NORTHEAST BIG DATA INNOVATION HUB

Seed Fund Program Success Stories

2020 & 2021 COHORT REPORTS

www.nebigdatahub.org/seed-fund-about/

Table of Contents

NEBDHub Seed Fund Program

The Northeast Big Data Innovation Hub (NEBDHub) is a community convener, collaboration hub, and catalyst for data science innovation in the Northeast Region. Established in 2015 and extended in 2019 with grants from the US National Science Foundation, the Hub amplifies successes of the community, and shares credit across the community to encourage collaboration and mutual success in data science endeavors. As of January 2024, the NEBDHub has grown to include 9,645 individuals and 1,431 organizations spanning academia, non-profits, industry, and government across the 50 U.S. states, Puerto Rico, and 62 other countries around the world.

The NEBDHub Seed Fund program was established in 2020 to support the development and cross-pollination of tools, data, and ideas, leveraging data science innovations, across disciplines and sectors including academia, non-profits, industry, government, and communities.

The Seed Fund Steering Committee (SFSC) established in 2020 evaluated all proposals across the 2020 and 2021 seed fund cohorts for their adherence to the following criteria:

- Alignment with Hub mission, goals and focus areas including significance toward data science
- Intellectual Merit
- Broader Impact
- Potential for award to seed an expected larger effort
- Commitment / Ability to Execute

In 2020 and 2021, the Northeast Big Data Innovation Hub awarded 25 Seed Fund grants from a pool of 72 proposals. Over $600,000 was granted to awardees across the NEBDHub region. The projects were a great success, reflecting the following results:

- Education + Data Literacy (12 awards) was the most common focus area, followed by Health (10 awards), Responsible Data Science (7 awards), and Urban to Rural Communities (5 awards). Some awards spanned multiple Hub focus areas.
- 65 of the proposals were from academic institutions and 7 were from non-profits
- 23 of the seed fund grants were awarded to academic institutions and 2 to non-profits
- 5 of the grantees were from Minority Serving Institutions – 5 Asian American and Native American Pacific Islander-Serving schools and 3 Hispanic Serving Institutions.
- The broader impact of these awards was substantial, reaching over 36,000 individuals, including PIs, Co-PIs, graduate and undergraduate students, participants, and collaborators.

This Seed Funding was designated to support researchers in the northeast region, aligned with the Goals and Focus Areas of the Northeast Big Data Innovation Hub, including Education + Data Literacy, Health, Responsible Data Science, and Urban to Rural Communities.

The NEBDHub is grateful to all Seed Fund awardees for providing the following outcomes reports and research summaries for this publication.

Award Overview

PI & Institution	Seed Fund Project Title	Award Year	NEBDHub Focus Areas	MSI Status
Kaushallya Adhikari, University of Rhode Island	Convolutional Neural Network Facilitated Functional Cortical Mapping using tEEG Signals	2020	HEALTH; EDUCATION + DATA LITERACY	
Babak D. Beheshti, New York Institute of Technology	Data Literacy as an Enabler to Broaden the Participation of Underrepresented Minorities in STEM Careers	2020	EDUCATION + DATA LITERACY	AANAPISI
Todd Brothers, University of Rhode Island	CritCOVIDView: A Critical Care Visualization Tool for COVID-19	2020	HEALTH	
Judy Cameron, Pittsburgh Data Works	Expanding the Reach of DataJam: Introducing High School Data Science to More Diverse Youth, Communities and Regions	2021	EDUCATION + DATA LITERACY; URBAN TO RURAL	
Changqing Cheng, State university of New York	Nonlinear Dynamics and Machine Learning for Accurate Detection of Early-stage Atrial Fibrillation	2020	HEALTH; EDUCATION + DATA LITERACY	
Seth Cooper, Northeastern University	Location-based Citizen Science in Augmented Reality Image Categorization	2020	URBAN TO RURAL	
Yusuf Danisman, Queensborough Community College	Data Science Research and Training Program	2020	EDUCATION + DATA LITERACY	AANAPISI HSI
Laura Dietz, University of New Hampshire	Forecasting Salinity in Rivers during Storm Events	2020	URBAN TO RURAL	
Eli Fenichel, Yale University	Contact patterns during the 2020 COVID-19 pandemic	2020	HEALTH; URBAN TO RURAL	
Trent Gaugler, Lafayette College	DEFLAB:Data Education and Feminism at Lafayette and Beyond	2020	EDUCATION + DATA LITERACY	
Amanda K. Greene, Lehigh University	Curricular Structures to Blend Data Science & the Digital Humanities	2020	EDUCATION + DATA LITERACY; RESPONSIBLE DATA SCIENCE	
Cathie LeBlanc, Plymouth State University	Development of a Data Analytics Learning Community	2020	EDUCATION + DATA LITERACY	
Ho-Joon Lee, Yale School of Medicine	A landscape of virus-host protein-protein interactions in SARS-CoV-2 infection in humans by machine learning	2020	HEALTH	

PI & Institution	Seed Fund Project Title	Award Year	NEBDHub Focus Areas	MSI Status
Garrett Dash Nelson, Norman B. Leventhal Map and Education Center at Boston Public Library	Building Tools and Training for Public & Educational Use of Geospatial Big Data	2020	EDUCATION + DATA LITERACY	
Chirag Patel, Harvard Medical School	Using data-driven approach to study health disparities and secular trends in the chemical and individual exposome in the NHANES	2021	HEALTH; EDUCATION + DATA LITERACY	
Aunshul Rege, Temple University	Using Data Science to Study Environmental Racism, Justice, and Policy	2021	RESPONSIBLE DATA SCIENCE	
Steven Skiena, Stony Brook University	Knowledge Graph Embedding Evolution for COVID-19	2020	HEALTH	
Mona Sloane, New York University	How to Innovate AI Procurement?	2020	RESPONSIBLE DATA SCIENCE	
Seung Woo Son, University of Massachusetts, Lowell	Improving Data Integrity Awareness in HPC Datasets using Sparsity Profiles	2021	RESPONSIBLE DATA SCIENCE	AANAPISI
Julia Stoyanovich, New York University	All Aboard – Developing Protocols for Accessible AI Education	2021	EDUCATION + DATA LITERACY; RESPONSIBLE DATA SCIENCE	
Shuting (Ada) Wang, Baruch College	Home-Bias as a Double Edged Sword? Existence and Influence of Patients' Preference for Local Physicians on Virtual Health Platforms	2021	HEALTH	AANAPISI HSI
Robert Wieder, Rutgers University	Harnessing Data to Predict and Prevent Cancer Treatment Adverse Events through Artificial Intelligence	2020	HEALTH	AANAPISI HSI
Beverly Woolf, University of Massachusetts, Amherst	Building the Community to Address Data Integration of the Ecological Long Tail	2020	EDUCATION + DATA LITERACY	
Shanchieh (Jay) Yang, Rochester Institute of Technology	Teaching Responsible Data Science through Cybersecurity Analytics	2021	EDUCATION + DATA LITERACY; RESPONSIBLE DATA SCIENCE	
Hongyu Zhao, Yale University	A scalable computational pipeline to develop polygenic risk scores from biobank data	2020	HEALTH	

NEBDHub Seed Fund Steering Committee

David Bader
Chair, Seed Fund Steering Committee; Distinguished Professor, Department of Computer Science and inaugural Director, Institute for Data Science, New Jersey Institute of Technology

Florence Hudson
Ex Officio, Seed Fund Steering Committee; Executive Director and Co-Principal Investigator, Northeast Big Data Innovation Hub

Jenni Evans
Director, Institute for Computational & Data Sciences, and Professor Meteorology & Atmospheric Science, The Pennsylvania State University; Centennial President, American Meteorological Society

Josh Greenberg
Director, Alfred P. Sloan Foundation's Digital Information Technology Program

Chris Hill
Principal Research Engineer, Department of Earth, Atmospheric & Planetary Sciences, MIT

Jill Jemison
Assistant Dean for Technology and CIO, The Robert Larner, M.D. College of Medicine, The University of Vermont

Daniel Lopresti
Professor of Computer Science and Engineering, Lehigh University

Renée Miller
University Distinguished Professor of Computer Science, Northeastern University

FOCUS AREAS

HEALTH

EDUCATION + DATA LITERACY

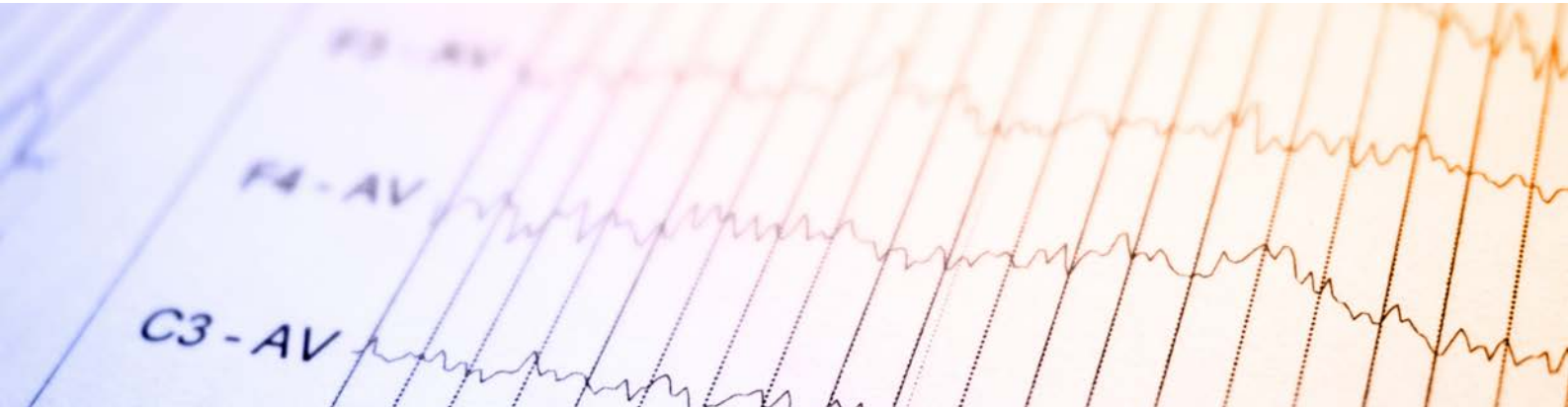
Convolutional Neural Network Facilitated Functional Cortical Mapping using tEEG Signals

Lead PI: Kaushallya Adhikari, University of Rhode Island

Kaushallya (Kay) Adhikari is an Assistant Professor of Electrical, Computer and Biomedical Engineering at the University of Rhode Island College of Engineering.

