content extraction module, a content transformation module, a discovery layer generator, and an output delivery module. A prototype is already available at https://www.thellmstore.com/. It currently features an input module, where users can enter a website URL. The tool then deploys a web crawler and a content extraction module to analyze the site and provide an AI visibility score, which includes a screenshot of the homepage. The next phase of development requires an engineer to build the remaining modules. This includes transforming the website content into a format and schema optimized for Large Language Models (LLMs). The discovery layer will be generated in JSON-LD or Markdown, making it easily discoverable by AI chatbots. Finally, the output delivery module will provide companies with the discovery layer and instructions on how to use it via an API to increase their visibility. After the AEO/GEO conversion tool is completed, the next phase will include the development of an AEO analytics dashboard. The AEO dashboard will allow companies to monitor the visibility of their brand across keyword searches in AI chatbots. For example, companies within a certain sector may wish to be affiliated with certain prompts or search terms, such as “running shoes.” The AEO search tool will generate a series of prompts related to “running shoes” and will query popular AI chatbots, including ChatGPT, Perplexity, Gemini, and Claude. The AEO conversion tool will parse the responses, identifying the brands, websites, and companies associated with “running shoes.” The tool will then assess the content on those websites and report on an AEO analytics dashboard the results. Results may include competing brands, sentiment towards those brands, associated key terms, visibility rankings for each of the brands, and prescriptive recommendations to a company on ways to further improve their AI visibility score. The types of visualizations the dashboard will include are spider charts, time series line charts, bar graphs, heat maps, geographical maps, among others. To evaluate the efficacy of the tool, we will employ split a/b testing experiments to understand the optimal design for dashboards, displayed content, and more. User satisfaction surveys will be administered, which will require the development team to make weekly changes to the tools. Importantly, we will employ an agile web development approach, and new features will be adjusted or included on a weekly basis.

Rationale for assistance in data analytics and visualization: Data analytics and visualization assistance are required for the following reasons: (1) the LLM conversion tool requires natural language processing to extract and convert a company’s website content into an AI visibility score. This will require extensive programming and mathematical computation to generate a feasible and meaningful visibility score. (2) The AEO dashboard will require a series of charts and visualizations to allow business owners to quickly assess their AEO performance and to understand ways to enhance their visibility score. (3) The

prototype currently is primitive, and the code base is available through GitHub. However, when the product is fully deployed, companies should have their own member's page, requiring their unique login and authentication. Thus, a Tableau-type deployment may be required to create unique visualizations for every member of the website.

Statement of benefit to the student: Students will learn crucial LLM skills as they deploy AI chatbots using APIs through OpenAI. They will learn to fine-tune LLM models to improve the response quality of AI chatbots. In addition, they will generate a novel conversion tool that converts website content into LLM-accessible material, helping businesses transition from e-commerce to AI commerce. Students will learn best practices in data visualization, as they develop dashboards for business owners. Before final coding of the dashboard, usability testing involving prospective business owners will be performed for quality assurance. Students will learn to develop AI tools using a sprint methodology, where features are developed on a weekly basis. Importantly, AEO consulting is an emergent field, and students who understand the technical aspects of improving companies' AI visibility can increase the appealability of their job market candidacy, especially when commerce moves toward agentic/AI commerce as envisioned by OpenAI, Visa, Amazon, and more.

Specific competencies required, including programming languages if applicable: Project Skills: (1) Data visualization, (2) Natural language processing, (3) Statistics, (4) Web Front-end and Back-end development, (5) LLM fine-tuning, and (6) AI chatbots. Programming Languages and Tools: (1) Python and (2) SQL.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1002: AI methods to uncover disease pathways for Dementia

Department: School of Medicine, Biostatistics and health data Science, Indianapolis

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: Network science is interdisciplinary. Focusing on improving current analytical strategies, I try to absorb influences of other disciplines to inform and improve understandings about human behaviors with applications in neuroscience, psychiatry, psychology and others. Broadly speaking, I model real-world networks with statistically sound principles and methods—reduce network complexity, perform intuitive visualization and test critical hypotheses with improved statistical power and controlled type I errors. I am particularly interested in the topological structures of networks and the roles they play in (neuro)degeneration, (neuro)development and resilience. The lab focuses on applying and developing scalable statistical and computational approaches for high-dimensional biomedical data with emphases on network science, high dimensional data analysis, neuroimaging, data integration and statistical machine learning. Depending on prior experiences, selected candidates will have the opportunity to contribute to substantive neuroscience research, method innovations or software development and documentation.

Rationale for assistance in data analytics and visualization: Brain imaging data are inherently high-dimensional and characterized by complex spatial, temporal, and network-level correlation structures. These challenges are compounded by distributed anatomical variation associated with physiological traits, behavioral phenotypes, clinical conditions, and the dynamic patterns of brain activity observed during both resting-state and task-based paradigms. Traditional mass-univariate approaches often fail to account for these dependencies, risking inflated false-positive rates and reduced power to detect biologically meaningful effects. My research program develops advanced statistical models and scalable computational tools that explicitly incorporate spatial dependence, hierarchical network structure, and multimodal data integration. These methods are designed to improve sensitivity and specificity in detecting brain–behavior associations, to robustly characterize large-scale network topology, and to accommodate the heterogeneity of clinical populations. By combining Bayesian hierarchical modeling, generative embedding frameworks, and high-performance computing, my work enables reproducible, interpretable, and computationally efficient analyses of large-scale neuroimaging datasets, directly supporting precision neuroscience and the development of network-based biomarkers for neuropsychiatric and neurodegenerative disorders.

Statement of benefit to the student: Depending on prior experiences, selected candidates will have the opportunity to contribute to substantive neuroscience research, method innovations or software development and documentation.

Specific competencies required, including programming languages if applicable: If you possess the following qualifications, I strongly encourage you to consider applying: Have a background or prior

experience in neuroimaging, network analysis, or computational neuroscience. Skilled in programming and/or statistical method development.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1003: The Digital Mirror: Analyzing Social Media Behavior to Address Mental Health in Young Adults

Department: Kelley School of Business, Department of Operations & Decision Technologies, Bloomington

Project requires Human Subjects? Yes

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: According to the 2022 National Survey on Drug Use and Health (NSDUH), a staggering 59.3 million adults in the U.S. experience mild to moderate mental illness. The highest prevalence, at 36.2%, is found among young adults aged 18-25, a demographic largely comprised of college students and their peers. This generation, having grown up with technology, relies on social media as a primary source for communication, information, and emotional support, often viewing these platforms as safe, non-judgmental spaces. However, this reliance also brings serious negative consequences, including sleep disruption, addiction, social isolation, and a heightened risk of suicide. As information technologies like wearable devices, mobile apps, and health sensors become more advanced, it's essential that we use signals from these digital sources to detect mental health issues, raise awareness, and promote mental wellness. This research will focus on the mental health and well-being of college students, exploring how their social media behavior can offer insights into their mental health conditions. Our ultimate goal is to develop actionable strategies for early diagnosis, intervention, and cost-effective behavioral treatments. We will generate evidence-based recommendations that demonstrate what works and how it works, offering practical insights to improve mental health outcomes at both the individual and societal levels. To achieve this, we will develop an empirical approach that combines artificial intelligence (AI) and cognitive science to analyze students' mental states and social media activities. By examining online social interactions, AI will be used to identify key content traits—such as topics, emotion, intent, and urgency—that shape digital conversations. This analysis will help us map linguistic patterns and situational characteristics, allowing us to assess the impact of social media on college students' mental health and vice versa. This project expands on my current research, which focuses on the relationship between college students' mental health and their academic performance. With existing IRB approval to access students' academic data, analyzing their social media activities will provide a more holistic understanding of their well-being. The anticipated outcome is an integrated framework that connects students' online and offline activities with their mental health status. Using empirical insights, we will design a system that incorporates data from various digital sources and leverages intrinsic motivation to promote mental wellness.

Rationale for assistance in data analytics and visualization: For this particular project, I need assistance with data collection and analysis. The focused social media platform is Reddit. I have funding to subscribe to Reddit's paid service, but I need help collecting data regularly to construct a panel database. Moreover, Reddit has a loose platform structure, and its data often contains a lot of noise. It will be time-consuming to map, extract, and clean the data to reach the desired quality for conducting research. After the initial data collection process, I need assistance with NLP or a generic AI framework to adapt or develop an AI-based measure for mental health condition detection. If the working relationship continues, the student(s) will help design mobile apps to test the effectiveness of behavior

interventions guided by the empirical findings. For the generalizability of the research, we will collect data from other social media platforms or healthcare-focused social groups on popular sites such as X or Facebook (I have received grant funding to cover subscription fees to these platforms). This requires continuous technical support.

Statement of benefit to the student: Based on my prior participation with FADS, I have learned the students' capabilities and have hired students as a Graduate Assistant in my teaching semester and Research Assistant in the other semesters.

Specific competencies required, including programming languages if applicable: The student(s) need to use Python to collect data, conduct data analysis, and perform language processing. Student(s) with project skills in Web & Social Media Mining are expected. To engage in data collection, cleaning, preparation, and analysis tasks, the student(s) need to have a good understanding of machine learning, Natural Language Processing, and statistics (e.g., Stata). The student(s) with experience in Cloud & High-Performance Computing are preferred.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1004: AI-Powered Drug Discovery for Deadly Pediatric Brain Tumors

Department: IU School of Medicine - Bloomington, Medical Sciences Program, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: Diffuse intrinsic pontine glioma (DIPG) and diffuse midline glioma (DMG) are universally fatal pediatric brain tumors, with median survival of less than one year despite aggressive multimodal therapy. The urgent unmet clinical need is compounded by the limited number of effective compounds that cross the blood–brain barrier (BBB) and target the unique vulnerabilities of DIPG/DMG, particularly those associated with tumor stemness and epigenetic dysregulation. Traditional drug development pipelines are too slow and costly to meet the needs of these children, making computationally driven drug discovery and repurposing an attractive and highly impactful strategy. This project aims to build an AI-enhanced drug discovery framework that integrates biological knowledge with advanced machine learning to identify new therapeutic opportunities for DIPG/DMG. Specifically, we will construct a heterogeneous knowledge graph (KG) that captures drug–target–gene–disease–pathway interactions from publicly available datasets (e.g., DrugBank, ChEMBL, LINCS, DepMap, and pediatric cancer resources). Using this KG, we will train graph neural networks (GNNs) to predict novel drug–target associations, prioritize drug repurposing candidates, and identify rational drug combinations. In parallel, we will apply generative deep learning models to design optimized analogs with improved BBB penetration, pharmacokinetic properties, and reduced pediatric toxicity. To ensure biological and translational relevance, we will anchor our computational pipeline with glioma stem cell (GSC) signatures that we have already derived from transcriptomic profiling of patient-derived DIPG/DMG lines. Candidate compounds predicted by the models will be validated through in silico docking and filtered using pharmacological rules tailored for central nervous system penetration. The most promising agents and combinations will then be prioritized for in vitro testing in our established patient-derived DIPG/DMG GSC models, with functional readouts including viability, sphere formation, and synergy with radiation or temozolomide. The ultimate deliverables of this project will include: (1) A ranked list of repurposable FDA-approved or investigational compounds predicted to target DIPG/DMG vulnerabilities. (2) Predicted rational drug combinations with mechanistic annotations. (3) A prototype decision-support framework that can be updated with new biological and pharmacological data, serving as a living resource for ongoing translational studies. By integrating cutting-edge AI methods with the biology of DIPG/DMG, this project has the potential to dramatically shorten the timeline for identifying actionable therapies. If successful, it will lay the groundwork for early-phase clinical trial concepts, while also providing students with exposure to advanced machine learning applications in biomedicine, graph analytics, and translational oncology.

Rationale for assistance in data analytics and visualization: The success of this project depends on advanced data integration, machine learning, and visualization capabilities that exceed the expertise of my laboratory. Constructing and analyzing a heterogeneous drug–gene–target knowledge graph requires skills in large-scale data wrangling, graph database design, and implementation of graph neural networks. Similarly, developing generative AI pipelines for drug design involves deep learning

frameworks that require specialized technical training. Beyond analytics, data visualization is critical to interpret and communicate the results. Clear visual representations of high-dimensional graphs, compound prioritization, and drug–target interaction networks will be essential for both hypothesis generation and publication-quality figures. For example, intuitive dashboards or interactive network visualizations will allow us to explore predicted drug–gene relationships in the context of DIPG/DMG biology. Partnering with a data science student will provide the necessary computational and visualization expertise, ensuring rigor and scalability of the analysis while also offering the student an opportunity to apply cutting-edge AI/ML methods to a real-world biomedical challenge. Without this program, the project would be significantly limited by the computational expertise available in my lab, slowing progress toward translational impact.

Statement of benefit to the student: This project offers the student a unique opportunity to apply data science in a context with direct, real-world impact: the urgent search for new treatments for children with DIPG/DMG, one of the deadliest pediatric brain cancers. The student will gain hands-on experience working with biomedical datasets and AI tools to identify potential drug candidates, while also learning how computational discoveries can be translated into biological testing and, ultimately, patient care. Beyond technical training, the student will benefit from close mentorship in a highly interdisciplinary environment that bridges data science, cancer biology, and clinical medicine. They will learn how to communicate their results to both scientific and non-technical audiences, strengthening their ability to work across fields. The project is structured to give the student ownership of a defined research question, ensuring they develop independence and confidence in applying their skills to complex, high-impact problems. By the end of the internship, the student will not only expand their portfolio with an innovative application of AI to medicine but also see how their efforts could contribute to meaningful advances in pediatric cancer research.

Specific competencies required, including programming languages if applicable: Data Visualization, Deep Learning, Machine Learning, Statistics, Excel, Python, R

Is there anything else you would like us to know about your project's time frame or work schedule?: I hope to start this project as soon as possible

Proposal Title: #1005: Global Muse-Gen: A Cross-Cultural AI Model for Museum Layouts

Department: Eskenazi School of Art, Architecture + Design, Interior Design, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: This project addresses the dominance of Western-oriented training data in current Generative Artificial intelligence (Gen AI) models—a bias that sidelines cultural and vernacular elements vital to architecture and interior design. Such elements, rooted in local practices and contexts, risk being erased from a tool that is rapidly reshaping these two professions. This risk is acute in museums—cultural hubs that, even before GenAI, often mirrored the design norms of well-funded agencies. By creating a dataset rich in cultural and vernacular elements and developing a Gen AI architecture attuned to these differences, this project will (1) ignite creativity, innovation, and expression in museum design and (2) bridge art, humanities, and technology, amplifying dialogues that redefine and enrich design norms. This project builds on Muse-Gen, funded by the Eskenazi Technology Innovation Lab in summer and fall 2024, which reached its first milestone: creating block diagrams—simplified spatial representations of museum layouts—and assembling a core dataset of 50 publicly available, predominantly Western-oriented sources. With continued support from the 2024 College Arts and Humanities Institute (CAHI) Faculty Grants in Support of Research + Creative Activity and the 2025 Eskenazi Community Impact Fund, the work now evolves into Global Muse-Gen. This next phase requires a richer dataset and a more advanced model architecture—one that follows architectural drawing conventions such as wall thickness and door swings, while accurately interpreting area boundaries and labeling the functions of different spaces. We are seeking support from the Faculty Assistance in Data Science (FADS) program to reach the next milestones: (a) expand the dataset to include non-Western museum layouts, and (b) finalize a new model architecture that delivers accurate area calculations (in square feet), clear functional labels, and handles complex design elements such as curvatures. Students will contribute to both data procurement and model development. Because the project uses synthetic data with no human subject interaction, IRB training is not required; however, students must sign a data management plan to ensure compliance with data stewardship and confidentiality. Initial work will involve using LibreCAD to create synthetic museum layouts from publicly available non-Western sources—an open-source choice that ensures ethical, IP-compliant training data and offers an intuitive interface for beginners. An undergraduate research assistant in interior design will provide hands-on guidance in software use and ensure adherence to architectural drawing conventions. Next, students will help build and train the model on the new dataset, maintaining a private GitHub repository with all code, documentation, and an interface for user testing of the complete Global Muse-Gen model. They will also document data sources to ensure transparency and reproducibility, create targeted test cases to as

Rationale for assistance in data analytics and visualization: The project now enters a phase requiring specialized technical expertise beyond the scope of our current undergraduate design students. Our work demands advanced skills in cloud and high-performance computing, computer vision, and deep learning to train a custom model on an expanded dataset of global museum floor plans—shifting from a Western-oriented base to a culturally diverse collection. We also require precision in architectural data

processing, including adherence to drawing conventions, accurate area calculations, and complex geometry handling. While our design students excel in spatial thinking, they lack the programming, statistical, and AI modeling experience needed to build and optimize this next-generation architecture. Graduate students with data science expertise will be pivotal in developing the model, managing a secure GitHub repository, creating test cases, conducting error analysis, and building an interface for user testing. With prior work completed using limited research funds, the FADS program will allow us to fully realize the model's capabilities, ensuring accuracy, transparency, and ethical dataset management. The students' contributions will directly amplify the dialogues between art, humanities, and technology while enriching design practices in the age of Gen AI.

Statement of benefit to the student: This project offers graduate students hands-on experience at the intersection of architecture, interior design, and advanced data science. By contributing to the expansion of Global Muse-Gen—including building a culturally diverse museum layout dataset and developing a custom AI model—students will gain expertise in applying generative AI to spatial design problems, an emerging and in-demand skillset in both academia and industry. Students will work with real architectural data, honing data management, documentation, and ethical stewardship practices while learning to use open-source tools like LibreCAD for dataset creation. They will engage directly in model training, test case development, and error analysis, deepening their knowledge of computer vision, deep learning, and architectural drawing conventions. Collaboration within a multidisciplinary team will strengthen project management, problem-solving, and communication skills. Students will also develop portfolio-ready technical outputs, from a secure GitHub repository to a functional user-testing interface. The workload (10–15 hours per week) will be designed to complement academic commitments. All contributing students will be credited in publications, with authorship order reflecting the extent of their contributions; significant but non-authorship contributions will be formally acknowledged. This experience will equip students with both technical and professional skills highly valued in design, data science, and AI-d

Specific competencies required, including programming languages if applicable: Cloud & High Performance Computing, Computer Vision, Deep Learning, Machine Learning, Natural Language Processing, Network analysis, Statistics, C/C++, Excel, Java, MatLab, Python, R, SQL

Is there anything else you would like us to know about your project's time frame or work schedule?: The project is due at the end of March 2026 but faced a major setback when Autodesk—an industry leader in architecture and interior design—banned any AI use of its software outputs, with no exceptions for research. Our core dataset, created in Autodesk software before this ban, can no longer be used. We are pivoting to LibreCAD, an open-source and ethically sound alternative, and rebuilding from scratch. This challenge highlights a growing risk: industry gatekeepers restricting data access in ways that could stall even public-benefit research. Support from the FADS program is critical to ensure this work continues—advancing ethical, accurate AI research in architecture and interior design for the benefit of the public.

Proposal Title: #1006: Database of Pulsed-Laser Testing Facilities for Microelectronics Radiation Effects Research

Department: Luddy School of Informatics, Computing and Engineering, Intelligent Systems Engineering, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Summer 2026

Summary of project: Microelectronic devices power modern technology from smartphones to satellites. In space and other harsh environments, radiation can cause sudden errors or failures. To understand and mitigate these effects, researchers use specialized testing. Pulsed-laser testing, which uses short laser pulses to emulate radiation, has become an increasingly important alternative to accelerator-based methods because it is more affordable and accessible. Although pulsed-laser testing facilities now exist in government labs, universities, and industry, no central resource catalogs their locations, capabilities, or access policies. New researchers and industry partners often struggle to identify appropriate facilities or compare their capabilities. This project will develop the first searchable, global database of pulsed-laser testing facilities. The database will capture standardized information including location, contacts, laser specifications (wavelength, pulse width, repetition rate, energy range), supporting equipment (microscopes, probe stations, chambers), areas of expertise, and representative publications. The outcome will be an interactive, web-based platform with three features: a user-facing interface with search, filters, and maps; a secure facility-facing portal for sites to manage their own information; and an intake form for new facilities to submit entries. These tools ensure the database is accurate at launch and sustainable over time. The database will reduce barriers for students and early-career researchers, provide industry and government with transparency and redundancy, and enable global analyses of pulsed-laser testing capacity. For the data science intern, the project offers the chance to apply technical skills to a resource with international visibility. The student will design the database structure, implement search and visualization tools, and help build a functional prototype used by the research community. Their contributions will form the basis of a peer-reviewed paper submission to the IEEE Nuclear and Space Radiation Effects Conference in 2026, giving them direct impact on both scientific infrastructure and the published literature. In short, the project combines advanced data science with a high-impact application, creating a lasting resource for researchers while giving the student practical experience in database design, visualization, and web application development.

Rationale for assistance in data analytics and visualization: The proposed database requires expertise in database architecture, data wrangling, and visualization that goes beyond the faculty investigator's primary training in microelectronics and reliability testing. While the PI can define the technical parameters needed to describe pulsed-laser facilities, the challenge lies in translating these requirements into a scalable and sustainable data system. Without data science support, the effort would remain at the level of a static spreadsheet, limiting its value to the community and making it difficult to maintain. FADS assistance is essential to design and implement a relational database structure, create efficient data pipelines, and build interactive visualization tools such as searchable tables, filters, and maps. A data science student can also implement a facility-facing web interface for ongoing self-management and an intake form for new entries, ensuring accuracy and sustainability over

time. This project will move from a concept to a robust, community-facing resource by pairing domain expertise with advanced data analytics and visualization skills. The partnership with an MSDS student will ensure that the platform is technically sound, user-friendly, and positioned for long-term growth.

Statement of benefit to the student: This project will give the student an opportunity to apply data science skills to the field of microelectronics research by developing a solution with international visibility. The student will work closely with the faculty investigator to convert domain requirements into robust data solutions, building valuable skills in cross-disciplinary communication. They will strengthen their ability to design technical systems from program needs, create user-focused products, and gain experience in web development. Their work will move beyond classroom exercises to a functional prototype that is actively used by researchers, industry, and government programs. The results will support a peer-reviewed paper submission to the IEEE Nuclear and Space Radiation Effects Conference in 2026, giving the student the opportunity to receive credit as an author. This project will strengthen their portfolio, highlight their ability to deliver usable data products, and demonstrate their impact on advancing research infrastructure.

Specific competencies required, including programming languages if applicable: Data Visualization; Database Management; Web Front-end Development; SQL;

Is there anything else you would like us to know about your project's time frame or work schedule?: A paper will be submitted in January 2026 for peer-review at the IEEE Nuclear and Space Radiation Effects Conference with the results of a survey of facilities. This survey will form the input data to the database. A Spring 2026 timeline allows students to work on the project and, if the paper is accepted, they will be added as co-authors in the final submission of the paper in July.

Proposal Title: #1007: Faculty Assistance in AI Applictaion

Department: Kelley School of Business, Accounting, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: This proposal requests data science assistance for several research projects that leverage artificial intelligence to analyze financial disclosures by companies and financial intermediaries. A unifying feature across these projects is the integration of both open-source models (e.g., Qwen, DeepSeek, Llama) and commercial AI platforms (e.g., OpenAI, Anthropic, Gemini). Open-source models provide transparency, flexibility, and reproducibility, while commercial systems deliver scalability and cutting-edge performance. This dual approach ensures robust, cost-effective, and academically rigorous analysis. Analyzing Roadshow Video Presentations Before Initial Public Offerings (IPOs) Roadshow presentations are pivotal communication events that shape investor sentiment before a company goes public. This project applies multimodal AI to video and audio data from these events. Open-source models such as Qwen-VL and Whisper-style speech encoders will extract linguistic, visual, and paralinguistic features, while commercial large multimodal models (LMMs) will provide contextual interpretation and higher-order pattern recognition. The goal is to uncover systematic cues in how firms present themselves and assess how these features influence IPO pricing and aftermarket performance. Parsing Sell-Side Equity Analyst Reports Sell-side analysts' stock recommendation reports are central to the information environment of capital markets. This project classifies the topics analysts emphasize—such as growth prospects, risk factors, or ESG considerations—and links those themes to their valuation models and price targets. Examining Vocal Features of Executives During Earnings Calls Executives' vocal delivery during quarterly earnings calls conveys information that extends beyond the transcript. This project analyzes tone, pitch, pacing, and other acoustic features using open-source speech and audio processing models, complemented by commercial multimodal models for contextual interpretation. By connecting vocal dynamics to short-window market reactions, the project will provide new insights into how communication style affects investor confidence, perceived credibility, and the informativeness of disclosures.

Rationale for assistance in data analytics and visualization: Assistance in data science is essential for building AI pipelines that can efficiently process and analyze the large volumes of audio, video, and textual data across these projects. Each project involves handling gigabytes of unstructured data, which requires the use of IU's supercomputing resources to ensure scalability and performance. Students supporting this work will need to design and maintain distributed ETL (Extract, Transform, Load) pipelines that run across multiple GPUs, making effective use of both parallel processing and shared memory. For roadshow video presentations, this means constructing pipelines that can distribute advanced video and audio models across GPU nodes to extract visual and vocal features at scale. For sell-side equity analyst reports, NLP models must be efficiently deployed to parse and classify thousands of long-form documents, requiring strategies to balance GPU throughput with RAM management. For executive earnings calls, audio signal processing and multimodal embedding models will need to be run across large datasets, with pipelines optimized to extract subtle acoustic features while minimizing

computational bottlenecks. By leveraging IU's supercomputing infrastructure, these pipelines will enable reproducible, large-scale analysis of financial disclosures that would not be possible on standard computing resources. The student must therefore have a strong foundation in distributed computing, GPU cluster management, and memory-efficient mo

Statement of benefit to the student: Participation in this project will give students direct experience applying machine learning and artificial intelligence techniques to research questions in capital markets. They will work with large and complex datasets—including video, audio, and text—that mirror the unstructured information increasingly central to financial analysis. Students will gain practical skills in designing and deploying scalable data pipelines, applying natural language processing and audio signal processing, and harnessing advanced AI models, including both open-source and commercial large language models. In addition, they will learn to run these pipelines on IU's supercomputing resources (Quartz and BigRed), developing expertise in distributing workloads across multiple GPUs and managing computationally intensive tasks efficiently. This training will not only deepen their understanding of how AI and ML can be used to study financial disclosures but will also prepare them for careers in data science, finance, or academic research.

