THE INSTITUTE FOR DATA-INTENSIVE ENGINEERING AND SCIENCE

JOHNS HOPKINS
INSTITUTE *for*
DATA-INTENSIVE
ENGINEERING & SCIENCE

2025

OCT 1, 2024 - SEPT 30, 2025

ANNUAL REVIEW

THE INSTITUTE FOR DATA-INTENSIVE ENGINEERING AND SCIENCE

2025 IDIES ANNUAL REVIEW:

TABLE OF CONTENTS

We want to share your accomplishments!

There's still time to have your content included.

Please email publications, awards, and achievements from November 1, 2024–September 30, 2025 to kstevens@jhu.edu by October 23rd for inclusion in the final addition and check back later to learn more about what other IDIES members are up to.

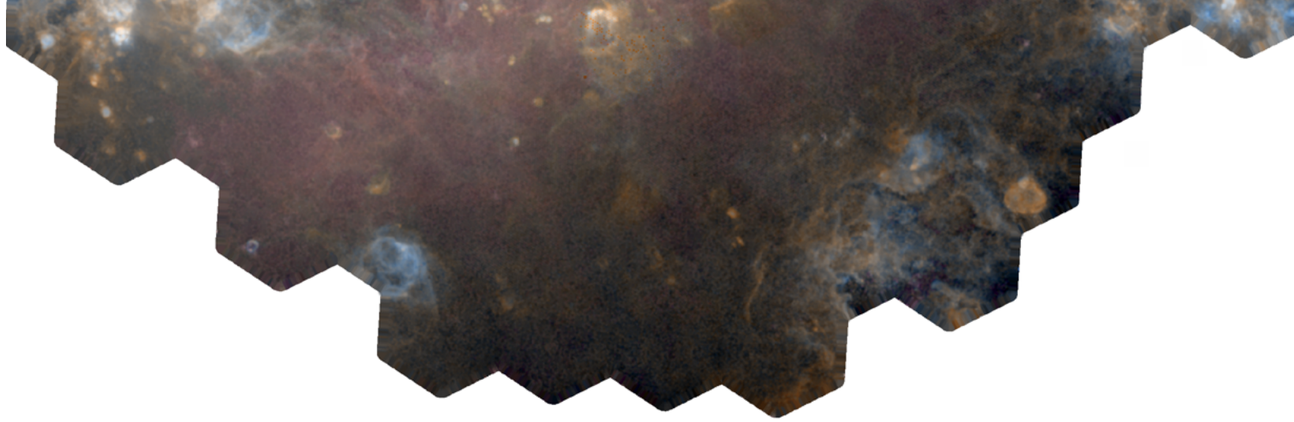


IMAGE CREDIT: SDSS/Héctor Javier Ibarra-Medel (IA-UNAM), and the SDSS-V

ON THE COVER

Image of the Large Magellanic Cloud (LMC), a satellite galaxy of our Milky Way, from the Sloan Digital Sky Survey (SDSS). IDIES hosts the complete SDSS data catalog and makes it available through the SciServer.org website.

This mosaic image is generated from more than two million spectra collected by the SDSS Local Volume Mapper (LVM) survey. Unlike most astronomy surveys, LVM focuses on the space between the stars, on the gas and dust from which new stars are born. Each pixel in the image shows an area just a few light-years across, more than 100,000 light-years away. Mapping spectra at this scale allows astronomers to analyze the distribution and physical conditions of ionized gas in unprecedented detail, gaining insight into the history and future of the Milky Way and its companions.

For this visualization, an RGB composite was created following the Hubble palette convention: the [O III] □5007 emission line is mapped to the blue channel, H□ to the green channel, and [S II] □□6716,6731 to the red channel. To emphasize the contrast between oxygen- and sulfur-dominated regions, the green component corresponding to H□ has been partially desaturated, thereby enhancing the visibility of the structures traced by [O III] and [S II]. This approach highlights the complex ionization structure of the LMC, revealing intricate nebular complexes, extended ionized shells, filamentary structures, and regions shaped by massive stellar feedback across the galaxy.