The goal for the project was to perform functional cortical mapping using tripolar electroencephalography (tEEG) and EEG data with convolutional neural networks (CNNs). The project used data from previous research efforts. In collecting the data, participants were shown a sequence of 50-drawings, each 3,500 milliseconds long. They were shown a grey color image between 2 drawings for 2,500 seconds. The participants identified each image under two conditions: overt (verbalize aloud the image name) and covert (silently name the image). Each condition was run twice (5 minutes long and total of 4 runs). During the experiment, their brain signal data was collected by tEEG and EEG, focusing on two important language areas in the brain, which are Broca and Wernicke.

This Northeast Big Data Innovation Hub Seed Fund project focused on classifying tEEG and EEG signals into right hemisphere and left hemisphere using CNNs. The researchers considered various input formats: spectrogram, raw two dimensional data, and two dimensional energy data. The results indicated that the CNNs cannot correctly classify left handed patients as right-hemisphere dominant and right-handed patients as left-hemisphere dominant based on tEEG and EEG signals.



The researchers also analyzed tEEG and EEG signals' energy levels before and during language stimulation. After filtering the signals with a 60Hz notch filter (fs=2000Hz), the project compared the mean energy during the rest (or baseline) period with the mean energy during the stimulation period for different channels by using p-values test. Project leaders also analyzed the changes in the following frequencies: Delta (0.1–4Hz), Theta (4–8Hz), Alpha (8–13Hz), Beta (13–30Hz), and Gamma (30–100Hz). This analysis showed that the same subject can have different active brain frequencies and brain areas at different recording instants. These results were not identical for EEG and tEEG signals. Some right handed patients had active areas in the left hemisphere while other right-handed patients did not. For left-handed patients, either the right hemisphere or both hemispheres were active. For every part of the brain (as indicated by the electrode placement), there was always at least one active frequency content.

In November 2021, Adhikari's team submitted a proposal to NSF Smart and Connected Health (SCH) based on this seed fund project.

FOCUS AREA

EDUCATION + DATA LITERACY

Data Literacy as an Enabler to Broaden the Participation Of Underrepresented Minorities in STEM Careers

Lead PI: Babak D. Beheshti, New York Institute of Technology

Babak D. Beheshti is a professor and dean of the College of Engineering and Computing Sciences (CoECS) at New York Institute of Technology. He received his Ph.D. in Electrical Engineering at the University of Massachusetts, Dartmouth, and his master's and bachelor's degrees in Electrical Engineering at Stony Brook University in New York.

The objective of this project was to expand data literacy and broaden the participation of underrepresented minorities and women in disciplines (and ultimately careers) in which an understanding of data science is foundational.

The development and delivery of the asynchronous "Introduction to Data Science" course had two main goals. The first was to increase data science capacity and talent by creating a sustainable pipeline from high schools and community colleges to universities, focusing on students interested in computer science and data science. The second aim was to increase the accessibility of data science in the broader community.

The project's ultimate goal was to make this course accessible to high school and community college students, as well as to the general public, by converting it to a fully asynchronous online mode. The project leveraged partnerships with local high schools and community colleges to advertise and, through a competitive vetting process, provide scholarships for a group of students to take this course free of charge. This allowed students from lower-income communities and students underrepresented in data science fields to have access to these educational resources.

The project earned tremendous appreciation from the course participants from across Long Island in New York State. Participants included 14 high school students (10 women and 4 men), 3 students from community colleges (men), and 8 students from New York Institute of Technology (NYIT) (3 women and 5 men).

Participants came to the program from the following high schools*:

- Hempstead High School – Nassau County (1,739 students; Minority enrollment: 98%; economically disadvantaged: 72%)
- Elmont High School – Nassau County (1,575 students; Minority enrollment: 98%; economically disadvantaged: 35%)
- Holy Trinity – Nassau County (1,300 students; Minority enrollment: 48.7%)
- St. John the Baptist Diocesan High School – Suffolk County (1,394 students; Minority enrollment: 28%)
- Central Islip High School – Suffolk County (2,343 students; Minority enrollment: 94 %; economically disadvantaged: 71%)
- Queens High School for Information Research and Technology – Queens, NY (500 students; Minority enrollment: 96 %; economically disadvantaged: 80%)

*US News & World Report Data

Participants came to the program from the following community colleges:

- Suffolk County Community College, ~27,000 participants
- Island Drafting and Technology Institute Enrollment, ~ 86 participants

Many students were reached through this project via recruitment efforts, including 6 high school students from Nassau County and 7 from Suffolk County who were enlisted to support the program. Also involved were students from Queens, New York City schools. Each of these students came from schools with high enrollment rates for minority and underrepresented communities. In total, the high school student body touched by this project included 8,851 high school students and 27,086 community college students. As

a result, the project team received an enormous number of applications and was unable to accommodate all requests due to limited funding.

With the above participants, the project's goals were achieved by building, developing, and advancing partnerships with the Long Island high schools, community colleges, as well as New York City public schools. This facilitated the creation of a sustainable educational pipeline for underrepresented and minority students interested in data science and related fields. Through these fundamental program offerings, these communities and participating students gained exposure to available resources and learned data science skills that can be leveraged to achieve economic advancements.

This project's impact could broaden with additional funding. This Seed Fund allowed the team to build the coursework foundations. With additional support, this project can successfully target more students from underserved and underrepresented communities, expanding the reach of data science education which supports high school graduates with career advancement. This would also support regional cybersecurity goals and provide STEM opportunities to women and minorities in particular. As of now, no additional grant funding has been secured, but the project team is considering several sustainability strategies.

The NEBDHub's seed funding allowed awardees to build partnerships and collaborate with local educational institutions. Beheshti and his team hope to expand the program in the future as it has sparked interest among high school and community college students in data science education and programs.

FOCUS AREA

HEALTH

CritCOVIDView: A Critical Care Visualization Tool for COVID-19

Lead PI: Todd Brothers, University of Rhode Island

Todd Brothers is a graduate of the Massachusetts College of Pharmacy and Health Sciences (MCPHS – Boston), is a Clinical Assistant Professor of Pharmacy Practice in the College of Pharmacy at the University of Rhode Island and is a board-certified clinical pharmacist, specializing in pharmacotherapy, pharmacogenomics, and critical care medicine. His professional experience has ranged from ambulatory and community based care to his current position in academia with an acute care focus.

The main goal of this project was to develop a cutting-edge tool, CritCOVIDView, for bedside clinicians interpreting individualized patient data through the development of an interactive dashboard. First, the team needed to develop a data mining algorithm to understand the prescribed medication patterns and analyze patient treatment modality complexities prior to and during the COVID-19 crisis. The team developed a custom association rule mining algorithm that efficiently discovered associations among the prescribed medications given the health status of critically ill patients. Stratified analysis was conducted by patient gender, race, ethnicity, and comorbidities.

Next, the team developed an interactive critical care dashboard to visualize prescribed medication patterns, lab results, and vital signs to facilitate prompt decision making. Then, the team developed a real-time interactive dashboard using an R-Shiny platform.

Dr. Abdullah Al-Mamun, PhD., the former (PI) on this project, is an Assistant Professor in the Department of Pharmaceutical Systems and Policy in the School of Pharmacy at West Virginia University. He was an essential member of the research team providing support and execution of the data science implementation, analysis, and assessment for the project.

Many students were involved in the development and execution of this award, including: D. Sabatino, a doctoral candidate (Pharm D), data acquisition; W. Cao, doctoral candidate (Ph.D.), data curation; J. Strock, doctoral candidate (Ph.D.), data curation, data science methods, and analysis; Allie Lindo, doctoral candidate (Pharm D), data curation; James Farrell, doctoral candidate (PharmD), data curation; and Martina Boda, doctoral candidate (PharmD), data curation.

The project team acquired a robust electronic health data set from Roger Williams Medical Center containing demographics, vital signs, laboratory indices, medication use, and treatment modalities (e.g. mechanical ventilation). The data science methods which supported this project were logistic and Cox regression(s), survival, and predictive modeling.

This seed grant will further support the development of an NIH grant to NIDDK [PAR-20-140], "Catalytic Tool and Technology Development in Kidney, Urologic, and Hematologic Diseases".

Publications

- Brothers T.N., Strock J., LeMasters T.J., Pawasauskas J., Reed R.C., & Al-Mamun M.A. (2022). Survival and recovery modeling of acute kidney injury in critically ill adults. SAGE Open Medicine, 10. https://doi.org/10.1177/20503121221099359

- Brothers, T., & Al-Mamun, M. (2021). 1322: Clinical characteristics and outcomes of acute kidney injury in critically ill adults. Critical Care Medicine, 50(1), 662–662. https://doi.org/10.1097/01.ccm.0000811612.78334.26

- Brothers T., Strock J., & Al-Mamun M. (2022). CO62 Evaluating the Clinical Characteristics of Acute Kidney Injury in the ICU Setting. Value in Health, 25(7):S315. https://doi.org/10.1016/j.jval.2022.04.160

- Al-Mamun M., Strock J., & Brothers T. (2022). CO54 Medication Regimen Complexity in the Critical Care Unit: Association with Length of Stay, Need for Invasive Mechanical Ventilation, and ICU Mortality. Value in Health. 25(7):S314. https://doi.org/10.1016/j.jval.2022.04.160

Poster presentations

- Predicting Inpatient Mortality Using Medication Regimen Complexity Score for Critically Ill Adult Patients. Brothers T., Strock J., Cao W., Sabatino D., Sikora-Newsome A., Al-Mamun M., ACCP Virtual Symposium

- "Evaluation of Seizure-Lowering Medication Use in the Intensive Care Unit Setting", Boda M., Brothers T., ASHP Midyear Conference, Virtual

- "Evaluation of Neuromuscular Blocking Agents in Patients with COVID-19 Infection", Lindo A., Brothers T., ASHP Midyear Conference, Virtual

- "Correlation of Severity of Acute Kidney Injury and Anti-infective Use in the Critical Care Setting", Farrell J., Brothers T., ASHP Midyear Conference, Virtual

Abdullah Al-Mamun is an Assistant Professor in the Department of Pharmaceutical Systems and Policy in the School of Pharmacy at West Virginia University. He received his Ph.D. in Computing and Information Sciences from the University of Northumbria at Newcastle, UK.

