THE UNIVERSITY OF CHICAGO
DATA SCIENCE
INSTITUTE

# Winter 2023 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.

# American Family Insurance

*Resilient Futures*

**Background:**
The insurance industry touches nearly every American's life as a crucial part of our economy. In 2014, the Data Science & Analytics Lab was born as part of American Family Insurance's commitment to leading the industry's response to new challenges. We're part of AmFam's Business Development division focused on evaluating, understanding, and fully utilizing new information sources.

Weather-related damage makes up a significant percentage of claims paid by American Family Insurance. Even so, homeowners or prospective homeowners are not always knowledgeable or aware of the specific types of weather risks faced by homes in their location.

By leveraging historical records of storm event type, location and severity, you will identify weather patterns and changes in patterns over time across regions of the country. You will use this data to derive insights and information to educate the public about storm and weather hazards faced by homes in their location.

**Mentor:**
Kayla Robinson is a data scientist at American Family Insurance, and an alumnus from the University of Chicago (PhD '19). Chris Billman is a senior data scientist at American Family Insurance with a background in telematics and data science product development. Jo Hadera is a data engineer at American Family Insurance, and an alumnus from the University of Wisconsin – Madison.

**Technology:**

- Python
- Plotly
- geopandas
- pySAL

# Argonne National Laboratory

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

**Background:**
Argonne is a multidisciplinary science and engineering research center, 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 do by itself.

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

This project will deliver an extended knowledge graph-based model of the information contained in the Prime Contract and additional policy documents, such as external DOE requirements documents, federal requirements (examples include the Federal Travel Regulation and the Federal Acquisition Regulation) and internal Argonne manuals, policies and procedures.

Development will build on and evaluate foundational progress recently completed on a proof-of-concept simulating the change impacts between the Prime Contract content and updated DOE orders.

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. If the Laboratory does not remain in compliance with changed requirements, then corrective action plans can be implemented to enforce returns to compliance. 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 this project proposed here will extend our existing knowledge graph prototype of the Prime Contract by scaling out our natural language processing (NLP)-driven construction pipeline to additional Argonne policies and procedures. 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.

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 is a software developer 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 Illinois Tech investigating advanced HPC scheduling algorithms and an Adjunct Instructor in Computer Science at UIS.

Kim has been with the laboratory for 32 years and has worked in multiple operations divisions throughout her career. Kim is responsible for coordinating activities that support the AI for Operations initiative.

**Technology:**

- Python

# 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 University of Chicago Data Science Institute (DSI) and Inclusive Development International (IDI) have begun creating a suite of free and open-source online tools. The Development Bank Investment Tracker (DeBIT), launched in May 2022 with support from Accountability Console and the 11th Hour Project, provides quick access to key information on more than 250,000 development projects financed by 17 finance institutions, as well as complaints filed against these projects. The Shareholder Tracker compiles stock purchases from 80 of the largest institutional investors in the world, reported quarterly through the U.S. Securities and Exchange Commission's (S.E.C.) 13F form. This past summer, DSI and IDI partnered with Netherlands-based charity BankTrack to begin the construction of a new data pipeline for commercial loans.

Two previous teams made a first pass at a pipeline that fetches 8K forms from the S.E.C., extracts and cleans relevant form sections, and then parses loan issuance dates, amounts borrowed, debtor and lender names, etc. from the text, before saving it in a database. This quarter, we will write scripts and train different NLP models to retrieve loan information from 6K and 20F forms—reports filed by international companies outside the United States. Our corpus of documents has the potential to be multilingual.

**Technology:**

- Data engineering
- NLP
- Information retrieval
- Docker

# Blue Ocean Gear

*Fishing Gear Anomaly Dashboard*

**Background:**
Blue Ocean Gear, a startup incorporated in California in 2019, has developed a Farallon Smart Buoy System™ that tracks fishing gear in the open ocean for commercial fleets. The product is a buoy that is tied to fishing gear (e.g., nets, cages, and pots) with a rope. The buoy floats on the surface of the water and reports its location, surrounding temperature, and other metrics via radio or satellite on a schedule configured by the fishers.

The company is dedicated to preventing lost fishing gear and the subsequent negative impacts of ghost fishing, while also improving operational efficiency for fishers through better data tracking. Thus far, roughly 100 buoys have been deployed in Alaska, California, Massachusetts, Maine, and several areas around Nova Scotia; about a dozen more were deployed by two customers off the coast of British Columbia this past year and nearly 150 are currently operating in New Brunswick.

