October 1, 2020 - September 30,

2021

idies AnnualReview

Johns Hopkins University | Institute for Data Intensive Engineering and Science

MESSAGE FROM THE DIRECTOR

At IDIES, faculty and students work together to solve amazing data-intensive problems, from genes to galaxies, including new projects in materials science, and urban planning. Over the last few years, our members have successfully collaborated on many proposals related to Big Data, and we have hired new faculty members, all working on different aspects of data-driven discoveries. Together, we are successful in building a collection of unique, large data sets, giving JHU a strategic advantage in making new discoveries.

COVID has left a mark on our abilities to work together, but we quickly adjusted to Zoom meetings. IDIES had a significant role in helping the COVID related genomic projects, by standing up a GPU infrastructure in a matter of a few hours.

Our collaborations across the many schools of Johns Hopkins have intensified, now include the School of Advanced International Studies, School of Education, the Mathematical Institute for Data Science (MINDS), the multi-institutional Paradim project in materials science (HEMI), the 21st Centuries Cities Initiative (21CC), cancer immunotherapy projects through the Bloomberg Kimmel Institute, and various large-scale studies in genomics and in medicine. Our technology provides the analytics engine at the heart of PMAP. More and more classes are using our SciServer as an interactive tool to support homework assignments in Python.

Over the past two years, IDIES has looked beyond JHU to establish new partnerships with outside organizations. We have active collaborations with the Lieber Institute for Brain Development, and the Kennedy Krieger Institute. We are involved with various space science projects: collaborating with the Space Telescope Science Institute on the WFIRST Space Telescope Data Archive; providing the data system for a new European X-ray satellite, eRosita; and we are working with NASA's Goddard Space Flight Center on hosting High Energy Astrophysics. Most recently, we started a new project with the National Institute of Standards and Technology (NIST) to build an aggregator for data from Puerto Rico about the effects of hurricane Maria and lately also helping the National Additive Manufacturing Initiative. We are starting a new collaboration with Brookhaven National Laboratories.

We welcome the new members of our Executive Board: Paulette Clancy(WSE), Alex Baras (JHMI), Brian Caffo and Jeff Leek (BSPH). Postdocs and graduate students are working with IDIES faculty on AI-related projects, from materials science to astronomy and cancer biology. Machine learning, in particular Deep Learning, has revolutionized how industry handles Big Data. IDIES and MINDS have decided to work together towards these very important emerging goals, joining our expertise to lead to greater new discoveries. JHU is moving towards creating a major effort in Artificial Intelligence, the AI-X project. IDIES is working very closely with AI-X on its infrastructural aspects.

The NSF has funded a new multi-institutional effort in Turbulence. JHU is a co-lead in a new NSF funded project to create the prototype for the National Science Data Fabric. The Open Storage Network project is gaining nation-wide acceptance.

IDIES aims to accelerate, grow and become more relevant across the University by providing more intensive help in launching and sustaining data intensive projects in all disciplines. We seek new ideas and new directions but cannot do this alone: we need your help and initiative. Please send us your ideas, big or small, on how we can improve our engagement with your research community.

CONTENTS

On the cover: The cover image is a plot of salinity distribution at the ocean surface produced from high-resolution global ocean circulation model data. The data, along with visualization and analysis tools, are made available through the Poseidon project (https://poseidon.idies.jhu.edu)

AGENDA

START	DUR	SESSION	SPEAKER	PRESENTATION
Thursday, October 21				
1:00 PM	15	Opening Remarks	Alex Szalay	
1:15 PM	40	Keynote Address	Judith Mitrani-Reiser	Disaster Programs at the National Institute of Standards and Technology Strengthen National Resillience
1:55 PM	20	IDIES Seed Award	Nadia Zakamska	The Search for Elusive Progenitors of Type I Supernovae
2:15 PM	10	SciServer	Gerard Lemson	SciServer Update
2:25 PM	20	IDIES Seed Award	Jonathan Ling	Comprehensive Analysis of Public Sequencing Archives to Uncover Novel Mechanisms of Pathogenesis in Amyotrophic Lateral Sclerosis
2:45 PM	10	Break		
2:55 PM	40	Keynote Address	K. T. Ramesh	AI-X: Bringing Johns Hopkins Together Through AI
3:35 PM	10	Robbins Award Introduction	Paulette Clancy	
3:45 PM	30	Robbins Future Faculty Award	Sai Pooja Mahajan	Towards Deep Learning Models for Target-Specific Antibody Design
4:15 PM	20	Robbins PhD Award	Karthik Menon	Computational and Data-Driven Analysis of Aeroelastic Flutter
4:35 PM	20	Poster Madness		
4:55 PM	30	Poster Session		
Friday, October 22				
1:00 PM	15	Opening Remarks	Alex Baras and Paul Nagy	
1:15 PM	40	Keynote Address	Rebecca Lindsey	Data Science and Machine Learning for Materials Under Extreme Conditions
1:55 PM	20	IDIES Seed Award	Thomas Lippincott	Developing Datasets and Infrastructure to Facilitate Translating Humanistic Data and Hypotheses into Computational Inquiry
2:15 PM	10	ARCH	Jaime Combariza	ARCH Update
2:25 PM	10	Break		
2:35 PM	40	Keynote Address	Ryan Abernathey	Pangeo: A Model for Cloud Native Scientific Research
3:15 PM	30	Invited Talk	Janis Taube	AstroPath: Mapping Cancer as if it were the Universe
3:45 PM	5	IDIES Student Fellow Awards		
3:50 PM	10	IDIES Student Fellow	Jaxon Wu	Humainzing Our Data: Proposal on Integrating Social and Behavioral Determinants of Health into Population Health Analytics
4:00 PM	10	IDIES Student Fellow	Shengwei Zhang	Using Machine Learning to Predict Surgical Case Duration in Operating Room Scheduling Optimization
4:10 PM	10	Closing Remarks	Ani Thakar	

KEYNOTES

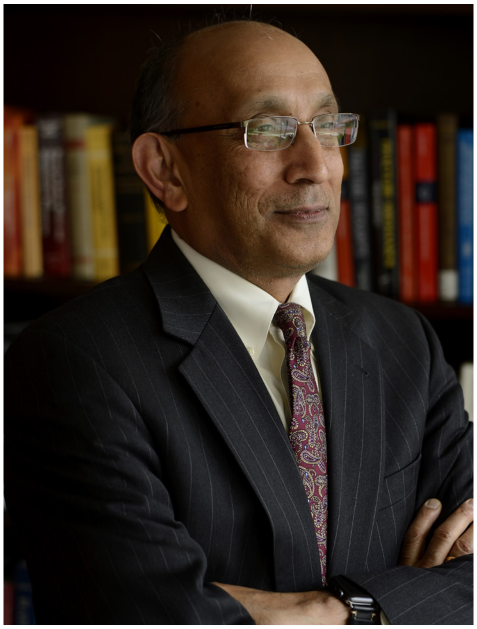
JUDITH MITRANI-REISER, PhD
Associate Chief, Materials and Structural Systems Division, National Institute of Standards and Technology (NIST)

Dr. Judith Mitrani-Reiser is at the lead of the NCST technical investigation of the collapse of the Champlain Towers South in Surfside, Florida and the leader of the mortality project of the NCST investigation of Hurricane Maria's impact on Puerto Rico. Judy's responsibilities at NIST also extend to managing and providing oversight to two other disaster statutory programs—the National Windstorm Impact Reduction Program and the National Earthquake Hazard Reduction Program—focused on interagency coordination to reduce losses in the U.S. from disasters and failures. Judy is Vice President of the Earthquake Engineering Research Institute (EERI), serves on the Executive Committee of the U.S. Collaborative Reporting for Safer Structures (CROSS-US), co-founded the American Society of Civil Engineers' (ASCE) Multi-Hazard Risk Mitigation Committee, and was elected to the Academy of Distinguished Alumni of UC Berkeley's Civil and Environmental Engineering Department. Judy earned her B.S. from the University of Florida, M.S. from the University of California at Berkeley, and Ph.D. from the Caltech.

Disaster Programs at National Institute of Standards and Technology Strengthen National Resilience

Extreme events, such as tornadoes and fires, test buildings and infrastructure in ways and on a scale that cannot be easily replicated in a laboratory. Therefore, actual disasters and failure events provide important opportunities for scientists and engineers at the National Institute of Standards and Technology (NIST) to study these events, and improve the safety of buildings, their occupants, and emergency responders. NIST has studied and investigated more than 50 earthquakes, hurricanes, building and construction failures, tornadoes, and fires since 1969 under several authorities. Current ongoing technical investigations include Hurricane Maria's impacts on Puerto Rico and the Champlain Towers South building collapse in Surfside, Florida. The talk will provide an overview of the disaster research conducted at NIST in the Materials and Structural Systems Division informed by recommendations and national strategic plans developed by national disaster statutory programs: Disaster and Failure Studies (DFS) Program, National Earthquake Hazard Reduction Program (NEHRP), and National Windstorm Impact Reduction Program (NWIRP).
The talk will highlight a collaboration with Hopkins' Institute for Data Intensive Engineering and Science (IDIES) scientists on the use of cyberinfrastructures that combine traditional data management and access services with computing resources. This collaboration leverages services provided by SciServer—such as storing/accessing large disaster data and server-side analysis using Jupyter notebooks—to make NIST disaster discoverable, inform critical disaster response activities, and enable necessary collaborations across stakeholders. The talk will also provide an overview of NIST Professional Experience Program (PREP) and how it enables the collaboration between NIST disaster scientists and IDIES data scientists.