FOCUS AREAS

EDUCATION + DATA LITERACY

URBAN TO RURAL COMMUNITIES

Expanding the Reach of DataJam: Introducing High School Data Science to More Diverse Youth, Communities and Regions

Lead PI: Judy Cameron, Pittsburgh Data Works

Judy Cameron, Ph.D., is the Director of The DataJam, a member of the Advisory Committee for The DataJam, and a Professor at the University of Pittsburgh. She has a long history of translating science to the public.

The DataJam is an academic competition administered by Pittsburgh Data Works. The DataJam runs throughout the school year and is designed to introduce high school students to data science tools and techniques, enabling them to formulate and answer research questions. Students work in teams of 5–7 members to develop a research question, find publicly available data sets, analyze their data, make data visualizations, and present their findings to a panel of judges. Students learn skills pertaining to the scientific method, data analysis, and scientific communications.

The first DataJam was held in 2014 to broaden participation in data science education and meaningfully address the need for educational equity in marginalized student communities in the Pittsburgh, Pennsylvania region and beyond. Many students face economic, cultural, and pedagogical barriers to data literacy. The DataJam attempts to overcome those barriers. With support from the NEBDHub, the DataJam was able to serve other students in the northeast region. DataJam administrators made use of the Northeast Big Data Innovation Hub (NEBDHub) network of educators and collaborators to bring other local communities and high schools into the program.

Particular emphasis was placed on working with diverse learners from immigrant, rural, indigenous, and economically-disadvantaged communities. To enable a successful data science journey, the DataJam mentorship program ensures that students have consistent access to high-quality data science leadership and support.

A formal evaluation of this NEBDHub Seed Fund-supported project showed a lot of promise for national expansion of the DataJam as an effective educational program to engage underserved communities learning about data science. Of the 24 students in the 2022 cohort, 83.3% identified as Hispanic, and 58.3% of students identified as female, nonbinary, or gender nonconforming.

The Seed Fund project was very effective in engaging women and youth from underserved populations. After participating in the DataJam, 64% of student participants expressed an interest in STEM careers, indicating the potential long-term impact of this program. Although the pre-survey showed that the majority of students wanted to explore data science, there was a marked increase in interest in STEM after participating in the DataJam. The students demonstrated a greater understanding of data science concepts and applications after participating in the DataJam.

Overall, the DataJam continues to reach a diverse range of communities at a national level, engaging youth of all backgrounds in data science.

FOCUS AREAS

HEALTH

EDUCATION + DATA LITERACY

Nonlinear Dynamics and Machine Learning for Accurate Detection of Early-stage Atrial Fibrillation

Lead PI: Changqing Cheng, Binghamton University, State University of New York

Changqing Cheng is an Assistant Professor in the Department of Systems Science and Industrial Engineering at State University of New York at Binghamton.

The overarching goal of this Seed Fund project was to develop an integrated platform to integrate nonlinear dynamics analysis and data science for incipient-stage Atrial Fibrillation (AF) detection.

With the support from this Seed Fund award, the PI and his team accomplished the following: (1) designed data science algorithms to process electrocardiogram (ECG) data and unbalanced data learning; (2) developed a Mobile App to monitor cardiac dynamics, which was ranked the second place in the Mobile App Competition organized by the Institute of Industrial and Systems Engineers (IISE) in May 2021; and (3) engaging in development of a master's thesis to integrate the developed algorithm on AF detection.

This project has kickstarted other initiatives, including the formation of a new research team and the beginnings of a new NSF proposal, which will mention the preliminary results of this Seed Fund initiative.

Publications

- Shu Y., Dan W., & Cheng C. (2021). "Spatiotemporal regularization in effective reconstruction of epicardial potential," IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). https://doi.org/10.1109/bhi50953.2021.9508525

- Che Y., Guo Z., & Cheng C. (2021). "Generalized polynomial chaos-informed efficient stochastic kriging," Journal of Computational Physics, 445, 110598. https://doi.org/10.1016/j.jcp.2021.110598

- Che Y., & Cheng C. (2021). "Active learning and relevance vector machine in efficient estimate of basin stability for large-scale dynamic networks," Chaos: An Interdisciplinary Journal of Nonlinear Science, 31(5), 053129. https://doi.org/10.1063/5.0044899

- Shu Y., Cheng C., & Smith T.G., "A novel mobile application for Tele-ICU monitoring using electrocardiographic imaging (ECGI)," 21st Triennial Congress of the International Ergonomics Association, Vancouver, June 13 – 18, 2021.

FOCUS AREA

URBAN TO RURAL COMMUNITIES

Location-based Citizen Science in Augmented Reality Image Categorization

Lead PI: Seth Cooper, Northeastern University

Seth Cooper is an Assistant Professor at the Khoury College of Computer Sciences at Northeastern University. His work combines scientific discovery games (particularly in computational structural biochemistry), serious games, and crowdsourcing games.

The main goal of this project was to integrate location-based images into an augmented reality (AR) citizen science game toolkit by interfacing with APIs provided by other citizen science projects. The toolkit, called Tile-o-scope AR, uses an AR mobile app to display images onto a set of physical tiles, with which a variety of image matching games can be played.

Dr. Cooper's team integrated the Tile-o-scope AR app with iNaturalist's API so that it is possible to dynamically create image sets of animals found around a particular city on request. They then developed a user interface that allows the user to specify a city, and then a set of images around that city are requested from iNaturalist and assembled into an image set. The image set can be used to play games in the app. Improvements to the initial interface design and a more formal user study of location-based citizen science applications are pending.

As a result of this project, Dr. Cooper's team have included Tile-o-scope AR and related AR technologies in proposals around citizen science to the NSF (related to coastal climate change research and resilience planning) and NASA (related to citizen science in Louisiana), which could also incorporate the location-based features developed into educational applications. They expect to include location-based AR technology in future grant proposals (e.g. planning an NSF proposal related to collaborative AR

exploration of map and location data such as air quality). Dr. Cooper's team plans to use this feature for climate change adaptation planning in an upcoming research proposal for coastal communities and would like to carry out a more formal user study of the technology specifically developed during this seed project and the impact and use cases of location-based image labeling in AR. A recent study run on the general Tile-o-scope AR app indicates it may be most effective in settings such as museums, which could integrate well with location features.

Also involved in the development of this project was Colan Biemer, a Ph.D. student at the Khoury College of Computer Sciences, Northeastern University.

Collaborator:

Sara Wylie is an Assistant Professor at Northeastern University with a joint appointment in sociology/anthropology and health sciences, where she is a member of the Social Science Environmental Health Research Institute. Wylie seeks to develop new modes of studying and intervening in large-scale environmental health issues through a fusion of social scientific, scientific and art/design practices and is engaged in developing open-source research projects on low-cost thermal imaging, low-cost imaging of water pollution, and community-based methods for detection of hydrogen sulfide among other civic science projects.

FOCUS AREA

EDUCATION + DATA LITERACY

Data Science Research and Training Program

Lead PI: Yusuf Danisman, Queensborough Community College

Yusuf Danisman is an Assistant Professor at Queensborough Community College. He received his Ph.D. in Mathematics at The Ohio State University and worked at the University of Oklahoma before joining the City University of New York (CUNY). He is currently working on fractal geometry and its applications to the stock market. Yusuf is also a member of the National Student Data Corps' Founding Committee and in 2021 became a member of the Northeast Big Data Innovation Hub Steering Committee.

The aim of this Seed Fund project was to improve programming and data science skills of community college students from different majors and backgrounds. At the conclusion of the project, the fully online program consisted of: orientation, 11 weeks of online lectures, 3 weeks of individual/group project meetings, attending panels, and a talk given by an expert. Google Collab lecture notes, which include Python, Data Science subjects and projects, were created. In this program, students learned the basics of data science, how to write medium-size codes in Python, and to work on a project as a team and prepare presentations. Supervised learning algorithms, Linear Regression, k-nearest neighbors (kNN), decision tree, random forest, and unsupervised learning algorithms kMeans, Principal Component Analysis (PCA) were also covered.

Thanks to the Seed Fund, Yusuf Danisman's application to Queensborough Community College to initiate the Data Science Lab for Undergraduate Research was accepted. Further, they are planning to apply for an NSF grant to continue the project with a budget amount around $100,000.

FOCUS AREA

URBAN TO RURAL COMMUNITIES

Forecasting Salinity in Rivers during Storm Events

Lead PI: Laura Dietz, University of New Hampshire

Laura Dietz is an Associate Professor at the department of Computer Science at the University of New Hampshire. Her research focus is text-based machine learning and information retrieval and data science on watersheds.

Every winter, vast amounts of road salts are scattered onto streets across the northeastern U.S. While important for public safety, ice-melt and rainfall wash this road salt into river systems where it damages our ecosystems. Often, the negative impacts of high salt concentrations in river systems only become evident after extended periods of time. For example, when Hurricane Sandy flooded New York City with salt water in 2015, the disastrous effects on trees became visible three years later.

