THE UNIVERSITY OF CHICAGO
DATA SCIENCE
INSTITUTE

Winter 2025 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.

Project List

Argonne ⚙

Extending Argo with LLM-driven Agents and Workflows

Background:

Argonne is a multidisciplinary science and engineering research organization where talented scientists and engineers work together to answer the biggest questions facing humanity, from how to obtain affordable clean energy to protecting ourselves and our environment. The laboratory works in concert with universities, industry, and other national laboratories on questions and experiments too large for any one institution to approach alone.

Surrounded by the highest concentration of top-tier research organizations in the world, Argonne leverages its Chicago-area location to lead discovery and power innovation in a wide range of core scientific capabilities, from high-energy physics and materials science to biology and advanced computer science.

Project Description:

This year, Argonne National Laboratory launched a secure, internal generative AI interface for the Argonne community called Argo. The initial implementation was a text-based conversational chatbot with access to OpenAI's GPT models hosted on private Azure instances. A key feature of our approach is not to store user queries and large language model (LLM) responses to enable those within the community who work with highly sensitive data the opportunity to leverage generative AI technology in a secure environment. Since its debut, Argo now features a document upload tool, an API for science developers, and multiple GPT-based LLMs and embedding models. The next phase for Argo is to integrate a retrieval-augmented generation (RAG)-based information retrieval mechanism supported by a knowledge graph to provide Argo users with direct access to Argonne-specific domain knowledge, such as operational policies, procedures, science records and data, and user facility and equipment documentation. This feature is currently under development at Argonne, which has been significantly supported by previous DSI student efforts. Much of their prior work, investigation, analysis, and prototype code are incorporated into our strategy and implementation design.

As we continue to enhance Argo with RAG techniques, we are excited to partner once again with DSI students. This collaboration will help us understand the latest techniques and lay the groundwork for our future design and development strategies. To support students' work, we plan to provide a stand-alone template of Argo that can be run locally within the UChicago network. While we are unable to provide access to our internal LLM endpoints, the specific LLMs utilized are of minor importance for the development process of agents and workflows. This approach will also allow a straightforward transfer of student-developed prototypes into the production version of Argo deployed at Argonne.

The prospects and performance-improving opportunities suggested by LLM-based agents have grown in recent months, and we are excited to consider how we can incorporate this latest engineering technique into Argo for the specific use cases of interest to the operations and science users at Argonne National Laboratory. Because LLM agent and workflow frameworks are relatively new, we have yet to develop a predefined strategy or prescription that must be followed. We are especially interested in students exploring, learning, and offering recommendations through a prototype local implementation of Argo with agents that we can incorporate into future feature enhancements.

The University of Chicago DSI students will be challenged in learning, planning, and prototyping the latest LLM-based engineering techniques into an existing generative AI interface that leverages vector and graph databases for supporting information retrieval. There may be multiple approaches to developing LLM-driven agents and process workflows, which must be explored and evaluated to recommend the most appropriate techniques for the Argo implementation and future expectations.

Mentor:

Matthew Dearing is a software engineer and Technical Lead for the AI for Operations initiatives at Argonne, with a Joint Appointment at UChicago. Matthew is also a Ph.D. student at the University of Illinois Chicago investigating advanced HPC management algorithms and digital twin modeling and an Adjunct Instructor in Computer Science at the University of Illinois Springfield.

Building Decarbonization Coalition ⚙

LLM/Chatbot for Thermal Energy Networks

Background:

The Building Decarbonization Coalition (BDC) aligns critical stakeholders on a path to transform the nation's buildings through clean energy, using policy, research, market development and public engagement. The BDC and its members are charting the course to eliminate fossil fuels in buildings to improve people's health, cut climate and air pollution, prioritize high-road jobs, and ensure that our communities are more resilient to the impacts of climate change.

Mentor Information:

Matt Rigg, Lead Software Engineer and Srishti Bal, Senior Data Associate

Matt has been creating delightful and thoughtful consumer experiences for over 10 years. He previously worked at Amazon, Noom, and TIME CO2, launching engaging products that now serve millions of users and help corporations accelerate their journey to net-zero and nature positive impact.

Srishti graduated from Cornell Tech with a Master's in Operations Research & Information Engineering. Her graduate work focused on improving urban infrastructure with an emphasis on the impact of the built environment on climate, particularly heat vulnerability. She previously worked at Tesla in the Autopilot department for 3 years, gaining experience in project management and data analysis.

Project Information:

The goal of this project is to build an LLM/Chatbot that can answer questions about Thermal Energy Network Pilots in New York.