A few anomalous events have occurred where fishing gear broke free from a Smart Buoy due to strong ocean conditions or hurricanes. Anomalous events can also include long periods of submergence or frequent exit/entry from the water. Blue Ocean Gear would like to leverage available datasets on ocean conditions to better predict when anomalies occur and where buoys that have broken free will travel ("drift").

Previously, students conducted an exploratory data analysis of buoys located in fisheries off the coasts of Maine, Massachusetts, and New Brunswick to describe their typical motion patterns and compare their recorded temperatures with those from third-party sources like satellites and weather stations. They also completed a first pass at an LSTM (long short-term memory) model that predicted the motion paths of drifting Blue Ocean Gear buoys after being trained on both synthetic and actual drifter data.

Students this quarter will focus on modeling. Their first task will be to improve the current deep learning model and experiment with additional models like transformers to predict the locations of drifting buoys as well as buoys that are still attached to their fishing gear. They will also use clustering algorithms (e.g., k-means, DBSCAN, and Isolation Forest), simple heuristics, and deep learning methods like auto-encoders to identify potential location outliers.

**Mentor:**
Will Morton is a cloud engineer with 25 years experience hacking, building and supporting internet applications. Before joining Blue Ocean Gear, he worked previously at Apple and Beats Music.

**Technology:**

- Python
- Anomaly detection
- Deep learning
- Forecasting
- Simulations
- Docker
- Pytorch

# Business and Professional People for the Public Interest (BPI)

*Chicago Police Department Traffic Stops*

**Background:**
BPI is a non-profit, public interest law and policy center that utilizes a combination of legal tools, policy research, advocacy, organizing, and convening to work towards transformational change. BPI is dedicated to advancing nuanced solutions to pressing racial, economic, and social justice issues within and across its program areas of Criminal Legal Systems/Police Accountability and Housing.

Law enforcement traffic stops are the most common interaction that individuals have with law enforcement. In Chicago, the number of those interactions have skyrocketed since 2015 and are concentrated among Black and Latinx communities and drivers. Chicago Police Department has said that traffic stops—often for common, low-level offenses—are one of the ways it deters and detects crime in the city. Prior DSI students have reviewed and analyzed data from Illinois and Chicago public agencies to investigate racial disparities in how traffic stops are conducted by Chicago Police Department. This quarter's project will build on that prior work by:

1. Conducting a preliminary analysis into whether or to what degree traffic stops have had an impact on crime levels in Chicago in order to test CPD's contention that they are an effective tool in addressing crime in the city.
2. Developing data visualizations for our future website, which will help communicate to the public, politicians, and officials the high number of traffic stops and the disproportionate impact on the South and West sides and their Black and Latinx residents. Students would get practice in pulling out compelling lessons from data as well as translating large amounts of data into visualizations that are accessible and useful to the general public.
3. Drafting explanations of the data analysis and visualization choices and process for the website. This will be useful not only to help people better understand what the visualizations are saying but also to help advocates in other cities develop ideas about how to analyze and present traffic stops data in their cities.

This work will help support our coalition of community and advocacy organizations in Chicago called Free to Move, which seeks to create a safer system of transportation in Chicago by investing in racially equitable transportation infrastructure and decreasing police enforcement. The coalition is in the process of building out a policy platform of the ways in which racially inequitable traffic and transportation systems in Chicago should be addressed; this project will directly inform our broader public education campaign about traffic stops and our ability to effectively respond to opposition.

**Mentor:**

Amy Thompson joined BPI's Criminal Legal Systems/Police Accountability Team in 2019 as Staff Counsel and a legal fellow. Amy's work at BPI involves conducting research and providing analysis on local and state policing policy matters, engaging with community partners and coalitions, and supporting the Criminal Legal Systems/Police Accountability team's ongoing efforts to reimagine public safety. Her work has included researching and crafting recommendations to strengthen Illinois' police officer decertification system, advocating for strict limitations on police foot pursuits, and contributing to a coalition's efforts to reform Chicago's police union contracts. Amy received her BA in 2013 from Lewis & Clark College and her JD in 2019 from the University of California Los Angeles School of Law.

**Technology:**

- Python
- pandas
- Plotly
- Data visualization
- Geospatial analysis

# Citizen Data

*Estimating Political Ideology from Voter Files*

**Background:**
Citizen Data maintains a national voter file in-house. National voter files start with data from state voter registration and layer on consumer, Census, and inferred demographic data. This data is then used to inform campaign strategies, research, and is leveraged as inputs to other machine learning models. Among the fields of interest in this dataset is a person's political ideology (how liberal or conservative they are) which has a significant impact on political strategy. However, this data is incomplete.

