HDR² From Harnessing the Data Revolution to Harvesting the Data Revolution

Wednesday 26 October 2022 - Thursday 27 October 2022

Book of Abstracts

Contents

Keynote talk: Let's make Data Science more like Science 1

Keynote Speaker - An integrative, university-wide approach to data science 1

Lightning talks 1

Poster Session 2

Intro Talk - AI expert (Adam Smith BU) 2

Panel Board 2

Open Mic 2

Breakout Session 3

Breakout Session #1: Research ML Challgnes 3

Breakout Session #2: Education and Outreach 3

Breakout Session #3: Data and Cyberinfrastructure 3

Welcome and HDR Update from NSF 3

Open Mic Session+Discussion 3

Concluding Remarks 3

Intro 4

Scenario discussion 4

Report out and Closing 4

MLCommons Research: An Open-Source Machine Learning Ecosystem Linking Industry, Government and Academia to Democratize AI Technologies for Everyone 4

Data and Cyberinfrastructure Panel: HDR Institute Imageomics 5

Open mic 5

Welcome and Intro 5

Lightening talk 5

Poster session 5

Modernizing Water and Wastewater Treatment Through Data Science Education and Research . . . 5

Data science training and practices: preparing a diverse workforce via academic and industrial partnership . . . 6

Professor . . . 6

Drift vs Shift: Decoupling Trends and Changepoint Analysis . . . 7

HDR DSC: Data Science for Social Good in Urban Areas . . . 8

Data Science Career Pathways in the Inland Empire (DS-PATH) . . . 9

Engaging Undergraduates in Data and Decisions Research at the Engineering/Biology Interface . . . 9

AI Across the Statewide Curriculum . . . 10

Collaborative Research: HDR DSC: Building Capacity in Data Science through Biodiversity, Conservation, and General Education . . . 11

D4 (Dependable Data Driven Discovery) Institute . . . 12

The National Data Mine Network . . . 12

Establishing the Metropolitan Chicago Data Science Corps (MCDC) . . . 13

Deletion Resilient Group Testing . . . 13

Data Science Corps: Wrangle, Analyze, Visualize –Experiential Learning in Local Community Organizations . . . 14

Multiscale Basis Dictionaries on Higher-Order Networks . . . 15

T-Tripods: Supporting the Foundation of a Data Intensive Studies Center . . . 15

The Earth Data Science Corps: a model for teaching & learning environmental data science skills . . . 16

Central Coast Data Science Partnership: Training a New Generation of Data Scientists . . . 16

Training next generation data scientists for Energy Transition . . . 17

Infusion of data science and computation into engineering curricula . . . 17

DAMAD Science CorpsThe Delaware And MiD-Atlantic Data Science Corps . . . 18

Interdisciplinary Traineeship for Socially Responsible and Engaged Data Scientists (iTREDS) . . . 19

Cost-aware Generalized alpha-investing for Multiple Hypothesis Testing . . . 19

HDR DSC: Collaborative Research: Transforming Data Science Education through a Portable and Sustainable Anthropocentric Data Analytics for Community Enrichment (ADACE) Program . . . 20

Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) . . . 21

The I-GUIDE Cyberinfrastructure Platform . . . 22

Convergence Curriculum for Geospatial Data Science . . . 22

iHARP- Harnessing Data and Model Revolution in the Polar Regions . . . 23

Institute for Data-Driven Dynamical Desgin . . . 24

HDR IDEAL - The Institute for Data, Econometrics, Algorithms, and Learning . . . 24

TRIPODS Phase II: Institute for Data, Econometrics, Algorithms, and Learning (IDEAL) . . . 25

Lightening talk I . . . 26

DIMACS DATA-INSPIRE at Rutgers: Conditioned Weiner Processes and a Rigorous Probabilistic Analysis of Dynamics . . . 26

DATA INSPIRE: Morse Graphs can effectively estimate the Regions of Attraction (RoAs) of dynamical systems, including closed-box ones. . . . 27

A3D3: Community Engagement, Education, Outreach . . . 27

A3D3: Accelerating AI Algorithms . . . 28

Research: Inductive Bias in Learning . . . 28

Complex Data Structures . . . 28

Data and Cyberinfrastructure around the HDR ecosystem - the I-GUIDE perspective . . . 28

Imageomics: Images as the Source of Information about Life . . . 29

Keynote Speakers / **1**

Keynote talk: Let's make Data Science more like Science

Corresponding Author: juliana.freire@nyu.edu

Data-driven exploration has revolutionized science and led to the establishment of Data Science as a new discipline that integrates approaches from computer science – including data management, visualization, machine learning – statistics, applied mathematics, and many application domains. I will give my perspective of how the field emerged and evolved over the past decade, and the virtuous cycle it has enabled which fuels interdisciplinary research that derives new problems and solutions for multiple areas.

A critical challenge in data science is how to empower domain experts to engage in data-driven exploration. While computing and storage are essentially free and data is abundant, we need humans in the loop to generate insights from data. Toward this end, there have been many efforts that aim to democratize data science, and today it is relatively easy to derive results. But it is also easy to derive incorrect results. I will give examples of common mistakes and problems that can affect results and that are hard to detect, and argue that, akin to natural systems, we must experiment with and observe data science pipelines to understand their behavior, assess the validity and properly explain their results. In essence, we need to make Data Science more like science and work towards democratizing trust and robustness.

Data & Cyberinfrastructure:

Education and Outreach:

Research:

Keynote Speakers / **2**

Keynote Speaker - An integrative, university-wide approach to data science

Corresponding Author: sstone3@uw.edu

Experience of Data Science education and Ecosystem building as a center/institute leader within big university, and forward looking about advise to build a successful national-wise HDR Ecosystem.

Data & Cyberinfrastructure:

Education and Outreach:

Research:

Poster session I / **4**

Lightning talks

Each poster speaker can deliver a 2 min highlight.
In addition, a 10 min video is expected to be uploaded to the agenda by Oct 14.

Poster session I / **5**

Poster Session

Research Panel / **6**

Intro Talk - AI expert (Adam Smith BU)

Data & Cyberinfrastructure:

Education and Outreach:

Research:

Research Panel / **7**

Panel Board

Corresponding Authors: berger-wolf.1@osu.edu, hassani@seas.upenn.edu, etoberer@mines.edu

Chair: Prof. Adam Smith

- Prof. Eric Toberer, Colorado School of Mines, HDR Institute: Institute for Data Driven Dynamical Design
- Prof. Tanya Berger-Wolf, Ohio State University, HDR Institute: Imageomics: A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning
- Prof. Hamed Hassani, University of Pennsylvania, TRIPODS Phase II: EnCORE: Institute for Emerging CORE Methods in Data Science

Data & Cyberinfrastructure:

Education and Outreach:

Research:

Research Panel / **8**

Open Mic

12

Breakout Session

Breakout session Summaries / **15**

Breakout Session #1: Research ML Challgnes

Corresponding Author: philip.coleman.harris@cern.ch

Breakout session Summaries / **16**

Breakout Session #2: Education and Outreach

Breakout session Summaries / **17**

Breakout Session #3: Data and Cyberinfrastructure

Data & Cyberinfrastructure:

Education and Outreach:

Research:

18

Welcome and HDR Update from NSF

Corresponding Author: awalton@nsf.gov

Closeout / **19**

Open Mic Session+Discussion

Closeout / **20**

Concluding Remarks

Corresponding Author: schsu@uw.edu

Data & Cyberinfrastructure:

Education and Outreach:

Research:

Education and Outreach - CARE & FAIR principles / **21**

Intro

Corresponding Author: naqu1888@colorado.edu

Education and Outreach - CARE & FAIR principles / **22**

Scenario discussion

Panelist:
Michelle Holko @Google

Divide audience into small groups, each of which discusses a scenario
- who controls the data?
- how do you handle the data?
- how do you communicate about the data?
- where are the data stored?