The data consists of unstructured PDF documents and it can also include Microsoft Excel and Word documents. Common types of documents are utility filings, public comments, motions, petitions, plans and proposals, press releases, reports, rulings, and other publicly available data about Thermal Energy Network pilots in New York. Most documents are smaller than 10MB. The main repository we are aware of has a dataset containing over 1,000 such documents comprising 11 proposed TEN pilots. We might also explore other data sources which could include other unstructured as well as structured data.

Examples of this data can be found at New York's Public Service Information site.

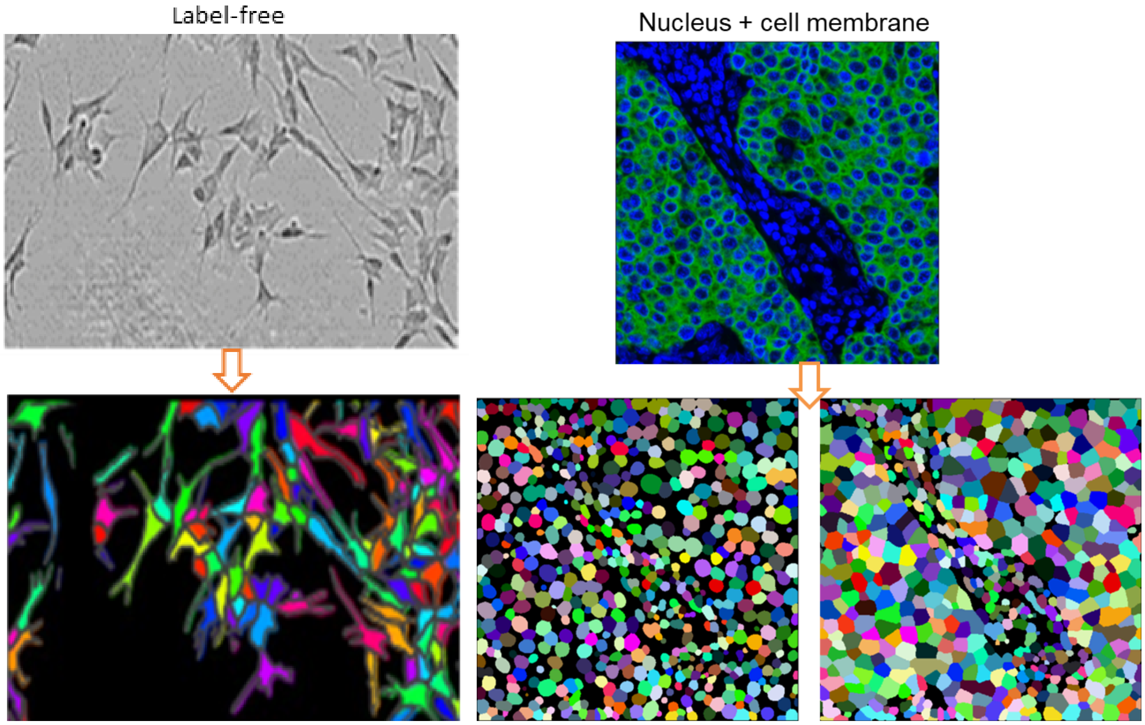
Center for Living Systems ⚙

AI and Cell Segmentation

Background:

Cells are the fundamental building blocks of all living organisms. Over the past few decades, advancements in microscopy techniques have revolutionized our ability to study cellular structures in detail. One critical step in this process is cell segmentation is essential for analyzing cell features, dynamics, migration, and morphology. Computer vision techniques, particularly AI-driven approaches fueled by big data, have driven significant progress in the development of cell segmentation tools. Two of the most advanced and widely used software tools for this purpose are CellPose and StarDist, both of which leverage deep learning techniques to enhance segmentation accuracy and efficiency. They are capable of cell segmentation for nucleus image, cell membrane image and label-free images.

Figure 1. Examples of cell segmentation based on nucleus, cell membrane and label-free images.



In this project, we would like to explore the applicability of these two software and study the performance over various cell types and image modalities. The purpose of this project is to get these two algorithms running on our internal data science cluster and then compare the results using cell segmentation evaluation metrics against ground truth data. Understanding the relative performance of these algorithms is crucial for determining which tool is best suited for specific conditions.

A stretch goal of the project is to train data-specific models for data from UChicago's Gardel Lab. 2D or 3D models can be trained by fine-tuning pre-trained models with these datasets. By comparing the performance of pre-trained models with data-specific models, we aim to gather valuable insights on addressing challenging cell segmentation problems in real biological research scenarios.