The goal of this project will be to build a model which leverages our self-reported data to estimate each voter's political ideology using modern statistical methods. This data will be used later downstream to offer improved segmentation, filtering, and modeling in the future.

Analysis is expected to take place in Python. Students will be expected to perform data cleaning, feature engineering on the provided data, matching the data to other data sources where relevant, and train machine learning models using stable Python packages.

**Mentor:**
Kyle Redfield is Citizen's Lead Data Scientist with experience working in and consulting for the federal government. He received his master's degree in Data Science from the University of California Berkeley and was the founding data hire for Citizen. He has experience in traditional economic research, machine learning, data engineering, and the intersection of those disciplines as they apply to providing actionable insights to clients.

**Technology:**

- Python
- pandas
- scikit-learn

# DRW Holdings

*Realized Volatility Patterns and Option Pricing*

**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.

A great deal of research has attempted to relate realized volatility to implied volatility, a key determinant of option prices. For example, we expect that the prices of call and put options on AAPL stock should be related to the recent volatility of AAPL stock returns - in this example, Apple stock is the "underlying." Yet, the relationship remains elusive.

Realized volatility is usually defined as quadratic variation of underlying returns, but we can extend the concept to encompass all the information in the history and pattern of underlying prices. We will review previously used approaches to modeling realized volatility and its relationship to option prices and build baseline models to benchmark these previously published results. Then we will attempt to develop new models that use the available realized return information patterns better, and investigate whether these beat previous approaches.

**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.

**Technology:**

- Python
- numpy
- scikit-learn

# Enrico Fermi Institute (EFI)

*Application of Machine Learning to the Development of Ultra-low-Dose Positron-Emission-Tomography With Large-Area Pico-second Photodetectors (LAPPD)*

**Background:**
The EFI is in the Physical Sciences Division of the University of Chicago. The Institute supports multi-disciplinary research and provides world-class technical support.

The project proposed here is to employ Machine Learning to develop ultra-low-dose Time-of-Flight Positron Emission Tomography (TOF-PET) using the LAPPDTM large-area psec photodetectors we have developed.  The goals of the PET development are: 1) higher resolution by at least a factor of 10; 2) a reduced dose to the patient by at least a factor of 100;  3) a substantial decrease in  cost, complexity, and infrastructure; and 4) access to PET scans in rural and third-world locations distant from major medical centers.

The new PET detector would produce two kinds of `raw' data: spatial images of energy clusters in the detector medium, and a map of locations and precise times of optical photons at the LAPPDTM photodetectors. We want to use both kinds of data to time-order the energy clusters to find the earliest one. The two data sets are correlated, as each optical photon is emitted at the location of one of the energy clusters. The task for Machine Learning is to find the most probable assignment of the photons to the clusters to produce a time-ordered list. We have a detector simulation running in the TOPAS/XCAT framework that produces both simulated data and the `truth' information for optimization studies.

**Mentor:**
Henry J. Frisch

Born Rural Sandoval County (Los Alamos) N.M., Aug. 1944
BA Harvard 1966
PhD Berkeley 1971
Faculty member Physics and EFI, Univ. of Chicago; 1971-present
Home page: hep.uchicago.edu/~frisch

**Technology:**

- Python
- Pandas/numpy
- scikit-learn

# Fermi National Accelerator Laboratory

*Automated anomaly detection in data acquisition systems*

**Background:**
Fermilab is a DOE National Laboratory, focused on particle and accelerator physics. Fermilab collaborates with researchers and institutions around the world, probing the fundamental nature of the universe at the highest energy particle collisions, with intense neutrino sources from accelerators, and through measurements of the cosmos.

As particle physics detectors grow in size and complexity, monitoring and understanding the flow of data through their data acquisition systems becomes increasingly necessary but also increasingly complicated. This project would include the development of visualization tools to help detector experts understand the operation and health of data acquisition systems. Using data from currently operating detectors, we can test the use of these tools, and how they scale with an increasing number of nodes. A further goal of this project is to use the data available to develop and test anomaly detection techniques, which would include the use of AI/ML. Automated anomaly detection in data acquisition systems will help lead to early-warning signs of potential problems, increasing operational efficiency.

**Mentor:**
Wesley Ketchum: I graduated with a Ph.D. in Physics from the University of Chicago in 2012, and have been engaged in particle physics experiments at Fermilab throughout my Ph.D. and since. Currently, I primarily work on neutrino physics experiments that use Liquid Argon Time Projection Chambers (LArTPCs) to take detailed images of interactions of neutrinos with matter. I specialize in data acquisition systems for these detectors, along with the processing and analysis of data.