Specific competencies required, including programming languages if applicable: These projects require advanced data wrangling, feature extraction, and modeling, all of which are computationally intensive and demand expertise in programming, machine learning, and scalable AI deployment. A key element is the ability to integrate both open-source models (e.g., Qwen, DeepSeek, Llama) and commercial AI platforms (e.g., OpenAI, Anthropic, Gemini) into robust pipelines that can be distributed across IU's supercomputing resources. Data science assistance is therefore essential for efficiently managing large, multimodal datasets and enabling reproducible, high-scale analyses. Proficiency in Python is critical, along with competencies in: High-Performance & Distributed Computing (leveraging IU Quartz, Big Red, and multi-GPU clusters) Pipeline Development for scalable AI workflows Database Management for storing and querying structured disclosure data Deep Learning model training and fine-tuning Audio Signal Processing for vocal feature extraction Machine Learning & Natural Language Processing for text classification and topic modeling Large Language Models (open-source and commercial) for text and multimodal understanding API Integration to access and operationalize commercial AI platforms

Is there anything else you would like us to know about your project's time frame or work schedule?: I prefer to work with one student at a time, as this ensures that project management, reporting, and communication remain clear and efficient—an especially important consideration given the complexity of these research tasks. Equally important, the student's proficiency and experience in working with AI models and large language models (LLMs) is crucial for my projects.

Proposal Title: #1008: Responsible AI in Mental Health Peer Support

Department: Kelley School of Business, Operations & Decision Technologies, Bloomington

Project requires Human Subjects? Yes

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Mental health disorders affect nearly one billion people worldwide, with anxiety and depression being the most common. According to the World Health Organization, recent years have seen sharp increases—26% for anxiety and 28% for depression. In the United States, the 2022 National Survey on Drug Use and Health reported that 59.3 million adults experience mental illness, with highest prevalence (36.2%) among young adults aged 18–25. Growing up in the digital era, this generation increasingly turns to online peer support communities, viewing them as safe, non-judgmental spaces for connection, understanding, and healing. The integration of AI agents into these platforms introduces new possibilities—and questions—about how technology shapes human connection in mental health settings. While AI can provide around-the-clock support and reach users who might otherwise go unheard, little is known about how it affects peer-to-peer dynamics. Does AI strengthen or weaken community support? How do users change their sharing behavior in the presence of AI? These questions are critical, as peer support platforms serve vulnerable individuals, and poorly designed AI systems risk disrupting the very communities they aim to support (Rayland and Andrews 2023). This study involves handling real-world data from mental health platforms implementing LLM-based AI agents. Our approach combines cutting-edge computational methods with qualitative insights. We employ advanced topic modeling (BERTopic, dynamic models) to track evolving mental health discourse and AI's influence on conversations. We also employ GenAI for mental health text analysis, detecting changes in emotional expression and support quality. Key techniques include prompt engineering, RAG systems for contextual analysis, and fine-tuning LLMs for sensitive domains. The anticipated outcome of this study is a robust, data-driven understanding of how AI integration transforms community dynamics, user behavior, and support outcomes in mental health settings. Our analysis will distinguish conditions under which AI enhances peer support from those where it may undermine community cohesion, offering clear, evidence-based guidance for optimal deployment. The result will be validated frameworks for responsible AI integration—designed to maximize technological benefits while preserving authentic human connection, particularly within vulnerable populations.

Rationale for assistance in data analytics and visualization: Mental health data presents unique analytical challenges requiring specialized support. The project involves two critical phases: first, we need to build comprehensive data collection infrastructure—developing web scraping pipelines, API connectors, and automated extraction systems to gather posts, responses, and AI interactions from online mental health platforms. This collection phase demands expertise in data engineering, ethical scraping protocols, and managing sensitive health information at scale. Once collection systems are operational, the analysis phase begins with processing millions of complex temporal, linguistic, and behavioral patterns. Advanced NLP techniques are needed to analyze mental health discourse, such as identifying emotional states, support quality, and community dynamics. The project requires sophisticated topic modeling (BERTopic), GenAI applications for pattern detection, and sentiment analysis tools specifically calibrated for mental health contexts. Visualization also serves the second

phase, which evolves displaying user engagement patterns, topic evolution, sentiment trajectories, and AI response effectiveness. These visualizations enable dynamic strategy adjustments while revealing behavioral shifts. The sensitive nature of mental health data and the complexity of processing streaming data from multiple sources requires expertise in real-time analytics, data validation, and representing high-dimensional embeddings from GenAI models.

Statement of benefit to the student: Students will gain highly marketable skills at the intersection of data science, AI, and social impact. Through hands-on experience building end-to-end data pipelines, students will master the complete research workflow, from designing data collection systems to deploying advanced analytics. Students will develop expertise in cutting-edge technologies including GenAI/LLM applications, advanced NLP techniques, and topic modeling specifically calibrated for sensitive health data. Working with real-world mental health platforms provides portfolio-worthy projects demonstrating both technical sophistication and social responsibility Beyond technical skills, students will develop critical competencies in research ethics, privacy-preserving methods, and responsible AI deployment, which is essential for careers in health tech, social computing, or AI safety. The mentorship provided includes career guidance, and recommendation letters highlighting unique combination of technical excellence and social impact focus. This experience positions students at the forefront of responsible AI development.

Specific competencies required, including programming languages if applicable: (1) Programming Languages: Python (primary language for data collection and NLP), JavaScript (optional - for scraping JavaScript-heavy sites), R (statistical analysis) (2) Data Collection: Web scraping (BeautifulSoup, Scrapy, Selenium), API integration and RESTful services (3) NLP and Text Analysis: Topic modeling (BERTopic, LDA, dynamic topic models), Sentiment analysis for mental health text, GenAI/LLM fine-tuning and prompt engineering

Is there anything else you would like us to know about your project's time frame or work schedule?: We will submit an IRB application once the specific data collection sites and methods are finalized. We anticipate seeking IRB exempt given the public nature of the data and our privacy-preserving collection methods.

Proposal Title: #1009: Human rights, violence, and conservation at the global scale: Web platform development and automated data collection

Department: Paul H. O'Neill School of Public and Environmental Affairs, The O'Neill School at Indianapolis does not have departments., Indianapolis

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: The Conservation, UnRest, and Violence (CURV) Project is a global research initiative that systematically documents violent events and human rights abuses associated with conservation enforcement in the areas of wildlife, forestry, fishing, and land and sea protection. The project compiles event-level data on actors, tactics, locations, and outcomes, with the aim of shedding light on the most pressing human rights implications of conservation initiatives. In a recently prepared working paper using information gathered from four global databases, the CURV project has shown that violence and human rights abuses are widespread in the conservation sector, especially in the regions of Sub-Saharan Africa and South Asia, that events frequently involve torture or other extreme tactics, and that these problems have grown dramatically in recent years. Through the FADS program, I propose to involve MSDS students in two critical tasks that will expand the accessibility and robustness of CURV: Public-Facing Web Platform Development: The CURV project findings are urgent, especially in light of international commitments to roughly double the extent of the world's land protected for conservation over the next five years under the global "30x30" initiative. There is a need to develop an online tool to publicize and track these problems in near-real time. Thus, students will design and implement a web-based interface that makes CURV data accessible in near-real time. This will include interactive data visualization (maps, timelines). The platform will democratize access to conservation-conflict data for scholars, journalists, NGOs, and policymakers. Web Mining and Data Augmentation: While the current dataset has already produced several insights, statistical analyses suggest that the data only capture a fraction of reported events at the global scale. Expanding the data coverage will expanded research directions for the project team. Thus, students will develop automated web-mining pipelines to supplement the current CURV dataset. This will involve web mining to identify relevant events from online media, NGO reports, and government sources.

Rationale for assistance in data analytics and visualization: The CURV project has reached a stage where scaling and public access require technical skills beyond the capacity of the current research team. Specifically: • Database + Web Development: Building a stable, interactive platform requires expertise in database architecture, web development, and interactive visualization. • Automated Web Mining: Developing processes for automated web mining. While the CURV team includes social scientists with expertise in governance and conservation, we lack the advanced computational capacity required to implement these components. The assistance of MSDS students will allow us to leverage cutting-edge data science techniques, ensuring both technical rigor and long-term sustainability of the project.

Statement of benefit to the student: This project offers MSDS students a unique opportunity to apply data science skills to a high-impact, policy-relevant global research initiative. Students will gain

experience in: • Designing a public-facing platform that merges database engineering, web development, and visualization. • Tackling real-world challenges of automated data collection. • Collaborating directly with faculty and international researchers engaged in human rights research. In addition, students will be credited as contributors to the CURV platform and pipelines, with potential opportunities for co-authorship on academic publications and presentations. The project thus provides both technical training and professional development in ways that extend beyond a typical classroom or course assignment. Finally, some students may take satisfaction from contributing to a project with clear social impact, where their work will inform policy and advocacy, help to publicize human rights abuses at the global scale, and potentially encourage policy change to make conservation policy more peaceful and humane. While the project may be an especially good fit for students interested in careers that use data science for advocacy, this is not a requirement.

Specific competencies required, including programming languages if applicable: Database development and management; Web development frameworks; Data visualization; Web mining and automated data collection

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1010: Investigating biological embedding of stigma and the healthspan of sexual and gender minority (SGM) populations

Department: College of Arts and Sciences, Anthropology, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: My research program investigates the biological embedding of stigma and the healthspan of sexual and gender minority (SGM) populations. Using longitudinal cohort datasets, biospecimens, and survey data, I study how adverse social conditions (e.g. discrimination, food insecurity, early life stressors), shape physical health outcomes across the life course. This project seeks assistance from a Master of Science in Data Science (MSDS) student to build robust data analysis pipelines that will streamline and expand my analytic capacity in examining these relationships. Specifically, this project involves working with large, multi-wave datasets (e.g., the National Longitudinal Study of Adolescent to Adult Health (Add Health), Canadian Longitudinal Study of Aging (CLSA), Demographic and Health Survey (DHS), Canadian Community Health Survey (CCHS), the National Health and Nutrition Examination Survey (NHANES), and the Cebu Longitudinal Health and Nutrition Survey (CLHNS)). These datasets include complex survey designs, longitudinal or repeated cross-sectional biomarker measurements, and sensitive sociodemographic variables. To effectively analyze these data, reproducible and scalable pipelines are needed for: •Cleaning and harmonizing variables across multiple datasets and waves. •Implementing advanced statistical models (e.g., structural equation modeling, latent profile analysis, and survival analysis). •Automating diagnostic and visualization steps for quality assurance. •Preparing results for integration into manuscripts, grant proposals, and conference presentations. Support from a FADS intern will enable the creation of generalized, modular scripts that can be re-applied across multiple projects. The intern will collaborate on developing code for data wrangling (R and Python), writing functions to automate model fitting and diagnostics, and producing high-quality visualizations to communicate findings. These efforts will free substantial time for conceptual development, manuscript writing, and mentoring of junior researchers, while also increasing transparency and reproducibility in my work. This assistance will catalyze new directions in my research by allowing me to move beyond one-off analyses and instead establish scalable workflows. The proposed pipelines will be directly applicable to upcoming grant submissions, including an NIH R01 on mechanistic pathways linking HIV infection to accelerated epigenetic aging. In addition, the project's focus on health disparities among SGM populations provides a unique opportunity to apply advanced data science methods to socially urgent questions.

Rationale for assistance in data analytics and visualization: The scope and complexity of the datasets I work with demand expertise in advanced data science techniques that extend beyond traditional statistical training. While I am proficient in R and experienced in epidemiological and anthropological analysis, building robust, automated pipelines requires dedicated time and programming specialization. Without targeted support, much of my time is spent on manual data wrangling and repetitive coding tasks, slowing the pace of discovery and limiting opportunities to expand my research. A FADS partnership will provide the technical expertise to transform my workflow. Assistance with

programming, database structuring, and data visualization will allow for reproducibility across multiple datasets and greater efficiency in handling complex analytic tasks. This will not only accelerate ongoing projects but also enable more ambitious analyses that are otherwise infeasible due to technical bottlenecks. Support in developing modular pipelines for longitudinal analysis and visualizations of biomarker trajectories will substantially advance my capacity to address central questions in health disparities research.

Statement of benefit to the student: The MSDS student will gain hands-on experience in applying data science methods to pressing issues in public health and health disparities research. This project offers the opportunity to work directly with large, complex, and sensitive datasets that require advanced methods in data cleaning, harmonization, and visualization. The student will be exposed to real-world challenges in managing longitudinal and multi-source data, a skillset highly relevant to careers in data science, biostatistics, or health informatics. The project will also give the student experience in interdisciplinary collaboration. They will work with a principal investigator trained in anthropology, epidemiology, and data science, as well as a broader research team that spans public health, sociology, and molecular biology. By contributing to manuscripts, grant proposals, and visual outputs, the student will see the direct application of their work to scientific knowledge production and public health. The mentorship structure will emphasize skill-building, reproducibility, and professional development, ensuring that the student's internship provides a valuable and marketable set of experiences. Moreover, the student will have opportunities to be included as a coauthor on subsequent manuscript resulting from this work.

Specific competencies required, including programming languages if applicable: The student should bring expertise in the following areas: • Programming Languages: Proficiency in R and Python for data cleaning, wrangling, statistical modeling, and visualization. • Project Skills: o Data Visualization: Ability to produce publication-ready graphics using R (ggplot2, Shiny) or Python (matplotlib, Plotly). o Database Management: Structuring and harmonizing longitudinal and multi-wave datasets. o Statistics: Applying regression models, structural equation modeling, clustering/latent profile analysis, and survival analysis. o Machine Learning & Data Mining: Experience with multivariate approaches and classification methods desirable. o Reproducible Workflows: Use of GitHub and/or R Markdown to ensure transparency and scalability. • Preferred Experience: Handling survey data with complex sampling weights (e.g., NHANES, DHS, Add Health, CCHS) and familiarity with health or social science applications.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1011: Examines statistical norms and cultural assumptions about "normal" bodies

Department: College of Arts and Sciences, Anthropology; Human Biology Program, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: My research program investigates biological normalcy, stigma, and their connections to health outcomes across the life course. Drawing on survey and biospecimen data, my work examines how statistical norms and cultural assumptions about "normal" bodies, together with lived experiences of weight-based stigma and discrimination, shape physiological processes and long-term health trajectories. Current collaborations may also overlap with topics in maternal and infant health. This project seeks assistance from a MSDS student to develop robust analytic pipelines that will expand my capacity to evaluate these complex relationships. Specifically, this project involves working with large, multi-wave datasets (e.g., the National Longitudinal Study of Adolescent to Adult Health (Add Health), Demographic and Health Survey (DHS), The Behavioral Risk Factor Surveillance System (BRFSS), the National Health and Nutrition Examination Survey (NHANES), etc). These datasets contain biomarker measurements, anthropometric data, sensitive sociodemographic variables, and repeated measures of weight-related attitudes and experiences. To analyze these data effectively, reproducible and scalable pipelines are needed for: ●Cleaning and harmonizing weight- and health-related variables across multiple datasets and waves. ●Cleaning and harmonizing maternal and infant health-related variables across multiple datasets and waves. ●Implementing advanced statistical models (e.g., latent profile analysis of body weight categories, structural equation models linking stigma and stress biology, and survival analyses of long-term health outcomes). ●Automating diagnostic and visualization steps for quality assurance. ●Preparing results for integration into manuscripts, grant proposals, and conference presentations. Support from a FADS intern will enable the creation of generalized, modular scripts that can be re-applied across projects in the research program. The intern will collaborate on data wrangling in R and Python, write functions to automate model fitting and diagnostics, and produce high-quality visualizations to communicate findings. This will free substantial time for conceptual development, manuscript writing, and mentoring, while increasing transparency and reproducibility in my work. This assistance will also catalyze new directions, including the development of analytic strategies for testing biological normalcy theory in large datasets and evaluating the physiological consequences of weight stigma across populations. The proposed pipelines will be directly applicable to upcoming grant submissions, including projects on stigma, stress biomarkers, and chronic disease risk.

Rationale for assistance in data analytics and visualization: The datasets used in this research are both methodologically complex and substantively sensitive. They include survey designs with stratification and weighting, multiple waves of biomarker and anthropometric data, and nuanced indicators of stigma, discrimination, and cultural beliefs. Analyzing these data requires advanced programming and analytic pipelines that extend beyond traditional statistical approaches. While I am proficient in Stata and experienced in anthropological and public health data analysis, building automated and scalable workflows requires dedicated technical expertise. Without targeted support, much time is lost to

manual wrangling and repetitive coding tasks, which slows the pace of discovery and limits the ability to broaden the scope of analyses. A FADS partnership will provide the technical expertise to transform my workflow. Assistance with programming, database structuring, and visualization will allow for reproducibility across multiple datasets and greater efficiency in handling analytic challenges. This support will enable more ambitious analyses, such as cross-cohort comparisons of weight stigma's effects on inflammation and metabolic outcomes, that would otherwise be infeasible due to technical bottlenecks.

Statement of benefit to the student: The MSDS student will gain hands-on experience applying data science methods to socially urgent questions in public health, stigma, and health disparities. This project offers opportunities to work directly with large, complex, and sensitive datasets that require advanced methods in data cleaning, harmonization, and visualization. The student will gain experience in managing longitudinal and multi-source data, a skillset that is highly relevant to careers in data science, biostatistics, or health informatics. The project also provides experience in interdisciplinary collaboration. The student will work with a principal investigator trained in anthropology and public health, and a broader research team spanning sociology, epidemiology, and biology. They will contribute to manuscripts, grant proposals, and visual outputs, and will see the direct application of data science to critical issues in health equity. Mentorship will emphasize skill-building, reproducibility, and professional development. Students will also have opportunities to be included as coauthors on subsequent manuscripts resulting from this work.

Specific competencies required, including programming languages if applicable: Specific Competencies Required The student should bring expertise in the following areas: ● Programming Languages: Proficiency in Stata, SAS, R, and Python for data cleaning, wrangling, statistical modeling, and visualization. ● Project Skills: ○ Data Visualization: Ability to produce publication-ready graphics using R (ggplot2, Shiny) or Python (matplotlib, Plotly). ○ Database Management: Structuring and harmonizing longitudinal and multi-wave datasets. ○ Statistics: Applying regression models, structural equation modeling, clustering/latent profile analysis, and survival analysis. ○ Machine Learning & Data Mining: Experience with multivariate and classification methods desirable. ○ Reproducible Workflows: Use of GitHub and/or R Markdown to ensure transparency and scalability. ○ Qualitative Data: Interest in computational approaches to qualitative and thematic textual analysis (preferred not required) ● Preferred Experience: Handling survey data with complex sampling weights (e.g., NHANES, Add Health, CCHS) and familiarity with health or social science applications.

Is there anything else you would like us to know about your project's time frame or work schedule?: This work is analyzing secondary data analysis

Proposal Title: #1012: Interpretable Deep Learning for Sleep Disorders: Informative-Region Detection in Functional Time Series

Department: School of Science, Department of Mathematical Sciences, Indianapolis

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Sleep disorders such as obstructive sleep apnea, insomnia, and hypersomnia are prevalent, heterogeneous, and underdiagnosed. We will leverage deep learning–based functional data analysis (DL-FDA) with informative-region detection to extract clinically meaningful structure from four complementary datasets: the NCH Sleep DataBank, Future of Families and Child Wellbeing Study (FFCWS), Sleep Heart Health Study (SHHS), and Cleveland Children's Sleep and Health Study (CCSHS). Our premise is that models that both predict and localize the signal segments that drive those predictions can accelerate discovery and translation in sleep medicine. Methodology. We will develop interpretable sequence architectures—1D CNNs/temporal convolutional networks (TCNs), transformer encoders, and multiresolution scattering–based hybrids—tailored to polysomnography (PSG) and actigraphy. Models will be trained primarily for sleep staging, with auxiliary heads for apnea-related outcomes. To exploit abundant unlabeled segments, we will use self-supervised pretraining (masked-segment prediction and contrastive objectives) followed by cohort-aware fine-tuning. Informative-region detection will combine temporal class-activation maps (t-CAM) for CNN/TCN models, attention rollout for transformers, and gradient-based attributions (integrated gradients/DeepLIFT) to produce contiguous region proposals. Visualization and evaluation. We will deliver a clinician-facing visual analytics toolkit that pairs accuracy with interpretability: • A synchronized raw-signal viewer (EEG/EOG/EMG/airflow/SpO_2/actigraphy) where users brush a time window to reveal local CAM/attribution values. • Time–frequency scalograms (STFT/wavelet) with highlighted informative regions and tooltips showing spectral energy and attribution. • Cohort-level "importance-density" heatmaps that aggregate where (by clock time or stage) informative regions concentrate in cases vs controls. • Model evaluation panels: calibration curves, ROC/PR, per-stage confusion matrices, threshold-sweep plots, and subgroup/fairness slices (age, sex, cohort). Aims. (1) Train and validate DL-FDA models that jointly predict sleep stage and apnea/disorder risk while localizing informative regions. (2) Quantify how localized patterns vary across age, sex, and cohort, and test whether region features add discrimination beyond stage labels alone. (3) Release a reproducible pipeline (tested code, attribution modules, and interactive dashboard) for IU investigators. Outcomes. Interpretable models linking stage-specific dynamics to disorder risk; stable, quantitative summaries of informative regions to guide hypotheses; and publication-ready visualizations and software assets that enable broader sleep-research use at IU Indianapolis.

Rationale for assistance in data analytics and visualization: This project hinges on scalable DL-FDA plus rich, clinician-friendly visualization. Assistance is needed to: (1) harmonize PSG/actigraphy signals and metadata across cohorts; (2) implement training pipelines with self-supervised pretraining, stratified cross-validation, GPU orchestration, and experiment tracking; (3) engineer informative-region detection (t-CAM/CAM, attention rollout, integrated gradients/DeepLIFT, Shapley attributions) with stability checks and unit tests; and (4) build interactive visuals that make localized evidence legible to non-

technical users. A dedicated data-science student will help construct reproducible modules in R/Python (PyTorch-Lightning, scikit-learn), curate cohort dictionaries, and develop dashboards (Shiny/Plotly Dash/Bokeh) featuring: linked hypnogram–signal viewers with region masks; time–frequency scalograms; attribution heatmaps; cohort-level importance-density maps; and evaluation panels (calibration, ROC/PR, per-stage confusion, threshold sweeps). The student will convert exploratory notebooks into tested packages, standardize figure templates (for manuscripts and talks), and ensure outputs meet publication-quality and clinical communication standards.

Statement of benefit to the student: The student will gain end-to-end experience building interpretable deep-learning systems for functional time series in health. Core competencies include: (i) signal processing for PSG/actigraphy; (ii) training and tuning 1D CNN/TCN/transformer models; (iii) informative-region detection (t-CAM/CAM, attention rollout, integrated gradients/DeepLIFT, Shap values) with stability assessment; and (iv) visual analytics design—linked views, brushing, tooltips, and uncertainty ribbons—implemented in Shiny or Plotly Dash. They will practice reproducible research (git, environment management, Quarto), privacy-aware data stewardship, and rigorous evaluation (cross-validation, calibration, subgroup fairness checks). Mentorship will include weekly meetings for code reviews, methods discussion, and milestones culminating in a dashboard demonstration. Professional outcomes include possible co-authorship on IU abstracts/manuscripts and deployable visualization components that will strengthen future IU sleep-research projects and the student's competitiveness for graduate study or data-science roles.

Specific competencies required, including programming languages if applicable: Cloud & High Performance Computing; Data Mining; Data Visualization; Database Management; Deep Learning; Internet of Things, sensors, wearables, gadgets; Machine Learning; Signal Processing; Statistics; Security / Privacy Management; Python; R; SQL

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1013: The Profile of Early Life Adversity and Psychosocial Well-being among Older Cancer Survivors

Department: Social Work, Social Work, Indianapolis

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: This pilot study aims to provide valuable real-world data on the prevalence and profiles of early life adversity among older cancer survivors and examine how different profiles of early life adversity are associated with their psychosocial well-being across the life course. Leveraging data from the Health and Retirement Study (HRS), a nationally representative cohort of older adults, this study has the following aims: Aim 1: To describe the prevalence and identify different profiles of early life adversity among older cancer survivors. Specifically, we will use descriptive statistics to report the prevalence of the early life adversity indicators and conduct latent class analysis (LCA) to identify different profiles. Aim 2: To examine how different profiles of early life adversity identified by LCA are associated with psychosocial well-being (i.e., anxiety, life satisfaction, purpose in life) among older cancer survivors over time. Specifically, we will use mixed-effects models to assess whether these profiles differentially influence psychosocial outcomes across the different time points. Aim 3: To explore potential multilevel protective factors (e.g., individual-, microsystem-, and exosystem-level) that may buffer the impact of different adversity profiles on psychosocial well-being. We will conduct moderation analyses examining the buffering effects of psychological resilience (individual), social support (microsystem), and neighborhood cohesion (exosystem) on the negative association between early life adversity and psychosocial well-being. Completion of this study will provide essential preliminary data to support a future National Cancer Institute (NCI) R21 application. Building on the identification of early life adversity profiles and protective factors, we aim to develop a trauma-informed, evidence-based intervention and evaluate its feasibility, acceptability, and preliminary impact in a randomized controlled pilot trial. We will submit 2-3 manuscripts to high-impact journals in aging and cancer, such as the Journal of Cancer Survivorship (IF=3.8) and the Journal of Geriatric Oncology (IF=2.7). We plan to present our study findings at a national or international conference, such as the International Society of Geriatric Oncology.

