THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Winter 2024 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.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

American Family Insurance

Data Augmentation and Balancing via Generative AI

Background:
The Data Science & Analytics lab at American Family Insurance uses AI and machine learning to re-envision insurance in light of modern technical and consumer changes. While our work is varied, we generally create the most value at the intersection of large proprietary data sets and difficult prediction problems. Most of all, we're always learning. We work with the latest tools, technologies, and open source tools to stay ahead of the curve — and on top of our game.

Recent breakthroughs in Generative AI show a lot of realism, and hold a lot of promise for various use-cases. Perhaps one of the most appealing use-cases is to bolster datasets in support of outside AI applications. This project is meant to explore that use-case within a constrained image-type, namely external pictures of houses. Students will perform a literature review of available models, and select a model that they will implement.

Students will begin by training a simple classifier to predict house style (i.e. ranch, craftsman, cottage), or similar, using unbalanced data which will be provided or jointly curated. Generative approaches should then be implemented to augment the training data, and the same baseline classifier trained on the new dataset to assess the value-added of this methodology.

Mentor:
Tim Rouse is a Data Scientist at DSAL with a background in Computer Vision and Engineering. Sharath Thirunagaru is a Data Scientist at DSAL with an emphasis in Computer Vision and Project Leadership. Kayla Robinson is a Data Scientist at DSAL, and an alumnus from the University of Chicago (PhD '19). Qing Huang is a Data Scientist at DSAL and has a background emphasis in traditional Machine Learning and Statistical Learning.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Argonne/Fermi National Laboratories

Lessons learned analysis

Background:
The purpose of this project is to apply the NLP tools and techniques currently being used on Argonne work process documents to similar style documents at Fermi. The data that is being used is process and work planning documents across both labs which contain a series of "lessons learned" that are important for safety and planning purposes. When drafting a work plan making sure to find appropriate lessons learned from other work plans is a critical goal of both labs.

The Work Planning and Control team at Argonne did an assessment of work control documents in 2019 and found that nearly 50% did not contain relevant lessons learned. Incorporating lessons learned into new and ongoing work activities is part of DOE's Integrated Safety Management model. Finding relevant lessons learned is a difficult process because there are thousands that exist, and it can feel like trying to find the needle in the haystack. Using artificial intelligence and machine learning simplifies this process by sending relevant lessons learned directly to key players associated with the work activity, so they can incorporate the lessons learned into their documents, pre-job briefings, and safety shares.

This project will leverage the models built by Argonne and apply them to the similar (but not the same) documents used at Fermi. Once the models are created and verified there will be opportunity to apply NLP tools and techniques to increase their accuracy against search terms.

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.

Argonne National Laboratory

Simulating operational requirements management with a knowledge graph-based digital twin

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.

Argonne does not have a comprehensive process for identifying, collecting, and communicating requirements (e.g., statutory, regulatory, and contractual) applicable to the operation of the Laboratory. Disparate, complex, and manual processes exist for handling and applying changes to these policies and procedures, many of which revolve around understanding the impact on or from the Argonne Prime Contract. As a component of a future Argonne Digital Twin, we envision a broad-scope and largely automated operational requirements management system that can map requirements changes to relevant policies and procedures and even recommend implementations of these changes to augment the review process and final decision-making.

Modeling Argonne internal and external policy documents and standard procedures as an interconnected knowledge graph enables exploring the complex operational relationships and requirements spanning lab-wide policies. This deep level of understanding can then be integrated into our vision of a digital twin simulation of transmitting modifications, recommending missing relationships, and establishing an understanding of contextual similarities. In addition, an Argonne operations knowledge graph will support a chat bot-style question-and-answer user interface currently in development that will enable an intuitive interaction for information extraction, as well as drive future advanced analytics that could automatically predict policy changes or identify gaps in procedures that may require review and updates.

The next phase of the project proposed here will adapt our existing knowledge graph prototype of the Prime Contract by refactoring our natural language processing (NLP)-driven construction pipeline to integrate with a Neo4j graph database structure. We may also explore improvements in our document similarity measures by leveraging state-of-the-art embeddings generated from OpenAI models.

As we incorporate more operational documents into this system, a networked model of relationships between operations across the laboratory will provide the framework for information extract simulations to better understand the dependencies and interactions of

the policies and procedures. This framework will especially enable the automatic identification of possible impacts—at a granular context level—from implementing requirements mandated by the DOE or virtual "what-if" simulations to support decisions by laboratory leadership.

The University of Chicago students will be challenged in advanced data curation strategies, including building graph-style data structures and working with state-of-the-art NLP approaches necessary for this project while engaging in a rich and complex real-world business data set.

