UNIVERSITY OF DELAWARE

DATA SCIENCE INSTITUTE

DELAWARE Data Science SYMPOSIUM

NOV 17, 2021

#datascienceUD21

DELAWARE Data Science SYMPOSIUM NOV 17, 2021

UNIVERSITY OF DELAWARE DATA SCIENCE INSTITUTE

SYMPOSIUM AGENDA

Registration	8:15-9:00am
Welcome Remarks: Anshuman Razdan and Cathy Wu	9:00-9:10am
Morning Keynote Speaker: Ben Shneiderman	9:15-10:05am
Ethics Panel Session	10:10-10:55am
UD Students and Postdocs Lightning Talks	10:55-11:10am
Coffee Break and Poster Session 1	11:10-11:35am
UD Faculty and Researchers Talks: Session 1	11:35am-12:20pm
Lunch	12:20-1:10pm
Industry Panel Session	1:10-1:50pm
Afternoon Keynote Speaker: Natalie Nelson	1:55-2:45pm
Coffee Break and Poster Session 2	2:45-3:05pm
Education Panel Session	3:05-3:50pm
UD Faculty and Researchers Talks: Session 2	3:55-4:40pm
Closing Remarks and Awards	4:40-5:00pm
Industry Tables *(All Day in the Atrium)*	9:00am-5:00pm

WI-FI ACCESS:

Guests can log onto the UDel_Guest Wi-Fi network.
You can register yourself for network access, then temporary credentials will be sent to your email account or smartphone.

MORNING KEYNOTE SPEAKER

Ben Shneiderman, PhD

Professor, Department of Computer Science & UMIACS, University of Maryland

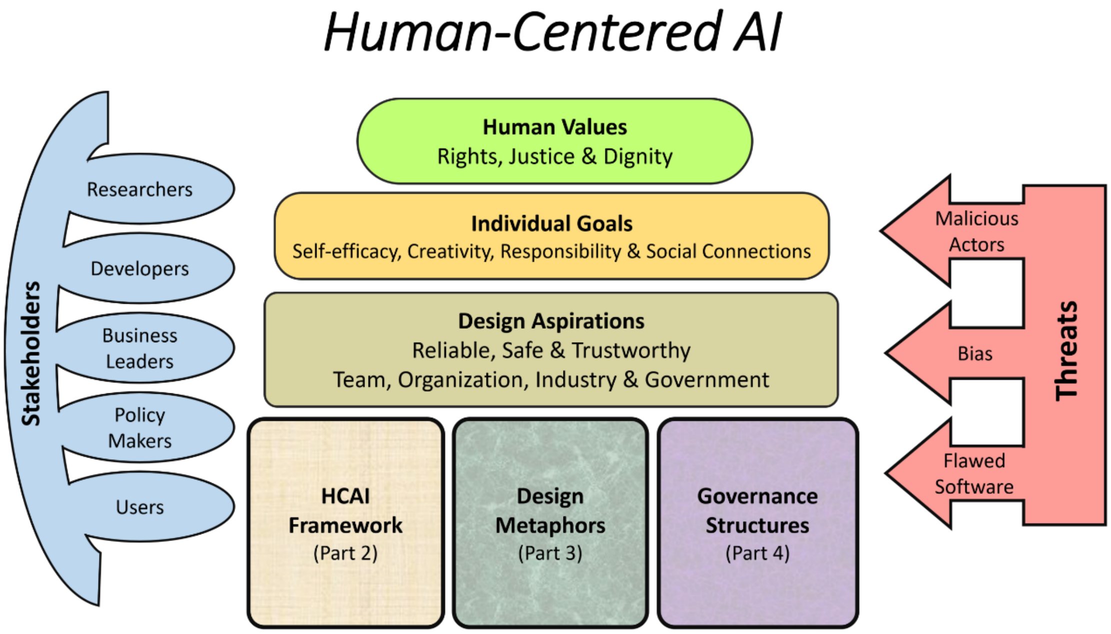
Founding Director,
Human Computer Interaction Lab

Dr. Ben Shneiderman is an Emeritus Distinguished University Professor in the Department of Computer Science, Founding Director (1983-2000) of the Human-Computer Interaction Laboratory (http://hcil.umd.edu), and a Member of the UM Institute for Advanced Computer Studies (UMIACS) at the University of Maryland. He is a Fellow of the AAAS, ACM, IEEE, NAI, and the Visualization Academy and a Member of the U.S. National Academy of Engineering. He has received six honorary doctorates in recognition of his pioneering contributions to human-computer interaction and information visualization. His widely-used contributions include the clickable highlighted web-links, high-precision touchscreen keyboards for mobile devices, and tagging for photos. Shneiderman's information visualization innovations include dynamic query sliders for Spotfire, development of treemaps for viewing hierarchical data, novel network visualizations for NodeXL, and event sequence analysis for electronic health records.

Ben is the lead author of *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed., 2016). He co-authored *Readings in Information Visualization: Using Vision to Think* (1999) and *Analyzing Social Media Networks with NodeXL* (2nd edition, 2019). His book *Leonardo's Laptop* (MIT Press) won the IEEE book award for Distinguished Literary Contribution. *The New ABCs of Research: Achieving Breakthrough Collaborations* (Oxford, 2016) describes how research can produce higher impacts. His forthcoming book on *Human-Centered AI*, will be published by Oxford University Press in January 2022.

http://www.cs.umd.edu/~ben

Human-Centered AI: Reliable, Safe & Trustworthy


Human-Centered AI
Human Values
Rights, Justice & Dignity
Individual Goals
Self-efficacy, Creativity, Responsibility & Social Connections
Design Aspirations
Reliable, Safe & Trustworthy
Team, Organization, Industry & Government
HCAI Framework (Part 2)
Design Metaphors (Part 3)
Governance Structures (Part 4)
Stakeholders
Researchers
Developers
Business Leaders
Policy Makers
Users
Threats
Malicious Actors
Bias
Flawed Software

Abstract: A new synthesis is emerging that integrates AI technologies with HCI approaches to produce Human-Centered AI (HCAI). Advocates of this new synthesis seek to amplify, augment, and enhance human abilities, so as to empower people, build their self-efficacy, support creativity, recognize responsibility, and promote social connections.

Researchers, developers, business leaders, policy makers and others are expanding the technology-centered scope of Artificial Intelligence (AI) to include Human-Centered AI (HCAI) ways of thinking. This expansion from an algorithm-focused view to embrace a human-centered perspective, can shape the future of technology so as to better serve human needs. Educators, designers, software engineers, product managers, evaluators, and government agency staffers can build on AI-driven technologies to design products and services that make life better for the users. These human-centered products and services will enable people to better care for each other, build sustainable communities, and restore the environment. The passionate advocates of HCAI are devoted to furthering human values, rights, justice, and dignity, by building reliable, safe, and trustworthy systems.

The talk will include examples, references to further work, and discussion time for questions. These ideas are drawn from Ben Shneiderman's forthcoming book (Oxford University Press, January 2022). Further information at: https://hcil.umd.edu/human-centered-ai

AFTERNOON KEYNOTE SPEAKER

Natalie Nelson, PhD

Assistant Professor, Department of Biological and Agricultural Engineering, NC State University

Principal Investigator, Biosystems Analytics Lab

Dr. Natalie Nelson is an Assistant Professor of Biological and Agricultural Engineering and Faculty Fellow in the Center for Geospatial Analytics at NC State. She is the Principal Investigator of the Biosystems Analytics Lab, studies from which take a data-intensive and management-focused approach to the study of environmental system dynamics. Natalie and her team pursue questions related to estuarine and coastal water quality, land-sea connectivity, and the influence of climate and land use change on agroecosystem productivity in the Atlantic-Gulf Coastal Plains through the use of statistical, process-based, and machine learning models. This work involves analysis of a wide range of data types – from in situ monitoring observations to satellite imagery. A key goal of Natalie's work is to advance the use of predictive models in natural resources management.

Natalie earned a B.S. degree in Agricultural and Biological Engineering from the University of Florida, and her Ph.D. in the same department through the support of a NSF Graduate Research Fellowship. During her Ph.D., she focused on hydrologic sciences and water quality modeling, and investigated drivers of harmful algal blooms in fresh and coastal waters of Florida. She joined NC State in 2017. While at NC State, Natalie has been named a Goodnight Early Career Innovator and University Faculty Scholar. In 2021, she received a NSF CAREER Award from the Environmental Engineering program to study the effects of sunny-day floods on coastal water quality.

https://www.bae.ncsu.edu/people/nnelson4/

Sensing to Sense-Making: The Role of Data Science in Advancing Environmental Sustainability

Abstract: From extra-terrestrial satellites to field instruments, the environmental landscape is increasingly outfitted with novel sensors designed for high-resolution monitoring. The advent of cost-efficient sensors now allows for a huge variety of biological, physicochemical, and socioeconomic factors to be observed at unprecedented rates and scales. In principle, these novel observations should revolutionize our understanding of environmental dynamics by revealing how systems interact and evolve over time and space. Yet, in practice, researchers and practitioners working in environmental management often struggle to make sense of vast and ever-growing volumes of data. Moreover, many mechanistic approaches and programs traditionally used in environmental systems analysis cannot handle "big data" and are becoming obsolete in the present data paradigm. As our capacity to observe environmental system dynamics through diverse sensor technologies grows, so does the demand for analyses that convert data into actionable information. In this talk, I will outline barriers that partly explain why we often observe abundant sensing without sense-making, and describe opportunities for advancing the use of data science to support sustainable environmental decision-making.