**Technology:**

- Python
- Pandas/numpy
- scikit-learn

# Fermi National Accelerator Laboratory

*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.

Fermilab is partnering with the University of Cincinnati and Northwestern University with the goal of developing a GNN for particle reconstruction in Liquid Argon Time Projection Chamber (LArTPC) neutrino experiments. Our development focuses on wire-based LArTPC, and is targeting the far detector of the future Deep Underground Neutrino Experiment (DUNE) as well as current smaller-scale experiments, such as MicroBooNE. Wire-based LArTPCs typically utilize three readout planes, each made of sense wires that measure the charge of drifting electrons produced by the ionization of the argon by charged particles. Wires in each plane have different orientation, so 3D information can be inferred combining the 2D measurement of different planes.

We have developed a message-passing graph neural network (GNN) that is used to classify the nodes, defined as the charge measurements or hits, according to the underlying particle type that produced them. Thanks to special 3D edges, our network is able to connect nodes both within and across wire planes, and achieves 94% accuracy with 97% consistency across planes. In this project we plan to expand the network so that hits are also classified in terms of additional properties related to the different particle instances in the interaction, thus clustering together hits from the same particle and identifying their start and end points. The planned activities and tasks are:

1. Complement ground truth labels including an instance counter for all nodes, and node categories corresponding to neutrino interaction point, particle instance start and end positions.
2. Modify the network loss function and potentially the model to perform object condensation and classification of endpoints. Integrate the development into the original GNN, thus extending its classification targets.
3. Evaluate the performance of the new network under different data set configurations, including growing complexity in terms of the background from cosmic ray data.

**Mentor:**
Giuseppe Cerati: 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).

**Technology:**

- Python
- Graph neural network
- PyTorch

# Fermi National Accelerator Laboratory

*Real-time Tagging and Triggering with Deep Learning AI for next generation particle imaging detectors*

**Background:**
Fermilab is a particle physics and accelerator laboratory in the United States. Since 1967, Fermilab has worked to answer fundamental questions and enhance our understanding of everything we see around us. As the United States' premier particle physics laboratory, we do science that matters. We work on the world's most advanced particle accelerators and dig down to the smallest building blocks of matter. We also probe the farthest reaches of the universe, seeking out the nature of dark matter and dark energy.

The current and future programs for accelerator-based neutrino imaging detectors feature the use of Liquid Argon Time Projection Chambers (LArTPCs) as the fundamental detection technology. These detectors combine high-resolution imaging and precision calorimetry to allow for study of neutrino interactions with unparalleled capabilities. However, the volume of data from LArTPCs will exceed 25 Petabytes each year and event reconstruction techniques are complex, requiring significant computational resources. These aspects of LArTPC data make utilization of real-time event-triggering and event filtering algorithms that can effectively distinguish signal from background essential, but still challenging to accomplish with reasonable efficiency, especially for low-energy neutrino interactions.

At Fermilab, we are developing a machine learning based trigger and filtering algorithm for the flagship experiments at Fermilab (Deep Underground Neutrino Experiment) to extend the sensitivity of the detector, particularly for low-energy neutrinos that do not come from an accelerator beam. Building off of recent research in machine learning to improve artificial intelligence, this new trigger algorithm will employ software to optimize data collection, pre-processing, and to make a final event selection decision. Development and testing of the trigger decision system will leverage data from the MicroBooNE, ProtoDUNE and Short Baseline Neutrino (SBN) LArTPC detectors, and will also provide benefits to the physics programs of those experiments.

While collaborating with DSI, we propose to first apply a Convolutional Neural Network (CNN) to the MicroBooNE data and study the performance metrics such as memory usage and latency. We would also like to deploy a Semantic Segmentation with a Sparse Convolutional Neural Network (Sparse CNN) on the same data and compare the performance of the two algorithms. The images produced in detectors are ideal for the application of a Sparse CNN which could improve the performance of the algorithm in terms of both memory and timing. While the addition of semantic segmentation would extend the capabilities of the trigger algorithm to allow for different data streams and pipelines.

**Mentors:**
Dr. Michael Kirby (Senior Scientist) : My scientific work has concentrated on Electroweak and Higgs physics during the last 15 years, but I am now focused on Neutrino Physics and the exciting new measurements possible with Liquid Argon Time Project Chambers.

Dr. Meghna Bhattacharya (Research Associate): As a graduate student I worked on the Muon g-2 experiment at Fermilab. I joined the MicroBooNE and DUNE experiments as a postdoc within the Computational Science and AI Directorate at Fermilab. My diverse background in physics ranges from working on hardware upgrades, analyzing physics data in both muon and neutrino experiments to the betterment of computing aspects for large scale experiments at Fermilab.

