UNIVERSITY OF DELAWARE

DATA SCIENCE INSTITUTE

DELAWARE Data Science SYMPOSIUM

NOV 15, 2019

#datascienceUD19

SYMPOSIUM AGENDA

Registration	8:15-9:00am
Welcome Remarks: Charlie Riordan and Cathy Wu	9:00-9:10am
Morning Keynote Speaker: Hendrik Hamann	9:10-9:55am
UD Students and Postdocs Lightning Talks: Session 1	10:00-10:15am
Coffee Break and Poster Session 1	10:15-10:35am
UD Faculty and Researchers Talks: Session 1	10:35-11:35am
Education Panel Session	11:40am-12:20pm
Lunch	12:20-1:35pm
Afternoon Keynote Speaker: Lauri Goldkind	12:50-1:35pm
Industry Panel Session	1:40-2:20pm
UD Students and Postdocs Lightning Talks: Session 2	2:25-2:40pm
Coffee Break and Poster Session 2	2:40-3:00pm
Tech Demo: Marvin Andujar	3:00-3:30pm
Provost Morgan Remarks	3:35-3:45pm
UD Faculty and Researchers Talks: Session 2	3:50-4:50pm
Closing Remarks and Award	4:50-5:00pm
Reception *(In the Atrium)*	5:00-5:30pm
Industry Tables *(All Day in the Atrium)*	9:00am-5:00pm

MORNING KEYNOTE SPEAKER

Hendrik Hamann, PhD

Distinguished Researcher and Senior Manager
Physical Analytics
Thomas J. Watson Research Center, IBM

Dr. Hendrik F. Hamann is a Distinguished Researcher at the IBM T.J. Watson Research Center, Yorktown Heights, NY. He received his PhD from the University of Göttingen in Germany. In 1999 he joined the IBM T.J. Watson Research Center, where has been leading the Physical Analytics and cognitive Internet of Things program. Hamann's interest includes sensor networks, sensor-based physical modeling, machine-learning, artificial intelligence, big data technologies and geospatial analytics. Hamann has authored and co-authored more than 120 peer-reviewed scientific papers and holds over 120 patents and has over 130 pending patent applications. Dr. Hamann is an IBM Master Inventor, a member to the IBM Academy of Technology and has served on governmental committees such as the National Academy of Sciences, the National Science Foundation and as an industrial advisor to Universities. He won several awards including the 2016 AIP Prize for Industrial Applications of Physics. He is a fellow of the American Physical Society (APS), a senior member of The Institute of Electrical and Electronics Engineers (IEEE), a member of the Optical Society of America (OSA) and the NY Academy of Sciences.

https://researcher.watson.ibm.com/researcher/view.php?person=us-hendrikh

AFTERNOON KEYNOTE SPEAKER

Lauri Goldkind, PhD

Associate Professor
Graduate School of Social Service
Fordham University

Dr. Dr. Goldkind has a longstanding interest and practice background in nonprofit leadership, capacity building, and organizational development. At Fordham she teaches across the foundation and advanced years. Her practice experience has been centered in the youth development, education, and juvenile justice realms.

Prior to joining the faculty, she served as the Director of New School Development and the Director of Evaluation at The Urban Assembly (UA), a network of new specialized public schools located in the Brooklyn, the Bronx and Manhattan. At UA she supported principals through the new school process, helping them earn start-up grants valued at over $500,000 per school; additionally, she provided technical assistance to principals and school-based staff on data-driven decision making, development and maintenance of data management structures and the effective use of data to improve student achievement. She has had the privilege of working with youth in NYC at organizations such as CASES, the Posse Foundation and the DOME Project.

Dr. Goldkind holds an M.S.W. from SUNY Stony Brook with a concentration in planning, administration, and research and a PhD from the Wurzweiler School of Social Work at Yeshiva University

https://www.laurigoldkind.net/

TECH DEMO

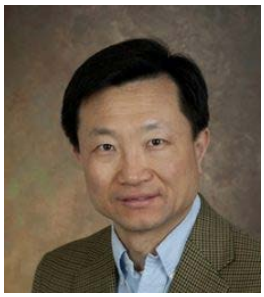
Marvin Andujar, PhD

Assistant Professor of Computer Science
University of South Florida

Dr. Marvin Andujar (Dr. A) is an Assistant Professor and Director of the Neuro-Machine Interaction Lab in the Department of Computer Science and Engineering at the University of South Florida. He received his PhD in Human-Centered Computing from University of Florida. During his Ph.D. studies, Dr. Andujar was recognized as a National Science Foundation Graduate Research Fellow, a GEM Fellow, a Generation's Google Scholar, and an Intel Scholar. His research concentration is on Affective Brain-Computer Interfaces where he focuses on measuring and decoding the user's affective state from the brain during human-machine interaction. His dissertation work focused on measuring the effectiveness of quantified-self attention feedback from the brain towards user's attention improvement. Dr. Andujar was the researcher who started the Brain-Computer Interface initiative in the Computer \& Information Science \& Engineering Department at the University of Florida. His effort led towards multiple publications in journals and conferences, obtain external funding of $300,000 from the CEO of Intel along with his colleagues, and co-founded the world's first Brain-Drone Race. The race was featured in more than 550 media news outlets including New York Times, Associated Press, Discovery Channel, the Verge, and Engadget. He is a member of ACM, IEEE Computers, and the Brain-Computer Interface Society.

https://www.marvinandujar.com/

EDUCATION PANELISTS



Bintong Chen, PhD *(Moderator)*

Professor of Operations Management
Director - Institute for Financial Services Analytics
Faculty Director - Financial Services Analytics Ph.D. Program
University of Delaware

Dr. Bintong Chen's cross-disciplinary expertise includes knowledge of both business management practices and systems engineering. He has consulted for companies in diverse industries including: fashion (Nordstrom, Inc.); utilities (Cascade Natural Gas Corporation); agriculture (North West Dairy Farm); telecommunications (AT&T) and transportation (Burlington North Railroad). His consulting work centers on improving management of inventory and customer service, improving processes and efficient scheduling (arrivals and departures of trains). He also studies game theory. Dr. Chen received his PhD in Decision Science from the University of Pennsylvania.

https://lerner.udel.edu/faculty-staff-directory/bintong-chen/

Marvin Andujar, PhD

Assistant Professor of Computer Science
Florida State University

Dr. Marvin Andujar (Dr. A) is an Assistant Professor and Director of the Neuro-Machine Interaction Lab in the Department of Computer Science and Engineering at the University of South Florida. He received his PhD in Human-Centered Computing from University of Florida. During his Ph.D. studies, Dr. Andujar was recognized as a National Science Foundation Graduate Research Fellow, a GEM Fellow, a Generation's Google Scholar, and an Intel Scholar. His research concentration is on Affective Brain-Computer Interfaces where he focuses on measuring and decoding the user's affective state from the brain during human-machine interaction. His dissertation work focused on measuring the effectiveness of quantified-self attention feedback from the brain towards user's attention improvement. Dr. Andujar was the researcher who started the Brain-Computer Interface initiative in the Computer & Information Science & Engineering Department at the University of Florida. His effort led towards multiple publications in journals and conferences, obtain external funding of $300,000 from the CEO of Intel along with his colleagues, and co-founded the world's first Brain-Drone Race. The race was featured in more than 550 media news outlets including New York Times, Associated Press, Discovery Channel, the Verge, and Engadget. He is a member of ACM, IEEE Computers, and the Brain-Computer Interface Society.

https://www.marvinandujar.com/

EDUCATION PANELISTS (CON'T)

Renée Hložek, PhD

Assistant Professor of Physics
Dunlap Institute for Astronomy & Astrophysics
University of Toronto

Dr. Renee Hložek studies a variety of problems in theoretical and observational cosmology. Her research focuses on understanding what the Universe is made of, its structure and how it is changing with time. She uses data from the Atacama Cosmology Telescope and the upcoming Simons Observatory and Large Synoptic Survey Telescope in the Chilean desert. Hložek hails from South Africa, where she studied at the University of Pretoria and the University of Cape Town. She received her DPhil from the University of Oxford in 2011, where she was a Rhodes Scholar. She was a Lyman Spitzer Jr. Postdoctoral Research Fellow in the Department of Astrophysics at Princeton University and the Spitzer-Cotsen Fellow in the Princeton Society of Fellows. She is also a TED Fellow.

https://www.dunlap.utoronto.ca/dunlap-people/prof-renee-hlozek/

Wendy Chan, PhD

Assistant Professor of Education
Human Development and Quantitative Methods Division
Graduate School of Education
University of Pennsylvania

Dr. Wendy Chan is an Assistant Professor of Education in the Human Development and Quantitative Methods Division at Penn GSE. She received her Ph.D. in statistics from Northwestern University, where she was a graduate research assistant for the Institute for Policy Research. She began her career in education as a member of Teach for America, where she taught sixth- and eighth-grade mathematics in a large middle school in New York City. Her research interests include improving generalizations from experimental studies, developing methods to improve precision from small sample sizes, and experimental design.