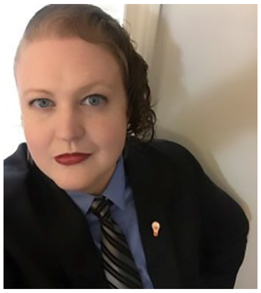
ETHICS PANELISTS

Thomas Powers, PhD *(Moderator)*

Associate Professor, Department of Philosophy, University of Delaware
Director, Center for Science, Ethics & Public Policy

Thomas M. Powers was born in Honolulu and received degrees in philosophy from the College of William and Mary (BA) and the University of Texas at Austin (PhD). He has been a DAAD-Fulbright dissertation fellow at the Ludwig-Maximilians-Universität in Munich, Germany, an NSF research fellow in the School of Engineering and Applied Science at the University of Virginia, and a visiting researcher at the informatics laboratory of the Sorbonne University, Paris, France. He is currently Associate Professor in Philosophy and in the Biden School of Public Policy and Administration, and the director of the Center for Science, Ethics, and Public Policy, at the University of Delaware. He is a faculty affiliate of the Delaware Biotechnology Institute, the Data Science Institute, the Sociotechnical Systems Center, and the Center for Autonomous and Robotics Systems, all at the University of Delaware. His research focuses on ethics in information technology.

https://www.udel.edu/faculty-staff/experts/thomas-powers/



Sarah Shugars, PhD

CDS Moore-Sloan Faculty Fellow, Center for Data Science (CDS), NYU
Research Fellow, School of Media & Public Affairs, George Washington University

Dr. Sarah Shugars is a computational social scientist who develops new methods in natural language processing, network analysis, and machine learning in order to better understand how people talk and reason about political issues. They are currently a Faculty Fellow at NYU's Center for Data Science (CDS) and a Research Fellow at George Washington University's School of Media & Public Affairs

https://sarahshugars.com/

ETHICS PANELISTS (CON'T)

Lauren F. Klein, PhD

Winship Distinguished Research Professor and Associate Professor,
Departments of English and Quantitative Theory and Methods,
Emory University
Director, Digital Humanities Lab

Dr. Klein is the Winship Distinguished Research Professor and Associate Professor in the Departments of English and Quantitative Theory and Methods at Emory University, where I also direct the Digital Humanities Lab. Before arriving at Emory, she taught in the School of Literature, Media, and Communication at Georgia Tech. She received her PhD in English and American Studies from the CUNY Graduate Center, and my AB in Literature (English and French) from Harvard University. In 2017, she was named one of the "rising stars in digital humanities" from *Inside Higher Ed*. She is the author of two books. The first, *Data Feminism* (MIT Press), co-authored with Catherine D'Ignazio, is a trade book that explores the intersection of feminist thinking and data science. The second, *An Archive of Taste: Race and Eating in the Early United States* (University of Minnesota Press), shows how thinking about eating can help to tell new stories about the range of people, from the nation's first presidents to their enslaved chefs, who worked to establish a cultural foundation for the United States. With Matthew K. Gold, she edits *Debates in the Digital Humanities* (University of Minnesota Press), a hybrid print/digital publication stream that explores debates in the field as they emerge. The most recent book in this series is *Debates in the Digital Humanities 2019*.

https://lklein.com/

Kristene Unsworth, PhD

Director, Center for Science, Technology and Society, Drexel University

Dr. Kristene Unsworth is an assistant teaching professor in the Department of Criminology and Justice Studies, and serves as the director of the Center for Science, Technology and Society at Drexel University. She received her PhD in Information Science from the University of Washington in 2010. She is an expert in ethics, information technologies and policy. Her research centers on information justice, which includes the ethics of data access and use as well as policy development that addresses the equitable use of information across society. She was co-Principle Investigator for a National Science Foundation research project on the Ethics of Algorithms [NSF EESE1338205], which led to the creation of case studies that are used in ethics education in computer science and engineering. She is an active faculty member in the Justice Informatics program in Criminology and Justice Studies where courses and research focus on the intersections between information technology, people, and justice.

https://drexel.edu/coas/faculty-research/faculty-directory/unsworth-kristene/

INDUSTRY/AGENCY PANELISTS

Patrick Callahan *(Moderator)*

Co-Founder, CompassRed

Patrick Callahan is the founder of CompassRed, a data science and analytics company. The CompassRed Team is made up of data scientists and technologists tackling some of the world's most fascinating problems on behalf of their global clients. As CEO, he represents the company's vital, core interests and growth as it develops in the changing world of Artificial Intelligence. Patrick ensures that CompassRed's data analytics and data science culture is made up of the most curious and brightest minds, from former scientists and engineers to seasoned corporate strategists. With 20+ years of big agency and enterprise experience, Patrick works to ensure the team understands their clients' data, uses that data to make predictions, and that the team delivers the recommendations on what steps to take next.

Evidenced by his involvement in developing the region's community of data scientists through the founding of the Delaware Data Innovation Lab and the "Philadelphia Data Jawn", Patrick devotes his time to making sure that we educate our team, our clients, and the community on the art of the possible as it relates to customer analytics and data science.

He is the Organizer of Social Media Data and Analytics Groups in San Francisco, New York, Washington DC, Boston, Milan, and London. He is also the co-founder of The Data Lab podcast, and co-author of "Engage Your Brand: How smart companies are using Social Media to Drive their Business Forward". He is on the board of the Delaware Prosperity Partnership and the Delaware Data Innovation Lab. Patrick resides with his wife and two children in Centreville Delaware.

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

Héctor Maldonado-Reis

Director, Research Development & Analytics – Data Innovation Lab, Tech Impact

Héctor "Héc" Maldonado-Reis (he/they) is a Director of Tech Impact's Data Innovation Lab. Tech Impact is a non-profit with the mission to use data and technology to better serve the world. There, Héc directs Research Development and Analytics leading the technical innovation and application of data science towards improving social good. At The Lab, their work focuses on leveraging data to convene industry, government, and community partners across multiple domains, collaborating towards enhancing population level well-being.

https://ddil.ai/

INDUSTRY/AGENCY PANELISTS (CON'T)

Mia Papas, PhD

Corporate Director of the ChristianaCare Institute for Research on Equity and Community Health (iREACH)
Principle Investigator for the Delaware Clinical and Translational Research (DE-CTR ACCEL) Program

Dr. Mia Papas is the Corporate Director of the ChristianaCare Institute for Research on Equity and Community Health (iREACH) where she is responsible for oversight of research and operations. She is also the site Principle Investigator for the Delaware Clinical and Translational Research (DE-CTR ACCEL) Program focused on building a research structure that allows for the rapid translation of research into practice targeting improved health outcomes for all Delawareans. Dr. Papas' expertise is the design and conduct of research that utilizes real-world evidence to understand disease, evaluate treatments, and demonstrate the impact of healthcare innovations and interventions with a focus on health equity. She is actively engaged in research that leverages data from electronic health records, administrative claims, registries, and government databases to answer critical questions about disease epidemiology, burden, and costs. Through a focus on real world evidence, she works toward bridging the gap between clinical and translational research, advancements in the quality of care, and improved population health outcomes.

Dr. Papas received her PhD in Epidemiology from the Bloomberg School of Public Health at Johns Hopkins University, her Master of Science in Biostatistics and Epidemiology from the Arnold School of Public Health at the University of Massachusetts, Amherst and her Bachelor of Science in Mathematics from Fairfield University. She has authored over 70 peer-reviewed articles, obtained funding from the National Institutes of Health and the Centers for Disease Control and Prevention, and presented her research at over 100 national and international conferences. She was recently elected to the Board of the American College of Epidemiology, has been an active member of the American Public Health Association and is a Founding Board Member for the Delaware Public Health Association. She has taught the principles of epidemiology throughout the world and is committed to developing good scientific practice in medical research. She is a native Delawarean and lives in Newark with her husband and two children.

https://research.christianacare.org/ireach/people/mia-papas-ph-d-ms/

INDUSTRY/AGENCY PANELISTS (CON'T)

Luke Rhine

Director of the Career and Technical Education (CTE) and STEM Workgroup, Delaware Department of Education

Luke Rhine is the Director of the Career and Technical Education (CTE) and STEM workgroup at the Delaware Department of Education (DDOE). He is responsible for leading the development, implementation, and continuous improvement of the statewide system of CTE in Delaware's secondary and postsecondary institutions as well as STEM initiatives in grades k through 12. Luke is also responsible for developing and implementing educational policy. Prior to working at the DDOE, Luke was a Program Specialist in CTE and STEM with the Maryland State Department of Education. He has also worked as a high school and middle school teacher. Luke has received several state and national awards for educational leadership and was a Fulbright scholar

https://leadershipdelaware.org/teams/luke-rhine/

Matthew Parks

Vice President, CRA Officer and Retail Banking at Discover Financial Services

Matthew joined Discover Bank in 2001 as Manager of Deposit Operations, after ten years of banking operations experience, including consumer, commercial and mortgage lending, credit and collections. In 2003, he began leading Discover Bank's Community Reinvestment Act (CRA) program and has originated and managed over $1.5 Billion in community impact investments. Investments under management include affordable housing, new markets, and historic tax credits and a variety of fixed income instruments. While in this role, he was also integral in the building of a national wholesale banking business; and became responsible for the Discover's sole retail branch. During his tenure, he also assisted with the development of the Bank's Affinity relationships, starting with AAA in 2007 with the origination of billions of consumer deposits which he continues to manage.