Mentor

Liya Ding is a Data Scientist at the Data Science Institute (DSI). She also contributes to the Center for Living Systems, under the leadership of Prof. Gardel. She earned her Ph.D. from The Ohio State University in 2009. Her experience includes postdoctoral roles at various institutions, a position as a scientist at the Allen Institute for Cell Science, and an associate professorship at Southeast University. Her research interests include computer vision, data science, and computational cell biology, with a particular focus on microscopy image processing and quantitative data analysis.

Chicago Metropolitan Agency for Planning (CMAP) ⚙

NE Illinois Stormwater Storage and Site-Scale Green Infrastructure Inventory

Background:
CMAP, the regional planning organization for northeastern Illinois, engages with local governments and stakeholders in seven counties to improve water resource management. However, a comprehensive inventory of stormwater storage and green infrastructure assets is needed. Such an inventory is crucial for maintenance, enhancing water quality, and strengthening stormwater management against climate change impacts. This knowledge gap offers an opportunity for stakeholders to use data in watershed and resilience planning more effectively.

Last year students built a baseline model using satellite images. Now that the baseline model is complete there are additional data sources and tuning that can be done to significantly increase the accuracy of these models!

Mentor:
Holly Hudson is a Senior Aquatic Biologist at CMAP and has more than 30 years' experience in lake and watershed monitoring, planning, and management. In addition to conducting lake and watershed studies, she provides technical assistance to the public and local governments and organizations on lake and watershed monitoring, management, and grant application development, and has overseen numerous Clean Lakes Program and Nonpoint Source Pollution Control Program implementation projects.

Food System 6

Visualizing the Economic Infrastructure of the US Poultry Industry

Background:

Food System 6 is a non-profit that envisions a future food system that scales sustainable solutions, fosters connectivity, restores biological and cultural diversity, and positively impacts health outcomes for all while nurturing both soils and spirits. The current food system, marked by the consolidation of wealth and power, has led to negative outcomes for communities, the planet, and people, including rising food insecurity, ecological degradation, and the erosion of human health and community wealth. This system has stripped farmers, workers, and consumers of their rights to innovation, ownership, and autonomy. In contrast, Food System 6 advocates for community-based innovation as a crucial element in diversifying food system solutions. However, the innovators in frontline communities often lack the necessary resources and support to scale business models that democratize wealth and restore health, ecology, and justice to the food system.

Project Description:

A major impediment to regenerative farming in the United States is that public and private financial supports for conventional farming are both large and often hidden. These are not only direct subsidies, but also hidden costs, such as financial programs whose requirements preclude regenerative farming efforts.

Over the last year Food System 6 ("FS6") has worked to understand the scope and magnitude of the financial ecosystem that underpins the US poultry industry and are looking to leverage this knowledge into building a visualization and set of dashboards to highlight the financial asymmetries between conventional poultry production and emerging poultry production systems that embrace regenerative principles, e.g., pastured poultry. These visualizations will be used to provide greater insight and education to the growing segment of stakeholders, funders, investors, and actors in the food system reform space.

We want this visualization to capture the economic mechanisms and cultural capital norms spanning from "land to plate" highlighting the benefits received by conventional poultry growers at each step of the process. We have identified about 20 different financial mechanisms, offerings, and programs. For each we need to understand how it financially impacts farmers from both a conventional and regenerative context. While many poultry producers may be able to access federal insurance products, for example, regenerative practices may automatically preclude a poultry grower from accessing those programs.

Our research has also identified various “capital cultural norms” that drive the practices and performance of conventional poultry. For example, poultry integrators use lopsided contracts to pit poultry growers against one another, lead them into obtaining multi-million-dollar financing to construct poultry houses, and allow them to cancel the contract for a wide variety of reasons with no consideration or renumeration to the poultry grower.

Mentor Information:

David LeZaks, Ph.D. is the Co-Managing Director of Food System 6. He is an environmental scientist and financial activist whose work is centered around developing innovative mechanisms for financing the transition to regenerative farming and food systems. David completed his Ph.D. in Environment and Resources and an M.S. in Land Resources at the University of Wisconsin – Madison. He is based in Madison, Wisconsin, where he is active in a number of community organizations and spends his spare time gardening and participating in a variety of silent sports.

Lauren Manning, Esq., LL.M., is an attorney, law professor, and farmer. Before her current role as Co-Managing Director of FS6, Lauren was a venture capital investor with food and ag-focused VC firm AgFunder. Lauren began with AgFunder in 2015 as part of AgFunderNews media and research team reporting on issues involving finance, agriculture, climate change, and more. From 2019 to 2021, Lauren supported the Sacred Cow documentary and book project discussing the nutritional, environmental, and ethical case for (better) meat production. At the University of Arkansas, Lauren serves as an adjunct law professor across multiple departments teaching courses on farm animal welfare, food safety, farm succession planning, agricultural cooperatives and local food systems, and more. Lauren raised grass-finished beef, lamb, and goat meat in NW Arkansas for eight years. In 2023, the Regenerative Food Systems Investment (RFSI) Forum recognized her as one of 15 Women Leading Investment in Regenerative Food Systems.