https://scholar.gse.upenn.edu/chan

INDUSTRY PANELISTS

Ryan Harrington *(Moderator)*

Data Science & Open Data Professional
CompassRed

Ryan leads the Data Science team at CompassRed, a Delaware data analytics start-up. Always looking for ways to better lift insights and tell stories from data, Ryan became a data scientist because he was driven by an insatiable desire to solve problems and explore solutions. He constantly draws on his experience as a high school math teacher, which he uses today to help make some of the complexities of data science simple and relatable for our clients. Ryan loves to solve and explore outside of work as well. He spends most of his free time working on Open Data Delaware, a civic technology community in Delaware that he founded and organizes. When he has a free moment, he loves understanding the world through travel - from Zion to Zanzibar.

https://www.compassred.com/our-team

Ali Ahmadzadeh

Manager, Structuring & Commercial Analytics
Constellation, an Exelon Company

Ali is responsible for deal valuation/structuring of complex commodity transactions and development of quantitative models which are used in pricing, hedging, risk management and portfolio optimization in electricity, natural gas and renewable markets. Ali received his M.S. and Ph.D. in Electrical and Systems Engineering from the University of Pennsylvania.

https://www.msds.udel.edu/people/leadership/detail/uql20/

INDUSTRY PANELISTS (CON'T)

Claudine Jurkovitz, PhD

Senior Physician Scientist, Christiana Care Value Institute
Lead BERD Core with DE ACCEL-CTR
Director, CRSN with DE-INBRE

Dr. Jurkovitz is Senior Physician Scientist in the Christiana Care Value Institute where she leads the Data Analytics core. As such she helps Physicians, Residents and young Investigators to develop their research projects and analytical plan and to work with our Data Analytics team. She also leads the Biostatistic Epidemiology Research Design (BERD) core of the Delaware ACCEL Center for Translational Research (CTR). The BERD was founded in 2013 and is composed of a group of highly skilled experts in Biostatistics, Bioinformatics, and Epidemiology. In 2018 she was appointed Director of the Centralized Research Support Network (CRSN), a newly developed INBRE core, which goal is to develop mechanisms to leverage existing infrastructure such as the Delaware ACCEL-CTR BERD and to make services provided by the BERD available to the INBRE network's biomedical investigators. She is also a member of the steering committee of the INBRE Bioinformatics Network of Delaware (BiND). As a Nephrologist Epidemiologist, Dr. Jurkovitz has actively developed her own research interests, mostly in the field of chronic kidney disease (CKD) and health services research. She was the PI of the ACCEL-CTR funded Big Data Pilot grant: "Linking Data for Kidney Care", which goal was to build a CKD registry in the State of Delaware to predict hospitalizations in patients with CKD. This project has been completed in May 2017 and resulted in several publications. She completed in May 2018 a project funded by an Eugene Washington PCORI award "Engaging Stakeholders for a Patient-Centered Research Agenda for Chronic Kidney Disease in Delaware", which goal was to build a network of patients that can provide guidance to researchers in the field of kidney disease.

https://research.christianacare.org/valueinstitute/people/claudine-t-jurkovitz-m-d-mph/

Brian Jelenek

Executive Director
JP Morgan Chase

https://www.linkedin.com/in/brianjelenek/

FACULTY TALKS: SESSION 1

1. **Jing Gao, Assistant Professor, Geography (CEOE)**
   Abstract: I plan to present my recent work simulating global spatiotemporal patterns of urban land expansion over the 21st century, using best available global data sets (quantitative and qualitative), including a 40-year time series of fine-spatial-resolution urban land observations.
2. **Jeff Buler, Associate Professor, Entomology & Wildlife Ecology (CANR)**
   Abstract: The USA network of weather surveillance radars have been actively archiving digital data of meteorological and biological phenomena in the airspace since the mid 1990's. At >3 petabytes in size, this archive is arguably one of the largest remotely-sensed biological datasets in existence. My lab has pioneered radar data processing methods to comprehensively quantify animal distributions as they leave terrestrial habitats for migration and provide novel insight into their terrestrial ecology at multiple scales and across large regions. I will present recent and ongoing studies in my lab that elucidate how anthropogenic modification of the terrestrial landscape (e.g., artificial light at night and land use) and climate (e.g., hurricanes and general wind patterns) are changing the terrestrial distributions and populations of flying animals over various time scales. Future automation of our methods by data scientists at UD would be a great advance for the field of aeroecology.
3. **Melinda Duncan, Professor, Biological Sciences (CAS)**
   Abstract: Cataract was the most prevalent cause of human blindness until the development of surgical therapies. However, not all cells can be removed, and these undergo a wound healing response leading to the deposition of a mixture of opaque scar tissue and dysgenic lens fiber cells in the visual axis. The resulting visual loss is defined as posterior capsular opacification (PCO), the major complication of cataract surgery, while other complications may also occur including glaucoma and the exacerbation of macular degeneration. However, the population wide risks of cataract surgery are unknown, a gap that has the potential to be filled by using Medicare data to determine the age stratified rates of cataract surgery and its acute post-surgical side effects, the age dependent risk of developing posterior capsular opacification, and the risk of developing vision-compromising sequelae.
4. **Kyle Davis, Assistant Professor, Geography (CEOE)**
   Abstract: Tropical forests are vital for global biodiversity, carbon storage, and local livelihoods, yet ongoing globalization has meant that remote factors exercise growing influence over forests and their resources. Large-scale land acquisitions (LSLAs) by multinational investors have emerged as an important mechanism linking the demands of distant actors to forests in the Global South. Yet it remains unclear if LSLAs have preferentially targeted forested areas or influenced rates of forest loss. We perform a global assessment of the effects of LSLAs on forests by combining georeferenced maps of individual land deals with satellite data on annual forest cover. We find that LSLAs were granted in areas of higher forest cover in 81% of cases and enhanced forest loss in 45% of cases. We observe no significant influence in all other instances. The disruptive nature of these deals will likely produce substantial impacts on forest systems and the availability of their resources for local communities.
5. **John Gizis, Professor, Physics & Astronomy (CAS)**
   Abstract: The LSST project will measure the properties of 40 billion stars and galaxies. I discuss how the LSST Stars, Milky Way & Local Volume collaboration is working on data science solutions to solve key problems in understanding the history and formation of our Galaxy.
6. **Andreas Muenchow, Professor, Marine Science (CEOE)**
   Abstract: Greenland impacts climate on a global scale, because it modulates the ocean, atmosphere, and cryosphere. Warm deep ocean waters move heat towards glaciers that extend deep into the water. These glaciers are melted by and thus add fresher waters into the oceans. The melt raises sea level and drives a more intense ocean circulation. Time scales range from hours to millennia as tides, weather, and orbital cycles all change rates of mass and heat exchanges. New sensor technologies, platforms, and modeling applied to Greenland all generate data that overwhelms most investigators: several ships operate around Greenland every year while satellites sense surface features at 15 meter scales every day. Aircraft probe ice with laser and radar while dropping sensors over the open water to profile the ocean. Merging these diverse data, I find challenging as the synthesis requires manual coding labor to just bring the diverse data sets into a common frame. I dream of a code that writes such code.

7. **Teomara Rutherford, Assistant Professor, School of Education (CEHD)**
   Abstract: Within this talk I give an overview of my program of research focused on understanding student decision-making within digital learning contexts. In particular, I focus on my NSF-funded work studying the elementary mathematics software, Spatial Temporal (ST) Math. My prior work on ST Math has involved educational data mining and more traditional social science research methods, both involving large-scale data. Upcoming work includes the embedding of experience sampling toward understanding factors that guide decisions in-the-moment. This work will be placed in the context of a research program examining the potential of digital tools for intervention and for understanding learning and motivational processes.
8. **Christopher Rasmussen, Associate Professor, Computer & Information Sciences (COE)**
   Abstract: How many sea scallops in the sea? Using Deep Learning and Underwater Robots to understand the health and distribution of the sea scallop fishery. In this talk we detail our efforts to improve the management of underwater resources. Utilizing autonomous underwater vehicles as a platform to collect color images of the seabed present a new found challenge, how to analyze hundreds of thousands of images. With a combination of human annotation and deep learning using the YOLOv3 Convolutional Neural Network we present resent results and outline additional data challenges for expanding the research.
9. **A.R. Siders, Assistant Professor, Disaster Research Center (CAS)**
   Abstract: Computational text analysis enables text-based research at scales not possible with manual methods. To illustrate the benefits of computational text analysis for social or environmental studies, I present a project to assess the adaptive capacity of social systems to climate change. Research on adaptive capacity has focused on identifying the traits and conditions that make actors or systems able to flexibly respond to change. By focusing on the relationships between traits and conditions at within-sentence and across-field scales, we can improve our understanding of the causal pathways that increase capacity. Using collocation and network analysis (centrality, criticality, and betweenness), we can analyze the existing literature (n=275) to develop a function-based model of adaptive capacity (Adaptive Capacities Framework) as well as a new theory for how capacity is built that has implications for adaptation, development, and risk reduction investments.
10. **Nina David, Assistant Professor, Public Policy& Administration (CAS) (Biden School)**
    Abstract: It is generally difficult to envision changes to the built environment. This is why planners produce renderings, maps, and other visualizations when they ask the public to think of planning related changes to the spaces around them. What if we could do better? If space indeed is the planner's laboratory, then, instead of simulated changes, could planners actually experiment in the field? Could national models like "better block" and other tactical urbanism projects be used to allow the public to experience changes to the built environment in an immersive sense? And what if we could capture peoples' interaction with the built environment on camera? I propose to present findings and ideas from a Better Block project that I implemented in Wilmington DE in 2014 together with West Side Grows Together, a coalition of residents, businesses, and community organizations in the west side of Wilmington. This project presents opportunities for a discussion of all of the above-mentioned ideas.
11. **Ali Poorani, Associate Professor, Lerner (BE) & Bill Sullivan, Managing Director of Courtyard by Marriott at the UD**
    Abstract: HR Analytics: Big Data and Financial Return: A Case of Courtyard Marriott. Big Data and related analytics are revolutionizing the way we see and process the world. McMillan (2016) stressed that use of analytics and big data proposes implementing a system of accountability, including measuring success with analytics and big data, and delivering value that will be credible to top executives. In this case study we combine Big Data and human capital analytics, more specifically Human Capital Return on Investment (ROI), Productivity Index, and Profit Sensitivity at the Newark Courtyard Marriott Hotel for a 3-year period; and, investigate if such analysis adds new outlooks beyond the usual metrics used by lodging enterprises. The data, in this single property, provided strategic level understanding of the use of Big Data. The study has implications to other properties across the board, as well as other industries.
12. **Vince D'Amico, Adjunct Faculty, Department of Entomology and Wildlife Ecology (CANR)**
    Abstract: Forests in the Present Day BosWash Megalopolis. In 1961 Jean Gottman published Megalopolis (Gottmann 1961), a treatise on the densely populated counties between Boston MA and Washington DC on the eastern coast of North America. The book included a chapter on Megalopolitan forests describing their past, present, and future prospects. Almost 60 years have passed since Megalopolis was published, and the trend toward reforestation has reversed as urban centers have gained favor with the generations following the post-WWII "Baby Boomers". We present our analysis of modern Megalopolitan temperate deciduous forests of 1 ha and larger: there are currently approximately 250,0000. By including census data, we show that, unsurprisingly, for most citizens in the Megalopolis the nearest forest is 5 ha or smaller, and 97.5% of all temperate deciduous forests are under 50 ha.