KEYNOTES



K. T. RAMESH, PhD
Alonzo G. Decker, Jr., Professor of Science & Engineering, Senior Advisor to the President for AI, Johns Hopkins University

Dr. Ramesh is known for research in impact physics and the failure of materials under extreme conditions. Ramesh also is a professor in the Department of Mechanical Engineering, and holds joint appointments in the Department of Earth and Planetary Sciences and the Department of Materials Science and Engineering. He is the founding director of the Hopkins Extreme Materials Institute (HEMI), which addresses the ways in which people, structures and the planet interact with and respond to extreme environments. Ramesh's current research focuses on the design of materials for extreme conditions, the massive failure of rocks and ceramics, impact processes in planetary science, and impact biomechanics. In one project, his lab is developing a detailed digital model of the human brain to help address how brain injury results from head impacts. Other current projects include the use of laser shock experiments to study the deformation and failure of protection materials for the U.S. Army, the use of data science approaches in materials design, the development of a hypervelocity facility for defense and space applications, and modeling the disruption of asteroids that could hit the Earth. He has written over 250 archival journal publications, and is the author of the book "Nanomaterials: Mechanics and Mechanisms."

AI-X: Bringing Johns Hopkins Together through AI

We will discuss the broad outlines of a university-wide initiative in the AI space, considering processes, prospects, and priorities. The intent is to begin a conversation and encourage collaboration, while seeking input on opportunities and challenges.

KEYNOTES

REBECCA K. LINDSEY, PhD

Materials Science Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory (LLNL)

Dr. Lindsey is a Staff Scientist in the Materials Science Division at Lawrence Livermore National Laboratory. Her research focuses understanding how materials evolve under extreme conditions, and how aging impacts their performance. She leverages machine learning and data science to generate next-generation reactive interatomic models enabling quantum accurate simulation of phenomena including shockwave-driven nanocarbon formation, and to build diagnostic models for complex systems capable of predicting device performance from materials characterization data. Her efforts are underpinned by a strong interest in developing tools enabling work in previously inaccessible problem spaces. Dr. Lindsey's work, which has had implications for nanomaterial fabrication, civil engineering, defense applications, and possible origins for life, were recently recognized through a LLNL Physical and Life Sciences Directorate Research Award.

Data Science and Machine Learning for Materials Under Extreme Conditions

For the past several decades physics-driven advances in experimental and simulation capabilities have served as the primary force enabling improved understanding of material evolution under extreme conditions. The pace of advancement has largely been limited by the complicated, highly dynamic, and inherently multiscaled nature of this phenomenon. However, data driven approaches are providing a path forward. From high throughput experiments to improved physical and reduced order models, this new paradigm has had a transformative effect on research in the physical sciences. In this presentation, I will discuss two broad machine learning efforts that aim to improve our understanding of the microscopic phenomena governing material evolution under extreme conditions via atomistic simulations and enable prediction of age-related changes in material performance from experimentally derived characterization data.

KEYNOTES

RYAN P. ABERNATHY, PhD
Associate Professor of Earth and Environmental Sciences, Columbia University, Lamont Doherty Earth Observatory

Dr. Abernathey is a physical oceanographer who studies large-scale ocean circulation and its relationship with Earth's climate. He received his Ph.D. from MIT in 2012 and did a postdoc at Scripps Institution of Oceanography. He has received an Alfred P. Sloan Research Fellowship in Ocean Sciences, an NSF CAREER award, The Oceanography Society Early Career Award, and the AGU Falkenberg Award. He is a member of the NASA Surface Water and Ocean Topography (SWOT) science team and Director of Data and Computing for a new NSF Science and Technology Center called Learning the Earth with Artificial Intelligence and Physics (LEAP). Prof. Abernathey is an active participant in and advocate for open source software, open data, and reproducible science.

Pangeo: A Model for Cloud Native Scientific Research

As a result of advances in remote sensing and computer simulation, geoscientists are now regularly confronted with massive datasets (many TB to PB). While such datasets have great potential to move science forward, they require a new approach to data sharing and computing infrastructure. The Pangeo Project aims to empower geoscientists to work painlessly with such datasets using open source software and infrastructure. In this talk, I will describe the architectures and best practices that have emerged from this project which form a foundation for future "cloud-native" science. These include the use of object storage for building analysis-ready, cloud-native data repositories, data-proximate computing with Jupyter, and on-demand scale-out distributed computing with Dask. I will demonstrate these tools in action with real science workflows from oceanography and climate science. I'll also discuss some technical and social challenges our project is facing as we try to transition from promising prototypes to sustainable infrastructure for our field.

IDIES and NIST Researchers Develop Automated Predictions for Wind Hazard Response

Arik Mitschang, Associate Research Scientist, IDIES

Each year in the United States tropical cyclones cost billions of dollars and result in substantial loss of life. These disasters are expected to get worse in decades to come as a result of climate change. Improved building codes and their adoption, along with better preparedness could reduce these impacts. The National Institute of Standards and Technology (NIST) has a number of authorities under which disaster and failure studies are conducted, including: (1) the NIST Organic Act, (2) the National Construction Safety Team (NCST) Act, (3) the National Earthquake Hazard Reduction Program (NEHRP), and (4) National Windstorm Impact Reduction Program (NWIRP). The Disaster and Failure Studies (DFS) Program coordinates and manages all disaster studies and investigations at NIST, which may include on-site deployments to damaged areas.

A critical component in determining NIST's response activities to a windstorm or other hazard event is the NCST Preliminary Reconnaissance Decision Criteria, which scores domestic and international events based on four broad categories: physical damage, mortality, storm severity and response. This decision rubric produces a numeric score between 1 and 5 to gauge the overall impact of the event, with 5 being the most devastating. At present, score sheets are populated manually by a NIST researcher after-the-fact, using a variety of inputs including non-standardized sources (e.g. news reports, email exchanges with other agencies, community websites, etc.). The post-event timing, manual nature of scoring and use of non-standardized sources introduces uncertainties and delays in decision making. This process also precludes scoring all events impacting buildings and infrastructure, which limits improvements to the score metrics themselves due to lack of data. Standardizing and automating portions of the rubric with machine accessible data sources can help reduce the effort by subject matter experts and improve data sharing across federal agencies and other stakeholders. Additionally, in the case of tropical cyclones, if predictions for the scores could be generated in the pre-landfall phase of the storm, preparations for deployment activities could begin early to optimize time spent in the field collecting data and evidence. Increasing the number of events that are scored will facilitate and simplify comparative analyses, ultimately leading to improvements to the scoring methodology.

In collaboration with NIST's Materials and Structural Systems Division, carried out under the JHU/NIST PREP program, IDIES researchers have been developing an automated workflow to score tropical cyclones against the NCST decision criteria. The workflow models the computation of each score from standard authoritative input data sources - including those published by the National Hurricane Center (NHC), the US Census Bureau and the Centers for Disease Control (CDC), among others - as a directed acyclic graph.

