THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Spring 2025 Data Science Clinic

Clinic Overview

The Data Science Clinic is a project-based course where students work in teams as data scientists with real-world clients under the supervision of instructors. Students are tasked with producing deliverables such as data analysis, research, and software along with client presentations and reports. Through the clinic course, Affiliate members gain access to undergraduate or graduate student teams to work on data science projects and explore proof of concepts while identifying top student talent. Projects are tailored and scoped to address company objectives with all deliverables overseen by the Clinic Director.

These unique collaborations allow Affiliate members to supplement their internal data science teams with outside support and perspectives, enlarging their capacity to experiment with new ideas. They also give students a window into a data science career, learning how companies build and use these tools internally.

Clinic Structure

Data Science Clinic runs during Fall, Winter and Spring quarters. Clinic projects are generally scoped to run for two full quarters. Each student works between 10 to 15 hours a week. Each team has a weekly 1-hour meeting with their assigned mentor and must submit a weekly progress report. Mentors are drawn from research staff, postdoctoral fellows and the faculty, subject to availability, interest and needs of the project. The mentor provides intellectual guidance, direct feedback to students and serves as a sounding board for both challenges and direction. The mentors will also provide support and guidance on any gaps in data science knowledge by providing literature and resources. Regular meetings are scheduled as it suits the client needs and to provide feedback to students.

What does the ⚙ mean?

If you look at the project descriptions below you will see that many of them have a gear/cog icon. These projects require a deeper knowledge of computing and preference will be given to those students who have demonstrated that capability.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Project List

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Argonne ⚙

Extending Argo with LLM-driven Agents and Workflows

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

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

Argonne is partnering with the DSI to enhance its internal LLM system, Argo. The system serves dual purposes: as a general LLM chatbot and as a specialized tool for answering questions about Argonne's internal policies. In previous quarters, Argo was developed to include basic agent functionality with tool usage capabilities and a Retrieval Augmented Generation (RAG) system, enabling it to answer questions using internal Argonne documents. This quarter, we aim to expand its features by implementing recent LLM techniques, focusing on:

- Enhanced tool usage capabilities
- Development of a multi-agent framework
- Implementation of LLM benchmarks to guide feature development