Rationale for assistance in data analytics and visualization: This study will use secondary longitudinal data from the Health and Retirement Study (HRS), a nationally representative survey of U.S. adults aged 50 and older. We will analyze data from the 2006–2020 interview waves, during which each respondent was surveyed multiple times. Based on prior research, the estimated sample size of cancer survivors in this dataset is approximately 3,747. The large size, repeated measures, and breadth of the HRS make it an excellent resource but also create challenges. Careful data cleaning and organization are needed to prepare the dataset for analysis, and specialized data analysis methods are required to achieve study aims. Assistance with data analytics will help ensure that the study findings are accurate, reliable, and clearly presented. We also need support in visualization. For example, graphics generated from latent class analysis (LCA) will help identify different profiles of early-life adversity, making findings easier to interpret and enabling results to be shared more effectively with audiences.

Statement of benefit to the student: As the mentor, I will provide individualized training and guidance throughout the project. Students will engage in the full research process, from conceptualizing research questions and operationalizing constructs to conducting analyses and interpreting results. Students will gain hands-on experience with advanced statistical methods, including latent class analysis, mixed-effects longitudinal models, and moderation analyses. Students will also develop proficiency in statistical software such as SPSS, SAS, Stata, or R, strengthening data analysis skills, and will learn to manage and analyze large secondary datasets, including data cleaning, coding, and handling missing data. In addition, students will build project management and communication skills that are essential for careers in research, academia, or applied settings. Finally, students will have the opportunity to co-author conference presentations and manuscripts, further enhancing their academic portfolio.

Specific competencies required, including programming languages if applicable: Students should be familiar with statistical data analysis methods, such as descriptive statistics and mixed-effects longitudinal models, and proficient in using data analysis software such as SPSS, SAS, or Stata. It would be preferable if the student also has experience working with large secondary datasets.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1014: Philanthropy Deserts: Identifying Mismatches between Philanthropic Funding and Community Need

Department: O'Neill School, NA, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Summer 2026

Summary of project: Are foundations distributing funds to communities proportional to their needs? Which communities have the greatest needs and receive the least amount of philanthropic funding? Philanthropic foundations distribute over $100 billion in grants each year, enabling them to address critical social needs. However, evidence suggests a mismatch between the distribution of funds and community needs. Our project will analyze county-level data on community needs and the distribution of 2.5 million grants awarded by foundations to identify "philanthropy deserts"—communities with the greatest needs that receive the least amount of philanthropic funding. To address these questions, we compiled data on 2.5 million grants distributed by foundations to grantees in 2023. Using these data, we constructed a foundation-grantee network dataset that maps the distribution of philanthropic funds throughout the U.S. nonprofit sector. At the same time, we compiled county-level data on community outcomes from government agencies including the USDA, HUD, HHS, and CDC that measure factors including food security, affordable housing, education outcomes, workforce training, substance use, and mental health. For this project, we are combining these data to identify mismatches between the distribution of foundation grants and community needs. This project will identify communities with the greatest needs that receive the least amount of philanthropic funding, providing valuable insights for funders, community leaders, and policymakers, and transforming how foundations allocate resources to address community needs. Data and Sources Foundations, Grants, and Grant Recipients [Source: Internal Revenue Service (IRS)] When foundations file their annual tax return (a Form 990), they need to list all of the grants they distributed in the prior year and include the grant amount, the recipient, and its address. In 2023, approximately 125,000 foundations distributed at least one grant and collectively they distributed 2.5 million grants for a total of over $100 billion in grants. I have extracted and cleaned all of these records and geocoded the location of each grant/recipient. These data will provide the basis for mapping and analyzing the geographic distribution of philanthropic funding. Community Characteristics [Source: U.S. Census Bureau] We will compile data on community characteristics from the U.S. Census including population distribution, demographic composition, and economic indicators. Grantees' Charitable Activity Area [Source: IRS Business Master Files] To identify each grant recipient's primary charitable activity area, such as education, healthcare, and housing, I obtained the grantees' National Taxonomy of Exempt Entities (NTEE) code from the IRS Business Master Files. These data will help identify the types of community needs the grants are allocated to address. Social Determinants of Life Outcomes [Source: Various] To measure community conditions, I will u

Rationale for assistance in data analytics and visualization: Data analytics and visualization expertise are needed to analyze the funder-grantee network we constructed and map the geographic distribution of charitable funding on top of the layers of the community's characteristics and its needs. Desert Identification: For each county, we will calculate standardized measures of community needs and the per capita amount of philanthropic funding it received. Regression analyses that control for county

characteristics including ethnoracial composition, population density, and economic indicators will identify counties that received less funding than predicted by their needs. Bias Analysis: We will examine whether funding patterns systematically disadvantage communities based on their 1) ethnoracial composition, 2) economic characteristics, 3) geographic proximity to foundations, and 4) historical patterns of disinvestment. Intervention Mapping: For identified desert areas, we will create comprehensive directories of nonprofits, assess their capacity and focus areas, and provide contact information to facilitate foundation connections.

Statement of benefit to the student: For all of my research assistants, I am committed to developing them as researchers and helping them achieve their career goals. This particular project is ready for the student step in and start working immediately. The project analyzes a commonly known field of organizations that are familiar to both experts and non-experts, and its scale will give the student experience working with big data and demonstrating its real-world applications. The project also will give the student mastery of a dataset that is known, valued, and used by people in both the public and private sectors. This project is straightforward such that the student could easily describe it to potential employers and explain their contributions and the value-added deliverables they generated. In addition, many components of this project will provide an intellectual and technical challenge that will stretch the student and enhance their problem solving capacity. Also, if I obtain additional external funding, I am open to hiring the student to continue working for me as a research assistant during the academic year.

Specific competencies required, including programming languages if applicable: Network Science/Analysis, Data Visualization, Python, and Alteryx

Is there anything else you would like us to know about your project's time frame or work schedule?: Thank you for continuing to make this program available. I've had great experiences working with the students, and they have advanced my research in meaningful ways, which has helped me obtain external funding. To coordinate scheduling, I prefer to have fellows who will be located in one of the U.S. time zones during their fellowship. Also, if possible, I welcome the opportunity to also have fellows during both the spring semester.

Proposal Title: #1015: Developing an App to Streamline Mosque Building Projects in the U.S.

Department: Eskenazi School of Art, Architecture + Design, Interior Design, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: MosqUS Design, the project I am seeking data science assistance for, is a website and mobile app that will act as an extension of a digital humanities project that is currently being developed on omeka.net. MosqUS Designs will help direct architects, designers and mosque community members through a community-engaged research-based design process that would streamline the building project, taking into consideration financial constraints, zip-code studies, community needs, pain points, navigating zoning challenges, project phasing, and future considerations. The website and mobile app will also connect mosque communities with architects who have experience building U.S. mosques as well as vendors, artists, and craftsmen who provide custom design services including art installations and wood carving of special design features. In doing so, this app bridges the gap between interior design research and practice, while also translating academic research to accessible and usable information for Muslim community members. Whether the mosque is an adaptively reused building, makeshift, storefront, or purpose-built, MosqUS Design is going to assist decision makers and stakeholders in achieving their goals regardless of their needs. This application is based on years of research on mosque designs and Muslims place making in the United States. My previous research proposed a framework for holistically studying U.S. mosques, considering the internal, local, regional, and global dimensions that impact the development of mosques based on a review of all publications on U.S. mosques before 2022, highlighting gaps and strengths. Other projects have also explored the experience of minority groups in mosques and third spaces, providing recommendations on creating inclusive places for Muslims. In addition, the ongoing MosqUS digital humanities platform, is a living repository of oral history interviews, architectural drawings, photos, and public records on the founding and design of mosques across the United States, documents the history of place making practices in the U.S. Together, these projects shape my research trajectory that culminates in an impactful public-facing version of this work. This project is particularly significant as mosque numbers have increased significantly between 2010 and 2020, bringing mosques to more than 3,000 facilities in the U.S. These spaces will continue to grow to accommodate the doubling population of U.S., as projected by the PEW Research Center (2017). Because Muslim communities are geographically spread and not every town has a local mosque, this app will guarantee that knowledge from precedent mosque projects are used to develop best practices for future mosques which will preserve community resources and aid in community development, an area that is unexplored in current interior design and architecture scholarship.

Rationale for assistance in data analytics and visualization: While I have substantial experience in design and qualitative research methodologies, I acknowledge that my expertise in the specific data science techniques that could transform the collected data and narratives to practical guidelines in the thumbs of communities is limited. Visualizing the data in an interactive web/app form is where the student's assistance becomes invaluable. The student's proficiency in text mining, data visualization and app development will allow the systematic extraction and meta-analyses of data from the digital

humanities website to come up with specific insights that would then be turned back to me to develop digestible guidelines for the users. Taking these guidelines, the students will develop an app that visualizes them in an accessible user-interface. The student's expertise in these areas will significantly contribute to the accessibility, quality, and visualization of the research. The student's involvement will yield a tangible product, the app and website, that will help bridge the gap between design theory, data-driven insights, and practice, paving the way for healthier and more community-driven mosque environments in the U.S. Their ability to navigate and conduct a preliminary synthesis of the data will facilitate the development of the design process and will later shape its presentation through the interactive website and app.

Statement of benefit to the student: Interning with a faculty member on a research paper exploring the design of mosques in the United States presents an exceptional opportunity for a data science student to enrich their professional skill set and communication abilities. This internship expects multifaceted benefits. Firstly, the student will apply their data science skills, particularly app development and data visualization, to a real-world context. Analyzing existing data will provide them with practical experience, honing down their ability to extract meaningful patterns from data and visualize them through UX best practices. Secondly, by delving into the architecture and design fields, the student will gain expertise in a domain they may not have explored otherwise, broadening their knowledge and problem-solving capabilities and communication skills. Given the topic of exploration, I expect that working on the project will also enhance the student's awareness of a minority group in the U.S. Furthermore, while the study design is developed and the majority of the data is mostly collected and ready for analysis, the student will be empowered to provide guidance on the technical side of the analysis. I expect that through this interaction students will build their leadership and collaboration skills. In sum, this internship presents an exciting opportunity for a data science student to grow professionally, culturally, and academically.

Specific competencies required, including programming languages if applicable: Data Mining, Data visualization, Web Front-end Development, and Smartphone App Development.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1016: Distance from the tip of the nose to the base of the tail

Department: Science, Biology, Indianapolis

Project requires Human Subjects? No

Project requires Laboratory Animals? Yes

Project requires Biosafety Review? Yes

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Down syndrome (DS or Trisomy 21) affects approximately 1 in 700 live births and individuals with DS experience learning and memory deficits, skeletal abnormalities, and altered gait. Mouse models of DS have been used to further understand the underlying genetic causes of DS-related traits. To understand the effects of triplicated human chromosome 21 genes on cerebellar development and gait, we are assessing the body length, stride length, stance width, and angle of step of Ts65Dn DS mice as compared to normal euploid littermates. We hypothesize that trisomic as compared to euploid animals will have smaller body length and size, and increases in gait stance width and angle. Preliminary data support our hypotheses on stance width and step angle. Because this is a tedious process, we are looking to automate the evaluation of body length, stride length, stance width, and angle of step in our mouse models. We envision that combinations of computer vision and machine learning AI can be used to perform much of the scoring of these variables. For testing we take a five-minute video of each animal from below using a plexiglass corridor and a Logitech camera, making sure a ruler is visible in the video. To manually quantify the gait parameters, we then go through each video to identify points at which each animal successfully walked across the corridor without stopping or turning (4 or more consecutive straight-line alternating steps with the hindlimbs), identifying a minimum of three time points per animal marking the points of each successive step in which the hindpaw is in contact and flat on the floor of the corridor. We select these points by inspecting the video frame-by-frame, taking screen captures of each successive step in these periods of uninterrupted movement. We overlay the selected frames of three consecutive steps for a given traversal using Photoshop, then use ImageJ software to measure the stride length, stance width, and step angle for each step during the period of constant movement. Each measurement is taken from the point between the first and second digits of each hind foot, which is typically distinguishable in the captured images from the video. Body length (distance from the tip of the nose to the base of the tail) measurements are taken from at least six separate points within the video and averaged. Stride length is the distance a single foot moves between two steps. The stance width is the distance between the two feet across two steps. The step angle is the angle of the three positions of the feet across two steps. We have over 110 videos of different animals where gait needs to be quantified. We hypothesize that combinations of computer vision and machine learning AI can be used to automate much of the scoring of these variables. The outcomes of this project will allow us to better understand how changes in the cerebellum of DS mouse models affect gait alterations and eventually lead to therapies to improve the lives of indiv

Rationale for assistance in data analytics and visualization: Our project with over 110 videos with well defined points of collection would benefit greatly from automated quantification using computer vision and machine learning AI. We have scored a complete set of separate videos without AI help, and it is extremely tedious to do. Assistance with visualization would establish a protocol that we and others in the DS field will use to quantify changes in gait in mouse models. Our present manually scored set of data could be used to compare the effectiveness of the automated scoring. Additionally, with the

publication of this new methodology, other individuals in various fields that are interested in gait analysis would be able to use this methodology.

Statement of benefit to the student: . We believe that the student would benefit from engaging with individuals in our laboratory to learn more about Down syndrome and the research that we do. This engagement may lead to a better understanding of the life sciences field and opportunities to engage with life scientists as future employment is sought. Our laboratory has weekly lab meetings, and bi-weekly zoom meetings with students to talk about DS and the progress in our work, that this student would be invited to join. We envision that this work, if successful, will lead to co-authorship on a developmental biology manuscript, and a methods paper to be submitted for peer review and will contribute to future empirical studies in DS mouse models. The outcomes of this project will allow us to better understand how changes in the cerebellum of DS mouse models affect gait alterations and eventually lead to therapies to improve the lives of individuals with DS. This is an exciting field for a student to be associated with.

Specific competencies required, including programming languages if applicable: computer vision and machine learning

Is there anything else you would like us to know about your project's time frame or work schedule?: We have the data from this project already collected. I would prefer a student in Spring, but would also welcome a student in Summer if that is not available.

Proposal Title: #1017: ArchIvory: A Database for Understanding the Science and History of Ivory

Department: College of Arts and Sciences, History, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: My team is creating an open-access database that can rapidly, non-destructively, and affordably identify the provenance of elephant ivory. Trained in Chinese and Mongolian environmental history, I have spent my career researching texts in East Asian and European archives. My recent work on China and the global ivory trade, however, revealed gaps in the textual record: documents failed to show where the ivory circulating in China came from in different eras. Beyond texts lies an alternative archive—museum collections—large enough to reconstruct, with unprecedented precision, the evolution of human-elephant networks across time. Museums hold an astonishing variety of historical ivory carvings. Unlocking the data within them requires a non-destructive method of detecting biochemical signatures in ivory, and a database that can correlate those signatures to specific elephant populations. In partnership with a zooarchaeologist and an elephant scientist, we are building the world's largest open-access database of ivory objects, integrating curatorial metadata, high-resolution images, and scientific analyses (DNA, isotopes, proteomics, XRF). With over 50,000 records already compiled and a target of 100,000, the database—called ArchIvory—will allow researchers and students to track the global circulation of ivory across time and see how heirloom objects connect to larger historical patterns. I am seeking student assistance with two components: (1) using the database to identify and visualize patterns of change in ivory objects, and (2) developing and refining image and text recognition tools to support public and scholarly use. For the first, students will analyze metadata to detect patterns in ivory carving—such as shifts in forms, workshops, trade hubs, and stylistic conventions—and help design visualization tools that make these patterns legible to researchers and the public. This work requires skill in data cleaning, exploratory analysis, and time-series visualization. Outputs will include interactive graphics showing how ivory fans, figurines, or other objects proliferated in different regions at different moments. For the second, students will join our team developing an object-recognition and recommendation engine. We aim to create a portal where individuals can upload photographs and short descriptions of heirloom ivory objects. The system will return predictions about where and when the object was carved, as well as recommendations of visually and contextually similar museum objects. Students will refine recognition pipelines by curating training data, testing deep learning models, and improving natural language processing for short curatorial descriptions. They will also help design a recommendation algorithm that integrates both modalities to suggest relevant comparisons.

Rationale for assistance in data analytics and visualization: The ArchIvory database has reached a scale where advanced computational methods are required to unlock its potential. With tens of thousands of ivory objects spanning centuries and continents, it is no longer feasible to rely solely on manual inspection to identify historical trends. Data analytics are essential for detecting patterns—such as changing styles, geographic clustering, or sudden surges in certain object types—that can illuminate the dynamics of the ivory trade and its environmental impact. Equally important, the image- and text-

recognition component demands machine learning expertise. Developing a recommendation engine that can return “nearest neighbors” across large museum datasets requires skills in training and testing neural networks, embedding multimodal data, and optimizing model performance. Student assistants trained in data science are therefore critical to the project’s next phase. Their work will also provide crucial groundwork for forthcoming applications to the National Endowment for the Humanities (NEH) and the National Science Foundation (NSF). Both programs emphasize interdisciplinary teams and proof-of-concept outputs. Students will help us produce precisely those pilots—visualizations, prototype recognition pipelines, and recommendation algorithms—that will make ArchIvory competitive for large-scale external support.

Statement of benefit to the student: The project affords students the opportunity to work in an interdisciplinary team and to participate in the team’s conversations about how best to make sense of our data. We want students who will take an active interest in the project. We do not expect all the applicants to be equally strong in all of the above-mentioned competencies. We hope, though, that will use their experience on the team to hone all of the above skills, whether it be database management, visualization of data for non-experts, or analysis of complicated datasets—skills that would serve them well academically and professionally moving forward. If interested, we welcome students to contribute to publications that will come out of this project, and, if possible, to continue working on the team past the spring. Additionally, students can strengthen key skills through the project through learning deep learning frameworks, advanced visualization, and multimodal machine learning, all while developing the ability to contextualize computational results within historical, archaeological, and museum-based frameworks, and collaborating effectively with humanities scholars.

Specific competencies required, including programming languages if applicable: • Programming (Python): Strong proficiency with Python, including experience using pandas and scikit-learn for data cleaning, exploratory analysis, and basic machine learning. • Data handling: Ability to work with large, messy datasets (CSV, JSON, API pulls); basic familiarity with SQL for structured queries. • Visualization: Competence with matplotlib and Plotly for producing clear exploratory and explanatory graphics. • Machine learning fundamentals: Coursework or prior exposure to classification, clustering, and embeddings; ability to apply pre-trained models for natural language processing or image recognition tasks.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1018: Building a global visualization platform for tracking floating solar energy growth and environmental tradeoffs

Department: O'Neill School of Public and Environmental Affairs, N/A, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: The deployment of floating solar photovoltaic (FPV) panels on the surface of water bodies—especially reservoirs and other artificial systems—is burgeoning worldwide. This trend is driven by two forces: global commitments to replace fossil fuels with climate-friendly renewables like solar power and the fact that solar requires large areas of land, which can create conflicts in densely populated or land-scarce regions. Leveraging the underutilized surface of reservoirs offers a practical solution to alleviate these conflicts because reservoirs are everywhere – over 4.4 million have been identified worldwide, covering close to 400,000 km2, an area roughly the size of Germany. However, installing solar panels across large portions of a reservoir reduces the amount of light that enters the water. Light is the most fundamental driver of aquatic ecosystems, shaping oxygen production, temperature patterns, and ultimately food web dynamics. It remains poorly understood how much surface coverage is enough to cause major ecological changes, or what thresholds might disrupt ecosystem health. My lab is addressing these knowledge gaps through globally relevant research published in leading journals such as Nature and Nature Sustainability. Currently, a postdoctoral researcher in my group (Dr. Aline Valério), with expertise in remote sensing, is leading a project with two goals: (1) to track the pace and geography of global growth of FPV using satellite imagery, and (2) to identify coverage levels that trigger ecosystem change using state-of-the-art remote sensing analysis. So far, we have identified more than 1,000 FPV sites worldwide. These sites are now being analyzed in time series to retrieve key water quality parameters, such as chlorophyll-a (a proxy for algae growth, which depends on light) and water temperature (which responds directly to shading). These metrics will be used to quantify both immediate and cumulative ecological responses to shading. Our next step is to create the world's first dynamic, open-access inventory of FPV installations through an interactive visualization tool. This platform will display the location of FPV systems and their associated environmental tradeoffs, accompanying a manuscript currently in preparation. It will support interdisciplinary research, enable long-term monitoring, and inform evidence-based decision making for low-conflict FPV deployment. This is where the FADS program can make a decisive contribution. We are seeking support from a data science student to help develop the visualization tool, working closely with me and the project's postdoctoral researcher. The student's expertise in data visualization, data analysis, and tool development could transform our dataset into an accessible and impactful platform for both science and practice. The tool will feature an interactive map with dropdown site selection, integrated data layers, and export options.

Rationale for assistance in data analytics and visualization: Our lab has strong expertise in aquatic ecology, remote sensing, and the environmental dimensions of renewable energy. We have a project well underway where we use freely available satellite imagery archives and are conducting time-series analyses to track deployment trends and changes in water quality indicators. Our database under construction includes attributes such as geolocation of FPV systems, area of the water body, area of the

FPV arrays, year of installation, type of water body, and trends in chlorophyll and temperature (percent change relative to baseline). Where we lack capacity is in translating these results into an accessible, interactive platform. The development of a dynamic visualization tool requires specialized skills in dashboard design and scalable web-based visualization that extend beyond the expertise of our environmental science-focused team. Without this assistance, dissemination of the work will remain limited to static figures in publications, constraining the broader impact of the project. Engaging an MSDS student through FADS will provide the technical expertise needed to build a global, open-access visualization tool. This tool will not only enhance the reach of the current project but will also establish reusable infrastructure for monitoring the future expansion and ecological tradeoffs of FPV.

Statement of benefit to the student: Participation in this project will give the student an opportunity to apply their technical skills to pressing environmental and energy challenges. They will engage directly with large-scale satellite datasets, developing interactive maps, visualizations, and automation workflows that make remote sensing outputs accessible to diverse users. In doing so, they will gain practical experience with platforms such as Google Earth Engine and NASA Earth Data, as well as programming tools. The student will join an internationally recognized research group that is spearheading global work on floating solar energy and aquatic ecosystems. This environment will provide mentorship in both data science and interdisciplinary collaboration. Importantly, the student's contributions will be recognized through co-authorship in peer-reviewed publications under development, enhancing their academic credentials and providing a tangible product of their efforts. The PI has successful experience collaborating with computer scientists, having worked closely with Institute for Computational Sustainability at Cornell University as a postdoctoral scientist (outputs include publications in Nature Communications, Science, and a visualization tool accompanying those papers: https://www.cs.cornell.edu/gomes/udiscoverit/amazon-ecovistas/).

Specific competencies required, including programming languages if applicable: Specific project skills essential for successful implementation include data visualization, database management, data mining, web front-end development, and cloud & high-performance computing. Programming language and tools that may be applied include Python, R, SQL, and JavaScript.

Is there anything else you would like us to know about your project's time frame or work schedule?: I indicated Spring 2026 because my postdoc has funding until the end of 2026. The MSDS student can then work together with her during the Spring, giving enough time to wrap up associated publications.

Proposal Title: #1019: Ensuring Medical Device Safety with AI

Department: Kelley School of Business, Operations and Decision Technologies, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: The process for clearing new medical devices for market, particularly through the FDA's 510(k) pathway, relies on a manufacturer's ability to demonstrate substantial equivalence to a previously cleared device, known as a predicate device. This system, while intended to be efficient, faces significant challenges. Current studies have highlighted a critical vulnerability: the broad discretion given to applicants in selecting their predicate device can lead to suboptimal outcomes. This can result in device lineages built on outdated technology or, in some cases, on recalled predecessors, which poses potential risks to public health. The lack of an objective, data-driven approach in predicate selection allows manufacturers to potentially choose a convenient predicate rather than one that is truly representative and technologically sound. In response to this need, this project proposes a comprehensive framework that harnesses data and algorithms to improve the predicate selection phase of device clearance. We will develop machine learning (ML) and large language model (LLM) approaches to build a system that can algorithmically identify a robust and representative set of similar predicate devices. This framework will act as a smart safety net, ensuring that both regulators and manufacturers consider the full network of relevant prior devices. By analyzing the technical specifications, performance data, and regulatory history of existing devices, our algorithms will move beyond a simple "one-to-one" comparison and provide a more holistic view of a device's place within the market. Furthermore, we will create an interactive tool that complements our algorithmic framework. This tool will allow FDA reviewers and manufacturers to upload a PDF of a device's summary document. Upon submission, the system will process the document and provide a visual representation of the network of similar predicates. This graphical interface will not only display a list of potential predicates but also show their relationships and historical context, providing a transparent and comprehensive view that can aid in more informed and safer regulatory decisions. By providing this data-integrated, interactive support system, our project aims to enhance the efficiency, safety, and integrity of the medical device clearance process.

Rationale for assistance in data analytics and visualization: Assistance with data analytics, machine learning, LLM Prompt Engineering, and an interactive tool design in Python is crucial for the success of this project. The FDA's device clearance data is vast and complex, encompassing millions of regulatory documents, technical specifications, and historical records. Collaborating with graduate students in data science will allow us to effectively manage and analyze this large dataset. By leveraging advanced data analytics and machine learning, we can extract meaningful patterns from the data, identifying connections and chains of lineage that are not apparent through manual review. This analytical approach is vital for building the predictive models that will power our framework. Furthermore, data visualization is key to making our findings understandable. Creating a visual, interactive tool will demonstrate the project's potential impact, which is essential for securing future funding and encouraging adoption by regulatory bodies and manufacturers.

