NORTH CAROLINA
AGRICULTURAL AND TECHNICAL
STATE UNIVERSITY

NSF HBCU-UP ACE DSA Program and Department of Mathematics and Statistics

2021 Fall Research Symposium on

Trends, Applications and Career Opportunities in Data Science and Statistical & Machine Learning

1:00-5:10 PM, Friday, November 12, 2021

Zoom Vritutal Conference

Zoom Meeting Link:

https://ncat.zoom.us/j/95575093026?pwd=RkIwY2ZqNGNVNWJRSmdtRHhFQkNUUT09

Meeting ID: 955 7509 3026

Passcode: 099245

2021 Fall Research Symposium on

Trends, Applications and Career Opportunities in Data Science and Statistical & Machine Learning

Hosted by the

National Science Foundation (NSF) Funded HBCU-UP ACE Implementation Project: Data Science and Analytics (DSA) Advancing STEM Education at North Carolina A&T State University (NSF Grant Award No. 1719498)

Department of Mathematics and Statistics
North Carolina A&T State University

TABLE OF CONTENTS

Overview of the Symposium

The NSF HBCU-UP ACE Data Science and Analytics Advancing STEM Education Program and the Department of Mathematics and Statistics at North Carolina A&T State University are jointly hosting a one-day research symposium on **Trends, Applications and Career Opportunities in Data Science and Statistical & Machine Learning** on Friday, November 12, 2021.

The purpose of this research symposium is to help the audience become knowledgeable about current trends, research frontiers, and academic preparations for career opportunities in the emerging fields of data science, big data analytics and statistical & machine learning. Symposium will include

- A plenary session
- A business and academic panel
- A faculty presentation session and
- A student presentation session

Symposium will feature data scientists/analysts from Facebook and Syngenta, and faculty and graduate students from North Carolina A&T State University and Texas A&M University as speakers and panelists. This research symposium is open and free to anyone who is interested in the topics.

Overview of the NSF HBCU-UP ACE Data Science and Analytics Advancing STEM Education Program

This NSF HBCU-UP Achieving Competitive Excellence (ACE) Implementation Project at North Carolina Agricultural and Technical State University seeks to develop and implement an innovative data science and analytics education and research program to train the next generation of undergraduate students to confront the challenges in computational and data-enabled science. The goal of this integrated, multidisciplinary undergraduate program is to use DSA to create an incubator for engaging URM students in computationally intensive tools training, data-enabled critical-thinking skills development, and global research and education experiences. Further, this project will transform institutional teaching and learning through evidence-based STEM gateway course reform and data-informed assessment and intervention. The project will be accompanied by an educational research study which will study the impact of faculty development and structures of support through communities of practice on faculty at the HBCU. The project is guided and informed by an on-going evaluation.

Specific objectives of the project are: to create an integrated education and research training environment; to increase faculty competitiveness in research and scholarship and prepare student readiness for graduate studies and professional careers in data science and analytics; to prepare globally competitive students; to facilitate and support faculty scholarly inquiries in research, teaching and learning; and to reach out to and collaborate with other institutions. The program will develop a data science and analytics certificate program, a statistical computing laboratory, and an analytical consulting center to broaden the participation and enhance preparation of students as well as support education and health analytics research development and outreach. The project will facilitate faculty research growth by contributing to professional development in research and by promoting collaborative research opportunities. Research and study abroad experiences through partnerships with two Chinese research universities will expose students to the data science research communities in China. Faculty engagement in scholarly inquiries in education analytics and development through a fellowship program, semi-annual education research symposia, workshops and research retreats will transform pedagogy and institutional teaching and learning. The program has the potential to be a replicable model in undergraduate data science and analytics education, research and training for other similar institutions.