Education and Outreach - CARE & FAIR principles / **23**

Report out and Closing

Data and Cyberinfrastructure Panel / **24**

MLCommons Research: An Open-Source Machine Learning Ecosystem Linking Industry, Government and Academia to Democratize AI Technologies for Everyone

Corresponding Author: gcfexchange@gmail.com

MLCommons Research is described as a community to collaborate with and as a model for similar communities. Its working groups cover Algorithms, Datasets, Platforms, Storage and Science and Medical applications. MLCommons involves 62 companies, 6 DOE laboratories, 11 Universities with a flagship benchmark set MLPerf and the mission of "Accelerating machine learning innovation to benefit everyone."MLCommons thrusts are Best Practice (software and ontologies), Datasets (Identify and package good ones; develop new ones) and Benchmarks. We describe 7 Science benchmarks with open data and models and a possible direction of Foundation AI models for Science. We note the importance of Clouds, Supercomputers and their integration HPC Clouds.

Data & Cyberinfrastructure:

Education and Outreach:

Research:

Data and Cyberinfrastructure Panel / **25**

Data and Cyberinfrastructure Panel: HDR Institute Imageomics

Corresponding Author: stewart@rpi.edu

Data and Cyberinfrastructure Panel / **26**

Open mic

27

Welcome and Intro

Corresponding Author: schsu@uw.edu

28

Lightening talk

Each poster speaker can deliver a 2 min highlight.
In addition, a 10 min video is expected to be uploaded to the agenda by Oct 14.

Poster session I / **29**

Poster session

Poster session I / **30**

Modernizing Water and Wastewater Treatment Through Data Science Education and Research

Authors: Amanda Hering[1]; Tzahi Cath[2]; Doug Nychka[2]; Michael Poor[1]; Greg Hamerly[1]

[1] *Baylor University*

[2] *Colorado School of Mines*

Corresponding Author: mandy_hering@baylor.edu

The field of water and wastewater treatment (W/WWT) is brimming with data analysis opportunities, but many working in the field lack the skills needed to navigate and extract knowledge from this data. This project began in 2019 with the development of a prerequisite-free course in data science and a five-week summer undergraduate research program. Both were populated with real problems and data from our W/WWT industry partners. The program has evolved to include a workshop in data science training for industry professionals, an advanced internship program, and a summer workshop in data science for environmental engineering students. A textbook from the course material is being developed, along with a dataverse that archives and publicizes all of the datasets that we have received from our stakeholder partners. Furthermore, appropriate but rarely used statistical methods in W/WWT have been identified and are being developed for these special types of problems.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 31

Data science training and practices: preparing a diverse workforce via academic and industrial partnership

Author: Babak Shahbaba[None]

Corresponding Author: babaks@uci.edu

We have developed a program comprising of curricular, training, and mentoring components to build a diverse community of learners. Our first cohort included 32 student fellows, including 28 (87%) women/underrepresented minority students, recruited from the three participating institutes: UC Irvine, CSU Fullerton, and Cypress College (representing the three-tiered structure of higher education at the state of California). During the academic year, all students took data science related courses, including four new courses (across the three institutes) developed by our program. Over the summer, our fellows participated in a bootcamp and research experience at UC Irvine. The bootcamp included: 1) curricular content on topics such as exploratory data analysis, modeling, inference, and prediction; 2) software and computing components including R and GitHub; and 3) student engagement opportunities with guest speakers on ethics in data science, a graduate student panel, workshops on career development, and team-science activities. For the research component, the fellows worked on a wide range of research projects related to neuroscience, cancer research, environmental studies, education, and health. Finally, through seminars and symposiums, we have ensured that our fellows are exposed to real-world problems and have established a network with non-academic partners in order to quickly assimilate in the data science workforce after graduation.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 32

Professor

Author: Jeffrey Errington[None]

Co-authors: Bina Ramamurthy [1]; Erin Rowley [1]; Kristen Moore [1]; Liesl Folks [2]

[1] *University at Buffalo*

[2] *University of Arizona*

Corresponding Author: jerring@buffalo.edu

There is significant demand for a workforce that is proficient in data science and analytics. Employers seek graduates with an ability to (1) understand, interpret, and analyze data, (2) effectively communicate results that stem from the analysis of data, (3) practice the ethical use of data, and (4) apply data science concepts to solve practical problems with real-world relevance. While the dissemination of data science competencies has been emphasized in some disciplines (e.g., computer science), the broad delivery of these skills to college graduates has been slow to evolve. The aim of this project is to develop and implement a scalable, innovative program, termed "Connecting the Dots", for delivery of data science competencies to students pursuing an undergraduate engineering degree.

Research:

Education and Outreach:

Yes

Data & Cyberinfrastructure:

Poster session I / **33**

Drift vs Shift: Decoupling Trends and Changepoint Analysis

Author: David Matteson[1]

Co-authors: Haoxuan Wu [1]; Sean Ryan [1]

[1] *Cornell University*

Corresponding Author: matteson@cornell.edu

Distinguishing between global or macro patterns and local or micro fluctuations helps summarize the evolution of complex non-stationary dynamic systems. Herein, we focus on making distinctions between drift and shifts. Drift describes the micro-level evolution of a process. This may appear as variation about gradual trends. In contrast, shifts refer to discontinuities, rapid changes, or major breaks in trend. These represent macro-level changes in a process. Both trends and shifts might be mechanistically or stochastically generated and/or modeled. However, the underlying causes of shifts are typically different from those of drift. While understanding such differences is a prime objective, this first requires distinguishing shifts from drift.

We introduce a new approach for decoupling trends (drift) and changepoints (shifts) in time series. Our locally adaptive model-based approach for robustly decoupling combines Bayesian trend filtering and machine learning based regularization. An over-parameterized Bayesian dynamic linear model (DLM) is first applied to characterize drift. Then a weighted penalized likelihood estimator is paired with the estimated DLM posterior distribution to identify shifts. We show how Bayesian DLMs specified with so-called shrinkage priors can provide smooth estimates of underlying trends in the presence of complex noise components. However, their inability to shrink exactly to zero inhibits direct changepoint detection. In contrast, penalized likelihood methods are highly effective in locating changepoints. However, they require data with simple patterns in both signal and noise. The proposed decoupling approach combines the strengths of both, i.e. the flexibility of Bayesian DLMs with the hard thresholding property of penalized likelihood estimators, to provide

changepoint analysis in complex, modern settings. The proposed framework is outlier robust and can identify a variety of changes, including in mean and slope. It is also easily extended for analysis of parameter changes in time-varying parameter models like dynamic regressions. We illustrate the flexibility and contrast the performance and robustness of our approach with several alternative methods across a wide range of simulations and application examples.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **34**

HDR DSC: Data Science for Social Good in Urban Areas

Author: Aryya Gangopadhyay[1]

[1] *UMBC*

Corresponding Author: gangopad@umbc.edu

The goal of this project is to develop a team-based data science corps program for undergraduate students from Computer Science, Information Systems, and Business integrating both academic training as well as hands-on experience through real-world data science projects. This project is a collaborative effort with the University of Maryland Baltimore County as the coordinating as well as an implementing organization, and the University of Baltimore, Towson University, and Bowie State University as implementing organizations. This project focuses on the city of Baltimore as an exemplar for other cities in the US and across the globe. The project team is collaborating with a number of communities in the city of Baltimore to integrate real-world data science projects into classroom instruction in data science. The specific objectives of this project are as follows: (i) Develop the technical, analytical, modeling, and critical thinking skills that are key to success as a data science professional; (ii) Connect a cohort of students to communities, organizations, and projects that can benefit from the power of data science; (iii) Nurture and support innovative thinking in solving some of the key challenges facing the real world; (iv) Promote a better understanding of the power and pitfalls of data-driven discoveries to improve the quality of life in urban communities; (v) Increase the data science workforce capacity to support this critical area that is of growing importance in society; and finally, (vi) Evaluate the effect of the proposed data science corps on student learning.