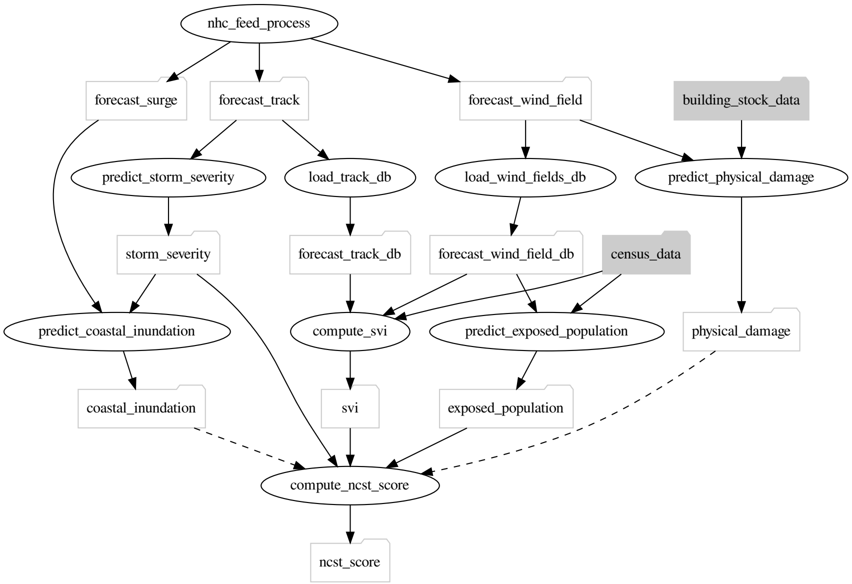
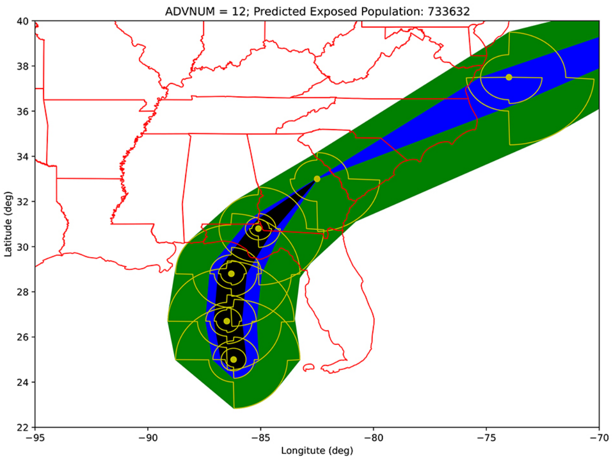
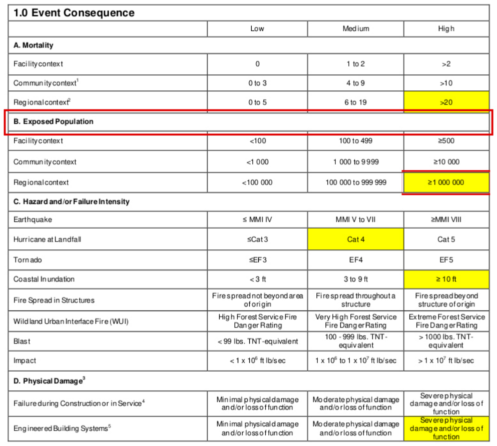


Fig. 1: The circular nodes represent notebooks that perform score predictions, other nodes are types of data sources. The arrows represent dependencies among nodes; those with dashed lines are optional. The shaded folder-shaped nodes represent externally pre-loaded archival data.

Each computational component is a Jupyter notebook that contains python codes for prediction and analysis, figures, and documentation all in a single document (see Figure 1). The components communicate with each other through output data products and synthesize the final score and report via a node which consumes these products. At the time of writing, all nodes have an implementation except for prediction of physical damage.

Fig. 2: The left hand side shows a portion of the completed score sheet for Hurricane Michael, which made landfall in Florida in 2018. Highlighted is the section related to exposed population. The right hand side shows a screenshot of the notebook used for automated prediction of exposed population for a particular NHC advisory, showing the predicted windfields of the storm several days out.





1.0 Event Consequence			
	Low	Medium	High
A. Mortality			
Facility context	0	1 to 2	>2
Community context[1]	0 to 3	4 to 9	>10
Regional context[2]	0 to 5	6 to 19	>20
B. Exposed Population			
Facility context	<100	100 to 499	≥500
Community context	<1 000	1 000 to 9999	≥10 000
Regional context	<100 000	100 000 to 999 999	≥1 000 000
C. Hazard and/or Failure Intensity			
Earthquake	≤ MMI IV	MMI V to VII	≥MMI VIII
Hurricane at Landfall	≤Cat 3	Cat 4	Cat 5
Tornado	≤EF3	EF4	EF5
Coastal Inundation	< 3 ft	3 to 9 ft	≥ 10 ft
Fire Spread in Structures	Fire spread not beyond area of origin	Fire spread throughout a structure	Fire spread beyond structure of origin
Wildland Urban Interface Fire (WUI)	High Forest Service Fire Danger Rating	Very High Forest Service Fire Danger Rating	Extreme Forest Service Fire Danger Rating
Blast	< 99 lbs. TNT-equivalent	100 - 999 lbs. TNT-equivalent	> 1000 lbs. TNT-equivalent
Impact	< 1 x 10^6 ft lb/sec	1 x 10^6 to 1 x 10^7 ft lb/sec	> 1 x 10^7 ft lb/sec
D. Physical Damage[3]			
Failure during Construction or in Service[4]	Minimal physical damage and/or loss of function	Moderate physical damage and/or loss of function	Severe physical damage and/or loss of function
Engineered Building Systems[5]	Minimal physical damage and/or loss of function	Moderate physical damage and/or loss of function	Severe physical damage and/or loss of function

As an example, Figure 2 shows a section of the notebook for the “predict exposed population” node (right) alongside the manually completed score sheet for the 2018 Hurricane Michael (left). The new automated approach uses predicted wind swath data from NHC at three distinct max wind speeds interpolated and overlaid with US Census data to estimate the population that may be affected by hazards related to the strongest winds. A historical analysis of past scored storms shows that in several cases the automated predictions matched well with the post-landfall score, whereas in other cases factors such as numeric boundary sensitivity combined with an integer score rubric led to divergent results (see Figure 3). In other cases, factors such as rainfall induced flooding were relevant in the manual score but are lacking in the automated workflow. There are already indications that the score can be improved rapidly based on a deeper analysis of the historical data. In addition, the ability to easily add relevant nodes to the workflow graph and evolve the score computation enables this critical tool to be implemented for tropical cyclone response in seasons to come and adapted for other event types in the future at NIST.

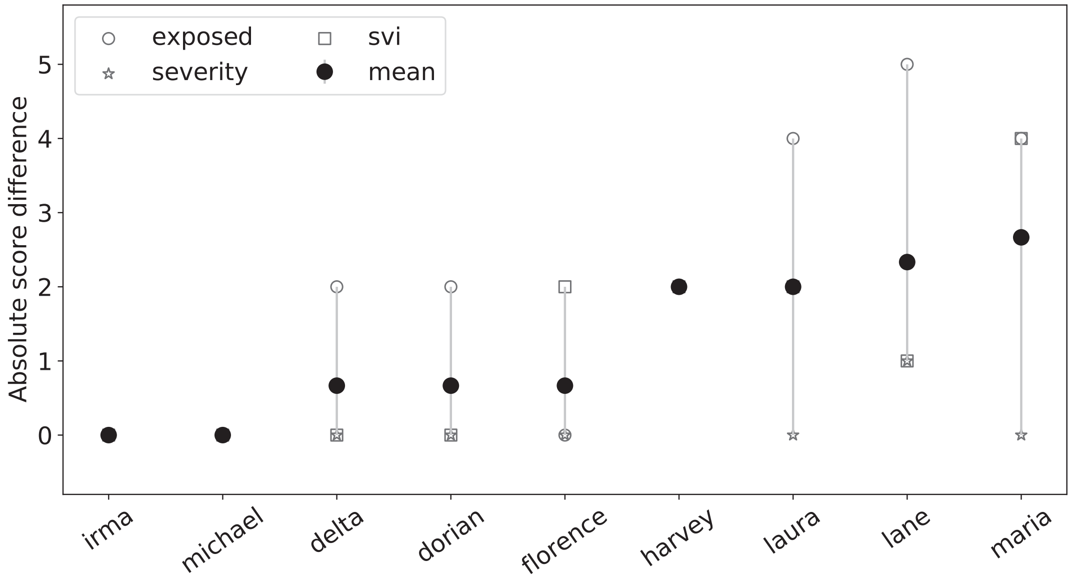


Fig. 3: For each storm we compare the automated score across advisories for each category with that of the manual score sheet completed post-landfall. For each category we plot the minimum difference between automated and manual scores and the mean of those across categories. The error bars illustrate the range of score differences. The data are sorted along the horizontal axis by mean score difference.

Scaling Data Science Education from Baltimore to the World with the Johns Hopkins Data Science Lab

Carrie Wright, Assistant Scientist, Bloomberg School of Public Health, JHU
Ava M. Hoffman, Research Associate, Bloomberg School of Public Health, JHU
Ashley K. G. Johnson, Program Administrator for DataTrail, JHU
Frederick J. Tan, Bioinformatics Research Faculty, Embryology, Carnegie Institution
Jeffrey T. Leek, Professor, Biostatistics, Director of DaSL, JHU

There has been an unprecedented explosion in data generation across all fields of science over the past decade. Entire fields have had to rapidly evolve to use new and exciting data and computing platforms. Yet, practitioners in these fields struggle to keep up with this dramatic evolution as training in data science and computing literacy remains a bottleneck for further advancement. More common frameworks for computing need to be developed and implemented to make collaboration across groups and domains more feasible. Furthermore, democratizing data science literacy provides an avenue for ultimately creating more equitable employment opportunities and health care for people of all backgrounds.