2025 IDIES

Layout and Design KATE STEVENS
Layout and Editing JORDAN RADDICK

Johns Hopkins University
3400 N. Charles St. Baltimore, MD21218
https://www.idies.jhu.edu/

MESSAGE FROM THE DIRECTOR

2025 IDIES ANNUAL REVIEW

Alex Szalay, PhD is the founding director of IDIES and the SSEC, a Bloomberg Distinguished Professor, Alumni Centennial Professor of Astronomy, a member of the National Academy of Sciences, and a professor of Computer Science. As a cosmologist, he workson the use of big data in advancing scientists understanding of astronomy, physical sciences, and life sciences.

THE **INSTITUTE** FOR **DATA-INTENSIVE ENGINEERING** AND **SCIENCE**

PHOTO CAPTION (right): IDIES director Alex Szalay at the 2024Symposium; PHOTO CREDIT: Alec Zabrecky

THE CHALLENGING YET CRITICAL ROLE OF IDIES WITHIN DSAI

Alex Szalay

It has been a challenging year for IDIES, as we have tried to better understand and define our role within DSAI. Part of the challenge we faced this past year was due to the DSAI leadership being in transition, but the DSAI infrastructure has also been evolving at the same time.

This year, we achieved major milestones with our flagship projects - 25 years with the Sloan Digital Sky Survey (SDSS), and 20 years with the Johns Hopkins Turbulence Database (JHTDB). In both cases, this has proven the trust that large user communities have placed in IDIES and the services that we provide,

This is our core competency and key to our role within DSAI, i.e., that we provide reliable and efficient access to the largest scientific datasets that represent “high value large datasets” for the worldwide scientific community, and in the process we earn the trust of our user communities. Data is the lifeblood of DSAI, and IDIES excels in making it available to the world.

As we sit on the brink of the dramatic changes that AI and AI-assisted discovery

ALEX SZALAY

are sure to bring about, IDIES is uniquely positioned to assume this critical role of being the DSAI Data Center.

In the coming year, we will continue to seek to increase IDIES’ profile within DSAI and JHU, in particular by engaging with the education community in collaboration with the JHU Library.

2025 IDIES ANNUAL SYMPOSIUM

AGENDA

Session	Time	Speaker
Session 1 | Morning (9:00–11:15)	**8:30–9:00**	Check-in, breakfast
	9:00–9:15	Opening Remarks by the Director, **Alex Szalay**
	9:15–9:45	**Stuart Feldman**—Keynote speaker (1 of 3) | **Scientific Software, Software Engineering, and Philanthropy**
	9:45–10:00	**Edward Hunter** | Updates on DSAI and SSECResearch Software Engineering
	10:00–10:05	Intro to IDIES Seed Funding program
	10:05–10:30	**Jung-Hee Seo** on behalf of the team of Rajat Mittal—Seed project (1 of 4) | eStomach – A Multiphysics In-Silico Model of the Human Stomach
	10:30–10:55	**Jiafang Song** on behalf of the team of Abhirup Datta—Seed project (2 of 4) | Correcting for Bias and Uncertainty in AI Algorithmic Outputs Used in Global Child Mortality Estimates
	11:00–11:15	15 min Morning Break
AM Break		
Session 2 | Morning (11:15–2:00)	**11:15–11:30**	**Gerard Lemson** speaking on SciServer
	11:30–12:15	Intro to Summer Student Fellowship (SSF) program followed by SSF projects
		Alex Larson | AI-Driven Cardiac Telemetry Classification for AFib Detection
		Srisha Nippani | A Non-Parametric Approach for Learning Interaction Laws in Agent-Based Systems
		Jooyoung Ryu | AI-driven Echocardiographic Model for Early Identification of Stress Cardiomyopathy
		James D'Alleva-Bochain | Understanding Learning: An Exploratory Analysis of the Geometry of Neural Network Weights and Activation
	12:15–12:20	Intros to Poster Gallery, Poster Lighting Talks
	12:20–12:40	Lightning Talks
	12:40–2:00	Lunch & Poster Gallery
Lunch & Poster Gallery		
Session 3 | Afternoon (2:05–4:05)	**1:30–2:00**	Poster entrants attend posters for gallery walk, Q&A
	2:00-2:05	Announcement of poster contest winner
	2:05–2:20	**David Elbert** speaking on the NASA IMQCAM Space Technology Research Institute and the Army Research Laboratory (ARL) Data at the Speed of Extreme Materials Discovery (DSEMD) Center
	2:20–2:50	**Beth Willman**—Keynote speaker (2 of 3) | **Creating the Conditions for Discovery: Rubin Observatory's LSST and a New Era of Data-Intensive Science**
	2:50–3:15	**George Liu** on behalf of the team of Francis Creighton—Seed project (3 of 4) | Artificial Intelligence-aided Video-based Assessment of Mastoidectomy Surgical Skills
	3:15–3:40	**Megan Gelement** on behalf of the team of Joel Bader—Seed project (4 of 4) | Solving the Phasing Problem of Infection Complexity in Polygenomic Malaria: Classifying Recurrent Infection
	3:40—4:05	25 min Afternoon Break
PM Break		
Session 4 | Afternoon (4:05 – end of event)	**4:05–4:25**	**Paulette Clancy** speaking about research computing infrastructure
	4:25–4:55	**Francois Lanusse**—Keynote speaker (3 of 3) | **Towards Scientific Foundation Models and How They Could Change ML In Survey Astronomy**
	4:55–5:00	Closing Remarks by the director, **Alex Szalay**
	5:00–6:00	Reception