Fermi National Accelerator Laboratory (GNN) ⚙

Graph Neural Networks for Liquid Argon Time Projection Chambers

Background:
Fermilab is America's particle physics and accelerator laboratory. Host of several particle physics experiments and international collaborations aiming at solving the mysteries of matter, energy, space and time.

Neutrinos are the lightest matter particles in the Universe. They are electrically neutral and interact with other particles only via the weak nuclear force. This makes the study of neutrinos very challenging, as they interact very rarely, thus requiring large detectors and intense beams. This also means that some properties of neutrinos are not fully known yet. Indeed, neutrinos may play a key role in the dominance of matter over antimatter in the Universe and can potentially give insight into the nature of Dark Matter. Current and future experiments at Fermilab are tasked to measure neutrino's properties with unprecedented precision using Liquid Argon Time Projection Chamber (LArTPC) detectors. These are large volumes of liquid argon equipped with three readout planes, each made of wires that measure the ionization charge produced by charged particles traveling through the argon.

Fermilab partners with the University of Cincinnati and Northwestern University with the goal of developing a Graph Neural Network (GNN) for particle reconstruction in LArTPC neutrino experiments. Goal of our GNN is to provide precise inputs to the study of neutrinos through the measurement of particles produced by the neutrino interaction in the detector. For instance, the identification of the neutrino type relies on this measurement. We have developed a message-passing GNN that is used to classify the nodes, defined as the charge measurements on the LArTPC wires, according to the type of the particle that produced them. Our network connects nodes both within and across measurement planes in the detector and achieves 94% accuracy with 97% consistency across planes.

We recently expanded our GNN to include multi-modal information. Previously only information from the LArTPC wires were used as input to the GNN. However, the experiment also includes additional detectors, such as an array of photomultiplier tubes (PMTs) and a cosmic ray tagger (CRT). PMTs, in particular, detect the scintillation light from the de-excitation of argon atoms that were excited by the passage of charged particles produced in the neutrino interaction. PMTs provide complementary measurements of position, energy, and time with respect to the wires. Including this information in the GNN as heterogeneous graph nodes provides additional information for improved accuracy and consistency in the measurement. A first version of the network including PMT information was developed over the summer. In this project we will work to optimize this first version by tuning the choice of hyperparameters and by exploring different network configurations in terms of information flow between the LArTPC wires and the PMT detectors.

Mentor:

Giuseppe Cerati received his Ph.D. in Physics and Astronomy at Università degli Studi di Milano – Bicocca in 2008. At Fermilab since 2016, currently working as Scientist. Working on collider experiments such as CMS and neutrino experiments such as MicroBooNE, ICARUS, DUNE, with focus on physics analysis and data processing algorithms (both traditional and machine learning).

Fermi National Accelerator Laboratory (Simulations) ⚙

Improving ML-based simulations of particle physics experiments

Background:
Simulation plays a crucial role in analyzing the data from particle physics experiments. High quality physics-based simulations have a significant computational cost, which will become impractical as dataset sizes grow. An alternate approach is generative machine learning, in which we train a model that matches the quality of the physics-based simulator but produces samples much more quickly. We have an initial version of such a model, based on diffusion, a state-of-the-art approach in image generation. This model approximates the physics-based simulation with very high accuracy, but the generation speed is not yet fast enough for many use cases. This project will focus on applying a method to speed up generation, by first compressing the inputs to a lower-dimensional representation before running the diffusion process. Students will have the opportunity to learn about and employ multiple cutting-edge techniques in generative machine learning. There is opportunity for exploring other approaches to speed up or improve the model, depending on student interest.

Mentor:
Oz Amram and Kevin Pedro are physicists at Fermilab working as part of the CMS Collaboration which is attempting to understand the very basic laws of our universe. Oz received his PhD from Johns Hopkins while Kevin received his from UMD. Their research focuses on the use of AI and ML techniques in high energy physics. Oz is also a writer for ParticleBites, a reader's digest of high energy physics news and papers designed for readers with an undergraduate level of physics knowledge.

IDI -- Palm Industry Grievances

Mapping Human Rights Violations in the Palm Oil Industry

Background:

Palm oil is a popular and versatile vegetable oil found in animal feed, biofuels, and nearly 50 percent of packaged supermarket goods and 70 percent of cosmetics. However, its production has caused significant environmental and social harm.