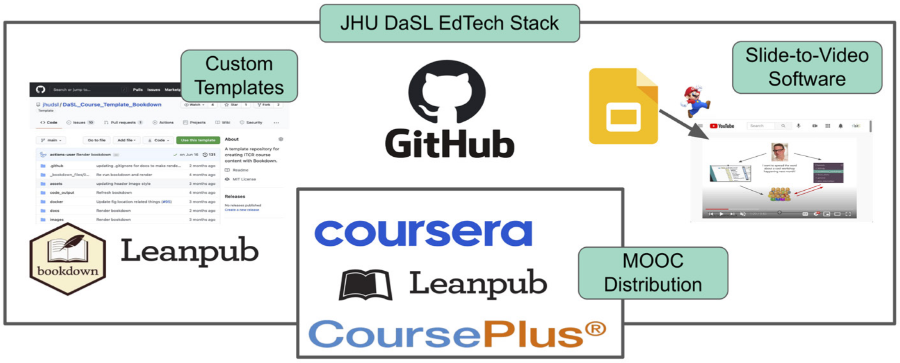


Figure 1. EdTech stack leveraged by JHU Data Science Lab for creating outreach content
The Johns Hopkins Data science Lab approach to data science education content creation involves creating templates using infrastructure like GitHub and GitHub Actions to allow for simple conversions of content into multiple formats that are supported by various publishing platforms for distribution as Massive Open Online Courses (MOOC). The lab has also created software to make it easy to convert content in Google Slides into videos with little effort.

The Johns Hopkins Data Science Lab (DasL - pronounced “dazzle”, jhudatascience.org) has the mission of training the world in data science practices - having taught more than 8 million people across online and offline platforms. The DaSL creates education content, technology, and infrastructure to provide training for individuals to become and stay up-to-date with data science literacy (Figure 1). The DaSL philosophy is to make content accessible, scalable, maintainable, and collaborative. Content is written once and published on multiple platforms to give greater access to as many individuals as possible. Borrowing ideas from software development, educational content is maintained on GitHub, using GitHub actions and packages developed by the team to make content updates seamless and easy including modifying video content without rerecording. All DaSL content uses open licensing to enable easy sharing, collaboration, and reuse, so that others can find materials used to create content for free.

The DaSL is actively applying this educational technology in several projects (Figure 2) aimed at democratizing data science literacy.

Figure 2. Communities served by JHU Data Science Lab projects include the NCI ITCR Training Network (ITN), NHGRI AnVil, and DataTrail

The Johns Hopkins Data Science Lab works on a variety of projects that provide data science education and resources to multiple communities, including the National Cancer Institute (NCI) Informatics Technology for Cancer Research (ITCR) Training Network (ITN) serving the cancer research community (including historically black colleges and universities), the National Human Genome Research Institute (NHGRI) Analysis Visualization and Informatics Lab-space (AnVIL), serving genomics and bioinformatics communities (including community colleges and institutes serving people of color), and DataTrail, serving young-adults of color from underserved communities.

- The Informatics Technology for Cancer Research (ITCR) Training Network (ITN) (itcrtraining.org) aims to increase access to informatics training and resources for the cancer research community and beyond. This involves creating courses to make researchers more aware of informatics best practices, resources, and tools, creating infrastructure to make it easier for cancer informatics tool developers to publish their own courses, and providing live education events to further disseminate course content (especially among faculty at historically black colleges and universities) and to encourage community engagement in cancer research.
- The National Human Genome Research Institute (NHGRI) Analysis Visualization and Informatics Lab-space (AnVIL) (anvilproject.org) provides cost-effective and secure cloud-based computing resources, data access, and workflows for genomic research. This project promotes research using 3+ petabytes of data from 300+ thousand human participants. Educational material includes step-by-step guides for setting up accounts using a persona-based approach; launching analyses using Jupyter, RStudio, Galaxy, and WDL Workflows; a series of exercises showcasing genomics analysis on the Cloud; and a guide to using AnVIL in the classroom. To promote institutes serving people of color and community colleges to participate more in genomics and bioinformatics research, the DaSL has helped facilitate the creation of a network called the Genomic Data Science Community Network (GDSCN) (gdscn.org) to provide support and education about how to work with AnVIL resources.
- DataTrail (datatrail.org) is a fresh take on the concept of workforce training in which prospective talent and industry professionals engage in mutually intensive coursework. Data science scholars—young adults of color recruited by our community-based partners—participate in guided technical and soft skills training, which are prerequisites of joining the industry. IDEA (inclusion, diversity, equity, and antiracism) scholars—who currently serve as industry professionals—engage in guided antiracism and mentorship training required to serve as informed advocates and hiring managers. The social connections established during the program will fuel job opportunities for learners and foster a more equitable and inclusive climate at the experts' institutions.

The goal of the JHU DaSL is to make data science accessible to all scientists and the broader worldwide community while creating equitable opportunities in this exciting new field for people from all backgrounds.

Seed Awardee Updates

NADIA ZAKAMSKA, PhD

Associate Professor, Physics & Astronomy, JHU

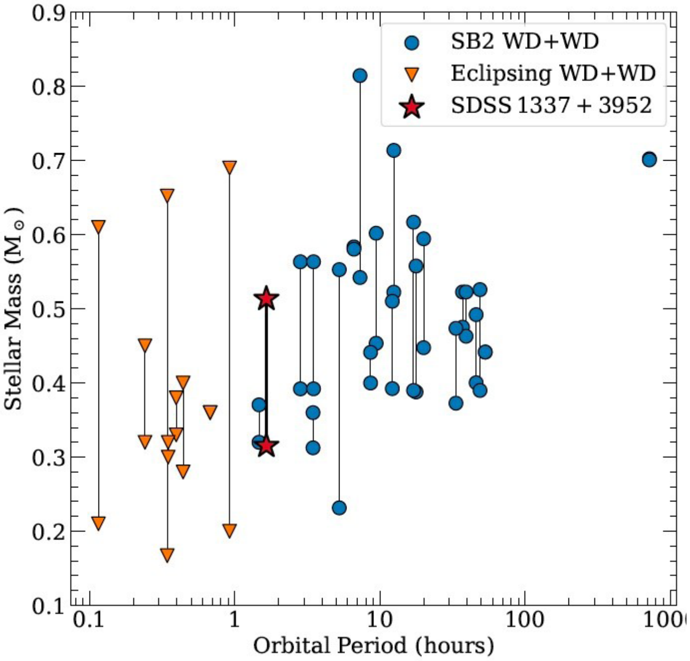
The Search for Elusive Progenitors of Type Ia Supernovae

Arguably one of the most enduring mysteries of modern astrophysics is that of the origin of type Ia supernovae, the cosmological standard candles that were used in the Nobel-prize-winning discovery of the accelerated expansion of the universe and are important in the evolution of chemical abundances of galaxies. The most likely scenario is that type Ia supernovae arise as a result of a merger of two white dwarfs -- compact remnants of evolution of stars like our Sun.

Although the supernovae themselves are routinely observed as bright astronomical transients out to great distances, to date no binary white dwarfs on track to become type Ia supernovae have been identified. This is due in part to the extreme difficulty of finding such objects. A convincing discovery would require high-quality, time-series spectroscopy and excellent photometry and distance measurements -- all for intrinsically faint stars -- in order to prove that the binary of white dwarfs exceeds the necessary critical mass for the type Ia explosion and that it would merge in less than the lifetime of the universe.

The field is now on the verge of a major break-through. The 20-year-old Sloan Digital Sky Survey (SDSS), in which JHU has long been a major partner, has just entered its fifth phase which will last about five years and will acquire high-quality spectra of 4-5 million stars across the Milky Way. In particular, the survey will obtain spectra of 200,000-300,000 white dwarfs. Furthermore, European satellite Gaia, operating since 2013, has allowed unprecedented measurements of distances to millions of astronomical sources. The combination of data now emerging from SDSS-V and Gaia may enable the long-sought discovery of type Ia progenitors, but no existing analysis tools can quickly identify the subtle features of white dwarf binaries in this large volume of data.

Our group is exploring a variety of methods to identify white dwarf binaries over a wide range of masses and periods. One of the most promising emerging avenues is to detect rapid radial velocity variations. SDSS spectra are obtained as a series of 15-minute exposures which



Our newly discovered binary white dwarf (Chandra et al.; red stars) is one of only a couple dozen known and is the first of many that will be revealed by the SDSS-V survey. A census of such objects and especially their mass measurements are essential for solving the long-standing puzzle of progenitors of type Ia supernovae.

could be on the same night or on different nights. If a white dwarf is in a binary system with a period of a few hours or less, its orbital velocity may noticeably change from one 15-minute exposure to another. The variations can be very subtle and especially difficult to identify because of the highly pressure-broadened absorption lines in white dwarf spectra.

We are developing a variety of modern statistical techniques to identify these subtle clues and therefore the most promising candidate white dwarf binaries. This program is already bearing fruit, and our paper on the discovery of a 99-minute white dwarf binary (Chandra et al. 2021) recently became the first scientific result of SDSS-V. With the unprecedented amount of white dwarf data from SDSS-V -- an order of magnitude more than from the previous surveys -- we are aiming to conduct a complete census of binary white dwarfs in the Solar neighborhood and to elucidate the nature of type Ia progenitors.