The project team is creating a core set of knowledge for developing solutions for real-world urban settings. The core set of knowledge includes data collection and cleaning, data analysis using machine learning and deep learning techniques, data visualization including geospatial data and virtual reality, data privacy and security, and infrastructure for smart cities including IoT-based sensor networks. The data science corps program will have three main phases: instructional phase, academic research, and real-world team projects, spanning one calendar year. Examples of team projects include: (i) developing community-based indicators that are compiled from open data portals and parametric and non-parametric statistical techniques to understand the relationship between urban sustainability and a range of factors including cleanliness and environment, crime and safety, business and economics, social and political, housing, health, and education; (ii) combining AI models in edge devices with autonomous systems for crime detection and prevention, accident prediction in road networks, and emergency response; (iii) combining sensor data and social media for automated information extraction, validation, and quality checks that can be beneficial to both citizens and emergency managers in crisis situations such as flash floods; and (iv) developing augmented reality-based systems that leverage systems such as Microsoft HoloLens and mobile devices for building evacuation.

Each year, around 25 undergraduate students participate in the data science corps program, many of whom subsequently work in internships and/or full-time data science positions. The data science

corps students are conducting research on AI on the edge combined with autonomous systems including unmanned ground and aerial vehicles for monitoring and real-time intervention in smart cities.

The outreach programs include presenting research papers in various peer-reviewed conferences and publishing in peer-reviewed journals. The project team is also engaged with community partners through the Baltimore Neighborhood Indicator Alliance (https://bniajfi.org/). We are also discussing possible collaborations with the Maryland Department of Transportation and the Department of Homeland Security, Prince Georges County Fire Department, and FEMA.

Research:

https://datasciencecorps.umbc.edu/publications-resources

Education and Outreach:

https://datasciencecorps.umbc.edu/training-modules

Data & Cyberinfrastructure:

https://datasciencecorps.umbc.edu/baltimore-data-week

Poster session I / **35**

Data Science Career Pathways in the Inland Empire (DS-PATH)

Authors: Paea LePendu[1]; Mariam Salloum[1]

[1] *UC Riverside*

Corresponding Author: paea.lependu@ucr.edu

The Data Science Career Pathways in the Inland Empire (DS-PATH) is a partnership that brings together 4-year and 2-year Universities and Colleges with a common goal of creating flexible pathways that will equip underrepresented students to become skilled and knowledgeable professionals in Data Science (DS). The partnership consists of six Hispanic Serving Institutions and covers all three segments of California's public higher education system, namely, the University of California, the California State University and the Community College System. So far, in its first year, DS-PATH has launched a minor in DS, a Masters degree in Computational DS and a Bridge Program for non-computing majors, and hosted 38 undergraduate students for a summer research experience. DS-PATH has also engaged four local school districts for grades 6-12 via an HDR Supplement aiming to design, launch and train teachers for a new DS curriculum for high school students along with expanding the successful DS Academy outreach program for middle school.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **36**

Engaging Undergraduates in Data and Decisions Research at the Engineering/Biology Interface

Author: David Schmale[1]

Co-authors: Birol Ozturk [2]; Eddie Red [3]; Michael Wolyniak [4]; Shane Ross [1]

[1] *Virginia Tech*

[2] *Morgan State University*

[3] *Morehouse College*

[4] *Hampden Sydney College*

Corresponding Author: dschmale@vt.edu

The ultimate goal of our program is to provide interdisciplinary education and research opportunities in data and decisions science for undergraduate students who are experts in a core discipline of engineering or biology, but who are also proficient in the alternate discipline. We are training students with complementary disciplinary expertise that can address problems at the engineering/biology interface, and show them, via stimulating grand challenge problems, the utility of data science techniques, thereby promoting data literacy and providing basic training in data science to key members of the science and engineering workforce. We have launched a unique HDR DSC program at Virginia Tech (coordinating organization), Morehouse College (HBCU for men, Georgia, implementing organization), Morgan State (HBCU for men and women, Maryland, implementing organization), and Hampden-Sydney College (all-male college, Virginia, implementing organization). A new collaborative, multi-university capstone course 'Solving Big Problems with Big Data'is being taught simultaneously at all four universities. Through this course, multi-university teams of students work together to identify relevant broad social, global, economic, cultural and technical needs and constraints, and determine the ways in which their complementary technical skills contribute to addressing complex data science grand challenges at the engineering/biology interface. In Fall, 2021, the project provided data science educational opportunities for 26 undergraduate students and one graduate student through the first course offering. Students were introduced to professionals in various sectors through 8 different stakeholder presentations, where they learned challenges of those sectors. In Summer, 2022, the project provided paid summer undergraduate data science research opportunities for 21 participants in 10 labs at Virginia Tech. This Fall, 2022, the project is providing educational opportunities for an additional 22 undergraduates that are enrolled in our undergraduate course for its second course offering.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 37

AI Across the Statewide Curriculum

Authors: Aavudai Anandhi Swamy[1]; Christina Gardner-McCune[2]; James Hoover[2]; Jennifer Drew[3]; Raquel Dias[2]; Satyanarayan Dev[1]; Sebastian Galindo[2]

[1] *Florida Agricultural and Mechanical University*

[2] *University of Florida*

[3] *University of Florida*

Corresponding Authors: satyanarayan.dev@famu.edu, jdrew@ufl.edu

An interdisciplinary team from the University of Florida and Florida Agricultural and Mechanical University are leading a project to enhance diversity, access, impact of a strong AI curriculum. Artificial intelligence is poised to make unprecedented impacts across all aspects of our society. Developing technical expertise in AI or relegating AI education to the computer and data science disciplines is not sufficient to develop a diverse and prepared AI workforce. The more pressing challenge is

to fully embrace the interdisciplinarity of effective AI by building flexible, inclusive learning pathways that allow students of diverse educational backgrounds and technical maturity to engage these emerging technologies and help solve the complex, real-world problems affecting our communities. Meeting this grand challenge requires acknowledging the importance of and intentionally building cross-institutional and cross-disciplinary pathways to ensure that our future AI-enabled workforce has the diversity of backgrounds, experiences and expertise necessary to engage our most difficult problems ethically and equitably.

The objective of this project is to jump-start our long-term efforts to facilitate unfettered access to cutting-edge technologies, expertise and experiential learning resources being developed at UF to diverse students across Florida and beyond, providing the next-generation workforce with the skills they need to work across institutional, disciplinary and historical disparity boundaries to solve real-world problems affecting our communities. To achieve this objective, our team of researchers and educators from UF and our partnerinstitution, Florida Agricultural and Mechanical University (FAMU) will pursue four aims:
1. Facilitate virtual AI curriculum across the state.
2. Expand participation of diverse students.
3. Address real-world AI needs with experiential learning.
4. Assessment of student gains and project evaluation for program sustainaibility

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 38

Collaborative Research: HDR DSC: Building Capacity in Data Science through Biodiversity, Conservation, and General Education

Author: Kathleen Prudic[1]

Co-authors: Greta Binford [2]; Ellen Bledsoe [1]; Ethan Davis [2]; Jeremy McWilliams [2]; Jeffrey Oliver [1]; Mila Pruiett [2]; Jill Williams [1]

[1] *University of Arizona*

[2] *Lewis and Clark College*

Corresponding Author: klprudic@arizona.edu

The workforce demand for data analysts and data scientists exceeds the current capacity for higher education to produce this skilled workforce. Our overall goal is to develop scalable, portable data science education that can be readily incorporated into existing programs concentrating on STEM with ecology, biodiversity, and conservation. We will do this by creating multiple curricular data science on-ramps for a broad range of students early in their undergraduate training through general education courses and foundational major courses using inclusive and expansive pedagogy techniques more common in liberal arts education. The expected outcomes from these activities are (1) development of reusable data science modules and courses that can be deployed into existing undergraduate general education and major curricula, (2) the ability for a broad range of conservation interested students to access real-world data science training they are passionate about at an early stage of their education, and (3) training and support mechanisms for undergraduate educators who wish to add data science to their curricula. The products of this proposed multi-institutional Data Science Corps program are designed to be generally extensible to other higher educational institutions and majors through open data and open science, providing capacity to rapidly deploy data science training.