**Technology:**

- Python
- Deep learning

# First Republic Bank

*Project Toffee – How sticky are pandemic deposits likely to be?*

**Background:**
Founded in 1985, First Republic is a publicly traded (NYSE: FRC) institution with over $200B in assets, offering private banking, business banking, and wealth management services. First Republic specializes in delivering exceptional, relationship-based service. We've experienced tremendous growth over the past 7 years, more than tripling in size, and our customers love us! Each year 50% of our growth comes from existing clients with another 25% from direct referrals by these clients. Additionally, our Net Promoter Score (measuring client loyalty and likelihood to refer) exceeds that of the U.S banking industry by a factor of 2, and even exceeds vaunted luxury brands such as Apple and Nordstrom. This project is being led by the bank's treasurer and some of the Treasury department's data science / engineering colleagues.

We are looking to build a model that contemplates the response of non-interest bearing deposits, based on a variety of customer and account characteristics, to changes in key interest rates, and other monetary policy drivers such as the size of the federal reserve balance sheet, and the presence of other fed programs such as the reverse repurchase program and the size of such programs. This project would act as a challenger model to existing models used by the bank to forecast the path of non-maturing deposit balances. In times of rapidly rising rates this kind of analytics can play an important role in managing the balance sheet of a bank. We foresee the need to move beyond simple regression analysis to capture nuanced interactions between micro and macroeconomic variables over time, and we also hope to examine if the model and findings are generalizable to the overall industry controlling for various differences between banks.

**Mentor:**
Mark Woodworth is Head of Treasury Engineering and works on building and enhancing systems which support the treasury team's reporting and analytics capabilities. Mark's background includes 8 years within treasury focusing on liquidity stress testing, and 6 years as a high yield debt analyst. Mark has a B.S.c. in Finance from the University of Texas at Dallas, and is a CFA charterholder.

Chris Csiszar is a Senior Data Scientist focusing on model research, design, and development for deposit forecasting and various liquidity regulatory compliance efforts. He's been doing econometrics or some type of data analytics for 6 years now and has a B.Sc. in Mathematics & Economics from UCLA and a M.Sc. in Data Science from the University of San Francisco.

Xu Liu is a Data Scientist focusing on unfunded commitment stress tests, various automation jobs, and data visualization. She has a B.Sc. in Computer Science and in Finance and a M.Sc. in Data Science from the University of San Francisco.

**Technology:**

- Python
- Pandas/numpy
- scikit-learn

# GreenWave

*Machine Vision on Kelp Farms*

**Background:**
GreenWave is a nonprofit focused on regenerative ocean farming. Regenerative agriculture in general is a set of practices that aim to make agriculture systems less extractive and more beneficial for the ecosystems they exist in. Kelp farms are a key component of regenerative ocean farming due to kelp's vital role in ocean ecosystems. GreenWave's goal is to support 10,000 farmers, catalyzing the scaled planting of regenerative ocean crops to yield meaningful economic and climate impacts. To scale, they focus on two program areas: Training and Innovation.

One of the ways GreenWave supports farmers is through organizing the Kelp Climate Fund. A recent study 2 sought to calculate the market value kelp provides to the surrounding environment. The Kelp Climate Fund promotes the environmental benefits of macroalgae by paying farmers for the positive externalities created by kelp farms based on how much kelp they grow. Currently farmers send in photographs of their kelp growths at different times of the year. We have a dataset of kelp images taken on flat surfaces with a consistent background, but in practice pictures will come in from varying angles, with messy backgrounds, and on cameras with varying resolution. The goal of this project is to help create a model that can predict the mass of kelp grown based on images.

**Technology:**

- Python
- Deep learning
- Computer vision
- PyTorch
- Docker

# Invenergy

*Turbine Blade Object Detection*

**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.

Drones are utilized across Invenergy's wind fleet to inspect turbine blades for damage caused by wear and tear, lightning strikes, or other issues. The drone inspection results in ~200 images per turbine, which then takes significant time to review for damage. Invenergy has developed a damage detection model, which is able to pinpoint any damage on the image, but the model currently evaluates the entire image, which includes significant chunks of background (sky, ground, neighboring turbines, etc.). The process would be greatly improved if we could segment the blade from the background in the drone image, and only inspect the blade for damage.

**Mentor:**
Zoe Kimpel is the Data Science Engineering Manager at Invenergy and has 8 years of energy experience. She has an engineering degree from the University of Oklahoma and a data science master's degree from Northwestern University. She also was a 2019 Data Science for Social Good Fellow through the University of Chicago.