Statement of benefit to the student: Participating in this project offers students a unique opportunity to contribute to enhancing the safety and integrity of the medical device clearance process. By being directly involved in the development of a data-integrated decision support framework for the FDA, students will gain hands-on experience using machine learning and natural language processing to analyze complex regulatory data. Through this experience, students will enhance their skills in predictive modeling, data science, and developing interactive tools, all of which are highly sought after in healthcare, public policy, and technology sectors. They will also have the chance to work with real-world regulatory data, providing them with invaluable exposure to the practical applications of data-driven policy development and safety monitoring. This project will help students develop skills that directly address significant public health challenges.

Specific competencies required, including programming languages if applicable: Python, Machine Learning, Interactive Tool Design in Python

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1020: Bidirectional Influences Between Social Media Use and Mental Health

Department: Kelley School of Business, Marketing, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Adolescent social media use and mental health are dynamically interlinked, yet existing research often falls short in capturing the nuances of this relationship. Most studies rely on cross-sectional or retrospective designs that focus solely on the amount of time spent online, rather than examining the specific content being consumed. Moreover, it is clear that mental health is not only influenced by online exposures but also shapes subsequent online behavior, creating a complex feedback loop that can either amplify risk factors or promote resilience. To effectively address these bidirectional dynamics, it's essential to employ ecologically valid and high-frequency measures that capture the intricate interplay between social media use and mental health. Our long-term goal is to create a cutting-edge system for real-time monitoring and analysis of adolescents' social media exposure and mental health, enabling tailored prevention and intervention strategies.

Rationale for assistance in data analytics and visualization: We need data science students to capture and analyze social media data using NLP and machine learning methods to quantify the content of social media. We will also develop a web application or APP to track users' social media use.

Statement of benefit to the student: Students will learn the application of data science and computer science skills in solving business problems, which will help them stand out in job applications from other students with no exposure to business applications. Outstanding students will receive Professor Li's recommendation letter and opportunities for paid RA or TA positions.

Specific competencies required, including programming languages if applicable: Machine Learning, Statistics, Natural Language Processing, Web & Social Media Mining, Data Visualization, Web Front-end Development, Smartphone App Development (plus but optional); Python, Excel, R or SPSS

Is there anything else you would like us to know about your project's time frame or work schedule?: I have worked with FADS students in the past two years and students were successfully placed to leading employers such as Amazon. Students spoke highly of the benefit of this program.

Proposal Title: #1021: AI-Driven Nowcasting of Extreme Precipitation Using Geostationary Satellite Data

Department: The College of Arts & Sciences, Geography, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: The prediction of extreme precipitation events (e.g., heavy rainfall) has become increasingly critical in the face of climate change and its associated impacts on weather patterns. As global temperatures rise, the frequency and intensity of extreme precipitation events are expected to increase, posing significant risks to human life, infrastructure, and ecosystems. In this context, the ability to predict such events with high accuracy and lead time is crucial for effective disaster preparedness and management. A growing body of literature addresses various methods of nowcasting extreme precipitation in near real-time. According to the World Meteorological Organization (WMO), nowcasting refers to forecasting that yields local weather details across the mesoscale and small scale, covering a period from the present up to 6 hours ahead. The nowcasting algorithms have evolved tremendously over the past few decades, driven by advances in remote sensing, computational resources, and algorithm innovation. Extrapolation-based methods form the foundation of precipitation nowcasting. These methods rely on the assumption that the motion of precipitation patterns in the immediate future can be inferred from their recent movement. The key advantage of extrapolation-based methods lies in their computational efficiency and the ability to produce reliable forecasts over short lead times (i.e., in the order of tens of minutes). However, they do not incorporate the physical dynamics of the atmosphere to predict future developments, which limits their accuracy in capturing evolving weather systems and decreases their prediction skill rapidly with increasing forecast length. NWP models simulate atmospheric processes using fundamental physical equations, allowing them to capture the initiation, growth, and decay of precipitation systems. A common way for NWP-based methods is to assimilate observational data (e.g., radar imagery) into NWP models, as it offers three-dimensional high-resolution observations of the atmosphere at the convective scale. However, NWP models still face challenges in nowcasting due to their computational complexity, inability to capture small-scale variations (e.g., convective storms with scales less than ~30 km), and scarcity of surface observational network. With the rise of big data and artificial intelligence, machine learning (ML) and deep learning (DL) techniques have emerged as powerful tools for precipitation nowcasting. Unlike NWP models, which are based on physical principles, AI-based nowcasting methods learn patterns from historical weather data, allowing them to capture the nonlinearities and complexities of precipitation processes without making strong assumptions about the physical system. This project will evaluate and develop AI-based nowcasting methods based on geostationary satellite (GOES-R), comparing with existing nowcasting algorithms, including a simple extrapolation method, a Spectral Prognosis (S-PROG) model, and a S

Rationale for assistance in data analytics and visualization: This project requires advanced data analytics and visualization support due to the complexity and novelty of the proposed research. Extreme precipitation nowcasting involves integrating multi-temporal high-frequency geostationary satellite

imagery (GOES-R). Processing this data in near real-time demands expertise in high-performance computing and advanced geospatial analytics pipelines (e.g., reprojection). Moreover, evaluating AI-based nowcasting methods requires sophisticated statistical analyses, including error decomposition, probabilistic skill assessment, and comparative benchmarking against established methods such as extrapolation, S-PROG, and STEPS. These tasks necessitate robust data management, efficient algorithm implementation, and reproducible analytical frameworks. Equally important is the need for effective visualization of complex spatiotemporal data. Extreme precipitation forecasts must be communicated not only in scientific publications but also in forms accessible to policymakers, disaster managers, and broader stakeholders. High-quality visualization, such as interactive dashboards, uncertainty maps, and time-lapse animations, can enhance interpretability, highlight actionable insights, and bridge the gap between technical outputs and end-user needs. Assistance in data analytics and visualization will ensure that the project's methodological innovations are matched by clarity in communication, ultimately maximizing the scientific, societal, and policy rel

Statement of benefit to the student: Participation in this project will provide the student with significant academic and professional development opportunities in the growing fields of data science, climate analytics, and geospatial visualization. The student will gain hands-on experience working with state-of-the-art geostationary satellite datasets (e.g., GOES-R), learning how to preprocess, manage, and analyze near-real-time data streams using advanced computational techniques. Through exposure to novel machine learning and deep learning models for precipitation nowcasting, the student will develop technical expertise in model implementation, evaluation, and comparison against traditional methods such as extrapolation and ensemble-based forecasting systems. In addition, the student will acquire skills in advanced visualization, including creating interactive dashboards, uncertainty maps, and animations to communicate complex spatiotemporal forecasts to diverse audiences. These experiences will not only strengthen the student's analytical and technical capabilities but also enhance their ability to translate data-driven insights into actionable knowledge for policymakers and practitioners in disaster risk reduction. By contributing directly to a high-impact research project, the student will build a strong foundation in interdisciplinary research, preparing them for graduate study or careers in atmospheric science, remote sensing, or data science applications in real-world issues.

Specific competencies required, including programming languages if applicable: Cloud & High Performance Computing, Data Visualization, Deep Learning, Machine Learning, Statistics, and Web Front-end Development

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1022: Time Series Analysis of the Influences Between Social Media Use

Department: Arts and Sciences, Economics, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Summer 2026

Summary of project: Adolescent social media use and mental health are dynamically interlinked, yet existing research often falls short in capturing the nuances of this relationship. Most studies rely on cross-sectional or retrospective designs that focus solely on the amount of time spent online, rather than examining the specific content being consumed. Moreover, it is clear that mental health is not only influenced by online exposures but also shapes subsequent online behavior, creating a complex feedback loop that can either amplify risk factors or promote resilience. To effectively address these bidirectional dynamics, it's essential to employ ecologically valid and high-frequency measures that capture the intricate interplay between social media use and mental health. Our long-term goal is to create a cutting-edge system for real-time monitoring and analysis of adolescents' social media exposure and mental health, enabling tailored prevention and intervention strategies.

Rationale for assistance in data analytics and visualization: We need data science students to capture and analyze social media data using NLP and machine learning methods to quantify the content of social media. We will also develop a web application or APP to track users' social media use.

Statement of benefit to the student: Students will learn the application of data science and computer science skills in solving business problems, which will help them stand out in job applications from other students with no exposure to business applications. Outstanding students will receive Professor Li's recommendation letter and opportunities for paid RA or TA positions.

Specific competencies required, including programming languages if applicable: Machine Learning, Statistics, Natural Language Processing, Web & Social Media Mining, Data Visualization, Web Front-end Development, Smartphone App Development (plus but optional); Python, Excel, R or SPSS

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1023: Expanding Healthcare Access through Drone Technology and Its Welfare Effects on Rural Households

Department: O'Neill School of Public and Environmental Affairs, O'Neill School of Public and Environmental Affairs, Indianapolis

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: In many low- and middle-income countries, weak road networks and unreliable supply chains impede timely access to essential medicines and vaccines. Unmanned aerial vehicles (UAVs, or "drones") have emerged as a promising logistics innovation capable of delivering temperature-sensitive products rapidly to remote facilities. This project examines the causal impacts of drone-enabled delivery on public health and socioeconomic welfare in Ghana. While the emerging literature documents operational gains, such as reductions in stock-outs and increases in the variety of medical products available in rural clinics, we still lack evidence on whether, and through which channels, these supply-chain improvements translate into broader population welfare. The need for timely access is acute: a national assessment of health-care accessibility found that 45% of surveyed households cited poor road conditions as the principal barrier to reaching health centers. Our study focuses on Zipline's drone distribution model, which delivers blood, vaccines, anti-venoms, antibiotics, and other time-critical supplies to rural facilities where conventional transport is slow or unreliable. By shortening delivery times and strengthening cold-chain reliability, drone services have the potential to affect not only clinical outcomes but also household decisions that depend on health status. Building on established links between health and human capital accumulation, we structure the analysis around three outcome domains. First, we estimate direct effects on child and maternal health, concentrating on conditions that require prompt treatment or refrigerated inputs. Second, we assess indirect effects on two pillars of socioeconomic development that are sensitive to health: schooling and agricultural productivity. Improved child health may reduce absenteeism and raise school attendance and progression, while healthier adults may reallocate time toward farm labor and management, potentially increasing productivity. Third, we study local labor-market adjustments generated by the opening and operation of drone distribution centers. These facilities and associated residential camps employ pharmacists, engineers, operators, cooks, and maintenance staff, creating nonfarm jobs that may influence employment composition and seasonal migration decisions among nearby households, including farm workers who typically migrate to cities during the dry season. Taken together, the project fills a critical gap by moving beyond logistics performance metrics to evaluate population-level consequences of drone delivery. By quantifying effects across health, human capital, productivity, and employment, the study will clarify whether drone-enabled supply chains generate measurable welfare improvements and under what conditions these gains materialize.

Rationale for assistance in data analytics and visualization: We seek a graduate data-science assistant to build a clean, analysis-ready panel linking baseline and follow-up data for the impact evaluation. Much of the raw data and documentation (including partially cleaned files) are already assembled; the assistant will complete harmonization, implement reproducible pipelines, and generate clear, policy-

relevant descriptives and visualizations. Scope of work • Review documentation for the Ghana Socioeconomic Panel and become familiar with the existing, partially cleaned datasets and code. • Harmonize and finalize variable construction for the most recent follow-up, ensuring consistency across waves (IDs, units, missing-data protocols). • Audit and update existing Stata code; refactor into a transparent, reproducible workflow (modular .do files, organized folders, log files). • Produce a data dictionary/codebook and a concise README describing sources, transformations, and quality checks. • Create descriptive statistics and visualizations (baseline balance checks, trends, and key outcomes by treatment/exposure and geography). • Maintain a version-controlled repository (GitHub) and communicate progress in regular check-ins with the research team.

Statement of benefit to the student: This project offers an exceptional applied learning experience at the frontier of public-interest technology. The student will work on an innovative drone (UAV) delivery program in Ghana that rapidly transports blood, vaccines, and other time-critical medical supplies to remote clinics. As part of a FADS fall or summer internship, the MSDS student will develop advanced data analytics, visualization, and database skills while building a clean, analysis-ready panel from household and facility data and producing policy-relevant figures and dashboards. The work is ideally suited to the FADS mission: the student will design reproducible pipelines (version control, documentation), implement quality checks, harmonize survey waves, and generate descriptive analyses that will guide a rigorous impact evaluation. The internship provides close mentorship from a cross-disciplinary team (economics, data science, public health), regular feedback on code and design choices, and exposure to the research cycle, from research questions to analytic outputs used by decision-makers. Because the project evaluates a novel logistics technology with meaningful social implications, it offers a rare chance to connect technical skills to real-world outcomes. Strong performance may lead to extended collaboration and recognition (e.g., acknowledgments or co-authorship, as appropriate).

Specific competencies required, including programming languages if applicable: • Excellent Stata programming (data reshaping, merges/appends, labeling, efficient loops, QC). • Survey data harmonization experience (cross-sectional and panel). • Data visualization skills (publication-quality figures of descriptive trends). • Database management and documentation (codebooks, READMEs, reproducible pipelines). • Applied statistics for descriptives and diagnostics (balance tables, missing-data audits). • GitHub for version control (branching, pull requests, issues) highly desired. (Familiarity with R/Python for visualization is a plus, but Stata will be the primary environment).

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1024: Cultural Indicators

Department: Media School, Communications, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Cultural Indicators is a project that is using AI to code culturally relevant information from US TV programs. It is an extension of a data collection effort that began in 1967. Topics analyzed include violence on TV, sexrole representation, race and other issues.

Rationale for assistance in data analytics and visualization: We've had success with two previous FADS cohorts, and would like to continue that trajectory.

Statement of benefit to the student: Students have enjoyed working on the project, they have been able to add it to their CVs ad publicize some of its outputs on social media.

Specific competencies required, including programming languages if applicable: R, Python, AGI

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1025: Measuring Uncertainty in Entrepreneurial Ecosystem

Department: Kelley School of Business, Business Economics and Public Policy, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: My research examines the economic mechanisms underlying entrepreneurship mentorship by analyzing detailed discussion transcripts from one of the world's leading startup accelerators. This accelerator operates unique mentorship sessions where experienced entrepreneurs and investors provide guidance to early-stage ventures, creating an ideal laboratory for studying how expert uncertainty affects strategic decision-making and venture outcomes. I investigate three fundamental research questions in entrepreneurship economics: (1) How does mentor uncertainty predict strategic pivoting recommendations? (2) Do different types of uncertainty, technical feasibility versus market viability concerns, have different effects on venture success? (3) What is the causal effect of uncertainty-driven strategic pivots on venture outcomes? These questions address core theoretical debates about the value of expert advice, the role of strategic flexibility in entrepreneurship, and the mechanisms through which mentorship creates value. My dataset comprises panel data on technology ventures across multiple cohorts, including transcribed mentorship discussions, detailed venture characteristics, structured advice, and long-term outcomes (program graduation, fundraising success, ...). Initially, I plan to investigate correlations between mentor uncertainty language and strategic pivoting behavior. The project require initial data tasks. First, measuring uncertainty requires advanced natural language processing beyond simple keyword counting. I need contextual understanding, semantic analysis, and machine learning classification of uncertainty types. Second, the high-dimensional nature of text data requires dimensionality reduction techniques. Third, data mining and developing ML based predictive models is needed for pattern discovery in the data. The broader impact extends beyond academic research. Findings will inform the design of entrepreneurship programs, mentor training protocols, and entrepreneur decision-making frameworks. The computational methods developed will be applicable to other contexts including medical consultations, financial advising, and policy deliberations. One of the contributions of this project will be introducing a data driven way to measure uncertainty from text data.

Rationale for assistance in data analytics and visualization: The availability of digital data and advancement of computational methods have fundamentally transformed the landscape of economic research, enabling investigation of previously intractable questions about human behavior and market mechanisms. Large-scale text datasets now provide insights into decision-making processes, information transmission, and strategic interactions that were historically unobservable to economists. Simultaneously, machine learning techniques, natural language processing, and advanced statistical methods have evolved to extract meaningful economic insights from these high-dimensional, unstructured data sources. This enables scholars to measure previously latent concepts such as uncertainty, sentiment, and attention at scale. In my research, Natural Language Processing is essential for uncertainty measurement, sentiment classification, and uncertainty type identification using transformer models and deep learning architectures. Machine Learning techniques are necessary for

building predictive models of venture success using high-dimensional text features, and developing classification algorithms for uncertainty categorization. Data Visualization is crucial for exploring complex relationships between discussion patterns, strategic changes, and outcomes across time, ventures, and mentor networks through interactive dashboards and data visualizations. Database Management expertise is needed for efficiently handling large-scale text datasets.

Statement of benefit to the student: This project offers exceptional learning opportunities. The student will gain hands-on experience with real-world business data, working with actual venture outcomes and high-stakes entrepreneurial decisions. Technical skill development includes cutting-edge data cleaning, natural language processing, data mining, and data visualization are highly valued competencies in technology companies, consulting firms, and research institutions. The student will learn to translate practical questions into computational problems, bridging data science methodology with real-world problem solving. Interdisciplinary approach combines computer science methods with economic theory and business strategy, providing unique perspective on computational social science applications. Professional development includes mentorship in academic research methods, exposure to university research culture, and networking with economics faculty and graduate students. The project's focus on innovation and entrepreneurship provides insights into startup ecosystems, technology commercialization, and venture capital. The student will develop a portfolio demonstrating ability to tackle complex, real-world problems using sophisticated analytical methods, positioning them for success in their future careers.

Specific competencies required, including programming languages if applicable: Highly Preferred Competencies: 1- Natural Language Processing: Python libraries (NLTK, spaCy, transformers, Hugging Face), text preprocessing, feature extraction, semantic analysis, sentiment analysis, Topic modeling (LDA, BERTopic), named entity recognition, word embeddings, document similarity 2- Machine Learning: scikit-learn, classification algorithms, dimensionality reduction, model validation, cross-validation techniques 3- Data Visualization: matplotlib, seaborn, plotly, interactive dashboards, network visualization, time- series plotting 4- Database Management: merging complex datasets, handling missing data, panel data structures. / Preferred Additional Competencies: 1- Statistical Programming: Python (statsmodels, pandas, numpy), econometric analysis, causal inference methods, instrumental variables estimation 2- Stata

Is there anything else you would like us to know about your project's time frame or work schedule?: I prefer Spring 2026, but summer would also work.

Proposal Title: #1026: Genes, worms, and forever chemicals: Genetic and genomic approaches to investigate per- and polyfluoralkyl substances toxicity

Department: O'Neill School of Public and Environmental Affairs, Environmental Science, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Per- and polyfluoroalkyl substances (PFAS) are chemical pollutants associated with detrimental health outcomes, yet their mechanisms of toxicity are unknown. There are tens of thousands of PFAS in the environment that have not been tested for safety. Hence, we face an urgent challenge to develop novel technological approaches to investigate molecular mechanisms of toxicity. We aim to harness natural variation and the awesome genetic and genomic tools of the model organism, C. elegans to investigate molecular mechanisms of PFAS toxicity. There are multiple parallel "Big Data" projects that are ready for bioinformatic and statistical analyses. Project 1) Effects of PFAS on gene function. We performed transcriptomics to investigate the effect of exposure on gene expression across 10 different PFAS chemicals in the laboratory reference strain. Our hypothesis is that PFAS that share similar structural attributes (such as chain length or functional moiety) disrupt similar genes and gene pathways. Tasks will include analyzing RNA-sequencing datasets, including functional enrichment analysis. Project 2) Statistical genetics approach to investigate PFAS susceptibility. We previously discovered that two different wild strains of C. elegans had opposite responses to two different PFAS, PFOA and its "regrettable" replacement, GenX. Strain JU258 was significantly more sensitive to GenX relative to CB4856, while JU258 was more resistant to PFOA relative to CB4856. We made Recombinant Inbred Lines (RILs) by conducting reciprocal crosses of these two strains which results in the mixing of genetic material in the offspring. 180 new strains (RILs) were then exposed to either GenX or PFOA for toxicity testing. We conducted whole genome sequencing of these 180 RILs. The goal of this project is to analyze the whole genome sequencing datasets and to conduct fine mapping to identify quantitative trait loci that are associated with variation in toxicity to GenX and PFOA.

Rationale for assistance in data analytics and visualization: Genomics datasets are complex. Assistance in working through this project will provide accurate and rigorous results at a much faster pace than what I can do on my own. This support, under my full supervision, will provide the lab with exciting preliminary data for future grant proposals that I aim to apply for next year. Furthermore, I am looking to expand the lab and recruit new students since I am just starting my lab. I am excited to explore these datasets with new students. A long-term career goal is to inform those responsible for decision-making about the safety of PFAS and other chemicals. It is a very exciting time to explore new ways to visualize trends and patterns in complex datasets, or even merge these into a multi-omics project in the future. This could have a large impact on chemical safety regulations if we can make evidence more accessible. I welcome students that have training in data science that will bring in creative, diverse and new perspectives to these data.

Statement of benefit to the student: I am highly devoted to mentoring and training curious and motivated students. These datasets are large and complex. With this opportunity, you will learn how to

manage data, analyze data, and explore new ways to visualize results. We are studying the effects of chemical exposures on health. Our goal is to develop innovative methods for toxicity testing, of which you will be a part of creating. You will deepen your understanding of biology, environmental and public health issues, but in the context of analyzing and presenting evidence. This is a transferable skill, despite what tools you will learn in the future. This is an excellent opportunity for students who enjoy working with quantitative data but are interested in learning and applying their skills to protect environmental health.

Specific competencies required, including programming languages if applicable: Cloud & High Performance Computing, R, Statistics, Data Visualization, Database Management, Network analysis

Is there anything else you would like us to know about your project's time frame or work schedule?: The goal is to use these datasets for a manuscript and for preliminary data for a NIH grant next fall cycle. The timeframe would be to complete these projects during the spring semester to have time to prepare for these deadlines.

Proposal Title: #1027: Defining connectivity of great lakes smallmouth bass populations using genomics and telemetry

Department: Paul O'Neill School of Public and Environmental Affairs, Environmental Science, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: This project will focus on conducting multi-variate statistical analysis and data visualization of a population genetic dataset containing tens-of-thousands of genetic markers genotyped in 300+ smallmouth bass from Green Bay, Lake Michigan, and the Western Basin of Lake Erie. The primary objectives are to (1) quantify and describe the spatial genetic variation among smallmouth bass in these two key fisheries and (2) investigate genetic markers associated with bass migratory behavior. Starting in 2022, smallmouth bass from seven locations in Green Bay (four rivers and three bays) and nine locations in Lake Erie (three rivers and six open-water areas) were collected, fin clipped, and tagged with acoustic tags, and then released. Fin clips were used to conduct genetic sequencing (RAD-seq) and genotyping on all individuals while acoustic tags provided daily information about individual movement. This study design was selected to investigate genetic association of river-run and river-resident behavior and provide information about day-to-day movement and gene flow among smallmouth bass habitats that are important to recreational fishing in the Great Lakes. Anglers have reported that while some smallmouth bass stay as residents in rivers all year round, they observe large numbers of smallmouth bass in the rivers each spring and hypothesize that these bass are migrating into the rivers from the Great Lakes. Our study will therefore determine (1) how common this behavior is in Green Bay and Lake Erie, (2) whether resident bass are genetically distinct from migratory bass, and (3) if there are any genetic markers that determine river-run and river-resident behavior consistent between two distant populations. To investigate this, we will conduct a variety of population genetic analyses, including multivariate statistics such as principal component analysis and discriminant analysis of principal component (PCA and DAPC), network analysis, Bayesian statistical analysis to identify the probable genetic ancestry of individuals, and various outlier detection approaches to identify which of the tens-of-thousands of markers are most likely under selection for river-run vs river-resident behavior. Along with statistical analyses, the researcher involved with the project will be responsible for general data management, the creation and documentation of the analysis pipeline, and for generating publication-quality figures that summarize results for use in reports, manuscripts, presentations to the Great Lakes Fishery Commission, Ohio Department of Fish and Wildlife, and Wisconsin Department of Natural Resources.

Rationale for assistance in data analytics and visualization: This project requires a basic set of population genetic and data science analyses that I think would be accessible for students without significant background in genetics or evolution but interested in an opportunity to work on real data that will have both a research impact and impact on public science outside of IU. The analysis and visualizations conducted by the student will facilitate my research because I do not currently have a project lead for this study, therefore their work will contribute to grant requirements in my lab. However, because I am familiar with all the types of analyses and visualizations needed, I will be able to

work closely with the student and guide them through the process, making this project better suited for the FADS program than a fulltime position or graduate student position.

Statement of benefit to the student: The student will benefit from this project by honing their skills in conducting project-oriented data science that from start to end. They will learn the basic properties and procedures of working with genetic sequencing data which is universal in both environmental and human/health sciences. Furthermore, the data analysis they do will contribute to presentations of these results to other scientists and to state and federal agencies working to manage smallmouth bass in the Great Lakes. The student will get to work closely with the PI of the lab but also will have opportunities to meet with and contribute to meetings with project collaborators outside of academia working for the USGS, Ohio Department of Fish and Wildlife, and Wisconsin Department of Natural Resources. This role will provide real-world experience and connections outside of Indiana University that provides exposure to how research projects are conducted and completed at large scale. The ultimate goal the project is to publish the results in a peer reviewed journal, therefore there will be opportunity for co-authorship depending on the outcome of the project and the students' contributions.

Specific competencies required, including programming languages if applicable: R, shell scripting, IU high-performance Research computing, slurm job submission

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1028: Linking Sub-Daily Rainfall Distribution and Mean Annual Water Availability

Department: O'Neill School of Public and Environmental Affairs, N/A, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Summer 2026

Summary of project: Understanding how climate change reshapes water availability is a central challenge in environmental science. Climate change intensifies the hydrological cycle, changing the total precipitation as well as causing disproportionate increases in extreme rainfall. The rarest and most intense rainfall events tend to grow faster than moderate events, which changes the distribution of sub-daily precipitation. While previous studies have linked long-term water availability to mean precipitation changes, the role of shifting sub-daily precipitation distributions in influencing streamflow remains poorly understood. Streamflow provides water supplies for ecosystems and societies, yet how intensifying sub-daily precipitation impacts the stability of water supplies remains unclear. This project, Linking Sub-Daily Rainfall Distribution and Mean Annual Water Availability, will quantify the historical dependence of water-year runoff on sub-daily rainfall distributions across seasons. We hypothesize that historical water-year runoff is significantly influenced not only by precipitation totals but also by the seasonal distribution of hourly rainfall intensities. The analysis will focus on approximately 581 watersheds across the contiguous United States. We will combine in-situ USGS streamflow data (1979–2022) with hourly precipitation datasets from multiple sources, including reanalysis products, radar-based observations, and satellite-based products. To characterize precipitation patterns, hourly rainfall will be binned into categories (drizzle, light, moderate, heavy, extreme) to reflect shifts in intensity distributions. Using these data, we will develop a log-linear fixed-effects panel regression model to relate annual median streamflow to rainfall distribution characteristics. The resulting coefficients will provide a quantitative metric of how shifts in precipitation distributions affect long-term water availability. The project will involve TB-scale of data analysis, which requires substantial computational resources and data analytics expertise. The Faculty Assistance in Data Science (FADS) Program will play an important role in this research. The FADS students will help with accessing USGS streamflow records and preprocessing and harmonizing multiple precipitation datasets to watershed scales. This project will provide a quantification of how sub-daily rainfall distributions shape long-term water availability across the United States. The results will inform water resource management and highlight the value of data science in addressing environmental problems in a warming climate.

Rationale for assistance in data analytics and visualization: A MSDS student is important to this project because of the extensive data pre-processing demands. We will access time series of streamflow data (1979-2022) across about 581 watersheds in the U.S. and match them with sub-daily precipitation from multiple datasets, each ~2 TB in size. Managing, harmonizing, and analyzing data at this scale exceeds the capacity of a single faculty member. The MSDS student contribute in several key areas: (1) collecting data: downloading in-situ USGS streamflow datasets; (2) pre-processing: screening for quality, filling or filtering missing values, and homogenizing formats; (3) aggregation: averaging gridded precipitation to watershed scale for integration with streamflow; (4) analysis: binning precipitation into intensity

categories including drizzle, light, moderate, heavy, and extreme, and help developing the log-linear fixed-effects panel regression model. (5) visualization: generating interpretable figures to summarize rainfall distributions, streamflow responses, and model outputs. My research group has already downloaded several precipitation datasets and analyzed their intensity, duration, and frequency on IU's HPC. With MSDS support, we can extend the scope to streamflow-precipitation linkages. Given the dataset size and computational complexity, the project is unlikely to succeed without the assistance in data analytics and visualization.

Statement of benefit to the student: The project will provide students with hands-on training, mentoring, and professional development in data science applied to real-world environmental challenges. First, they will gain direct experience working with hydroclimate datasets, including U.S. Geological Survey streamflow records and precipitation data from radar, satellite, and reanalysis products. They will learn to interpret these diverse datasets and apply advanced data analytics to address questions of water availability and climate change, issues with important societal implications.

Second, as a climate modeler and physical hydrologist, I will provide domain knowledge in hydroclimatology. This ensures that they understand not just how to perform tasks, but also the scientific rationale behind their work. Through regular check-ins, I will guide them in analyzing and visualizing datasets, which will help prepare them for their future careers, such as in climate risk assessment, flooding insurance, civil engineering, and geospatial analysis. Finally, the student will strengthen their communication and presentation skills. I work closely with FADS fellows on preparing for the end-of-program research showcase, where my students have won the best overall presentation in Summer 2025. Past fellows have also broadened their interests. For example, one developed expertise in HPC and was accepted into the competitive ACI Student Fellows Program.

Specific competencies required, including programming languages if applicable: An ideal FADS student would have some experience in Statistics, Data Visualization, or Cloud and High-Performance Computing, as these skills are considered essential. Proficiency in Python is strongly preferred because of the existing legacy code, making it easier to integrate new analyses with previous work. Given the specific focus of my project on spatial and temporal data analysis, I will provide training to the selected student in working with multi-scale geospatial datasets. Based on my experience working with MSDS students before, most are familiar with tabular data but have not worked with multidimensional datasets in netCDF format, which is commonly used in climate and earth science. I will provide students with tutorials on libraries such as xarray, cartopy, matplotlib, etc. If the student lacks prior experience, I will also assist them in using IU's High-Performance Computing (HPC) System for tasks such as data collection, processing, and management.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1029: Examining AI and Purpose: how advances in artificial intelligence affect people's sense of meaning towards their work

Department: Kelley School of Business, Management and Entrepreneurship, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: This research project examines how advances in artificial intelligence affect people's sense of meaning and purpose towards their work. In recent years, we have seen very strong developments in artificial intelligence technologies, particularly in the domain of generative artificial intelligence, exemplified by GenAI tools like ChatGPT, Claude, Copilot, etc. On one hand, these technologies have the potential to boost productivity and augment work productivity. However, the is also growing concern regarding the ability of artificial intelligence to replace human altogether. Such displacement of labor has raised attention towards not just the viability of employment in the future, but also the meaning of work altogether. A natural question that arises is – how do people think about the meaning of their work as artificial intelligence capabilities become more and more powerful? According to some of the leading AI experts, this is indeed one of the most pressing challenges that we face. In a recent interview, Dr. Geoffrey Hinton who is often called "the Godfather of AI" and won the 2024 Nobel Prize in Physics, when asked what the biggest threat to human happiness is, said: "I think that joblessness is a fairly urgent short-term threat to human happiness. I think if you make lots and lots of people unemployed, even if they get universal basic income, they are not going to be happy … because they need purpose … they need to feel that they are contributing something, that they are useful." Similarly, AI safety expert Dr. Roman Yampolskiy in a recent interview said: "We as a humanity, when we all lose our jobs, what do we do? What do we do finically? Who's paying for us? And what do in terms of meaning? What do I do with my extra 60, 80 hours of week? … The hard problem is, what do you do with all that free time? For a lot of people, their jobs are what gives the meaning in their lives. So they would be kind of lost." In this research project, we will attempt to establish the relationship between AI and purpose. We will establish a database that records all the major developments of AI advances, primarily based on media report outlets. We will then use text analysis techniques to codify these media articles related to AI. We will then link this database with an existing survey data repository from Gallup, which has collected data from millions of respondents on their attitudes regarding their jobs. We will then conduct analysis on how those AI advancement events impacted the way in which people think about their jobs.

Rationale for assistance in data analytics and visualization: This project involves collecting and processing large scale data. It also involves using complex text analysis techniques to analyze news articles. The visualization will also be a challenge as we will be working with a large-scale survey data.

Statement of benefit to the student: This project tackles a state-of-the-art research question that is relevant to the latest technological developments. Many AI companies nowadays have a research unit that aims to understand the impact of AI tools on the broader labor market as well as the future of work. For instance, Anthropic published a paper that established an Economic Index for the impact of GenAI on different types of knowledge work. OpenAI also recently released a report on how ChatGPT

users are using the tool. The project will help students to establish analytical and research skill sets that are valuable to understand and examine the social and economic impacts of GenAI.

Specific competencies required, including programming languages if applicable: High Performance Computing, Database Management, Natural Language Processing, Statistics, Python, Stata

Is there anything else you would like us to know about your project's time frame or work schedule?: Must be willing to commit to regular research project meetings. Some but not all meetings will in-person.

Proposal Title: #1030: Archival data sourcing and analysis for behavioral science / management research

Department: Kelley School of Business, Management & Entrepreneurship, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: I am seeking a data-savvy student to assist with several stages of my behavioral science research in management. My work focuses on micro-organizational behavior—studying how individuals think, feel, and act in organizational settings. The student will contribute to projects in four main ways, in sequence: 1. We will begin with a meeting to discuss some of my ongoing projects. The student will read selected manuscripts and works in progress to gain a clear understanding of the hypotheses and research questions I am pursuing. 2. Once familiar with the research questions, the student will identify relevant archival data sources that could be used to test the hypotheses. 3. The student will analyze the archival data to evaluate whether it supports the hypotheses. They will walk me through their analyses so I can incorporate the findings into manuscripts. 4. Finally, the student will review the written results sections of manuscripts, providing feedback on whether the results are reported accurately and clearly. To illustrate the type of work involved, here are two examples: • A colleague's published study used natural language processing to examine over 43,000 documents (1.23 billion words). The authors found that hiring female chief executive officers and board members was associated with shifts in organizational language, such that the semantic meaning of "woman" became more similar to that of "agency." This showed how organizational decisions can influence broader associations about gender and leadership. • In my own work, I examined the effect of progress steadiness on motivation. To test this, I scraped publicly available longitudinal running-time data and compared people who showed steady, consistent improvements with those who had fluctuating patterns (high highs and low lows). This allowed me to investigate whether steady progress leads to better long-term performance or if variability can yield similar outcomes. This position differs from a traditional role in that it is not tied to a single, predefined dataset. Instead, the student will play an active role in thinking creatively about how to find or construct datasets that address my hypotheses. Some hypotheses that I'm working on are related to motivation (progress steadiness), goal pursuit, leadership, and prosocial behaviors in the workplace. They may also generate new research questions from the data they uncover if the archival data is interesting. Experience with methods such as natural language processing, LIWC, and data scraping will be especially useful. Familiarity with software such as Stata, R, and SPSS is strongly preferred, as is familiarity with behavioral science data and analytical approaches commonly used in journals such as Academy of Management Journal, Journal of Applied Psychology, and Journal of Personality and Social Psychology.

Rationale for assistance in data analytics and visualization: Although I am a quantitative researcher and typically conduct most of the statistical analyses for my experimental studies on my own, I have not yet successfully sourced and analyzed an archival dataset that directly fit the needs of my projects. My training and experience have centered on designing experiments, collecting survey and lab data, and running analyses such as regressions, moderated mediation, and multilevel models. However, identifying, cleaning, and analyzing large-scale archival data—whether scraped from public sources,

drawn from organizational records, or processed through computational methods like natural language processing—requires a different skillset. This is where I see the student playing a critical role: bringing creativity and technical expertise to locate appropriate data sources, structure them for analysis, and generate insights that can meaningfully extend my experimental research.

Statement of benefit to the student: I hope students will feel a great deal of freedom in this work. While I have specific hypotheses and research questions in mind, they will have the opportunity to think creatively about how to identify and source archival datasets that could be used to test these ideas. This means the student is not limited to working on a single, predefined dataset but can actively contribute by finding new data sources, applying analytic techniques, and shaping how the research unfolds. In addition, if the student develops their own hypothesis and can identify a compelling archival dataset that supports it, I would be open to exploring that project together as a potential paper. This creates a unique chance for the student not only to strengthen their technical and analytical skills but also to exercise intellectual creativity and potentially co-develop original research. By working with me, students will gain hands-on experience in sourcing, preparing, and analyzing archival data, while also learning how these analyses connect to theory and hypotheses in behavioral science. This experience will be especially valuable for students considering graduate school or a career in research.

Specific competencies required, including programming languages if applicable: • Understanding of behavioral science research • Ability to identify and source relevant archival datasets • Ability to analyze archival datasets • Ability to clearly explain analytic procedures and results • Proficiency in STATA, SPSS, or R (preferred analysis software)

Is there anything else you would like us to know about your project's time frame or work schedule?: There is no fixed due date for these projects. The scope is flexible, as I do not have a single predefined dataset requiring analysis. The pace and work schedule can be set by the student, though ideally they will work a consistent number of hours each week on sourcing and analyzing interesting datasets.

Proposal Title: #1031: Using the best-available science to reduce bias in estimates of forest carbon uptake and storage

Department: O'Neill School of Public and Environmental Affairs, Environmental Science, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: Forests dominate the land carbon sink, and forest-based carbon removal strategies have the potential to confer many benefits for the environment and the bio-economy. However, at scales ranging from individual sites to the entire globe, estimates of forest carbon uptake and storage (FCUS) vary by 50-100% (or more). Much of this uncertainty stems from a misalignment between our state-of-the-art understanding of FCUS and the decades-old tools used to estimate it in practice. Our overall objective is to use the best-available science to confront these discrepancies through a robust intercomparison of ground, tower, and satellite-based measurement approaches. We will test an overarching hypothesis that the limitations of traditional measurement approaches have led to a systematic underestimation of how much carbon is removed from the atmosphere by undisturbed forests. Our workflow for this project will begin with a comprehensive synthesis of information from environmental observation networks including NSF's National Ecological Observatory Network (NEON), the FLUXNET tower network, and the USDA Forest Inventory and Analysis (FIA) network. This "big-data" synthesis will also leverage multiple next-generation satellite remote sensing datasets to allow us to assess the ecological similarity between the nodes of these disparate networks, and to permit us to upscale ground observations to the wall-to-wall maps that are most useful to policymakers and forest managers.

Rationale for assistance in data analytics and visualization: The project workflow requires the analysis and synthesis of large datasets from multiple, high-profile environmental observation networks. Moreover, the project requires us to match and upscale point-scale data to the vast amounts of information made available from satellite remote sensing data products. Luddy MSDS students have many of the core skills necessary to execute the analysis, data management, and data visualization steps necessary to accomplish these research goals. We have previously enjoyed the opportunity to work with a MSDS student on a related project. After that student graduate, he joined our lab group as a full time data scientist for one year. During that time, he laid the groundwork for a journal article on which he will appear as co-first-author. This student recently landed a HIB eligible position as a software engineer for Amazon Robotics. From this perspective, we are happy to consider the potential for a longer-term appointment for students who join our team through the FADS, and who have an interest in continuing work with our group for more than a few months.

Statement of benefit to the student: The student will have the opportunity to apply their Data Science skillsets in support of a large, collaborative, NSF-funded project. The student will have the opportunity to work closely with multiple IU faculty, graduate students, and postdoctoral scholars. The work has the potential to lead to co-authorship on peer-reviewed publications. The project aims are closely tied to

the terrestrial carbon bio-economy (e.g. carbon markets), and thus this student will also gain knowledge and experience relevant to private-sector sustainability and climate opportunities.

Specific competencies required, including programming languages if applicable: Cloud & High Performance Computing, Data Visualization, Database Management, Deep Learning, Machine Learning, Statistics

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1032: Detecting Cognitive Vulnerability in Aging via Large Language Models

Department: College of Arts and Sciences, Psychological and Brain Sciences, Bloomington

Project requires Human Subjects? Yes

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Depression is a well-established risk factor for cognitive decline and Alzheimer's disease and related disorders. However, there remains no scalable, sensitive method to detect early cognitive vulnerability, or preclinical decline, in older adults, particularly in those with depression histories. While clinical screening tools exist, they have low sensitivity in preclinical stages. While a variety of novel digital tools are emerging, from passive sensing to wearables, they are often expensive and not scalable. Broadly, traditional assessments are too resource-intensive or infrequent to detect opportunities for early intervention. Using online language offers a simpler, lower burden, scalable option to detect cognitive vulnerability, while considering the important role of depression. We have previously developed a cognitive distortion lexicon to detect maladaptive thinking patterns associated with depression in online language. In this project, we propose to compare this lexicon-based matching approach to cutting-edge large language models (LLMs) (e.g., SBERT, GPT, LLaMA) to evaluate their ability to identify cognitive distortions and markers of cognitive vulnerability in real-world samples from midlife and older adults from BlueSky and Reddit. We will also conduct perplexity analyses to measure how predictable vs. disfluent a person's language appears to an LLM. Our theory is that this could serve as a potential marker of subtle cognitive decline. Increased perplexity may reflect disorganized or atypical language patterns that precede overt impairment and decline. By integrating natural language processing and deep learning, this work will test whether language-based markers (cognitive distortions, perplexity scores) are enriched among older adults at elevated depression risk, positioning them as candidate indicators of cognitive vulnerability for future validation. This approach is innovative, scalable, and leverages naturally occurring online data to identify possible early warning signals of decline in at-risk populations.

Rationale for assistance in data analytics and visualization: To execute this work, I am seeking 1–2 data science graduate students with expertise in: • Natural Language Processing (transformer-based models, embeddings, fine-tuning) • Deep Learning • Machine Learning (classification, feature extraction, predictive modeling) These students would help build and benchmark models that (a) classify texts for cognitive distortions using our existing lexicon vs. fine-tuned LLMs, and (b) compute and analyze perplexity metrics across large samples of older adult language. While I have experience designing the overall framework and conceptual models, I lack hands-on support for implementing and optimizing large-scale NLP and deep learning analyses. The FADS mechanism would allow me to closely mentor technically skilled students who can accelerate this project from concept to reality.

Statement of benefit to the student: The student(s) will: • Gain experience applying cutting-edge NLP and deep learning to real-world mental health and aging data • Contribute to model development, evaluation, and visualization pipelines • Potentially earn authorship on manuscripts and presentations • Receive close, individualized mentorship from me (weekly meetings) and access to my

collaborative network across psychology and informatics. This experience would be especially valuable for students interested in digital mental health or clinical applications of AI.

Specific competencies required, including programming languages if applicable: Natural Language Processing , Deep Learning, Machine Learning, Data Visualization

Is there anything else you would like us to know about your project's time frame or work schedule?: I'd like to finish this by early summer 2026. I have the data already and need data scientists for analytic support with the LLM.

Proposal Title: #1033: Early African movies database

Department: College of Arts and Sciences, Department of French and Italian, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: Founded in early 2024, the Early African Cinemas Lab (EACL) at Indiana University investigates the diverse film practices developed in West Africa since the 1960s, particularly during the sociopolitical transformations surrounding African independence. As contemporary audiovisual industries in West Africa – first in Nigeria, now increasingly in Ghana, Senegal, and beyond – gain global visibility, there is an urgent need to contextualize these developments through a deeper understanding of African cinema's foundational decades. This requires concrete, archival-based analyses of production methods, funding sources, and reception contexts that shaped the early cinematic landscape. To support this mission, EACL initiated the development of a centralized database in collaboration with a FADS team in 2024. This database was designed to organize and share archival findings from IU's unique collections, including the Ousmane Sembène papers at the Lilly Library and the Paulin S. Vieyra papers at the Black Film Center & Archive. Unfortunately, due to unresolved hosting issues – specifically the lack of an approved IU vendor for data storage – the database remains inactive. The proposed 2025 project seeks to reprogram and relaunch the EACL database using Indiana University-approved data storage infrastructure. The current version, built in PostgreSQL, will need to be adapted or migrated while preserving its core functionalities and design logic. These include: - Controlled editing access: The database must be editable only by designated members of the Lab and trusted collaborators within the EACL network. This requires an authentication system that restricts modification privileges to approved users. - Public-facing data export: Non-restricted data should be accessible to external visitors, necessitating a mechanism for exporting and displaying selected content in a user-friendly format. While the specific technologies used may be adjusted to suit the FADS team's recommendations, the database must retain its capacity to store detailed metadata on individual films – covering technical production, artistic contributors, distribution pathways, and reception history. Each film entry should also include geographic visualization tools to map production and post-production locations, funding sources, and screening circuits.

Rationale for assistance in data analytics and visualization: To fulfill its scholarly and pedagogical mission, the database must be reprogrammed and hosted within IU's approved infrastructure, while preserving its core functionalities—controlled editing access, public-facing data export, and geographic mapping of film-related activities. Assistance from a FADS student will be essential in adapting the existing PostgreSQL framework to IU's hosting standards. Moreover, the student's expertise will help ensure that the database remains both technically sound and accessible to non-specialist users, including researchers, students, and the broader public.

Statement of benefit to the student: This project offers a unique opportunity for a data science student to apply their technical expertise within a digital humanities context, working closely with a non-specialist client who has specific research needs. The student will gain valuable experience in translating

those needs into a functional and sustainable digital infrastructure, while learning to manage expectations and communicate effectively across disciplinary boundaries. The challenge lies in balancing simplicity and professionalism: the final product must be technically sound and visually engaging, yet intuitive enough to be maintained by non-specialists over time. This will require the student to not only build the system but also guide the PI's team in basic maintenance tasks, offering a chance to develop and practice soft skills such as instruction, documentation, and collaborative problem-solving. Additionally, the student will engage with archival materials and metadata related to African cinema, gaining exposure to a growing field of interdisciplinary research. Their contribution will directly support an international network of scholars and help shape a resource that will be used in workshops, classrooms, and future research projects. In short, this is a real-world application of data science that bridges technology, history, and global collaboration.

Specific competencies required, including programming languages if applicable: We are seeking students with demonstrated skills in database management, data visualization, and web front-end development. The current version of the Early African Cinemas Lab (EACL) database was built using PostgreSQL, and while we are open to adapting the platform, the core functionalities must be preserved. These include controlled editing access for trusted collaborators and public-facing data export capabilities. As we are not specialists in this domain, we welcome the student's input on the most appropriate and scalable technologies for this project.

Is there anything else you would like us to know about your project's time frame or work schedule?: This advanced database is essential to the success of our research. It will enable us to centralize, share, and process data collected from archives, publications, and testimonies related to films produced in West Africa during the 1950s–1970s. Beyond its function as a repository, we envision this digital platform as a kind of virtual lab – a collaborative space where experts in the field can contribute, exchange insights, and build new knowledge together. To realize this vision, it is imperative that we adapt the database's programming language and infrastructure to meet Indiana University's approved data storage protocols. This transition is not merely technical; it is foundational to ensuring the long-term accessibility, security, and sustainability of the platform within IU's research ecosystem.

Proposal Title: #1034: Harnessing Social Media to Monitor Emerging Drug Trends

Department: Public Health, Epidemiology & Biostatistics, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: Substance use disorder (SUD) and overdose remain urgent U.S. public health crises, with nearly 100,000 annual deaths and 24 million people affected by non-alcohol, non-tobacco SUDs. Despite proven interventions, major gaps persist—naloxone is missing in most fatal opioid overdoses, and few patients receive recommended medications. New, real-time approaches are needed to detect trends, reduce harms, and improve responsiveness. Social media provides an unprecedented lens into candid, real-time drug-related discourse. Platforms such as Reddit, YouTube, TikTok, Bluesky, and X (formerly Twitter) host millions of conversations about drug use, recovery, stigma, and misinformation. Users often communicate more openly online than in clinical or survey settings, employing slang, emojis, memes, and coded terms that evolve quickly and vary across platforms. These environments can surface early warnings of emerging slang, shifts in behavior, or misinformation (e.g., the “rainbow fentanyl” scare) that traditional data sources miss. The long-term goal of this project is to develop AI-enabled, cross-platform surveillance tools to support timely and effective public health messaging. Our central hypothesis is that advanced computational methods—natural language processing (NLP), machine learning, and large language models (LLMs)—can detect and contextualize drug-related discourse across platforms, generating actionable insights for prevention and intervention. We propose two specific aims: Build and validate a cross-platform semantic framework for drug-related language. Starting with the National Institute on Drug Abuse (NIDA) drug inventory, we will expand this ontology using embeddings, clustering, and similarity measures applied to millions of posts from multiple platforms. A panel of people with lived drug use experience will validate emerging terms to ensure accuracy and relevance. The result will be a scalable, adaptive ontology capturing slang, code words, and platform-specific language. Track and model the evolution of drug-related slang and topics. Using topic modeling, network science, and semantic seeding, we will cluster posts into coherent themes, trace how drug-related topics emerge and shift, and map how slang and narratives spread across platforms over time. Analyses will identify which platforms incubate new language and how discourse evolves across drug classes, contexts, and communities. Expected outcomes include: A validated ontology for detecting drug-related discourse across platforms. Real-time, AI-enabled surveillance tools capable of adapting to new slang, platforms, and modalities. Insights into how digital communication shapes behaviors, stigma, and norms. Evidence to inform responsive, culturally sensitive public health messaging. This research will strengthen national capacity for timely overdose and SUD interventions by bridging human expertise with computational power while aligning with NIDA’s priorities.

Rationale for assistance in data analytics and visualization: This project analyzes millions of posts across Reddit, YouTube, TikTok, Bluesky, and X, requiring specialized expertise to manage, process, and interpret large-scale, heterogeneous data. While the investigative team has strong backgrounds in public health, NLP, and machine learning, dedicated support in data analytics and visualization is

essential to ensure rigor, efficiency, and impact. First, preprocessing and standardizing diverse datasets (e.g., slang, emojis, video-linked comments) demands skilled analysts to reduce noise and prepare data for modeling. Second, advanced analytic methods—including embeddings, clustering, topic modeling, and network analysis—require optimization and coding support to handle evolving, cross-platform discourse. Equally critical is visualization. Mapping discourse trends, slang evolution, and network diffusion into clear, interpretable outputs (e.g., co-occurrence networks, Sankey diagrams, semantic maps) is vital for both internal analyses and stakeholder communication. Effective visualization will facilitate validation by advisory panels of people with lived experience and ensure findings are accessible to public health practitioners and policymakers. Dedicated analytics and visualization support will strengthen reproducibility, adaptability, and translation of findings, ensuring the project delivers timely, AI-enabled tools for real-time surveillance and responsive public health strategies.

Statement of benefit to the student: Students will be collaborating closely with public health faculty who have expertise in epidemiology, implementation science, NLP/AI integration, and quantitative measurement. As part of their internship with our team, we expect students to contribute to, or directly lead, peer-reviewed manuscripts, peer-reviewed conference proceedings, and data analyses to support grant proposals as needed. As students prepare to enter the job market, these activities are readily transferrable to CVs and resumes, ensuring students are able to demonstrate sufficient work and internship experience necessary to securing their first job post graduation.

Specific competencies required, including programming languages if applicable: Cloud & High Performance Computing, Data Mining, Data Visualization, Machine Learning, Database Management, Natural Language Processing, Network analysis, Statistics, and Web & Social Media Mining. Languages: Python, R, SQL

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1035: Cracking the Code of Air Pollution: Automating Great Lakes Data Analysis

Department: O'Neill School, SPEA, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: The Integrated Atmospheric Deposition Network (IADN) is a long-standing binational program that has monitored persistent organic pollutants (POPs) in the Great Lakes atmosphere for more than thirty years. The resulting dataset is one of the most comprehensive of its kind worldwide. To fully leverage its potential, we aim to automate workflows that currently limit the efficiency and reproducibility of advanced source apportionment analyses. Two useful approaches for source apportionment include Positive Matrix Factorization (PMF) and the Potential Source Contribution Function (PSCF). PMF is capable of identifying source contributions from ambient concentration data while PSCF integrates air mass trajectories to locate potential geographic source regions. Although these methods are relying on different analytical tools, they are both currently executed through manual and time-consuming steps. This project will deliver a user-friendly application tailored to IADN data that automates PMF and PSCF analyses. The tool will: (1) streamline data preparation and input formatting through an intuitive interface, (2) run models and diagnostic checks automatically, (3) generate standardized visual outputs such as source contribution profiles and back-trajectory maps, and (4) ensure reproducibility and long-term usability through built-in documentation and transparent workflows. This project represents a unique opportunity to combine advanced data science with pressing environmental monitoring needs in the Great Lakes.

Rationale for assistance in data analytics and visualization: PMF and PSCF analyses require specialized coding, statistical modeling, and data visualization skills that extend beyond the expertise of our environmental chemistry research group. While we have deep knowledge in atmospheric monitoring and pollutant fate, developing computational tools for automated analysis requires programming and reproducible workflow design. We can bridge this gap by partnering with a data science graduate student. The student's expertise will help us automate our data analysis workflows thus enabling us to move beyond manual analyses with new scalable tools.

Statement of benefit to the student: This project offers a unique, applied learning experience at the intersection of environmental science and data science. The student will gain hands-on experience in developing workflows for real world scientific datasets, applying advanced statistical techniques (e.g., PMF) and geospatial modeling (e.g., PSCF) to address questions of atmospheric pollution sources. The project will also allow the student to practice reproducible coding practices, develop data visualizations, and collaborate in an interdisciplinary environment. The deliverables (i.e., user-friendly application) will provide work products that the student can include in their portfolio. In addition, the student will gain experience in the process of applying data science methods to environmental monitoring programs, which are directly relevant to careers in environmental consulting, applied research, and data analytics.

Specific competencies required, including programming languages if applicable: Specific competencies required include: proficiency with Microsoft Excel and programming using Python or R. Required project skills include statistical modeling, data visualization, code development, and workflow documentation

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1036: Whether multi-agent AI interviewing systems can generate richer and more actionable qualitative insights

Department: Kelley School of Business, Operations and Decision Technologies, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: This project tests whether multi-agent AI interviewing systems can generate richer and more actionable qualitative insights than traditional methods, such as human-led interviews. Current interviewing practices, though effective, are time-consuming, costly, and difficult to scale. Multi-agent systems, by contrast, promise scalable, specialized probing that could accelerate the discovery of tacit knowledge. However, their validity and research value remain untested. Guided by Cognitive Load Theory (CLT), the study investigates how multi-agent systems shape participants' ability to generate meaningful insights. CLT suggests that distributing inquiry across specialized AI agents (e.g., probing feasibility, customer perspective, equity implications) may reduce cognitive effort and increase deep reflection (germane load). At the same time, poorly managed multi-agent questioning may increase extraneous load, overwhelm participants, and reduce insight quality. Testing these competing dynamics is central to the study. The research employs a controlled experiment with three conditions: (1) human-led interviews, (2) single-agent interviews, and (3) multi-agent interviews with two to three role-differentiated AI agents. Participants will evaluate product concepts, responding to structured questions designed to elicit diagnostic reasoning, creative solutions, and design tradeoffs. Data will include both subjective and objective measures. Cognitive load will be assessed using NASA-TLX surveys and response-time analysis. Insight richness will be measured by blinded expert coders on semantic diversity, novelty, and actionability, supported by lexical diversity and response complexity indices. Participant demographics, AI experience, and task familiarity will serve as control variables. Analysis will involve (a) between-group comparisons of insight richness across conditions, and (b) mediation testing to determine whether perceived cognitive load explains differences in outcomes. This design ensures both methodological rigor and theoretical contribution. The project's contributions are threefold. First, it offers the first empirical test of multi-agent interviewing systems, establishing whether they represent a valid new method for qualitative research. Second, it extends CLT into AI-mediated settings, clarifying how cognitive load shapes insight generation in human–AI collaboration. Third, it develops a practical framework for evaluating the quality of AI-generated qualitative data, equipping researchers and practitioners with scalable methods for gathering actionable insights. By integrating system design, cognitive theory, and rigorous analytics, this work will advance the science of AI agent interviewing while producing a prototype platform for future research and application.

Rationale for assistance in data analytics and visualization: The success of this project depends on building a robust multi-agent interviewing platform and analyzing complex qualitative data at scale. A data science student will provide essential expertise in three areas. First, platform development requires technical support to design and deploy the agent-based interview system as a web or app interface. The student will help integrate multiple AI agents with distinct roles, ensure smooth interaction flow, and establish reliable data capture pipelines. Second, rigorous data management is critical. The student will construct and maintain databases to store transcripts, cognitive load measures, and metadata. They will also support preprocessing and organization of interview outputs to enable reproducible analysis. Third,

advanced analytics and visualization are necessary to evaluate outcomes. The student will apply natural language processing, deep learning, and statistical modeling (e.g., mediation analysis) to assess insight richness and cognitive load. They will also develop clear, interpretable visualizations to communicate findings to scholarly and practitioner audiences.

Statement of benefit to the student: This project provides a unique training opportunity for a data science student to apply advanced computational skills in a cutting-edge research setting. By contributing to the design and evaluation of a multi-agent interviewing platform, the student will gain hands-on experience at the intersection of artificial intelligence, data analytics, and human–computer interaction. The student will develop and refine competencies in natural language processing, deep learning, data mining, and web mining, while also managing large-scale qualitative datasets and building robust databases. They will engage in advanced analytics and strengthen their skills in data visualization by creating interpretable outputs for both scholarly and applied audiences. Beyond technical development, the student will benefit from direct mentorship and collaboration on a faculty-led research project with clear theoretical and applied contributions. This experience will demonstrate how data science methods can be integrated into innovative business research, producing results that are both academically rigorous and practically impactful. Participation will enhance the student's professional portfolio, provide exposure to interdisciplinary research methods, and prepare them for competitive careers in academia, industry, or applied data-driven research.

Specific competencies required, including programming languages if applicable: Required competencies include data mining, database management, deep learning, natural language processing, web mining, and data visualization. Proficiency in C/C++, Python, and R will enable the student to deliver a scalable, technically sound platform and robust analytical results.

Is there anything else you would like us to know about your project's time frame or work schedule?: I will apply the IRB approval in Fall 2025 and make sure I get the approval before we start this project. My goal is to have the system built in spring 2026 with the help of the data science student.

Proposal Title: #1037: Geopolitical controversy and digital entrepreneurs' strategic adaptation

Department: Kelly School of Business, Management and Entrepreneurship, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Background Geopolitical risks and nationalist sentiments increasingly shape the global business environment. Multinational corporations have long been studied in this regard, with research examining how trade protectionism, sanctions, and techno-nationalism influence firms' foreign investment decisions, supply chain structures, or market entry strategies. Yet, less is known about digital entrepreneurs—bloggers, influencers, podcasters, and content creators—who operate across borders through international platforms such as YouTube, TikTok, and Instagram. These individuals are highly exposed to sociopolitical dynamics: their audiences are global, their business model relies on sponsorship and advertising revenues, and their visibility depends on continuous engagement from viewers with diverse political or cultural orientations. Unlike multinational firms, digital entrepreneurs are not constrained by physical or formal institutional barriers, but their dependence on online platforms renders them especially vulnerable to shifts in public sentiment. Geopolitical disputes can transform everyday commercial activities—such as showcasing a fashion brand—into acts laden with political meaning. This study seeks to understand how such entrepreneurs adapt strategically when confronted with these pressures. Research Focus Our project examines how fashion bloggers who publish on international platforms respond to geopolitical tensions. Fashion bloggers are an ideal focus group because they form one of the largest categories of digital entrepreneurs and are closely tied to global brands through advertising and sponsorships. Decisions about which products or brands to feature can simultaneously serve commercial goals and signal political alignment, placing them at the intersection of business opportunity and sociopolitical controversy. Specifically, we ask: 1. How do digital entrepreneurs adjust their content in response to rising geopolitical tension? 2. Do these adjustments differ based on the size, diversity, and loyalty of their audiences? 3. What are the commercial and reputational consequences of strategic adaptation? Data and Methods The study leverages a large dataset of fashion blogger activity before and after a major geopolitical controversy. Data will include: • Content features: which brands are featured, frequency of mentions, and shifts over time. • Engagement metrics: likes, comments, shares, and view counts. • Commercial signals: indications of brand sponsorships or promotional partnerships. • Audience responses: textual data from comments, to be analyzed using sentiment analysis. Expected Contributions This project will generate insights into how digital entrepreneurs adjust content in the face of geopolitical controversy. At the same time, we hope to understand how these changes affect engagement across different communities of viewers. At last, we hope to shed light on whether adaptation comes at the cost of sponsorship opportunities or conten

Rationale for assistance in data analytics and visualization: This project relies heavily on analyzing large-scale, unstructured digital trace data from international platforms. The data include video metadata, brand mentions, audience engagement statistics (likes, shares, comments), and thousands of viewer comments in text form. Making sense of this information requires advanced data processing, natural language processing (NLP), and visualization skills. Graduate assistance in data analytics will be

essential for systematically cleaning and structuring raw data, implementing sentiment analysis of audience comments, and applying statistical models to assess how bloggers adjust their strategies under geopolitical pressure. Beyond text, the project requires integrating numeric engagement metrics and categorical brand features into robust models that capture content adaptation over time. Visualization support is equally critical. Clear and effective data visualizations will allow us to track brand inclusion/exclusion dynamics, shifts in audience sentiment, and engagement patterns across diverse viewer communities. Network visualizations will further help illustrate how bloggers' audiences fragment or converge in response to their strategic choices. Such analytic and visualization expertise ensures that findings are not only rigorous but also interpretable and communicable to both academic and practitioner audiences, making graduate data science assistance indispensable to th

Statement of benefit to the student: This project offers the student hands-on experience in analyzing large-scale digital data in an international business context, using Python for data collection, cleaning, and analysis, API interaction, and potentially Stata for data management. They will gain practical skills in structuring, preprocessing, and integrating complex datasets, as well as advanced data visualization to present findings clearly. The student will also develop experience applying data science methods to real-world organizational questions, enhancing both technical and critical thinking skills. Collaborating with faculty on a live research project provides insight into the full research workflow—from hypothesis development to analysis and interpretation—preparing the student for future roles in data analytics, research, or consulting.

Specific competencies required, including programming languages if applicable: This project requires proficiency in Python for data processing, text analysis, and visualization, experience with API interaction for data collection, and familiarity with Stata for dataset management and integration. Some of the data could be in Chinese. So being able to read basic Chinese is preferred but not required.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1038: Verifiable Training for Bayesian Tree Models

Department: College of Arts and Sciences, Statistics, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: My research in statistics involves devising machine learning models, adding tools to such models to aid in interpretation, and to provide open-source reference implementations for use by the wider scientific and industry communities. I am the primary author of an influential open-source codebase known as OpenBT (https://bitbucket.org/mpratola/openbt/wiki/Home). OpenBT implements various tree-based regression models, namely the Bayesian Additive Regression Tree model and variants that are popular in data science and data analytics and used by the wider scientific and industry communities. OpenBT and earlier versions of that code base have provided reference implementations of key techniques for the last 10 years, and directly contributed to software packages that are widely used by practitioners such as the BART package for R (https://cran.r-project.org/web/packages/BART/index.html) and the Taweret package in the BAND software framework (https://bandframework.github.io/software/), among many others. This project will continue the work originally started in the FADS 2024 cycle. Last year we successfully rewrote the base class (brt) of OpenBT in Rust and the goal this year is to rewrite OpenBT's primary high-level model classes (mbrt, ambrt) in Rust. The primary reason for this research is the ability to run Rust code in a Zero-Knowledge environment (ZKvm) that provides a cryptographic proof of code execution (https://risczero.com). The goal is a new reimplementation of OpenBT that enables reproducibility and verifiability of data analyses, which becomes possible with Rust. Other languages do not admit this possibility, so transitioning the C++ codebase to Rust enables an entirely novel capability that will bring a lot of value to users. For example, the ability to trust and verify a model and its results becomes feasible. In an era where AI will increasingly make fictional predictions trivial to produce and impossible to validate, creating the theoretical methodology and computational tools to attach quality and validity to results of data analyses becomes a critical primitive necessary for data analyses to have tangible value. A secondary benefit is the strict typing system of Rust, which helps eliminate memory-based bugs. It is well known that memory bugs account for roughly 70% of errors in C/C++ codes, so a Rust implementation will benefit the community from a general code quality standpoint. The FADS program is an excellent initiative to support this work, as evidenced with our success last year. We will focus on transitioning two sub-components of OpenBT to Rust and evaluate the performance in ZKvm. The assistance of talented MSDS graduate student(s) supported by FADS would be an ideal opportunity for those with some experience in Rust, or a strong desire to learn. This project will also provide a low-risk environment since it will be an initial exploration of the idea, yet the potential impact is high.

Rationale for assistance in data analytics and visualization: The FADS program is designed to provide access to expertise and assistance available via the skills, training and talents of M.S. students in Data Science (MSDS) for research internships. The skillset of such a student is a perfect fit for what I aim to achieve in this research project, and I clearly have the experience for this to be an excellent opportunity for the student, to leverage the support of FADS in furthering my research goals, and for this project to contribute to the wider scientific and industry communities. For instance, my group has published 17

papers on or related to this machine learning model, I have attracted significant grant support for my research in this area over the years, and the OpenBT codebase has served as a reference implementation for many other open-source packages. Moreover, this area of machine learning model development has been experiencing explosive growth in recent years, confirming the opportunity for high impact (starting in 2010, the cumulative number of publications in Google Scholar with "Bayesian Additive Regression Trees" in the title are [4,4,5,8,13,15,23,31,39,58,71,88,107,126,147]).

Statement of benefit to the student: The benefit of this opportunity for the student is tremendous. First, I am a recognized leader and expert in the development of software codes for tree-based models, and therefore have a lot of knowledge and experience the student can gain from. Second, this is a highly novel project with the potential for high impact and further ongoing work. To my knowledge, I am not aware of other researchers in statistics attempting to add proof of execution/training capabilities to statistical models and so this could be a first in the field and the experience could result in a publication. Finally, this project offers a tremendous opportunity to learn and apply cutting-edge techniques at the intersection of statistics, computer science and cryptography that provides a unique opportunity to the student(s), who will gain a rare and valuable skillset through this research. To summarize, this project supports all the stated goals of FADS, including a valuable learning experience for the student(s), a promising/new and high-impact research direction for the faculty, and I am unlikely able to achieve this alone without the help of an ambitious and talented MSDS student(s).

Specific competencies required, including programming languages if applicable: (i) Some experience in Rust or a desire to learn Rust. (ii) Experience in at least one of C, C++ or Python is desirable. (iii) An interest in blockchain, zero-knowledge proofs, cryptography and verifiable computing are a plus but not necessary. (iv) Knowledge of tree-based models, such as Random Forests, CART, BART, Boosting, CatBoost, etc., and experience applying them in data science/data analytics are desirable.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1039: Forecasting Industry Transformation from Job Posting Data

Department: Kelley School of Business, Marketing, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Firms' hiring decisions provide important signals about their future priorities. Job postings often reveal where organizations are investing resources, such as in artificial intelligence, digital tools, or sustainability. Unlike traditional labor statistics, which are slow and retrospective, job postings are real-time and forward-looking, offering a unique lens into how industries evolve. This project aims to develop a Transformer-based model to analyze large-scale job posting data and forecast industry trends. Transformers, originally designed for language processing, are well suited to this task because they can capture patterns in sequences and integrate information from both structured attributes (such as job titles or industries) and unstructured text (such as job descriptions and skill requirements). The goal of the project is to demonstrate how these models can be used to identify shifts in industry growth, detect emerging skills and roles, and capture differences in how firms adapt to technological or regulatory changes. The approach focuses on forecasting and pattern recognition, showing how advanced data science methods can be applied to management and strategy research. This project will provide an opportunity to work with cutting-edge machine learning techniques while addressing questions of practical importance to business, policymakers, and society.

Rationale for assistance in data analytics and visualization: This project requires careful preparation and analysis of large-scale job posting data that combine structured information and unstructured text. Student assistance will be valuable for cleaning and organizing the data, creating features that reflect hiring patterns, and preparing inputs for Transformer-based models. Equally important is the visualization of results. Forecasts of industry growth and signals of emerging skills must be communicated clearly to diverse audiences. Students will help create charts, dashboards, and other visuals that translate complex model outputs into accessible insights. Through this work, they will gain practical experience in data science, text analytics, and visualization for management research.

Statement of benefit to the student: This project will give students practical experience in applying data science to real-world business questions. By analyzing large-scale job posting data, students will learn how firms' hiring decisions reflect industry priorities and strategic shifts. They will gain skills in data preparation, text analysis, and feature engineering, as well as exposure to advanced machine learning methods such as Transformer models. Students will also practice developing visualizations that translate complex model outputs into clear insights about business and industry trends. This combination of technical training and applied problem-solving will prepare students to use data science not only for academic research but also for business decision-making. Participation will strengthen their readiness for graduate study or careers in data science, analytics, and consulting.

Specific competencies required, including programming languages if applicable: Python; Machine learning (PyTorch or TensorFlow); Natural language processing (text handling, embeddings); Data visualization; Reading proficiency in Mandarin (for Chinese job posting data)

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1040: A data-driven approach to mapping the complexities and challenges of online book reviewing

Department: Luddy School of Informatics, Computing, and Engineering, Department of Information and Library Science, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: This is an ongoing interdisciplinary research project on social reviewing of books (i.e., various forms of online book reviews, such as Amazon book reviews, book pod casts, "BookTok" on YouTube, and "Bookstagram" on Instagram). My lab has worked on a variety of sub-projects, from exploring the impacts of incentivized reviews on the visibility of books, to comparing online book reviews in multiple languages and across platforms. We adopt a variety of methods to a spectrum of questions, such as texting mining with large language models, qualitative content analysis of online discussions, and conducting surveys and semi-structured interviews among stakeholders of online book reviews. Our research on online book reviews has been published in top-tier journals and conference proceedings in critical data studies, digital humanities, and library science venues, such as Big Data & Society, Digital Scholarship in Humanities, the ACM-IEEE Joint Conference on Digital Libraries, and iConference (Best Paper awarded). Currently, my lab focuses on two sub-projects that require computational data analysis. The first project investigates how book content generated or mediated by artificial intelligence (AI)—hereafter AI-book-content—is received, negotiated, and contested by everyday readers, and how such content is reshaping contemporary book culture. The AI-book-content includes AI-generated book texts, cover art, illustrations, as well as paratexts such as book reviews and fanfiction. These research questions arise amid heated debates on AI-mediated content. For example, some readers have called for boycotts of books with AI-generated covers—viewed as unethical derivatives of human labor—while others argue that such backlash unfairly harms authors, particularly when production decisions are made by publishers. These increasing tensions demand a critical investigation into the expansion of generative content in the cultural sphere, and the sociotechnical challenges it has posed. The second research project aims to investigate the evolving dynamics of social reviewing online through the lens of backlash activities (e.g., reviewing bombing, boycotts, and book banning online). In recent years, social reviewing and book campaigns online have gained phenomenal success in book marketing. However, they have also caused broad concerns due to their role in reinforcing inequality in the publishing industry, and intensifying tensions around book banning, polarized discourse, platform moderation, and cybersecurity concerns. These incidents highlight an urgent need for new insights for both book industry stakeholders and everyday readers to navigate the evolving landscape of social reviewing. For both projects, we will conduct large-scale text mining and other computational analysis on social reviewing data, which requires skills in data analytics and visualization.

Rationale for assistance in data analytics and visualization: I am an assistant professor at the Department of Library and Information Science at Luddy School of Informatics, Computing, and Engineering. I joined Luddy last year and I don't have any PhD advisees yet. While I have experience in

working with large language models, massive datasets, and computational models, most students in our department come from humanities backgrounds and are not familiar with these computational approaches. In addition, due to great uncertainty associated with the primary funding agencies in my areas (e.g., Institute of Museum and Library Services, and National Endowment for the Humanities), I only have limited funding which restricts my ability to work with students outside of my own department through regular hourly paid jobs. Due to these factors, my research will significantly benefit from Faculty Assistance in Data Science. I have years of experience in working with computer science and data science graduate students since my doctorate study. After joining Luddy, I have worked with multiple computer science and data science graduate students, as well as computer science and intelligent system engineering undergraduates, through independent studies and short-term research assistantships. I am confident that I can provide effective guidance and valuable research experience for students working on my projects.

Statement of benefit to the student: Students working on this project will have the opportunity to practice their data analytics and visualization skills and apply mixed methods to solve real-world sociotechnical problems. They will also receive training in fundamental research ethics and research development. In addition, they will gain unique experience to work in highly inter-/multi-disciplinary lab environments, with collaborators (e.g., faculty members in library science, computer science, and informatics) and lab mates from a variety of backgrounds, including doctoral students, master students, and undergraduates. Students who make significant intellectual contributions to the project will be considered for co-authorship of publications in top tier venues.

Specific competencies required, including programming languages if applicable: Project skills: Data Mining, Data Visualization, Database Management, Natural Language Processing, and Web & Social Media Mining; Language: Python

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1041: INSPECT-AI: Evaluating the Development and Impact of AI-Assisted Integrity Assessment of Randomized Trials in Evidence Syntheses

Department: School of Public Health, Epidemiology and Biostatistics, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: The integrity of scientific research is crucial to scientific progress, yet a growing awareness of fraudulent publications undermines the trustworthiness of many areas of research. This is especially critical in biomedical research, where systematic reviews of Randomized Clinical Trials (RCTs) inform guidelines used by organizations like the National Institute for Health and Care Excellence (NICE) and The Cochrane Library. The inclusion of fraudulent RCTs, containing fabricated, falsified, or plagiarized data, can have severe consequences; members of our team have shown that just 27 retracted RCTs negatively affected the conclusions of over half of the 88 systematic reviews they were included in. This can lead to ineffective or harmful patient treatments, misdirect future research, and waste public funds. To combat this problem, research integrity experts have developed comprehensive checklists, such as 'INSPECT-SR', to help reviewers identify problematic trials. However, the manual application of this 26-item checklist is extremely labor-intensive and time-consuming, which severely limits its widespread adoption and slows the production of trustworthy clinical evidence. Our project, INSPECT-AI, aims develop a user-friendly, AI-assisted software tool to semi-automate and scale the integrity assessment of RCTs. The core of our approach is to combine the data and text extraction power of large language models with symbolic AI approaches. This will create a "smart data layer," using knowledge graphs to link integrity-related information across a network of publications and enabling automated, rule-based assessments. For example, the tool will be designed to perform checks such as: • Verifying if a paper has already been retracted or has an expression of concern listed in a database like Retraction Watch. • Identifying inconsistencies between the published article and its clinical trial registry data regarding participant eligibility, outcomes, or study dates. • Flagging mathematically improbable or impossible numbers in data tables, such as implausible baseline trial data or impossible variances for integer data. A key feature of our methodology is a co-design process involving key stakeholders. We are working directly with Project Partners from five leading evidence synthesis organizations, including NICE and Cochrane, to ensure the prototype is effective, efficient, and tailored to their real-world needs. By developing this tool, we aim to empower researchers (in particular systematic reviewers), journal editors, and funding bodies to more effectively safeguard the integrity of the scientific record, ultimately benefiting patients and society.

Rationale for assistance in data analytics and visualization: Our team possesses deep expertise in research integrity, alongside strong technical skills in core pipeline development and the application of LLMs. To accelerate our project, we seek assistance with several time-intensive challenges that well suited for a focused internship. These tasks are key to improving our project's performance. We would direct students to several key areas. A primary task is extracting and harmonizing data from multiple clinical trial registries, which lack a standardized format and present a significant data wrangling challenge. Another is evaluating different table extraction methods from PDFs to determine the most robust approach for our pipeline. Finally, we need to integrate existing statistical checks written in R into

our primary Python environment, or enhance the speed at which R can run in our Python environment. Assistance with this task would greatly streamline our workflow and increase the overall efficiency of our pipeline.

Statement of benefit to the student: This internship offers a good opportunity for a student to apply their data science skills to a project that we believe will have significant real-world application, given that we are co-developing this software with partners who will implement it in their workflows. The student will contribute directly to important tasks that will influence the development of INSPECT-AI, a tool designed to safeguard the integrity of biomedical research and protect patient health. They will tackle complex data engineering challenges, such as extracting and harmonizing data from unstructured PDFs and varied web sources, providing invaluable hands-on experience in NLP and data mining. The student will gain practical experience in evaluating machine learning models for table extraction and will see their work, such as porting R code to Python, potentially integrated directly into our software pipeline. Working alongside an international, interdisciplinary team of experts in AI and research integrity, the student will not only enhance their technical portfolio but also gain exposure to the emerging field of metascience.