Research:

Education and Outreach:

HDR-Data Science Corp

Data & Cyberinfrastructure:

Poster session I / 39

D4 (Dependable Data Driven Discovery) Institute

Authors: Hridesh Rajan[1]; Pavan Aduri[1]; Daniel Nettleton[1]; Eric Weber[1]

[1] *Iowa State University*

Corresponding Author: hridesh@iastate.edu

Data-driven discoveries are permeating critical fabrics of society. However, unreliable discoveries lead to decisions that can have far-reaching and catastrophic consequences on society, defense, and to individuals. This makes the dependability of data-science lifecycles producing discoveries and decisions a critical issue that requires a new holistic view and formal foundations. Furthermore, while the notion of dependability is well-studied in the computer-systems literature, challenges in data science push the boundary of existing knowledge into the unknown. This project, the Dependable Data-Driven Discovery (D4) Institute at Iowa State University, is advancing foundational research on ensuring that data-driven discoveries are of high quality. The D4 Institute advances the theoretical foundations of data science by fostering foundational research to enable understanding of the risks to the dependability of data-science lifecycles, formalizing the rigorous mathematical basis of the measures of dependability for data science lifecycles, and identifying mechanisms to create dependable data-science lifecycles. The institute is facilitating transdisciplinary training of a diverse cadre of data scientists through activities such as the Midwest Big Data Summer School and the TADS Lunch-n-Learn. Phase I focuses on a subset of the Data Science lifecycle and 4 risks (i.e., complexity, uncertainty, resource constraints, and freshness).

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 40

The National Data Mine Network

Author: Mark Daniel Ward[1]

[1] *Purdue University*

Corresponding Author: mdw@purdue.edu

The National Data Mine Network launched in August 2022. Our students work on data-driven projects with our Corporate Partners and with faculty members. The Corporate Partners working with NDMN students this year include Bayer (2 projects), Convo, John Deere (2 projects), Indiana Family and Social Services Administration, Inogen, Lockheed Martin, Merck, Raytheon (2 projects), Sandia National Laboratories, and the USDA US Forest Service. We also have a partnership with The University of Virginia's Data Justice Academy. The undergraduate student participants this year are studying at 38 Minority Serving Institutions during the 2022-23 academic year, while they are working on their research projects with The National Data Mine Network.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **41**

Establishing the Metropolitan Chicago Data Science Corps (MCDC)

Authors: Suzan van der Lee[1]; Michelle Birkett[1]; Mark Potosnak[2]

Co-authors: Eunice Santos [3]; Pascal Paschos [4]; Nadja Insel [5]; Yoo-Seong Song [3]; Arend Kuyper [1]; Francisco Iacobelli [5]; Sara Woods [1]; Denise Drane [1]; Bennett Goldberg [1]; Matthew Sperry [1]

[1] *Northwestern University*

[2] *DePaul University*

[3] *UIUC*

[4] *Chicago State University*

[5] *Northeastern Illinois University*

Corresponding Author: suzan@northwestern.edu

We established the Metropolitan Chicago Data science Corps (MCDC) in the Fall of 2021. MCDC is a partnership between five Chicago-area universities and local not-for-profit organizations. It serves data science needs of the organizations and provides real world data science questions, data sets and experience for data science students. Goals of MCDC are to advance data-driven decision making, build community, facilitate data utility and access, develop the data science workforce, and implement a learning system for dissemination. MCDC works at the intersections of data science with environmental sciences, health sciences, and social sciences. Data science students who participate in MCDC take foundational data science course work and enroll in a practicum. During the practicum students develop skills in 9 different areas of data acumen and are exposed to the multifaceted aspects of applying data science with community partners. Subsequent summer internships with community partners (Data science Application for Undergraduates (DAU)) allow more intensive collaborative applications of data science in service of the goals of the community partner.
During the 2021-2022 academic year about seventy students participated in practica and nine summer interns participated in a DAU, partnering with a total of eleven different non-profit organizations. Faculty served as practicum instructors and community-specific project mentors. Seventy five MCDC community partners, students, and faculty gathered at the inaugural MCDC workshop in June and shared best practices, experiences, results, ideas, and goals. Faculty development was supported by the workshop, an interactive on-line collaboration platform, committees' activities, and a collection of resources curated by MCDC researchers with expertise in community-academic partnerships.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **42**

Deletion Resilient Group Testing

Authors: Haodong Yang[1]; Nikita Polyanskii[None]; Venkata Gandikota[1]

[1] *Syracuse University*

Corresponding Author: vsgandik@syr.edu

Group testing is the study of pooling strategies that allow the identification of a small set of k defective items among a population of n using a small number of pooled tests. State-of-the-art testing schemes have shown that \Theta(k log n) schemes are both necessary and sufficient for the purpose which provides large gains when k is small (sublinear in n). However, these schemes are not resilient to deletion noise. In this work, we explore group testing algorithms resilient to deletion channels. We provide lower bounds, sufficient conditions, construction of matrices that meet the sufficiency condition and a decoding algorithm to recover the set of all k defective items.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **43**

Data Science Corps: Wrangle, Analyze, Visualize –Experiential Learning in Local Community Organizations

Author: Valerie Barr[None]

Co-authors: Benjamin Baumer [1]; Brian Candido [2]; Ileana Vasu [3]; Jaime Davila [4]; Matthew Rattigan [5]; Nicholas Horton [6]; Randi Garcia [1]

[1] *Smith College*

[2] *Springfield Technical Community College*

[3] *Holyoke Community College*

[4] *Hampshire College*

[5] *UMASS-Amherst*

[6] *Amherst College*

Corresponding Author: vbarr@bard.edu

While coursework provides undergraduate data science students
with some relevant analytic skills, many are not given the rich experiences with data and computing they need to be successful in the workplace. Additionally, students often have limited exposure to team-based data science and the principles and tools of collaboration that are encountered outside of school. The DSC-WAV program is a data
science workforce development project in which teams of undergraduate sophomores and juniors work with a local non-profit organization on a data-focused problem. To help students develop a sense of agency and improve confidence in their technical and non-technical data science skills, the project promotes a team-based approach to data science, adopting several processes and tools intended to facilitate this collaboration. Evidence from the project evaluation, including participant survey and interview data, shows the degree to which the project has successfully engaged students in team-based data science, and how the project changed the students'perceptions of their technical and non-technical skills. The project also supports development of a data science workforce pipeline by helping community colleges establish data science programs and align with 4-year institutions in order to facilitate student transfer as data science majors.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **44**

Multiscale Basis Dictionaries on Higher-Order Networks

Author: Naoki Saito[1]

Co-authors: Stefan Schonsheck [1]; Eugene Shvarts [1]

[1] *University of California, Davis*

Corresponding Author: saito@math.ucdavis.edu

We have generalized the multiscale basis dictionaries (e.g., the Haar-Walsh wavelet packet dictionary and local cosine dictionary), which were developed for digital signals and images sampled on regular lattices and have a proven track record of success (e.g., audio/image compression, feature extraction, etc.), to for the graph setting. Our previous such basis dictionaries (e.g., Generalized Haar-Walsh Transform, Hierarchical Graph Laplacian Eigen Transform, Natural Graph Wavelet Packets, and their relatives) were developed for analyzing data recorded on nodes of a given graph. In this work, we propose their generalization for analyzing *data recorded on edges or on faces* (i.e., triangles) of higher-order networks, in particular, *simplicial complexes* (e.g., a triangular mesh of a manifold). The key idea is to use the *Hodge Laplacians* and their variants for hierarchical partitioning of edges or faces, and then build localized basis functions on those subsets. We demonstrate their usefulness for data approximation on simplicial complexes generated from a co-authorship/citation dataset and an ocean current/flow dataset. This is a vertically-integrated collaboration among a faculty member, a postdoc, and a graduate student.