Companies eager to secure fertile, tropical soil on which to build palm fruit plantations have violated local and Indigenous communities' right to land and self-determination through land seizures, trespassing, coercive tactics, bribes, and failures to fully consult with communities' chosen representatives during land sales. Following their displacement, these communities are often unjustly compensated and spend years in land disputes. The Consortium for Agrarian Reform (KPA) documented 2,047 such conflicts in Indonesia from 2015 to 2019. In addition, industrial farming and palm oil refinement have reduced biodiversity, increased carbon emissions, and polluted nearby water sources due to run-off from fertilizers, pesticides, and palm oil effluents (POME). Communities have borne the brunt of these effects in the form of increased health risks and a loss of food sources, economic livelihoods, and cultural heritage.

To empower communities in their fight for land preservation, the Data Science Institute and the nonprofit **Inclusive Development International (IDI)** are building a tool called **PalmWatch** to identify "pressure points" of leverage in the palm oil supply chain—i.e., companies that fund, operate, or buy from palm oil mills and could therefore be held accountable by association, especially those that have made explicit public commitments to sustainable development. To date, however, no human rights data has been incorporated into the tool.

This Data Clinic will expand PalmWatch by integrating grievances filed by community members and nonprofit watchdogs as another data source. Expected tasks include:

1. Locating, documenting, and evaluating grievance datasets
2. Identifying relevant fields to extract and mine from the datasets
3. Mapping fields from each dataset to a common data model
4. Writing web and/or PDF scrapers to extract raw data from the datasets
5. Writing scripts to clean and standardize the scraped data in accordance with the model
6. Performing an exploratory data analysis of the cleaned data to describe and visualize patterns in complaints over time and across geographies, consumer brands, palm oil suppliers, and plantation companies

By doing so, students will increase the transparency of the number and type of complaints associated with brands and palm oil suppliers and allow PalmWatch's users to target those companies for reform.

Mentor:

Launa Greer is a software engineer at the University of Chicago Data Science Institute. Through a grant provided by The Schmidt Family Foundation's 11th Hour Project, she helps social impact organizations around the world investigate difficult research questions and communicate data to larger audiences through innovative technical projects. Prior to her current role, she worked as an adult education instructor and software consultant at a Microsoft partner company. She holds a bachelor's degree from Princeton University and a master's degree from the University of Chicago.

International Rescue Committee

Malnutrition Prediction

Background:

The International Rescue Committee (IRC) helps people whose lives and livelihoods are shattered by conflict and disaster to survive, recover, and gain control of their future. The Research and Innovation (R&I) team designs, tests, and scales life-changing, cost-effective solutions for people affected by conflict and disaster. Nutrition is a team comprised of technical advisors and specialists providing technical assistance to country programs and organizational thought leadership in nutrition. Together, R&I and Nutrition have identified *Tackling Child Malnutrition: Scaling innovations that improve access, coverage and cost-effectiveness of acute malnutrition treatment in children under five* as a Global Research and Innovation Priority (GRIP) to focus energy over the coming years in order to generate a set of breakthroughs to radically improve client outcomes and change the humanitarian sector.

Project Information:

In our pilots, we track key indicators on each patient undergoing malnutrition treatment, and we use dashboards that update in near real-time to review that data. Using the data we have, we would like to determine which demographic, anthropometric and environmental characteristics are predictive of whether a child will default (skip treatment before being discharged).

We are looking to bring together different datasets to develop predictive models: Data from IRC and publicly available geolocated data at the village or district level, including distance to health center, local weather, crop growth, conflict events, and other factors that may contribute to missing treatment, declining health, and mortality.

Through the use of predictive analytics, our goal is to improve outcomes through identification of high-risk patients, as well as reduce treatment costs without lowering effectiveness of treatment and improve program targeting to reduce mortality, malnutrition relapse, and recovery time per patient. Each of these elements is vital to increasing the number of patients we can admit and their potential for recovery.

Mentor:

Zach Tausanovitch is the Data Science Lead for Nutrition at IRC. He holds the MACRM degree from Harris at the university of Chicago, and has 10 years of professional experience across monitoring, evaluation, data science and international development. He currently leads a team that focuses on supporting research and introducing new initiatives to improve malnutrition treatment and research.

Internet Equity Initiative

Milwaukee Internet Equity and Device Testing

Background:
The Internet Equity Initiative ("IEI") at the University of Chicago is an interdisciplinary research initiative which aims to produce datasets, toolkits, and actionable research and insights to support communities across the United States.

You can find additional information about the initiative at our website.