In the Delaware community, Matt serves on the Board of the Delaware Bankers Association and was appointed the Chairman of the Board for the Milford Housing Development Corporation, a nonprofit multifamily and single-family housing developer and the Stepping Stones Community Federal Credit Union. In a national capacity, he was on the board for the Affordable Housing Investors Council and served on the American Bankers Association's Housing and Economic Development Committee. Matthew obtained an MBA from the University of Delaware and Bachelor of Science degree in Finance from Lemoyne College.

EDUCATION PANELISTS

Cherese Winstead *(Moderator)*

Interim Dean, College Agriculture, Science & Technology (CAST) and Professor, Department of Chemistry, Delaware State University

Dr. Winstead has recently accepted an appointment as Interim Dean of the College of Agriculture, Science & Technology (CAST) and has previously served as the Department Chair of Chemistry at Delaware State University (DSU) for over 8 years. During academic career she has continued to make strides in the areas of research, teaching, and service. Since joining the faculty in the Department of Chemistry in Fall of 2008 at Delaware State University, she has secured over $18.4 million in funding on projects for which she has serve as Principal, co-investigator, or key personnel on projects surrounding the areas of material and data science. Combined with proposals that have not been funded and pending, her grant-seeking efforts over the past five years total nearly $33.1 million. Dr. Winstead has worked with numerous regional, national and global agencies, non-profit organizations, revealing an important inter-institutional and interdisciplinary approach toward the integration of research & education. In the area of service, Dr. Winstead is Founder and President of the Young Chemists Society (YCS) and the Delaware Association for the Advancement of Science (DAAS), both non-profit organizations dedicated to the early education of students in the sciences.

https://cast.desu.edu/about/faculty-profiles/cherese-winstead-phd

Amarilis Lugo de Fabritz

Master Instructor for Russian, Department of World Languages & Cultures, Howard University

B. Amarilis Lugo de Fabritz, Ph. D., is the Master Instructor for Russian at Howard University's Department of World Languages and Cultures. Only thing to add: Howard University is the only HBCU with a Russian program – the Russian Language and Literature Minor. Amarilis was born in Ponce, Puerto Rico. She attended high school in Acton, Massachusetts, where she started her studies in Russian language. She received her Bachelor Degree in Russian Language and Literature with Honors from Brown University. She received a Master of Art in International Studies, in Russian, East European, and Central Asian Studies from the Jackson School of International Studies at University of Washington, Seattle, Washington. She then completed her doctorate in Slavic Languages and Literatures at the University of Washington, Seattle, Washington. She has been involved in a number of data science- and cyber security-related undergraduate education and training initiatives, including a recent Howard University - National Security Agency - University of Delaware data science collaboration.

https://profiles.howard.edu/profile/40511/brunilda-amarilis-lugodefabritz

EDUCATION PANELISTS (CON'T)

Roghayeh (Leila) Barmaki, PhD

Assistant Professor, Department of Computer and Information Sciences,
University of Delaware
Data Science Institute Resident Faculty

Leila Barmaki is a computer-science faculty at the University of Delaware (UD), where she directs the Human-Computer Interaction Lab (HCI@UD). Leila received her PhD in Computer Science, and MSc in Artificial Intelligence in 2016 and 2012 respectively. Before joining UD in 2018, she was a postdoctoral fellow at Johns Hopkins University. Her research is at the intersection of data science and human-computer interaction for healthcare and education applications. In particular, she combines embodied cognition with technological advancements in augmented and virtual realities, multimodal machine learning, and human-computer interaction to design high-impact medical and educational interventions. Leila's research has been supported by Amazon Research and INBRE Pilot awards. Leila enjoys working with graduate and undergraduate students on solving interdisciplinary research problems, and she introduces these real-world problem-solving skills to her students in the classrooms, too.

https://sites.udel.edu/rlb/

Jordan Harrod

Ph.D. Candidate in Medical Engineering and Medical Physics,
Harvard-MIT Health Sciences and Technology program

Jordan Harrod (she/her) is a PhD Candidate in Medical Engineering and Medical Physics at the Harvard-MIT Health Sciences and Technology program. She works at the intersection of non-invasive brain-machine interfaces and machine learning for pain and anesthesia under Dr. Ed Boyden and Dr. Emery Brown. Jordan received her Bachelor of Science in Biomedical Engineering from Cornell University in 2018, where she worked on interfacial tissue engineering, medical image analysis, and machine learning for MRI reconstruction. In her spare time, Jordan is actively involved in science communication via her YouTube channel, which focuses on engaging the public on artificial intelligence, as well as on Twitter, Tiktok, and Instagram. Outside of research, you can find her reading the latest V. E. Schwab book or practicing her Olympic lifts at the gym.

https://www.jordanharrod.com/

EDUCATION PANELISTS (CON'T)

Michael Ayewoh, PhD

Chief Research & Sponsored Programs Officer, Office of Sponsored Programs, Lincoln University

Dr. M. Ehi Ayewoh, Graduated from Tennessee State University, Nashville with a bachelor of science degree, with honors, in animal science. He attended Pennsylvania State University and earned a master of science degree in poultry science and technology, followed by a master of education in extension education (adult education), and then a doctorate in agricultural and extension education, also at Penn State, where his emphases were on planning, evaluation, administration, youth programs, supervision of government and nonprofit organizations, research and development, and leadership for social change. Dr. Ayewoh holds post-doctoral certificates in Advanced Leadership Development from The Chair Academy-Worldwide Leadership Training for Post-Secondary Leaders, Generating Expectations for Student Achievement from the New York State Department of Education, and Advanced Grantsmanship from David G. Bauer and Associates and The Foundation Center.

Dr. Ayewoh is now the Chief Research & Sponsored Programs Officer at Lincoln University. In addition, Dr. Ayewoh provides comprehensive oversight for all sponsored program initiatives that are inclusive of grants, contracts, and cooperative agreements.

https://www.lincoln.edu/faculty-and-staff/directory/michael-ehi-ayewoh

FACULTY TALKS: SESSION 1

1. **Pinki Mondal, Department of Geography & Spatial Sciences; Plant & Soil Sciences; DSI**
   Do You Have Salt on Your Land? *Abstract:* Inland movement of seawater can change soil salinity in coastal farmlands, leading to loss in agroecosystem productivity and reduced farmland resilience. Exacerbated by sea-level rise, drought, and storm surges, such incremental seawater intrusion can lead to new ecosystems. Here we report a means of early detection using a newly developed method of identifying surficial salt deposits. Our image classification scheme relied on a range of input bands including visible and near infrared bands from the very high-resolution aerial imagery (1m) collected through the National Agriculture Imagery Program, in addition to multiple satellite-derived spectral indices, and seasonal thermal bands from Landsat (30m). Using a Random Forest algorithm with 100 trees and over 87,500 reference points, we developed high-resolution (1m) datasets for two time-steps: 2011-13 and 2016-17. The geospatial datasets are classified images for each time-step and have eight categories: forest, marsh, salt crust, cropland, other vegetation, bare soil, built, water. We further quantified changes between these two time-steps by developing a change trajectory map to document the magnitude and direction of these changes. The geospatial datasets will be released through an open data repository and will be available to the stakeholders at no cost. The team also developed an app prototype that will allow the users to examine the changing landscape. Eventually, the app will allow the users to report salt deposits on their land. This project demonstrates the effectiveness of a tiered mentoring approach where a group of faculty, graduate students, undergraduate students and a high school student contribute towards research, teaching and outreach.
2. **Rahmat Beheshti, Computer & Information Sciences; Health Sciences; Nemours Children's Health System; DSI**
   Responsible AI to combat childhood obesity *Abstract:* Childhood obesity is a major public health problem affecting many families across the globe. Besides its own challenges and complications, childhood obesity can lead to numerous other short and long-term diseases across children's lives and continue into adulthood. Because of the complex nature of this disease, engaging in early interventions is a key factor to succeed to prevent and control of the disease. Similarly, once developed, improving the overall outcomes of the interventions to treat the disease is critically needed. In this talk, I will discuss our multi-year and multi-party project aimed at improving prevention and treatment interventions of childhood obesity using state-of-the-art artificial intelligence techniques. I present the steps we have taken to use large-scale electronic health records datasets from Nemours Children's health to develop a comprehensive package of childhood obesity comprising of prevention- and treatment-centric models. As these tools are designed to be used in clinical settings, I will also discuss numerous steps that we have taken or are planning to take to address the concerns about the responsible development of our tools, including addressing biases, accountability, interpretability, and actionability.
3. **Kayla Abner, UD Library**
   Library Support for Data Science *Abstract:* The Digital Scholarship and Publishing team at the Library, Museums, and Press are available to assist with computational and digital methods in research and teaching. We partner with faculty and students to provide instruction and guidance for the employment of digital techniques for data management, visualization, analysis, and publishing, among others. Our focus is building a culture of sustainability, accessibility, and ethical use of data. Librarians in Digital Scholarship and Publishing consult on the scope and scale of proposed projects, balancing the need for projects to be big enough to accomplish research goals, but managing scale so that projects are sustainable and ethical. As experts in data management, analysis, and visualization, we can consult on best practices, offer workshops customized for your course or working group, and can collaborate to develop teaching materials or other deliverables. Find out more about our team at.
4. **Alexei Kananenka, Department of Physics & Astronomy**
   Machine learning for long-time quantum dynamics *Abstract:* Accurate simulations of quantum dynamics in complex condensed-phase systems are our gateway to understanding many physical, chemical, and biological processes. Exact numerical simulations often require computational resources that scale exponentially with the number of simulated time steps and the size of the system. Approximate perturbative or quantum-classical methods are often only reliable at short simulation times, rendering such methods inapplicable to study long-time quantum phenomena. Such phenomena are known to occur during photosynthesis. Experiments showed that quantum coherence between electronic states of light-harvesting complexes can persist for several hundreds of femtoseconds even at physiological temperature. The physical mechanisms underlying the long-lasting quantum coherence and the role of a protein environment are still not fully understood. I will show that artificial neural networks can efficiently and accurately simulate complex quantum dynamics across different regimes. Such methods reduce the required computational time for long-time simulations by at least two orders of magnitude and provide new routes for simulating quantum dynamics for arbitrarily long times, starting with computationally feasible short-time dynamical information.