13. **Karen Hoober, Associate Director, Graduate Education & Outreach, The Center of Bioinformatics and Computational Biology (CBCB)**

Abstract: The Center of Bioinformatics and Computational Biology (CBCB) academic initiatives build on the core curriculum of Bioinformatics Data Science (BDS). CBCB offers degrees including: Bioinformatics Master, Graduate Certificate in Bioinformatics and PhD in Bioinformatics & Systems (now BDS), Online Certificates in Applied Bioinformatics and Biomedical Informatics and DS. Each initiative trains the next-generation of researchers/professionals playing key roles in multi- and inter-disciplinary teams, bridging life- and computational-sciences. Experts in the CBCB fields are housed in several Colleges: Engineering, Arts&Sciences, Agriculture&Natural Resources, Health Sciences and Earth, Ocean&Environment: the BDS degrees are university-wide interdisciplinary programs with emphasis on professional skills and immersive Internships opportunities (e.g., Christiana Care, Delaware Health Information Network, and DSU), preparing graduates for careers in industry, government agencies, or non-profits.

FACULTY TALKS: SESSION 2

1. **Behnam Abasht, Associate Professor, Animal and Food Sciences (CANR)**
   Abstract: Differential abundance of allelic transcripts in a diploid organism is referred to as allele specific expression (ASE) or allelic bias, which may caused by cis-regulatory mutations or epigenetic modifications that influence gene expression. ASE can be examined using single nucleotide polymorphisms (SNPs) from RNA-seq data. We performed ASE analysis using our custom analysis software called VCF ASE Detection Tool (VADT), which detects ASE using a 2-dimensional test. On average ~174,000 SNPs in each tissue passed our filtering criteria and were considered informative, of which ~24,000 (~14%) showing ASE. The overlap of ASE SNPs among the 3 tissues was only 3.7%, with ~83% of SNPs showing tissue specificity, but if ASE genes (genes containing ASE SNPs) were compared between the tissues, tissue overlap increases to 20.1%. Overall, we found that ASE genes show enrichment for tissue specificity, but all three tissues showed enrichment for pathways involved in mRNA translation to protein.
2. **David R. Legates, Professor, Geography (CEOE)**
   Abstract: The High-Resolution Weather Data System is a real-time meteorological data system that applies advanced Doppler radar technology and spatial interpolation methodologies to enhance water resource management by providing an improved estimate and spatial distribution of rainfall and other weather variables over a region or watershed. Developed for Duke Energy, this system was designed to ingest real-time data from radar and surface weather stations and provide spatially-distributed estimates of various meteorological fields at an hourly time-step. Sophisticated spatial interpolation methodology which includes topography and other spatial gradients is employed in the representation of surface fields such as air temperature and humidity. Vector data of wind speed and direction also are well-represented.
3. **Medina Jackson-Browne, Assistant Professor, Epidemiology (CHS)**
   Abstract: Early childhood exposure to perfluorooctanoic acid (PFOA), perfluorooctane sulfonic acid (PFOS), perfluorononanoic acid (PFNA), and perfluorohexane sulfonic acid (PFHxS) may affect the immune system to increase the risk of allergic and respiratory diseases. Thus, we examined the cross-sectional associations of serum PFOA, PFOS, PFNA, and PFHxS concentrations with childhood asthma. We used data from 607 US children aged 3-11 years who participated in the National Health and Nutrition Examination Survey (2013-2014). After covariate adjustment, PFOA (1.07; 95% CI: 0.85, 1.34), PFOS (1.15; 95% CI: 0.8, 1.65), PFNA (1.12; 95% CI: 0.76, 1.66), and PFHxS (1.08; 95% CI: 0.89, 1.34) were weakly associated with increased odds of asthma. In this cross-sectional study, we observed no evidence that serum PFAS concentrations were weakly associated with increased asthma prevalence in US children.
4. **Li Liao, Associate Professor, Computer & Information Sciences (EG)**
   Abstract: Comorbidity of diseases presents great challenges to accurate diagnosis and treatment. To help better understand the genetic causes of comorbidity, we have developed computational methods to predict comorbid diseases and detect missing common genes for comorbid diseases. Two diseases sharing common genes tend to increase their comorbidity. We formulate the search for missing common genes as an optimization problem to minimize network based module separation from two subgraphs produced by mapping genes associated with disease onto the interactome. To predict whether two diseases are comorbid, we embed the interactome into a high dimensional geometric space with weights assigned to the network edges and uses the projection onto different dimension to “fingerprint” disease modules to discriminate comorbid diseases versus non-comorbid diseases. Remarkable performance is reported when the methods are tested in cross-validation with a benchmark dataset of more than 600 disease pairs.
5. **Austin Cory Bart, Assistant Professor, Computer & Information Sciences (EG)**
   Abstract: Instructional design is a systematic and iterative process that inherently encourages evaluation and revision. However, this kind of tracking can be challenging in complex, dense courses - for example, my introductory Computer Science curriculum involves hundreds of interactive experiences distributed throughout the semester. In this talk, I will describe the ways I collect data about my learners and how I use this data to revise my curricular materials. The protocols and formats that I use come from emerging standards in Computer Science Education and highlight the growing consensus on the critical need to understand our classrooms more clearly. Modeling student knowledge, motivation, and self-regulation is challenging but increasingly important as we seek to tie curricular design decisions to outcomes. Beyond my descriptions, I will also highlight opportunities that I see for future research projects in Learning Analytics and Data-Driven Instructional Design.