2025 IDIES ANNUAL SYMPOSIUM

Keynote Speaker Bios

IDIES is honored to welcome three keynote speakers for the 2025 Annual Symposium.

Stuart Feldman

President of Schmidt Sciences; former VP of Engineering, East Coast, at Google; and former VP of

Schmidt Sciences financially supports the Scientific Software Engineering Center (SSEC) at Johns Hopkins University as part of their Virtual Institutes for Scientific Software initiative (VISS). Feldman is known for creating Make—the CLI tool—as well as the first Fortran compiler.

Talk title:

Scientific Software, Software Engineering, and Philanthropy

Beth Willman

CEO of the Legacy Survey of Space and Time (LSST) Discovery Alliance

Beth Willman, is the CEO of the LSST Discovery Alliance, the org aiming to ensure that any scientist with an excellent question for Rubin Observatory's LSST has access to the resources needed to answer it. She discovered the first known ultra-faint dwarf galaxies, including Willman 1, now known to be the most common type of galaxy in the Universe.

Talk title:

Creating the Conditions for Discovery: Rubin Observatory's LSST and a New Era of Data-Intensive Science

François Lanusse

Cosmologist and Astrostatistician at the French National Centre for Scientific Research (CNRS), and guest researcher of the Flatiron Institute

Lanusse's Polymathic AI project for astronomical observations was selected as one of the inaugural projects to utilize France's Jean Zay supercomputer.

Talk title:

Towards Scientific Foundation Models and How They Could Change ML In Survey Astronomy

SUPPORTING USE OF COMPUTATIONAL METHODS ON THE UCSF-JHU OPIOID INDUSTRY DOCUMENTS ARCHIVE

KEVIN S. HAWKINS, JHU PROGRAM DIRECTOR

The UCSF-JHU Opioid Industry Documents Archive (OIDA) collects, organizes, preserves, and provides free online access to millions of previously internal documents made public through legal settlements to enable multiple audiences to explore and investigate information which shines a light on the opioid crisis. This groundbreaking digital archive of opioid industry documents advances understanding of the root causes of the U.S. opioid epidemic, promotes transparency and accountability, and informs and enables evidence-based research and investigation to protect and improve public health.

For example, materials related to pharmaceutical distributors speak to the methods that they used to monitor the opioid supply chain, and the degree to which indicators of potential high-risk opioid distribution were acted upon. Policy analyses might examine how manufacturers engaged with advocacy organizations to achieve their policy objectives and strategies that manufacturers may have used to respond to regulatory concerns regarding opioid safety. The varied nature of the documents, which include corporate e-mail chains and internal company documents in connection with brochures and pamphlets, allow researchers to compare internal marketing strategies against the claims of safety and due diligence presented to practitioners and regulatory bodies. Because the litigation has focused on abatement of the opioid crisis, the documents also contain extensive information regarding how to best prevent further harms, and at what cost.