Symposium Staff

NSF HBCU-UP ACE DSA Program Leadership Team

Guoqing Tang, Professor of Mathematics, Chair of the Mathematics and Statistics Department, and PI/PD, NSF HBCU-UP ACE DSA Program

Margaret Kanipes-Spinks, Professor of Chemistry, Director of the University Honors Program, and Co-PI, NSF HBCU-UP ACE DSA Program

Seong-Tae "Ty" Kim, Associate Professor of Statistics, Director of Statistics and Analytics Consulting Center, and Co-PI, NSF HBCU-UP ACE DSA Program

NSF HBCU-UP ACE DSA Program Steering Committee

Mohd Anwar, Professor of Computer Science
Lauren Davis, Professor of Industrial & System Engineering
Tamer Elbayoumi, Assistant Professor of Statistics
Kossi Edoh, Professor of Mathematics
Albert Esterline, Associate Professor of Computer Science
Scott Harrison, Associate Professor of Biology
Seongtae Kim, Associate Professor of Statistics
Margaret Kanipes-Spinks, Professor of Chemistry
Sayed Mostafa, Assistant Professor of Statistics
Guoqing Tang, Professor of Mathematics
Hong Wang, Professor of Management
Xiaohong Yuan, Professor of Computer Science

Symposium Agenda

NSF HBCU-UP ACE DSA Program and Department of Mathematics and Statistics

2021 Fall Research Symposium on Trends, Applications and Career Opportunities in Data Science and Statistical & Machine Learning

Friday, November 12, 2021

12:50—1:00 PM Check-in

1:00—1:55 PM Plenary Session

- **Speaker**: Dr. Lauren Davis, Professor of Industrial & System Engineering and PI, NSF NRT Program, NCAT
  Improving equity and access in hunger relief supply chains: models for prediction and distribution of uncertain supply
- **Moderator**: Dr. Guoqing Tang, Professor of Mathematics and PI, NSF HBCU-UP ACE DSA Program, NCAT

2:00-2:55 PM Business and Academic Panel Discussion

- **Panelists**: Dr. Jami Mulgrave, Facebook; Mr. Jonathan Fabish, Syngenta; Dr. Scott Harrison, NCAT; Dr. Yongli Zhang, NCAT and formerly Symantec; Ms. Chinenye Ifebirinachi, NCAT
- **Moderator**: Dr. Margaret Kanipes-Spinks, Professor of Chemistry and Co-PI, NSF HBCU-UP ACE DSA Program, NCAT

3:00 – 4:00 PM Faculty Presentation Session

- **Speakers**:
  3:00-3:30 PM***:*** Dr. Mohamed Ahmed, Assistant Professor of Geosciences, Texas A&M
  Filling temporal gaps within and between GRACE and GRACE-FO records: A machine learning approach

  3:30-4:00 PM: Dr. Mohd Anwar, Professor of Computer Science, NCAT
  Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning
- **Moderator**: Dr. Ty Kim, Associate Professor of Statistics and Co-PI, NSF HBCU-UP ACE DSA Program, NCAT

4:05-5:05 PM Graduate Student Presentation Session

- **Speakers**:
  4:05-4:20 PM: Sade Wilson Davenport, PhD Student in Computational Data Science and Engineering Program, NCAT
  Analysis of fungal genetics in the ocean using generative adversarial networks

  4:20-4:35 PM: Arbaaz Mohideen, MS Student in Applied Mathematics Program, NCAT
  Modeling the relationship between police funding and fatal police shootings in the United States

  4:35-4:50 PM: Steve Chesney, PhD Student in AST-Data Science & Analytics Program, NCAT
  A Comparison of ML & DL techniques for anomaly detection in MQTT-based IoT networks

  4:50-5:05 PM: Brett Hunter, PhD Student in AST-Data Science & Analytics Program, NCAT
  Classifying GRACE data for suggesting group interpolation for missing data

- **Moderator**: Dr. Sayed Mostafa, Assistant Professor of Statistics and Senior Scientist, NSF HBCU-UP ACE DSA Program, NCAT