References: Chandra et al. 2021, Astrophysical Journal, in press, https://ui.adsabs.harvard.edu/abs/2021arXiv210811968C/abstract

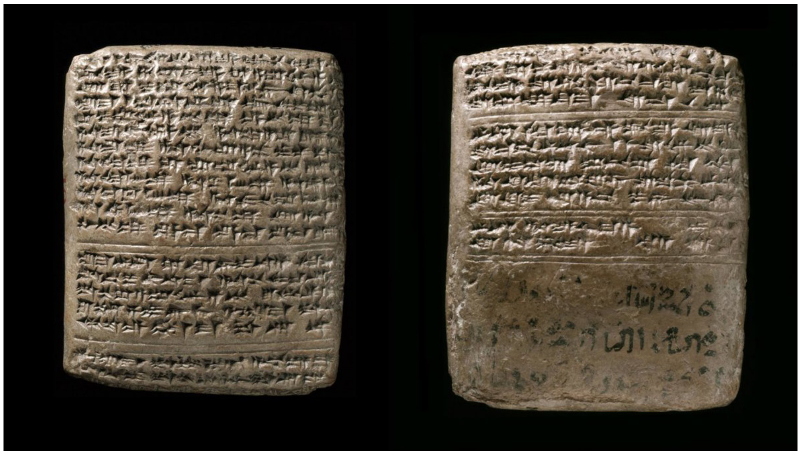
THOMAS LIPPINCOTT, PhD

Assistant Research Professor, Computer Science, JHU

Developing Datasets and Infrastructure to Facilitate Translating Humanistic Data and Hypotheses into Computational Inquiry

The initial phase of our research has focused on assembling suitable datasets for the specific subdomains of interest in English literature and the Ancient Near East, working with faculty and graduate students to refine the humanistic questions to be asked, and focusing our list of engineering goals to maximize impact for this and future collaborations.

With the Department of English, we have assembled a corpus of approximately 60,000 documents from the 16th through 19th centuries published in England, Scotland, and Ireland. Our goal is to consider how extracted linguistic patterns reflect the evolving English attitudes towards Ireland, and how this evolution aligns with major events (e.g. acts of Parliament, insurrections) and across time/space. We are starting with simple statistical tests targeting specific terminology and contexts annotated by graduate students.

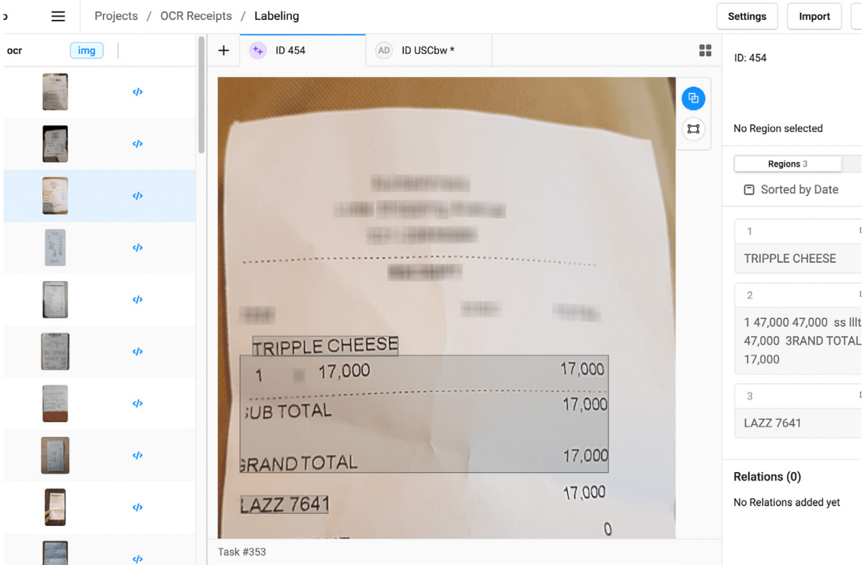


An example of the El Amarna letters showing correspondence between Egypt and the Mitanni, a people at the intersection of Near Eastern and Indo-Aryan Cultures who spoke Hurrian language, which belongs to neither.

CONTINUED FROM PRIOR PAGE (LIPPINCOTT)

With the Department of Near East Studies, we are testing, and potentially refining or expanding, a hypothesis regarding the scribal hands responsible the El Amarna Letters. This cache of tablets, written in the Cuneiform script, contains correspondence between the Egyptian Pharaoh Akhenaten and client kingdoms in the years preceding the Bronze Age Collapse. By encoding the existing hypothesis, linking it to images and transliterations, and gathering expert annotation from graduate students focused on this area, we set the stage for bringing techniques from natural language processing and computer vision to bear on questions typically answered via close manual scrutiny.

In response to the common needs of these applied studies, we are focusing engineering efforts on extending the Turkle annotation framework to perform annotation of images, combined temporal and geographic visualization of arbitrary features, and a flexible web editor to guide humanists in specifying and validating descriptions of their domains. These tasks are all against the backdrop of adopting JSON-LD as the canonical underlying format for linked data, and deploying a production-quality server under the JHU domain that will facilitate access and consolidate the public-facing aspects of our research.



1 TRIPPLE CHEESE

2 1 47,000 47,000 ss lllt 47,000 3RAND TOTAL 17,000

3 LAZZ 7641

The open-source Javascript libraries powering the LabelStudio annotation system will provide rich OCR and image-labeling capabilities to the JHU-developed Turkle framework.

NATALIA TRAYANOVA, PhD

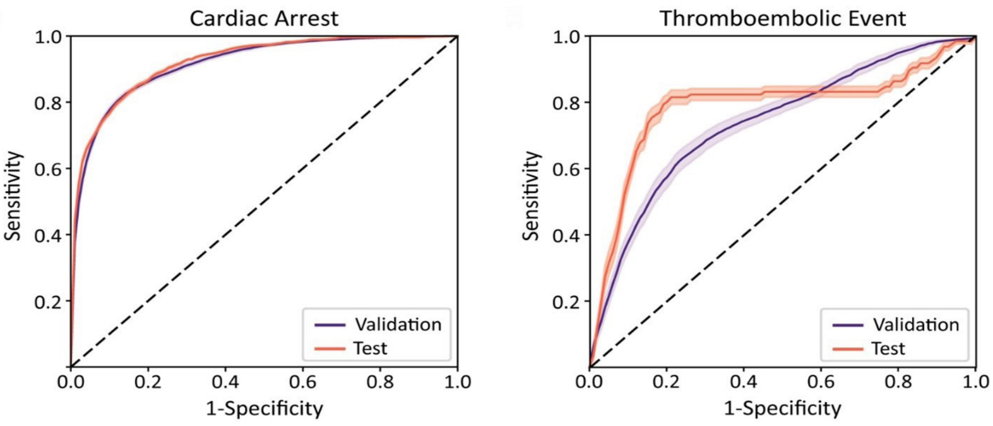
Professor, Biomedical Engineering & Medicine, JHU

Real-Time Prediction of Long-term Cardiovascular Complications in Hospitalized Patients with COVID-19

Patients with COVID-19, the disease caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), often present with cardiovascular (CV) manifestations such as myocardial infarction, thromboembolism, and heart failure. Clinically overt cardiac injury or cardiomyopathy is reported in 8 to 33% of hospitalized patients and is associated with up to 50% mortality, but imaging studies suggest the true incidence of cardiac involvement in all persons infected with SARS-CoV-2 could be as high as 60%. Thromboembolic events are also frequently reported in severe COVID-19 and are associated with mortality; one study found that 70.1% of non-survivors and 0.6% of survivors met criteria for disseminated intravenous coagulation. Furthermore, thromboembolic complications are more pronounced in acute COVID-19 infection than in other viral illnesses, and include pulmonary embolus and ischemic stroke, which can be fatal and are a significant cause of morbidity even as the infection resolves.

Despite the prevalence of no approach currently exists to forecast adverse CV events in COVID-19 patients in real time. In this study, we develop and validate the first prognostic ML model to forecast the real-time risk of CV complications in hospitalized patients with COVID-19. We term the model the COVID-HEART predictor. We focus on predicting two clinically important CV outcomes in COVID-19: in-hospital cardiac arrest and thromboembolic events. In-hospital cardiac arrest is a clearly identifiable outcome and is often CV-related, thus it was selected to demonstrate the potential utility of COVID-HEART. Thromboembolic events are more difficult to identify and require imaging confirmation, thus, this outcome was selected to demonstrate the versatility of COVID-HEART in analyzing real-world clinical data and handling CV-specific outcomes. Finally, the predictor is tested in two different ways. First, it is tested with data from patients hospitalized after the end of data collection for patients in the development set, to ascertain that COVID-HEART can accurately predict risk in real time for new patients in the face of rapidly changing clinical treatment guidelines. The predictor is next tested with leave-hospital-out nested cross-validation to assess its performance when training and testing is done with data from different populations.

