UNC CHARLOTTE
School of Data Science

Seed Grants for Data Science Spring and Summer 2021 Recipients

Summer 2021

1. **Mechanism of Interplay between Genetic Factors and Social Determinant Health on Gender Disparities in Lung Cancer among Non-smokers.**

Dr. Shan Yan (PI) - College of Liberal Arts & Sciences/Department of Biological Sciences.
Dr. Yuqi Guo (PI) - College of Health and Human Services/School of Social Work
Dr. Yaorong Ge (Co-PI) - College of Computer and Informatics/Department of Software and Information Systems; School of Data Science

Increasing gender disparities in lung cancer incidence among non-smokers has become a prominent public health concern in recent years. Whereas genetic mutations in some penetrant genes may play an important role in non-smoker's lung cancer incidence susceptibility, female non-smokers experience a disproportionately high burden of lung cancer for reasons that remains to be fully understood. The purpose of our project is to determine the mechanism of gender disparities in lung cancer incidence and survivorship among non-smokers through analyzing the interplay between genetic factors and social determinant of health. The objectives of this proposed study aim to: 1) to determine the mechanisms of interplay between genetic factors and social determinants of health on lung cancer incidence in female non-smokers and male non-smokers; 2) to develop a machine learning workflow to predict therapy response to lung cancer treatment among non-smokers and examine the role of gender on therapy response.

Data for this study will be retrieved from the Southern Community Cohort Study. Several types of data-driven machine learning (ML) model including random forest, neural network, and support vector machine will be developed to identify the mechanism of lung cancer incidence among non-smokers.

2. **An Investigation Into Destructive Leadership Using Data Science and Social Science Methods.**

Dr. Scott Tonidandel (PI) - Belk College of Business/Department of Management; School of Data Science
Dr. George Banks (Co-PI) - Belk College of Business/Department of Management
Dr. Wenwen Dou (Co-PI) - College of Computer and Informatics/Department of Computer Science; School of Data Science
Dr. Janaki Gooty (Co-PI) - Belk College of Business/Department of Management

Destructive leadership is a less commonly occurring form of social influence that yet can have strong adverse effects for stakeholders across a variety of spectrums. As with other leader approaches, destructive leader behaviors (DLBs) have been sometimes confused with follower evaluations of those behaviors, which may vary by perspective. These evaluations can be prone

to retrospective bias as followers attempt to remember behaviors. Moreover, it is difficult to study DLBs dynamically over time if investigations are limited to annual cross-sectional surveys. We propose that a solution to these problems is the development of a Natural Language Processing (NLP) algorithm. We will qualitatively identify DLBs through interviews and then label text for analysis using BERT (Bidirectional Encoder Representations from Transformers) to develop a model that can automatically score DLBs in new text. We will complete one complementary experiment that would allow for causal inferences needed to conceptually defend this technique when applying for external funding.

3. **Objective Measures of Structural Racism Using Spatial Analytics**

Dr. Rajib Paul (PI) - College of Health and Human Services/Department of Public Health Sciences; School of Data Science
Dr. Ahmed Arif (Co-PI) - College of Health and Human Services/Department of Public Health Sciences
Dr. Jean-Claude Thill (Co-PI) - College of Liberal Arts & Sciences/Department of Geography & Earth Sciences; School of Data Science

The Aspen Institute for Community Change and Applied Research Center of UC Berkley define Structural Racism (SR) as the normalization and legitimization of policies and procedures that disproportionately benefit White people and put people of color under chronic adverse situations. SR is a long-standing problem in the U.S., however, the lack of objective and quantitative measures of SR makes the planning of effective intervention approaches and efficacy assessment of existing awareness and mitigation programs challenging. Using disparate structured and unstructured databases on socioeconomic variables, health status, delinquency and incarcerations, state and local laws and policies, voting records, etc., this research project aims to develop and validate quantitative measures of SR at the county and neighborhood level using Bayesian and spatial factor analysis approaches that incorporate knowledge on epidemiology, population health, biostatistics, big data analytics, urban analytics, socio-economic geography, and environmental systems. The developed SR index will be compared and validated against the University of Wisconsin's Area Deprivation Index (ADI) and The Centers for Disease Control and Prevention's Social Vulnerability Index (SVI). The knowledge gathered from this research will be used to prepare a research grant proposal to be submitted via the National Institutes of Health's R15 (AREA) mechanism to the National Institute of Minority Health and Disparities (NIMHD). The objectives of this research are aligned with NIMHD's current funding priorities.

4. **Data-Driven Estimation of Human Intent in Semi-Automated Steering Control.**

Dr. Minwoo Jake Lee (PI) - College of Computer and Informatics/Department of Computer Science; School of Data Science
Dr. Amirhossein Ghasemi (Co-PI) - The William States Lee College of Engineering/Department of Mechanical Engineering & Engineering Science

This project aims to develop a data-driven approach to estimate and predict human intent in a haptic shared control paradigm wherein a human interacts with a robotic partner through a physical object. With recent advances in artificial intelligence and robotics, conflicts may arise in which a co-robot can make decisions different from the human partner's. While Human teams can be exceptionally efficient at resolving conflicts using shared mental models, the ability of co-robots for negotiating and resolving conflicts is significantly underdeveloped. In this project, we propose to study the interaction between a human and a co-robot when a conflict arises. Specifically, we consider two forms of conflicts in terms of control commands: (i) the human and co-robot detect an obstacle but choose a different path to avoid it and (ii) either human or co-robot doesn't detect the obstacle. By performing a series of human-subject studies and employing dilated temporal convolutional neural networks, we seek to develop a data-driven approach for estimating and predicting the human's intent. While the fundamental approaches and models proposed in this research can be applied to a wide range of physical-human robot systems, we select steering control of semiautomated vehicles as a setup for exploring the proposed research. The experiments and evaluation studies will be supported by a fixed-base driving simulator at UNC Charlotte.