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

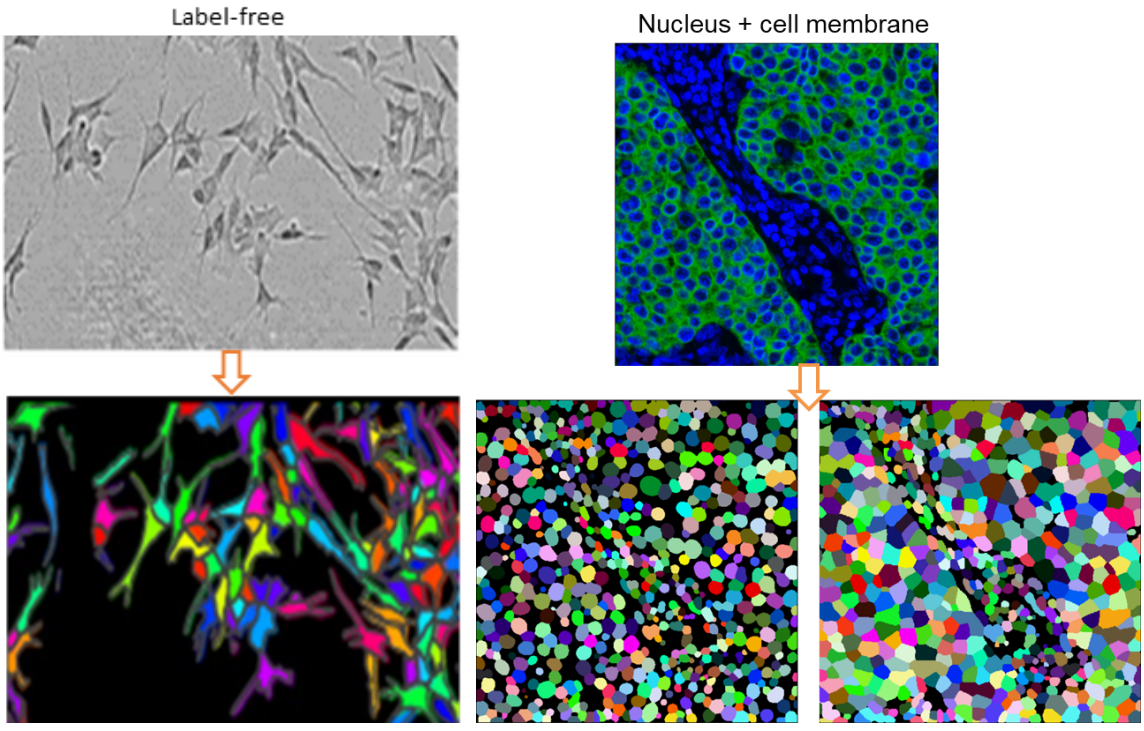
Center for Living Systems ⚙

AI and Cell Segmentation

Background:

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

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



In this project, we would like to explore the applicability of these two software tools and study the performance over various cell types and image modalities. The purpose of this project is to get these two algorithms running on our internal data science cluster and then compare the results using cell segmentation evaluation metrics against ground truth

data. Understanding the relative performance of these algorithms is crucial for determining which tool is best suited for specific conditions.

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

Mentor

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Chicago Metropolitan Agency for Planning (CMAP) ⚙

NE Illinois Stormwater Storage and Site-Scale Green Infrastructure Inventory

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

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

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Climate Cabinet

Campaign Finance Networks

Background:

A key part of advancing important climate related policy is getting legislators elected and a key part of getting legislators elected is money. Climate Cabinet hopes to understand the relationship between campaign finance and climate legislation by identifying where candidates are receiving funds and which donors have relationships to fossil fuel or clean energy groups.

The DSI is helping Climate Cabinet develop an open source pipeline for collecting and analyzing state campaign finance data. This quarter, we wish to understand:

- the relationship between campaign donations and electoral results
- the relationship between campaign donations and legislators' climate scorecards
- clustering patterns in campaign finance networks

Mentor:

Trevor Spreadbury is a Software Engineer II at the DSI. He helps social impact organizations to enhance their operations, research, and communication by utilizing software engineering and data science tools. His work focuses on agriculture, human rights, energy, and marine technology. Before DSI, Trevor worked as a research assistant at Argonne National Laboratory. Trevor has a BS in Computer Science from MIT.

College Financial Health Data Repository

Background:
The higher education landscape is experiencing unprecedented financial pressures, making it crucial for prospective students to understand the financial health of institutions they're considering. While comprehensive financial data is available through the Integrated Postsecondary Education Data System (IPEDS), this information isn't easily accessible or interpretable for most students and families making college decisions.

This project aims to create a modern, accessible platform for understanding college financial stability, building upon methodologies similar to those used by TuitionTracker.org (https://tuitiontracker.org/fitness/methodology.html and https://tuitiontracker.org/fitness) but with updated data and improved visualization techniques. Our primary objective is to develop a comprehensive database using the IPEDS API to collect ten years of financial and operational data for all U.S. colleges and universities. Key metrics we'll track include:

- Enrollment trends
- Average tuition revenue per student
- Student demographic breakdowns (undergraduate, graduate, international)
- Endowment levels
- Operating margins
- Faculty and staff expenses
- Institutional aid levels

Using this dataset, we will develop a risk assessment framework to identify institutions that may be facing financial challenges. The project's culmination will be an interactive Streamlit dashboard that presents this complex financial data in an accessible format for prospective students and their families. This tool will help users better understand the financial stability of institutions they're considering, ultimately contributing to more informed college selection decisions.

The final product will serve both as a practical tool for student decision-making and as a valuable resource for understanding broader trends in institutional financial health across American higher education.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Data & Democracy

Historical Election Datahub

Background:
America's electoral history (pre-1998) is currently trapped in thousands of PDF documents, limiting access to this crucial historical data. Building on last quarter's successful development of data extraction systems, this project now focuses on creating a comprehensive validation framework and public-facing website for U.S. historical election results.