6. **Rodrigo Vargas, Associate Professor, Plant & Soil Sciences (CANR)**
   Abstract: Data science of SPAM: generating knowledge across the Soil-Plant-AtmMosphere continuum
   With the exponential growth of satellite-derived information, environmental sensor networks, and national-to-global environmental observatory networks, we have entered an era of unprecedented information related to Ecoinformatics. This talk will summarize recent advancements and applications of data science across the soil-plant-atmosphere continuum from local-to-global scales. Examples include forest cover dynamics, soil moisture patterns, soil organic carbon and nitrogen estimates, greenhouse gas emissions to the atmosphere, and design of environmental networks. This presentation will end describing how knowledge from new information is being applied for policy relevant research from national-to-global scales.
7. **Cencheng Shen, Assistant Professor, Applied Economics and Statistics (CANR)**
   Abstract: Understanding and developing a correlation measure that can detect general dependencies is not only imperative to statistics and machine learning, but also crucial to general scientific discovery in the big data age. In this talk I will discuss recent advances and development on using distance-based correlation for dependency discovery. It is easy and intuitive to use, achieves universal consistent testing against all relationships, very powerful in small dataset with nonlinear and high-dimensional dependency, and can be fast and scalable on big data.
8. **Xiao Fang, Professor, Management Information Systems (BE)**
   Abstract: Machine Learning for Social and Business Good: From Social Network Analytics to Financial Technology
   In this talk, I give a overview of my group's research on social network network analytics and financial technology. Specifically, we focus on a unique phenomenon in social networks – the diffusion of contents, information, and adoption behaviors from one social entity to another. In particular, we investigate three critical and related problems concerning this phenomenon: adoption, persuasion, and link recommendation. We are also interested in designing state-of-art analytical tools to solve fundamental problems of Financial Technology (FinTech). One project aims at developing a better industry classification system by designing a novel deep learning algorithm to analyze firms' annual 10K reports.
9. **Rudi Eigenmann, Professor, Computer Engineering (EG)**
   Abstract: Cyberinfrastructure - critical enabler of data-intensive and computational science: this talk will give a brief overview of hardware, software, data, and people infrastructure enabling DSI projects, including a recent NSF MRI award for a new compute and storage cluster, and plans to create a research software engineering team in support of data-intensive and computational research.
10. **Stan Owocki, Professor, Physics & Astronomy (CAS)**
    Abstract: Hot, massive, luminous stars show evidence of eruptive "Luminous Blue Variable" phases, characterized sudden ejections of up to solar mass or more. This talk will discuss how such eruptions represent a kind of bridge between steady mass loss in stellar winds, and supernovae explosions that completely destroy the star. One key goal of large data surveys in astronomy is to identify such "supernova imposters" and distinguish them from actual supernovae. An overall aim here is to connect theoretical models with observational properties for making this distinction and so help clarify the role of such eruptive mass loss in massive-star evolution.
11. **Randy Wisser, Associate Professor, Plant & Soil Sciences (CANR)**
    Abstract: Exotic populations of crops harbor unique adaptations and beneficial alleles for agricultural improvement, but they tend to be poorly adapted to target production environments. In order to capitalize on this broader array of diversity and mine alleles from exotics, a new plant breeding method is being developed for rapid environmental adaptation. Using maize as a target crop and model system, this effort builds off of unique germplasm resources, emerging knowledge on the genetic basis of flowering phenology and advances in genotyping and genomic selection methodology. It is anticipated that this will open a new pathway for capitalizing on diverse germplasm collections and lead to new developments in genomic-assisted plant breeding.
12. **Tracy DeLiberty, Associate Professor, Geography (CEOE)**
    Abstract: Incorporating cover crops into agricultural production has long been recognized as a management practice that could reduce not only soil erosion but also the leaching of nutrients, and as a consequence, government agencies and research groups have long promoted its adoption. Currently, nationwide field surveys are based on farmer's responses and statewide dashboard surveys which both have some drawbacks. The use of remote sensing is investigated to detect cover crop fields based on monthly average vegetation indices during selected months that are spectrally different from summer crops and native vegetation. Our model is tested in Google Earth Engine to leverage the use of Landsat imagery over the last 10 years with the results compared to ground truth data where cover crop adoption is known.

13. **Harvey White, Professor, Public Policy& Administration (CAS) (Biden school)**
Abstract: "Equitable Use of data" for the delivery of public services. Big data, AI and emanating automation, and robotics have the potential to revolutionize governance at all levels. Data technologies will empower citizens to indicate their service preferences in real-time and allow governments to tailor service delivery to both community and individual needs. In the process, public services can be provided more effectively and economically. The automation of routine activities can save billions of dollars through increased efficiencies and cost reductions. New jobs will be created and many current jobs eliminated. This, in turn, will affect how and for what jobs public servants are trained. The unfolding revolution in governance has major implications for equity and ethics in the delivery of public services, which will be addressed including a discussion of moral and social justice issues surrounding projected disparities, data biases, and accountability for automated decision making.
14. **Jodi Hadden-Perilla, Assistant Professor, Department of Chemistry & Biochemistry**
Abstract: Hepatitis B virus (HBV) is a major cause of liver disease, for which there is no cure. The HBV capsid, an icosahedral protein shell that encases the viral genome, plays an essential role in HBV infection. Drug molecules known as capsid protein allosteric modulators (CpAMs) disrupt capsid assembly in vitro to prevent packaging of the genome. Here, all-atom molecular dynamics simulations of intact, drug-bound capsids are applied to investigate the mechanisms by which two classes of CpAMs disrupt capsid structure, dynamics, and assembly. The capsid's response to drug-induced perturbation reveals new details regarding its biophysical properties and biological function. Microsecond simulations of the capsid in its native environment at chemical resolution enable characterization of drug binding modes far beyond what has been achieved through experimentally-determined structures. The results of this study are relevant to the development of new therapeutics that target the capsid of HBV.

STUDENT POSTER SESSION 1

1. **Jamie Doyle, Undergraduate Student (CHS)***
   Abstract: A secondary data analysis was conducted using the 2018 Health Information National Trends Survey (HINTS). This research explored how perceived quality of healthcare and information-seeking behaviors are associated with knowledge about the Human Papillomavirus and vaccination behavior. Our subset consisted of 671 females between the ages of 18-49. We had three hypotheses and two research questions.
2. **Rita Rawal, PhD Student, Bioinformatics and Systems Biology (CBCB)***
   Abstract: Dietary diversity (DD) is recognized key component of a healthful diet. This study determined DD scores for count, evenness, and dissimilarity across adulthood in a diverse sample. Participants were from the Healthy Aging in Neighborhoods of Diversity across the Life Span study, a longitudinal NIA study, which included 3,720 African American and White adults. DD measures were calculated using 2 24-hr recalls for 3 study waves (2004-2017). The count was based on consumption of ≤50% of an equivalent of food from 21 subgroups. Evenness was derived using Berry-Index adjusted by health value of food; dissimilarity, by Mahalanobis Distance. Two sample t-test was used to analyze difference between groups for each wave. Only count and dissimilarity scores significantly differ by sex and race ($p<0.001$). All three DD measures were statistically different between income groups (>125% vs <125% poverty). White women and persons with higher income had better DD.
3. **Haobai Zhang, PhD Student, School of Education (CEHD)***
   Abstract: Practitioners of finite mixture models often have difficulty selecting the correct number of classes due to conflicting fit indices. Recent empirical and simulation studies advised cross validation as an alternative way to measure the accuracy of class enumeration with growth mixture models, because it iteratively tests models while controlling for Type I error rate. Such studies are limited to hold-out and k-fold cross validation and conversely suggests that the k-fold approach only performs well when classes have large separation. Our current project expands this area of research to four types of class enumeration methods: k-fold, leave-out-one, hold out, and bootstrap. We also compare the performance of cross validation and bootstrap methods on two mixture models, latent class and latent profile analyses (LCA and LPA).
4. **Hongzhe Zhang, PhD Student, Graduate Interdisciplinary Program***
   Abstract: Predictive analytics, an essential tool for firms to survive and thrive in competitive marketplaces, learns models from historical data to predict the future. Most predictive analytics methods assume that entities are isolated. However, in reality, entities are connected through various social networks. As a result, the behaviors of an entity's network neighbors could be predictive of the entity's behavior. Therefore, an effective predictive analytics method should leverage both intrinsic and network characteristics of entities. In our study, we address this challenge methodologically by proposing a novel latent network-enhanced predictive analytics method that considers both observable intrinsic characteristics of entities and latent (unobservable) network ties among them. Our method predicts an entity's behavior using both her intrinsic characteristics and the behaviors of her network neighbors identified through latent network ties.
5. **Keshab Subedi, Graduate Certificate, Agriculture and Natural Resources (CANR)***
   Abstract: Project 'Engage ED' is a program using embedded Peer Engagement Specialists (PES) to connect Emergency Department (ED) patients with substance use disorder (SUD) to community-based treatment. This study evaluates the effect of project 'Engage ED' on length of stay (LOS) and one year ED readmission. For analysis we created control group using 1:1 propensity match. We used generalized linear Gamma models to evaluate the PES intervention's impact on index visit LOS, and negative binomial models to evaluate the PES impact on one year ED readmission. Compared to controls, patients with an opioid use disorder who were successfully connected to community SUD treatment experienced an increase in the EDLOS but a significant reduction of ED visits whereas patients who were provided information had reduced EDLOS but the number of ED visits did not significantly decrease.
6. **Peng Su, PhD Student, Computer & Information Sciences (EG)***
   Abstract: Significant progress has been made in applying deep learning on natural language processing tasks recently. However, deep learning models typically require a large amount of annotated training data while often only small labeled datasets are available for tasks in biomedical domain. Building large-size datasets for deep learning is expensive since it involves considerable human effort and usually requires domain expertise in specialized fields. In this work, we consider augmenting manually annotated data with large amounts of data using distant supervision to help model generalization. However, data obtained by distant supervision is often noisy, we first apply some heuristics to remove some of the incorrect annotations. Then, we explore the methods inspired from transfer learning to combine the data from distant supervision and human-labeled data, we show that the resulting models outperform previous models

