THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Spring 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.

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

Blue Ocean Gear

Buoy Anomaly Detection

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").

Over the past year, students have conducted exploratory data analysis to describe Smart Buoys' typical motion patterns in different fisheries. They have also trained deep learning models (LSTMs and transformers) on both real and synthetic drifter data to predict the motion paths of drifting Smart Buoys.

Students this quarter will complete the motion analysis by detecting anomalies using a variety of techniques, including: physics-based heuristics, clustering algorithms (k-means, DBSCAN, and Isolation Forest), and deep learning (autoencoders and computer vision using CNNs). They will also determine if anomalies can be identified with respect to the positions predicted by the location models.

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
- Computer vision
- Docker
- Pytorch

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

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Hawai'i Alliance for Progressive Action

Studying Pesticide Use in Hawai'i

Background:
The Pesticide Action Network (PAN) North America is one of five regional centers worldwide. PAN works to link local and international consumer, labor, health, environment and agriculture groups into an international citizens' action network. PAN works to challenge the global proliferation of pesticides, defend basic rights to health and environmental quality, and ensure the transition to a just and viable food system.

Hawai‘i Alliance for Progressive Action (HAPA) is dedicated to fighting for social, economic, and environmental justice in Hawai‘i. Whether protecting communities from toxic pesticide drift or advocating for a living wage, HAPA strives to improve the unfair conditions forced upon locals and the ‘āina through community organizing, advocacy, and education.

Throughout the thousands of years of human agriculture, pesticides have only come to widespread use in the past sixty years. Since their wide adoption, many important critiques and risks have been identified. Pesticides are a major health risk and a threat to the environment at large. The focus of this project will be on pesticide use in Hawai'i. Starting in 2013, HAPA and partner organizations successfully pushed for the passage of a new law regulating pesticides in Hawai'i. As a result, chemical companies must report on the Restricted Use Pesticides (RUPs) they use. This project will work with the data from the first round of reporting along with other publicly available data to analyze the effects of pesticides in Hawai'i. Visualizations will be produced to convey results with frontline organizations and communities.

Technology:

- Exploratory Data Analysis (EDA)
- Web scraping
- pandas, plotly
- data visualization

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

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

Perpetual

Foodware Flow Model

Background:
Perpetual partners with cities 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 requires modeling the locations for optimal collection, distribution and washing points for reusable foodware packages in a city and estimating the economic feasibility of the entire operation. It will use multiple datasets on Food-ware Using Establishments (FUEs), residential density, foot traffic and behavioral patterns. Clinic students will be performing geospatial analysis and spatial clustering to identify optimal distribution, washing and pickup points. Finally, students will also implement routing 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

- Geospatial Analysis (Geopandas/PySAL/ArcGIS)
- Clustering
- Route optimization

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Prudential Financial

PGIM Real Estate Market Prediction

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. 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.

PGIM is the investment arm of Prudential and is interested in leveraging their infrastructure to predict where apartment demand is heading in different metro areas. The goal of this project is to leverage internal and external data sources to provide directional information regarding the residential rental market in different metro areas. The project will begin by building simple models of apartment demand and then increase the complexity of these models to increase their accuracy. The information from these models will be summarized in dashboards that PGIM will leverage when making real estate holding decisions.

Mentor:
Mentors for this project will include multiple members from the PGIM Innovation Team including Dave Power, Director of Innovation and Dean Deonaldo, AVP of PGIM Real Estate

Technology:

- Python
- Data visualization
- scikit-learn

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

University of Chicago – Center for the Science of Early Trajectories (SET)

Retrospective Neonatal EEG Analysis Protocols

Background:
The Center for the Science of Early Trajectories (SET) founded a diverse network of basic scientists, physician scientists, and clinicians in order to transform infant development research. SET's ground-breaking collaboration will establish a biologic map of the development of infants at the cellular and molecular levels, thereby defining the optimal trajectory for infants. This new discipline of research will translate into nuanced, patient-centered care for our most vulnerable babies in the community we serve, improving their health outcomes for their entire lives.

Premature infants (infants born before 37 weeks gestational age) are at risk of several health problems when compared to infants born full-term. These risks include infections, necrotizing enterocolitis (NEC), and seizures. Unfortunately, clinicians do not have a way to predict which patients are most likely to have a seizure, limiting our ability to provide the best care. Our hope is that by conducting machine-learning analysis on the MRI images, we will be able to better predict when an infant is at highest risk of experiencing a seizure.

Mentor:
Henry David, MD, is a pediatric neurologist who specializes in neurocritical care and treats children of all ages with serious neurological conditions, spanning from acute care to long-term treatment. Dr. David is also an expert in cerebrovascular disorders, including ischemic strokes and hemorrhagic strokes, and he provides comprehensive, compassionate care to his patients and their families.

Technology:

- Python
- Machine Learning
- Docker

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

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