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.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

BankTrack

S.E.C. Commercial Loan Disclosure Pipeline

Background:
Internationally financed projects like dams, mines, and oil pipelines are notorious for environmental and human rights abuses, including forced displacement of Indigenous people, poisoned water sources, child labor, and physical and sexual abuse by foreign workers. Tracing the investment and supply chains associated with these projects allows researchers and community activists to identify the "pressure points" most responsive to advocacy, such as highly-public organizations concerned with their institutional reputations or organizations with prior expressed commitments to accountability and sustainability.

To increase transparency within the finance industry and empower advocates, the DSI and Inclusive Development International (IDI) have begun creating a suite of free and open-source online tools, starting with the Development Bank Investment Tracker (DeBIT), launched in May 2022. Over the past year, DSI and IDI have partnered with Netherlands-based charity BankTrack to begin the construction of a new data pipeline for commercial loans.

Data on investment chains is available in publicly accessible forms that are required to be filled out with the SEC by many companies in the United States. This data, however, is in unstructured text. To be able to extract information on corporate debt, previous teams have created a labeled dataset entity relations in SEC 8-K forms. Students this quarter will work on understanding and expanding the current dataset, perform analysis of the available data, and use natural language processing (NLP) tools and models to automate the extraction of commercial debt data from forms.

Mentor:
Ryan Brightwell: Ryan's mission at BankTrack is to lead their 'Banks and Human Rights' campaign and ensure that their research and communications are in solid shape. Ryan joined BankTrack in October 2012, and is Director of Communications & Research. He also coordinates BankTrack's campaign work on human rights. Before moving to Nijmegen he lived in Manchester where he worked at The Co-operative Group as a specialist in ethical finance. He holds a BSc in Mathematics and Management Sciences from The University of Manchester and a MA in Sociology from Manchester Metropolitan University.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Chicago Metropolitan Agency for Planning (CMAP)

NE Illinois Stormwater Storage and Site-Scale Green Infrastructure Inventory

Background:
CMAP is the region's comprehensive planning organization and serves the 7 counties of northeastern Illinois (Cook, DuPage, Kane, Kendall, Lake, McHenry, and Will). CMAP works with local governments and stakeholders to conserve and restore the region's water resources. CMAP also conducts policy and data analysis and shares relevant data with stakeholders to inform local land use and transportation decision-making.

Students will be tasked with using computer vision (deep learning) to identify existing stormwater storage locations in aerial photography. Students will train a model to identify different types of locations (for example, wet ponds, dry-turf bottom, dry-mesic prairie, and constructed wetland detention basins) and then use this model to identify other areas of the region with these attributes.

No comprehensive inventory of stormwater storage and green infrastructure (GI) assets exists across northeastern Illinois. Understanding the location of these assets is critical to ensuring proper maintenance as well as building a better understanding of the potential impacts to water quality and stormwater management. An inventory could help county and municipal stormwater engineers, public works officials, and others ensure proper maintenance. The data could also inform the development of watershed-based plans and resilience plans.

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.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Chicago Trading Co

Sentiment Analysis of social media postings to use in stock price predictions

Background:
Chicago Trading Company ("CTC") is a cutting-edge proprietary trading firm with a long-term vision and a clear focus on helping the world price and manage risk. Our fun and trusting culture inspires us to solve the industry's most challenging problems and take calculated risks in a collaborative environment. Started in 1995 by a team of forward-thinking traders, we are proud to call ourselves an industry leader that keeps making markets and each other better.

Social media postings have been exceedingly powerful in conveying positive or negative sentiment influencing stock price changes. Project aims to derive "sentiment score" to use in stock prices predictions. In the first phase of the project implement and evaluate documented theoretical approaches.

Once those theoretical approaches are implemented, the second phase of this project is to follow up with independent research to improve the previous results. We are looking to leverage contemporary research on this subject to generate sentiment scores with an acceptable degree of accuracy.

Mentor:
Natasha Pekelis is a Head of AI/ML Lab at Chicago Trading Company. Prior to that she ran Technology and Data Analytics at CTC. Prior to that she had senior leadership roles at various Global Markets Technology organizations.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Climate Cabinet

Campaign Finance Tracking

Background:
Climate Cabinet is a nonprofit that provides policy analysis for politicians looking to run on a climate platform or introduce climate legislation. In smaller, local elections, politicians typically do not have the resources to do in-house analysis. Climate Cabinet is like a climate staffer for these politicians. Since 2018, when it began as a volunteer team for a Texas state legislature candidate, it has grown to support hundreds of political campaigns across the country.

