THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Spring 2024 Data Science Clinic

Clinic Overview

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

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

Clinic Structure

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

Argonne National Laboratory

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

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

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

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

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

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

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

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

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

Argonne/Fermi National Laboratories

Lessons learned analysis

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

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

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

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

The Center for Investigative Reporting (CIR)

OSHA Workplace Injuries

Background:
The Center for Investigative Reporting (CIR) is the umbrella organization for Reveal, a weekly investigative radio show and podcast, CIR Studios, which makes full-length documentaries, and Mother Jones, a reader-supported investigative newsroom producing a print magazine and website. Our nonprofit newsroom goes deep on the biggest stories of the moment, from politics and criminal and racial justice to education, climate change, and food/agriculture.

CIR has obtained a vast repository of US workplace injury data obtained through FOIA requests aimed at the Occupational Safety and Health Administration (OSHA). This data is crucial for enhancing public understanding of workplace safety. The project aims to supplement a web platform that enables policy makers, journalists, and community members to explore the prevalence and context of workplace injuries within their locales.

To make this data accessible and informative, the project encompasses several critical tasks:

1. Industry Segmentation Algorithm: Development of an algorithm to categorize workplaces into industries accurately. The challenge lies in the evolving nature of OSHA's industry definitions, which may not always resonate with the public's perception. The goal is to create a classification system that is both accurate and intuitive for users.
2. Contextual Analysis Model: Understanding workplace injury rates demands a nuanced approach, considering various influencing factors. Although a preliminary model exists, it first, requires validation and, given sufficient time, possibly enhancement to ensure it effectively captures the complexities impacting injury rates, thus providing deeper insights.

The ultimate objective is to unveil a comprehensive web-based resource that sheds light on workplace injuries, fostering greater awareness and informed discussions among stakeholders about occupational safety. This resource will be publicly facing as well as intentionally distributed to other journalists who can use it to build local labor coverage in their communities.

Mentor:
Melissa Lewis is a data reporter for Reveal. Prior to joining Reveal, she was a data editor at The Oregonian, a data engineer at Simple and a data analyst at Periscopic. She is an organizer for the Portland chapter of the Asian American Journalists Association. Lewis is based in Oregon.

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.

Students will take on the challenge of filling this gap by applying deep learning techniques to aerial photographs to identify stormwater storage facilities. They will develop a model distinguishing between various green infrastructure types, such as wet ponds and constructed wetland detention basins. This project will expand our green infrastructure inventory and play a vital role in bolstering the region's climate resiliency by improving water management strategies in response to climate change.

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Chicago Trading Co

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

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

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

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

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Climate Cabinet

Campaign Finance Tracking

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

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

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

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

Fermi National Accelerator Laboratory (GNN)

Graph Neural Networks for Liquid Argon Time Projection Chambers

Background:
Fermilab is America's particle physics and accelerator laboratory. Host of several particle physics experiments and international collaborations aiming at solving the mysteries of matter, energy, space and time.

Neutrinos are the lightest matter particles in the Universe. They are electrically neutral and interact with other particles only via the weak nuclear force. This makes the study of neutrinos very challenging, as they interact very rarely, thus requiring large detectors and intense beams. This also means that some properties of neutrinos are not fully known yet. Indeed, neutrinos may play a key role in the dominance of matter over antimatter in the Universe and can potentially give insight into the nature of Dark Matter. Current and future experiments at Fermilab are tasked to measure neutrino's properties with unprecedented precision using Liquid Argon Time Projection Chamber (LArTPC) detectors. These are large volumes of liquid argon equipped with three readout planes, each made of wires that measure the ionization charge produced by charged particles traveling through the argon.

Fermilab partners with the University of Cincinnati and Northwestern University with the goal of developing a Graph Neural Network (GNN) for particle reconstruction in LArTPC neutrino experiments. Goal of our GNN is to provide precise inputs to the study neutrinos through the measurement of particles produced by the neutrino interaction in the detector. For instance, the identification of the neutrino type relies on this measurement. We have developed a message-passing GNN that is used to classify the nodes, defined as the charge measurements on the LArTPC wires, according to the type of the particle that produced them. Our network connects nodes both within and across measurement planes in the detector and achieves 94% accuracy with 97% consistency across planes.

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

Mentor:
Giuseppe Cerati received his Ph.D. in Physics and Astronomy at Università degli Studi di Milano – Bicocca in 2008. At Fermilab since 2016, currently working as Scientist. Working

on collider experiments such as CMS and neutrino experiments such as MicroBooNE, ICARUS, DUNE, with focus on physics analysis and data processing algorithms (both traditional and machine learning).

Fermi National Accelerator Laboratory (Simulations)

Improving ML-based simulations of particle physics experiments

Background:
Simulation plays a crucial role in analyzing the data from particle physics experiments. High quality physics-based simulations have a significant computational cost, which will become impractical as dataset sizes grow. An alternate approach is generative machine learning, in which we train a model that matches the quality of the physics-based simulator but produces samples much more quickly. We have an initial version of such a model, based on diffusion, a state-of-the-art approach in image generation. This model approximates the physics-based simulation with very high accuracy, but the generation speed is not yet fast enough for many use cases. This project will focus on applying a method to speed up generation, by first compressing the inputs to a lower-dimensional representation before running the diffusion process. Students will have the opportunity to learn about and employ multiple cutting-edge techniques in generative machine learning. There is opportunity for exploring other approaches to speed up or improve the model, depending on student interest.