5. **Isa Haskologlu, Political Science & International Relations at UD; American University at Washington, D.C.**
An Interdisciplinary Approach: Teaching Data Science in Political Science *Abstract:* The age of the Big Data opens new opportunities for the empirical studies in the field of Political Science. Examples include analyzing and utilizing the data for research such as data on census records, campaigns, international trade, climate, social media posts, race, religion, conflict, etc. With the knowledge in the political science, Big Data can make two significant contributions to the political science research: 1) identify instrumental variables by making previously unobservable variables observable, and 2) helps hypothesis generation and theory building. As a political scientist, I designed "Applied Political Data Science" course to provide students a tool to benefit from the advantages of the data environment and boost their research methods. Students have been developing skills on data retrieval, data cleaning and wrangling, Exploratory Data Analysis (EDA), data visualization, Feature Engineering &amp; Feature Selection.
6. **Karen Hoober, Center for Bioinformatics & Computational Biology; Computer & Information Sciences**
Graduate Programs at The Center of Bioinformatics and Computational Biology *Abstract:* The academic program for The Center of Bioinformatics and Computational Biology (CBCB) is built on the core curriculum of Bioinformatics Data Science (BDS). CBCB offers two online graduate certificates in Applied Bioinformatic and Biomedical Informatics & Data Science, an on-campus certificate in Bioinformatics & Computational Biology, an MS degree in Bioinformatics and Computational Biology, a PSM in Bioinformatics, and a PhD in Bioinformatics Data Science. CBCB trains the next-generation of researchers/professionals to play 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 internship opportunities (e.g., Christiana Care, Delaware Health Information Network, and DSU), preparing graduates for careers in industry, government agencies, or non-profits.
7. **Eric Best, Homeland Security and Public Policy, School of Public Affairs, PennState Harrisburgh; Center for Applied Demography & Survey Research at the Biden School**
Delaware Emergency Management Agency COVID-19 Predictive Modeling Group *Abstract:* In March 2020, UD researchers developed a specialized COVID-19 hospital capacity model for the State of Delaware and the hospital systems in the state at the request of the Delaware Emergency Management Agency. This lightning talk will discuss the rapid creation of a new data collection and validation network across hospital systems required to create accurate short-term capacity models and the simultaneous development of specialized predictive tools put into production for the Delaware Emergency Management Agency and the hospital systems, a project that continues today.
8. **Arijit Bose, Department of Physics & Astronomy**
Data-driven avenues in high-energy-density laboratory astrophysics *Abstract:* High-energy-density laboratory astrophysics involves studies of matter at extremely high temperatures and densities, like in planetary and stellar interiors, produced in a controlled laboratory environment. Scaled down terrestrial experiments contribute to our understanding of many exotic astrophysics phenomena that are typically inaccessible. Machine learning models and data-driven methods are in the process of reshaping our exploration of these extreme systems, through models that enable rapid discovery of complex interactions in large datasets. The newest generation of extreme laboratory astrophysics facilities can perform experiments multiple times a second (as opposed to approximately daily), this presents a need to move away from human-based control towards automatic control based on real-time interpretation of diagnostic data and updates of the physics model.
9. **Maryam Rahnemoonfar, College of Engineering and Information Technology, University of Maryland, Baltimore County, iHARP**

FACULTY TALKS: SESSION 2

1. **Greg Dobler, Biden School, Department of Physics & Astronomy, DSI**
   Better cities through images *Abstract:* With millions of interacting people and hundreds of governing agencies, urban environments are the largest, most dynamic, and most complex macroscopic systems on Earth. I will describe how persistent, synoptic imaging of an urban skyline by the "Urban Observatory" (UO) can be used to better understand the urban system, in analogy to the way persistent, synoptic imaging of the sky can be used to better understand the heavens. The UO is a multi-city facility consisting of a network of observational platforms that combines data collection and analysis techniques from the domains of astronomy, physics, computer vision, remote sensing, and machine learning to provide new insights into cities as living organisms that consume energy, have environmental impact, and display characteristic patterns of life, and how that new understanding can be used to improve city functioning and quality of life for its inhabitants. UO undergraduate exchange student Stephanie Gómez-Fonseca will describe her research on the signal processing techniques required to leverage UO observations for the study of urban air quality and vegetation health.
2. **Beth Willman, NSF's NOIRLab, Department of Physics & Astronomy**
   Innovative Facilities + Data Science = The Future of Astrophysics *Abstract:* Billions of dollars are being invested in observatories that will generate astrophysics datasets of a volume, velocity, complexity, and/or precision unrivaled by today's facilities. To effectively apply these data to fundamental questions about the Universe, new facilities are now planning for the data science aspects of operating a modern observatory before construction of the observatory even begins. By applying principles of open data and research inclusion to those plans, high-level curated datasets can democratize astronomy and act as an enormous force multiplier of scientific impact. This talk highlights examples of combining innovative facilities with data science within NSF's NOIRLab (the imminent Rubin Observatory and of the proposed US Extremely Large Telescope Project)."
3. **John Callahan, Department of Geography**
   Prediction of Storm Surge in the Delaware Inland Bays using Machine Learning Methods *Abstract:* It has been widely documented that Delaware is highly vulnerable to the impacts of coastal flooding along its Delaware Bay, Atlantic Ocean, and Delaware Inland Bay shores. It has been hit hard by tropical cyclones Hurricanes Irene in 2011 and Sandy in 2012, and extratropical cyclones the Ash Wednesday Storm of March 1962 and the Mother's Day Storm of May 2008. Flooding occurs from surge due to major and minor coastal storms as well as astronomical high tides, both of which are predicted to worsen as relative mean sea levels increase in a warming world. The Delaware Inland Bays (DIB) region, located in the southeast portion of the state, is an area particularly vulnerable to coastal flooding. It's an area of high population density and growing residential and commercial development. Population in the DIB watershed doubled from 1990 to 2010, with expected increases of 15% between 2010 and 2020 and 46% between 2010 and 2040, leading to gains in residential and commercial development and open water areas balanced by losses in agriculture, forests, and wetlands.

   It's also a complex, hydrodynamically-connected system. The DIB encompasses all the tidal waters and the diched and unditched wetlands of the Indian River Bay, Indian River, Rehoboth Bay, and Little Assawoman Bay, covering approximately 292 square miles of land that drains to 35 square miles of bays and tidal tributaries. Water enters the system from the Delaware Bay (via the Roosevelt Inlet on the Broadkill River to the north), the Atlantic Ocean (via the Indian River Bay Inlet to the east), and the Big Assawoman Bay (via canals and the Ocean City Inlet to the south). This flow through the narrow inlets is restricted, and combined with shallow waters and complex surface water pathways, makes the magnitude and timing of surge and tidal ebb/flow at these vulnerable locations difficult to estimate using hydrodynamic models.

   Our current project attempts to develop a predictive statistical model at multiple communities within the DIB that often experience coastal flooding, ultimately to be incorporated into the Delaware Coastal Flood Monitoring System (CFMS). The Delaware CFMS is an online, coastal flood early warning system that is focused on Delaware Bay communities and does not include the DIB. Water level sensors were installed at 13 locations located the DIB region in late 2015, maintained by the Delaware Environmental Observing System (DEOS), part of the Center for Environmental Monitoring and Analysis (CEMA) in the Department of Geography and Spatial Sciences. Water level data from these stations will be combined with data from 7 USGS and 2 NOAA tide gauges to generate predicted astronomical tides, non-tidal residuals, and peak high tides from early 2016 through 2020, approximately 4 years of data.