7. **Tanya Nesterova, Undergraduate Student, Chemistry & Biochemistry (CAS)***
   Abstract: The human immunodeficiency virus, or HIV, causes a viral illness that suppresses cells that are vital for the human immune response . Without proper treatment, HIV will develop into AIDS, or autoimmune deficiency syndrome, which leads to death by fully suppressing the immune responses of its host. The HIV virus is a complex system of which the main components are the viral envelope, capsid, and RNA. Unlike other viruses, the HIV capsid follows an irregular asymmetrical shape. This conical shape varies and has frequent mutations and inconsistencies. To collapse the capsid prior to infection, we must understand its geometric properties and any potential patterns in structure. However, the missing wedge effect from CryoET makes it difficult to fully characterie the structure of HIV capsids, which cannot be analyzed by single particle analysis. We propose a mathematical solution that heals the missing wedge effect from CryoET through a C++ source code that utilizes Poisson Disk Sampling.
8. **Rayanne Luke, PhD Student, Mathematical Sciences (CAS)***
   Abstract: The tear film (TF) protects the ocular surface and promotes clear vision. Tear film breakup (TBU) occurs when a dry spot forms on the eye. Elevated osmolarity in TBU is likely a key factor in developing dry eye syndrome (DES). We infer TBU events from video recordings of the TF imaged using fluorescein (FL) dye, which glows green under blue light; physical and chemical quantities of interest cannot be measured directly at this time. We build and analyze nonlinear PDE models for TF height, osmolarity and FL concentration for evaporation-driven TBU. Theoretical FL intensity is fit via nonlinear least squares optimization to circular and linear TBU instances from experimental imaging data gathered from normal subjects' TFs to estimate parameters such as TF thinning rates. Our estimates for variables that cannot be measured in vivo during TBU fall within accepted experimental ranges, help clarify the mechanisms for TBU, and help medical professionals better understand TF function and DES.
9. **Xin Guo, PhD Student, Department of Electrical and Computer Engineering (EG)***
   Abstract: This paper presents a hybrid deep learning network submitted to the 6th Emotion Recognition in the Wild (EmotiW 2018) Grand Challenge [9], in the category of group-level emotion recognition. Advanced deep learning models trained individually on faces, scenes, skeletons and salient regions using visual attention mechanisms are fused to classify the emotion of a group of people in an image as positive, neutral or negative. Experimental results show that the proposed hybrid network achieves 78.98% and 68.08% classification accuracy on the validation and testing sets, respectively. These results outperform the baseline of 64% and 61%, and achieved the first place in the challenge.
10. **Colleen Mueller, Undergraduate Student, Public Policy& Administration (CAS) (Biden School)***
    Abstract: After the federal ban on TV, radio, and billboard advertising in 1998, the tobacco industry has focused on point-of-sale (POS) advertising at locations like convenience stores, gas stations, and local store fronts. While there has been extensive national research on the impact of tobacco marketing and likelihood of future tobacco use, there is a need for more research at a level local to Delaware. The goal of this research project is to look for correlations between locations of POS tobacco advertising, use of tobacco products, and assessment of overall health. The first step was to create a map identifying all locations where business retailers are licensed to sell tobacco within the city of Wilmington. The next step was to collect images of tobacco advertising and overlay that data on the locations of all licensed tobacco retailers. The future aim of the study is to determine how the exposure to POS tobacco advertising impacts public health trends of community members in Wilmington.
11. **Menolin Sharma, PhD Student, Agriculture and Natural Resources (CANR)***
    Abstract: Bacteriophages are predicted to be the most numerous biological organisms on earth. Yet, tracking their diversity is difficult due to a lack of marker gene in phages like 16S genes in prokaryotes. Similarly, difficulty in isolating and culturing viruses have made their phenotypic studies difficult. More than 25% of dsDNA viruses are shown to have DNA polymerase genes. Given the important role of DNA polymerase enzymes in viral replication, identifying viral DNA polymerase genes and their signature patterns will help us know more about their diversity as well as phenotypic featues. In the following study, viral DNA polymerase sequences from two viral metagenome datasets, GOV and SERC are mined. We have tried to identify differential residues and patterns at equivalent positions between viral and bacterial polymerases, and also among viruses. We aim to identify links between the patterns and phenotypic differences between viruses and bacteria as well as within viruses.
12. **Maxim Bazik, PhD Student, Computer & Information Sciences (EG)**
    Abstract: An efficient, fully automatic method for 3D face shape and pose estimation in unconstrained 2D imagery is presented. The proposed method jointly estimates a dense set of 3D landmarks and facial geometry using a single pass of a modified version of the popular "U-Net" neural network architecture. Additionally, we propose a method for directly estimating a set of 3D Morphable Model (3DMM) parameters, using the estimated 3D landmarks and geometry as constraints in a simple linear system. Qualitative modeling results are presented, as well as quantitative evaluation of predicted 3D face landmarks in unconstrained video sequences.

13. **Agnijit DasGupta, PhD Student, Computer & Information Sciences (EG)**
Abstract: It is known that disparity maps are generated by matching similar features in stereo image pairs. The quality of the stereo analysis depends on the quality of these features. Because ice images have very low texture, it is challenging to find and match such features. In this work, we hypothesize that a group of pixels in an image originates from different but unknown distributions. Gaussian Mixture Model on different correspondence metrics is used to identify and classify the aforementioned distributions. Multiple algorithms are tested to determine which produces better analysis for these groups. We analyze the most effective stereo measures however, none of those individually are capable of performing the stereo analysis of ice images accurately. For each stereo pair, four state-of-the-art stereo generating algorithms from our study are finally applied. The winning algorithm is recorded for each group based on the least RMS error. The results show that our stereo matching is better than just using one algorithm for the entire stereo pair.
14. **Riza Li, PhD Student, Computer & Information Sciences (EG)**
Abstract: From 1999-to-2017 the US obesity rate increased by 29.8% while Delaware's obesity rate mushroomed by 81.7%. The 1999-2017 Center for Disease Control (CDC) death records disclosed a 60.53% increase in crude Delawarean mortality rate when obesity was listed as a single underlying cause of death. However, when any mention of obesity was documented on the death certificate, Delaware's crude mortality rate rose by 75.69% and its age-adjusted rate increased by 53.18%. Diabetes, circulatory system diseases, and neoplasms (cancer), are three common obesity comorbidities. For these three conditions, Delaware's 1999-2017 mortality rate figures mimic the falling national patterns of mortality rate averages, when each disease is listed as the single underlying cause of death, including observations where there are disproportionate numbers of cases that affect the African American/Black race.
15. **Xiaolong Li, PhD Student, Physics & Astronomy (CAS)**
Abstract: How do you prepare to discover the unknown? Scheduled to begin in 2022, the Large Synoptic Survey Telescope (LSST), with its unique survey capability, will have the potential to make unexpected discoveries. Facilities and data reduction pipelines are being finalized to process 20Tb of LSST images each night in real time. But how can we assure that the LSST observing strategy under design do not prevent the discovery of the unknown? The operations of LSST are evaluated by metrics under the Metric Analysis Framework (MAF) API. Although many metrics have been designed, no metric is designed to evaluate LSST's ability to discover unknown phenomena. We present such a metric by mapping LSST planned observations to a phase space defined by the brightness, color and the magnitude change. Based on the distribution in phase space we will be able to tell which regions support detection and which do not. Our results are expected to provide improvement on LSST's observation strategies.
16. **Pinar Selimoglu, Undergraduate Student, Physics & Astronomy (CAS)**
Abstract: Observational and theoretical work has now established that the fossil fields of magnetic massive stars are surviving remnants from an earlier event, or an earlier evolutionary phase. However, many important questions remain regarding the effects of these fields on the late stages of stellar evolution, as well as their impact on the core-collapse mechanism and the formation of exotic compact objects such as magnetars and gravitational wave progenitors. There is currently a critical need to incorporate the impact of fossil fields in models of the structure and evolution of magnetic stars, and to determine the evolutionary history of magnetic massive stars. We present a preliminary population study of a cluster of co-evolving stars based on MESA evolutionary tracks that account for the effect of magnetic wind quenching.
17. **Melinda Kleczynski, PhD Student, Mathematical Sciences (CAS)**
Abstract: The development of quantitative methods for characterizing addiction from medical imaging poses a fundamental challenge for health professionals. Functional MRI potentially provides a global picture of how the brain's function changes due to addiction. However, the high volume of data in fMRI recordings, as well as the multifaceted nature of brain activity involved in addiction, have impeded the use of classical analytic tools. So we are applying data analysis techniques from the emerging field of applied algebraic topology. We are examining data from the National Institute on Drug Abuse consisting of functional MRI scans from rats in both nicotine exposure and control conditions. Our preliminary results suggest that identifiable changes arise as addiction develops. We are beginning to evaluate these signatures in a larger experimental sample and across various analytic parameters. We ultimately aim to provide more explicit information about the neural mechanisms driving addiction.
18. **Gregory Richards, Post Doctoral Researcher, Physics & Astronomy (CAS)**
Abstract: The universe is filled with a rich variety of particles that propagate through the cosmos at speeds close to the speed of light, and the sources of these particles remain mostly unknown. Astroparticle physics is a branch of physics and astronomy aiming to better understand these mysterious particles via studying cosmic particle accelerators and their resulting multi-messenger emission. The big, collaborative experiments in this field typically end up with huge data sets on hand (tera or petascale) which must be heavily mined given the difficulty in finding signals that are vastly overwhelmed by background contamination in the data. In this contribution, I will briefly review a few of the machine-learning methods that scientists in astroparticle physics have recently employed to better solve these types of problems over traditional algorithms.