The transport of salt through waterways is only poorly understood by biogeochemists and hydrologists. The "Forecasting Salinity in Rivers during Storm Events" project takes a data science approach in forecasting the salt concentration in rivers across New Hampshire. The purpose was to analyze 'what-if' scenarios regarding salinity at particular river sites in order to estimate the impact of changing weather patterns (such as rain-on-snow, drought, or intense rainfall) and different road treatment events. For example, if a dry period in winter is followed by multiple severe storms, would we observe a sudden spike of salinity? What if frequent rain-after-snow events wash smaller amounts of salt into the environment on a consistent basis? Would the salt aggregate near roads or wash down the river? How are sudden changes in expected weather events affecting the river systems' ability to buffer salinity? Answering these questions allows us to quantify the resilience of different riverine ecosystems with respect to salt.

In this project, the team analyzed riverine data collected across multiple river sites in 15-minute intervals over the course of five years. The objective was to develop models that predict the expected salinity development over time, for a given series of temperature and flow rates. The team can query the model to predict how salinity is expected to change with weather patterns by varying the input flow rates and temperatures. Of particular interest are storm events and other abnormal weather patterns.

In earlier work, the team analyzed both traditional and deep learning models for time series for salinity forecasting. Prediction performance is quantified by 'root mean squared error' (RMSE) to the actual salinity. The vast majority of models can accurately predict near futures, but for long horizons deep neural GRU (Gated Recurrent Units) models perform best (3.8 RMSE). However, the accuracy of salinity prediction drops drastically during storm events. With 14.5 RMSE, the best model for storm events is a neural CNN (Convolutional Neural Network), which amounts to a five-fold error-rate increase. Since we are particularly interested in modeling storm events accurately, this performance loss is not acceptable for our purposes.

While the team focuses on salt concentration in rivers, the algorithms are designed to generalize to other solutes of scientific interest, such as dissolved organic carbon (DOC), nitrogen, and phosphorus. Here, focus is on aqueous sensors that are submerged in the river, but the algorithms transfer to sensor data collected in soil and air.

One novel contribution is the "Adjustable Context-aware Transformer model," which is designed to improve the quality of multi-horizon forecasts over several state-of-the-art methods, including the Transformer model, which has obvious shortcomings in the temporal context. The model overcomes the challenge that fast dynamics sometimes experience (such as during a rainstorm), where slow processes can dominate. The trick is, for every time point, to adaptively choose the right context-granularity, which is then used in the forecast. This approach was presented at the American Geophysical Union (AGU)

Fall 2021 meeting and published in an ECML European Conference on Machine Learning and Data Mining) workshop on Temporal Data.

A second contribution is a process model to learn the non-linear dependency between covariates (such as flow rate) and target variables (such as salinity). This work is still in progress as of September 2022.

Two students were supported on the seed award: Ph.D. student Sepideh Koohfar (forecasting of solute concentrations) and M.S. student Sarah Hall (process learning of flow-rate solute relationships).

The seed fund project intensified the relationship with researchers in the Natural Sciences, both at the team's home institution, the University of New Hampshire, and across the United States. Follow-up conversations after their presentation at the American Geophysical Union's (AGU) Fall 2021 meeting led to data exchange and a grant proposal that is currently under preparation in 2022.

Publications

- Koohfar S., Wymore A., McDowell W., & Dietz L. (2021). "Temporal Context Transformers for Multi-Horizon Prediction of Solute Concentration Responses." AGU Fall 2021 Meeting Abstracts.

- Koohfar S. & Dietz L. (2022). "Adjustable Context-aware Transformer." 7th Workshop on Advanced Analytics and Learning on Temporal Data. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML).

- Koohfar S. (2022). "Adaptive Temporal Pattern Matching." Women in Machine Learning Workshop at the Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS).

FOCUS AREAS

HEALTH

URBAN TO RURAL COMMUNITIES

Contact patterns during the 2020 COVID-19 epidemic

Lead PI: Eli Fenichel, Yale University

Eli Fenichel is the Knobloch Family Professor of Natural Resource Economics at Yale University. His research approaches natural resource management and sustainability as a portfolio management problem by considering natural resources as a form of capital.

The goal of the project was to bring together mathematics, data science, economic epidemiology, and public health, contributing to all fields by analyzing smart device contact data to infer behavioral responses to COVID-19 risk and COVID-19 policy. Smart device data have attracted a lot of attention during the COVID-19 pandemic. This project found good uses for these data and also explored their limitations. The NEBD Hub Seed Fund supported the acquisition and analysis of a large smart device dataset, and supported students learning to tackle real-world data problems, data analysis tools, and cloud computing at a large scale. It also supported several studies of pandemic-associated behavior patterns.

The project resulted in an undergraduate thesis (with two more underway), one master's thesis, and two other students were involved at a lower level of commitment.

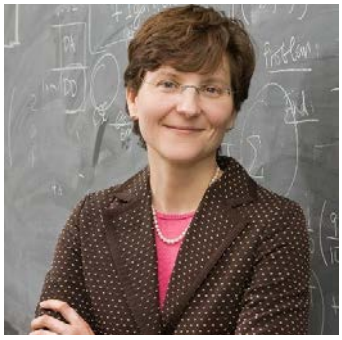
Publications

- Gonsalves G. S., Copple J. T., Paltiel A. D., Fenichel E. P., Bayham J., Abraham M., Kline D., Malloy S., Rayo M. F., Zhang N., & Faulkner D. (2021). Maximizing the Efficiency of Active Case Finding for SARS-CoV-2 Using Bandit Algorithms. Medical Decision Making, 41(8), pp.970-977.

A follow-on grant of about $40k was obtained by the project team to continue related work.

Collaborators:

Anna Gilbert is the John C. Malone Professor of Mathematics and Statistics & Data Science at Yale University. Her research interests include analysis, probability, discrete mathematics, and algorithms. She is especially interested in randomized algorithms with applications to harmonic analysis, signal and image processing, and massive datasets.



Roy Lederman is an Assistant Professor of Statistics & Data Science at Yale University. Lederman is interested in the organization and analysis of data, and he is working on computational and modeling problems in cryo-Electron Microscopy, a technology for mapping molecular structures.

FOCUS AREA

EDUCATION + DATA LITERACY

DEFLAB: Data Education and Feminism at Lafayette and Beyond

Lead PI: Trent Gaugler, Lafayette College

Trent Gaugler is an Associate Professor in the Department of Mathematics at Lafayette College. After completing his BS in Mathematics at Bucknell University in 2003, Trent received his Ph.D. in Statistics from Penn State University in 2008 under the supervision of Michael Akritas, and stayed on there for three years as a fixed-term assistant professor working primarily in the Statistical Consulting Center. Trent then moved on to a visiting assistant professorship in the Department of Statistics at Carnegie Mellon University.

The goals of this Seed Fund project were to introduce students to the fundamentals of data science through socially relevant projects, to enhance Lafayette College students' ability to design data science projects and communicate data science methods to other students, and to incorporate principles of "data feminism" into the entire project. The aim of this last goal was to facilitate students' learning of data science with awareness of gender and other social inequities and strategies for promoting gender equity through the use of data — in other words, to show students that they can use data science to effect social change.

Overall, the project was a great success. The team was able to purchase several copies of Klein and D'Ignazio's book "Data Feminism" to facilitate a community reading at Lafayette College. Two virtual community discussions (splitting the book in half) were held on the project and the events were quite well-attended. Team members were also able to bring Prof. Klein to campus (via Zoom) for a wonderful lecture on the book. In addition to this public lecture, Prof. Klein also agreed to a smaller meeting with the Seed Fund grant PI and several students who wanted to design their own data feminism-informed projects to share with peer students. She listened to their ideas and provided valuable feedback to sharpen their approaches.

These student projects were initially proposed to be delivered to community college students, but COVID forced us to alter those plans. In lieu of that venue for presenting the projects, the students instead visited an introductory statistics course at Lafayette and delivered the projects as lab assignments. They ran the course session entirely and developed their own assignment documents and guidelines. The statistics students eagerly engaged with these projects.

Collaborators:

Chris Phillips is Professor of English at Lafayette College, specializing in American and transatlantic literatures of the 18th and 19th centuries, history of the book, religion and literature, historical poetics, and the digital humanities. He is the author of numerous articles on the above subjects, including as the books Epic in American Culture, Settlement to Reconstruction, (Johns Hopkins, 2012) and The Hymnal: A Reading History (Johns Hopkins, 2018), and editor of The Cambridge Companion to the Literature of the American Renaissance (2018).

Jason Sims is the Manager, Research and High-Performance Computing at Lafayette College.

FOCUS AREAS

EDUCATION + DATA LITERACY

RESPONSIBLE DATA SCIENCE

Curricular Structures to Blend Data Science & the Digital Humanities

Lead PI: Amanda K. Greene, Lehigh University

Amanda K. Greene has been a researcher at the University of Michigan Medical School's Center for Bioethics and Social Sciences in Medicine (CBSSM) since August 2022. She works on projects exploring career development, gender issues, and women's representation in science and academic medicine, as well as projects exploring the ethical, legal, and social implications of health data sharing. Before joining CBSSM, she spent three years as an Andrew W. Mellon postdoctoral research associate at Lehigh University where she co-directed The Humanities Lab, a university-wide center for interdisciplinary research and teaching.

The goal of this project was to develop pedagogical resources that integrate humanist perspectives, ethics, and data science by supporting a collaborative working group of academics and industry professionals. The resulting working group developed and redesigned courses at the intersection of the digital humanities and data science. In emphasizing the role of data science in forwarding social justice initiatives and prioritizing data science ethics, these pedagogical structures targeted and cultivated students' technical capabilities. The result is that students were able to apply socially-conscious humanities skills during all phases of the data lifecycle. This project incorporated critical data competencies into coursework to effectively prepare students for digital humanities and data science careers outside of academia that center social justice.

The working group developed five model lesson plans with case studies and data sets that can be adapted and incorporated into the classroom at Lehigh University and Saint Vincent College in Pennsylvania. Two workshops were held to present these lesson plans. The first, at Saint Vincent College, reached an interdisciplinary audience of

twenty faculty members and introduced them to teaching tools at the intersection of data science and digital humanities. This workshop also promoted involvement in a new Digital Humanities undergraduate major.