**Technology:**

- Python
- Deep learning
- Computer vision
- PyTorch
- Docker

# mBio

*Analysis of African Biotechnology Networks*

**Background:**
The 11th Hour mBio Project is a collaboration between the University of San Francisco, the University of Cambridge, and UChicago. mBio works at the nexus of energy, food & agriculture, and human rights. For us, growing beyond a linear economy based on extraction and waste towards a regenerative economy means advancing frameworks that value healthy ecosystems, active civic engagement, and social fairness. We do this by looking through a systems lens, finding new opportunities for a healthier, more reciprocal relationship with the resources of our natural world. The 11th Hour Project addresses issues impacting the health of our planet by providing grants to qualified 501(c)3 organizations that align with our unique strategic priorities.

Since at least 2000, a global network of private industries, development agencies and philanthropic donors have promoted biotechnologies on the African continent; now nearly a dozen countries have genetically modified organisms (GMOs) in some stage of research or commercialization. African civil society organizations (CSOs) have raised important critiques, noting that biotechnologies pose threats to African sovereignty and the environment. They have called for greater transparency from the institutions ushering in these new technologies, as well as monitoring of the global flow of finance sponsoring this work.

African CSOs lack the resources of biotechnology promotion networks and importantly also do not have access to data and analyses that could inform and enhance their advocacy work. mBio will work to build public datasets and analytical tools which empower civic engagement and deepen social impact on a global scale. Previous work has developed and performed analysis on three dataset: media data (2 million articles published about biotechnology in Africa since 1997), financial data (IRS 990 forms and other financial disclosures), and crop data (a database of GM crops that are or have been under development in Africa). Current results are available on mbioproject.org. A pipeline has been created that extracts quotes and their speakers from articles, analyzes the sentiment of quotations, and extracts networks from this information. We will be expanding on this pipeline and analyzing results.

**Technology:**

- web scraping
- Graph theory
- PyTorch
- pandas
- NLP
- SpaCy

# Morningstar, Inc.

*NLG for 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.

The Morningstar Medalist Rating unites two forward-looking rating systems – the Morningstar Analyst Rating and the Morningstar Quantitative Rating – into one. The combining of our quantitative and qualitative research will make it even simpler for investors to research and select best-in-class managed investments.

To compliment the Medalist Rating, Morningstar provides analyst-like auto generated text for funds. We produce 200,000 of these text-based reports for managed products monthly. The algorithm used to generate the text is producing narratives which sound clunky and like a computer wrote them.

For this project, we'd like someone to leverage human-in-the-loop AI to create and train a model that generates reports with Morningstar's style of writing, given analyst written text and accompanying Ratings Notes.

**Mentor:**
JoshCharney is a Quant Research Manager and has been with Morningstar for 12 years. He holds a CFA and a Master's in Computer Science from UChicago. Tom White Law is a Global Director for Equity Ratings and has been with Morningstar for 15 years. He is based out of the United Kingdom. He received his degree in business from Sheffield Hallam University. Lidia Breen is an Associate Product Manager and has been with Morningstar for almost 2 years. She received a Master's in Engineering from Lehigh University.

**Technology:**

- Python
- Pandas/numpy
- NLP

# Pediatric Cancer Data Commons (PCDC)

*Cancer Knowledge Graph*

**Background:**
The Pediatric Cancer Data Commons is an international leader in data harmonization and democratization. Headquartered at the University of Chicago, we work with groups around the world to expand access to pediatric cancer data for oncologists and researchers.

Cancer research, like many other biomedical domains, suffers from what has been called a "sea of standards". Several public data models exist to bind concepts to standardized definitions and codes—yet there is much overlap across these models. The PCDC is leading an effort to link knowledge across these models and create a searchable graph. This will enhance these resources and facilitate proper use.

Common cancer models will be converted from their native format into a common format. Shared concepts will then be linked using a mixture of automated and manual methods. These linkages will be searchable and will constitute the cancer knowledge graph. A similar undertaking that may provide context is the Mondo Disease Ontology which links together disease definitions from multiple sources.

The anticipated deliverables include six LinkML schemas (one for each cancer model) that are linked through internal LinkML mappings and a collection of SSSOM mappings. The work will be documented in an academic publication and the resulting knowledge graph will be widely disseminated throughout the scientific community.