The immediate priority is developing robust error-checking algorithms to verify the accuracy of our extracted data. We need to implement automated checks for vote total consistency, identify anomalous results that require human verification, and cross-reference results across multiple source documents where available. These validation systems are crucial before we make the data publicly available.

Once our quality control systems are in place, we'll build an interactive web platform to host and visualize the validated election data. The platform will feature tools to explore the data through novel visualizations. This resource will help deepen our understanding of American democratic processes while ensuring the highest standards of data accuracy.

Mentor:

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Data Science Institute

Agents on a Budget

Background:
Over the last few years there has been much hype around the use of AI agents and the DSI is interested in evaluating the technology for internal use. The goal of this project is to build a proof-of-concept agentic chatbot capable of answering complex queries about the DSI. Students will be responsible for developing a cutting-edge AI agent capable of navigating complex environments, accessing a variety of information sources, and making informed decisions. Particular emphasis will be placed on developing an agent that is:

- User friendly
- Free (uses only open source tools and consumer hardware)
- Accurate

Mentor:
TBD

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Data Science Institute

HPC Cluster Performance: Time Series Analysis and Visualization

Background:
The Data Science Institute operates a high-performance computing (HPC) cluster to support the computational needs of students and researchers. Efficient resource allocation and performance monitoring are critical to maximizing the cluster's effectiveness. This project focuses on visualizing and predicting HPC usage over different time scales—daily, weekly, and quarterly—using time series analysis.

By collecting and analyzing CPU, GPU, memory, and storage usage data, this project aims to:

- Visualize trends in resource utilization using interactive dashboards.
- Identify peak demand periods and patterns of underutilization.
- Predict future usage using machine learning models for better scheduling and resource allocation.

The insights from this analysis will help optimize job scheduling, improve user experience, and guide future infrastructure planning.

Mentor:
TBD

Food System 6

Visualizing the Economic Infrastructure of the US Poultry Industry

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

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

Over the last year Food System 6 ("FS6") has worked to understand the scope and magnitude of the financial ecosystem that underpins the US poultry industry and are looking to leverage this knowledge into an interactive simulation game to highlight the financial asymmetries between conventional poultry production and emerging poultry production systems that embrace regenerative principles, e.g., pastured poultry. This game aims to provide greater insight and education to the growing segment of stakeholders, funders, investors, and actors in the food system reform space.

We want the simulation to convey the economic mechanisms and cultural capital norms spanning from "land to plate" highlighting the benefits received by conventional poultry growers at each step of the process. We have identified about 20 different financial mechanisms, offerings, and programs. By integrating these elements into the game, we aim to educate players on how they financially impact farmers in both a conventional and regenerative context. While many poultry producers may be able to access federal insurance products, for example, regenerative practices may automatically preclude a poultry grower from accessing those programs.

Our research has also identified various "capital cultural norms" that drive the practices and performance of conventional poultry. For example, poultry integrators use lopsided

contracts to pit poultry growers against one another, lead them into obtaining multi-million-dollar financing to construct poultry houses, and allow them to cancel the contract for a wide variety of reasons with no consideration or renumeration to the poultry grower.

Mentor:

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

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Groundwork Bridgeport - Urban Forest Equity

Computer Vision Analysis of Urban Forestry Over Time

Background:

Groundwork Bridgeport (Groundwork) is a community based 501(c)(3) non-profit organization with a mission to bring about the sustained regeneration, improvement, and management of the physical environment, by empowering people, businesses, and organizations to promote environmental, economic, and social well-being. Since its inception in 1998, Groundwork's main goal has been to convert blighted areas into gardens, parks, playgrounds, and other open spaces that instill pride in the community. To accomplish these goals, the organization develops education, community service, and other programming as a means of improving the quality of life for those it serves. Groundwork's programs primarily serve and impact populations that are marginalized and lacking readily accessible resources. Their work with the DSI Open Spatial Lab utilizes urban forestry data to understand the distribution of tree species and benefits to assess racial equity in tree distribution and other key metrics.