19. **Millisen Dill, Undergraduate Student, Department of Human Development and Family Sciences (CEHD)**
Abstract: The Whole Earth Telescope is an international collaboration of astronomers dedicated to continuous observation of pulsating stars. We will present an overview of data reduction and analysis, highlighting issues in working with data from multiple observatories.
20. **Xuerong Wan, PhD Student, Department of Electrical and Computer Engineering (EG)**
Abstract: This poster presents a neural network based time series forecasting framework . We used monitoring cpu usage data from computers and servers. The purpose of this NN model is forecasting the behavior and identifying peak usage times, and making cpu usage projections. The model used sequential training method to increase the topology of the network, with part of the network has fixed memory of the short-term pattern from past training, and grows the network training with longer time series to identify long-term patterns, such as daily or weekly patterns. We will demonstrate the implementation and results of this model in this poster.
21. **Zach Schreiber, PhD Student, Bioinformatics and Systems Biology (CBCB)**
Abstract: This project aims to shed light on what unclassified metagenomic viral proteins may consist of under the assumption that gene neighbors often share complementary function. The overall concept is driven by a network based analysis on predicted peptide open reading-frames or (ORFs) that are positioned along assembled contiguous reads (contigs). By clustering these ORFs at a low sequence identity ranging between thirty and fifty percent, analysis can then be made on how often the predicted ORFs share a co-occurring, adjacent, or proximal relationship between cluster groups. Various statistical metrics such as social networking analysis (SNA) will then be used to make predictions on the characteristics of the known and unknown virome clusters.
22. **Alexander Bryer, Master Student, Chemestry and Biochemestry (CAS)**
Abstract: Two applications of convolutional neural networks to computational biochemistry and biophysics are presented. One application is DCTAP, the deep chemical shift and torsion angle predictor, which takes atomic coordinates of proteins in the ubiquitous PDB format and predicts chemical shifts – values classically obtained via Nuclear Magnetic Resonance (NMR) experiments or prohibitively expensive quantum chemistry calculations – by using the theory and framework of a neural network designed to process RGB images. The second application is a two-phase classifier, designed to automatically detect HIV-1 viral particles from experimental electron micrograph data and determine their individual maturities.
23. **Matthew Walter, PhD Student, Geography (CEOE)**
Abstract: Understanding the impact that humans have on natural environments is critical when making environmental planning and policy decisions. Wetland ecosystems are important resources, providing great economic benefits for surrounding communities. We utilize remotely sensed and geographic data for large scale mapping of wetland stress through the establishment of metrics that are commonly used to measure humans' impact on wetlands. We apply these techniques to the coastal state of Delaware to understand the spatial scale and patterns of stress to Delaware wetlands. Through our analysis we better understand how humans' impact on land cover, vegetation, and hydrology is effecting wetlands in Delaware. By providing quantifiable and visual results, our analysis can be used to identify the wetlands in Delaware that may need further protections.
24. **Xinjie Lan, PhD Student, Department of Electrical and Computer Engineering (EG)**
Abstract: We propose a Bayesian regularization learning algorithm to prevent Deep Neural Networks (DNNs) from overfitting. Above all, we propose a probabilistic representation for explaining the architecture of DNNs, which demonstrates that the hidden layers close to the input formulate prior distributions, i.e., explicit regularizations, to guarantee the generalization performance of DNNs. Moreover, we demonstrate that because of the degradation problem of DNNs, the backpropagation algorithm cannot learn these prior distributions precisely, which results in overfitting. Based on the proposed probabilistic representation, we propose a novel regularization learning algorithm for DNNs, which independently learns these prior distributions from the training dataset, thereby generating precise regularizations and preventing DNNs from overfitting. Simulations validate the superiority of the proposed regularization for image recognition on the CIFAR-10 benchmark dataset.
25. **Andrew Hill, PhD Student, Plant & Soil Sciences (CANR)**
Abstract: Within anthropogenic environments natural carbon cycle controls and drivers are modified thus altering carbon cycling. The current study presents 4.3 years of continuous soil chamber measurements from urban forest and suburban lawn within a Mid-Atlantic suburb. Carbon cycle dynamics are explored at multiple temporal scales with examination of environmental drivers. Considering all years in the study with complete data (2015-2018), findings show that urban forest floor, devoid of vegetation, is a constant carbon source generating 587 g C m-2 annually. Residential lawn intermixed with large shade trees fixed 471 g C m-2 while respiring 1,220 g C m-2 resulting in a greater annual net carbon source (749 g C m-2). The strength of environmental drivers and observed respiration rates changed dramatically depending on surface type and meteorological season.

26. Samuel Fordin, PhD Student, Physics & Astronomy (CAS)

Abstract: 25 Years of WIND: Automatic Characterization of Waves in the Solar Wind. Space weather encompasses the study of conditions in near-Earth space, with a focus on the solar wind. The solar wind is composed of highly-energetic charged particles that stream outward from the Sun, which can interfere with Earth- and space-borne electronics. The behavior of the solar wind is tied to the collective behavior of these energetic particles, often exhibiting periodic fluctuations (waves) in the measured magnetic field. The properties of these waves provide clues to the processes that shape the solar wind. Decades of solar wind observations by the WIND spacecraft has yielded a rich array of time series magnetic field data—too much to examine manually. Therefore, machine learning is a practical approach to handling the WIND dataset. To this end, a subset of WIND data will be labeled and used to train a machine learning algorithm aimed at classifying waves, with the goal of classifying the full WIND dataset.

***Denotes Participant – Student Lightning Talks Session 1**

STUDENT POSTER SESSION 2

1. **Shaquann Seadrow, PhD Student, Physics & Astronomy (CAS)****
   Abstract: Massive stars are key elements in forming galaxies as we know them. These stars chemically enrich galaxies via thermonuclear reactions, create neutron stars and black holes from their energetic deaths as supernovae, and usher in the next stellar generation from their remnants. We study the rare but exciting massive stars that manage to maintain dipolar magnetic fields at their surfaces. Surveying these stars could tell us how these magnetic fields were created as well as their impact on stellar astrophysics. Young massive stars will still be embedded in clouds of interstellar dust, making them difficult to observe in the optical wavelengths. We utilize SPIRou, a new near-infrared spectropolarimeter at the Canada-France-Hawaii Telescope, given that the dust is not opaque to starlight at these wavelengths. We examine the potential of SPIRou data to be used in measuring the magnetic fields in the infrared by using magnetic massives stars that were studied at other wavelengths.
2. **Mario Guevara, PhD Student, Dept. of Plant and Soil Sciences (CANR)****
   Abstract: Recent reviews have highlighted the potential of multiple forms of statistical learning research (i.e., machine learning, deep learning, ensemble learning) for better understanding soil distribution in space and time, and for enabling the monitoring of soil condition using digital soil mapping (DSM) across scales. We use a data-driven DSM conceptual framework as an example to demonstrate the potential of statistical learning for the optimal production of digital soil maps. From the plot to the continental scale, we present a state of the art review on statistical learning and show multiple real case prediction problems on DSM. We compare and test multiple forms of statistical learning and identify why different statistical learning methods generate different results (e.g., under which level of data availability, scale and type of spatial configuration). We provide recommendations through open source code and data sharing principles for ensuring reproducible results.
3. **Elizabeth Smith, Master Student, Behavioral Health and Nutrition (CHS)****
   Abstract: An environmental correlation approach was performed to upscale soil nitrogen concentration data across conterminous United States to provide a quantitative understanding of its spatial distribution, variation across depths, and its relationship with total nitrogen deposition and biomass. We predicted soil nitrogen concentration at 6 standardized depths and included model prediction error estimates. We used a random forest regression kriging statistical framework to explain, by the means of 10-fold cross validation, 38–42% soil nitrogen variability. We used Pearson correlation coefficient to determine the relationship of soil nitrogen with total nitrogen deposition and biomass. We found major variability and concentration of soil nitrogen at the surface, uncertainty tends to increase with depth, and relationships with biomass decrease with depth.
4. **Sachin Gavali, PhD Student, Computer & Information Sciences (EG)****
   Abstract: As a part of Delaware's effort to improve health care quality and access, the Delaware Health Information Network (DHIN) has developed a Health Care Claims Database (HCCD) as a repository for Delaware health care claims. To evaluate the feasibility of utilizing the HCCD, we have developed a dashboard to support data visualization and exploratory analyses. The dashboard provides functionalities to support use cases such as 1) Healthcare cost: amount paid by health insurance plans for specific conditions; 2) Patient demographics: patient demographics changes over time; 3) Time to an office follow-up visit after hospital discharge: a measure of health care quality; 4) Disease burden: a measure of the prevalence of multiple health conditions in the population. With this project, we have determined that the HCCD database can be a valuable resource in population research. We plan to extend the dashboard to include more use cases to facilitate data extraction leading to actionable knowledge.
5. **Shani Craighead, PhD Student, Animal and Food Sciences (CANR)****
   Abstract: Food safety policy relies on data-driven proactive approaches in hazard identification, prevention, and education rather than typical reactive mechanisms. Metadata collected at UD is analyzed and contextualized for policy development affecting produce growers across the country. Contributors to on-farm risks are manure (biological soil amendments of animal origin, BSAAO) and water; influencing pathogen survival and transfer to crops. BSAAO type affects bacterial survival in amended soils, but longer persistence does not correlate to greater transfer. Moreover, pathogenic bacteria, viruses and protozoa were detected throughout the harvest season in irrigation waters. Real-time and epidemiological data are utilized to develop interactives that enhance student engagement and critical thinking, where data clues are presented as manipulatives. This study illustrates collection, analysis, and implementation of data for developing policy and educational tools.