Climate Cabinet would like to empower local and state candidates to discuss the political influence of the fossil fuel industry relative to the clean energy industry vis-à-vis campaign financing. In 2010, the U.S. Supreme Court ruled in Citizens United v. Federal Election Commission that the First Amendment protected "money as speech" for labor unions, nonprofits, and for-profit corporations. Since then, federal campaign finance contributions from fossil fuel interests have more than doubled—from $35 million in 2010 to $84 million in 2018, with the vast majority of funding directed to re-elect candidates with a track record of voting against clean energy policies. However, there is currently no comparable analysis of campaign finance contributions at the state or local level.

To help Climate Cabinet move towards this goal, the DSI is creating a searchable online database of contributions from both industries and then analyzing the data to describe money flows over time. These deliverables will give candidates a fuller picture of what is happening in their state and allow them to craft more effective narratives on the campaign trail. Students this quarter will contribute to the project by writing scripts to standardize campaign finance datasets; visualize the results; and then analyze the cleaned data to describe historical trends and correlations.

Mentor:
Caleb Braun serves as lead data engineer at Climate Cabinet. He joined the team after working as a researcher at the International Council on Clean Transportation, where he ran models and analyses, and as a software developer at the Joint Global Change Research Institute. He holds a B.A. in computer science from Carleton College.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Compost Research & Education Foundation

Disposable Packaging Disintegration Analysis

Background:
The Compost Research & Education Foundation (CREF) researches the disintegration of compostable foodware and packaging to find correlations between different composting methodologies and the rate of disintegration. Through the Compostable Field Testing Program, facilities submit their composting results and CREF analyzes the data to find best practices in composting.

CREF is building an interactive dashboard that enables users to easily view analysis on different combinations of composting methods and materials. In order to analyze their composting data, CREF needs to standardize existing trial results into a consistent database structure and develop a process for facilities submitting new data. This database will facilitate analysis of existing data and will also serve as the back-end for a future website. Results will also be presented at conferences and in white papers. Additionally, CREF will use existing data in order to inform the design of future controlled experiments testing different composting methods.

Mentor:
Emily McGill is the Program Director of the Compostable Field Testing Program, an international research project gathering real-world disintegration data for compostable items break down in composting facilities across North America. With a background in Bioresource Engineering, she coordinates field tests and is instrumental in the advancement of standardized methods for field testing within ASTM. Her experiences span solid waste management planning at corporate and municipal levels and developing policy and education for zero waste and single-use plastic reduction. Since 2015 she has also fostered collaborative community-based projects in urban sustainability, circular economy and regenerative systems design in the respective areas of deconstruction, textile waste prevention, and urban agriculture.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

DRW

Predicting Short Term Volatility

Background:
DRW is a diversified trading firm with decades of experience bringing sophisticated technology and exceptional people together to operate in markets around the world and across many asset classes.

Very short term options (often called "0DTE" for zero days to expiration) have recently surged in popularity. In response the CBOE created a new index for one day volatility called VIX1D, disseminated starting on April 24, 2023. The VIX1D index is a short-dated version of the venerable VIX volatility index, which underlies some of the most liquid volatility derivatives available to trade.

Our project is to study and attempt to predict the VIX1D and related volatility indexes, using historical patterns in the index and in underlying or related securities.

Mentor:
Ian Adam has been a senior quantitative strategist in DRW's US equity and index options group since 2015. Before joining DRW he was a quant strategist in a high-frequency options trading firm in New York since 2008. He holds an AB in Physics from Princeton University and a PhD in Physics from Columbia University.

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 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 additional decoders to perform further classifications and regressions. In particular, over the summer we expanded the network to regress the neutrino interaction point in 3D. This is crucial information for a precise measurement of the neutrino interaction. The network is able to learn how to perform this task, but its performance has room for improvement, both from the physics and computing point of view. In this project we will work to improve the interaction point decoder to reduce the computing resource usage and to improve the accuracy by studying different network configurations and hyperparameter values. The results will be evaluated on the MicroBooNE open data sets.

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.

International Rescue Committee

aprendIA: Personalized education for last mile learners

Background:
The mission of the International Rescue Committee (IRC) is to help people whose lives and livelihoods are shattered by conflict and disaster, including the climate crisis, to survive, recover and gain control over their future. At the Airbel Impact Lab, we design, test, and scale life-changing solutions for people affected by conflict and disaster. Our aim is to find the most impactful and cost-effective products, services, and delivery systems possible. We work to develop breakthrough solutions by combining creativity and rigor, openness and expertise, and a desire to think afresh with the experience of a large-scale implementing organization.

IRC is interested in leveraging data that is currently being generated by Audio-Class, an app that provides lessons to students via their mobile devices. This project will attempt to measure and analyze engagement and product usage. The Clinic team will:

- Analyze usage patterns around when and how students use Audio-Class.
- Define retention and engagement metrics that reflect students' behavior.
- Build models to identify predictors of student success.
- Understand how students use these tools and what qualities of their usage predict success.

Mentor:
Atish Gonsalves leads the Education Global Research and Innovation Priority (GRIP) team at the IRC. Atish is also the founder of Gamoteca, a platform and app to easily create human-to-human, collaborative learning experiences. With a background in software engineering, AI and human-computer interaction, Atish's experience includes leadership roles at technology and international non-profit organizations and multilateral institutions, including the United Nations.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Internet Equity Initiative

Location Fabric analysis

Background:
The FCC has been working on a map of broadband access across the United States that they call their "Location Fabric". This map will be used to identify underserved areas and direct investment in broadband infrastructure. As part of their map building process they allow individuals and institutions to submit "challenges" (basically places where the map may be incorrect). Over the last year the FCC has collected these challenges and adjudicated them while releasing the data to the public.

The purpose of this project is to understand and analyze the geographic patterns of where challenges were successful and where they were not. Are there specific features of challenges that make them more or less likely to be successful? What states have the highest challenge success rate? Which is the lowest?

The first step of this project would be to collect and build a data pipeline to get the challenge information for all 50 states. We would then want to link this to demographic information at the county level for the purposes of trying to build models to understand which features drive challenges and their success. This project will require using python and jupyter notebooks to build a data pipeline. It will also use geospatial analysis to build maps and understand the underlying challenge data.

Mentor:
Jonatas A. Marques joined DSI as a postdoctoral scholar in July 2023, and was previously a PhD student in Computer Science at the Federal University of Rio Grande do Sul (UFRGS, Brazil), advised by Luciano Paschoal Gaspary. His current research interests are on the intersection of machine learning and computer networking, with focus on programmable networking and network management. Jonatas is part of the Internet Equity Initiative at DSI, with the goal of measuring and analyzing Internet performance and reliability to address inequity in U.S. communities.

Invenergy

DSI - Icing

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:
Turbine blades can ice up during environmental conditions which affects the operation of the wind turbine. Power output is impacted significantly, meaning the turbine can produce less when the blades have ice build up. Some turbines are required to shut down if ice is detected, due to the risk of ice throw impacting the environment (close to roads, people, etc.).

Invenergy currently does not have a model to forecast or detect ice minus simple physics-based methods using temperature data and turbine underperformance. If we had a model to forecast icing ahead of time, we could use this information for day ahead market and planning purposes since we wouldn't expect the turbines to produce their full capacity. If we had a model to classify ice, we could calculate total lost production due to ice, which is a common ask from our asset management team and customers.

Mentor:
Zoe Kimpel is a Senior Data Science Manager with Invenergy.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Morningstar

Narrative Integrity – Automated Morningstar Reports

Background:
Morningstar, Inc. is an American financial services firm headquartered in Chicago, Illinois and was founded by Joe Mansueto in 1984. It provides an array of investment research and investment management services. Our mission is to empower investor success. We've empowered investors all over the world, and we're continuing to look for new ways to help people achieve financial security.

Millions of investors globally rely on Morningstar's fund rating and reporting to make investment decisions that impact their financial and personal well-being. Written reports are one of the ways we contextualize how investors should think about these investments and make decisions. In the absence of clear quality controls, a rogue natural language generator could provide bad financial advice resulting in bad outcomes. The objective of this project is to use NLP to fact check ratings reports generated by a Chat-GPT process.

Mentor:
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 UChicago. 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.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Perpetual

Reusable foodware system design

Background:
Perpetual is a non-profit organization that aims to reduce plastic waste through city-wide, reusable foodware systems. The DSI is currently collaborating with Perpetual to implement a model which will help make foodware distribution feasible, sustainable, and convenient for users. Upon completion, Perpetual plans to pilot the model in four U.S. cities.

Clinic students will work on the following set of tasks:

- Compile and analyze supplementary datasets, such as foot traffic and population density data, to pinpoint the most strategic locations for placing collection bins.
- Analyze and extrapolate data from the on-ground surveys conducted by Perpetual to accurately estimate the capacity and type of reusable foodware required by different establishments, and to discern usage patterns.
- Build upon prior work on vehicle routing and distribution. Develop sophisticated algorithms and incorporate the latest advancements in routing mechanisms to construct an optimal and resilient routing strategy.
- Conduct a feasibility study to determine the optimal set of parameters, such as the number of vehicles, coverage area, and bin placement.

Mentor:
Ellie Moss is the Founder and Executive Director of Perpetual. She possesses experience in strategy consulting and developing environmental and social impact strategies for corporations, investors and nonprofit organizations, with particular expertise in circular economy solutions and food and agricultural systems. Ellie has an MBA from Wharton.

Prudential

[TBD]

Background:
Prudential Financial, Inc. (NYSE: PRU), a global financial services leader and premier active global investment manager with more than $1.5 trillion in assets under management as of March 31, 2022, has operations in the United States, Asia, Europe and Latin America. Prudential's diverse and talented employees help to make lives better by creating financial opportunities for more people.

This project is still under discussions, but Prudential is a valued partner and the projects under discussion are all very interesting. In the past, projects have included:

- Using NLP techniques to analyze news and financial filings
- Using machine learning tools and techniques to predict stock performance based on signals generated from publicly traded stocks
- Determining geographic areas to engage in multi-family real estate housing investments.
- Using earnings call information and NLP techniques to build portfolios of stocks to invest in.

Mentor:
Amol Tembe leads the Corporate Functions Data Science team within Prudential's Chief Data Office. He obtained his Ph.D in Computer Science (2004) and Executive MBA in Global Business (2012).

Robert Huntsman is the Chief Data Scientist for Prudential's US Businesses. Robert is a graduate of Stanford, the UCLA Anderson School and a CFA charter holder and has over 20 years of experience as a senior executive with leading financial services and technology firms.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Rural Advancement Foundation International (RAFI)

Poultry Packaging Consolidation

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.

Previous work on this project used historical business data to show that many states have highly concentrated poultry packaging industries. This analysis was completed using revenue estimates for specific poultry packaging plants. However, RAFI would like to frame the impact of this market concentration in terms of the actual farmers affected. However, finding accurate business records for the number of farms in a given area is not plausible. Microsoft built a computer vision model that is able to recognize poultry barns from aerial photography images available from the USDA. While this model shows proof of concept, it has a lot of false positives — recognizing roads, fields, shorelines, and warehouses as poultry barns. The DSI will implement rules-based filtering to eliminate false positives. We may also retrain Microsoft's model to recognize some of these other structures. By recognizing poultry barns from aerial photography, we will be able to reasonably estimate the number of poultry barns in monopsony-captured areas.

Mentor:
Aaron is the Program Manager for the Challenging Corporate Power Program at RAFI-USA, working to end trends of concentration and extraction in the meat industry and build resilient community-rooted animal agriculture economies. Prior to joining RAFI-USA, he worked as a program designer and evaluator for several youth asset-building nonprofits in Durham, NC. Originally from Illinois, Aaron grew up spending his summers on his grandfather's hog and dairy farm, and later working as a farmhand on his uncle's hog farm. These formative experiences shaped Aaron's passion for fighting for more ecologically just food systems.

University of Northern Iowa

How Do Fictional Characters Feel? A Sentiment Analysis of Characters' Emotional Status in YA Novels

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 recognize the experience of African American fictional characters, examining how their emotional status might reveal a recurring pattern. Identifying the most frequent emotions used in Young Adult novels helps to explore affective patterns throughout the stories and identifies multicultural characters' critical times, detecting their emotional state as well as affectual shifts.

In this project, we explore how sentiment analysis can be used as a tool to quantify basic emotion words and identify emotional patterns in novels. There are different approaches for emotions analysis and sentiment analysis. The methods can be simple or more sophisticated. It is preferable to analyze the words in the context not as isolated words.

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.

WBEZ

Illinois Traffic Stops

Background:
WBEZ is Chicago's NPR news station, one of the largest and most respected public media stations in the country. Throughout our history we have had a legacy of innovation as the birthplace of the most iconic shows in public media, such as *This American Life*, *Serial* and *Wait, Wait…Don't Tell Me*. And today, we are more dynamic and forward-looking than ever before.

We have collected, cleaned and standardized over 19 years of traffic stop data from across the state of Illinois. This is nearly 100 columns of data that we are excited to analyze and understand. The data includes information about the stop, such as location, what entity did the traffic stop and some basic information about the person being pulled over. This is an incredibly unique dataset which can be analyzed in several different ways.

Our first project is to try to build a relative risk model of being pulled over. We would like to compare people who are pulled over against the state at large and understand the relative frequencies of being pulled over. This would involve doing some research on statistical methods for analyzing similar data. The end goal of this project would be to build a set of visualizations to communicate the results to the public at large.

Mentor:
Matt Kiefer is an editor at WBEZ, specializing in acquiring, analyzing and visualizing data for investigative journalism.