Research:

x

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **45**

T-Tripods: Supporting the Foundation of a Data Intensive Studies Center

Author: Lenore J Cowen[1]

Co-authors: Bert Huang [1]; Misha Kilmer [1]; Eric Miller [1]; Abani Patra

[1] *Tufts University*

Corresponding Author: lenore.cowen@tufts.edu

The Tufts University T-Tripods Phase I Tripods institute supports interdisciplinary research and learning in the foundations of data science, fostering collaboration among researchers in computer science, mathematics and electrical and computer engineering departments at Tufts, as well as connecting to scientists and scholars in a wide range of application domains.

The three focus areas for the institute are R1) Graphs and Tensor representations of data; R2) learning from data with spatio- and temporal aspects, and R3) Data Guarantees: quality, transparency, privacy, fairness, FAIRness, and trust. In research, we spotlight some progress in area R1 on inference for biological networks. In education, we describe our project for designing a collection of

case studies for teaching ethics of data science to undergraduate students and invite collaborators. In the area of broadening participation, we warmly invite undergraduates from all group, including those traditionally under-represented in STEM fields, to apply to our summer DIAMONDS REU experience.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 46

The Earth Data Science Corps: a model for teaching & learning environmental data science skills

Authors: Nathan Quarderer[1]; Jennifer Balch[1]

[1] *CU Boulder/CIRES/Earth Lab/ESIIL*

Corresponding Author: nathan.quarderer@colorado.edu

The Earth & Environmental Sciences (EES) produce vast amounts of data at a pace and on a scale that precipitate a need for EES researchers who are equipped with the technical data analytic skills required to work with large EES data sets. There are currently limited opportunities to learn these critical earth and environmental data science (EDS) skills leading to a gap between the demand for and supply of well trained data analysts, and contributes to a lack of diversity in the workforce. One model for meeting these demands is the NSF-supported Harnessing the Data Revolution (HDR) Earth Data Science Corps (EDSC) which has engaged with 60 students and 8 faculty partners from Minority Serving Institutions (MSIs) and Tribal Colleges and Universities (TCUs) in the 3 years of the program. Through online instruction and a 12-week paid internship that includes training in fundamental Python programming and geospatial science, we have demonstrated significant growth across different aspects of participants' technical Python and data science skills, as well as their science identity and sense of belonging to a larger population of data scientists. These findings will be discussed along with implications for teaching EDS to members from historically underrepresented communities.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 47

Central Coast Data Science Partnership: Training a New Generation of Data Scientists

Author: Alex Franks[None]

Corresponding Author: amfranks@ucsb.edu

Our collaborative program establishes pathways for data science training through coursework and real-world projects, connecting three main public higher education institutions in California. Students learn the underlying principles of data science, including data-generating processes and the

role of measurement, ethics and privacy, information-processing tools for harnessing the power of big data, and the oral and written communication skills necessary for pursuing effective professional careers in the field. Training culminates in a year-long capstone course synthesizing and applying previously learned tools and techniques in a large-scale project sponsored by a company or academic lab. To date, at the three institutions our project has supported 80 distinguished Data Science Fellows, including 12 UR and 18 Hispanic students. The project has also let to the development of several new data science courses at each of the institutions. These new data science courses at UCSB have served more than 1,000 students in their first two years.

Research:

Education and Outreach:

Yes

Data & Cyberinfrastructure:

Poster session I / **48**

Training next generation data scientists for Energy Transition

Author: Mikyoung Jun[None]

Corresponding Author: mjun@uh.edu

This project (started in October 2021) plans to train next generation workforce in data science for energy industry, ranging from traditional oil and gas energy to renewable energy and energy transition. We are a team of five universities in the greater Houston region: University of Houston (UH) as the leading institution with UH-downtown, UH-Victoria, UH-clear lake, and Sam Houston State University. Each year, the program trains a cohort of 40 students with diverse background in undergraduate or Master program from the five participating universities and beyond, through a year long program. The program consists of a 5 weeks summer camp (on various topics on statistics, data science fundamental, geophysics, energy policy, computer science and engineering) and a semester long research team projects on data provided by industry partners. At the end of the program, students will do summer internship at various industry and also some of them will work as junior scholar for our program to help with the next cohort. We just finished a summer camp for cohort 1 successfully in summer 2022. This project has created opportunities for collaborations across participating universities in teaching as well as research.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **49**

Infusion of data science and computation into engineering curricula

Authors: Rebecca Napolitano[1]; Wesley Reinhart[None]

Co-authors: Ryan Solnosky ; Allen Kimel

[1] *Penn State University*

Corresponding Author: rjn5308@psu.edu

The goal of this project is to develop a curricular framework for data science education and workforce development that is transferable between diverse institutions, so STEM-related programs can "plug and play"data science lessons with existing curricula without much overhead. These lessons will be created in conjunction with community stakeholders and industry partners to ensure a focus on real-world problem solving and include student organizations in course development to promote flexible learning pathways. The proposed additions to undergraduate STEM education will provide an evidence-based blueprint for best practices in integrating data science with existing engineering curricula. Implementation across multiple engineering departments will result in a significant impact on society through the training of a diverse, globally competitive STEM workforce with high data literacy. The objectives of this project are to (1) facilitate data science education and workforce development for engineering and related topics, (2) provide opportunities for students to participate in practical experiences where they can learn new skills in a variety of environments, and (3) expand the data science talent pool by enabling the participation of undergraduate students with diverse backgrounds, experiences, skills, and technical maturity in the Data Science Corps. This work will support the Data Science Corps objective of building capacity for education and workforce development to harness the data revolution at local, state, and national levels. The institutions gathered for this project will develop training programs and curate datasets that will be made available so they can be included in undergraduate instruction nationwide. Furthermore, the training materials will be shared with industry partners, facilitating workforce development. The project team will develop a website to house data science training programs, didactic datasets, and other resources for educators. These resources are intended to reduce barrier to entry for faculty seeking to incorporate data science into their instruction, as recruiting and retaining faculty to create and teach integrated introductory courses in data science has been recognized as a significant hurdle by the National Academies.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 50

DAMAD Science CorpsThe Delaware And MiD-Atlantic Data Science Corps

Authors: Federica Bianco[1]; Jing Gao[1]; Gregory Dobler[1]; Claude Tameze[2]; Hacene Boukari[3]

[1] *University of Delaware*

[2] *Lincoln University*

[3] *Delaware State University*

Corresponding Author: jinggao@udel.edu

The Delaware And MiD-Atlantic Data Science Corps (PI Bianco) is an NSF HRD-sponsored, regional partnership between the University of Delaware (UD), Lincoln University (LU), and Delaware State University (DSU) aimed at creating an equitable, accessible program for undergraduate data science education that: (1) is accessible to students of any background with a focus on STEM preparation level; (2) supports students' education and career goals across disciplines inside and outside of STEM; (3) employs equity-focused educational practices that support all students'identities, (4) provides job-ready skills including data-ethics training; (5) builds capacity for data science training at LU and DSU, two Historically Black Colleges and Universities (HBCUs) Lu with emerging interests in Data Science education by leveraging the established pedagogical experience of the UD in Data Science; and (6) improves equity-focused educational practices at UD, a Primarily White Institution (PWI), to improve our recruitment and retention of students from groups traditionally marginalized from STEM and higher education. The program has/will create new classes, events (hackathons), and research

opportunities for students at the three partnering institutions. It will generate a joint certificate program for students at each partnering institution. It will produce templates for boot camps, introductory Data Science classes, and advanced topical classes in time series analysis machine learning methods, geospatial data science, and AI-based image analysis. We will also document our successes and failures and produce a roadmap and "best practices" for other institutions interested in integrated educational partnerships across Research1, PWI, and minority-serving institutions.