The second workshop brought together Lehigh University faculty from the Anthropology, Computer Science, Mechanical Engineering, English, Women's Studies, and Physics departments, alongside colleagues from the College of Health. This workshop's focus was to disseminate and revise the lesson plans. The faculty considered ways to integrate elements of these lesson plans into their existing courses alongside Lehigh University's Humanities Lab and supported the development of a full syllabus for an undergraduate course titled "Algorithms and Social Justice" that will be co-taught by a faculty member in English, and Women, Gender, and Sexuality Studies with a faculty member in Engineering at Lehigh University in Fall 2022.

While students were not directly involved in this phase of the project, the groundwork it laid and initiatives it created will impact a large number of students in the coming years. It will directly impact students in the "Algorithms and Social Justice" course and students in other courses where faculty involved in the workshops are piloting new course structures. Additionally, the Digital Humanities minor at St. Vincent College (an institution where 43% of students are first-generation college students) will promote students' career development, providing them with sought-after data science skills so they can apply these skills in support of social justice.

Members from the working group at Lehigh University and St. Vincent College have submitted a proposal for the National Endowment for the Humanities (NEH) Institutes for Advanced Topics in the Digital Humanities grant, "A Humanities Toolkit for Data Science." The goal of this project is to extend the work of the seed grant in order to create opportunities for digital humanists and data scientists in the greater Pittsburgh area (where St. Vincent is located) and the greater Philadelphia area (where Lehigh is located) to collaborate on best practices for how to bring principles drawn from the humanities into the data science classroom.

Collaborators

Dominic DiFranzo (Lehigh University)
Edward Whitley (Lehigh University)
Annie Laurie Nichols (Saint Vincent College)
Lauren Churilla (Saint Vincent College)
Belle Lipton (Norman B. Leventhal Map & Education Center)
Catherine Nikolovski (CIVIC Software Foundation)

FOCUS AREA

EDUCATION + DATA LITERACY

Development of a Data Analytics Learning Community

Lead PI: Cathie LeBlanc, Plymouth State University

Cathie LeBlanc has been at Plymouth State University since 1998, first in the Computer Science and Technology Department, and since 2006, in the Communication and Media Studies Department, acting as chair of the department from 2011 to 2017. She currently serves as Plymouth State University's General Education Coordinator. She is the co-author (with Evelyn Stiller) of the textbook Project-Based Software Engineering: An Object-Oriented Approach and has created and/or edited several open educational resources including Creating Games and Tackling Wicked Problems, two texts that are used in classes at Plymouth State University.

Plymouth State University has a history of active faculty learning communities focused on various aspects of teaching. This latest learning community, funded by the Northeast Big Data Innovation Hub's 2020 Seed Fund Program, is focused on teaching data analytics content, particularly in classes where students are not expecting this content.

The activities of the learning community began in January 2021, when nineteen faculty participated in a week-long workshop related to teaching data analytics. As many members of the learning community had little experience with data analytics, Daniel Lee, Associate Professor of Economics and resident data analytics expert provided instruction about using R to analyze data. Cathie LeBlanc, Professor of Digital Media and General Education Coordinator, facilitated discussions about the science of learning and its implications for teaching data analytics content. In particular, we discussed the cognitive principles summarized in the 2015 Deans for Impact report called The Science of Learning and the relationship between those cognitive principles and teaching data analytics content.

As the goal of the grant was to increase the capacity of faculty at Plymouth State to teach data analytics content, the project team designed a co-teaching experience to learn more about the supports that help faculty to teach this content. Daniel Lee, Associate Professor of Economics, and Jonathan Couser, Teaching Faculty in History, co-taught a General Education interdisciplinary course called "Making Sense of 'Madness'." The course merged a humanities-focused examination of our cultural understandings of mental illness with a data analytics approach to gaining insights into those cultural understandings.

The course was divided into four modules, each focused on a theme related to mental illness: diagnosis, institutions, treatment, and cultural representations. For each module, students completed an assignment in which they analyzed some data and then wrote an essay about the ways in which the data confirmed, contradicted, or further illuminated the readings they had done related to that theme. For example, in the module about institutions, students compared census data from 1940 with census data from 1970 related to the demographics of institutionalized individuals. During this time period, the United States underwent a radical de-institutionalization process so that the overall number of people in mental health institutions declined dramatically. However, analysis of the census data showed that the numbers of people of color in mental health institutions during this time did not decline. The stability of the numbers of institutionalized people of color from 1940 to 1970 was not mentioned in any of the assigned readings. This insight from the data sparked a discussion that would not have happened otherwise and provided students with an understanding of the power of data analytics.

The experience of co-teaching data analytics content with Dr. Lee has given Dr. Couser the confidence that he can teach the class on his own the next time it is offered. He is also helping another faculty member learn data analytics concepts using R so that she can teach the class in the future. Each member of the learning community committed to including at least one module or assignment related to data analytics in one of their classes in either the Spring or Fall 2021 semester. The project team continued to meet monthly throughout the Spring 2021 semester to discuss the challenges we were facing in incorporating data analytics content and to shape the development of a Data Analytics minor. The minor will be proposed during this academic year and members of the learning community have expressed interest in continuing to meet throughout the Fall 2021 semester.

FOCUS AREA

HEALTH

A landscape of virus-host protein-protein interactions in SARS-CoV-2 infection in humans by machine learning

Lead PI: Ho-Joon Lee, Yale School of Medicine

Ho-Joon Lee is an Associate Research Scientist at the Department of Genetics and Yale Center for Genome Analysis, Yale School of Medicine. He obtained a Ph.D. in bioinformatics from Free University of Berlin and Max Planck Institute for Molecular Genetics in Germany and Master's degrees in theoretical physics and applied mathematics from Cambridge University and Swansea University in the United Kingdom. His postdoctoral training was in systems biology at Harvard Medical School. His research topics include single cell biology, systems/network biology, and biomedical machine/deep learning. In response to the SARS-CoV-2 pandemic, he initiated a voluntary COVID HASTE working group at Yale School of Engineering & Applied Sciences (SEAS) together with Dr. Prashant Emani (Yale University) to study molecular mechanisms of SARS-CoV-2 infection and drug repurposing.

The team's goal with this Seed Fund project was to first build machine learning classifiers for multi-level evidence prediction of virus-human protein-protein interactions based on protein sequence profiles of interacting proteins. By applying those classifiers, the team aimed to identify human proteins that are targeted by viral proteins of the novel coronavirus, SARS-CoV-2, that causes the COVID-19 disease, at the proteome level to offer insight into a SARS-CoV-2 interactome landscape.

The project team used tree-based ensemble learning models of random forests and XGBoost and deep learning models of GraphSAGE with protein sequence-based features for multi-level evidence prediction of virus-human protein-protein interactions. The large-scale public database of Viruses.STRING was used for model

development. This achieved respectable a performance of 74% AUC and 68% accuracy by the best XGBoost model. The team then made novel predictions of different evidence levels for SARS-CoV-2 virus-human protein-protein interactions in a comprehensive and unbiased way in silico, which could be considered as a new dataset of a virus-human protein-protein draft interactome. Human target proteins predicted with high evidence levels were also prioritized and functionally characterized for specific hypotheses, e.g. importance of cysteine and histidine in protein sequences and histone H2A as a target of multiple SARS-CoV-2 proteins.

As a result of this initial research, project leaders are now developing an NIGMS Technology Development Program R21 proposal to the National Institutes of Health (NIH). The proposal is to develop a more general analytical framework for virus-host protein-protein interactions using various machine/deep learning models, and to integrate with our two other projects of drug repurposing and network controllability to target and disrupt those interactions.

The methods and predictions by random forests and XGBoost classifiers are published as a preprint in bioRxiv. Results from GraphSAGE models will be incorporated in the next version.

Presentations

- Lee, H. (2022). "A landscape of virus-host protein-protein interactions in SARS-CoV-2 infection in humans by machine learning." COVID Information Commons (CIC) Lightning Talks & Research webinar.

Collaborators:

Prashant Emani is an Associate Research Scientist in the Department of Molecular Biophysics & Biochemistry at Yale School of Medicine.

Mark Gerstein is the Albert L Williams Professor of Biomedical Informatics and Professor of Molecular Biophysics & Biochemistry, of Computer Science, and of Statistics & Data Science at Yale School of Medicine.

Shrikant Mane is a Professor of Genetics, the Director of the MBB Keck Biotech laboratory, and Director of the Yale Center for Genome Analysis.

FOCUS AREA

EDUCATION + DATA LITERACY

Building Tools and Training for Public & Educational Use of Geospatial Big Data

Lead PI: Garrett Dash Nelson, Norman B. Leventhal Map and Education Center at Boston Public Library

Garrett Dash Nelson is a historical geographer who works extensively with geospatial technologies in the social sciences and humanities. He is interested in the relationship between geographic knowledge and civic questions like politics, community, and planning.

The primary goal of this Seed Fund project was to create a gateway for adult library patrons and K-12 educators to begin engaging with geospatial data at the Leventhal Map & Education Center at the Boston Public Library. This meant building both technical infrastructure and "social infrastructure" designed to facilitate access for non-specialist users to approach geospatial data from a critical perspective.

The first step was to design and launch a new Public Data Portal to host the Center's public-access data sets, based on a user-centered paradigm of human readable data. The Leventhal Center designed the Data Portal from scratch so it is easily maintainable by Center staff, and also to emphasize facilitated access to datasets for non-specialist users. This alpha version of our Data Portal has already become a major piece of the Center's data engagement infrastructure, and team members will post a white paper about its development, as well as making the codebase open access.

The second step was to create training materials to help patrons understand how to use the data portal and Geographic Information System (GIS) technologies as part of a broader introductory course on "Making Sense of Maps & Data." Together, these efforts are meant to help individuals critically evaluate maps and the data used to produce them, fostering public engagement with geospatial data literacy.