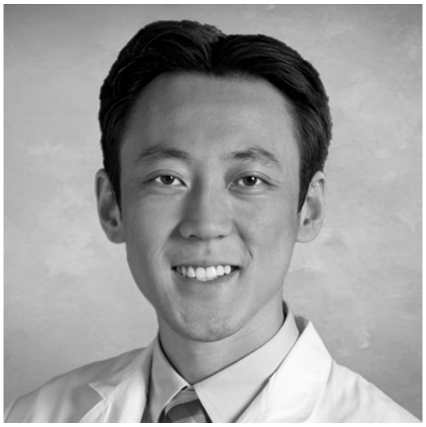
3,650 patients met eligibility criteria for prediction of cardiac arrest; 1100 (30.1%) were assigned to the test set according to the date cutoff. 2650 patients met eligibility criteria for prediction of thromboembolic events; 796 (30.0%) were assigned to the test set. Overall, 402 out of 3650 patients (11.0%) experienced cardiac arrests, 26 of whom experienced return to spontaneous circulation. Of these, 18 occurred in the intensive care unit (ICU), three occurred in a non-ICU inpatient unit, four occurred in intermediate care/ stepdown, and one occurred in long-term inpatient recovery care. 41 out of 2650 (1.5%) eligible patients experienced imaging-confirmed thromboembolic events. 36 additional patients had either an imaging-confirmed thromboembolic event within 24 hours of admission or had clinical suspicion of recent history of thromboembolic events prior to admission and were excluded for those reasons.



Cross-validation (purple) and testing (orange) receiver operating characteristic (ROC) curves for prediction of cardiac arrest using the optimal classifier configuration: a linear classifier with all feature types.

The optimized COVID-HEART predictor achieved AUROCs of 0.918 and 0.771, sensitivities of 0.768 and 0.500, and specificities of 0.903 and 0.879 for the full test set for prediction of cardiac arrest and thromboembolic events, respectively. Following the initial development-test split, the temporal development-test split was repeated and results over 20 iterations were aggregated to obtain 95% confidence intervals for the performance metrics. Mean cross-validation and test AUROCs were 0.917 (95% CI: 0.916-0.919) and 0.923 (95% CI: 0.918-0.927) for prediction of cardiac arrest and 0.757 (95% CI: 0.751-0.763) and 0.790 (95% CI: 0.756-0.824) for prediction of thromboembolic events, respectively.

For prediction of cardiac arrest, the mean test AUROC, sensitivity, and specificity for the left-out hospitals were 0.956 (95% CI: 0.936-0.976), 0.885 (95% CI: 0.838-0.933), and 0.887 (95% CI: 0.843-0.932). For prediction of imaging-confirmed thromboembolic events, the mean test AUROC, sensitivity, and specificity for the left-out hospitals were 0.781 (95% CI: 0.642-0.919), 0.453 (95% CI: 0.147-0.760), and 0.863 (95% CI: 0.822-0.904).



JONATHAN LING, PhD

Assistant Professor, Pathology, JHU

Comprehensive Analysis of Public Sequencing Archives to Uncover Novel Mechanisms of Pathogenesis in Amyotrophic Lateral Sclerosis

Amyotrophic Lateral Sclerosis (ALS) is a fatal adult onset motor neuron disease characterized by progressive loss of upper and lower motor neurons. Over 5,000 Americans die from ALS each year and with few approved treatments, the average life expectancy for patients is only two to five years after diagnosis. There is an urgent need to identify the genetic and environmental factors that underlie the onset and progression of ALS.

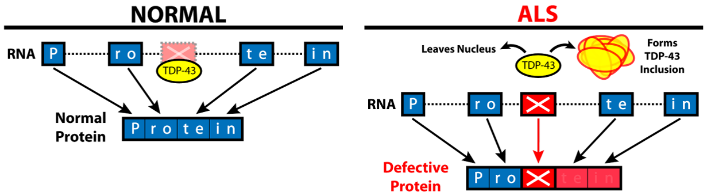


Figure 1

Work over the past decade has revealed that the RNA-binding protein TDP-43 is central to the pathogenesis of ALS. In postmortem brain tissue from ALS patients, TDP-43 forms pathological aggregates outside of the nucleus, where the protein normally resides. In 2015, we discovered that mislocalization of TDP-43 leads to the incorporation of deleterious, cryptic exons that disrupt protein synthesis (Fig. 1). Recent studies have further confirmed that cryptic splicing can induce the motor neuron loss observed in ALS. However, the mechanisms that initiate TDP-43 aggregation and loss-of-function are poorly understood. Various genetic and environmental factors have been proposed to explain how TDP-43 mislocalization and aggregation can occur in ALS, but these studies remain largely inconclusive.

With support from the IDIES Seed Funding Initiative, we have begun to leverage the vast publicly available RNA sequencing archives to uncover datasets that exhibit TDP-43 cryptic splicing. Our goal is to identify datasets that have no prior connection to ALS, with the hope of revealing experimental manipulations and environmental effects that can induce TDP-43 loss-of-function. Such findings would indicate novel mechanistic insights into ALS pathogenesis.

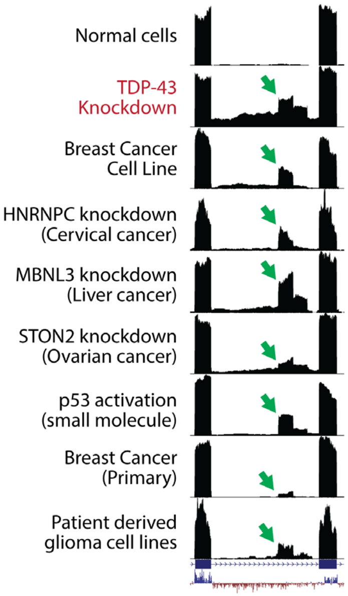


Figure 2

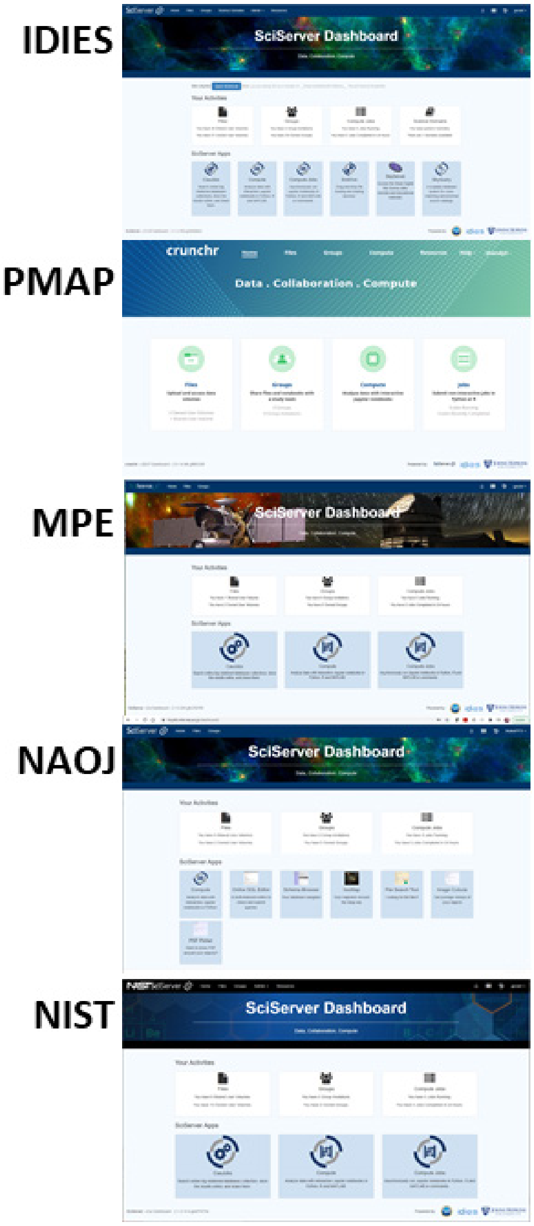
Preliminary analyses have successfully identified several hundred RNA-Seq samples that exhibit TDP-43 cryptic exons, filtered from the over 300,000 samples available in recount3. Many of these samples are related to experimental manipulations of TDP-43 or samples from ALS, which helps to validate our approach. Of the samples that have no obvious relation to TDP-43 or ALS, we have begun to replicate these computational results using in vitro model systems. Interestingly, we also find that RNA-Seq datasets from a variety of cancers appear to exhibit cryptic exons when certain cellular pathways are disrupted (Fig. 2). Further study is required to determine the relevance of cryptic exons in other human diseases beyond neurodegeneration.

Our cross-disciplinary effort to bridge neuropathology with big data science aims to answer questions that would otherwise remain inaccessible and to gain insights that may reveal novel therapeutic targets for preventing neurodegeneration.