5. **Data-Driven Approach for Smart On-Demand Public Transit in Charlotte in Underserved Communities- Pilot Study for User Acceptance and Early Data Collection.**

Dr. Hamed Tabkhi (PI) - The William States Lee College of Engineering/Department of Electrical and Computer Engineering
Dr. Mona Azarbayjani (Co-PI) - The College of Arts + Architecture

Lack of access to adequate public transportation is a significant component of the problem of inequity and socio-economic mobility in low-income communities. Low-income workers who rely heavily on public transportation face a spatial mismatch between home and work, resulting in higher unemployment, longer job search times, and longer commute times. This equity challenge will continue following the pandemic, demonstrating the broad societal importance of developing systems that enhance public transportation for workers experiencing home-work spatial mismatches. The overarching goal of this proposal is to get initial data that would result in creating a connected, coordinated, demand-responsive, and efficient public bus system that minimizes transit gaps for low-income, transit-dependent communities.

To create equitable metropolitan public transportation, this seed proposal makes initial steps evaluating CATS mobile app, collecting data to understand the demand and transit gap of a pilot study--Sprinter Line, which connects Charlotte City Center to Charlotte International Airport. Building on the success of the seed funding, the team will seek larger NSF funding to create a demand responsive public transit to efficiently meet dynamically changing daily bus service needs.

UNC CHARLOTTE
School of Data Science

6. **Code Switching for Information Manipulation in Online Communication**

Dr. Lina Zhou (PI) - Belk College of Business/Department of Business Information Systems and Operations Management (BISOM); School of Data Science
Dr. Monica Rodriguez (Co-PI) - College of Liberal Arts & Sciences/Department of Languages and Cultural Studies
Dr. Samira Shaikh (Co-PI) - College of Computing and Informatics/Department of Computer Science; School of Data Science

Code switching (hereafter CS) refers to moving from one language to another within the same communicative event. Since bilinguals account for a significant percentage of the U.S. population, understanding the use of CS in online communication for deception creates unique opportunities. The primary objective of this proposed work is to identify strategies and behaviors in CS that are used for information manipulation in online environments.

Three specific aims are addressed in this work: 1) Identifying the domains that are highly susceptible to information manipulation in online communication and collecting datasets in both English and Spanish based on systematic review of the literature in the selected domains; 2) Understanding the prevalence of CS for information manipulation in online communication by conducting a diary study; and 3) Identifying CS in online communication by extending and evaluating natural language processing techniques**.**

Spring 2021

1. **Detecting and Analyzing the Stance of the Public Toward the COVID-19 vaccine on Social Media in English and Arabic.**

Dr. Mary Lou Maher (PI) - College of Computing and Informatics/Department of Software and Information Systems
Dr. Frederico Batista Pereira (Co - PI) - College of Liberal Arts & Sciences/Department of Political Science and Public Administration; School of Data Science

The purpose of this seed grant project is to explore the potential for semi-automated stance detection from social media in 2 languages: English and Arabic. In addition to developing a methodology for stance detection, this project will study the differences in public stance towards the COVID-19 vaccine in English and Arabic and whether tweets posted by humans have statistically different stances than tweets posted by bots. The methodology is applied to tweets in two languages to explore the differences in the use of language in expressing stance as well as to explore the reliability of machine learning classifiers in non-english text. The analysis of the stance of tweets from humans vs bots is relevant to understanding public opinion separately from the opinion of those trying to influence public opinion (bots).

2. **"Text-to-Knowledge Graph" Machine Reading for Strategy Mapping: Developing a Prototype for the Financial Industry**

UNC CHARLOTTE
School of Data Science

Dr. Victor Zitian Chen (PI) - The Belk College of Business; School of Data Science
Dr. Razvan Bunescu (PI) - College of Computing and Informatics/Department of Computer Science
Dr. Wlodek Zadrozny (Co-PI) - College of computing and Informatics; School of Data Science
Dr. Gus Hahn-Powell (External Collaborator) - The University of Arizona/Department of Linguistics

Business decision-makers rely on a causal map between strategy and performance to make good decisions. However, the causal insights about a business or industry are often fragmented and disconnected across various reporting and research documents. For instance, the International Federation of Accountants estimates the efforts of integrating insights from different reporting documents for decision making may have cost the financial industry alone $780 billion annually. Our proposed research seeks to address this problem by automating the detection and integration of key insights from textual data into a causal knowledge graph for data analytics and decision-making. We propose to develop a prototype of machine reading to detect, deconstruct, and integrate causal propositions from scholarly and reporting texts into a knowledge graph data, showing all the causal pathways among strategy-relevant variables and enterprise performance outcomes. Specifically, we will develop a prototype combining multiple machine-reading models, and train them on a sample of recent SEC documents of S&P financial companies. With an expert from the financial industry, we are developing a proposal for NSF PFI-RP grant due July 14, 2021, as well as pursuing industry sponsorship.