**Mentor:**
Michael Watkins is the Data Standards team and academic research lead for the PCDC. He holds a Ph.D. in Biomedical Informatics from the University of Utah with a research emphasis in translational informatics, clinical genomics, and clinical data interoperability. Responsibilities include data modeling for partner disease consortia, harmonizing clinical data standards used throughout PCDC data models, and participating in several projects for research grants and contracts awarded to the PCDC.

**Technology:**

- Python

# Perpetual

*Foodware Flow Model*

**Background:**
Perpetual partners at the city level to design and implement immersive reuse systems, starting with foodware. As a non-profit, Perpetual establishes public-private partnerships in order to ensure that the reuse system benefits everyone, at a high cost to no one. Perpetual is currently working with four US cities to develop reuse systems that are viable from an economic, environmental, technical and social standpoint.

By leveraging a number of different data sources, Perpetual is trying to build a "Foodware Flow Model" to understand how to build and implement an immersive foodware reuse system. Specifically they want to build a model of how much silverware, plates, cups, etc. are used by customers within a geographic area. Using this information they will also identify locations where foodware users (think people who buy a cup of coffee) would be likely to return a foodware container so that it can be washed and then reused. Perpetual has already received multiple grants in order to investigate and design such systems and as a clinic student you will have a large impact on their business model. This project will scrape multiple data sources as well as build maps and implement algorithms to optimize the flow of foodware within a city.

**Mentor:**
Andy Rose is an experienced circular economy professional focused on shifting business to an ecologically & economically viable future. Andy is a mechanical engineer by education and started his career in software designing and implementing administrative software for financial institutions before pivoting his career to focus on sustainability. He has helped launch two reusable packaging companies and has held roles across program development, operations, reverse logistics, packaging design, strategy & brand management. At Loop, he onboarded the initial brand partners to launch the platform and then went on to manage the circular supply chain for North America. With Good Goods, Andy launched a reusable wine bottle program in NYC and consulted large wineries on their reuse strategy.

**Technology:**

- Python

# Prudential Financial

*Predict the earnings of publicly traded companies/PGIM Real Time Dashboard*

**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 opportunity for more people. Prudential's iconic Rock symbol has stood for strength, stability, expertise and innovation for more than a century. For more information, please visit news.prudential.com.

**Project 1:**
In this project students will develop an ecosystem of data and models used to predict corporate earnings line items. Leveraging public resources, students will develop a data set consisting of SEC filings, news releases, stock prices, dividends, macro-economic data (interest rates, inflation, employment) and derived data sets (sentiment) and one or more predictive models that will predict corporate earnings.

**Project 2:**
PGIM is the investment arm of Prudential and is interested in leveraging their infrastructure to build a number of near-real-time dashboards. The goal would be to leverage internal and external data sources in order to provide insight into PGIM's real estate business. Specifically, this would entail understanding PGIM's business, what effects it and investigating data sources that provide value to that core business. Examples might be using google trends style data to understand how rents are changing in different markets or building a sentiment analysis tool on news feeds.

**Technology:**

- Python
- Data visualization
- scikit-learn

# Remora Fishing Traceability

*Deep Learning for Computer Vision: Fishing Vessel Detection*

**Background:**
Remora Fishing Traceability is a Costa Rica based startup focused on ensuring the "traceability of seafood, from the sea to the plate." It recently approached the UChicago DSI with the idea of using computer vision to generate baseline estimates of the number of small-scale and industrial fishing boats operating off the coastlines of various countries.

Countries belonging to SICA (Central American Integration System), like Costa Rica, are required to report their fishery size while considering (1) the total number of fishermen, (2) the total number of boats, and (3) the total volume of catch. However, official estimates are often inaccurate. The Costa Rican national government, for example, reports that approximately 2,000 fishing boats exist based on the number of licenses, but informal estimates peg that number closer to 15,000 instead. Generating estimates would spur governments towards more accurate reporting and help them better allocate resources for fishery conservation.

The goal of this clinic is to use deep learning models to segment and count the number of boats found in high-resolution satellite imagery of 100+ designated fishing sites in Costa Rica.

**Technology:**

- Deep learning
- Image segmentation
- Object detection
- PyTorch
- Docker

# University of Chicago – Center for RISC

*Electronic Monitoring – Device Shielding Prediction*

**Background:**
We are an innovation lab for social change. Driven by curiosity, unfettered by orthodoxy, and grounded in the sciences of human behavior, we're investigating bold new ways to tackle the world's biggest problems. See more here: https://risc.uchicago.edu/.

Electronic Monitoring can be an effective, safe, and humane means of decarceration if management software can appropriately identify serious infractions, allowing for low-touch, passive monitoring of participants, the de-prioritisation of technical violations, and a program where people are not kept under effective house arrest but have free movement to go to work, school, and reintegrate into their communities.