6. **Mark Lundine, PhD Student, Marine Science (CEOE)****
   Abstract: For over a century, the enigmatic Carolina bays have captivated geologists and spurred contentious debate on their origin. These circular to ovate and shallow depressions span the Atlantic Coastal Plain from Florida to New Jersey, with total counts ranging between 10,000 and 500,000. Using 1 meter gridded, 1.7km by 1.7km LiDAR digital elevation models of Delaware as training images, a convolutional neural network (CNN) was trained to detect the highly abundant Carolina bays. Using Faster R-CNN within the TensorFlow Python library, a network was trained on 978 LiDAR images for 24 hours (42,450 iterations) on an Intel Core i7-4790K CPU at 4.00 GHz. This network automatically detects bays from LiDAR images with a bounding box and a confidence level. This method for detecting geomorphic features is a highly efficient process that will provide better means for mapping various types of abundant geomorphic features in the future.
7. **Arshiya Khan, PhD Student, Electrical Engineering (EG)****
   Abstract: Hackers have been trying to attack machine learning based image classification system by inputting adversarial images to the system and forcing it to misclassify images. The aim of my research is to protect machine learning based applications even when there is possibility of image perturbation. The images are put through lossy compression which removes adversarial properties of images and clean images are sent to the image classifier. Different number of iteration of lossy compression leads to interesting results.
8. **Eric Allen, Master Student, Geography (CEOE)****
   Abstract: DHave you ever wanted to search National Weather Service (NWS) forecasts for case studies, particular weather events, or other specific information? Now you can with the Forecast Search Wizard (FSW). Millions of weather forecasts have been issued by the NWS, but what happens to that information, or data, once it is no longer relevant? It has long been a forgotten source of data. The FSW aims to change that by turning text into data. To my knowledge this is the only free and OS tool available. I am aware of one Silicon Valley-based Carnegie Mellon University team that sells data to end-users. One of their customers, is a team from NASA's MSFC. To use: just install, complete the two "namelist" files, and run. Supported in Python 2 & 3. I will demonstrate how the FSW can be a useful tool, explain its capabilities, and discuss its limitations. I will examine two weather events 1) Sea Breezes and 2) Coastal Flooding to show just one way that the FSW can be used to turn text into data.
9. **Harrison Smith, Master Student, Marine Science (CEOE)****
   Abstract: The United States Mid Atlantic region is subject to extensive, improving observation efforts in the form of Autonomous Underwater Vehicle (AUV) surveys, fixed profiler and buoy records, terrestrial HF radar current measurements, and remotely sensed data products from satellites. Satellites provide large spatial coverage of ocean properties over regular time intervals, but these estimates are known to diverge from in-situ measurements, especially estimates of chlorophyll. Chlorophyll is the main photosynthetic pigment in phytoplankton, which form the base of all marine food webs, and is important for estimating global net primary productivity (NPP) and carbon budgets. By using AUV optical data to produce a regionally tuned remote sensing chlorophyll algorithm, a higher quality data product can be derived and included in Mid Atlantic efforts to observe and model an area of high importance to the ocean observation and ocean color community.
10. **Farid Qamar, Master Student, Arts & Science (CAS)****
    Abstract: There is a growing need to develop a better understanding of plant health in urban environments including the impact of air quality on the health of public green spaces. This need has led to the use of Hyperspectral Imaging (HSI) as a potential rapid and non-destructive sensor to assess plant health remotely and in an automated fashion. However, due to the geometric complexity and strong diversity of materials in urban areas, separating the vegetation from other materials using remote, ground-based imagery remains a challenge. Here we examine and compare various machine learning methods performing image segmentation on ground-based hyperspectral imagery in order to distinguish the materials and make use of their spectra for further analysis.
11. **Sergio Cabrera-Cruz, PhD Student, Agriculture and Natural Resources (CANR)**
    Abstract: Urban areas affect terrestrial ecological processes, but we know little about their effect on aerial ecological processes. We evaluated the effect of urbanization on flight altitudes of nocturnally migrating birds. We predicted that altitudes would be lower over urban than over non-urban areas. We used data from nine weather surveillance radars in the eastern US to estimate altitudes at five quantiles of the vertical distribution of birds migrating at night over urban and non-urban areas. We fit Generalized Linear Mixed Models by season for each of the five quantiles of bird flight altitude. We found that birds generally fly higher over urban areas compared to rural areas in spring, and marginally higher at the mid-layers of the vertical distribution in autumn. Our results suggest that the effects of urbanization on wildlife extend into the aerosphere, stressing the need of understanding the influence of anthropogenic factors on airspace habitat.

12. **Danielle Riser, PhD Student, Department of Human Development and Family Sciences (CEHD)****
Abstract: Latent class analysis (LCA) and latent profile analysis (LPA) are finite mixture models widely used among social and behavioral scientists. While both models are used to identify meaningful subgroups in cross-sectional analyses, LCA detects patterns in categorical observed variables, whereas LPA detects subgroups in continuous observed variables. Recent examples of LCA and LPA applications include identifying post-traumatic stress disorder patterns and predictors among refugees, and splitting latent classes into trees and relating them to external variables Mplus is a latent variable modeling program, widely used for running statistical analyses and mixture models such as LCA and LPA. However, because Mplus uses a one-input one-output approach, the program can only run one model at a time. In instances such as LPA or LCA, where researchers must run and compare output from several models at a time, analyses can be burdensome and time consuming. The Mplus Automation package in R was created to automate processes and decrease some of the burden associated with running large batches of models in Mplus. The package can reduce the effort required to conduct LCA/LPA analyses by automating the process of creating Mplus syntax, running batches of models, and creating tables to compare the output from each model. However, as this package was recently created, there are few researchers familiar with its' use and little online guidance for researchers looking to try the package. This poster aims to familiarize the audience with the Mplus Automation package by presenting an example of how the package can be used to facilitate LCA/LPA analyses. Using example data to conduct an LPA, the poster will display syntax, output, and results showing the use of the Mplus Automation package, as well as links to further resources and examples. Through this example, the author hopes to convey the benefits of using the Mplus Automation package and help researchers discover and use the package.
13. **Christi Erba, PhD Student, Physics & Astronomy (CAS)**
Abstract: Massive stars, the powerhouses of our universe, spark the next generation of stars. Their extreme radiation is strong enough to drive mass off their surfaces in a stellar wind, which can be studied using ultraviolet (UV) spectra. Magnetic massive stars have surface dipolar magnetic fields about one thousand times stronger than that of the Sun. They channel stellar winds into complex magnetospheres, creating spectral line profiles with atypical shapes. UV resonance lines reveal clues about the density and velocity structure of the magnetosphere, and help us study the wind/field interaction. To date, large-scale modeling of this phenomenon has been limited by the computational cost. Our analysis solves this problem by applying a simple analytic prescription to efficiently calculate synthetic UV spectral lines. Our model grid is directly compared to legacy Hubble Space Telescope UV observations of magnetic massive stars, enabling us to make better predictions of stellar wind parameters.
14. **Riddhi Bandyopadhyay, PhD Student, Physics & Astronomy (CAS)**
Abstract: Studies of solar wind turbulence traditionally employ high-resolution magnetic field data, but high-resolution measurements of ion and electron moments have been possible only recently. We report the first turbulence studies of ion and electron velocity moments accumulated in pristine solar wind by the Fast Plasma Investigation (FPI) instrument on board the Magnetospheric Multiscale Mission. Use of these data is made possible by a novel implementation of a frequency domain Hampel filter, described herein. After presenting procedures for processing of the data, we discuss statistical properties of solar wind turbulence extending into the kinetic range.
15. **Jacob Dums, Post Doctoral Researcher, Agriculture and Natural Resources (CANR)**
Abstract: Bacteriophage have been implicated as a major driver of biogeochemical processes in the ocean where a shift in phage population can modify the collective metabolic activity of the ocean biome. Previous studies exploring the fluctuations in phage populations often tracked populations at the resolution of seasons or years. Only recently have shorter time scale studies been performed that show rapid changes in phage populations over weeks, days, and hours. Twenty years of monthly surveys of phage populations in Narragansett Bay, Rhode Island have shown a great diversity of phage, especially phage of cyanobacteria. To leverage this breadth of historical data, viral metagenomes were generated at a finer temporal resolution of every high and low tide (~every 6 hours) for 48 hours to confirm the fidelity of the monthly surveys to describe viral diversity. The viral metagenomes will deepen our knowledge of phage diversity in these quickly oscillating systems.
16. **Bradley Altmiller, Undergraduate Student, Computer & Information Sciences (EG)**
Abstract: Our project focuses on applying machine learning algorithms to the clinical and genomic data of nearly 200 Leukemia patients to determine the molecular changes that drive childhood cancers. The goal is to find what mix of feature selection techniques and classification models provide the most accurate and generalizable results. Selecting which of the 20,000 gene groups to train our models on is a major question with many different approaches. Will the compare and contrast of different feature selection algorithms report a similar group of 'important' genes? Our project uses Decision Trees and progresses to more complex predictors i.e. Support Vector Machines (SVMs). We observe a prediction accuracy of 65% with Decision Trees, a 75% using Random Forests and XGBoost, and 88% with SVMs. We will also explore the sensitivity and specificity factors.