SciServer Update

Gerard Lemson, Ph.D., Research Scientist, Director of Science, IDIES, JHU

SciServer is a collaborative science platform developed at IDIES that provides online storage and computational capabilities to scientists from a range of disciplines. SciServer supports traditional fields of astronomy, cosmology and fluid dynamics to disseminate results from large scale observational surveys as well as simulations, and smaller groups have used SciServer to support their efforts with dedicated storage and computational environments shared among their collaborators.



In this past year we explored how SciServer might interact with commercial cloud providers. Using AWS credits we investigated the use of cloud compute resources for analyzing data obtained from SciServer. Two example projects involved GPU nodes for machine learning on SDSS data and an Elastic Map Reduce node to replicate analysis on time series data from the Zwicky Transient Facility (ZTF). Further investigations will be performed using Azure and a grant from Microsoft.

Our collaboration with NIST is extended for another year. Using SciServer, we have started development of a predictor for damage caused by hurricanes prior to landfall.

A second NIST project aims to prepare SciServer for publishing the data from the AMBench 2022 project. This project will produce data sets from a large variety of detailed measurements of objects created using various methods of Additive Manufacturing. Purpose of these is to provide benchmarks for modelers trying to create simulations of these processes. Whereas testing is performed on the NIST SciServer instance, the public release of the data is planned to use the SciServer at IDIES.

The JHU School of Medicine, Carnegie Institution, and Weill Cornell University used SciServer in their annual Practical Genomics Workshop this year. The four-day workshop consisted of virtual classes with hands on sessions where the participants learned how to analyze single-cell RNA sequencing data using publicly available tools, such as R and Bioconductor. SciServer provided a homogeneous online computing environment for all participants which removed many of the usual problems related to individual participant computing environment.

Continued on next page.

CONTINUED FROM PRIOR PAGE (LEMSON)

Separate instances of SciServer are now deployed at 4 locations outside of IDIES: Precision Medicine Analytics Platform (JHMI), MPE (Munich, Germany),   NIST in the AWS cloud, and NAOJ (Japan). The SciServer platform at NAOJ supports the HSC-PFS (Hyper-Suprime Camera - Prime Focus Spectrograph) project for the SuMIRe (Subaru Measurements of Images and Redshifts) collaboration. This is a custom deployment of SciServer whereby a subset of the full complement of the SciServer tools is deployed using Kubernetes for container orchestration and automated deployment. The SciServer Login Portal, a customized Dashboard with traditional analysis tools developed at NAOJ, and Compute are the currently deployed components at NAOJ and supported by IDIES. The screenshot in Fig 3 shows the customized Dashboard with Compute and the NAOJ tools as well as the dashboards for the other three deployments.

If anyone is interested in publishing their data sets through the SciServer platform, or to use it in the class room, please feel free to contact Gerard Lemson (glemson1@jhu.edu), Science Director of IDIES and project manager of SciServer.

Shared CyberInfrastructure for Advanced Computing at Hopkins

Jaime Combariza, Ph.D., Director, Advanced Research Computing at Hopkins (ARCH), JHU

Advanced computing, integrating traditional High Performance, Data Intensive (DI), and Artificial Intelligence (AI, machine and deep learning) to enable fast growing, challenging projects in science and engineering founded on data-driven research, has become a priority for funding agencies. In the last two years, the National Science Foundation (NSF) has invested over $100M to deploy several powerful Advanced Computing systems that provide break-through cyberinfrastructure capabilities. In order to remain competitive, academic institutions must plan for these constant new developments in advancing computational research.

JHU is implementing new models to provide, maintain and sustain a state of the art facility. For an essential core facility of this magnitude, no single entity's financial infrastructure can be relied upon to support refresh cycles every three to five years. Thus arose the need for a communal support system.

A three pronged plan (as described in Figure 1) relying on the contributions of separate yet interdependent support groups is being implemented with great success. The HPC core facility is supported by the University through each of its schools, by research groups that join forces and apply for large grants (MRI, MURI, DURIP), and by individual researchers that add small condos to better serve their computational needs. The results are community-shared resources, guaranteed sustainability and the creation of a core facility with enough capacity to enable desired research. It is important to stress the "shared-resource" model, which allows any one group to use additional resources above their individual contribution. Additionally, JHU deanery are contributing to the annual operation of the facility, minimizing costs for JHU faculty in exchange for 'sharing' resources.

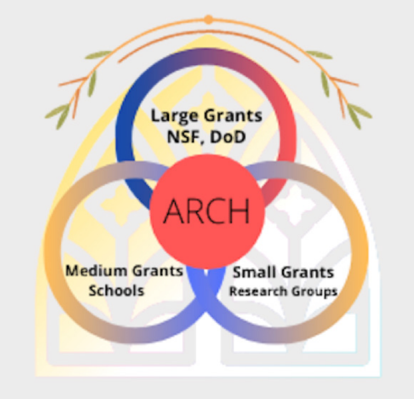


Figure 1. Rockfish shared infrastructure model.

Currently the new cluster, “Rockfish”, is growing fast: doubling the compute capacity and number of cores in a single year. Following the diagram described in Figure 1, the large grant contribution is composed of an NSF MRI grant that provided 388 compute nodes and 10PB of storage. This grant was also used to provide the shared infrastructure that will house other condos. Another large grant from DoD (DURIP) added 74 compute nodes, for a total of 462 nodes. The second contribution (first quarter 2022), provided by the JHU deans, will add 120 compute nodes plus 4 PB of storage. The final circle of contributions consists of over 150 compute nodes from 26 research groups and is expected to continue to expand over the next year. Rockfish will have over 730 compute nodes, 3.4PFLOPs theoretical peak and about 2.2PFLOPs sustained peak. This increased computational power makes resources at Hopkins comparable to those of peer institutions across the nation.

The new shared cluster, Rockfish, will have three sets of compute nodes: 680 regular memory compute nodes with 48 cores per node, 192GB RAM and a local NVMe SSD with 1 TB capacity; a set of 27 nodes with 48-cores and 1.5 TB of memory, and, finally, a set of 19 GPU nodes featuring the newest Nvidia technology (Ampera-100). GPU nodes will have either 2 or 4 A100, 40GB GPUs. All nodes and storage are connected via 100gbps Infiniband, allowing fast I/O and internode communication for parallel jobs. Rockfish has a parallel file system with 16 PB of storage and was placed in production in March 2021. Rockfish also supports two other important groups: Morgan State University, as a partner on the MRI, and the national community through the distribution of 20% of the computing resources via XSEDE.

The success of this three pronged approach continues to ensure the future of HPC at Hopkins. We would like to invite all current university research groups, regardless of their current research computing model, to contribute in this endeavor by procuring funding to add their own condos.

Mark O. Robbins Prize
in High Performance Computing

In recognition of a cherished friend and contributor to ARCH, IDIES and JHU, the Robbins Prize was instigated in 2020 to recognize outstandingly talented PhD students who reflect Dr. Robbins’ contributions to computational science and engineering. The Robbins Prize is made possible thanks to generous donations from the Department of Chemical and Biomolecular Engineering, Hopkins Extreme Materials Institute (HEMI), the Institute of Data Intensive Engineering and Science (IDIES), Department of Mechanical Engineering, and the Department of Physics and Astronomy.

Mark Robbins received his BA and MA degrees from Harvard University. He was a Churchill Fellow at Cambridge University, U.K., and received his PhD from the University of California, Berkeley. Dr. Robbins was a professor in Physics and Astronomy at JHU from 1986 until his untimely death in 2020. He was a renowned condensed matter and statistical physicist who played a key role in supporting the development of computational facilities at JHU, through his leadership for the Maryland Advanced Research Computing Center and the Institute for Data-Intensive Engineering and Science.

2021 Robbins Prize Recipients

The 2021 Robbins Prize awardees are: Dr. Sai Pooja Mahajan (Future Faculty Award), Dr. Karthik Menon (PhD Award), and Dr. Andrew Ruttinger (PhD Award).

Towards Deep Learning Models for Target-Specific Antibody Design

Dr. Sai Pooja Mahajan, Postdoc Fellow, Chemical & Biomolecular Engineering, JHU

Recent advances in machine learning, especially in deep learning (DL), have opened up new avenues in many areas of protein modeling. In the most recent Critical Assessment of Structure Prediction, a biennial community experiment to determine the state-of-the-art in protein structure prediction, DL-based methods accomplished unprecedented accuracy in the "difficult" targets category. Protein design is the inverse problem of protein structure prediction, i.e., the prediction of sequence given structure. Antibody design against an antigen of interest is a particularly challenging problem since it involves the design of the highly variable CDR loop regions to bind an antigen with reasonable affinity and specificity. In my talk, I will present some emerging DL-based methods to design proteins applied to the design of antibody CDR regions, and some of the early successes and important outstanding challenges in the use of current DL frameworks for protein, antibody and interface design.