5:05-5:10 PM Closing Remarks

Presentation Abstracts

Title: Improving equity and access in hunger relief supply chains: models for prediction and distribution of uncertain supply
Speaker: Dr. Lauren Davis, Professor of Industrial & System Engineering and PI, NSF NRT Program, NCAT

Abstract: During the past decade, an increasing number of natural disasters and humanitarian emergencies have prompted significant research in the area of relief chain logistics and supply chain management. Much of the research has focused on challenges associated with stocking and distribution of relief supplies in response to sudden-onset disasters. However, issues surrounding slow onset and persistent disasters (like food insecurity) present a unique set of challenges, particularly with respect to the management and distribution of donated supply. Based on a partnership with a local non-profit hunger relief organization, we describe the relief supply chain associated with the provision of food aid to populations suffering from hunger. We present predictive and descriptive models that quantify the availability of supply over time, characterize demand, and optimize the distribution of uncertain supply to ensure equity and improve access. Implications of our findings on operational efficiency and service delivery are discussed.

Title: Filling temporal gaps within and between GRACE and GRACE-FO records: A machine learning approach
Speaker: Dr. Mohamed Ahmed, Assistant Professor of Geosciences, Texas A&M University– Corpus Christi

Abstract: Temporal gaps within the Gravity Recovery and Climate Experiment (GRACE) (gap: 20 months), between GRACE and GRACE-Follow On (GRACE-FO) missions (gap: 11 months), and within GRACE-FO record (gap: 2 months) make it difficult to analyze and interpret spatiotemporal variability in GRACE- and GRACE-FO-derived terrestrial water storage ($\text{TWS}_{\text{GRACE}}$) time-series. In this talk, we share our recent results for a study in which we used an innovative approach that integrates three machine-learning techniques (deep-learning neural networks [DNN], generalized linear model [GLM], and gradient boosting machine [GBM]) and eight climatic and hydrological input variables to fill these gaps and reconstruct $\text{TWS}_{\text{GRACE}}$ data record at both global grid (1° × 1°) and basin (62 global watersheds) scales. Results from our robust and effective approach could be used to validate GRACE-FO datasets. By providing a continuous and uninterrupted $\text{TWS}_{\text{GRACE}}$ record, our research will promote additional and improved use of GRACE and GRACE-FO products by the scientific community, end-users, and decision makers.

Title: Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning
Speaker: Dr. Mohd Anwar, Professor of Computer Science, NCAT

Abstract: It has been over 21 months since the first known case of coronavirus disease (COVID-19) emerged, yet the pandemic is far from over. To date, the coronavirus pandemic has infected over 249 million people and has killed more than 5 million worldwide. The U.S. has the highest number of cases and deaths. This talk will present a study that aims to examine two questions: "*how useful is Reddit social media platform to surveil COVID-19 pandemic?*" and "*how do people's concerns/behaviors change over the course of COVID-19 pandemic in North Carolina?*" The purpose of this study was to compare people's thoughts, behavior changes, and discussion topics about the pandemic by applying natural language processing (NLP) and machine learning methods to COVID-19 related social media data.

Title: Modeling the relationship between police funding and fatal police shootings in the United States
Speaker: Arbaaz Mohideen, MS Student in Applied Mathematics Program, NCAT

Abstract: Fatal police shootings presents a major issue in the United States where on average nearly 1000 civilians are killed by the police annually. This issue has received considerable attention from many researchers who have attempted to investigate the various determinants that correlate with the occurrence of these shootings. Some of the factors that were found to be associated with the rate of fatal police shootings in the literature include firearm prevalence, urbanization/rurality, and the racial composition of the population. However, to the best of our knowledge, there does not exist any study in the literature that has investigated the association between the state's/city's level of spending on police and the rate of fatal police shootings. In this study, we use data on fatal police shootings from the Washington Post's "Fatal Force Database" (2015–2020) combined with data on police funding and many other ecologic variables (e.g., % of shooting incidents with investigation completed, crime rate, poverty rate, urbanization, population diversity, population density, etc.) to explore the association between police funding and fatal police shootings at both the state and city levels. Our extensive analysis shows that police funding is positively and significantly associated with the rate of fatal police shootings at the state level after accounting for the other covariates. At the city level, however; the positive association becomes insignificant when the other covariates are accounted for in the models.

