THE UNIVERSITY OF CHICAGO
DATA SCIENCE
INSTITUTE

Autumn 2022 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.

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

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

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

Business and Professional People for the Public Interest (BPI)

Racial Disparities in 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. The 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. This project will review data from Illinois and Chicago public agencies to investigate racial disparities in how traffic stops are conducted by the Chicago Police Department. This regression-adjusted benchmark analysis is needed, because while the raw numbers of traffic stops made on Black and Latinx drivers are significant, the more complex analysis will help us further understand what the data is telling us about racial bias in traffic stops by factoring in the impact of elements like residential segregation and police deployment.

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 which policy solutions we move forward with as well as our broader public education campaign about traffic stops and our ability to effectively respond to opposition.

Mentor:
Loren Jones joined BPI's Criminal Legal Systems/Police Accountability Team in 2021 as Staff Counsel. Her work includes a focus on community oversight of policing and reimagining effective strategies for creating authentic public safety for all communities. She also provides legal, policy, and technical advice to amplify advocacy efforts by organizations and coalitions. Prior to joining BPI, Loren worked at Dvorak Law Offices, LLC where she represented victims of police misconduct. Loren received her JD cum laude from the University of Illinois College of Law in 2018.

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
- R
- geopandas
- Plotly/Dash

Blue Ocean Gear

Fishing Gear Anomaly Dashboard

Background:
Blue Ocean Gear, incorporated in California in 2019, has developed a Farallon Smart Buoy System™ that tracks fishing gear in the open ocean for commercial fleets. 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, such as those from the National Oceanic and Atmospheric Administration (NOAA) and Fisheries and Oceans Canada (DFO), to better predict when anomalies occur.

In previous quarters, students collected buoy, weather station, surface current, and surface temperature data from BOG, NOAA, DFO, and NASA and then merged those records together based on distance and timestamp. Using this combined dataset, they developed Random Forest regression models for predicting buoys' location, depth, and battery temperature, as well as DBSCAN and Isolation Forest models to detect anomalous motion. Finally, they created a dashboard to visualize the results of anomaly detection and predictions on a map and through auto-generated reports.

This quarter, Blue Ocean Gear hopes to conduct a data analysis of the 150 buoys based in New Brunswick, which are currently configured to report data at a high temporal frequency. Students will improve the current ML models and experiment with new ones to (1) describe the typical motion patterns of buoys in different fisheries, (2) compare buoys' recorded temperatures with third-party readings, (3) flag anomalous motion and temperatures, and (4) predict buoy location and battery temperature, including during breakaway events. Final deliverables for the project will consist of a series of reports in the form of interactive Jupyter notebooks, as well as updates to the data pipeline, API, and dashboard.

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
- Time-series analysis
- Geospatial analysis
- React
- Docker
- Django

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.

A previous team made a first pass at a pipeline that fetches 8K forms from the S.E.C., extracts and cleans relevant form sections, and then uses pretrained Question Answering (QA) and Named Entity Recognition (NER) models to locate loan issuance dates, amounts borrowed, debtor and lender names, etc. This quarter, we will improve and productionalize the pipeline, ultimately creating a website that allows users to search and download the collected loan data.

Mentor:

Technology:

- Python
- NLP
- PostgreSQL
- Django
- React
- 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). A pipeline has been created that extracts quotes and their speakers from articles, analyses the sentiment of quotations, and extracts citation networks from this information. We will be expanding on this pipeline and analyzing results.

Mentor:

Technology:

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

Wallace Center

Predicting Regenerative Farm Conversion Likelihood

Background:
The Resilient Agriculture and Ecosystems team at the Wallace Center promotes regenerative agriculture and land management practices in order to achieve “balanced environmental, economic, and social solutions needed for our current challenges”. They have worked on perennializing agriculture and promoting regenerative grazing in order to improve water quality, but have started to intersect with other regenerative work. Previously, they did an analysis of grazing potential in Illinois and Indiana in order to identify the lands best suited for outreach and promotion of regenerative grazing. This project intends to conduct a similar analysis, allowing us to predict which crops and regenerative practices will be best suited for which regions. This will allow organizations to focus their outreach efforts and advise farmers accordingly. The goal is not to create a map that leads farmers to think “growing X won’t work for me because the map says it is less suitable”, but to create a map that helps inform outreach. UW-Madison has done similar research on grazing.

The Mississippi River/Gulf of Mexico Hypoxia Task Force of the United States EPA aims to reduce hypoxia in water bodies that feed into the Mississippi River and Gulf of Mexico. Hypoxia is the condition of water having too little oxygen and can cause major disruptions to aquatic ecosystems. This can be caused by farms through fertilizer runoff and soil nutrients eroding into water bodies. Because of this, midwestern states have been required to release semi-annual Nutrient Loss Reduction Strategy reports including information on the uptake of certain agricultural practices within their state.

Previous work for this project has compiled a list of relevant data sources on farm income, demographics, sizes, and types, water and soil quality, and bird and pollinator habitats. The goal of this project is to combine these datasets to evaluate the characteristics that lead a particular geographic area to become more likely to use certain regenerative practices.

Technology:

- Machine learning
- scikit-learn
- pandas
- React
- Django

Prudential Financial

Predict the Earnings of Publicly Traded Companies

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.

Predicting corporate earnings is the "Holy Grail" of investment research and management. The advent of advanced modeling techniques and computing power creates new opportunities for individuals and companies to solve this decades-old problem.

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.

The project will begin with a literature review and interviews with internal experts at Prudential. Students will align on one or more approaches to predicting corporate earnings and assemble the data required for predictive models. Students will perform a comprehensive data analysis and develop models for predicting line items from a company's income statement.

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.

Technology:

- Machine learning
- scikit-learn
- pandas
- React
- Django

University of Chicago – Neurocritical Care

National Trauma Database Analysis

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.

In addition to clinical work, the group runs an elaborate research operation in areas of computer vision, outcome modeling, brain death, intracranial pressure time series prediction, decision making under uncertainty and medical/neuro ethics.

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 ultimately train and validate a predictive model that can assist in selecting candidates for surgical management of severe TBI vs medical management.

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