One of the research focuses at the IEI is understanding the service received by home internet users. This study involves asking volunteers to hook up a device, at their house, which measures internet usage. This last year the IEI produced a study with the results of this data for Chicago and we would like to replicate this analysis for Milwaukee. This information is important for policy makers and regulators to understand the difference between advertised and actual speeds that customers see.

A second task associated with this project would be to verify and test the devices that people use. This would entail students taking the devices and connecting to their home internet and then sifting through the data reported to make sure that the data is being collected properly. Note that this would be optional, but it would be nice if at least one or two of the students on the team would be willing.

Mentor

Matthew Triano and Alexis Schrubbe would be the mentors of this project. Matthew is a Senior Data Scientist at the University of Chicago working at both the Crime Lab and the Internet Equity Initiative and Alexis is the Director of the Internet Equity Initiative. Alexis received her PhD from UT Austin doing research on digital inclusion while Matthew received his Master's in CS from Purdue and has worked on multiple high leverage projects across the University.

Invenergy ⚙

Gen AI for Energy News

Background:

Invenergy is the world's leading privately held sustainable solutions provider. We develop, own and operate large-scale renewable and other clean energy generation and storage facilities worldwide. Our home office is in Chicago, and we have regional development offices in North America, Latin America, Asia and Europe. To date, we have successfully developed 191 projects totaling more than 30,200 megawatts.

Project Description:

The energy industry is full of lengthy and complex documents which have meaningful impacts to renewable energy power plant profitability, development speed, and operational efficiency. However, there are too many documents to keep track of. New employees need resources to get up to speed on energy markets quickly, and experienced employees need a method to comb through detailed documentation in an expedient fashion.

The chatbot will be used across multiple teams including energy trading, risk and regulatory. It will answer questions about the details of Independent System Operators (ISO) rules and operations – the body that governs our electrical grid. Invenergy teams need information from these documents daily and the model could save weeks' worth of work in a given year.

There are three main objectives of this project:

1. Automatic ingest of documents via API or web scraping from multiple publicly available industry websites
2. Build custom energy document chat utilizing generative AI (using Azure OpenAI API).
    1. Leverage retrieval augmented generation (RAG) to create first chatbot solution
    2. Build accuracy assessment measurements collaborating with business subject matter experts
    3. Fine-tune model based (Invenergy will provide guidance and subject matter expertise on energy industry for fine-tuning)
        1. One or multiple of the following, depending on time:

1. Isolate paragraph(s)/page within longer document where information was sourced (e.g. section A paragraph 4) (could include human expert-generated examples)
2. Specify how to answer when it doesn't have the answer or when there are significant nuances related to the energy industry (could include human expert-generated examples)
3. Tune a small model for speed/cost

3. Web front-end interface which allows users to ask questions and receive answers from the trained model.
   1. Ideally, a way to collect responses/feedback on model responses to allow for future fine-tuning

The project’s main goal is to create the full pipeline for a generative AI chatbot on energy industry regulatory documents. If the full pipeline is built, we can focus on aspects to increase performance, such as including additional data or trying new fine-tuning techniques.

Kids First Chicago

Improving Data Accessibility and Transparency: Centralizing Chicago Public Schools Education Data

Background:
Kids First Chicago's (K1C) mission is to dramatically improve education for Chicago's children by ensuring high-quality public education is accessible to all families. One of the pillars supporting K1C's mission is data stewardship; we strive to improve data transparency and accessibility to empower parents and families to make informed decisions and take action in their children's education. In the spirit of this, we have begun a project that will centralize roughly 30 years of publicly available education outcomes and enrollment data, sourced from Chicago Public Schools (CPS) and the Illinois State Board of Education (ISBE). The end-goal is to create a queryable database accessible through the K1C website, where any person would be able to download filtered datasets related to their information of interest. Given that CPS education data is currently spread across multiple datasets, websites, and formats, it is our hope that this centralized database will improve general accessibility, allowing families, interested public figures, and researchers, alike, to be able to efficiently empower themselves with harmonized data and further progress education goals. We expect that these data will help in the investigation of several pressing research and policy interests, including the relationship between chronic absenteeism and education outcomes, and the impact of parent involvement on education outcomes.

To reach this end-goal, we will first need to harmonize the roughly 30 years of data. The majority of this project will focus on matching data formats, bridging gaps in variable name changes and values, creating crosswalks to match ISBE and CPS data, investigating variable definitions (and how they may change year-to-year), and creating a log of all variable name and definition changes over years. This work will be done using Python. If students are able to complete the harmonization aspect of the project within the available time, there is an opportunity to pursue exploratory analyses using the harmonized data, potentially looking at any number of publicly available variables and their relationship to academic outcomes.

Mentor:
Chris Poulos is a Senior Manager of Research and Policy at K1C. Micaelan Valesky is a Data Scientist at K1C.

Morningstar ⚙

LLM Performance Improvements

Background:

Many companies, including Morningstar, face significant challenges in effectively leveraging Large Language Models (LLMs) due to limited resources and budget constraints. The necessity for extensive prompt engineering often leads to a lack of control over the model's output, while full fine-tuning of LLMs is a resource-intensive process that requires substantial time and effort. Moreover, the creation of comprehensive datasets is a complex task, demanding attention to numerous contingencies. Unlike tech giants such as Google, Morningstar lacks a large user base, resulting in limited feedback—only a few dozen responses per day—which hinders the ability to implement Reinforcement Learning from Human Feedback (RLHF) effectively. These challenges impede Morningstar's ability to optimize LLM performance, necessitating an exploration of alternative approaches to better train and refine these models.

The objective is to improve the accuracy of our LLM outputs, which, despite generally effective prompt engineering, occasionally produce errors that can mislead clients or misinterpret technical terms. We aim to explore simpler, more streamlined methods that can nudge the LLM towards more reliable responses without the need for extensive fine-tuning or complex Reinforcement Learning from Human Feedback (RLHF).

Mentor Information:

Josh Charney is a Quant Research Manager and has been with Morningstar for 12 years. He holds a CFA, an MBA, and a master's in computer science from U Chicago.

David Wang is a Senior Principal Data Scientist and has been with Morningstar for 19 years. He holds a PhD in mathematics and computer science from the University of Illinois, Chicago.

RAFI – Grocery Atlas

MSA Analysis on Grocery Atlas

Background:

RAFI's Challenging Corporate Power initiative battles corporate consolidation in the food supply chain. While typically focused on issues closer to farming, consolidation in grocery stores impacts the rural communities that supply much of our food. RAFI engages in regular advocacy with legislators at the state and national level, and they need tools that tell a clear, visual story of market capture in our food supply chains. The grocery market has consolidated over time with large corporations buying up smaller regional chains. Additionally, these large conglomerates also merge with each other, as in the current Albertsons and Kroger merger. The DSI is helping RAFI visualize consolidation in the grocery market with an interactive time series map showing parent company ownership of grocery stores over time. In the future, this map will be publicly available on the web and will be used in conversations with legislators.

The current map can be found here: https://grocerygapatlas.rafiusa.org/

There are several features and analysis that we want to add to this map for this project. Specifically, we want to answer the following questions:

1. How do different metropolitan areas compare in their concentration and food access? Are there ranking mechanisms that we can use to identify which areas experience effects from market concentration?
2. Are there specific characteristics of different metro areas which make them more or less likely to see the effects of market concentration?

Currently the dataset does not have a metro area attached, so the first step will involve doing some data engineering work to add the metro area definition.

Mentor:

Dylan Halpern is the technical lead of the Open Spatial Lab at the University of Chicago. He has a Master's degree from MIT and a wealth of experience building technological solutions for human problems. He has been at UChicago working on a number of research and applied projects for the last four years.

RAFI – Poultry

Demographic overlay on Poultry Farms

Background:

RAFI's Challenging Corporate Power initiative fights consolidation throughout the food supply chain. The director of the Challenging Corporate Power initiative regularly engages legislators at the state and national levels to advocate for policy changes that prevent further consolidation. The DSI is helping RAFI build an interactive dashboard that shows consolidation in the poultry-packing industry. As large corporations gain control of more plants, farmers in monopsony-captured areas are vulnerable to exploitation. Short term, this map will be used in conversations with legislators to demonstrate areas of high market concentration. In the future, this map will be publicly available on the web.

Last year students worked on building a map of the capture areas of different regions. You can find a link to the map here.

One of the limitations of the current data is that there is no demographic information in the current map. The purpose of this quarters task is to add an additional geographic layer onto the map to identify any trends in what types of farmers are effected by this consolidation.

Specifically, we want to do the following:

- Identify a set of research questions and the variables that can answer them in the census data that will help us understand who is the most effected. Example variables could potentially include race/ethnicity and income. Identify which specific variables in the Census data represent this information.
- Update the current data pipeline to merge on the appropriate census information.
- Analyze how different capture area properties (such as number of integrators available) correlate with these indicators.

Mentor Information:

Trevor 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. Before DSI, Trevor worked as a research assistant at Argonne National Laboratory. Trevor has a BS in Computer Science from MIT.

University of Chicago Library

Optimizing Library Storage

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.

Project:
Physical book and journal publishing has not seen the downturn that many had expected with the rise of digital publishing. In fact, physical Library collections are growing as fast as they ever have. This puts pressure on libraries of all kinds to find sufficient space to house their collections on and off site, while facilitating fast access to those items for researchers and readers alike.

We propose a project to optimize storage of items across various locations to ensure continued access to materials without compromising service quality. Each storage location comes with different levels of service, particularly in terms of retrieval time, making it essential to strike a balance between efficient space utilization and timely access to resources.

The goal is to create a strategy that maximizes space efficiency while minimizing the impact on retrieval times for library patrons, as well as the Library's emissions and transport costs, all while ensuring that the library can continue to serve the academic community effectively.

Over the next academic year, a team of data science students will work closely with the University of Chicago Library to analyze the logistical dynamics of these storage options. The project will involve developing a comprehensive model to assess current storage usage and optimize the allocation of resources across different storage options. Students will employ advanced data analysis techniques, including predictive modeling and optimization algorithms, to recommend a storage strategy that aligns with the library's operational goals.

This project will require a deep dive into various data sets, including data about the physical items we hold and where they currently are ('collection' and 'holdings' data) and the anonymous requests and check-outs data, dating back 10 years (historical 'circulation' data), as well as data about electronic items (e.g., ebooks), especially where they overlap with or duplicate physical items.

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 the its physical collections.

University of Northern Iowa

Characterizing African American Young Adult Novels' Narrative Through Variations in Characters' Power Dynamics

Background:

The University of Northern Iowa (UNI) is a public university that offers more than 90 majors with a total enrollment of about 9000 students. UNI's main reputation is for its rich history in teacher preparation.

The present study aims to understand the dynamics of power within the narrative of African American YA novels, exploring how the presence or absence of power among characters in these novels shapes their stories. The purpose is to examine the dynamics of power, highlighting how characters may gain or lose influence throughout the narrative, and how these fluctuations impact their experiences and relationships. Identifying characters' power dynamic in Young Adult novels helps to explore stereotypes and negative patterns throughout the stories. Some important questions are: What makes a story powerful? Who holds or lacks power? How are power dynamics portrayed in these novels?

The main methodology for this study is to use a power-danger framework (Ousiogram suggested by Christopher M. Danforth and Peter Sheridan Dodds). To do that, it is important to leverage large scale language models for automated analysis of African American YA novels corpus to identify quantitative aspects in text that help define power across a character's narrative arc and how it alters for characters of different demographics.

Mentor:

Taraneh Matloob is an associate professor of children's literature at the University of Northern Iowa. She teaches multicultural children's literature as well as doctoral courses. Her scholarly interests are focused on multicultural children's literature, sentiment analysis, virtual reality, and augmented reality.

University of Rwanda

Climate change-induced multi-hazard risks: Flooding, Landslides, and Environmental Degradation in the Kaduha-Gitwe Corridor, Southern Province, Rwanda

Background:

The University of Rwanda (UR), established in 2013 through a merger of independent institutions, is the country's largest public university, serving over 31,000 students across multiple campuses. With a mission to become a globally recognized center for academic excellence and innovation, UR addresses local and global challenges through research, knowledge transfer, and community engagement. Its College of Science and Technology leads national efforts in sustainability, climate resilience, and disaster management.

The Data Science Clinic students will analyze climate-driven multi hazard risks in the Kaduha-Gitwe Corridor, focusing on flooding, landslides, and soil erosion. Using advanced data science, the project will identify key environmental and human factors contributing to these risks.

Expected outcomes include high-resolution vulnerability maps, predictive models, and policy recommendations aimed at improving disaster preparedness, sustainable land use, and environmental protection. Further, students will be mentored in data-driven decision-making through hands-on work with real-world data, fostering skills in disaster risk analysis and climate resilience.

Mentor:

Prof Aime Tsinda is an Associate Professor in Environmental Sciences at the University of Rwanda. He holds a Ph.D. in Environment and Sustainability from the University of Surrey (UK), a master's in urban planning from Université de Montréal, and a Bachelor's in Geography with Education from the University of Rwanda. Dr. Tsinda's interdisciplinary research focuses on sustainable water and sanitation, circular economy, environment and climate change, disaster risk management, and policy analysis.

Dr. Elias Nyandwi is a Senior Researcher and Lecturer at the University of Rwanda, specializing in geo-information sciences for environmental and sustainable development. He serves as the Director of the Centre for Geographic Information Systems and Remote Sensing (CGIS) at the university. Dr. Nyandwi earned his Ph.D. in Spatial Planning from the University of Twente, the Netherlands, and his research focuses on spatial analysis, environmental health, and natural resource management.