   This project will model the peak water levels for each high tide at each of the DIB communities up to 48 hours in advance. Peak high tides near locations of the water in and out flow to the DIB (i.e., USGS Indian River Bay Inlet, NOAA Lewes, and NOAA Ocean City) will be used as input to the model. These locations were chosen as they are within areas of operational hydrodynamic models currently operated by NOAA (i.e., Delaware Bay Operational Forecast System (DBOFS) and ET-Surge models).

Meteorological observations at a DOES weather station near the IRB Inlet site will also serve as input to the model. These include hourly wind speed, wind direction, atmospheric pressure, and precipitation in the hours preceding peak water levels at the DIB communities being modeled. Machine learning techniques will be applied through traditional statistical methods, such as multiple linear regression with time lags, as well as neural networks. Results of this project will expand the existing Delaware Coastal Flood Monitoring System to include communities in the Delaware Inland Bays. This information will provide improved forecast information that emergency managers and local officials can use to prepare for impending coastal flooding events in this vulnerable region.

4. **Leila Barmaki, Computer & Information Sciences; DSI**
   Multimodal Learning Analytics in Virtual Learning Environments *Abstract:* Virtual Learning Environments are reshaping today and tomorrow's classrooms, especially in the post-pandemic era. Multimodal Learning Analytics is a growing field, but its primary focus has been on co-located learning settings. In this presentation, we will discuss what are the possibilities of transferring co-located multimodal learning analytics findings into remote, virtual learning environments. Furthermore, visual data analysis methods for extracting multimodal behavioral features of the students, including gaze, posture, and social proximity will be introduced.
5. **Maiko Arichi, Department of Mathematical Sciences**
   Direct Mathematical Method For Real-Time Ischemic Detection From Electrocardiograms Using The Discrete Hermite Transform *Abstract:* A real-time automated identification technique is developed for the detection of ischemic episodes in long term electrocardiographic (ECG) signals using mathematical expansions involving the Discrete Dilated Hermite Transform. The discrete Hermite functions are generated as eigenvectors of a symmetric tridiagonal matrix that commutes with the centered Fourier matrix. The Discrete Hermite Transform (DHmT) values are computed from a simple dot product between an individual ECG complex extracted from the European Society of Cardiology (ESC) ST-T database and the corresponding discrete Hermite function. These values are found to contain information about the ECG shape, highlighting changes between ST segment and T wave alterations which are the features of ischemic episodes. This information from the discrete Hermite transform, based on an orthonormal set of n-dimensional digital Hermite functions that serve as shape-identification functions, can be used to identify ischemic episodes from the ECG. The performance was analyzed in terms of sensitivity, specificity and positive predictive value.
6. **Michael Crossley, Department of Entomology And Wildlife Ecology**
   Delving into historical changes in US agriculture *Abstract:* Agriculture has fed billions of people, and represents one of the most dramatic transformations of Earth's surface. How to feed the world without denuding and chemically sanitizing the planet is a grand challenge, and there is great uncertainty about how our agricultural landscapes will change as human preferences, plant and animal communities, and environmental constraints continue to evolve. One way to brace for the future is to peer into the past, to examine the causes and consequences of agricultural land use change over spatially broad and temporally deep scales. Mining archives of the US Census of Agriculture, I have mapped out the acreage and production of major crops at a county-level and roughly decadal time steps since 1840. Many examples of shifts in crop production in response to human innovations or natural upheavals are evident. Despite overall decreases in the total amount of land devoted to crops post-WWII, production has continued to climb thanks to concurrent increases in yield. Importantly, the overall diversity of crops grown has plummeted: in 1940, 88% of counties grew >10 crops, but only 2% did so in 2017, and combinations of crop types that once characterized entire agricultural regions are lost. Concomitantly, the acreages of many crops have become spatially concentrated, especially in the last two decades, with important consequences for the spread of crop pests, agrochemical use, and vulnerability of our food system to climate change and socioeconomic shocks. There is ample opportunity to examine causes and consequences of agricultural land use change using these data. For example, sociologists and economists might test how changing policies, domestic and international markets, and other social changes have shaped the amount and distribution of crop production. On the flipside, ecologists can see how strongly changes in land use have influenced shifts in occurrence and abundance of species of conservation or economic concern. Agricultural land use change can be rapid and sometimes appears to be unpredictable, but a deeper understanding of the processes underlying broad shifts in the past can bring us one step closer to anticipating the ways our landscapes will continue to evolve in the future.
7. **Yuqi Wang, Department of Computer & Information Sciences**
   Clustering of UniProt proteins using sequence embedding and Cloud computing. *Abstract:* Clustering of UniProt proteins using sequence embedding and Cloud computing Yuqi Wang 1* , Hongzhan Huang 1 , Chuming Chen 1 , Sachin Gavali 1 , Julie E. Cowart 1 , Cecilia N. Arighi 1 , Peter McGarvey 2 , Cathy H. Wu 1 1 Center for Bioinformatics and Computational Biology, University of Delaware, Newark, USA 2 Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC, USA *To whom correspondence should be addressed. Tel: +1 302-831-3234; Fax: +1 302-831-4841; Email: yuqiwang@udel.edu The UniProt Reference Clusters (UniRefs) provide clustered sets of protein sequences, which hides redundant

sequences while maintaining complete coverage of sequence space at several resolutions (UniRef100/90/50). The number of protein sequences underlying UniRef databases are growing exponentially and currently we are working with 330 million sequences. Clustering such a huge and fast-growing protein database is challenging and requires constant updates to computing software and hardware. Conventionally, protein sequence clustering is done through sequence alignment and sped up with k-mer filtering (CD-HIT, MMSeqs, etc). Now with advancements in artificial intelligence (AI) and machine learning (ML), sequence embedding shows real promise in data mining, protein clustering, and structure and functional analysis. Once the protein sequences are converted into a proper vector space, they can be processed and clustered with various leading data mining and machine learning techniques. We are developing a more advanced sequence embedding algorithm for protein sequence clustering with ML methods, looking forward to achieving clustering with better quality at a much faster speed. Currently, many protein embedding research focuses on feature prediction, using known protein annotations, even UniRef90/50 themselves, as ground truth in training/validation. SGT[1] and ProtVec[2] provide approaches that do not rely on external annotation information but only the sequences themselves and have shown promising results for small scale databases. We have expanded the embedding equation into a more general form, which potentially could yield advantages from both works. We have done some tests with a variety of parameters and visualized the vector space for the embedding. The preliminary results demonstrate even more distinguishable shapes and thresholds in comparison with the SGT and ProtVec. Sequence embedding and clustering with ML for such a huge protein sequence database relies heavily on computational resources. With the support of NIH STRIDE AWS cloud computing infrastructure, we are exploring the use of GPU clusters and Amazon SageMaker for large-scale protein sequence embedding and clustering. [1] Chitta Ranjan, S. Ebrahimi, Kamran Paynabar. Sequence Graph Transform (SGT): A Feature Extraction Function for Sequence Data Mining. arXiv:1608.03533 [stat.ML]. (https://arxiv.org/abs/1608.03533v6) [2] Asgari, E., McHardy, A. &amp; Mofrad, M.R.K. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). Sci Rep 9, 3577 (2019). https://doi.org/10.1038/s41598-019-38746-w. PMID: 30837494.

8. **Federica Bianco, Department of Physics & Astronomy, Biden School, DSI**
   Inclusive Data Science Pedagogy Across Domains and Student Populations *Abstract:* Data Science offers the opportunity to expand STEM education across social, racial, and domain lines and reach within groups historically excluded from STEM disciplines by directly connecting mathematical and statistical tools to domain applications. The University of Delaware is ramping up equitable and inclusive interdisciplinary programs at the undergraduate and graduate level in partnership with nearby Delaware State University, Lincoln University, Howard University, and many other local educational institutions. We are designing a new educational framework where Data Science can be taught to students with minimum coding experience and little quantitative inference training by focusing on applications so that students can leverage their domain expertise and develop a growth mindset through the validation of their background. Real-world datasets expose students to real-world problems. The inclusion of data ethics, collaborative work tools and practices, and visualization practices in the pedagogical path prepares students for the workforce. I will describe and showcase just a few of the exciting programs being developed at UD, including the Delaware And MidAtlantic coast Data Scientist Corps (DeMADScientistCorps), recently funded by the NSF Harvesting the Data Revolution program.