Another goal of the project was to engage three data empowerment interns from the Massachusetts Institute of Technology (MIT) Data + Feminism Lab to develop a set of reusable training materials for a new introductory course, "Making Sense of Maps and Data." The team offered this course to an initial cohort of 30 adult participants facilitated by both staff from the Leventhal Map & Education Center (LMEC) at the Boston Public Library and public data empowerment interns. The course was then offered a second time to a smaller group, facilitated solely by LMEC staff. An article on the success of this course has been published by the data empowerment interns.

In addition to the three student interns from the MIT Data + Feminism Lab, the project team members also worked with an advisory panel of high school teachers drawn from Boston Public Schools to ensure the portal and the instructional series are well-suited for use by educators and by high school students. Approximately a quarter of the participants in the pilot course of "Making Sense of Maps & Data" were current K-12 educators who can use these skills and course materials in their classrooms.

This initial version of the Public Data Portal will grow to become the Boston Public Library's primary repository for serving and describing open geospatial data sets. Additionally, the "Making Sense of Maps & Data" course will become the key gateway course for new adult patrons looking to work with geospatial data at the Center. The project team foresees this course as being a kind of "prerequisite" for other adult programs that we will offer to equip patrons with the ability to engage critically, creatively, and carefully with geospatial data.

Publications

McCann, T. (2021). "Making Sense of Maps and Data at the Leventhal Map and Education Center," Medium.

Collaborators:

Belle Lipton's background is in digital mapping and geospatial data librarianship. She is passionate about exploring the relationships between social issues, GIS data, and mapping technologies. She enjoys thinking about ways to use library best practices, web mapping tools and educational outreach together to engage and empower researchers with increased spatial literacy skills.

Michelle LeBlanc leads all aspects of teacher training, school programs and curriculum development at the Map Center. She has over 20 years of experience in museums and classrooms, teaching history and designing programming for varied audiences. She holds an M.A. in Public History from Northeastern University and is a licensed teacher for grades 5–8 in Massachusetts.

FOCUS AREAS

HEALTH

EDUCATION + DATA LITERACY

Using a data-driven approach to study health disparities and secular trends in the chemical and individual exposome in the NHANES

Lead PI: Chirag Patel, Harvard Medical School

Chirag Patel is an Associate Professor of Biomedical Informatics at Harvard Medical School. Chirag Patel's long-term research goal is to address problems in human health and disease by developing computational and bioinformatics methods to reproducibly and efficiently reason over high-throughput data streams spanning molecules to populations.

This research project considered the health challenges posed by environmental hazards across the U.S., with a particular focus on the health disparities experienced by different social groups. The research team considered chemical exposomes – the totality of chemical exposures and its secular trends – to understand the relationships between complex exposures, environmental inequalities, and health disparities. Earlier research by this group showed that geographic differences play a significant role in disease risk, potentially intersecting with other factors such as income, education, and individual behaviors.

Chemical environmental exposures may also be an important contributor to health disparities across U.S. populations (e.g. circulating dioxins and polycyclic aromatic hydrocarbons in blood). This project was designed to systematically investigate chemical co-exposures in the U.S. to identify disparities in the patterns, correlations, and temporal trends of exposures in the disadvantaged groups using chemical biomarker data (over 140 chemicals from 16 different classes) available in multiple National Health and Nutrition Examination Surveys (NHANES, 1999 to 2018).

Specifically, this project had three specific aims:

1) to estimate the temporal changes of the chemical mixtures in relation to the disadvantaged populations with simple and set-based statistics –- Spearman's rank correlation and canonical correlation analyses

2) to calculate a novel "chemical Gini index" to succinctly quantify the environmental inequalities between and within the disadvantaged groups

3) to develop a chemo-exposure risk score (CRS) to summarize the risks of particular health outcomes from the totality of chemical exposures and investigate how CRSs and their temporal trends differ in the disadvantaged groups.

The outcomes of this project demonstrate the value of using a data-driven approach to conduct exploratory analyses in disparity and inequality research. New insights that are previously hidden in the data could provide guidance for further studies with a traditional hypothesis-driven approach. Such a combination of approaches can speed up the discovery and translation of research findings to support science-based policies that are formulated to address environmental inequality and health disparity issues in the U.S.

Key results from this study were shared in a collaborative article titled "Spatio-Temporal Interpolation and Delineation of Extreme Heat Events in California between 2017-2021."

FOCUS AREA

RESPONSIBLE DATA SCIENCE

Using Data Science to Study Environmental Racism, Justice, and Policy

Lead PI: Aunshul Rege, Temple University

Aunshul Rege is an Associate Professor with the Department of Criminal Justice at Temple University, where she directs the Cybersecurity in Application, Research, and Education (CARE) Lab. She holds a Ph.D. and M.A. in Criminal Justice, an M.A. and B.A. in Criminology, and a B.Sc. in Computer Science. Her research examines the human, behavioral, and social aspects of cyberattacks and cybersecurity and has been funded by several NSF grants (CAREER, CPS, EAGER, SaTC EDU) and Department of Energy. She organizes the Social Engineering Event, which helps teach social engineering in a safe, fun, and ethical way. Dr. Rege serves on the Advisory Board for Raices Cyber, Black Girls Hack, and Breaking Barriers Women in Cybersecurity.

This project examined environmental injustice using a qualitative criminological lens. The project surveyed known case studies of environmental injustice in the United States to identify and rank harms along incidence and severity, and identify corresponding remediation processes, if any. Due to COVID-19, however, the team was unable to interview subject matter experts in the field.

The research team focused on the state of Pennsylvania, to provide a detailed and more manageable examination of environmental justice. The team filtered through the Environmental Protection Agency's (EPA) Superfund Site National Priority List to identify site types, EPA site scores, contaminants, harms count, impact level, and whether the issue was reconciled using any specific procedures for the state of Pennsylvania. The team then intersected these with the Census.gov datasets to identify poverty levels, racial and ethnic demographics, education backgrounds, and health issues. This intersection resulted in the creation of a dataset with 91 incidents

identifying and ranking superfund sites and corresponding racial and health disparities across the state of Pennsylvania. The team can share this dataset with anyone who wishes to study the data, use it for their research, and/or use it to inform the creation of similar datasets for other states in the United States.

This project supported Ph.D. candidate Katorah Williams, who was instrumental in data collection and analysis as well as the dataset generation and maintenance.

FOCUS AREA

HEALTH

Knowledge Graph Embedding Evolution for COVID-19

Lead PI: Steven Skiena, Stony Brook University

Steven Skiena is a Distinguished Teaching Professor of Computer Science at Stony Brook University. He was co-founder and the Chief Science Officer of General Sentiment, a social media and news analytics company. His research interests include algorithm design, data science, and their applications to biology.

With this award, the researchers at Stony Brook University built and made available to the research community a time-evolving knowledge graph associated with COVID-19 articles in Wikipedia, tracking changes since the beginning of the pandemic outbreak. With this dynamic visualization tool, researchers can observe how scientific understanding of the COVID-19 pandemic evolved between 2020 and 2022. Moreover, this model has potential future applications, as it can be used to demonstrate how knowledge is created and modified in broader contexts. This powers fundamental research into how embeddings evolve with time.

The approach is described in the paper "Subset Node Representation Learning over Large Dynamic Graphs" authored by Xingzhi Guo, Baojian Zhou, and Steven Skiena, and available on arXiv. As detailed in the paper, the team proposes a new method, namely Dynamic Personalized PageRank Embedding (\textsc{DynamicPPE}) for learning a target subset of node representations over large-scale dynamic networks. Additional information about the project, including datasets and coding, can be found on GitHub.

This project was developed with the support of Xingzhi Guo, a Computer Science Ph.D. student at Stony Brook University.

Conference Presentations

- Guo X. & Zhou B. (2022). "Subset Node Anomaly Tracking over Large Dynamic Graphs." 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining, Washington D.C.

- Skiena S. & Guo X. (2022). "Knowledge Graph Embedding Evolution for COVID-19." COVID Information Commons (CIC) Lightning Talks & Research webinar.

FOCUS AREA

RESPONSIBLE DATA SCIENCE

How to Innovate AI Procurement?

Lead PI: Mona Sloane, New York University

Mona Sloane is a sociologist working on inequality in the context of AI design and policy. She frequently publishes and speaks about AI, ethics, equitability, and policy in a global context. Mona is a Fellow with NYU's Institute for Public Knowledge (IPK), where she convenes the Co-Opting AI series and co-curates the The Shift series. She also is an Adjunct Professor in the Department of Technology, Culture and Society at NYU's Tandon School of Engineering, a Senior Research Scientist at the NYU Center for Responsible AI, and is part of the inaugural cohort of the Future Imagination Collaboratory (FIC) Fellows at NYU's Tisch School of the Arts.

Artificial intelligence (AI) systems are increasingly deployed in the public sector. As these technologies can harm citizens and pose risk to society, existing public procurement processes and standards are in urgent need of revision and innovation. This issue is particularly pressing in the context of recession-induced budget constraints and increasing regulatory pressures.

The AI Procurement Roundtables Project brought together leading experts in the public sector, data science, civil society, policy, social science, and the law to generate a structured understanding of existing public procurement processes and identify how they can best mitigate risk and support community needs. Three separate conversations focused on mapping data science solutions used by public institutions; algorithmic justice and responsible AI and governance innovation and procurement in the context of AI.

The report sets out to equip individuals, teams, and organizations with the knowledge and tools they need to kick-off procurement innovation as it is relevant to their field and circumstances. To do so, it first sets the scene by examining the histories and current issues related to procurement and AI.

It then outlines six tension points that emerge in the context of procurement and AI – definitions, process, incentives, institutional structures, technology infrastructure, and liabilities – each of which are paired with questions to help address these tension points.

The report also outlines five narrative traps that can hinder equitable AI innovation:

1. "We must engage the public."
2. "We must find simple definitions of 'X'."
3. "The main threat is the government use of data."
4. "One incentive shared across all actors can initiate change."
5. "We can create change in AI design and deployment through procurement alone."

Each narrative trap is presented with ways and strategies to avoid said trap. The report closes with four calls for action as concrete steps that can be taken to create environments in which AI procurement innovation can happen, namely to re-define the process, create meaningful transparency, build a network, and cultivate talent.

This seed fund project resulted in the development and announcement of a new global standard – the IEEE P3119 Standard for the Procurement of Artificial Intelligence and Automated Decision Systems.

Final Project DOI: AI and Procurement: A Primer
Associated Publication: A How-To Guide on Acquiring AI Systems, IEEE Spectrum

Collaborators
Rumman Chowdhury, Ph.D. (Parity)
John C. Havens (IEEE)

FOCUS AREA

RESPONSIBLE DATA SCIENCE

Improving Data Integrity Awareness in HPC Datasets using Sparsity Profiles

Lead PI: Seung Woo Son, University of Massachusetts, Lowell

Seung Woo Son is an Associate Professor in the Department of Electrical and Computer Engineering at the University of Massachusetts, Lowell. He was previously a postdoctoral researcher in the Electrical Engineering and Computer Science department at Northwestern University and the Math and Computer Science Division at Argonne National Laboratory.

As scientists conduct analyses that rely on large-scale simulations to achieve breakthroughs in many disciplines, their ability to trust the data produced is paramount. However, the data they generate, process, and transfer will be subjected to increasingly higher profiles due to various data anomalies, which may go undetected because of the lack of mechanisms to make scientists aware of data integrity compromises. The goal of this project was to exploit the existence of spatial sparsity profiles exhibited in scientific datasets for effective anomaly detection. A sparsity profile means that a few significant signal components could represent the given datasets concisely, thus minimizing the need for inspecting entire data points for anomaly detection.

The University of Massachusetts, Lowell team produced an evaluation framework to inject errors in various data formats (binary, CSV, netCDF, etc.) using a diverse error injection metric (point vs. relative, gaussian vs. uniform). The developed framework has been open-sourced and used for evaluating datasets for the PM2.5 prediction model. The Anomaly Detection with Sparsity Profile (ADSP) framework has also been used for evaluating various reference scientific datasets.

As a result of this research, the team has published a short paper titled "Anomaly Detection in Scientific Datasets using Sparse Representation" as part of the Proceedings of the First Workshop on AI for Systems, held in August 2023. Authors of this paper include Aekyeung Moon, Minjun Kim, Jiaxi Chen, and Seung Woo Son. Additionally, a master's student at the University of Massachusetts, Lowell was offered six credits of coursework for supporting this research project.

Additionally, these research findings were used to develop a grant proposal for NSF's OAC (Office of Advanced Cyberinfrastructure) program in December 2022. The proposal, titled "Improving Data Integrity for HPC Datasets using Sparsity Profile" (NSF #2312982), was awarded in June 2023.

FOCUS AREAS

EDUCATION + DATA LITERACY

RESPONSIBLE DATA SCIENCE

All Aboard – Developing Protocols for Accessible AI Education

Lead PI: Julia Stoyanovich, New York University

Julia Stoyanovich is an Associate Professor in the Department of Computer Science and Engineering at the Tandon School of Engineering, and the Center for Data Science. She is a recipient of an NSF CAREER award and of an NSF/CRA CI Fellowship. Julia's research focuses on responsible data management and analysis practices: on operationalizing fairness, diversity, transparency, and data protection in all stages of the data acquisition and processing lifecycle.

As artificial intelligence (AI) takes on a significant role in mediating our social lives, a democratic discourse around how this technology should be built, used, and regulated becomes increasingly important. To align the use of AI with broader social goals of equity, this discourse must include the voices of a diversity of stakeholders. This, in turn, requires that all stakeholders be appropriately informed about the basics of AI, to productively engage in critical conversations about the benefits and risks of this technology. In other words, AI education is needed! However, AI education itself is often inequitable, especially when it comes to accessibility and participation from people with disabilities. There is an urgent need to take AI education out of the classroom and into the public sphere, and to develop a culture and practice of accessibility. The mission of the All Aboard! project is to address this need.

The overarching goal of All Aboard! is to improve the accessibility of AI education. Towards this goal, the project team convened three roundtable discussions in Spring 2022, bringing together data scientists, disability scholars and activists, and social scientists. The group used material from the NYU Center for Responsible AI's public "We are AI" course as a concrete use case. Specifically, discussions focused on improving the accessibility of the videos and comics within the course.

The All Aboard! primer is the outcome of those discussions and provides concrete recommendations. In the primer, readers will find:

➢ Best practices and guidelines for making text-based and visual educational content accessible.

➢ A case study that illustrates how comics can be used to accessibly communicate AI concepts to the general public.

➢ Pointers to free resources you can use to improve the accessibility of educational content you are developing.

The All Aboard! project was the first of its kind, bringing together a diverse group of students, scholars and disability activists. In addition to convening this group and supporting participants in forging important connections among each other, the project succeeded in concretely improving existing public AI education in the form of the "We Are AI" course, and in producing ten key considerations for AI practitioners and educators to make their own pedagogies more accessible.

FOCUS AREAS

HEALTH

URBAN TO RURAL COMMUNITIES

Home-Bias as a Double-Edged Sword? Existence and Influence of Patients' Preference for Local Physicians on Virtual Health Platforms

Lead PI: Shuting (Ada) Wang, Baruch College

Shuting (Ada) Wang is an assistant professor of Information Systems at Zicklin School of Business, Baruch College, City University of New York. Her research interests include the spread of fake news, the impact of social media on society, and the information system design in different contexts such as online health, Fintech, and e-commerce.

In this study, the Baruch College team investigated the existence and influence of patients' preference for local physicians (i.e. home bias) on virtual health platforms by starting with two key questions:

1) Do patients, especially those from medically disadvantaged areas where there is a lack of experienced physicians, have home-bias when selecting online physicians?

2) What is the impact of patients' home-bias on the propensity of them to obtain definitive diagnoses?

Analysis yields several interesting findings. First, while results suggest the existence of home bias among patients by showing that they are more likely to select physicians from the local city than those from other cities, the magnitude of such home bias is very small among patients from medically disadvantaged cities. Surprisingly, only 3% of patients from medically disadvantaged cities chose local physicians, indicating that the lack of home bias may largely mitigate the healthcare demand away from medically disadvantaged areas and thus, threaten the survival of physician practices in these already underserved areas.

Second, results suggest that patients, especially those from medically disadvantaged areas, are less likely to obtain definitive diagnoses when consulting local physicians, demonstrating that the existence of home bias may prevent patients from choosing the best physicians for their diseases and thus lower the quality of healthcare services they receive on virtual health platforms.

Third, the researchers found that the availability of online reviews of physicians significantly moderates the existence and influence of home bias. Results suggest that the availability of online reviews can enhance the existence of home bias in medically disadvantaged areas and thus, help sustain healthcare demand in these areas, while largely mitigating the negative impact of home bias on reducing patients' chance to obtain definitive diagnoses.

Considering that millions of people are living in medically disadvantaged areas across the US, an examination of home bias in their choice of physicians on virtual health platforms yields important implications for policymakers, platform managers, and individuals. The team plans to extend this research by submitting a proposal to early career NSF Computer and Information Science and Engineering (CISE) grants (up to $180,000).

FOCUS AREA

HEALTH

Harnessing Data to Predict and Prevent Cancer Treatment Adverse Events through Artificial Intelligence

Lead PI: Robert Wieder, Rutgers University

Robert Wieder is an Attending Physician at Rutgers Cancer Institute of New Jersey at University Hospital and Provost of Rutgers Biomedical and Health Sciences Newark.

The goal of the project was to conduct an in-depth large-scale study of the comprehensive 1991-2016 Surveillance, Epidemiology, and End Results (SEER)-Medicare dataset*, which the investigators, Dr. Robert Wieder (PI), and Dr. Nabil Adam (Co-PI) obtained through a two-tiered review. The aim was to identify predictors of adverse events in patients treated for breast cancer, applying Machine Learning and Artificial Intelligence (ML/AI) techniques.

Task 1. The dataset

The dataset used for this project combined clinical data from population-based cancer registries with claims data from the Center for Medicare/Medicaid Services (CMS) Medicare program. The dataset consists of the following files:

- Patient Entitlement and Diagnosis Summary File (PEDSF), which has SEER data for breast cancer cases diagnosed between the 1990s and 2015
- Physician/Supplier (NCH) 1991-2015
- Outpatient 1991-2016
- MEDPAR 1991-2016
- Chronic Conditions Flags 1999-2016
- Part D (PDE) 2007-2016
- Part D Enrollment (PTden)
- Census Data by Zip Code
- Census Data by Census Tract

A part-time undergraduate senior, Aqsa Syed, was hired and was paid through this grant funding. Under the supervision of the investigators, the student helped set up the dataset.

Task 2. Data Fusion and Integration.
The project team developed an OMOP (Observational Medical Outcome Partnership) -based data model to fuse and integrate the SEER-Medicare dataset. OMOP provides a Common Data Model (CDM) for storing data within observational databases with common semantics (terminologies, vocabularies, coding schemes). The CDM cancer module provides the needed granularity and abstraction (e.g., recurrence, remission, end-of-life events, chemotherapy regimens, treatment cycles, response to treatments) of cancer data – diagnoses, treatments, and outcomes. A cancer diagnosis is defined by an assemblage of histology, site, stage, grade, and genetic biomarkers. Cancer treatments, which individual drugs cannot describe, are administered in specific order and cycles.

Task 3. Data Preprocessing and Analysis.

- Data Cleaning: renaming, sorting and recording, handling duplicate, missing, and invalid data, and filtering to the desired subset of data.
- Data transformation and standardization – Ensure all Features are Numeric and standardized.
    - Apply StandardScalar as part of the ML pipelines.
    - Apply Bucketizing for continuous values, e.g., age.
    - Apply encoding, e.g., one-hot encoding for categorical features.
- Tumours are characterized by site, laterality, stage, grade, ER and PR status, and Her2 status (human epidermal growth factor receptor 2).
- Patients are characterized for race/ethnicity, age, marital status, obese/overweight status, hypertension, comorbidities, syndromes, prior malignancies, and therapy.