We will review our plan, current activities, and the challenges we are facing in developing this unusual, but potentially transformative program.

Research:

Research activities will be conducted by students across domains

Education and Outreach:

Setting up an equity-focused data science educational program as a collaboration between an R1-Primarily White Institution, and two Historically Black Colleges and University with emerging Data Science interests

Data & Cyberinfrastructure:

nothing to report

Poster session I / **51**

Interdisciplinary Traineeship for Socially Responsible and Engaged Data Scientists (iTREDS)

Authors: Nitesh Chawla[None]; Thomas Mustillo[1]; Kristin Kuter[2]; Ron Metoyer[1]; Sugana Chawla[1]; Don Howard[1]; Ann-Marie Conrado[1]; Christopher Wedrychowicz[2]

[1] *University of Notre Dame*

[2] *Saint Mary's College*

The iTREDS program trains undergraduate students in data science through a lens of social responsibility and community engagement, including rigor and responsibility, ethics, society, and policy. The students also develop superskills in the areas of teamwork, working with stakeholders,ethics,communication, and entrepreneurship. The goal of this 15-credit program is to develop scholars with an in-depth data science background as well as communication, critical thinking, teamwork, and other skills necessary for professional development. Launched in 2020, the iTREDS program has successfully recruited and trained three cohorts of 72 students (67% female and 19% URMs) from University of Notre Dame and St. Mary's College. The students experience an enriched experiential learning framework involving coursework, summer internships, and capstone projects. Under faculty mentorship, the students develop their capstone project topics by thinking and working together with community partners and organizations on data-driven problems.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **52**

Cost-aware Generalized alpha-investing for Multiple Hypothesis Testing

Authors: Thomas Cook[None]; Harsh Dubey[None]; Ji Ah Lee[None]; Guangyu Zhu[None]; Tingting Zhao[None]; Patrick Flaherty[1]

[1] *University of Massachusetts Amherst*

Corresponding Author: pflaherty@umass.edu

We consider the problem of sequential multiple hypothesis testing with nontrivial data collection cost. This problem appears, for example, when conducting biological experiments to identify differentially expressed genes in a disease process. This work builds on the generalized α-investing framework that enables control of the false discovery rate in a sequential testing setting. We make a theoretical analysis of the long term asymptotic behavior of α-wealth which motivates a consideration of sample size in the α-investing decision rule. Using the game theoretic principle of indifference, we construct a decision rule that optimizes the expected return (ERO) of α-wealth and provides an optimal sample size for the test. We show empirical results that a cost-aware ERO decision rule correctly rejects more false null hypotheses than other methods. We extend cost-aware ERO investing to finite-horizon testing which enables the decision rule to hedge against the risk of unproductive tests. Finally, empirical tests on a real data set from a biological experiment show that cost-aware ERO produces actionable decisions as to which tests to conduct and if so at what sample size.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 53

HDR DSC: Collaborative Research: Transforming Data Science Education through a Portable and Sustainable Anthropocentric Data Analytics for Community Enrichment (ADACE) Program

Author: Yu Liang[1]

Co-authors: Dalei Wu ; Hemant Jain [1]; Jiang Li [2]; Lani Gao [1]; Lyn Potter [3]; Noman Saied [3]; Shaolei Teng [2]

[1] *University of Tennessee at Chattanooga*

[2] *Howard University*

[3] *Chattanooga State Community College*

Corresponding Author: yu-liang@utc.edu

Led by an interdisciplinary team from the University of Tennessee at Chattanooga, Howard University, and Chattanooga State Community College, the proposed Anthropocentric Data Analytics for Community Enrichment (ADACE) program will develop a sustainable education and research platform for human-centric data science, where humans are either considered as the research subjects or regarded as a component of data analytics.

The ADACE program includes the following activities: (1) advancing or innovating the data science curricula of the participating institutions; (2) recruiting undergraduate students in the program; (3) enriching the local communities by organizing a number of open seminars, workshops, or hackathons; and (4) developing cooperative community projects with the local business and research institutions. These community projects focus on topics pertaining to anthropocentric data

science, such as seamless human-machine interaction, interpretable neural networks, human-in-the-loop machine learning, social networks, and AI ethics, etc.

Participating institutions will take advantage of their connections to for-profit and nonprofit businesses in the area to expose students to practical data science issues and offer networking opportunities. Through this proposed program, students will acquire systematic data science knowledge, problem-solving abilities, practical experience, and rigorous research training. All curricular materials will be designed to be portable, sustainable, and easily disseminated to ensure their expanded impact. They are being evaluated for measurable outcomes and tailored to include non-traditional students, who form a large portion of the potential data science workforce in the regions surrounding the participating institutions.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 54

Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE)

Author: Shaowen Wang[1]

Co-authors: Deanna Hence [1]; Mohan Ramamurthy [2]; Carol Song [3]; David Tarboton [4]; Anand Padmanabhan [1]; Jiawei Han [1]; Bo Li [1]; Jianguo Liu [5]; Eric Shook [6]; Diana Sinton [7]

[1] *University of Illinois Urbana-Champaign*

[2] *University Corporation for Atmospheric Research*

[3] *Purdue University*

[4] *Utah State University*

[5] *Michigan State University*

[6] *University of Minnesota*

[7] *University Consortium for Geographic Information Science*

Corresponding Author: shaowen@illinois.edu

In today's interconnected world, disasters such as floods and droughts are rarely isolated events, and their cascading effects are often felt far beyond their locations of origin. The Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) creates an open platform for harnessing geospatial data to better understand interconnected interactions across diverse socioeconomic-environmental systems for enhancing community resilience and environmental sustainability. I-GUIDE nurtures a diverse and inclusive geospatial discovery community across many disciplines by bridging disciplinary digital data divides with broader impacts amplified through a well-trained and diverse workforce and proactive engagement of minority and underrepresented groups. I-GUIDE transforms geospatial data-intensive sciences through integration of artificial intelligence and cyberGIS (cyber-based geospatial information science and systems), reproducible data-intensive analytics and modeling, FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, and innovative education and workforce development programs. This transformation catalyzes new convergence science necessary to drive advances across many fields ranging from computer, data and information sciences to atmospheric sciences, ecology, economics, environmental science and engineering, human-environment and geographical sciences, hydrology and water sciences, industrial engineering, sociology, and statistics. Through synergistic advances of these fields, I-GUIDE is empowering diverse communities to produce data-intensive solutions to society's resilience and sustainability challenges such as biodiversity loss, and food and water insecurity.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **55**

The I-GUIDE Cyberinfrastructure Platform

Author: X. Carol Song[1]

Co-authors: Anand Padmanabhan ; Lan Zhao [1]; Rajesh Kalyanam [1]; Shaowen Wang ; Zhiyu Li [2]

[1] *Purdue University*

[2] *University of Illinois*

Corresponding Author: cxsong@purdue.edu

The I-GUIDE platform is designed to harness the vast, diverse, and distributed geospatial data at different spatial and temporal scales and make such data broadly accessible and usable to convergence research and education enabled by cutting-edge cyberGIS and cyberinfrastructure. The platform comprises composable and interoperable tools and cyberinfrastructure capabilities integrated through application programming interfaces and information exchange standards. An I-GUIDE science gateway has been created as the primary user environment of the I-GUIDE platform to find, explore, and share data and models with friendly access to software, end-to-end research workflows, hosted services, computational resources, and learning materials. The platform leverages several existing capabilities including services for simplifying access to high-performance computing (HPC) resources, the US national HPC infrastructure, reusable geospatial workflow building blocks, and scalable, interactive computing environments to provide users with a web-based platform to carry out research and education workflows. Notebooks for pilot use case workflows have demonstrated their seamless access to sophisticated geospatial data methods and tools and state-of-the-art computational resources, with significantly improved usability and reproducibility. The I-GUIDE platform is bridging the "missing middle"to enable and accelerate the institute's convergent research and education agendas as well as being made available to the broader community.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