STUDENT POSTER SESSION (IN-PERSON)

***Please visit the posters and recordings at: https://dsi.udel.edu/events/dsi-symposium-2021/posters/**

1. **Xiaolong Li, PhD Student, Astronomy, Data Science**
   Detection of Light Echoes based on YOLO framework. *Abstract:* Light Echoes (LEs) are the reflections of astrophysical transients on interstellar dust. They are fascinating astronomical phenomena that enable studies of the interstellar dust that reflects them as well as of the original transients. LEs, however, are rare and extremely difficult to detect as they appear as faint, diffuse, time-evolving features. Detection of LEs still largely relies on human inspection of images, a method unfeasible in the era of large synoptic surveys. In this work, we prepare a dataset from the ATLAS telescope and test an AI object detection framework, YOLO, to demonstrate the potential of AI in the detection of LEs. We find that an AI framework can reach human-level performance even with a size- and quality-limited data set. We explore and highlight challenges, including class imbalance, label incompleteness etc., and roadmap the work required to build an end-to-end pipeline for the automated detection and study of LEs in high-throughput astrophysical surveys.
2. **Tatiana Acero Cuellar, PhD Student, Astronomy, Data Science**
   There's no difference: CNN for transient detection without template subtraction. *Abstract:* We present a Convolutional Neural Network (CNN)-based model for the separation of astrophysical transients from image artifacts, a task known as real-bogus (RB) classification, that does not rely on template subtraction (or Difference Image Analysis, DIA). We compared the efficiency of two models with similar architectures, one that uses image triplets composed of template, search, and difference image, and one that takes as input the template and search only. We investigate what information is used by each model by exploring the models' maps of pixel importance. Our work demonstrates that CNNs models can perform RB classification relying exclusively on the imaging data, bypassing DIA, the most computationally expensive step in the detection of astrophysical transients.
3. **Van Huong Le, Postdoc, Engineering,Data Science**
   Copula-based dependency model for CO2 efflux prediction and its uncertainty quantification. *Abstract:* A new data-driven method for the prediction of temporally distributed CO2 efflux conditioned by environmental variables is presented. The method, namely Bernstein copula-based temporal stochastic cosimulation, is based on Bernstein copula for the estimation of the joint probability distribution function and simulated annealing for the temporal distribution simulation. The proposed method can model complex non-linear relationships between variables in a fully nonparametric approach. The main advantage is that it does not require linear dependence between variables nor any distribution constraint. This method is first validated in a time series and applied in another time series to predict CO2 efflux conditioned to temperature in a tidal marsh soil in Delaware State. The results show that this method reproduces very well the behaviors: univariate, joint and temporal of the phenomenon studied.
4. **Riley Clarke, PhD Student, Astronomy**
   Detection & Removal of Periodic Noise in Kepler K2 Photometry with Principal Component Analysis. *Abstract:* We present a novel method for detrending systematic noise from time series data using Principal Component Analysis (PCA) in Fast Fourier Transforms (FFT). This method is demonstrated on time series data obtained from Campaign 4 of the Kepler K2 mission. Unlike previous detrending techniques in time-domain astronomy that utilize PCA, this method performs the detrending in Fourier space rather than temporal space. The advantage of performing the analysis in frequency space is that the technique is sensitive purely to the periodicity of the unwanted signal and not to its morphological characteristics. This method could improve measurements of low signal-to-noise photometric features by reducing systematics.
5. **Willow Fortino, PhD Student, Astrophysics**
   Classifying Stripped Envelope Supernovae with Properly Synthesized Low-Resolution Spectra. *Abstract:* Stripped envelope (SE) core collapse supernovae (SNe) must be classified through spectroscopy. Therefore, when deciding the resolving power R of new spectrographs it is important to know the minimum R necessary to classify these SNe to arbitrary accuracy. This work attempts to identify a critical R at which spectral classification SE SNe becomes impossible. To classify the spectra we follow previous work of Williamson (2019) and first perform Principal Component Analysis (PCA) on spectra taken at similar phases of the SN's evolution. Subsequently, a Support Vector Machine (SVM) classifier is used on some of the principal components (PCs). The SVM score for each group of SNe is recorded as we artificially degrade the spectra. We confirm that even at R = 5, the SVM score remains at ~0.50, significantly above what would be expected for a random guess, ~0.25. Further work includes measuring the performance of different data-driven classifiers as a function of spectral resolution.

6. **Jiwon Nam, PhD Student, Political Science**
The Double-Edged Sword of Democracy: Topic Model Analysis of International Climate Change Speeches. *Abstract:* Despite the quarter century of continuous effort, climate change negotiations have failed in reaching a comprehensive agreement. What explains the gap between routine negotiations and states' inabilities to reach effective agreements? Do countries' levels of democracy influence reaching agreements through negotiation? To answer these questions I apply a Structural Topic Model to UNFCCC Conference of the Parties (COPs) speeches from 16th COPs to 25th COPs. After extracting 25 topics from the speeches, I focus on the deliberative dimension of democracy to evaluate whether it influences countries to negotiate in unique manners. Results suggest that high levels of deliberative democracy led to more state-centric bargaining priorities which indicates countries' willingness to participate in global climate change actions. However deliberative democracy happens to compel states to prioritize aspects that are beneficial to their individual interests instead of the global climate change agenda.
7. **Sajan Kumar, Postdoc, Astrophysics (Gamma-ray astronomy)**
Application of decision tree algorithm to classify signal and background events in IACTs. *Abstract:* Imaging Atmospheric Cherenkov Telescopes (IACTs) are designed to detect gamma rays from astronomical sources such as Supernova remnants, Pulsars, Active Galactic Nuclei etc. However, the sample of triggered events is dominated by cosmic-ray induced background events. Understanding and classify background events allows us improved signal to noise ratio for gamma-ray sources. This also allows for individual particle populations to be resolved within the background, such as cosmic ray electrons. In this poster, I will discuss the application of tree classification method Boosted decision tree to classify the signal and background events in IACTs, in particular for measuring the electron spectrum.
8. **Deepti Anand, Postdoc, Biological Sciences**
Single-cell sequencing of the developing lens. *Abstract:* The lens focuses light on the retina for optimal vision, and loss in its transparency is termed cataract. It has distinct cell populations, broadly classified as anterior epithelium (AE) and posteriorly located differentiated fiber cells (FCs). Further cell subpopulations are recognized in AE and FCs based on proliferative or differentiation status, respectively. Pathological changes in individual cells are hypothesized to cause human lamellar cataract. Thus, to gain insights into lens cell-specific transcript heterogeneity, I developed a workflow to obtain viable isolated mouse embryonic and newborn lens cell suspensions for single-cell RNA-sequencing (scRNA-seq). 10x Genomics tools were used to assign unique molecular identifiers to expressed transcripts and identify lens marker genes. These data show that scRNA-seq identifies distinct new cell populations in the lens. Further, its application to cataract animal models can identify disease-specific changes in individual cell types.respectively. Pathological changes in individual cells are hypothesized to cause human lamellar cataract. Thus, to gain insights into lens cell-specific transcript heterogeneity, I developed a workflow to obtain viable isolated mouse embryonic and newborn lens cell suspensions for single-cell RNA-sequencing (scRNA-seq). 10x Genomics tools were used to assign unique molecular identifiers to expressed transcripts and identify lens marker genes. These data show that scRNA-seq identifies distinct new cell populations in the lens. Further, its application to cataract animal models can identify disease-specific changes in individual cell types.heterogeneity across lens cells, I performed RNA-sequencing at single-cell resolution. I developed a protocol to dissociate mouse-lens for attaining a cell suspension and performed single-cell RNA-seq using 10X Genomics Chromium and Illumina NovaSeq 6000. The raw data was processed using the 10X genomics Cell Ranger and unique molecular identifiers (UMI) were assigned to the expressed transcripts. In addition, Seurat R-package was used for downstream analysis and to identify differentially expressed genes (DEGs) in the lens cell-population. These analyses identified lens epithelial and fiber-cell marker genes as well as several new cell populations. This new approach to study eye-lens at a single-cell level provides novel insights into molecular changes relevant for lens development and associated defects.
9. **XIbrahim Balogun, PhD Student, Infrastructure Systems Engineering**
Deep Learning Approach Towards Squat Isolation in a Multi-Embedded Track Geometry Defects. *Abstract:* Railroad defects is a major challenge that has received attention amongst railway specialists in recent years. All types of defects are observed to be deleterious to railroad safety if not attended to. In particular, squat defects are produced by wheel-rail dynamic impact, leaving a depression mark on the rail surface. In this research, deep neural networks are used to classify thousands of kilometers of railway track into two binary classes: squat(p) and no-squat(n). In order to mitigate large class imbalance, we consider several natural sampling and data augmentation methods. We demonstrate that data augmentation and segmentation techniques can yield a considerable improvement compared to traditional re-sampling approaches.