It currently includes over 5.4 million documents, with millions more expected to be added in the coming years. When dealing with such a massive corpus of documents, searching and browsing a digital library interface isn't always sufficient to find what you need. Therefore, the OIDA team at JHU has created the OIDA Toolbox, which provides a set of options for accessing the raw data behind the archive in order to support the needs of those who wish to use computational methods to study this invaluable corpus of data.

A key tool in the OIDA Toolbox is using SciServer to work with the OIDA corpus. Thanks to the generous support of the SciServer team, OIDA's raw data is available in IDIES storage, and we have developed a few Jupyter notebooks to allow users to quickly get started on accessing and analyzing OIDA documents in ways that you couldn't do through OIDA's main digital library interface. For example, a multi-disciplinary team of public health and legal researchers used SciServer to analyze all OIDA documents referencing veterans or the military to examine Janssen Pharmaceuticals' targeted marketing to veterans with chronic non-cancer pain that minimized addiction risks and exaggerated long-term benefits.

OIDA staff also use SciServer to analyze the quality of the archive's document metadata, which in most cases is automatically generated before the OIDA team receives the documents. Team members retrieve this inherited metadata and craft bulk corrections, sometimes to tens of thousands of documents at a time. These changes are reflected in archive's public interface to improve the experience of using filters and advanced search functionality. Pharmaceuticals' targeted marketing to veterans with chronic non-cancer pain that minimized addiction risks and exaggerated long-term benefits. OIDA staff also use SciServer to analyze the quality of the archive's document metadata, which in most cases is automatically generated before the OIDA team receives the documents. Team members retrieve this inherited metadata and craft bulk corrections, sometimes to tens of thousands of documents at a time. These changes are reflected in archive's public interface to improve the experience of using filters and advanced search functionality.

THE UNIVERSE IN A WEB BROWSER: IDIES HOSTS THE LATEST MAP OF EVERYTHING FROM THE SLOAN DIGITAL SKY SURVEY

JORDAN RADDICK
IDIES, SDSS INSTRUCTIONAL DESIGNER

IDIES has been the home of the entire catalog dataset of the Sloan Digital Sky Survey (SDSS) since the first SDSS public data release in 2001, and we are happy to announce that we are now hosting the latest public data release - Data Release 19 (DR19). The SDSS is a project launched in 2000 to map a significant fraction of the Universe in unprecedented detail, and has been publishing all of its catalogs - amounting to more than a half a Petabyte of data - through specialized services developed by IDIES for the past 25 years.

For the past 10 years, the SDSS catalog data has been served by the SciServer science platform developed by IDIES (www.sciserver.org), where you can use Jupyter Notebooks to develop complex workflows in Python or R to visualize, analyze, and publish results.

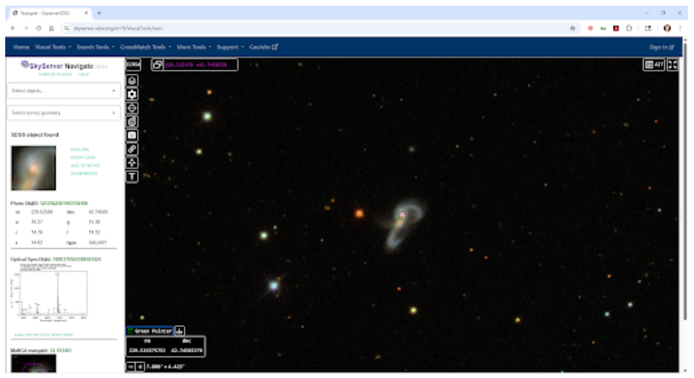


Figure 1. The new Navigate tool allows you to move through the sky smoothly. When you click on a star or galaxy in the main window, information about it appears in the left-hand frame.

There are Python notebook tutorials, also available on sdss.org, to help you learn how to work with DR19 data. The Getting Started data volume also contains Python notebooks for working with SDSS data, including interactive student notebooks for lecture and lab activities.