I-GUIDE Cyberinfrastructure Platform

Poster session I / **56**

Convergence Curriculum for Geospatial Data Science

Author: Eric Shook[1]

Co-authors: Anand Padmanabhan [2]; Bo Li [3]; Diana Sinton [4]; Giri Narasimhan [5]; Jayakrishnan Ajayakumar [6]; Mark Daniel Ward [7]; Mohan Ramamurthy [8]; Peter Darch [2]; Shaowen Wang ; Upmanu Lall [9]; Venkatesh Merwade [7]; Vetria Byrd [7]

[1] *University of Minnesota*

[2] *University of Illinois at Urbana Champaign*

[3] *University of Illinois Urbana-Champaign*

[4] *University Consortium for Geographic Information Science*

[5] *Florida International University*

[6] *Case Western Reserve University*

[7] *Purdue University*

[8] *UCAR*

[9] *Columbia Univ.*

Corresponding Author: apadmana@illinois.edu

The Convergence Curriculum for Geospatial Data Science is an integrative curriculum to prepare students, scholars, and professionals to build the necessary knowledge, skills, and competencies to solve convergent problems without having to go through a series of multi-week regular courses. This multi-tiered curriculum starts with 5 Foundational Knowledge Threads to establish a common basis for individuals coming from diverse backgrounds. Individual learners begin to integrate skills, knowledge, methods, and technologies as they move up through Knowledge Connections and Knowledge Frames. The pinnacle of the curriculum is Knowledge Convergence, which combines previous competencies with existing domain knowledge. Each component in the curriculum can be tailored to individuals at varying depths: 3 sentences, 3 slides, a 3-hour module, or a 3-week unit. This configuration allows learners to adapt their learning experience to match their own learning pathway. In this poster, we will share example curriculum materials that combine new materials with existing Open Education Resources (OERs) and the first draft of the Convergence Curriculum for Geospatial Data Science.

Research:

Education and Outreach:

Convergence Curriculum for Geospatial Data Science

Data & Cyberinfrastructure:

Poster session I / 57

iHARP- Harnessing Data and Model Revolution in the Polar Regions

Author: Vandana Janeja[1]

Co-authors: Aneesh Subramanian ; Jianwu Wang [1]; Mathieu Morlighem [2]; Shashi Shekhar [3]

[1] *UMBC*

[2] *Dartmouth*

[3] *UMN*

Corresponding Author: vjaneja@umbc.edu

The melting of the polar ice sheets contributes considerably to ongoing sea-level rise and changing ocean circulation, leading to coastal flooding and impacting tens of millions of people globally. However, we are yet unable to accurately predict how quickly the ice sheets will continue to shrink contributing to the sea level rise. In particular, we are still challenged by a limited understanding of transdisciplinary processes that determine ice sheet change, such as the role of subglacial topography and ice-atmosphere-ocean interactions.

iHARP research aims to (1) integrate heterogeneous, noisy, and discontinuous data in space and time, (2) integrate data with numerical and physical models via physics-informed machine learning and causal Artificial intelligence (AI), (3) develop spatial-temporal algorithms to forecast the

changes in the Arctic and Antarctic, and (4) build scalable algorithms to apply our solutions at a global scale.

iHARP serves as a research hub where experts in data science, Arctic and Antarctic science, and cyberinfrastructure in academia, government, and private sectors come together to develop transformative and integrative data science solutions to reduce uncertainties in projecting future sea-level rise and climate change.

iHARP champions multiple clusters of research-integrated educational initiatives across diverse backgrounds of students, with a specific focus on facilitating cross-disciplinary collaborations, training next-generation multi-disciplinary researchers and engaging the public in scientific inquiry as related to climate change and data science. In partnership with related diverse communities, i-HARP designs curricula, and offers hands-on community workshops, lecture series, conference tutorials, and training.

Research:

https://iharp.umbc.edu/

Education and Outreach:

https://iharp.umbc.edu/news-and-events/

Data & Cyberinfrastructure:

Poster session I / **58**

Institute for Data-Driven Dynamical Desgin

Author: Eric Toberer[1]

Co-authors: Jane Greenberg [2]; Steven Lopez [3]

[1] *Colorado School of Mines*

[2] *Drexel U.*

[3] *Northeastern*

Corresponding Authors: janeg@drexel.edu, s.lopez@northeastern.edu, etoberer@gmail.com

The NSF Institute for Data-Driven Dynamical Design (ID4) aims to transform how scientists and engineers harness data when designing materials and structures. From chemistry to civil engineering, we seek to create platforms that accelerate the discovery of new mechanisms and dynamics through the complementary union of human and machine intelligence. Cross-cutting these efforts are efforts to understand, predict, and control transition state pathways and collective dynamics. Throughout these activities we are committed to training the next generation and engaging with the broader data-driven community.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **59**

HDR IDEAL - The Institute for Data, Econometrics, Algorithms, and Learning

Authors: Aravindan Vijayaraghavan[None]; Jason Hartline[None]

Co-author: Varun Gupta

Corresponding Author: guptav@uchicago.edu

This IDEAL (Phase I) project involves the development of a multi-discipline and multi-institution collaborative institute in the Chicago area that focuses on key aspects of the theoretical foundations of data science. The institute leverages existing strengths across computer science, statistics, economics, electrical engineering and operations research across Northwestern University, Toyota Technological Institute at Chicago (TTIC) and University of Chicago to bear upon foundational problems related to machine learning, high-dimensional data analysis and optimization in both strategic and non-strategic environments.

The research thrusts center around three broad themes:

1. High dimensional data analysis: This theme addresses both algorithmic and statistical challenges in dealing with high dimensional data, and investigate topics like dimension reduction, metric embeddings, sketching, inference on networks and problems in unsupervised learning like clustering and probabilistic modeling.
2. Data Science in Strategic Environments: This addresses computational and information theoretic challenges in econometric models of strategic behavior. Complexity arises, for example, from high-dimensional parameter spaces, unobserved heterogeneity, and multiplicity of equilibria in games. Specific topics of interest include inference on structural parameters, algorithms to characterize boundary of sets, partial identification, and machine learning in econometrics.
3. Machine learning and optimization: This theme addresses foundational questions in both continuous and discrete optimization and its use in machine learning; topics include representation learning, robustness in learning, and provable bounds for non-convex optimization and deep learning.

There have been 5 special quarters (fall and spring of each year) so far where the institute has brought together investigators, postdocs and Ph.D. students to focus on the topic of the special quarter. We are currently in the middle of our special quarter on Data Economics (Fall 2022). In addition to the interdisciplinary foundational research and education through coordinated graduate courses, there have been several hybrid and virtual workshops, seminars (which are all recorded and posted on our public webpage) and other events that have allowed us to engage with the broader community.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 61

TRIPODS Phase II: Institute for Data, Econometrics, Algorithms, and Learning (IDEAL)

Author: Lev Reyzin[1]

[1] *University of Illinois at Chicago*

Corresponding Author: lreyzin@uic.edu

Our institute is a multi-institution and transdisciplinary collaborative Phase II Institute for Data, Econometrics, Algorithms, and Learning (IDEAL), which focuses on key aspects of the foundations

of data science. IDEAL will consolidate and amplify research devoted to the foundations of data science across all the major research-focused educational institutions in the greater Chicago area: University of Illinois at Chicago (UIC), Northwestern University (NU), Toyota Technological Institute at Chicago (TTIC), University of Chicago (UC), and Illinois Institute of Technology (IIT). Our team involves 55 faculty working on the foundations of data science from all the core TRIPODS disciplines of computer science, electrical engineering, mathematics, statistics, and related fields like economics, operations research, optimization, and law. Additionally, the team includes a group of 9 Google researchers, who add to our technical strength and provide a direct connection to industry.