10. **Lian Ming, Master's Student, Astrophysics**
    Prompt Identification of Rapidly Evolving Astrophysical Transients. *Abstract:* Although slower evolving transients have been studied extensively, the "fast transients" are extremely difficult to discover as they brighten and fade away swiftly. Vera C. Rubin Observatory is a synoptic survey telescope under construction in Chile and is planned to start a 10-year survey photometric survey in 2024. With its ability to detect extremely faint objects (~24 in magnitude) and huge field of view (~10 deg^2), the telescope can scan through all the visible sky in only 3 days. It will generate about 20 TB of data and discover more than 1000 astronomical transients, including fast and slower transients, each night. Identifying the fast transients and planning follow-up observations is a significant and challenging task. A probability classifier was created for the identification of the "fast transients" with only three images acquired within one night. This is crucial to enable the scheduling of follow-up observation before the objects fade. To do this, we used simulations of millions of well-studied transients to create a look-up table that enables an instantaneous classification, based on the observed color and brightness evolution.
11. **Matt Myers, PhD Student, Research Methodology**
    A data simulation program to assess machine learning algorithms for classification problems. *Abstract:* This paper presents a data simulation tool designed to provide meaningful data for classification algorithms (e.g., random forests) in the context of educational research. Using Python's PANDAS, NUMPY, and RANDOM modules, each simulation generates datasets derived from user-specified global parameters, which include constraints for sample size, the numbers of predictors and the degree of statistical noise, as well as the range of distributional characteristics for the predictors. Sample size is partitioned into two or more categorical outcomes. Dataset characteristics are defined via RANDOM's pseudo-random number generation. To emulate the covariation among predictors typical of educational research, the program transforms initial unadjusted predictor values using randomly defined covariance matrices. This paper details the algorithm, then demonstrates its potential usage.
12. **Christopher Russell, Postdoc, Astrophysics**
    Using virtual reality to analyze multi-dimensional data sets. *Abstract:* Virtual reality's fundamental feature -- immersion -- makes it ideal for analyzing complex, multi-dimensional data sets. With virtual reality (VR), researchers are immersed into virtual representations of their data that they then explore using VR goggles and controllers, allowing the researcher to move anywhere and look in any direction within their data. The researcher can then identify features that are not possible with traditional, flat-screen visualization methods, making VR research tools particularly useful for analyzing complex data sets. In this poster, we will present our two VR research tools, one for 3D data sets and or for arbitrary phase-space data. The former gives the researcher the feeling of being in their time-evolving simulations as it runs, while the latter allows up to 8-dimensional phase space at once. We aim to establish new collaborations with other researchers interested in analyzing their data in VR, as well as meet other researchers working in VR.
13. **Emma Stell, PhD Student, Plant and Soil Sciences**
    Spatial biases influence estimates of soil respiration: how can we improve global predictions? *Abstract:* Soil respiration (Rs), the efflux of CO2 from soils to the atmosphere, is a major component of the terrestrial carbon cycle, but is poorly constrained. The global soil respiration database (SRDB) is a compilation of in situ Rs observations from around the globe that has been updated with new measurements. It is unclear if the addition of data has produced better global Rs estimates. We compared two versions of the SRDB to determine how more data influenced global Rs annual sum and associated uncertainty. A quantile regression forest model using SRDBv3 yielded a global Rs sum of 88.6 Pg C year-1 and uncertainty of 57.9 (S.D.) Pg C year-1, but using SRDBv5 yielded 96.5 Pg C year-1 and uncertainty of 73.4 (S.D.) Pg C year-1. We tested an optimized global distribution of measurements, which resulted in a sum of 96.4 ± 21.4 Pg C year-1. These results support current global estimates of Rs but highlight spatial biases that influence model parameterization and interpretation.
14. **Akshay Bhosale, PhD Student, Computer Engineering**
    Measuring the impact of Automatic Program Parallelization Techniques in Cetus v2.0. *Abstract:* Cetus is a source-to-source translator for programs written in the C language. The primary use is as a parallelizing compiler, translating C programs to equivalent C code annotated with OpenMP parallel directives. Cetus is a research platform to study parallelization techniques and related program transformations. Cetus was created out of a need for a state-of-the-art automatic parallelizer for multicores, written in a modern language and capable of performing analyses and transformations for today's architectures. This poster presents an in-depth evaluation of the existing and newly added analysis and transformation techniques in Cetus on a set of benchmark applications.

15. **Rachel Keown, PhD Student, Biology / Biochemistry**
"Novel DNA polymerases mined from metagenomic sequences fill the phenotypic. *Abstract:* Viruses are the most abundant and diverse biological entities on the planet and constitute a significant proportion of the genetic diversity on Earth. The vast majority of this diversity is not represented by isolated viral-host systems and has only been observed through sequencing of viral metagenomes (viromes) from environmental samples, and this information is stored in large, commonly incomplete databases. Viromes provide snapshots of viral genetic potential, and a wealth of information on viral community ecology, yet phenotypic information is largely unknown. These data also provide opportunities for exploring the biochemistry of novel viral enzymes. In this study, the in vitro biochemical characteristics of novel viral DNA polymerases was explored to test hypothesized differences in polymerase biochemistry according to protein sequence phylogeny. Thirty bacteriophage DNA Polymerase I (PolA) proteins from Estuarine viromes and reference viruses spanning a broad representation of currently known diversity were synthesized, expressed, and purified. Novel functionality was shown in multiple PolA clones. Intriguingly, some of the estuarine viral polymerases demonstrated moderate to strong innate DNA strand displacement activity at high enzyme concentration. Strand-displacing polymerases have found important technological applications where isothermal reactions are desirable. Biochemical data elucidated from this study has the potential to expand our understanding of the biology and ecological behavior of unknown viruses. Moreover, given the importance of viral DNA polymerases to biotechnology, novel viral polymerases discovered within viromes may be a rich source of biological material for further technological advancements for in vitro DNA amplification. Moreover, this work has the potential to fill in the phenotypic data gap in many metagenomic databases.

16. **Somayeh Khakpash, Postdoc, Astronomy, Data Science**
Generating magnification maps for gravitational microlensing using Neural Networks. *Abstract:* Looking at the night sky, what we see is a modified representation of what luminous objects really are. The gravity from massive galaxies between us and a distant astronomical object can create multiple resolved images of that object in a phenomenon called "strong lensing". Additionally, the multiple images will be also affected by the individual stars and objects along their line of sight, which is called "microlensing". To model the microlensing variability, magnification maps should be generated that determine how the overall effects of the individual stars changes the brightness of the images. Many maps should be generated to find the best one describing the observed variability, but this is a computationally expensive process. Here, we are using a deep neural network, a "variational autoencoder", to create lower dimension representations of these maps. Using the lower dimensional maps, we will train a neural network that can generate the maps from their physical parameters.

17. **Isaac Lam and Austin Kuba, PhD Students, Materials Science & Engineering**
Machine learning approach for C-V-f fingerprint analysis of recombination in perovskite solar cells. *Abstract:* Capacitance measurement techniques are powerful methods for characterizing semiconductor devices, especially thin-film solar cells. An underutilized technique is voltage dependent admittance spectroscopy (C-V-f), which has recently been used to characterize electronic loss mechanisms in CIGS solar cells. This technique measures capacitance over a broad band of frequency and voltage conditions, creating a "loss map" fingerprint that is a convolution of multiple electronic loss mechanisms. Quantitative analysis of this dataset is challenging due to the wide variety of possible loss pathways. By simulating a large dataset of C-V-f loss maps using drift-diffusion modeling varying critical parameters, a machine-learning algorithm may be trained to identify dominant loss mechanisms in experimentally obtained fingerprint loss maps. We investigate the application of this technique for thin film perovskite solar cells and find that surface recombination is a significant loss mechanism.

18. **Md Mozaharul Mottalib, PhD Student, Data Science**
Study on the effect of COVID-19 on weight gain trajectory. *Abstract:* The COVID-19 pandemic has been associated with weight gain among adults, but little is known about the weight of US children and adolescents. To evaluate pandemic-related changes in weight in school-aged youths, we compared the body mass index (BMI; calculated as weight in kilograms divided by height in meters squared) of youths aged 2 to 18 years before the diagnosis of the disease with BMI trajectory in the period after the diagnosis. Characteristics were examined on different trendline categories of BMI on age-grouped clusters.

19. **Austin Kuba (Co-Presenting with #17)**

20. **Elizabeth Smith, PhD Student, Ecological data science**
Spatial Variability of Soil Nitrogen, GPP and Biomass Relationships in the CONUS. *Abstract:* We provide a quantitative assessment of the three-dimensional spatial distribution of soil N across the conterminous United States (CONUS) using a digital soil mapping approach. We used a random forest-regression kriging algorithm to predict soil N concentrations and associated uncertainty across six soil depths (0-5, 5-15, 15-30, 30-60, 60-100, 100-200 cm) at 5 km spatial grids. Across CONUS, there is a strong spatial dependence of soil N, where soil N concentrations decrease but uncertainty increases with soil depth. We also compared our soil N predictions with satellite-derived gross primary production (GPP) and biomass from the National Biomass and Carbon Dataset. Finally, we used uncertainty information to optimize locations for designing future soil surveys.

STUDENT POSTER SESSION (VIRTUAL)

***Please visit the posters and recordings at: https://dsi.udel.edu/events/dsi-symposium-2021/posters/**
****Note: Number denotes breakout room**

1. **Kyungmin Lee, PhD Student, Data Science**
Extraction of Air Conditioner End Use Indicators from Proximal Infrared Remote Sensing of Buildings. *Abstract:* To analyze dynamics in complex urban energy systems, we use infrared imaging of a residential building in Singapore to characterize energy consumption related to user behavior through image processing and machine learning techniques. By detecting short-time scale variations in pixels, we find that the different pixels have distinct time series characteristics depending on their types. To segment the scene based on sequences of IR images, we then apply and assess random forest (RF), convolutional neural network (CNN) models. We also analyze time series of AC units and find that the time-dependence of these sources has characteristic on/off transitions that correspond to AC operations and potentially user behavior. By applying state-change model and change point detection method, we determine precisely on/off timings of AC units. This work has implications for improving energy monitoring systems and providing empirical evidence as an input to decision-making processes in energy policy.

1. **Angie Stephanie Gomez Fonseca, Master's Student, Urban Physics**
A Bayesian Approach to Identify Uncertainties in Atmospheric Modeling on Ground-Based HIS. *Abstract:* A framework that integrates an open source spectral library (HITRAN-HAPI) and statistical sampling through Markov Chain Monte Carlo (MCMC) is presented in this work to identify trace gases concentrations in the terrestrial atmosphere, their covariances, and uncertainties from ground-based hyperspectral images. A proof of concept of a simplified atmosphere inverse model was carried out, stating that the model has practical potential to contribute to the understanding of emissions. It was found that H2O introduces correlations with other molecules for having high absorption in LWIR range. Additionally, when atmospheric parameters surpass the instrument noise, their uncertainties tend to get smaller.

1. **MattHardy, Master's Student, Wildlife Ecology**
A Remote Sensing Biosecurity Mechanism for the Poultry Industry. *Abstract:* Avian influenza virus (AIV) risk at poultry operations increases due to interactions with wild waterfowl. I'm examining the potential use of combined data streams to create real-time mapping products which mitigate that risk. First, using high frequency GPS/GSM-telemetry, I'll generate wintering waterfowl distribution maps for the mid-Atlantic and California. Further, I'll determine if differences exist in movement patterns as a function of environmental variables, resource depletion, and hunting pressure via data collected from telemetry marked waterfowl. Second, I'll validate a novel approach to quantify waterfowl density in the airspace within the mid-Atlantic and CA during Nov-Mar 2020–2022 to pinpoint areas of high potential AIV contact at poultry farms relative to overall waterfowl density in the airspace using data from the NEXRAD weather surveillance radar (WSR) network. WSR techniques provide a comprehensive assessment of bird activity both in the airspace, and on the ground.

1. **Kristina Holton, Certificate Student, Bioinformatics**
Exploring the influences of functional connectivity architecture on cortical thickness networks in patients with early psychosis. *Abstract:* Cortical thickness and functional connectivity are two parallel approaches that have been widely used to gain insights into psychotic disorders. Significant abnormalities in these modalities, even at the early stage of psychosis, have been shown in the literature. However, few have studied them together or explored the influences of functional connectivity on cortical thickness networks. Prior studies using gyral-based atlases reported that cortical thickness regions susceptible to thickness reductions are strongly interconnected1 and that brain tissue volume loss in schizophrenia is conditioned by structural and functional connectivity2. With data-driven approaches, we assessed: 1) How are cortical thickness networks organized, functionally or structurally? 2) What features drive this organization, and do features vary by diagnosis? 3) What are the relationships between cortical thickness reductions and clinical assessments?

2. **Sachin Gavali, PhD Student, Bioinformatics**
Exploring the kinome using graph representation learning. *Abstract:* The human kinome contains a vast network of interacting kinases and substrates. Some of these kinases are very well studied and have proven to be useful as therapeutic targets, but many are poorly understood and their biological roles unknown. In this work, we use graph machine learning methods to explore the biological roles of these understudied kinases. We use PTM data to build an interaction network, and the Gene Ontology functional annotation data to enrich this network. We then use the node2vec algorithm to learn vector representation of the kinases and substrates in this network, and use these representations to predict novel interactions for understudied kinases. We then perform a functional enrichment analysis of the predictions to understand the biological roles of understudied kinases. For two of the understudied kinases - Q9UEE5 (STK17A) and Q9H422 (HIPK3) we ascertained that they play an important role in cancer activity and mediating an inflammatory immune response.

2. **Ginnie Sawyer-Morris, PhD Student, Data Science**
Dimensionality reduction of autism data using a basic autoencoder. *Abstract:* Human behavior is complex. Information captured through linear modeling, on its own, is insufficient to explain the complex interactions that likely exist across psychological constructs. In autism spectrum disorder (ASD) research, behavioral and symptoms questionnaires are numerous, complex, and multifaceted. Furthermore, fields within ASD each have their own set of theoretical constructs (Ewen, 2020), and the relevant degree and redundancy of information across items within theorized constructs remain unknown. Large dynamic databases coupled with analytic tools such as unsupervised artificial intelligence offer the ability to explore complex, potentially nonlinear interactions, through methods such as dimension reduction. The current study will demonstrate how a nonlinear modeling technique (i.e., basic autoencoder) can be used to reduce the dimensionality of psychological data within a large cross-sectional sample of individuals with ASD.

2. **James Korman, PhD Student, Social Science**
State Capture and the Role of Political Parties in Latin America. *Abstract:* This study explores the impact of political parties on state capture in Latin America. First, the impact of political party in power [years] is analyzed for a sample of 19 different Latin American countries with data ranging between the years 1996-2017. Second, the impact of political party in power [years] on state capture at varying levels of economic development as measured by GDPPC is then examined. The analysis provides support for the negative impacts of political party in power [years] on state capture. However, the results demonstrate that the impact of political party in power [years] on state capture can be mitigated the more economically developed a country becomes. Overall, the results suggest that a lack of political competition and horizontal accountability that political parties are able to provide in a given country results in enhanced levels of corruption and state capture.

3. **Piyush Mehta, PhD Student, Global Environment Change**
Mapping Changes in Global Area Equipped for Irrigation. *Abstract:* The intensification of irrigated agriculture has increased global crop production but resulted in widespread stress to freshwater resources. Ensuring that increases in irrigated production only occur in places where water is relatively abundant is a key objective of sustainable agriculture, and knowledge of the evolving extent of irrigated land is important for crop modelling and integrated water management. Here we utilize the latest sub-national irrigation statistics from various official sources to develop a gridded global product of the area equipped for irrigation (AEI) for the years 2000, 2005, 2010, and 2015 and to quantify changes in global irrigated areas since the start of the century. We also merge this information with data on physical and economic blue water scarcity to examine to what extent irrigation expansion has occurred in places already experiencing water stress.

3. **Parinaz Barakhshan, PhD Student, Engineering**
iCetus: A Semi-automatic Parallel Programming Assistant. *Abstract:* The iCetus tool is a new interactive parallelizer, providing users with a range of capabilities for the source-to-source transformation of C programs using OpenMP directives in shared memory machines. While the tool can parallelize code fully automatically for non-experts, power users can steer the parallelization process in a menu-driven way. iCetus which is still in its early stages of development is implemented as a web application to support the user through all phases of the program transformation process, including program analyses, and optimization.

3. **Qitong Wang, PhD Student, Ecology and Geography**
Exploratory Inconsistency Analysis of Energy-critical Materials Public Datasets. *Abstract:* Energy-critical minerals (ECM) are required for emerging sustainable energy sources. The estimates of the total volume of illicitly-sourced minerals in ECM trade flows remain ad hoc, anecdotal, and incomplete. There are no global measurements of the licit-illicit composition of ECM trade flows or their evolution over time. We explore the licit-illicit composition if ECM trade by measuring the inconsistency in energy-critical minerals public datasets. Our researches exposes multiple types of inconsistencies in public datasets, which include significant inconsistency of source countries' import-export logs within a single dataset and across multiple datasets, large annual fluctuations of measurements of different commercializing materials, and potential inconsistencies between the traded amount and the traded currency. This work sets the stage for a large program of interdisciplinary research to be conducted at UD supported by NSF grant 2039857 (PI Klinger).

3. **Amin Boukari, High School Student**
Aragonite Saturation as an Indicator for Oyster Habitat Health in Delaware Inland Bays. *Abstract:* Oyster farming in Delaware is a crucial industry, bringing in $300,000 to $500,000 in sales every year. Oysters use calcium carbonate ions in the form of aragonite and calcite to form their shells. Ocean acidification can lead to a decrease in carbonate ions making forming these shells difficult. When aragonite saturation state falls below 3, calcifying organisms become stressed and when it drops below 1, their shells begin to dissolve. Therefore, measuring the aragonite saturation state yields crucial insight into the suitability of habitats to support oyster growth. This project aimed to calculate the aragonite saturation state from seven sites within Delaware Inland Bays to determine their feasibility in supporting the establishment of oyster farms. Monitoring was conducted biweekly from July to November 2020 and 2021.

DSI SYMPOSIUM PLANNING COMMITTEE

Cathy Wu, *Director, Data Science Institute*

Federica Bianco, *Co-Chair*

Benjamin Bagozzi, *Co-Chair*

Michael Blaustein
Richard Braun
Parinaz Barakhshan
Chuming Chen
Susan Conaty-Buck
Adam Davey
Kyle Davis
Greg Dobler

John Gizis
Medina Jackson-Browne
Somayeh Khakpash
Matthew Mauriello
Farid Qamar
Amy Shober
Claude Tameze (Lincoln U.)
Andrea Trungold

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

JOIN US ON TWITTER AND SLACK

@dsiudel

hashtag: #datascienceUD21

Slack channel:
bit.ly/UDDSIsymposium2021

THANK YOU TO OUR INDUSTRY PARTNERS

Please Visit Our Industry Partners in the Atrium