Demo data explorer: https://open-spatial-lab.github.io/groundwork-bridgeport/

This project aims to enhance this work with a machine learning / computer vision approach to mapping tree canopy. Existing data relies on observations of trees and their measurements, and so gaps in the data inevitably emerge. This project aims to:

1. Utilize and refine a model for detecting tree canopy from satellite imagery of Bridgeport, Connecticut
2. Use this model to examine shifts over time (2000, 2010, and 2020) and connect changes in tree canopies to observed census data
3. Hypothesize and test potential relationships between trends in tree cover and demographic trends
4. Package the workflows in a reusable for Groundwork organizations across the United States and globally

Mentor:

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Inclusive Development International (IDI)

Mapping Human Rights Violations in the Palm Oil Industry

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

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

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

This Data Clinic will expand PalmWatch by integrating grievances filed by community members and nonprofit watchdogs as another data source. Building off the work of previous Clinic teams, students this quarter will—

- Scrape websites and query APIs for palm oil supply chain actors (e.g., plantations, mills, suppliers, consumer goods manufacturers, retailers) and the relationships between them.
- Perform record linkage (i.e., entity resolution) to match scraped entities with those from other, authoritative supply chain datasets like the Universal Mill List.
- Leverage LLMs with few-shot prompting and Retrieval Augmented Generation (RAG) to identify complainants and respondents in the grievance text, as well as the relationships between them, while using the generated supply chain dataset as a source of truth.

- Analyze the extracted data to describe patterns in complaints across consumer brands, suppliers, plantations, etc. and visualize problematic chains with network graphs.

Through their work, students will increase the transparency of the number and type of complaints associated with companies from every stage in the palm oil supply chain—empowering PalmWatch's users to target those companies for reform.

Mentor:

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Internet Equity Initiative

Milwaukee Internet Equity and Device Testing

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

One of the research focuses at the IEI is understanding the quality of service received by home Internet users. This study involves asking volunteers to connect a device, at their house, which measures Internet performance. In the past, IEI produced a study with the results of this data for a Chicago deployment. In the previous quarter, the clinic focused on building a notebook to enable automatically replicating the analysis for other deployments and testing this notebook on a deployment in Milwaukee, WI. One of the limitations of the previous work is that it focuses on analyzing deployments (i.e., sets of devices) and little has been done to study measurement results obtained for individual devices. The purpose of this quarter's work is to introduce additional visualizations and tests to enable analyzing and drawing insights from individual devices, which also helps explain deployment-level results. Specifically, we want to do the following:

- Identify a set of visualizations and tests that can help characterize Internet performance experienced by individual devices in the study.
- Build an interactive dashboard that integrates these visualizations and tests to enable the exploration of measurement results for individual devices.

Mentor:
Jonatas Marques is a Postdoctoral Fellow at the University of Chicago working at the Internet Equity Initiative. He holds a PhD in Computer Science from UFRGS (Brazil) and has more than ten years of experience doing research in and around computer networking and Internet measurement.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Kids First Chicago

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

Background:
Kids First Chicago's (K1C) mission is to dramatically improve education for Chicago's children by ensuring high-quality public education is accessible to all families. One of the pillars supporting K1C's mission is data stewardship; we strive to improve data transparency and accessibility to empower parents and families to make informed decisions and take action in their children's education. In the spirit of this, we have begun a project that will centralize roughly 30 years of publicly available education outcomes and enrollment data, sourced from Chicago Public Schools (CPS) and the Illinois State Board of Education (ISBE).

The end-goal is to create a public website where any person would be able to download filtered datasets and visualize data related to their information of interest. Given that CPS education data is currently spread across multiple datasets, websites, and formats, it is our hope that this centralized database will improve general accessibility, allowing families, interested public figures, and researchers, alike, to be able to efficiently empower themselves with harmonized data and further progress education goals. We expect that these data will help in the investigation of several pressing research and policy interests, including the relationship between chronic absenteeism and education outcomes, and the impact of parent involvement on education outcomes.