Title: A comparison of ML & DL techniques for anomaly detection in MQTT-based IoT networks
Speaker: Steve Chesney, PhD Student in AST-Data Science & Analytics Program

Abstract: It is commonly understood that deep learning techniques have proven to be very effective in detecting malicious traffic in network architectures. While most research has been done against traditional networks, of late, several IoT (Internet of Things) network datasets have emerged for analysis with ML and DL algorithms, to include BoT-IoT, IoT-23, IoT Network Intrusion and MQTT-IOT-IDS2020. With the MQTT-IOT-IDS2020 dataset a Convolutional Neural Network (CNN) was used for evaluation with very positive results; however, traditional machine learning methods have proven to be just as effective. We will show a comparison between the two techniques and introduces the use of additional deep learning techniques for binary classification against the MQTT-IOT-IDS2020 dataset.

Title: Classifying GRACE data for suggesting group interpolation for missing data
Speaker: Brett Hunter, PhD Student in AST-Data Science & Analytics Program, NCAT

Abstract: Missing data in Gravity Recovery and Climate Experiment (GRACE) and GRACE Follow-On (GRACE-FO) time-series data causes errors in forecasting the time series. This presentation aims to group water basins to suggest interpolating missing data in GRACE-derived terrestrial water storage ($\text{TWS}_{\text{GRACE}}$) based on a detrended variability for the data. We are trying to capture key features of the time series data using K-means clustering on a zeroed mean or detrended data and standard deviation on a non-missing portion of data. The next step is to classify each group to have an optimal model to reconstruct the missing value for $\text{TWS}_{\text{GRACE}}$. Models will be constructed using the addition of exogenous variables previously identified to be correlated to $\text{TWS}_{\text{GRACE}}$.

Title: Analysis of fungal genetics in the ocean using generative adversarial networks
Speaker: Sade Wilson Davenport, PhD Student in Computational Data Science and Engineering Program, NCAT

Abstract: Our objective is to relate genetic variation of signal transduction pathways to different associated environments, specifically as would concern potential adaptations where pathways either diversify in function or become vestigial. For conducting this study, we would analyze genomic data across yeast lineages by identifying orthologous associations across these genomes based on annotated gene sequences originating from laboratory studies. Contrasting instances of yeast genomic data across 68 different environments were collected from Tara Oceans. In order to differentiate sequences, we have begun to utilize an unsupervised method of machine learning known as the Generative Adversarial Network (GAN). GANs are traditionally used for photographic image analysis. In our approach, we have oriented this capability upon genetic variation across environments. To more fully evaluate the potential of GAN, a gray box method was used to re-run our initial analysis by providing GAN with human-guided preprocessing of data. We have then more comprehensively modeled data from Tara Oceans with a goal to examine for pathway differences through logistic regression. Following this, we have implemented a semantic pathway analysis based on Resource Description Framework (RDF) constructs using Integrated Network and Dynamical Reasoning Assembler or INDRA. Our findings are that INDRA helps to construct and visualize pathways based on semantic biological phrases, and that we can utilize this information in interpreting genetic variation for yeasts found in both environmental and laboratory contexts. Our overall work demonstrates the use of data integration and modeling of Tara Oceans along with the GAN-based identifications to enrich the INDRA visualization.