SDSS DR19 includes the following new data products:

- Spectra of more than half a million stars across our Milky Way galaxy and its companions, including the Magellanic Clouds
- Spectra of more than 300,000 galaxies and quasars (giant black holes at the centers of galaxies on the other side of the Universe)
- Measurements for stars, including temperature, size, motion, and chemical composition measurements for more than 20 elements
- Detailed maps of the distribution of interstellar gas and dust in the Helix Nebula

All SDSS data releases are cumulative, so all previously released measurements are still available in DR19. As always, you can navigate through the sky in an enhanced Google Maps-like interface (SkyServer Navigate) and click on stars and galaxies for more information.

[FE/H] as well to see what evolutionary stage low metallicity stars in our sample are in.

In [4]:

```
#visualise the data on the sky, the alpha-fe plane, and the stellar parameters
plt.figure(figsize=(22,5))

plt.subplot(1,3,1)
plt.scatter(data['RA'][mask], data['DEC'][mask], c=data['FE_H'][mask],vmin=-2,vmax=0.5,s=0.1,cm
plt.xlabel('ra [deg]',fontsize=20)
plt.ylabel('dec [deg]',fontsize=20)
plt.colorbar(label='[Fe/H]')

#remove stars with bad stellar parameters
mask_sp = (data['TEFF']>2e3)&(data['TEFF']<7e3)&(data['LOGG']<6)&(data['LOGG']>-1)&(data['SNR']
    &(np.abs(data['MG_H'])<2)&(np.abs(data['FE_H'])<2)
plt.subplot(1,3,2)
mgfe = data['MG_H']-data['FE_H']
# plt.hist(data['FE_H'],bins=np.linspace(-2,0.5,50),histtype='step',lw=4,color='k')
# plt.hist(data['FE_H'],bins=np.linspace(-2,0.5,50),lw=4,color='k',alpha=0.3)
plt.hist2d(data['FE_H'][mask_sp&mask],mgfe[mask_sp&mask],bins=300,cmap='bone_r',norm=LogNorm())
plt.xlabel('[Fe/H]',fontsize=20)
plt.ylabel('[Mg/Fe]',fontsize=20)
plt.xlim(-2,0.7)
plt.ylim(-0.5,0.6)

plt.subplot(1,3,3)
plt.scatter(data['TEFF'][mask_sp],data['LOGG'][mask_sp], c=data['FE_H'][mask_sp],vmin=-2,vmax=0
plt.xlabel('T$_{eff}$ [K]',fontsize=20)
plt.ylabel('log(g)',fontsize=20)
plt.colorbar(label='[Fe/H]')
plt.xlim(7000,3000)
plt.ylim(6,-1)
```

Out[4]:
(6.0, -1.0)

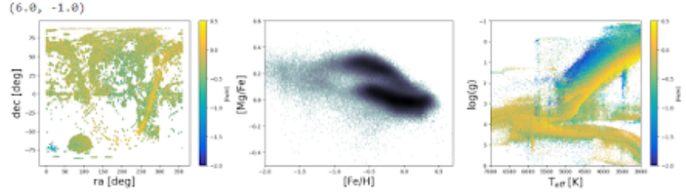


Figure 2. This example notebook is available through the SciServer SDSS SAS data volume and sdss.org

OPEN SCISERVER

ARIK MITSCHANG,
RESEARCH SOFTWARE ENGINEER, IDIES

Open SciServer is no longer a dream, but a reality! This update represents the second year of our NSF cyberinfrastructure sustainability grant, and we are happy to report that the primary components of SciServer are now available within a single public github repository. https://github.com/sciserver/opensciserver.

To make this all a reality, the SciServer team has spent effort in three key areas related to the software system: code cleanup, build and release automation, and documentation. With some exceptions, all of the SciServer web application components have been collected from disparate repos into the new single repo. At the same time, they have been updated to use the same Gradle build process and updated to Java 17 (from 8 in several cases), and stylistic changes and checks have been applied.

A streamlined build and release process is important to ensure the usefulness of any complex software system of many parts, such as SciServer. Collecting our components into a single repo was a decision we made early on to ensure that we can guarantee compatibility of any change across the system, and to simplify installation and upgrades. We have added automation such that each release version, merge to mainline branch, or pull request gets a full build, and all necessary artifacts for installation or upgrade are made available online via github's web serving and container registry services. This means we can create a fresh test sciserver installation for any given change with a single command! Updates to a production installation are as simple as updating a single version string. Importantly, to prove the utility of our new process, we have updated the official production SciServer here to use this exact process.