The following paper is under review:
Wieder R. & Adam N. (2022). "Drug Repositioning for Cancer in the Era of Big Omics and Real-World Data." Critical Reviews in Oncology/Hematology, 175, 103730. https://doi.org/10.1016/j.critrevonc.2022.103730

The following proposal was submitted to the National Cancer Institute (NCI) for funding as a result of this seed funding:
Deep intelligence Comprehensive Cancer Care 1.1 (Di3C1.1)
Nabil Adam (PI), Robert Wieder (Co-PI), Sita Kapoor (Co-PI), and Tarek Adam (Co-PI)

Collaborator:

Nabil Adam is a Distinguished Professor of Medicine and Computer & Information Systems and specializes in cybersecurity, machine learning, healthcare technology, and clinical/healthcare informatics at Rutgers University. Nabil is also the Co-founder & CEO of Phalcon, LLC. He has served as the Vice-Chancellor for Research & Collaborations at Rutgers University.

FOCUS AREA

EDUCATION + DATA LITERACY

Building the Community to Address Data Integration of the Ecological Long Tail

Lead PI: Beverly Woolf, University of Massachusetts, Amherst

Beverly Woolf joined the University of Massachusetts, Amherst, in 1992 and became a Research Professor in 2006. She is the Director of the Center for Knowledge Communication at the University of Massachusetts, Amherst.

Dr. Woolf was appointed a Presidential Innovation Fellow in 2013 and was on loan to the U.S. National Science Foundation. She is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI) and has delivered numerous international invited talks and served as Program Co-Chair or Executive Board Member for several conferences. She has received several awards for best paper, poster or video in a conference and served on several journal editorial, conference review, and advisory boards, including IEEE Computer and AAAI Spring Symposium.

This research project focused on Educational Data Mining, Learning Science, and Machine Learning. The team built upon current big data research and combined teacher inquiry and learning analytics to enhance teachers' ability to collect and utilize real-time data about their students. The research explored, evaluated, and applied machine learning techniques for optimizing and simplifying teachers' assessment of students' strengths, weaknesses, and socio-affective profiles to better create and adjust educational plans.

FOCUS AREAS

EDUCATION + DATA LITERACY

RESPONSIBLE DATA SCIENCE

Teaching Responsible Data Science through Cybersecurity Analytics

Lead PI: Shanchieh (Jay) Yang, Rochester Institute of Technology

Shanchieh (Jay) Yang is a Professor in the Department of Computer Engineering at the Rochester Institute of Technology. He is also the Director of Research for the ESL Global Cybersecurity Institute. His research focuses on advancing generative AI, data science, and simulation for predictive cyber intelligence and anticipatory cyber defense.

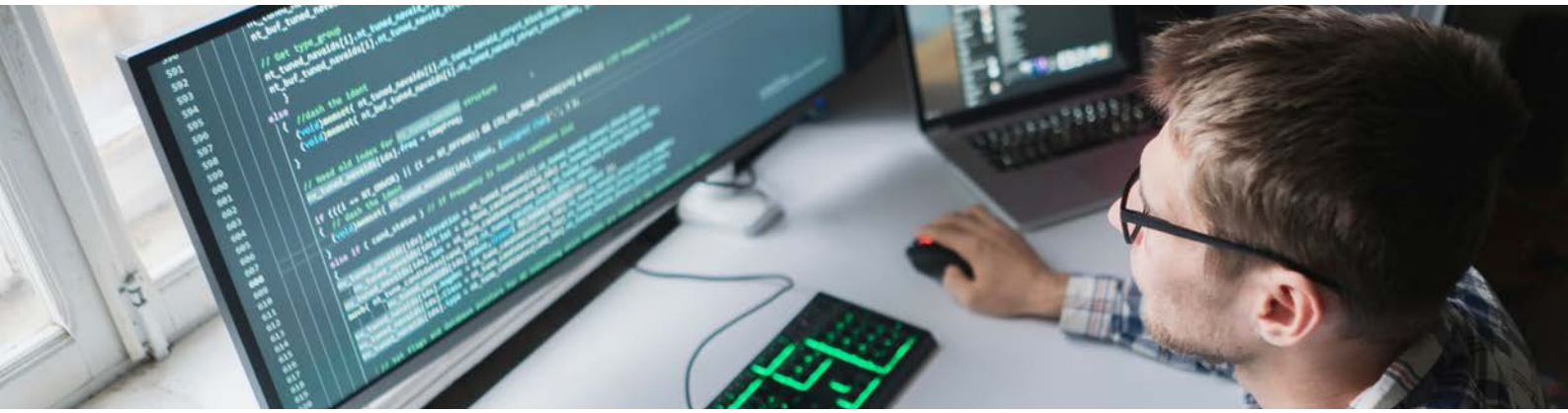
There were three primary objectives for this project:

1) To compile cybersecurity datasets from the open-domain for education and learning;

2) To develop education content for responsible data science from the cybersecurity perspective;

3) And to engage a broad community of students and professionals, with a particular emphasis on underrepresented minorities.

At the conclusion of this study, a list of continuously updated cybersecurity datasets was produced. This list contains network traffic data for cybersecurity analytics and was used in PI Yang's classes. The datasets collected represent varying quality and usability. Among the most commonly used open-domain repositories are: Canadian Institute of Cybersecurity Datasets, Czech Technical University, Stratosphere Research Laboratory Datasets, USB (University of Sannio, Benevento) Datasets, UNSW Dataset, and VizSec.



Additionally, PI Yang used this grant to develop and enhance a course on Machine Intelligence for Cybersecurity Analytics at RIT. The course is a semester-long project based course for senior undergraduate and graduate students. The key modules developed include the winsorizing of polarized feature values, upsampling, frequency encoding for categorical features, and the use of cross entropy to assess the effect of feature engineering. Some notable Git repositories developed by students in PI Yang's class include those from Vazgen Tadevosyan, Dan Popp, Anna Nicolais, and Varun Malhorta.

Students who worked with PI Yang in short-term research studies have examined a number of continual learning and transfer learning techniques for network traffic and malware analysis. In total, six students received partial support from this project (five undergrad, one master's student; two women; four international students). PI Yang and Chanel Cheng (RIT undergraduate computer science student) presented a poster on "Cross-Organizational Continual Learning of Cyber Threat Models" at the Annual Computer Security Applications Conference (ACSAC) in Austin, TX in December 2022. Other students, including Matthew Heller, Vazgen Tadevosyan, and Serena Yang, contributed to the various aforementioned feature analysis techniques across open-domain datasets.

At the conclusion of the project, PI Yang and his PhD student, Reza Fayyazi, worked with visiting international students, Pradumna, and Stavros Damianakis, to explore how Large Language Models (LLMs) may be applied to cybersecurity operations. This has led to the discovery of new research directions in investigating the tradeoffs of LLMs' creative imagination versus factually grounded capabilities for cyber-defense.

FOCUS AREA

HEALTH

A scalable computational pipeline to develop polygenic risk scores from biobank data

Lead PI: Hongyu Zhao, Yale University

Hongyu Zhao is the Ira V. Hiscock Professor of Biostatistics, Professor of Genetics and Professor of Statistics and Data Science in the Yale School of Public Health at Yale University. He received his B.S. in Probability and Statistics from Peking University in 1990 and Ph.D. in Statistics from the University of California at Berkeley in 1995. His research interests are the developments and applications of novel statistical methods to address scientific questions in genetics, molecular biology, drug developments, and precision medicine.

The goal of this project was to address the computational and implementation issues by developing a unified and user-friendly web platform for practicing PRS analysis and benchmarking most existing PRS methods.

Dr. Zhao's team developed a faster version of Annopred software, which significantly reduces the computational time (about 2X speed up) while achieving the best performance in the benchmark. The updated Annopred is freely available. A web-server presenting data processing details during PRS calculation and also visualizing benchmarking results across different algorithms is constructed. The web platform can also calculate genetic risks based on uploaded genotype files or new weights from users. Dr. Zhao's team intends to host the webserver on Google Cloud computing.

This project has established collaborations between Dr. Zhao's group and Dr. Robert Bjornson of the Yale Center for Research Computing. Together, they can develop efficient computation tools for polygenic risk scoring methods. An NIH application was submitted and ranked in the 3rd percentile based on some of the preliminary results

from this seed grant. The Yale portion of that R01 application was budgeted for $600K for four years.

A number of students supported this project, including Jerry Shan, an undergraduate at Yale University, Takintayo Akinbiyi, a postdoc at Yale University, and Wei Jiang, a postdoc at Yale University.

Collaborators:

Dr. Robert Bjornson, a Research Scientist in Computer Science at Yale University, has a background in parallel computing and bioinformatics. He has extensive experience with HPC algorithms and software in both academic and commercial environments. Dr. Bjornson manages the high-performance computing clusters for the Biological & Biomedical Sciences, and the Keck Biomedical HPC facility. In addition to providing training, consultation, programming support, and debugging assistance for user applications, he manages the installation and availability of a variety of HPC software applications, libraries, and tools. Dr. Bjornson holds a Ph.D. in Computer Science from Yale University.

Dr. Wei Jiang, an Associate Research Scientist in Biostatistics in the Yale School of Public Health at Yale University, received his Ph.D. in Electronic and Computer Engineering from the Hong Kong University of Science and Technology. His research interests lie in the fields of Bioinformatics and Biostatistics. His current research topic is to develop computational and statistical analysis methods in genome-wide association studies. His papers have been published in AJHG, Briefings in Bioinformatics, Bioinformatics, PLoS Computational Biology, etc. He received the best paper award in APBC2016 held in San Francisco.

NORTHEAST BIG DATA INNOVATION HUB

The Northeast Big Data Innovation Hub is supported by the National Science Foundation through awards #1916585 and #2333532, and previously by NSF awards #1550284 and #1748395.