Biographies of the Speakers and Panelists

Dr. Lauren Davis is a Professor in the Department of Industrial and Systems Engineering at North Carolina A&T State University. Her research focuses on decision-making under uncertainty primarily using stochastic optimization techniques (Markov Decision Processes, stochastic programming) and simulation. Her work has been applied to solve optimal stocking, transportation scheduling and distribution decisions in for-profit and non-profit supply chains. She has more than 40 peer-reviewed journal papers and refereed conference proceedings addressing issues related to inventory management, transportation scheduling, and emergency response in areas such as food supply chains, food security, port operations, and humanitarian relief. Her work has been supported by the National Science Foundation, Department of Homeland Security, and US Department of Agriculture totaling more than $4 million in grant funding. Additionally, her research examining hunger relief supply chains has been featured in CNN's Great Big Story and NSF's Discovery article series. She is currently the Principal Investigator for an NSF-funded National Research Traineeship grant that explores food security and hunger relief using computational data science.

Dr. Mohamed Ahmed is an Assistant Professor of Geosciences at Texas A&M University – Corpus Christi. He applies an integrated (geophysics, remote sensing, hydrogeology, modeling, GIS) approaches to investigate a wide range of geophysical, geological, hydrological, and environmental problems. Before attaining a Ph.D. in Geosciences at Western Michigan University in 2012, he received a M.Sc. (2009) and B.Sc. (2004) in Applied Geophysics from Suez Canal University, Egypt. He offers more than a decade of research and teaching experience and an excellent record of publication. His portfolio includes a number of awards and myriad articles published in professional national and international journals. Dr. Ahmed has collaborated in research funded, among others by, NASA, NSF, NOAA, SEG, and NATO. He has presented papers and served as a discussant at several national and international conferences and symposia.

Dr. Mohd Anwar is a professor of Computer Science at North Carolina A&T State University. He is an interdisciplinary computer scientist with research expertise in two main areas: (1) cybersecurity and (2) smart and connected health. The former is focused on intrusion/malware detection, usable security, cyber identity and differential privacy, and the latter is focused on mHealth technology-based individual-level health monitoring and health service delivery as well as AI-powered, secondary data-driven (e.g., social media data) public health monitoring. Towards pursuing his research goals, he uses *AI, Data*

Science, Human-Computer Interaction techniques as well as apply theories from Social Sciences to design solutions.

Dr. Jami Jackson Mulgrave is a research scientist in the People Analytics and Workplace Strategy group at Facebook. She was a National Library of Medicine postdoctoral research fellow at Columbia University in the Bioinformatics department. She earned a PhD in Statistics at North Carolina State University in 2018. She was a recipient of a National Science Foundation graduate research fellowship and received a $10,000 scholarship as a runner-up for a Pepsi-Co "Hidden Figures" search.

Mr. Jonathan Fabish is a Statistics Consultant at Syngenta RTP. He is a native of Greensboro, North Carolina where he has earned four undergraduate degrees in philosophy and Spanish from the University of North Carolina at Greensboro and in biological engineering and applied mathematics, and a master's degree in applied mathematics from North Carolina Agricultural and Technical State University. He completed most of PhD in Statistics coursework and obtained a master's degree in Statistics from North Carolina State University. He is very passionate about yoga, nature, science, math, and cats, and dedicated his life to the betterment of society through science and sustainability.

Dr. Scott Harrison is an Associate Professor in the Department of Biology at North Carolina A&T State University. Dr. Harrison manages a bioinformatics cluster within the College of Science and Technology where bioinformatics tools are used to investigate different types of biological systems with a goal to increase the speed and frequency by which computational analysis interfaces with experimental investigation. Specific research areas have focused on methods for genomic analysis, the modeling of persistence and survival in diverse types of populations, and host-microbial interactions. His laboratory hosts students whose interests are interdisciplinary and transdisciplinary, including undergraduates with interests in data science and analytics, and graduate students from the Department of Computational Science and Engineering. He is a core member of a multi-laboratory collaboration in genomics research and training