Computational and Data-Driven Analysis of Aeroelastic Flutter

Dr. Karthik Menon, Postdoc Fellow, Cardiovascular Biomechanics Computation, Stanford University

The interaction of fluid flows with flexible and moving surfaces is a problem of wide applicability and exhibits highly non-linear responses of the fluid as well as the immersed surface. A particular source of complexity in these flows is the generation of several vortices, their interactions, and the non-linear forces they induce on immersed surfaces. In this talk, I will discuss our efforts in dissecting the flow physics of aeroelastically pitching wings using computational modeling and data-driven methods. I will demonstrate a novel energy-based tool to analyze, predict and control the often non-intuitive oscillation response of such systems. I will also describe data-driven techniques we have developed to analyze the vortex dynamics that drive the physics of such problems.

Dr. Andrew Ruttinger, Policy Analyst (PARDP), Natural Resources Canada

(Unavailable to speak at this year's symposium)

Andrew's work focused on modeling and simulation of a broad array of problems associated with sustainable energy. He used molecular simulation and density functional theory to study novel ways to selectively extract lithium from brine to offset current shortages of Li for batteries as well as creating the foundational rules affecting the design of "phase change materials" that could be used to store energy from intermittent sources. Finally, he worked with Dr. Sarah Jordaan (SAIS) to provide a techno-economic assessment (TEA) of processes that could move the U.S. towards "net-zero carbon" production of fuels, chemicals, and other renewable products in the context of a carbon tax."

IDIES Summer Student Fellowships

The IDIES Summer Student Fellowship program invites JHU undergraduate students to submit a 10-week summer research project with a focus in data science, and guidance from an IDIES faculty mentor. These projects are meant to provide an opportunity for students to participate in a full-time data science focused project, and encourage further interest in research while rounding out their undergraduate experience. In addition to a data-intensive computing focus, student projects must: relate to the IDIES mission, encompass the potential to advance knowledge, challenge and seek to shift current research/practice by utilizing novel concepts, approaches or methodologies, and benefit society and contribute to the achievement of specific, desired societal outcomes.

2021 IDIES Summer Student Fellows

The 2021 student fellowship recipients are:

Optimizing Resource Distribution Based on Sales Price Data Through Machine Learning

Chengkai Tian, Applied Mathematics & Statistics and Public Health Studies (Mentor: Jian Ni, CBS)

In the face of the COVID-19 pandemic's increasing severity, the goal of this research is to provide signals to indicate inefficiencies in medical resource allocation, and thus to help control the impact of the pandemic on people. The original design is to collect data on medical equipment prices and study their relationship with the severity of the pandemic in different locations. Due to the difficulty in collecting actual medical equipment prices, the research instead attempts to construct a shadow price to indicate the relation with medical resource allocations. While the research could not complete its original goal, it showed that statistical signals can become significant tools in controlling a disease. In this experiment specifically, the twitter data ends up being chosen as the signal, and the next step could be to study the sensitivity of this signal for better prediction and control.

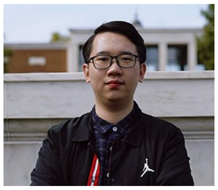
Humanizing Our Data: Proposal on Integrating Social and Behavioral Determinants of Health into Population Health Analytics

Jaxon (JunBo) Wu, History of Science, Medicine and Technology (Mentor: Jonathan Weiner, Chintan Pandya, BSPH)

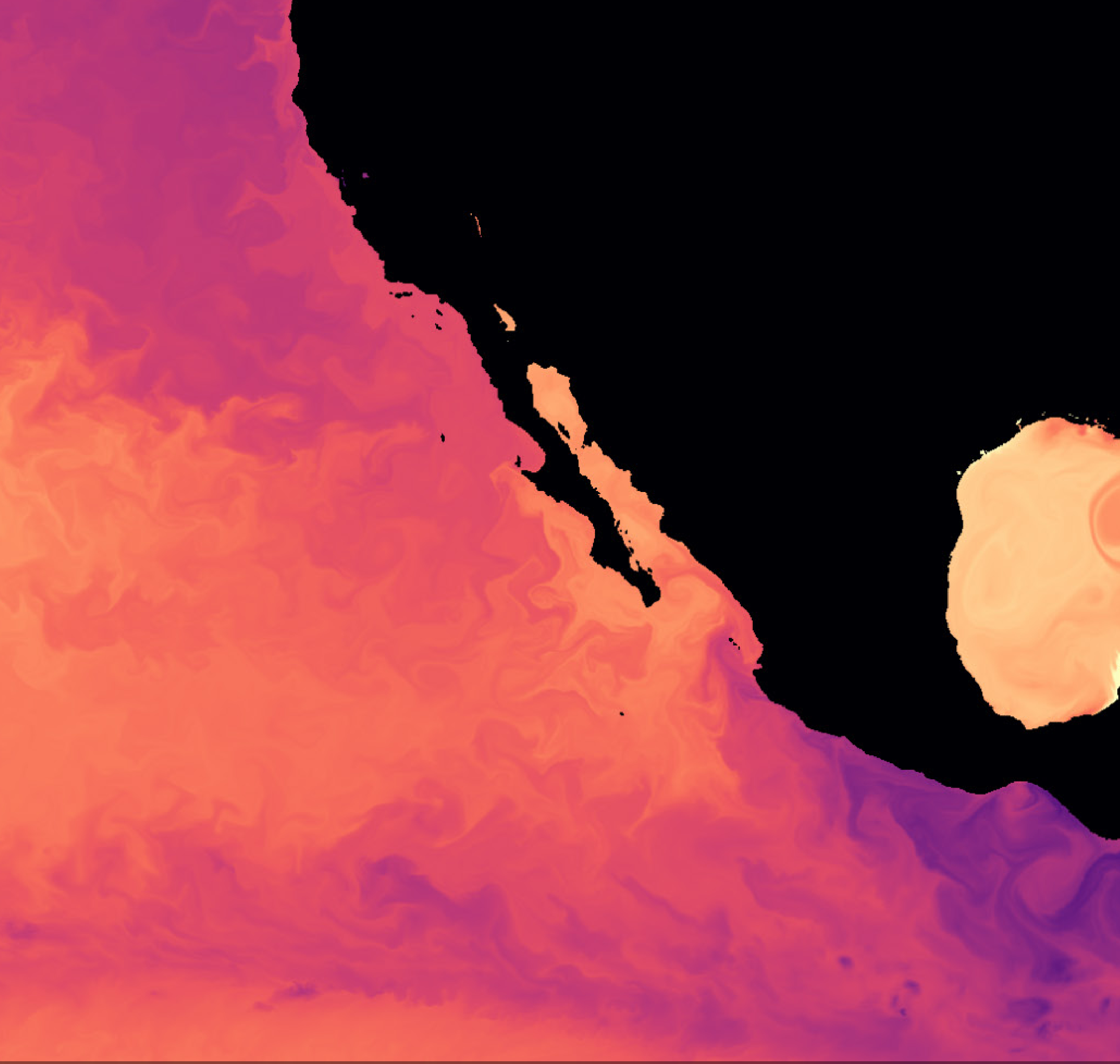
The aims of my project can be clearly defined as the following: first, I sought to explore the prevalence of SDoH needs in administrative claims, EHR, and claims-EHR combined data of a Medicaid population and extract the most widely documented ICD-10 Z-codes; second, I assessed the impact of ICD-10 social needs on healthcare utilization—narrowly defined as emergency department (ED) visits and inpatient hospitalizations—and health care expenditures—seen through total healthcare, pharmacy, and medical costs—of our Medicaid population. For each regression, we ran five different models that each had different combinations of independent variables. The base model contained our covariates and models 1 and 2 and models 3 and 4 respectively captured social needs markers and domains alongside ACG count versus ACG scores.

Using Machine Learning to Predict Surgical Case Duration in Operating Room Scheduling Optimization

Shengwei Zhang, Applied Mathematics & Statistics (Mentor: Tinglong Dai [CBS], Kimia Ghobadi [WSE])



Operating rooms (ORs) are the most expensive and financially productive resource in a hospital, and any disruption in their workflow can have a detrimental effect on the rest of the hospital operations. The goal of this project is to develop various machine learning models that can effectively predict the surgical case duration and compare the predictive power of these models to surgeon's empirical estimation. For the all-inclusive model, the comparison of three modeling algorithms' result shows that Random Forest and XG Boosting have a better predictive capability than Linear Regression and XG Boosting works much faster than Random Forest. For the service-specific model, the comparison of three modeling algorithms' result shows similar prediction accuracy on many OR services. The service-specific model obviously performs better than the all-inclusive model, but it also has some limitations. Since it only trains and tests on instances of a specific OR service, the data size is a huge confounder. In the dataset, 16 services have less than 100 instances, which make Random Forest and XG Boosting not suitable for these services. While in the all-inclusive model, since services are not treated respectively, the model performance is consistent for each service, regardless of the number of instances in the dataset.



JOHNS HOPKINS

INSTITUTE FOR
DATA-INTENSIVE
ENGINEERING & SCIENCE

IDIES • Johns Hopkins University • 3400 N. Charles St • Baltimore, MD 21218