17. **Prasanna Joglekar, PhD Student, Biology (CAS)**
Abstract: Bradyrhizobium spp. are soil dwelling Gram-negative Alphaproteobacteria that nodulate the roots of legumes including soybeans, a crop critical to global food security. These symbionts fix atmospheric nitrogen and provide 75% of the soybean plant's nitrogen requirements. Within a large culture collection (n=352) of soybean bradyrhizobia, several isolates produced phage particles spontaneously or by chemical induction. While much is known about symbiotic efficiency of bradyrhizobia, scant data are available on the role of their lysogenic viruses. Multiple approaches were used to characterize the viruses produced by four bradyrhizobia accessions, including host genome sequencing and viral fraction sequencing to identify prophage regions. All the five identified prophages range from ~45 – 45 kb in size and two of them contain Gene Transfer Agent (GTA) like elements that may help bacteria in horizontal gene transfer and have an impact on bacterial symbiotic efficiency.
18. **Chaoyi Xu, PhD Student, Arts & Science (CAS)**
Abstract: Cryo-electron microscopy (CryoEM) has become a routine tool to determine the atomic structures of biomolecules. There are currently over 2,400 CryoEM structures deposited in the Protein Data Bank. However, many were found to have atomic clashes and/or bad rotamers. Here, we present a protein sidechain refinement method based on CryoEM densities with near-atomic resolutions. This method uses the protein structures as inputs. Followed by structure analyses based on Molprobity, the problematic residues in the input structures were identified and refined in the real-space. The refined structures created could be used as initial models for the next refinement. We applied this method to 20 protein structures in PDB. The results suggest that the overall Molprobity scores of the models are notably improved and most of the structural defects in the initial models are fixed.
19. **Imam Cartealy, PhD Student, Computer & Information Sciences (EG)**
Abstract: Determining whether a protein belongs to a metabolic pathways is an important annotational task, can provide context to the basic functional annotation and aid reconstruction of incomplete pathways. In this work, we develop a method for pathway membership inference based gene ontology (GO) similarity between a query protein and proteins that are known to the members of a given human pathway. By comparing with various existing GO term semantic similarity, we develop an effective and efficient way to take into both information content of individual GO terms and the whole GO hierarchy. We test the classifier using 10-fold cross validation for all metabolic pathways reported in KEGG database and demonstrate that our method either outperform with statistical significance or perform comparably with a suite of existing semantic similarity measures, as evaluated using ROC score. And our method outperforms other methods in running time by multiple orders of magnitude for long pathways.
20. **Alexandre David-Uraz, Post Doctoral Researcher, Physics & Astronomy (CAS)**
Abstract: In this poster, I will discuss some of the exciting science that can be done in the study of massive stars using space facilities (HST, TESS, etc.). In particular, I will touch upon some of the challenges related to the management of large datasets and some of the collaborative efforts that have been deployed to face those challenges.
21. **Lan Yu, PhD Student, Public Policy& Administration (CAS) (Biden School)**
Abstract: In this project, the researcher trying to evaluate the influence of green-space on foot traffic in urban environments by using New York City as an example. New York City is one of the greatest cities in the world, many people are living and working in this city. The more people living in one city, the more influence will they bring to the urban environments. In this project, different with other research projects which are focusing on the straight pollutions in urban environments. The researcher analyzing the influence of green-space on foot traffic in urban environments. Whether the larger area of green-space will increase the works/residents willingness to go outside? What influence will these increased foot traffic bring to the urban environments? These are the main topics that the researcher wants to stress in this project. The green space databases that used in this project are coming from NYC open data, NYC jobs database coming from United States Census Bureau.
22. **Xiaohang Zhao, PhD Student, Graduate Interdisciplinary Program**
Abstract: Industry classification systems (ICSs), aiming at identifying economically related firms as peer firms, play important roles in both academia and industry. The traditional expert-driven approach of designing ICSs have several limitations such as high development and maintenance cost as well as limited granularity. While the alternative text-based algorithm-driven approach is promising in circumventing the limitations, one need to solve the challenging problem of text representation. In this paper, we propose a novel text representation algorithm based on deep learning. We develop a gating mechanism to integrate information from contexts of different scopes. Based on the proposed algorithm, we further develop a novel ICS and validate its effectiveness empirically by showing that it generates better peer groups than existing ICSs.

23. **Ryan Moore, PhD Student, Bioinformatics and Systems Biology (CBCB)**
    Abstract: Inteins are mobile genetic elements found within the coding regions of genes. The protein equivalent of introns, they are transcribed and translated along with their flanking protein fragments (exteins) before splicing out from the precursor protein. Whereas inteins were previously thought to be parasitic genetic elements providing no benefit to the host organism, recent studies suggest that inteins may impact host ecology, providing a selective advantage to the organism in which they reside by exhibiting post-translational control on extein sequences. Additionally, this ability to post-translationally control other proteins has been leveraged in various biotech and genetic engineering applications. Given their utility, inteins have been the subject of an increasing number of studies aiming to identify novel inteins in large protein datasets. To aid in this endeavor, we introduce InteinFinder, a standardized, automated pipeline for screening inteins from large peptide datasets.
24. **Kexin Yin, PhD Student, Graduate Interdisciplinary Program**
    Abstract: Link recommendation, which recommends unlinked friends to users, is a core function provided by leading social network. Existing methods recommend links in terms of user structural or nodal proximity but overlook the benefit of friends' diversity. For each user in a social network, improving the diversity of friendship network can not only entrench diversified information channels but also enhance his/her social capital. Diverging from diversification methods in recommender systems, which try to improve the diversity while maintaining precision at an acceptable level, we define diversity preference to portray individual's demand of friend diversity and formulate a new research problem: diversity preference-aware link recommendation problem. We model the problem as a sum-of-ratio optimization programming to select recommendations from a candidate set by maximizing the diversity preference matching score. We proof the problem is NP-hard and propose an efficient heuristic solution.
25. **Amelia Harrison, Master Student, Marine Science (CEOE)**
    Abstract: Marine viruses are critical to biogeochemical cycling, microbial mortality, and horizontal gene transfer. Viral marker genes such as ribonucleotide reductase (RNR) can provide insight into phage ecology. RNRs are enzymes that reduce ribonucleotides to deoxyribonucleotides and are common in the genomes of diverse lytic dsDNA phage. RNRs are divided into several types based on their biochemical requirements. Thus, RNR type is indicative of the environmental conditions surrounding DNA replication. With the goal of studying the role of RNR in shaping environmental phage populations, we examined RNRs from 143 marine viral metagenomes. Overall, RNR-containing viruses were reflective of the total viral community. Additionally, viruses carrying different RNR types had different ecological distributions, indicating connections between viral RNRs and environmental gradients. The capacity of RNR to connect phage biology and ecology makes it an informative marker for marine viruses.
26. **Abdul Qadir, Master Student, Geography (CEOE)**
    Abstract: Watching through the clouds: A machine-learning based multi-temporal radar image classification for monsoon crops using cloud-computing platform. Monitoring the monsoon crops through optical satellite data in tropical regions is challenging due to many reasons including consistent cloud cover throughout the season, small field size, and highly dynamic cropping pattern through space and time. Radar data provide an alternative for monsoon crop mapping and monitoring due to the sensor's cloud penetrating capabilities. In this work, a machine-learning based random forest (RF) classifier was applied for monsoon crop monitoring in India. We processed 420 scenes of dual polarization (VH+VV) Sentinel-1 radar data on Google Earth Engine (GEE) platform. For validation of the proposed method, stratified random samples were collected in a contiguous region composed of eight different agro-ecological regions in India. The classification accuracy obtained using radar images was 87%.
27. **Mu He, Master Student, Graduate Interdisciplinary Program**
    Abstract: With the adoption of Learning Management Systems (LMSs) in educational institutions, many researchers have used logging data/clickstream from the LMS to predict student performance. However, due to the diversity in course formats and predictor variables crafted from the data, those findings and conclusions are not consistent, and the prediction models lack portability across different settings. In this research, we have collected and analyzed data from multiple online graduate programs over the past three years. By coupling the LMS data and institutional data, we first built various machine learning models based on different data source. Then we created a hybrid model by applying ensemble methods.

****Denotes Participant – Student Lightning Talks Session 2**

DSI SYMPOSIUM PLANNING COMMITTEE

Cathy Wu, *Director, Data Science Institute*

Greg Dobler, *Co-Chair*

Zachary Collier, *Co-Chair*

Benjamin Bagozzi
Federica Bianco
Richard Braun
Austin Brockmeier
Susan Conaty-Buck
Adam Davey
Vu Dinh
Xiao Fang
Jing Gao
John Gizis
Arshiya Khan
John McNutt
Pinki Mondal
Jenn Roth
Andrea Trungold
Timothy Webb
Xiugang Wu

Visit the University of Delaware Data Science Institute at https://dsi.udel.edu/

JOIN US ON TWITTER AND SLACK

@dsiudel

hashtag: #datascienceUD19

Slack channel: bit.ly/UDDSI19

THANK YOU TO OUR INDUSTRY SPONSORS

ChristianaCare™

Christiana Care Value Institute

Predictive Analytics Group
Analyze Tomorrow Today

Predictive Analytics Group