In Phase II, IDEAL's research goals center around three main thrusts –Foundations of Machine Learning, High-dimensional Data Analysis and Inference, and Data Science and Society. Specific topics include foundations of deep learning, reinforcement learning, ML and logic, network inference, high-dimensional data analysis, trustworthiness & reliability, fairness, and data science with strategic agents. The research activities are designed to facilitate collaboration between the different disciplines and across the five Chicago-area institutions, and they build on the extensive experience from our Phase I institutes. The activities include topical special programs, postdoctoral fellows, co-mentored PhD students, workshops, coordinated graduate courses, visiting fellows, research meetings, and brainstorming sessions.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 62

Lightening talk I

Poster session I / 63

DIMACS DATA-INSPIRE at Rutgers: Conditioned Weiner Processes and a Rigorous Probabilistic Analysis of Dynamics

Authors: Cameron Thieme[1]; Konstantin Mischaikow[2]

[1] *DIMACS, Rutgers University*

[2] *Rutgers University*

Corresponding Author: cameron.thieme@rutgers.edu

We study a Weiner process that is conditioned to pass through a finite set of points and consider the dynamics generated by iterating a sample path from this process. Using topological techniques we are able to characterize the global dynamics and deduce the existence, structure and approximate location of invariant sets. Most importantly, we compute the probability that this characterization is correct. This work is probabilistic in nature and intended to provide a theoretical foundation for the statistical analysis of dynamical systems which can only be queried via finite samples.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 64

DATA INSPIRE: Morse Graphs can effectively estimate the Regions of Attraction (RoAs) of dynamical systems, including closed-box ones.

Author: Ewerton Rocha Vieira[1]

Co-authors: Aravind Sivaramakrishnan [1]; Edgar Granados [1]; Konstantin Mischaikow [1]; Kostas Bekris [1]; Marcio Gameiro [1]; Yao Song [1]; Ying Hung [1]

[1] *Rutgers*

Corresponding Author: ewertonrvieira@gmail.com

This work proposes an integration of surrogate modeling and topology to significantly reduce the amount of data required to describe the underlying global dynamics of robot controllers, including closed-box ones. A Gaussian Process (GP), trained with randomized short trajectories over the state-space, acts as a surrogate model for the underlying dynamical system. Then, a combinatorial representation is built and used to describe the dynamics in the form of a directed acyclic graph, known as *Morse graph*. The Morse graph is able to describe the system's attractors and their corresponding regions of attraction (RoA). Furthermore, a pointwise confidence level of the global dynamics estimation over the entire state space is provided. In contrast to alternatives, the framework does not require estimation of Lyapunov functions, alleviating the need for high prediction accuracy of the GP. The framework is suitable for data-driven controllers that do not expose an analytical model as long as Lipschitz-continuity is satisfied. The method is compared against established analytical and recent machine learning alternatives for estimating \roa s, outperforming them in data efficiency without sacrificing accuracy. Link to code: \url{https://go.rutgers.edu/49hy35en}

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / 65

A3D3: Community Engagement, Education, Outreach

Authors: Mark Neubauer[1]; Mark Stephen Neubauer[2]

Co-authors: Philip Coleman Harris [3]; Shih-Chieh Hsu [4]

[1] *Univ. Illinois at Urbana Champaign (US)*

[2] *Univ. Illinois at Urbana-Champaign*

[3] *Massachusetts Inst. of Technology (US)*

[4] *University of Washington Seattle (US)*

Corresponding Authors: msn@uiuc.edu, msn@illinois.edu

A3D3 aims to be a nexus for exchanging new ideas, algorithms and tools between scientific domains, AI communities and industry partners for AI-Hardware co-design. In this presentation, we will show efforts based on strong foundation on the Fast Machine Learning (FastML) community efforts. Our on-going programs on Postbaccalaurate Fellowships, Training, Education, and strong connection with industry leaders in AI hardware, HPCs, and cloud computing via integration with the FastML community and university research partnerships.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Poster session I / **66**

A3D3: Accelerating AI Algorithms

Author: Philip Coleman Harris[1]

Co-authors: Mark Neubauer [2]; Mark Stephen Neubauer [3]; Shih-Chieh Hsu [4]

[1] *Massachusetts Inst. of Technology (US)*

[2] *Univ. Illinois at Urbana Champaign (US)*

[3] *Univ. Illinois at Urbana-Champaign*

[4] *University of Washington Seattle (US)*

Corresponding Author: philip.coleman.harris@cern.ch

A3D3 Institute, Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery, aims to pursue next generation AI Algorithms combined with next generation processor technology to develop AI algorithms that can be run fast to solve real-time scientific problems with AI Domains: High Energy Physics, Multi-Messenger Astronomy, and Neuroscience. We will present Hardware-Algorithm co-design and collaborative approaches within different science domains to achieve optimal low latency and performance for science. We are also working to make Machine Learning challenges to highlight low latency domain and aiming to connect with MLPerf science and other organization with similar computational challenges.

Research:

Education and Outreach:

Data & Cyberinfrastructure:

Breakout session Summaries / **67**

Research: Inductive Bias in Learning

Corresponding Author: jsulam1@jhu.edu

Breakout session Summaries / **68**

Complex Data Structures

Data and Cyberinfrastructure Panel / **69**

Data and Cyberinfrastructure around the HDR ecosystem - the I-GUIDE perspective

Corresponding Author: cxsong@purdue.edu

The Core Cyberinfrastructure (CI) Capabilities and Services is one of the six focus areas in the I-GUIDE, an NSF HDR Institute for Geospatial Understanding through an Integrative Discovery Environment. Its primary mission is to bridge a wide range of distributed, heterogenous and rapidly increasing geospatial datasets with convergence research to achieve a greater society resilience and sustainable development. The I-GUIDE CI team follows several best practices, including integration and interoperation to leverage significant prior investment and proven platforms, making data and CI useful and usable in lessening data wrangling, capturing workflows for reproducibility, and improving adoption by using familiar interfaces; identifying our audiences to deliver tools and content to meet the needs. A year into the formation of the institute, the team has released a seamless and scalable platform to support computational intensive and complex scientific workflows. The team is developing tools, methods, and educational curriculum and material around geospatial data and computation to be shared with the HDR community.

Data & Cyberinfrastructure:

Education and Outreach:

Research:

Poster session I / **70**

Imageomics: Images as the Source of Information about Life

Author: Tanya Berger-Wolf[1]

[1] *Ohio State University*

Corresponding Author: berger-wolf.1@osu.edu

Introducing the new NSF HDR DIRSE Institute Imageomics: A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning. The institute aims to establish a new field of science, imageomics: from images to biological traits using biology-structured machine learning.

Images are the most abundant, readily available source for documenting life on the planet. Ranging in resolution, scale, and subject, and coming from natural history collections, laboratory scans, field studies, camera traps, wildlife surveys, autonomous vehicles on the land, water, and in the air, tourists'cameras, citizen scientists'platforms, posts on social media, aerial surveys and high resolution satellites, there are millions of images of living organisms. But their power is yet to be harnessed for science and conservation. Even the traits of organisms cannot be readily extracted from images. The analysis of traits, the integrated products of genes and environment, is critical for biologists to predict effects of environmental change or genetic manipulation and to understand the significance of patterns in the four billion year evolutionary history of life.

Data science and machine learning can turn massive collections of images into high resolution information database about wildlife, enabling scientific inquiry, conservation, and policy decisions. I will share our vision of the new scientific field of imageomics.

Research:

Education and Outreach:

Data & Cyberinfrastructure: