THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Spring 2026 Data Science Clinic

Clinic Overview

The Data Science Clinic is a project-based course where students work in teams as data scientists with real-world clients under the supervision of instructors. Students are tasked with producing deliverables such as data analysis, research, and software along with client presentations and reports. Through the clinic course, Affiliate members gain access to undergraduate or graduate student teams to work on data science projects and explore proof of concepts while identifying top student talent. Projects are tailored and scoped to address company objectives with all deliverables overseen by the Clinic Director.

These unique collaborations allow Affiliate members to supplement their internal data science teams with outside support and perspectives, enlarging their capacity to experiment with new ideas. They also give students a window into a data science career, learning how companies build and use these tools internally.

Clinic Structure

Data Science Clinic runs during Fall, Winter and Spring quarters. Clinic projects are generally scoped to run for two full quarters. Each student works between 10 to 15 hours a week. Each team has a weekly 1-hour meeting with their assigned mentor and must submit a weekly progress report. Mentors are drawn from research staff, postdoctoral fellows and the faculty, subject to availability, interest and needs of the project. The mentor provides intellectual guidance, direct feedback to students and serves as a sounding board for both challenges and direction. The mentors will also provide support and guidance on any gaps in data science knowledge by providing literature and resources. Regular meetings are scheduled as it suits the client needs and to provide feedback to students.

What does the ⚙ mean?

If you look at the project descriptions below you will see that many of them have a gear/cog icon. These projects require a deeper knowledge of computing and preference will be given to those students who have demonstrated that capability.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Accountability Counsel

AI for Human Rights

Background:

Many international finance organizations maintain "accountability mechanisms"—systems for individuals and communities who have observed violations (environmental, labor, human rights) committed by internationally financed companies to submit a formal complaint to the finance organization. This offers communities an avenue to advocate for their rights, but the accountability mechanisms can be complex and difficult to utilize. Accountability Counsel's mission is to help local communities craft and submit successful complaints against human rights and environmental violations.

Over the years, Accountability Counsel has collected data on thousands of complaints, including the text of the complaint and whether the complaint was deemed eligible and ultimately successful, and they're looking to use the data they've collected to develop tools to help others write eligible complaints and advocate successfully. This project aims to develop AI models to characterize the likelihood that a given complaint will pass eligibility criteria, given the text of the complaint and relevant context about the accountability mechanism.

Mentor:

David Jacobson is the Data & Engineering Manager for the Data Science Institute Core Facility team. David specializes in leading machine learning and data science teams to develop practical solutions to complex real-world problems. His work is focused on data science projects for social good, especially related to climate, agriculture, human rights, and global health.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

AICE: AI for Climate ⚙

MonsoonBench

Background:
AI for Climate (AICE) is an interdisciplinary initiative that leverages advances in AI and expanding data availability to tackle critical challenges in climate prediction, impact modeling, and adaptation strategies. By uniting expertise from climate science, computer science, economics, physics, public health, and other fields, AICE develops novel tools—such as physics-informed predictive models and trustworthy datasets—to better understand climate dynamics and inform effective responses.

This project will transform a research benchmark for predicting the onset of the Indian monsoon into an open-source Python package. Accurate forecasting of the monsoon's spatiotemporal onset is critical for agriculture in India, yet the performance of new AI-based weather models on this task remains largely untested. Building on a methodology developed at the University of Chicago, the package will use precipitation forecasts to estimate onset dates, validate them against station-based rainfall data, and compute key skill metrics such as onset error, miss rates, and false alarms. It will also generate a visual scorecard to enable straightforward comparisons between AI-driven and traditional numerical weather prediction models. Deliverables include a PyPI package that implements the full benchmark using tools such as Xarray and Dask, providing researchers and stakeholders with a reproducible framework for evaluating monsoon prediction skill.

Mentor:
Pedram Hassanzadeh is an Associate Professor in Geophysical Sciences and Computational and Applied Mathematics at the University of Chicago and Faculty Director of AI for Climate. He leads the Climate Extremes Theory and Data Group, integrating theory, simulations, observations, and machine learning to study the dynamics and future of extreme weather. He earned his Ph.D. in geophysical turbulence and M.A. in applied mathematics from UC Berkeley and was a Ziff Environmental Fellow at Harvard. His honors include an NSF CAREER Award and an ONR Young Investigator Award.

Adam Marchakitus is a researcher in the Climate Extremes Theory & Data Group at the University of Chicago, where he develops and evaluates data-driven models for weather and climate prediction. A DSI Clinic alumnus, he graduated from UChicago in 2025 with a B.S. in Environmental Science and Data Science.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Becker Friedman Institute for Economics (BFI), University of Chicago

Healthcare Employment Geography Data Studio

Background:

The Becker Friedman Institute for Economics (BFI) serves as a hub for cutting-edge analysis and research across the entire University of Chicago economics community, uniting researchers from the Booth School of Business, the Kenneth C. Griffin Department of Economics, the Harris School of Public Policy, and the Law School in an unparalleled effort to uncover new ways of thinking about economics. Inspired by Nobel Laureates Gary Becker and Milton Friedman, BFI works with the Chicago Economics community to turn evidence-based research into real-world impact by translating rigorous research into accessible and relevant formats and proactively disseminating it to key decision-makers around the world.

The project this quarter will build upon previous quarters by working with the team at BFI in order to build visualizations and tools for papers written by researchers in the Economics department. In this past this has mean working on tooling to understand workforce

Mentor:

Eric Hernandez is the Senior Digital Media Manager at the Becker Friedman Institute for Economics. He brings extensive experience in digital content strategy and marketing, having previously strategized and created content for Organizing for Action, a non-profit organization that advocated for former President Barack Obama's political agenda. He also worked for Argonne National Laboratory developing marketing campaigns for divisions within the laboratory, giving him valuable experience in communicating complex research to diverse audiences.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Cook County Justice Advisory Council

Labeling the Law: Illinois Criminal Statute Data Enhancement

Background:

The Justice Advisory Council (JAC) coordinates and implements Cook County Board President Toni Preckwinkle's criminal and juvenile justice reform efforts and community safety policy development. The work of the Justice Advisory Council is guided by the county's Policy Roadmap, which identifies a central priority of building safe and thriving communities throughout Cook County. In collaboration with governmental and non-governmental stakeholders, the Justice Advisory Council devises, supports, and advocates for administrative reform within Cook County, as well as legislation which improves conditions and outcomes for individuals involved in the justice system. The JAC manages a portfolio of grants, primarily awarded to community-based organizations who work in geographic areas that have experienced historic disinvestment.

This project enhances the Administrative Office of the Illinois Courts (AOIC) "Offense Code Table" (OCT) by linking offense codes to the full statutory text from the Illinois Compiled Statutes (ILCS). The current OCT relies on truncated descriptions that require manual cross-referencing, limiting readability and analysis. By integrating OCT codes with complete ILCS descriptions, the project applies data science methods—such as text processing, clustering, and structured data integration—to produce clearer descriptions and analysis-ready groupings of offenses. This enables more robust examination of criminal legal trends, policy impacts, and charging practices across Illinois.

The enhanced dataset will reduce redundant lookups, establish standardized conventions for categorizing offenses, and support more precise studies of disparities, reforms, and legislative impacts. Designed for both operational and research use, it will provide court officials, attorneys, and policymakers with a machine-readable, scalable foundation for dashboards, models, and decision-support tools. At its core, the project creates a common analytical language for Illinois statutes, improving both clarity and reproducibility in criminal legal system research.

Mentor:

Whitney Key Towey, PhD, MPH, MSW – Director of Data and Research, JAC
Nico Marchio – Associate Director, Office of the Cook County Public Defender

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Data, Policy, and Innovation Centre

AI-Enabled Grievance Redressal

Background:

The Data, Policy & Innovation Centre (DPIC) is a first-of-its-kind partnership between the University of Chicago and the Odisha government in India. Odisha is a coastal state in eastern India with a population of over 45 million. DPIC, funded by the state government and located in the state capital, Bhubaneswar, leverages large administrative datasets and cutting-edge data science and research to address complex development challenges and drive evidence-based governance. DPIC's work spans the full data-to-policy cycle - from curating high-quality datasets, extracting analytical insights, conducting rigorous research and building data capabilities within government.

This project applies data science and natural language processing (NLP) techniques to modernize *Janasunani*, the Government of Odisha's grievance redressal platform, by improving efficiency in handling the 1.8 million citizen complaints filed between April 2021 and June 2025. The clinic will develop workflows to digitize unstructured English-language documents, extract entities such as schemes, departments, and locations, automatically summarize long-form grievances, and classify complaints into thematic categories. The outputs will include a structured dataset of summarized and categorized complaints and reproducible code pipelines. Beyond reducing manual entry and administrative burden, the work will generate actionable insights at scale, turning citizen feedback into a tool for governance improvement and demonstrating how data science can strengthen trust between governments and citizens.

Mentor:

Dr. Urmila Chatterjee is the Executive Director at DPIC and former Research Director at the Energy Policy Institute at the University of Chicago in India (EPIC India). Previously, she served over a decade as Senior Economist at the World Bank, leading policy dialogue, research, and lending projects across South Asia on topics including human development, firm growth, fiscal policy, energy reform, and climate adaptation. She has managed large teams, mentored junior colleagues, and published widely in policy journals, reports, and newspapers. Earlier in her career, Urmila worked at the Indian Institute of Management and Citigroup. She holds a Ph.D. in Economics from UC Berkeley, an M.A. in Economics from the University of Mumbai, and is a Chartered Financial Analyst.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Ecdysis ⚙

Deep Learning to Characterize Species Diversity

Background:
Ecdysis Foundation's mission is to support the evolution of a regenerative food system using science, education, and demonstration. The Ecdysis 1000 Farms Initiative partners with farmers to characterize the chemical and physical properties of soil, water dynamics, and species diversity on their land. To further expand this program, Ecdysis is utilizing AI and deep learning algorithms to characterize species diversity using audio and image data.

Currently Ecdysis collects data on avian species diversity on a small number of farms via visual observation—a.k.a. bird-watching. This process is too resource intensive to scale to 1000+ farms, so they want to develop a deep learning pipeline for identifying bird species and characterizing overall abundance and diversity based on hour-long audio recordings taken on the farms. The audio recordings contain a mix of human and natural sounds, including bird calls. This project aims to build a pipeline that will separate the relevant from irrelevant sounds, match bird calls to species, and characterize the likely abundance of each species.

Mentor:
Trevor Spreadbury is a Software Engineer II at the DSI. He helps social impact organizations to enhance their operations, research, and communication by utilizing software engineering and data science tools. His work focuses on agriculture, human rights, energy, and marine technology.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

The Impact Project

NLP for Federal Government Impact Mapping

Background:

The Impact Project collects and synthesizes data on how government change affects communities, overlaying information on local economies, industries, services, and demographics. At a time of rapid shifts in federal policy and funding, there is a critical need to track and explain how these changes ripple through states, districts, and neighborhoods. Reliable, timely information enables policy debates grounded in evidence and helps local leaders and advocates respond effectively.

This quarter, students will develop models to extract structured information about federal government funding, workforce, and policy changes from news articles, public reports, and other unstructured sources. The resulting data will expand the Impact Project's "Impact Map," allowing users to trace government decisions to their local consequences. Students may also investigate the trends in this impact database — identifying sectors most affected, geographic disparities, or emerging themes in government activity.

Mentor:

From The Impact Project:

- Abby André, Director @ The Impact Project
- Jonathan Gilmour, Data Lead @ The Impact Project

From UChicago DSI:

- Dylan Halpern builds tools with social impact partner organizations focused on low-code and no-code geospatial data science and visualization. His work includes open source software engineering, product management and design, and stakeholder engagement.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Karczmar Lab, Department of Radiology, University of Chicago ⚙

Predicting Breast Cancer Treatment Response Using Blood Vessel Networks

Background:
The Karczmar Lab in the University of Chicago Department of Radiology is a leading research group specializing in advanced MRI techniques for cancer imaging and treatment prediction. Led by Dr. Gregory Karczmar, Director of MRI Research, the lab has pioneered innovative MRI methods that are now used in clinical breast cancer screening. The lab combines expertise in MRI physics, medical imaging analysis, and computational methods to develop novel approaches that directly improve cancer detection and patient care.

This project aims to develop a novel approach to predict how breast cancer patients will respond to treatment by analyzing the patterns of blood vessels around tumors. Blood vessels form natural networks that change in response to treatment, and we hypothesize that these vascular patterns can serve as early markers of treatment effectiveness. The ultimate goal is to enable personalized treatment decisions and potentially develop risk assessment tools for earlier cancer detection.

The core innovation involves representing each patient's blood vessel structure as a mathematical graph, where vessel branch points become nodes and vessel segments become edges, annotated with properties like length, width, and curvature. Students will implement Graph Neural Networks (GNNs) to analyze these vascular networks, as GNNs can capture both local structural changes and global network patterns critical for understanding patient-specific responses. The project will involve building a complete pipeline from MRI data to graph representations, training and evaluating GNN models against traditional approaches, and identifying which vascular features are most predictive of treatment outcomes.

Mentor:
Dr. Gregory Karczmar is a Professor of Radiology and Medical Physics and Director of MRI Research, with more than 30 years of experience developing new MRI techniques now used in breast cancer screening. Dr. Milica Medved is a Research Associate Professor of Radiology with over 20 years of experience developing advanced MRI methods for cancer detection and risk prediction. Dr. Zhen Ren is a Research Assistant Professor of Radiology specializing in advanced imaging techniques, including dynamic contrast-enhanced MRI and image reconstruction for breast cancer research.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Metropolitan Water Reclamation District of Greater Chicago ⚙

Oak Park Roof Analysis for Stormwater Management

Background:
The MWRD manages wastewater and stormwater in Cook County, Illinois. Established in 1889, it operates one of the largest wastewater treatment systems in the world, serving over 5 million residents. The MWRD's mission is to protect public health and the environment by treating and reclaiming water, managing stormwater, and reducing pollution. It also plays a key role in flood control and water quality improvement.

This project uses high-resolution satellite imagery, machine learning, and GIS to identify and analyze the largest rooftops in Oak Park. By generating a distribution of roof sizes, the project will help answer practical questions such as how many roofs exceed a certain area and highlight properties suitable for downspout disconnection. The analysis will be integrated with property records and environmental data, creating a comprehensive view of stormwater management opportunities. Students involved will gain hands-on experience in data analysis, imaging technology, and environmental science. The outcomes include a software tool, detailed maps, and a final report with recommendations that can be scaled across the Metropolitan Water Reclamation District, advancing sustainable urban development and climate resilience.

Mentor:
Richard Fisher is a Principal Civil Engineer in the Stormwater Division of the MWRD's Engineering Department. He oversees stormwater master planning, pilot studies, and various stormwater programs. With over 30 years of experience, he has planned, designed, and managed numerous public and private capital improvement projects.

Mothers Out Front

School Lookup Tool for Environmental Action

Background:

Mothers Out Front is a grassroots movement of parents and caregivers organizing for healthy, climate-safe schools and communities. Local teams often confront fragmented, technical information spread across federal and state databases, district websites, and mapping tools—slowing advocacy and putting at a disadvantage schools with fewer resources. Volunteers need school-specific answers to practical questions: Are there grants or rebates for electrification or solar? Has the district adopted a climate or sustainability plan? Is there known lead risk in drinking water? What is the rooftop solar potential? By assembling these signals in one place and pairing them with clear next steps, the tool will help community members move quickly from data to action—supporting equitable, evidence-based campaigns at the school and district levels.

Project Objective:

Students will build an interactive lookup tool (search by school name, district or location) that assembles an actionable environmental profile for each school/district, including information such as:

- Documents describing climate-related policies and incentives, which could be parsed and categorized using LLMs.
- Drinking water lead/copper results from EPA ECHO.
- Potential for environment-friendly facilities improvements, including rooftop solar potential indicators and possible HVAC and electrification upgrades.
- Electric school bus adoption and opportunities.
- Public support in your district for climate-friendly investment.

The tool should give parents and advocates the information they need to prioritize and activate for meaningful change in their school and district.

Mentor:

- Camille Greer - Co-Executive Director @ Mothers Out Front
- Jenny Zimmer - Co-Executive Director @ Mothers Out Front

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Occidental Arts & Ecology Center ⚙

Gully Identification Algorithm for Erosion Prevention

Background:

California stands at the crossroads of two accelerating disasters: increasing wildfires and persistent, deepening drought. The Fuels 2 Flows campaign, reimagines these twin challenges as an opportunity to address both when re-connected. Instead of treating forest "slash" as a waste to burn or discard, Fuels 2 Flows harnesses this material as living infrastructure, converting it into vital, water-retentive biomass that heals eroded upland gullies.

Jim Pivarski (UChicago DSI) and Brock Dolman (OAEC) have collaborated on a Found Gully Mapping Tool to help landowners and project leaders identify high-impact opportunities to use the Fuels 2 Flows method to reduce both erosion and fire risk. For this mapping tool to be effective, we need to precisely identify the locations and paths of gullies—including smaller gullies that contribute heavily to erosion. In this project, students will develop a gully identification algorithm that uses LiDAR data in conjunction with state-of-the-art AI image analysis to identify the locations and paths of gullies in Sonoma County. The algorithm will then be used to target opportunities for erosion prevention.

Mentor:

- Jim Pivarski is a data scientist/engineer who has worked in and out of academia. He was trained as a particle physicist with a Ph.D. from Cornell and helped to commission and analyze first data from the CMS experiment at the Large Hadron Collider (LHC) in Geneva. He then worked as a data science consultant at Open Data Group, analyzing data from commercial clients, and then at Princeton, developing analysis software for physicists and promoting better integration with the world beyond academia. Jim is the original author of Awkward Array, a NumFOCUS affiliated project with thousands of users. Now he is analyzing data at U. Chicago's Data Science Institute for the 11th Hour Project and its philanthropic goals.
- Brock Dolman (he/him) co-directs the WATER Institute, Permaculture Design Program, and Wildlands Program. He has taught permaculture and consulted on regenerative project design and implementation internationally in Costa Rica, Ecuador, the US Virgin Islands, Spain, Brazil, China, Canada, Zimbabwe, Tanzania, the Democratic Republic of Congo, Cuba, and widely in the United States. He has been the keynote presenter at numerous conferences and was featured in the award-winning films The 11th Hour by Leonardo DiCaprio, The Call of Life by Species Alliance, and Permaculture: A Quiet Revolution by Vanessa Shultz.

Promega Corporation ⚙

Biological Organoid Development

Background:

Promega is a global biotechnology company, headquartered in Madison, Wisconsin, that provides products and solutions for life science research, drug discovery, and human identification. The company manufactures reagents, enzymes, and other biochemicals used in fields like genomics, protein analysis, cellular analysis, and forensics, serving academic, government, and industrial researchers worldwide. Promega is also committed to sustainability, employee development, and community engagement, striving to create innovative, science-driven solutions for real-world challenges.

This project leverages data science and AI methods to predict the suitability of organoids for scientific use by analyzing early-stage characteristics. Using a dataset of organoid images, chemical composition information, and survey data collected by Promega, the study aims to build supervised learning models that classify organoids as “good” or “bad” after 30 days of growth. The approach will begin with tabular data to establish baseline predictors and then extend to image analysis with computer vision techniques, enabling the identification of key morphological and chemical signals associated with successful development. The resulting models will improve efficiency in organoid research by reducing time, cost, and experimental waste.

Mentor:

Liya is a Data Scientist at the University of Chicago.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

University of Chicago Library

Optimizing Library Storage

Background:
In support of free inquiry and expression, the University of Chicago Library is transforming the global knowledge environment to be open, accessible, and equitable. We enable the University of Chicago and our greater community to create a better world through effective information services, a comprehensive connected collection, and a culture of innovation, respect, and partnership.

This project will develop predictive models to help the Library decide which single-volume print monographs should remain in on-site storage. Using ten years of anonymized circulation data combined with bibliographic metadata such as subject classifications, publication dates, and call numbers, students will analyze patterns of past usage and estimate likely future demand. The core deliverable is a tool that produces optimized on-site storage lists for different collection sizes (e.g., 30,000 to 3 million volumes) and allows adjustments by class or subclass to reflect curatorial priorities. If time allows, the team will also create a report that explores distinctive patterns of usage and collection strength within selected Library of Congress subclasses, highlighting areas of unusual demand or uniqueness.

Mentor:
David Bottorff is the Collection Management & Circulation Services Librarian for the University of Chicago Library and is leading the Library's strategic priority of developing a comprehensive collection management and storage plan for its physical collections.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Syllabus Project

Understanding Common Syllabus Language using LLMs

Background:

College syllabi serve as both a contract between instructor and student and a vehicle for communicating institutional policy. Many elements of a syllabus (Title IX notices, mandatory reporting disclosures, disability services statements, academic integrity policies) originate not from the instructor but from the institution. Despite their prevalence, there is little systematic understanding of how common these elements are, how they vary across institutions, or what other structural components (course calendars, late work policies, grading breakdowns) consistently appear. The purpose of this project is to use large language models to extract, classify, and analyze the common elements found in college syllabi, with the goal of helping faculty understand the prevalence and variation of standard syllabus components.

This project consists of two stages. In the first stage, students will work with a corpus of publicly available syllabi to develop a taxonomy of common syllabus elements — both institutional boilerplate and structural components (such as course calendars, attendance policies, and grading criteria). Rather than beginning with a fixed list, students will use LLMs via OpenRouter to iteratively identify and refine the categories that emerge from the data. In the second stage, students will apply their taxonomy at scale to characterize how frequently each element appears, how language differs across schools and institution types, and what patterns emerge in syllabus construction.

The goals of this project are to: develop a working taxonomy of syllabus elements, build effective LLM prompting strategies for information extraction from unstructured documents, and characterize the landscape of syllabus language across institutions. This is primarily a natural language processing and analysis project.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

College Financial Health

Factors causing college closure

Background:

Research into the predictors of college failure has grown as demographic forces have reshaped the economics of higher education in the United States. The purpose of this project is to present, in a visually appealing style, information about the economic stresses faced by colleges and universities in the United States.

This project consists of two components: The first is a paper replication for linking institutional financial data (IPEDS) with information about college closure. The paper to be replicated is Predicting College Closures and Financial Distress written by Robert Kelchen, Dubravka Ritter, and Douglas Webber. Once this replication is completed, the second stage of this study will be to implement a visually appealing representation of the results so that students making decisions about where to attend college have easy access to information about their college decision.

The goals of this study are to: identify characteristics of colleges and universities which make them more or less likely to fail and how do we present this information? This is primarily an analysis and visualization project.

Mentor:

Nick Ross, director of the data science clinic will serve as the primary mentor on this project.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Zero Foodprint

Parcel-Level Carbon Farming Candidate Identification in California

Background:
Zero Foodprint, a nonprofit mobilizing the food and beverage industry to fund regenerative agriculture, is partnering with a DSI team to identify California agricultural parcels with high potential for carbon farming interventions. The organization pools small contributions from restaurants and food businesses to provide grants directly to farmers implementing practices like composting, cover cropping, and reduced tillage. To scale this impact, Zero Foodprint needs a systematic approach to identify and prioritize farm parcels that could benefit from regenerative practice adoption.

Students will design data integration pipelines to combine California parcel data with land-use raster datasets, creating a comprehensive geospatial database of potential candidate farms. The team will develop reusable scripts that Zero Foodprint can deploy to systematically identify opportunities for farmer outreach and grant distribution, along with an interactive visualization platform (such as a Streamlit map) for exploring and filtering candidate parcels based on land-use characteristics and geographic location. If time permits, students can extend the analysis by recommending specific regenerative practices tailored to different land-use classifications and quantifying potential carbon sequestration impacts using USDA COMET data.

This project offers students hands-on experience at the intersection of climate action, agriculture, and geospatial data science. Students will gain practical skills in GIS data processing, large-scale data integration, and interactive visualization development, while contributing directly to Zero Foodprint's mission to restore the climate "one acre at a time" through regenerative agriculture funding.

Mentor:
Tim Hannifan is a Research Software Engineer at the University of Chicago Data Science Institute, where he develops software for interdisciplinary research projects spanning AI systems, data pipelines, and web applications. Previously serving as Assistant Data Science Clinic Director, he led software development for research partnerships with Morningstar, the City of Chicago, and Argonne National Laboratory, building solutions ranging from agentic AI systems to machine learning platforms. Tim specializes in translating research ideas into robust, scalable systems using Python, cloud infrastructure, and modern AI/ML frameworks. He holds an MS in Computational Analysis and Public Policy from the University of Chicago.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Palmwatch

LLMs for Extracting Data from Palm Oil Disclosures

Background:
Palm oil is widely used in consumer goods, animal feed, and biofuels, but its production has led to serious environmental and social harms, including land seizures, deforestation, carbon emissions, biodiversity loss, and water pollution. To support communities advocating for land preservation, the Data Science Institute has partnered with **Inclusive Development International** to build **PalmWatch**—a tool for discovering the companies that fund, operate, or source from mills. In doing so, the platform allows communities to target companies linked to harmful development projects in their advocacy campaigns.

PalmWatch currently relies on a manual, rule-based process to extract supply chain data from consumer brands' mill disclosure documents, which limits its scalability. This quarter, students will use Large Language Models (LLMs) to automate extraction from disclosures, significantly expanding the platform from 15 to 50+ brands. Potential tasks include: building a data pipeline to collect, extract, clean, and standardize mill data; comparing the results of extraction using LLMs and third-party APIs; performing entity resolution against authoritative datasets; and producing a comprehensive dataset of brand–mill associations for analysis and reporting.

Mentor: David Jacobson is the Data & Engineering Manager for the Data Science Institute Core Facility team. David specializes in leading machine learning and data science teams to develop practical solutions to complex real-world problems. His work is focused on data science projects for social good, especially related to climate, agriculture, human rights, and global health.