Mentor:
Oz Amram and Kevin Pedro are physicists at Fermilab working as part of the CMS Collaboration which is attempting to understand the very basic laws of our universe. Oz received his PhD from Johns Hopkins while Kevin received his from UMD. Their research focuses on the use of AI and ML techniques in high energy physics. Oz is also a writer for ParticleBites, a reader's digest of high energy physics news and papers designed for readers with an undergraduate level of physics knowledge.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Internet Equity Initiative

Location Fabric analysis

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

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

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

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Invenergy

DSI - Icing

Background:
Invenergy is the world's leading privately held sustainable solutions provider. We develop, own and operate large-scale renewable and other clean energy generation and storage facilities worldwide. Our home office is in Chicago, and we have regional development offices in North America, Latin America, Asia and Europe. To date, we have successfully developed 191 projects totaling more than 30,200 megawatts.

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

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

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

Morningstar, Inc.

Morningstar Codebase and Data Structure LLM Optimization

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 goal of this project is to finetune a Language Model (LLM) specifically tailored to comprehend Morningstar's intricate codebase and data structure. By optimizing its understanding of Morningstar's proprietary systems, the LLM aims to streamline various aspects of development and analysis within the organization.

1. Enhanced Efficiency: The refined LLM will expedite workflows by swiftly identifying ownership of different components within the codebase, saving valuable time for developers.
2. Code Suggestions: It will provide intelligent suggestions for code modifications, leveraging its comprehensive understanding of Morningstar's coding conventions and best practices.
3. Knowledge Repository: Capable of answering queries and providing detailed methodological insights, the LLM will serve as a reliable repository of institutional knowledge.
4. Bug Detection: With its deep comprehension of the codebase, the LLM may detect potential bugs or issues, contributing to improved software quality and reliability.

By harnessing the power of advanced language modeling techniques, this project aims to empower Morningstar's development teams with a sophisticated toolset for codebase navigation, optimization, and quality assurance.

Mentor:
Josh Charney is a Quant Research Manager and has been with Morningstar for 12 years. He holds a CFA, an MBA, and a master's in computer science from U Chicago. David Wang is a Senior Principal Data Scientist and has been with Morningstar for 19 years. He holds a PhD in mathematics and computer science from the University of Illinois, Chicago. Brad Boemmel is a Director of Technology for the Analytics group at Morningstar and has been with Morningstar for 7 years. He holds a bachelor's in computer engineering from the University of Illinois, Urbana-Champaign.

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Perpetual

Reusable foodware system design

Background:
Perpetual is a non-profit organization that partners with municipal governments, zero-waste organizations, business leagues, community groups, and other stakeholders to implement city-wide, reusable foodware systems. It envisions cities where consumers can borrow reusable cups from anywhere they would normally purchase food and drinks and then return those cups either at the same locations or one of many outdoor collection bins. A third-party fleet of trucks and bicycles would visit these "foodware-using establishments" and outdoor bins on a schedule to drop off clean foodware and/or pick up dirty foodware for washing at a local depot.

Optimization is a key component of Perpetual's system design process. Placing outdoor location bins in areas of high foot traffic will increase the likelihood of consumer participation, while minimizing the total time traveled along the dropoff and pickup routes will reduce the overall cost and environmental impact. For the past year and a half, the DSI has been collaborating with Perpetual to develop optimized models of foodware delivery for two "pilot" cities: Galveston, Texas, and Hilo, Hawaii.

This quarter, Data Clinic students will create a data pipeline for classifying points of interest as candidates for outdoor bin locations using a rule-based algorithm. Anticipated work involves analyzing patterns in pedestrian foot traffic through clustering and identifying and scraping data sources to accurately estimate building sizes. Students will iteratively improve the vehicle routing and distribution models and apply them to the remaining two cities in the pilot: Ann Arbor, Michigan, and Savannah, Georgia. In addition, students will conduct a feasibility study to determine the optimal set of parameters, such as the number of vehicles, bicycles, coverage area, and bin placement, within each city.

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

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Rural Advancement Foundation International (RAFI)

Poultry Packaging Consolidation

Background:
RAFI's Challenging Corporate Power initiative fights consolidation throughout the food supply chain. The director of the Challenging Corporate Power initiative regularly engages legislators at the state and national levels to advocate for policy changes that prevent further consolidation. The DSI is helping RAFI build an interactive dashboard that shows consolidation in the poultry-packing industry. As large corporations gain control of more plants, farmers in monopsony-captured areas are vulnerable to exploitation. Short term, this map will be used in conversations with legislators to demonstrate areas of high market concentration. In the future, this map will be publicly available on the web.

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

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

University of Northern Iowa

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

Background:
The University of Northern Iowa (UNI) is a public university that offers more than 90 majors with a total enrollment of about 9000 students. UNI's main reputation is for its rich history in teacher preparation.

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

In the next iteration of the project, Prof. Matloob would want to shift the computational analyses to understand the dynamics of power within the narrative of novels and their association with African American characters. We would leverage similar large scale language models for automated analysis of our corpus to identify quantitative aspects in text that help define power across a character's narrative arc and how it alters for characters of different demographics.

Mentor:
Taraneh Matloob is an associate professor of children's literature at the University of Northern Iowa. She teaches multicultural children's literature as well as doctoral courses. Her scholarly interests are focused on multicultural children's literature, sentiment analysis, virtual reality, and augmented reality.

WBEZ

Illinois Traffic Stops

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

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

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

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