We have begun a concerted effort to write developer documentation for all components - something which only existed sparsely prior to the Open SciServer grant funding. The primary target for these docs are maintainers of SciServer installations, and those that may develop SciServer code to make contributions. For this we use Sphinx, which uses the re-structured text format and builds into html, pdf, and epub formats. At the time of writing we have added documentation for the fileservice, resource and access control manager (RACM), the compute jobs manager (COMPM), the SciServer user client library (SciScript), and the future java-based CasJobs SQL web service replacement SciQuery.

There exist three independent external installations of SciServer that we know of: One at Max Planck Institute for Extraterrestrial Physics serving eRosita data, one at Brookhaven National Laboratories for internal use serving GPU-intensive workloads, and one at JHU Precision Medicine in a HIPAA-safe environment serving clinical use-cases. We have notified these stakeholders of our open repo, which was well received. Looking forward, we anticipate that all will benefit from the streamlined process, and will be able to leverage github facilities to get more involved, and even contribute code via pull requests!

^ Open Sciserver is supported by NSF award 2311791 - "Sustainability: Open SciServer: A Sustainable Data-Driven Science Platform"

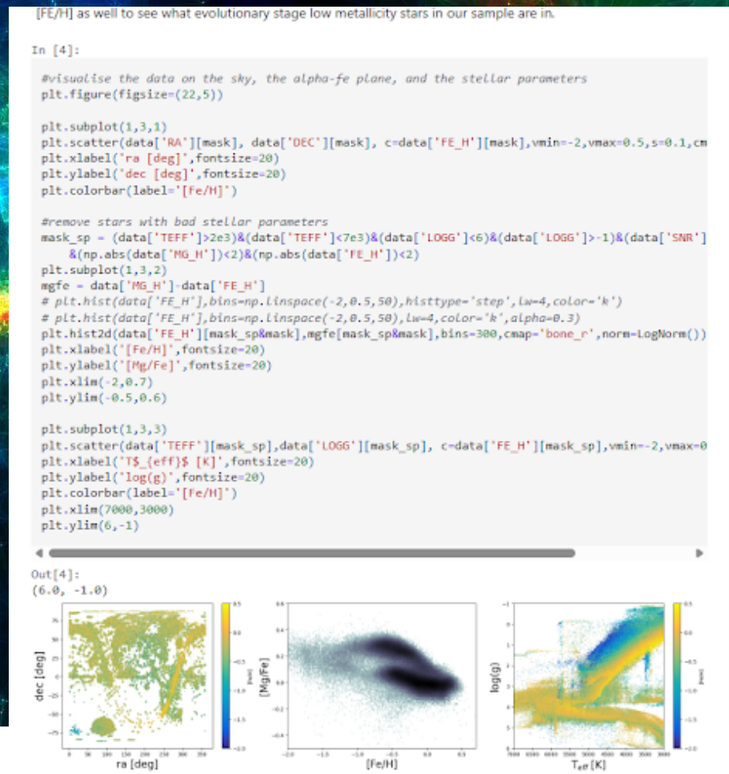
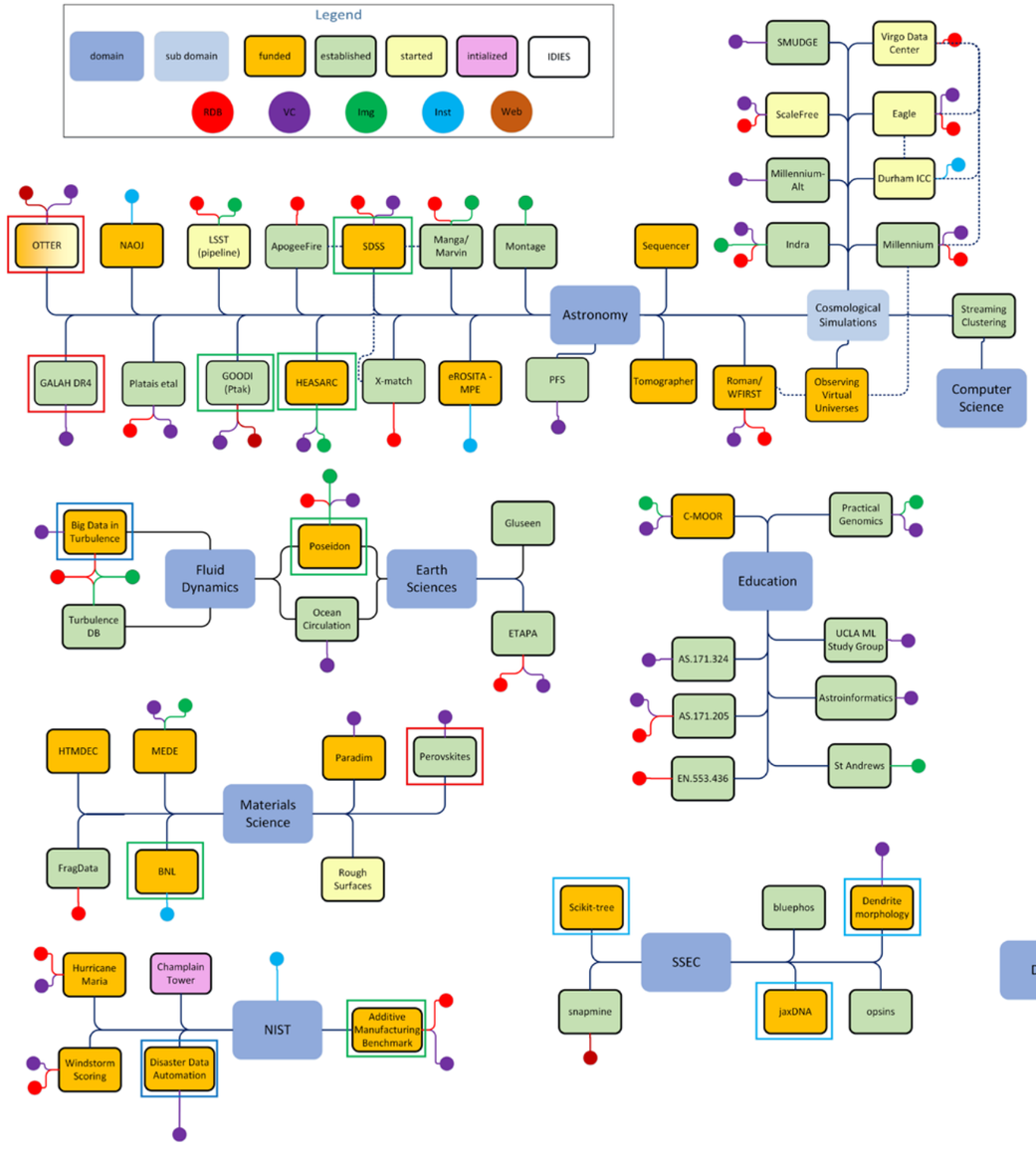