Building on the previous work completed in the clinic, students this quarter will develop and deploy a production-grade application to host the CPS and ISBE data.

Mentor:
Susan Paykin is the Senior Associate Director, Community-Centered Data Science, at the DSI, where she oversees social impact and strategic partnership initiatives across the organization's research, education, and engagement. She is also the Program Lead of the Open Spatial Lab where she leads geospatial data science projects and partner engagement. Susan was previously the Research Manager at the Center for Spatial Data Science at UChicago and has served in leadership roles for environmental and social impact organizations. She holds a Master in Public Policy (M.P.P) from the Harris School of Public Policy at University of Chicago and a B.A. from Brandeis University.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Pesticide Action Network - California People and Pesticide Explorer

Mapping harmful pesticide use in vulnerable communities

Background:

The California People and Pesticide Explorer is a tool developed by the DSI Open Spatial Lab and 11th Hour teams working with Pesticide Action Network (PAN) and Californians for Pesticide Reform (CPR). The tool makes the highly granular Pesticide Use Regulation (PUR) data more easily accessible and viewable in a dynamic interactive web application. This tool builds on existing workflows by enabling live filter, aggregation, and data downloads, along with incorporating key demographic data to help understand the social context of who gets exposed to pesticide use. CPR and PAN use this tool and others to help identify problematic usage of harmful chemicals - those with carcinogenic properties or posing reproductive risks. Demographic data and census spatial units (Tracts, School Districts, Zip Code Tabulation Areas) are important to help identify which communities might bear a disproportionate exposure to certain active ingredients, and which at risk communities (children, elderly) may need advocacy and support to reform pesticide use in their areas.

The current web application can be found here:
https://pesticideinfo.org/pesticide-maps/ca-pesticide-map

The existing API allows for easy querying of the data. This analysis will utilize the data to study specific active ingredient use in key communities:

1. Which communities have the most exposure to different active ingredients? Do any socio-economic factors suggest a disproportionate impact to BIPOC communities?
2. For certain harmful active ingredients (specifics TBD), which schools and school districts see the most use?
3. Over the past five years, have any potential trends and usage near schools emerged?

The current API allows for data querying as granular as ~ 1 mile x 1 mile “section” spatial units and can be disaggregated by month, active ingredient chemical, chemical class, application method, and a number of other available data filters.

Mentor:

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

RAFI – Grocery Atlas

MSA Analysis on Grocery Atlas

Background:

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

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

This phase of the project will focus on understanding the conditions over time that lead to a highly concentrated market. Data are available from 2000 to 2023.

We want to explore the following questions:

- How can we measure changes in the market landscape over time to better understand how areas become dominated by a single large company?
- What are some economic or demographic characteristics that areas/markets that have become highly denominated have in common?
- Can we identify areas that may be in the process of becoming highly concentrated to focus advocacy efforts and support small businesses?

This project involves some exploratory analysis to identify the right metrics to capture these complex changes, and then applying the method to generate actionable insights.

Mentor:

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Satellite-Based Detection of Ancient Water Systems⚙

Background:
Remote sensing and deep learning techniques are increasingly intersecting to unlock new possibilities in archaeological discovery. This project aims to develop a Python-based workflow that leverages modern deep learning architectures to identify ancient qanat water supply systems from satellite imagery.

Qanats - underground aqueduct systems that historically supplied water to arid regions - are often difficult to detect through traditional archaeological methods. However, these systems leave subtle surface signatures that can be visible in high-resolution satellite imagery. Our goal is to create a machine learning pipeline that can automatically detect and map these historical water management features across large geographic areas.

The project will involve processing satellite imagery, developing appropriate neural network architectures, and creating training datasets to identify the characteristic surface patterns of qanat systems. We'll build an end-to-end workflow that handles image preprocessing, model training, and prediction visualization. The resulting tool will help archaeologists and researchers efficiently survey large areas for these important historical water infrastructure systems, advancing our understanding of ancient water management practices and human settlement patterns.