Specific competencies required, including programming languages if applicable: Data Mining, Natural Language Processing, Database Management, Machine Learning, Python, R

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1042: Research-ready database of land and labour conflicts in Brazil from 1990 to 2025

Department: O'Neill School of Public and Environmental Affairs, Public and Environmental Affairs, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: This project will build a structured and research-ready database of land and labour conflicts in Brazil from 1990 to 2025, using scanned reports produced by the Comissão Pastoral da Terra (CPT). These annual bulletins contain detailed tables about conflicts, families affected, and violent events, but much of the data remains locked in non-digital formats. The project will use modern tools in Python, such as LayoutParser for table detection and OCR for Portuguese text recognition, to convert scanned tables into machine-readable form. Once digitized, the data will be cleaned, standardized, and linked to official municipality codes and geospatial boundaries from IBGE. The end result will be a reproducible database that researchers can use to study patterns of land conflict over time and space, and to connect with other datasets on environment, governance, and economic development. The MSDS student will help implement and document this pipeline, test extraction accuracy, design validation checks, and produce clear visualizations and summary statistics. Deliverables will include a set of structured datasets, a code repository, a quality assurance report, and example analysis notebooks. This work builds on the applicant's current research showing how digital land registries can reduce violent contestation and will enable extended analysis across the full 1990–2025 period.

Rationale for assistance in data analytics and visualization: This project requires advanced technical methods that go beyond routine research assistance. The scanned CPT tables are complex and vary in layout across years. Extracting these correctly demands experience with computer vision and optical character recognition in Python. In addition, standardizing place names, linking conflicts to IBGE municipality codes, and building a reproducible database requires database design skills and careful quality control. The assistance of a trained MSDS student will ensure that these steps are done in a systematic and reproducible way. Visualization expertise is also needed to check extraction accuracy and present patterns clearly. Without technical support, progress would be slow and error-prone, limiting the usefulness of the data. With FADS assistance, the project will not only succeed technically but will also produce tools and outputs that can be shared with the broader academic community.

Statement of benefit to the student: The student will gain valuable experience in applying cutting-edge data science tools to a real-world social science problem with global significance. They will learn how to use layout analysis and OCR for document digitization, handle multilingual text (Portuguese), and design structured datasets for long-term research use. The student will also gain practice in database construction, entity resolution, and data validation. In addition, they will contribute to scientific outputs with potential for co-authorship or acknowledgment in publications. This project offers a chance to bridge data science methods with questions of governance, inequality, and development. The mentoring plan includes weekly meetings, regular code review, and support for professional development. The experience will strengthen the student's portfolio for both academic and applied data science careers.

Specific competencies required, including programming languages if applicable: - Python programming - Optical Character Recognition (OCR) with Tesseract or PaddleOCR - Document layout analysis with LayoutParser - Data cleaning and transformation with pandas - Entity resolution and fuzzy matching (municipality names, actors) - Database design and SQL (PostgreSQL or SQLite) - Data visualization (matplotlib, seaborn, or Plotly) - Reproducible coding practices (Git, testing, documentation)

Is there anything else you would like us to know about your project's time frame or work schedule?: The project can be completed in either the Spring 2026 (January–April) or Summer 2026 (May–July) term. Workload is flexible, up to 20 hours per week. Early weeks will focus on setting up OCR and table parsing, mid-phase on entity resolution and database construction, and final weeks on quality assurance and visualization. Regular weekly check-ins will ensure steady progress.

Proposal Title: #1043: Transnational Social Movements and Backlash: A Case Study in Sentiment Analysis

Department: O'Neill School of Public and Environmental Affairs, Art Administration & Cultural Policy, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Summer 2026

Summary of project: The 2017 #MeToo Movement was a turning point in public discourse on sexual harassment and abuse. While the term "MeToo" was first employed by activist Tarana Burke in 2006, the 2017 Movement was sparked by reports of sexual harassment and abuse perpetrated by film producer Harvey Weinstein. Survivors of Weinstein's mistreatment, including former employees of his production company and Hollywood actresses, shared their experiences in the press and on social media. The #MeToo hashtag went viral and resulted in millions of people around the world sharing their stories of harassment and abuse (Saguy and Rees 2021). Recently, there has been a backlash against #MeToo and the alleged disruptions to the careers of people accused of sexual harassment in connection with the movement (Cossman 2021; Dodson et al. 2023). Although much of the scholarly and popular literature on #MeToo has focused on the U.S. case, a growing body of research uses computational social science methods to examine how the movement expanded globally and developed into a transnational phenomenon (Lee & Murdie 2021; Stubbs-Richardson et al. 2024; Suk et al. 2024; Sweeny 2020). Our project contributes to this comparative scholarship by utilizing the Global Database of Events, Language, and Tone (GDELT), a massive open dataset that monitors both news and social media in more than 100 languages to track global events, narratives, sentiment, and tone. Since 1979, GDELT has captured daily, society-wide patterns of behavior and is accessible via Google Cloud's BigQuery. Researchers have employed it to study protests, conflicts, disasters, media coverage, and predictive analytics. For this project, GDELT provides a unique opportunity to conduct sentiment tracking and examine the emotional polarity of: a)international media coverage of #MeToo and the subsequent backlash to the movement b)public statements made by #MeToo victims and accused perpetrators Such analyses will provide unique insight into how the media and civil sphere interpreted the experiences of victims and perpetrators in transnational contexts during and after #MeToo. This research will be part of a larger project on power and labor in Hollywood (Dessauer 2025) and the backlash against #MeToo. The GDELT data will not only provide a global context for this research, but also help measure how victims and accused perpetrators' statements affect their career trajectories.

Rationale for assistance in data analytics and visualization: The scale and complexity of the GDELT dataset make it essential to have dedicated support in data extraction, analysis, and visualization. GDELT spans multiple decades and includes rich sentiment and tone data, requiring advanced computational techniques to handle effectively. While Professors Julia Dessauer, a cultural sociologist, and Nilesh Shinde, an economist specializing in quantitative methods, provide strong supervisory expertise, the breadth of this dataset exceeds what can be managed without additional support. A data analyst will play a critical role in processing GDELT's high-volume data, applying natural language processing methods, and creating clear visualizations that connect sentiment patterns to media narratives. This expertise will ensure the project's methodological rigor while allowing the research team to focus on

interpretation and theoretical contribution. Without this additional assistance, the project's ambitious goal of situating #MeToo and its backlash within a global media context would be difficult to realize.

Statement of benefit to the student: This project offers an MSDS student a unique opportunity to gain experience with real world, large-scale datasets related to a timely research topic. The student will utilize and strengthen their skills in Natural Language Processing, SQL, and R, while also learning about contemporary social movements. The student will have the opportunity to become a co-author on related publications.

Specific competencies required, including programming languages if applicable: Natural Language Processing, SQL, R

Is there anything else you would like us to know about your project's time frame or work schedule?: The student will be supervised by both Julia Dessauer and Nilesh Shinde, assistant professors in O'Neill.

Proposal Title: #1044: The AI Public Management Map: Visualizing Adoption and Impact in Social Services Agencies

Department: O'Neill School of Public and Environmental Affairs, Public Affairs, Indianapolis

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: State and local social services agencies face increasing pressure to deliver programs more efficiently, equitably, and transparently in the face of resource constraints and rising citizen needs. Artificial Intelligence (AI) technologies have emerged as powerful tools for enhancing administrative capacity, from optimizing the allocation of scarce resources to improving decision-making and accountability in service delivery. Yet, little is known about how public agencies—particularly in the social services domain—adopt, implement, and govern AI in their day-to-day operations. This project will study the adoption and implementation of AI across state and local social services agencies in three key areas: • Planning program priorities – how agencies use AI to forecast needs, target interventions, and set strategic agendas. • Managing applications – how AI is integrated into intake systems, eligibility determinations, and administrative workflows. • Evaluating employee performance – how agencies use data-driven tools to monitor, reward, and discipline workers. The study will generate comprehensive empirical data of AI in public management by combining qualitative data (agency documents and reports) with quantitative data (administrative datasets, adoption records, and publicly available performance dashboards). A key innovation of this project is the creation of an AI Public Management Map—an interactive visualization that depicts where and how AI is being used in social services across the United States. This tool will highlight geographic variation, policy drivers, and program contexts of AI adoption. For scholars, it will provide comparative insights into the evolving role of technology in governance and empirical data for further research. For practitioners and policymakers, it will serve as a resource for identifying best practices and anticipating challenges such as bias, accountability gaps, or resistance among frontline workers. The study builds on my broader research agenda on representative bureaucracy, administrative burden, and public service governance. By focusing specifically on AI adoption in social services, the project advances theoretical debates on technology, equity, and democratic accountability in public administration. It also addresses pressing policy concerns, as AI decisions affect vulnerable populations who rely on state and local social programs. Through this project, I seek to answer three guiding questions: • What organizational, political, and contextual factors drive AI adoption in social services? • How do agencies implement AI tools in ways that affect citizens, employees, and managers? • What are the implications of AI use for equity, transparency, and accountability in public management?

Rationale for assistance in data analytics and visualization: The project requires advanced data analytics and visualization support to build the AI Public Management Map, an interactive platform that synthesizes diverse datasets on AI adoption across social services agencies. This tool will require integrating data from multiple sources (government reports, administrative datasets, agency dashboards, and survey data) into a user-friendly and visually intuitive format. Assistance from IU's data science consultants will be crucial in three areas: • Data integration and cleaning – consolidating unstructured and structured data from state and local agencies into a consistent, analyzable format. •

Advanced visualization – designing dynamic, interactive mapping and dashboard tools that allow users to explore variation across agencies, states, and policy domains. • Analytics – applying text mining, clustering, and geospatial techniques to uncover patterns in AI adoption and implementation. The support will enable me to move beyond traditional academic outputs and produce a publicly accessible resource with both scholarly and policy impact. By leveraging IU's expertise in data science, the project will model how applied analytics can advance public administration research and provide practical tools for practitioners.

Statement of benefit to the student: This project offers students a unique opportunity to engage at the intersection of public administration, social policy, and data science. By assisting in the development of the AI Public Management Map, students will gain hands-on experience in data collection, cleaning, integration, and visualization while working with real-world datasets drawn from government sources. Substantively, students will explore the range of social services provided by state and local governments, including programs in healthcare, child welfare, employment services, and income support. They will learn how AI is transforming administrative processes and consider the ethical, social, and equity implications of technology adoption in public management. Methodologically, students will strengthen their skills in data wrangling, geospatial analysis, and interactive visualization design—skills highly valued across academic, public sector, and private sector careers. Working alongside faculty and data science consultants, students will also gain exposure to interdisciplinary collaboration and applied research practices. Ultimately, the project will enhance student capacity to combine technical competencies with substantive policy knowledge, preparing them for leadership in data-driven public affairs and equipping them with a competitive skill set for both academic and professional trajectories.

Specific competencies required, including programming languages if applicable: Data Mining, Data Visualization, Database Management, Machine Learning, Natural Language Processing, Web & Social Media Mining, GIS, Python/R, Shiny(R) or Dash(Python) for interactive visualization

Is there anything else you would like us to know about your project's time frame or work schedule?: No specific timeframe.

Proposal Title: #1045: Integrating Social Media and Mobility Data for Real-Time Shelter Demand Forecasting

Department: Kelley School of Business, Operations & Decision Technologies, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Summer 2026

Summary of project: Natural disasters are becoming more frequent and severe due to climate change, posing serious threats to human health and safety. When disasters strike, response organizations must act quickly to provide critical resources to affected communities—chief among them, safe and accessible shelters. From an operations management perspective, the success of shelter operations depends heavily on accurately predicting demand. However, the high level of uncertainty during disasters—combined with unpredictable individual responses to evacuation guidance—makes accurate forecasting especially difficult as events unfold. This project aims to improve real-time shelter demand forecasting by integrating social media data with offline mobility information to better predict traffic to shelters as disasters evolve. Understanding how, when, and where people move in search of shelter is essential for effective planning and resource allocation. Current methods for predicting shelter needs typically consider factors like population density, disaster severity, and demographics. However, we believe social media activity also plays a crucial role, influencing shelter usage patterns and improving the accuracy of demand forecasts. In this research, we will collect and analyze data from social media platforms (such as X and Facebook), real-time foot traffic to shelters, and demographic information from disaster-affected areas. Using AI-powered models, we will examine online conversations to identify behavioral patterns and key factors that influence shelter selection, particularly among vulnerable populations (such as those from low- and middle-income groups or individuals with special medical needs). By combining these insights into an empirical model, our approach aims to significantly improve real-time forecasting and support disaster response organizations in making smarter decisions about resource allocation and service delivery.

Rationale for assistance in data analytics and visualization: This project requires advanced analytic and visualization expertise to integrate, model, and interpret large-scale, heterogeneous data. We will merge high-resolution mobility data with diverse social media datasets. Additionally, this research will require complex statistical methods, including time-series modeling to test whether social media activity predicts shelter demand and natural language processing to analyze tone and urgency of messages.

Statement of benefit to the student: This project offers the student a unique opportunity to develop advanced skills in data science and applied social research while contributing to a socially impactful study. The student will gain hands-on experience with large-scale mobility and social media datasets and learn techniques in data integration, statistical modeling, natural language processing, and geospatial analysis. In addition, the student will work closely with faculty to translate complex results into visualizations and decision-support tools, providing valuable exposure to interdisciplinary collaboration.

Specific competencies required, including programming languages if applicable: Python, AWS

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1046: Optimizing Deep Brain Stimulation in Parkinson's Disease to Minimize Side Effects

Department: Medicine, Neurological Surgery, Indianapolis

Project requires Human Subjects? Yes

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Deep brain stimulation (DBS) is a neurosurgical therapy used to treat movement disorders such as Parkinson's disease, essential tremor, and dystonia. In this therapy, electrodes are surgically implanted in specific areas of the brain, where they deliver electrical pulses to modulate abnormal brain activity and improve symptoms. In Parkinson's disease, DBS is applied to the subthalamic nucleus (STN), a small structure in the midbrain. STN DBS is effective in improving major motor symptoms, including tremor, rigidity, and bradykinesia. However, in some patients, it can also produce side effects such as dyskinesia and changes in cognition or mood. A typical DBS lead contains 4-6 electrode contacts. Stimulation can be applied through a single contact or multiple contacts and can be configured in either a monopolar or bipolar fashion. Furthermore, the stimulation current can be individually adjusted to tailor therapy for each patient. DBS exerts its therapeutic effect mainly by activating axonal fibers passing near the electrode. The internal capsule, a major white matter structure containing fibers that project to diverse regions such as the motor cortex, brainstem, and pedunculopontine nucleus, lies adjacent to the STN. Small variations in electrode placement or improper configuration can activate unintended sets of fibers, which may lead to side effects. This project will focus on patients who experience side effects from DBS, aiming to adjust DBS settings to minimize adverse effects while maintaining therapeutic benefit. Since DBS patients undergo CT and MRI scans, we can determine the electrode locations in the patient's head. By transforming standardized brain atlases into each patient's space, we can reconstruct various brain structures and fiber pathways specific to the individual. Using finite element modeling, we can also calculate the electric potential at each point in the brain. This allows us to estimate which fibers are likely to be activated by stimulation and to optimize stimulation settings accordingly.

Rationale for assistance in data analytics and visualization: Clear visualization of the target brain structures, passing fibers, electrode placement, and the surrounding electric field will be crucial in this project. Such visualizations will not only aid in understanding the spatial relationships between stimulation sites and nearby neural pathways but also provide actionable insights into how therapy can be optimized while minimizing side effects. Assistance in data analytics and visualization will therefore play a central role in transforming complex imaging and modeling output into interpretable and clinically meaningful representations. This work will also serve as an excellent opportunity for students to apply their data science and visualization skills to a real-world medical application. It provides the chance to work with clinical neuroimaging data, deepening their understanding of data standards and efficient representation of complex datasets. By developing visualizations that are both rapid to generate and computationally lightweight, students will gain experience balancing scientific rigor with practical usability.

Statement of benefit to the student: Participation in this project will provide students with hands-on experience in applying data science and visualization skills to real-world medical data. Students will work directly with advanced neuroimaging techniques, computational modeling, and cutting-edge visualization tools used in brain research. They will gain practical knowledge in areas such as medical image analysis, interface development, and the translation of scientific data into meaningful visual representations for clinical and research applications. Ultimately, this experience will prepare students for future careers in neuroscience, medical data science, and related fields, equipping them with valuable skills and insights highly sought after in both academia and industry.

Specific competencies required, including programming languages if applicable: • Expertise in Python and/or Matlab and knowledge of data formats. • Understanding of key linear algebra concepts, include matrix, transformation, and coordinate system. • Familiarity with neuroimaging datasets, visualization tools such as ParaView, 3D Slicer, and the Qt Framework will be an asset.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1047: The IU Policy Map

Department: SPEA, SPEA Other, Indianapolis

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: This research aims to advance public administration and public policy scholarship by identifying policy issues that are most prevalent at the local level in the United States. The objective of this project is to develop a large-scale, multi-state database of local-level issues (e.g., housing, disaster preparedness, economic development) across time and space sourced from publicly available meeting agendas and minutes. Meeting agendas and minutes offer a rich account of city government and agency priorities; however, they are currently a highly underutilized source of unstructured text data. This project, named The IU Policy Map, is the first of its kind, to our knowledge, that will collect, document, analyze, and map written records of local government public documents, and identify issues most prevalent at the local level using Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques across time and space. To implement this project, we have so far collected the general website and the direct link to meeting minutes of 1,004 municipal governments in the Midwest Census Region of the United States, including cities in the following states: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin.

Rationale for assistance in data analytics and visualization: This project requires data science assistance with building a database schema; developing code to (1) implement the initial retrieval of publicly available documents, (2) conduct routine automated retrievals of publicly available documents, and (3) develop code to identify and categorize policy issues in publicly available documents using Intelligent Document Processing (IDP) and Named Entity Recognition (NER); and, developing dashboards that highlight the most prevalent issues discussed at the local level across space and time.

Statement of benefit to the student: The benefit to the students is twofold. First, it provides hands-on experience with text analytics in a civic context, which is an area with growing demand for data science expertise in both public and private sectors. Second, the project emphasizes translating technical outputs into actionable insights by creating dashboards and spatial visualizations that highlight patterns in local policy agendas. The students will thus gain experience not only in back-end data engineering and coding, but also in front-end communication of results to academic and practitioner audiences. Finally, because this project sits at the intersection of computer science, public administration, and policy analysis, the students will expand their interdisciplinary knowledge base and learn how we can use data science to make meaning of underutilized publicly available resources. The resulting database and analyses will be among the first of its kind, positioning the students as contributors to a pioneering initiative with strong potential for academic publications and professional recognition.

Specific competencies required, including programming languages if applicable: The ideal candidates will have a background in SQL and Python, and familiarity with the Google Colab Jupyter Notebook service.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1048: Evaluating Inference Methods for Complex Dynamical Time Series

Department: Luddy School of Informatics Computing and Engineering, Computer Science, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: The project aims to test underexplored machine learning strategies on time series representing loosely related, complex dynamical systems such as in ecological population dynamics. The ultimate goal is to infer process models relating states at different times that exploit hidden commonalities between systems. However, the more near-term goal for this spring is to test some emerging analysis techniques on this kind of data, seeing if one can better extract predictively useful information. Commonly applied techniques have historically struggled with ecological data, often failing to outperform the 'naive' model that merely posits continuity over time. Furthermore, these data often come in short, numerous series that promote overfitting and are difficult for usual 'big data' approaches. When individually fitting short time series, simplistic models can outperform more complicated models even when the latter are more mechanistically accurate. Nonetheless, it's widely suspected that these systems do reflect complex underlying processes. Such data often appear to have visible non-random features, such as partial periodicities, that are not simple enough for elementary Fourier analysis or similar. The ultimate goal of proposed analyses is to discover underlying patterns that are masked by differences in magnitudes, periodicities, and other effective parameters. These patterns will reduce the effective parameter space, potentially escaping tendencies that would favor simplicity over mechanistic accuracy. Initial, biological data sources include the relatively small Global Population Dynamics Database and larger, more modern BioTIME database. Each contains populations over time for a large number of biological species, locations, and other characteristics. These data represent complex dynamical systems that are supposed to follow physics-like equations of motion, but with far more real-world complications than are common in traditional physics domains. Data might be combined with publicly available weather and other data that are expected to be statistically correlated with biological population data. Work may also analyze non-biological dynamical systems, such as in hydrology and weather. Finally, the project may also consider artificially simulated data to isolate specific hypotheses from the many complications of real-world, noisy time series. Specific methods to test initially include Fourier analysis networks, random features, sparsity promotion, and cross-learning. Further methods will combine information-theoretic ideas with dynamical systems. Proposed work will also attempt to classify data into distinct types that are susceptible to different analysis methods. Strategies to classify data by optimal analysis method will further seek correlations with biological features, such as a species's trophic role. A successful initial outcome connects quantitative gains in hindcasting performance to qualitative hypotheses about underlying mechanisms.

Rationale for assistance in data analytics and visualization: The PI's primary research area is quantum computing, combining the mathematical physics of underlying dynamics with computational and information-theoretic paradigms. Inferring models of dynamical systems presents an analogous setting, also combining physical and computational ideas. As quantum computing is a nascent technology that has yet to realize much of its potential, the PI's quantum research is primarily theoretical. Motivating

this proposed project is a desire to bridge lessons learned from quantum information to a more mature domain with available data and immediate practical implications. The PI brings domain knowledge in mathematical physics and information theory. The master's student is expected to complement this background with knowledge of state of the art machine learning and data analysis tools.

Statement of benefit to the student: Nvidia CEO Jensen Huang recently said, "For the young, 20-year-old Jensen, that's graduated now, he probably would have chosen ... more of the physical sciences than the software sciences." Sameer Samat of Google said of computer science, "It's definitely not learning to code… It is the science, in my opinion, of solving problems." As artificial intelligence increasingly automates common software development and data science workflows, the human's role more strongly emphasizes higher-level problem solving, cross-cutting ways of thinking, and understanding the physical environment. This research project will supplement the student's machine learning background with dynamical systems and more physically inspired ways of thinking. The student will learn how to translate between theoretical hypotheses and their applications and validations via real data. These additional skills will help the future graduate stand out in a field where entry level jobs are becoming more competitive.

Specific competencies required, including programming languages if applicable: Skills required: machine learning, deep learning, Python. Preferred: signal processing, data visualization. Note: could substitute Rust, Julia, or C++ for Python if the student is sufficiently skilled to implement deep learning and time series analysis in one of these and to replace common Python-based toolkits.

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1049: Data management and visibility support for the Connection Cafe Street Medicine Program

Department: School of Public Health, Department of Applied Health Science, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: The Connection Cafe Street Medicine (CCSM) Program is a recently funded academic-community partnership between Indiana University School of Public Health-Bloomington, the Fayette County Connection Cafe, and IU Health Positive Link HIV Services. This 4-year initiative aims to effectively respond to the needs of people with HIV and Hepatitis C (HCV) who are out of care and unsheltered in a 12-county eastern Indiana region, an area heavily impacted by the effects of substance use disorder. The Fayette County Connection Cafe is a leader in eastern Indiana for recovery services, and the CCSM Program will expand the Cafe's current work to serve more people with HIV in need of comprehensive medical, mental health, substance use, and social services through (1) adding medical staff and medical services to the current street outreach program and (2) expanding the geographic reach of the program by adding five counties to the seven currently served. The CCSM Program will build on current outreach efforts by bringing low-barrier, compassionate and client-centered HIV, HCV, and other needed medical care to individuals in unsheltered spaces where they live, spend time, and congregate. With a reliance on robust community partnerships, an established history of recovery and street outreach support in the community, evidence-based implementation science and evaluation frameworks, and rigorous scientific processes, we expect that the CCSM Program will have a measurable positive effect on a significantly disadvantaged region in rural Indiana. Further, through sharing program results as one of ten national demonstration sites, other organizations will be able to replicate the intervention, increasing the evidence base within the field and improving public health outcomes nationwide. We will be building the CCSM Program infrastructure during the 2025-2026 academic year. Part of this process is developing a secure system for recording, monitoring, and reporting participant data that can be used easily and reliably by multiple program staff across organizations and in the field. This system will be critical for tracking and assessing program activities and outcomes. Additionally, the Fayette County Connection Cafe website—created at the Cafe's inception—is now outdated and in need of significant revision. It would greatly benefit from updates including the addition of the CCSM Program so that anyone interested in seeking services or learning more about them will be able to do so.

Rationale for assistance in data analytics and visualization: The CCSM Program team is trained and experienced in public health, intervention development and evaluation, clinical medicine, and recovery and support services. However, we do not have the technical expertise required to build the robust database and website infrastructure described above. While we could create rudimentary systems, they would likely be inefficient and create obstacles to achieving our long-term goals. By contrast, systems designed by individuals with greater technical skills would deliver far stronger performance and better support for our team. We now realize that in our early planning we did not fully anticipate the need for deeper technical support. Since beginning the actual CCSM Program preparation and determining what will be necessary for our documentation and reporting, we better recognize how valuable having more

advanced technical systems will be. When we saw the call for FADS applications, we were extremely excited about this potential partnership opportunity that could both strengthen the CCSM Program while providing a meaningful and applied experience for students.

Statement of benefit to the student: We expect that this project will yield multiple substantive benefits for FADS interns. Students will gain hands-on experience directly aligned with their degree competencies, applying what they have learned to a real-world public health initiative. They will also contribute critical support to an impactful public health program and see firsthand how their skillset can be used for public good. Working with a multidisciplinary team of faculty researchers, clinicians, and non profit professionals will broaden their network and strengthen their professional and interpersonal skills. Moreover, mentoring students and treating them as part of the CCSM Program team is something we deeply value and commit to doing. Ultimately, we believe that the skills, experiences, and concrete deliverables that FADS interns will gain through this work will be essential for their professional development and will position them well in their future job applications and careers.

Specific competencies required, including programming languages if applicable: Data visualization; Database management; Security/privacy management; Web front-end development

Is there anything else you would like us to know about your project's time frame or work schedule?: N/A. Please note that we selected that this project does not involve human subjects because in the coming year we will only be establishing systems for the program and will not be working with participants directly. We appreciate you making this opportunity available.