Dr. Yongli Zhang is an Associate Professor in the Department of Mathematics and Statistics at North Carolina A&T State University. Dr. Yongli Zhang got his PhD from University of Minnesota and Joined NCAT as an associate professor of statistics in 2019. Before he joined NCAT, he was a senior data scientist at Symantec Corporation in Silicon Valley. Dr. Zhang's research ranges from statistics and econometrics to machine learning and causal inference. He has published in the Journal of Econometrics and the Journal of Business and Economic Statistics. Since he joined NCAT, Dr. Zhang helps build up the data science program and actively does research in the interdisciplinary field between machine learning and causal inference. Dr. Zhang was a recent awardee of an NSF EiR grant ($780k) as the leading PI.

Ms. Chinenye Ifebirinachi is a second-year Ph.D. student in the Applied Science and Technology Program with the Data Science and Analytics concentration. She is a Graduate Research Assistant at the North Carolina A&T State University’s Department of Mathematics and Statistics, under the mentorship of Dr. Seongtae Kim and Dr. Yongli Zhang, and a recipient of the 2021 ACM SIGHPC Fellowship. Prior to this role, she was the inaugural Ph.D. scholar at NCAT’s Center for Outreach in Alzheimer’s, Aging, and Community (COAACH), where she functioned as a student researcher. Chinenye holds a BSC in Pure and Applied Mathematics from the University of Benin, Benin City, Nigeria, and she is on track to earn an MSC in Data Science and Analytics, alongside her Ph.D. degree. Before moving to the United States for her graduate program, she worked in Nigeria's technology, education, and finance industries. Her research interests are network data analysis and the use of quasi-experimental methods for causal inference in observational studies involving health data.

Mr. Steve Chesney a PhD Candidate in Applied Science & Technology with a concentration in Data Science & Analytics at North Carolina A&T State University. His research has been on how ML & DL algorithms can be utilized to detect and classify cybersecurity attacks against IoT devices. He earned an MBA from Wake Forest University and is a Gulf War Veteran with over 30 years of experience in the Information Technology (IT) industry. Throughout his career, he has achieved 22 certifications, 9 of those certification are in cloud technologies. Steve is happily married to his high school sweetheart for 30 years and the couple have two wonderful children. His daughter is in graduate school pursuing her law degree and his son is a sophomore in college. He enjoys watching basketball, track and field and he is an avid swimmer.

Ms. Sade Wilson Davenport is a PhD student in the Computational Data Science and Engineering program at North Carolina A&T State University. She is originally from Chesapeake, VA, and earned a Bachelor of Science in Biology and a Master of Science in Computational Science and Engineering. Her work as a computational science and engineering professional is to develop effective approaches for data collection, predictive analytics, and the modeling of complex systems. She seeks to use computational methods to analyze the increasing scope and potential of Big Data for real-world systems analysis and risk factor modeling, and has had ongoing involvements in how this type of approach can be applied to the study of biological and environmental systems.

Mr. Brett Hunter is a devoted Aggie here at North Carolina A&T State University. He earned a BS degree in Mathematics in the spring 2011, and obtained his master's degree in Applied Mathematics in the spring 2016, both from NCA&T. He currently is a 3rd year Ph.D. student in the doctoral program in Applied Science and Technology with a focus in Data Science and Analytics. He has taught classes in the Mathematics and Statistics department for over four years. He has been a mentor and tutor since he started at NCAT. He has also served the department as a graduate research assistant aiding summer research programs for NSF-funded programs.

Mr. Arbaaz Mohideen is a graduate student at NC A&T pursuing a master's in applied mathematics. He has a bachelor's degree in applied mathematics with a concentration in statistics. He has research experience in areas such as Prediction Modeling, Regression Analysis, Sentiment Analysis, Topic Modeling and Generative Adversarial Networks (GANs). He loves mathematics and statistics and is always on a pursuit to excel in and contribute to the field.