This work combines modern computational techniques with archaeological research, demonstrating how AI can enhance our ability to study and preserve cultural heritage.

Mentor:
Dr. Mehrnoush Soroush and Dominik Lukas (https://camelab.uchicago.edu/team) who are both part of the center for ancient middle eastern landscape.

Society for the Protection of Underground Networks (SPUN) ⚙

Hyperspectral Image Analysis

Background:

The DSI is analyzing hyperspectral images taken by plane that may be able to monitor the health of underground fungal networks. This would broaden the impact of the direct soil samples taken by SPUN, the Society for the Protection of Underground Networks, to wide regions that can be covered by aerial photography. The key question is whether underground fungal networks, which have a profound influence on above-ground plant life, can be observed in precise changes of the spectrum (color) of light reflected from their leaves.

This quarter's work addresses that question using hyperspectral and LIDAR images taken by plane over forests in the U.S. from NEON, the National Ecological Observatory Network. If a significant correlation between SPUN ground-truth and NEON images can be established, this work can then progress to building predictive models to expand SPUN's coverage and/or focus the search for better measurement sites.

Mentor:

Jim Pivarski is a data scientist/engineer who has worked in and out of academia. He was trained as a particle physicist with a Ph.D. from Cornell and helped to commission and analyze first data from the CMS experiment at the Large Hadron Collider (LHC) in Geneva. He then worked as a data science consultant at Open Data Group, analyzing data from commercial clients, and then at Princeton, developing analysis software for physicists and promoting better integration with the world beyond academia. Jim is the original author of Awkward Array, a NumFOCUS affiliated project with thousands of users. Now he is analyzing data at U. Chicago's Data Science Institute for the 11th Hour Project and its philanthropic goals.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

University of Chicago Library

Optimizing Library Storage

Background:

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

The DSI is helping the library optimize which books to remove from the physical library collection. The data contains information about the collection, historical 'circulation' data dating back 10 years, as well as data about electronic items (e.g., ebooks), especially where they overlap with or duplicate physical items. We have been working with the library to develop a collection of features that rank the books in the collection to determine which books to move offsite.

The goals for the upcoming quarter are to develop a sense of model evaluation and build the current algorithm into a toolkit that can process different library collections. In particular we want to do the following:

- Analyze the weight sensitivity of the model
- Determine if our current pipeline is generalizable for recommendation and build up a toolkit for this process.

Mentor:

David Bottorff is the Collection Management & Circulation Services Librarian for the University of Chicago Library and is leading the Library's strategic priority of developing a comprehensive collection management and storage plan for its physical collections.

University of Chicago Transportation

UGo Shuttle Analysis

Background:

The UGo Shuttle program is a free shuttle service provided by the University of Chicago. Each UGo shuttle is equipped with a card reader, located just inside the door of the shuttle. All passengers are asked to tap their valid UChicago-issued ID on a card reader each time they board a UGo Shuttle. The card readers are used to collect information on the peak times and routes used by the various parts of the University community.

Working in collaboration with uchicagoshuttles.com, the DSI is helping UChicago Transportation analyze their rider data and make informed decisions about the routes, stops, and schedules of UGo shuttles. The purpose of this quarter's work is to better understand rider waiting patterns with respect to time-of-day and location-specific effects. Specifically, we want to do the following:

- Identify a set of research questions and the variables that can answer them in the shuttle data
- Develop a data pipeline to extract relevant information from the shuttle data
- Create visualizations or dashboards to explain your findings

Mentor:

Beth A. Tindel is the Director of Transportation & Parking Services at the University of Chicago. In this role, Beth oversees University-wide commuter, parking, traffic, and transit functions. She is responsible for parking functions in the University's parking structure and surface lots and she manages the University's agreement with the CTA, the daytime and NightRide shuttles, charter buses, and the University's various alternative transportation programs. Beth earned a Bachelor of Arts in Studio Arts from the University of Georgia.