Proposal Title: #1050: OmniSOC Data Liasion Support - Facilitating Access to Operational Cybersecurity Data for Advanced AI-enabled Analytics

Department: Kelley School of Business, Operations and Decision Technologies, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: This project involves the deployment and implementation of a modern, open-source data ingestion tool, such as Airbyte, to create automated and scalable data pipelines for Data Science and Artificial Intelligence Lab at Kelley School of Business. The primary goal is to extract data, transform and load security operations center data from the OmniSOC (higher education's first security operations center) into a designated database to support advanced Artificial Intelligence (AI)-enabled cybersecurity analytics research to be positioned for National Science Foundation (NSF) grants. This foundational structure will ensure that our AI/ML researchers have timely, reliable, and analysis-ready data to fuel their machine learning models and experiments. Students involved in this project will have significant experience working with operational cybersecurity environements that could facilitate adviance AI-enabled cybersecurity analytics. Specific areas of work will include data landing from OmniSOC's multi-source data spots, analytics on the landed data, and reporting/artifact development back to the OmniSOC environment. Selected student outputs will also be included into NSF-funded grant proposals targeted at the NSF SaTC, CICI, and SFS programs.

Rationale for assistance in data analytics and visualization: The success and speed of our AI projects are heavily dependent on the quality and accessibility of rich, real world cybersecurity data, a resource that is scarce within the cybersecurity community at the moment. There is currently no process or technology in place that would enable the sharing of the data to our researchers from OmniSOC (which was designed for) and would need to implement data pipeline from scratch. By dedicating a student to this project, we can provide the focused effort required to build a robust, automated data pipeline infrastructure. The student will play a critical role in establishing the backbone of our data operations with the help of the data engineer within OmniSOC and facilitate access to advanced AI-enabled cybersecurity analytics.

Statement of benefit to the student: This position offers a fantastic opportunity to gain practical, hands-on experience at the crucial intersection of data engineering and artificial intelligence. You won't be learning theory, you will be building, deploying, and managing the essential data infrastructure that powers cutting-edge AI research. By the end of the project, you will have learned to: Implement a modern data stack: Deploy and manage an industry- standard data ingestion tool like Airbyte to solve real-world data challenges Build Automated Pipelines: Design and orchestrate robust data pipelines that extract data from various sources and prepare it for AI applications Master In-Demand Skills: Gain significant practical experience with highly sought-after technologies like SQL, Python, and Docker within a professional research environment Understand the AI Data Lifecycle: Witness firsthand how data moves from a raw source to a clean, usable format that directly impacts the performance of machine learning models Collaborate with Experts: Work alongside AI researchers and data scientists,

gaining invaluable insight into their workflows and data needs This experience will provide you with a powerful skill set and a compelling portfolio piece, giving you a significant competitive advantage for future roles in data engineering, MLOps and data science.

Specific competencies required, including programming languages if applicable: Programming skills – Basic proficiency in Python and experience with basic scripting Database knowledge – Familiarity with SQL for querying and understanding of relational database concepts (schemas, tables, data types) Technical Aptitude: Comfortable working in a command-line environment (Linux/Unix) and a strong desire to learn new data tools Bonus Skills – Any exposure to Docker, cloud platforms and data lakes

Is there anything else you would like us to know about your project's time frame or work schedule?:

Proposal Title: #1051: Intelligent Knowledge Graph Development for Physics-First 3D Modeling in Architecture, Engineering, and Construction

Department: Eskenazi School, Comprehensive Design, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: I'm comfortable working with students in either term.

Summary of project: This project aims to develop an innovative knowledge graph system integrated with a physics-first 3D modeling software specifically designed for the Architecture, Engineering, and Construction (AEC) industries. The system will revolutionize how designers and builders approach digital fabrication by creating an intelligent archive that learns from past projects and informs future design decisions through advanced data analytics and machine learning. Research Background and Significance Drawing from over 17 years of experience in computational design (2008-2025), I bring unique expertise to this project. As co-founder of the Advanced Computational Group at Pelli Clarke Pelli Architects, I collaborated directly with computer scientists to develop innovative tools including smart campus planning software. This experience involved developing strategy, defining feature landscapes, and extensive bug fixing in collaborative software development environments. My recent academic includes: Leading teams in developing novel structural joint systems with capabilities beyond existing systems Pioneering the use of AI tools in architectural design, including machine learning style transfer processes Creating full-scale experimental structures that test new materials and technologies Publishing peer-reviewed research on computational design methodologies Exhibiting innovative work at international venues including the digitalFUTURES conference

Rationale for assistance in data analytics and visualization: The proposed knowledge graph system addresses a critical gap in current AEC software tools, which lack the ability to learn from project histories and integrate material properties with physics-based modeling at the foundational level. The system will be trained initially on my personal archive of over 20 years of project files, but the ultimate goal is to create a framework that can survey any user's project archive to provide personalized, AI-driven design assistance based on their own work history.

Statement of benefit to the student: Learning Opportunities for MSDS Student This project offers exceptional learning experiences across multiple data science domains: Real-world Application: Direct application of data science to solve industry-specific challenges Interdisciplinary Collaboration: Work with architecture, engineering, and design professionals Cutting-edge Technology: Experience with emerging technologies in AI and knowledge graphs Research Experience: Contribute to publishable research with clear attribution Industry Exposure: Potential for networking with AEC industry professionals and software developers

Specific competencies required, including programming languages if applicable: Primary Focus Areas: Knowledge Graph Development Design and implement graph database structures for AEC project data Develop ontologies for architectural elements, material properties, and construction methods Create automated data extraction pipelines from project archives Natural Language Processing Implement

NLP techniques to extract meaningful information from project documentation Develop semantic understanding of architectural and engineering terminology Create automated tagging and categorization systems for project elements Machine Learning Integration Develop recommendation algorithms that suggest design solutions based on project history Implement predictive models for material performance and structural behavior Create learning systems that improve accuracy based on user feedback Database Development and Management Design scalable database architectures for large project datasets Implement efficient querying systems for real-time software integration Develop data versioning and backup systems Secondary Areas: Data Visualization for project relationship mapping and performance analytics Web/App Development for user interface components of the knowledge graph system

Is there anything else you would like us to know about your project's time frame or work schedule?: Alignment with Faculty Research Trajectory This project directly supports my research goals in: Sustainable Building Innovation: Physics-first modeling enables better material optimization Proto-architectural Sculpture: Advanced tools support full-scale experimental construction Computational Design: Expanding the frontier of AI applications in architecture Community Engagement: Better design tools lead to more effective public installations Expected Deliverables Technical Deliverables: Functional knowledge graph database with sample AEC data API for integration with 3D modeling software Documentation and user guides Performance benchmarks and testing results Academic Deliverables: Conference paper submissions (target: CAADRIA 2027, CHI 2027, or UIST 2027) Technical report for IU research repository Progress report for future grant applications Educational Deliverables: Student learning outcomes assessment Best practices documentation for similar projects Recommendations for curriculum integration Timeline and Milestones Weeks 1-2: Project setup, data assessment, and architecture planning Weeks 3-6: Core knowledge graph development and initial data integration Weeks 7-8: NLP implementation for document processing Weeks 9-10: Machine learning model development and testing Weeks 11-12: Integration testing, documentation, and future planning Spring 2026 timeline (January 19 - April 24) provides optimal timing for proof-of-concept development leading into summer fabr

Proposal Title: #1052: REDACTED: THE USES OF SECRECY IN WAR CRIMES TRIALS PHASE III

Department: Maurer School of Law, Law, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Trials rely on transparent process for their authority. So do war crimes trials, which often add a goal or promoting reconciliation by creating definitive accounts. Yet curiously, one of the essential devices of international criminal tribunals is the disabling of narrative through secrecy. Techniques of secrecy – closed sessions, protected witnesses, restricted and redacted documents – are regular features of international trials, used much more than in domestic courts. There are compelling reasons for this, but secrecy also creates a challenge for these courts' reconciliatory goals. Redacted will study how secrecy is produced and used at the eight major war crimes courts of the contemporary era: the International Criminal Court, the International Criminal Tribunals for the Former Yugoslavia and for Rwanda and their successor the Residual Mechanism, the Special Tribunal for Lebanon, the Kosovo Specialist Chambers, and the Extraordinary Chambers in the Courts of Cambodia. How is secrecy created, what purposes does it serve, and what are its effect on the usefulness of war crimes courts? The project will result in a book, the first to study secrecy in war crimes trials, and as such will make a novel contribution to the rapidly growing literature on international courts. The book is now under contract with Cambridge University Press. Redacted's analysis proceeds in four parts: a taxonomy, a map, a context, and a critique – and the second in particular would benefit from quantitative analysis. The first part is a taxonomy of secrecy's forms. There are many: Witnesses testify under protective measures; parties file documents that only judges can see; states negotiate restrictive conditions on public use of intelligence. I describe each type, examining their use in particular trials. A second part measures the amount of material kept secret. With the help of FADS-supported researchers, I have been trying to estimate how much material on the courts' public websites includes some form of redaction. I hope to show which courts, cases, actors, and trial phases rely the most on redaction. This is the part of the project that FADS support will help develop. A third part considers the institutional context. Secrecy is essential to these tribunals' operations and does not necessarily make trials unfair. Secrecy is both strategic and routine. So, I lay out the rationales and operations of secrecy – how it is produced, and how its producers understand what they are doing. Much of this section builds on interviews with insiders at these tribunals: judges, prosecutors, defense attorneys, and archivists. Finally, in a fourth part, I critique those rationales and operations in light of these courts' broader goals. Secrecy literally reconfigures the narrative of trial: Authoritative judicial truths are supposed pave the way for reconciliation, but secrecy increases scepticism about judgments. Why should we believe a judgment whose reasons are hidden?

Rationale for assistance in data analytics and visualization: It is difficult to study what isn't visible, but that is what I am trying to do. To date, I have conducted my research through qualitative examination of public documents – access to non-public materials is routinely denied – and through open-ended interviews with court officials. But given the size of the documentary databases involved – over 100,000 documents, with millions of pages, many in non-digitized image formats – and the hidden or partly hidden nature of much of the material, there are serious limits to what this approach can achieve. My

expectation is that quantitative, statistical analysis could provide important, even unexpected insights into the frequency of redactions and secret documents and the circumstances in which they are used.

Statement of benefit to the student: The project offers supported students the opportunity to devise a complex research strategy from start to finish, in communication with a non-specialist. The students would have to develop a complex strategy for collecting, identifying and processing several very large datasets, including procedures for identifying evidence of non-public documents (which might be mentioned in public documents, for example) and for evidence of redactions within documents. Based on the experience I've had with other FADS-supported researchers who have worked on this project, this involves devising bespoke search strategies to deal with a highly heterogeneous set of document inputs – strategies for searching in one court won't work in another, for example. And, as noted, all of this would have to be worked out with a non-specialist – a valuable learning experience for student and faculty alike! It's a good project for a student who is interested in a creative, open-textured puzzle, and who likes looking for what isn't there.

Specific competencies required, including programming languages if applicable: The project requires the construction of a database or CSV files, strategies for extracting meta-data, and optical character recognition. Additional skills include data mining, data visualization, natural language processing, statistics, and possibly network analysis and machine learning. The problem could be analogized to the kinds of large language models used in AI training, though perhaps more akin to the sorts of projects that have been developed for digital humanities analysis. (The project involves secrecy, but we are not seeking access to the actual secret materials, so there would be no need for specific security and privacy management skills.)

Is there anything else you would like us to know about your project's time frame or work schedule?: I was fortunate enough to receive FADS funding for this project in two previous cycles. I had hoped the project would be completed, but in the nature of things it has taken longer than anticipated, for two reasons: expansion of the project from four to eight courts has increased the target document base (and the complexity of the searches required), and complications have (inevitably) arisen in solving the technical problems of how to search for redactions and scale the processing of documents at speed. But, based on the assessments of students who worked on the last iteration, I am reasonably optimistic that this application will allow me to complete the project and incorporate its findings into my book.

Proposal Title: #1053: GeoAI with 3D Computer Vision for Natural Disaster Monitoring and Response

Department: Luddy School of Informatics, Computing, and Engineering, Department of Computer Science, Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: This project applies cutting-edge 3D-aware AI and deep learning to monitor post-disaster environments using aerial, satellite, drone imagery, LiDAR point clouds, and terrain data such as USGS DEMs. Students will prepare labeled datasets by interactively annotating objects (e.g., fallen trees, debris, road segments) and "stuff" (e.g., flood water, fire scars, smoke) in 3D environments, supported by foundation models like SAM2 and SAM2Point that enable rapid annotation with minimal clicks and fine-tuning. With these datasets, students will train 3D computer vision models to detect, segment, and track critical features, turning unstructured pixels and point clouds into structured spatial data. On top of this, disaster response applications will be developed, such as identifying houses at risk near floods or wildfires, estimating water depth or damage to homes and facilities, recalculating evacuation and rescue routes that avoid blocked or flooded roads, and designing terrain-aware alternative routes within given constraints. This research will provide students with hands-on experience at the intersection of AI, geospatial data, and disaster resilience, with direct societal impact.

Rationale for assistance in data analytics and visualization: As a project in 3D spatial setting, visualization and annotation for data curation is an essential step, and students can even extend our in-house 3D annotation tool with new features to meet their needs. Once AI models are trained to turn unstructured vision data into 3D spatial data, applications on top heavily uses data analytics to bring values in disaster response. Moreover, training 3D computer vision models itself would be an interesting and valuable experience.

Statement of benefit to the student: Students will gain experience in 3D computer vision and its applications.

Specific competencies required, including programming languages if applicable: Python, Computer Vision, Data Visualization, Cloud & High Performance Computing, Database Management, Deep Learning

Is there anything else you would like us to know about your project's time frame or work schedule?: Preference is Spring 2026, but Summer 2026 is also OK if Spring is not possible

Proposal Title: #1054: Development and evaluation of novel photo and audio biomarkers of accelerated biological aging using deep learning models

Department: College of Arts and Sciences, Sociology, Bloomington

Project requires Human Subjects? Yes

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: Photo- and audio-derived data are a promising source for development of low-cost, noninvasive biomarkers of aging, but more work is needed to establish their predictive performance in relation to established biological aging measures and all-cause morbidity and mortality. Research on photo biomarkers and audio-derived biomarkers (i.e., analyzing patterns of voice and speech) shows that visual and acoustic features and deep-learning models can reliably predict chronological age and detect health-related states such as frailty, cognitive decline, hospitalization risk, and mortality. Photo and audio-predicted "biological aging" measures have been proposed, but direct validation against molecular biomarkers—especially DNA-methylation epigenetic clocks—is limited and results are preliminary (e.g., based on small convenience samples). In the current study, address this gap by completing the following specific aims: Aim 1. Develop photo-derived age acceleration estimates using human coding and deep learning and assess whether they associate with chronological age and DNA methylation-based epigenetic aging clocks. Aim 2. Develop audio-derived (i.e., from voice and speech) age acceleration estimates using human coding and deep learning and assess whether they associate with chronological age and DNA methylation-based epigenetic aging clocks. Aim 3. Leverage machine learning models and multimodal signals (photo, speech/voice, and DNA methylation biomarkers) to prospectively assess predictive performance of complementary aspects of biological aging on morbidity and all-cause mortality. We leverage a unique large state representative dataset (N=2,390) containing respondent photos, audio recordings of responses to an open-ended questions, extensive survey data, a range of morbidities, electronic health records, anthropometric data (e.g., blood pressure, resting heart rate), and DNA methylation data. In this pilot project, we will elicit human ratings of perceived age using photos and audio recordings for an age-, race-, and gender-stratified subsample of 400 respondents. We will seek to collect ratings for each set of data artifacts (photo and audio) from 15 independent raters of different ages, races, and genders. Average ratings across raters will be correlated with epigenetic aging clocks, anthropometric measurements, and a morbidity index.

Rationale for assistance in data analytics and visualization: Collaboration with FADS students will be critical to the success of this project, which requires competencies in data management and preprocessing, learning and AI development, and software and technical development not currently represented on the research team. First, this project requires managing, cleaning, and harmonizing diverse data sources, including preprocessing images (e.g., cropping, alignment, illumination correction, facial landmark detection, and augmentation for model training) and preparing audio data (e.g., denoising, segmentation, feature extraction). Second, this project requires the application of advanced machine learning methods to photos and audio data, including convolutional neural networks, transformers, and speech recognition models. Data science expertise is needed to build, train, and validate deep learning models for predicting chronological age and health outcomes while mitigating algorithmic bias. Additionally, the ability to integrate multiple modalities through ensemble learning and

multimodal fusion techniques will be critical for developing robust measures of biological aging. Third, the project requires technical proficiency in the software and infrastructure that support large-scale, multimodal AI research. This includes programming in Python and R, leveraging machine learning libraries such as PyTorch, TensorFlow, Hugging Face, and scikit-learn, and using tools like librosa and OpenCV for audio and image processing.

Statement of benefit to the student: Graduate students collaborating on this project will gain unique cross-disciplinary training at the intersection of data science, AI, and aging research. They will acquire hands-on experience with advanced machine learning methods for computer vision and speech processing, while also learning to integrate multimodal data sources—including photographs, audio recordings, survey responses, health records, and DNA methylation biomarkers. Students will build practical skills in preprocessing and managing complex, sensitive datasets, and designing and validating human annotation protocols. Importantly, they will also engage with critical issues in ethical AI, including bias detection and mitigation, privacy protection, and fairness in predictive modeling. Beyond technical expertise, graduate students will benefit from close collaboration with faculty across disciplines, gaining mentorship in framing interdisciplinary research questions, publishing in high-impact journals (if desired), and preparing competitive funding applications. This project provides a rare opportunity to train future scholars who are not only proficient in cutting-edge computational methods but also equipped to apply these skills responsibly to pressing scientific and societal challenges in aging and health.

Specific competencies required, including programming languages if applicable: deep learning, machine learning, signal processing, web development, database management, Python, R

Is there anything else you would like us to know about your project's time frame or work schedule?: If the students work out well, we would be happy to continue supporting them as hourly workers over the summer

Proposal Title: #1055: The Impact of Local-Global Identity on Consumer Behavior

Department: Kelley School of Business, Department of Marketing, Bloomington

Project requires Human Subjects? Yes

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: This research program, developed in collaboration with colleagues from the University of New Mexico, the University of Manitoba, and Indiana University, investigates how consumers' global-local identity shapes marketplace behaviors and decision-making. At its core, the program seeks to understand the psychological, social, and behavioral mechanisms through which cultural identity orientations—whether global or local—guide choices, preferences, and consumption strategies. The central premise is that consumers' identity orientation plays a pivotal role in shaping everyday decisions ranging from brand loyalty to willingness to try new products. Global identity, characterized by cosmopolitan values, openness to diversity, and an emphasis on interconnectedness with distant others, is hypothesized to increase exploration and experimentation in the marketplace. Individuals with strong global identities are therefore more likely to engage in variety-seeking, embrace novel offerings, and demonstrate greater receptivity to products emphasizing innovation or multicultural appeal. These behaviors are particularly relevant in today's dynamic marketplace, where new product introductions, global brands, and rapid shifts in consumer culture have become commonplace. In contrast, local identity highlights community ties, familiarity, modesty, and attachment to a specific cultural or geographic context. Consumers with strong local orientations are more likely to favor products aligned with local traditions, heritage, and practical value. One manifestation of this tendency is a heightened inclination toward private-label and value-oriented brands, which signal modesty, trust, and embeddedness within familiar environments. Understanding the psychological underpinnings of these behaviors allows marketers, retailers, and policymakers to design more effective positioning strategies that speak directly to the cultural frameworks guiding consumer choices. The research program contributes to both theory and practice in three key ways. First, it extends scholarly understanding of culture and identity by moving beyond global versus local consumption outcomes to identify the cognitive and emotional processes driving these patterns. Second, it demonstrates how identity interacts with product characteristics, positioning strategies, and marketplace cues, thereby revealing critical boundary conditions and contextual dependencies. Third, it offers actionable insights for managers and practitioners, including brand strategists, product developers, and retailers, who must navigate increasingly diverse and culturally complex consumer markets. Methodologically, the program adopts a multi-method approach to capture the richness and complexity of these dynamics. Across a sequence of studies, we integrate large-scale secondary datasets, field and laboratory experiments, and cross-cultural surveys to triangulate evidence from different sources.

Rationale for assistance in data analytics and visualization: The proposed research program requires advanced expertise in data analytics and visualization to fully realize its objectives. While the project team has strong theoretical and empirical knowledge in consumer behavior and cultural identity, the complex multi-method design demands dedicated skills in managing large, heterogeneous datasets, and applying sophisticated analytical techniques. Graduate student assistance in data science will be

essential for processing and integrating diverse data sources, including secondary datasets, experimental data, and survey responses. Moreover, this project hinges on uncovering subtle behavioral patterns and interactions, necessitating advanced statistical modeling techniques that go beyond standard methods. The student's proficiency in these areas will enhance the rigor and reproducibility of the analyses. Equally important is the ability to translate complex findings into clear, compelling visualizations that facilitate both scholarly interpretation and practical application by marketing practitioners. By leveraging expertise in tools such as SPSS, SAS, R, and interactive visualization platforms, the student will enable the project to generate actionable insights and communicate results effectively through dynamic dashboards and graphical presentations. This assistance will streamline data management and enrich the overall impact of the research, ultimately supporting novel theoretical contributions and managerial relevance.

Statement of benefit to the student: This internship offers the MSDS student a valuable opportunity to deepen their applied skills in widely used research tools such as SPSS, SAS, and Qualtrics, which are essential in both academic and industry research environments. The student will gain hands-on experience managing and analyzing complex survey and experimental datasets, applying advanced statistical techniques within SPSS and SAS to uncover meaningful consumer behavior patterns. Working with Qualtrics, the student will contribute to designing and conducting rigorous surveys and learn best practices in data collection and survey management. Through close collaboration with faculty, the student will enhance their ability to prepare clean, well-documented datasets, perform reproducible analyses, and create clear reports and visualizations tailored for scholarly and practical audiences. These skills will equip the student with a competitive edge in research-oriented roles and consulting positions in marketing, consumer insights, and social science research. Moreover, the student will gain valuable experience in interdisciplinary collaboration and research project management, including exposure to survey design, multi-method data integration, and interpretation of cross-cultural consumer identity data. This blend of technical expertise and substantive research training will significantly contribute to the student's professional and academic development.

Specific competencies required, including programming languages if applicable: The project requires students with foundational skills in data analysis and survey research tools. Proficiency in SPSS and/or SAS for statistical analysis is essential to handle consumer behavior data and perform standard and advanced statistical tests. Experience with Qualtrics is helpful for designing, managing, and analyzing survey data, which is central to our cross-cultural studies. Basic data wrangling skills to clean and prepare datasets within these platforms are also needed. While programming languages like Python or R are not required, familiarity with any data visualization capabilities within these tools or ability to generate clear reports will help in communicating findings effectively. These competencies will enable students with core knowledge in popular social science analytics software to contribute meaningfully to the project.

Is there anything else you would like us to know about your project's time frame or work schedule?: The application system did not allow me to select both the Spring 2026 and Summer 2026 terms. I have therefore selected Spring 2026. However, as noted earlier, I would like to be paired with one skilled student for each of the Spring 2026 and Summer 2026 semesters in order to maintain continuity and advance the project efficiently across both terms.

Proposal Title: #1056: Global Health Impact Project (GHI) automation and forecasting tool

Department: Hamilton Lugar School, , Bloomington

Project requires Human Subjects? No

Project requires Laboratory Animals? No

Project requires Biosafety Review? No

Please indicate which term you prefer to work with students on your outlined project?: Spring 2026

Summary of project: The Global Health Impact Project (GHI) proposes the development of an automation and forecasting tool to transform global health decision-making. Leveraging data from the WHO, UNICEF, and other global health sources, this tool will automate the calculation of disease impact scores and provide user-driven forecasting capabilities. The GHI automation/forecasting tool can help policy makers address a critical gap in global health decision-making by replacing manual, error-prone disease impact scoring with an automated, scalable solution. Currently, GHI calculations are done using Google Sheets, limiting usability and accuracy. This tool improves upon that by using data on disease burden, treatment efficacy, and access levels to estimate the potential health impact of expanding access to essential medicines. The tool empowers governments, nonprofits, and pharmaceutical companies to make faster, evidence-based decisions about resource allocation, program planning, and product development. Users will be able to run "what-if" scenarios—such as forecasting how impact scores shift if a company lowers a drug's price—supporting real-time, strategic planning. For example, a policymaker in Kenya could input local tuberculosis data and determine which treatment would yield the greatest impact within their national budget.

Rationale for assistance in data analytics and visualization: Ideally, the person who takes this position will have a strong analytic mind to be able to translate complicated models from Excel into a format that is available for programming into a data visualization. We could also use help with the front end visualization on the forecasting tool as well as enhancing aspects of our global-health-impact.org website

Statement of benefit to the student: A student working for our team will get experience in working with a small startup like research group. We currently have support from someone who works with Princeton, consultants and faculty at other institutions and the person taking on this role can help manage a team of other students if they have that capability.

Specific competencies required, including programming languages if applicable: The responsibilities of this position include programming software applications and building statistical models to automate data collection/analysis process for the Global Health Impact Index (global-health-impact.org), and may include programming websites of the Global Health Impact Project. Proficiency in front-end (Typescript or JavaScript, HTML/CSS, etc.) or Python is strongly recommended. The ideal person for this position would help collect and analyze data and build statistical models, as well as help lead a team complete work on our websites and projects. So, communication and teamwork skills are important for this role. To help with these tasks, proficiency in Pandas and Microsoft Excel are required and experience in R is desirable. Proficiency in Python, JavaScript, HTML/CSS, etc. is strongly recommended.

Is there anything else you would like us to know about your project's time frame or work schedule?:
Students working for our project commit to doing 10 hours a week over the course of at least a semester. We have a weekly meeting of 1 to 2 hours currently on Wednesday is at 9:30 AM. Student should provide video demos of their work and an update weekly and need to be responsive to email.