Device shielding – or the deliberate obstruction of devices with foil to prevent their location from being tracked may be a significant source of risk. Some participants are able to successfully do so and have been allegedly involved in criminal behavior while their signal was lost. Many other participants simply lose their signal due to environmental factors — living in a built-up environment, basement, or the like. RISC has an algorithm that attempts to differentiate between the two, but it remains largely untested with little ground truth to rely on — few participants have actually been confirmed as shielding their device.

We propose conducting extensive field tests of EM devices to build sufficient ground truth on deliberate shielding and environmental signal loss, and building predictive models to tell the two apart.

**Mentor:**
Ben Thevathasan  is Lead Data Scientist at the Center for RISC, where he leads or advises RISC's stable of data projects in domains like healthcare, criminal justice, and animal welfare. He is technical lead for an ongoing criminal justice reform partnership, where RISC designs and deploys algorithms and behavioral tools to reduce the incidence of spurious law enforcement encounters with released detainees.

**Technology:**

- Python
- pandas
- scikit-learn
- Docker

# University of Chicago – Neurocritical Care

*National Trauma Database Analysis - Penetrating Brain Injury*

**Background:**

The Neurocritical Care section at the University of Chicago in an intensive care unit that caters to patients who suffer severe neurological or neurosurgical injury. Such Injury includes severe Traumatic Brain Injury(TBI), Gunshot wounds to the head, Intracranial hemorrhages, Large strokes (malignant stroke), and status epilepticus amongst other conditions. The Neuro-ICU offers a primary service as well as a consulting service for other ICUs that may house patients whose injuries include an injury to the brain. It is staffed by 4 board certified neuro-intensive care physicians.

The National Trauma Data Bank® (NTDB®) is the largest aggregation of U.S. trauma registry data ever assembled. We have access to the registry's data between the years of 2010 and 2019. This includes hundreds of thousands of patient encounters in the context of trauma. We plan to extract data relevant to severe traumatic brain injury and explore variables relevant to outcomes following severe traumatic brain injury. The goal is to isolate patients with penetrating brain injury and describe variables related to survival particularly within the cohort with undifferentiated GCS. We would also like to describe early parameters associated with survival (blood products, blood pressure) and potentially develop a survival model that can then be validated on our local data set.

**Mentor:**

Ali Mansour, MD, is a neurologist specializing in neurocritical care. Dr. Mansour has a background in signal analysis, advanced neuroimaging (fMRI and DTI) as well as bio-informatics. Currently, his research emphasizes the management and prognosis following penetrating brain injury (gunshot wounds to the head). He is also evaluating the role of neuroimaging in prognosis following neurocritical illness and cardiac arrest. Dr. Mansour is also interested in neuroinformatics; he and a multidisciplinary team of experts aim to optimize data capture and analysis in neurological and neurocritical illness to improve patient outcomes.

**Technology:**

- Python
- pandas
- SQLite
- scikit-learn
- Docker

## University of Chicago – Internet Equity Initiative

*National Urban Digital Divide*

**Background:**
The Internet Equity Initiative aims to realize equitable, resilient, and sustainable Internet solutions that benefit all communities. As society increasingly relies on the Internet for work, education, health care, recreation, and many other aspects of daily life, the prevalent and persistent inequity in people's ability to access, adopt, and use the Internet is more evident than ever. In the wake of the COVID-19 pandemic, these inequities have become apparent at the global, national, municipal, and neighborhood scales. The IEI has three goals: Developing measurement techniques and datasets that directly address unknown questions and evaluate the effectiveness of different interventions; creating data-driven collaborations with communities that are underserved by current Internet infrastructure to develop and test different options for infrastructure investments, the effectiveness of which can depend critically on the specific characteristics and needs found in different communities; and producing better data and analysis about how Internet connectivity relates to the social and individual conditions that contribute to whether and how the Internet actually improves people's lived experience.

While disparities in broadband access have received increasing national attention for years, pandemic-induced remote work/school and massive federal broadband investment make questions of internet access particularly salient today. Understanding the digital divide is the first step toward its mitigation, enabling the government and policymakers to effectively target the limited resources to the least connected areas. In spring 2022, the DSI Data Clinic provided an analysis of the digital divide in Chicago, looking at differences in Internet connectivity rates by neighborhood, and seeing how those rate differences correlated with socioeconomic characteristics of neighborhoods. This project builds on that analysis (including its existing code base) to perform the same analysis for cities across the country.

**Technology:**

- Python
- pandas
- geospatial analysis
- visualization
- Docker