Figure 2. This example notebook is available through the SciServer SDSS SAS data volume and sdss.org.

SCISERVER SCIENCE DOMAINS



SEED FUNDING PROJECT UPDATE

SOLVING THE PHASING PROBLEM OF INFECTION COMPLEXITY IN POLYGENOMIC MALARIA: CLASSIFYING RECURRENT INFECTION

JOEL BADER
MATTHEW IPPOLITO
MEGAN GELEMENT

The frequent occurrence of polyclonal malaria parasite (e.g., Plasmodium falciparum) infections in moderate and high transmission areas can confound efforts to identify the source of recurrent infection in longitudinal patient data. Recurrent infections may be recrudescent (arising from one or more of the same clones as the initial infection) or new (representing an infection from a new clone or set of clones). This distinction is clinically relevant, as the knowledge is required to identify therapeutic failures and detect drug-resistant parasites.

This difficulty in infection classification is a direct result of the phasing problem in complex malaria infection, which refers to the need to deconvolve polyclonal infection data into individual clonal genotypes. Malaria parasite genomes are repetitive with high AT content, parasite sequencing is error-prone, and existing computational approaches for polyclonal infection analysis are optimized for molecular epidemiology questions—such as estimating transmission networks or mapping population structure—rather than for the clinically-focused task of identifying recrudescent clones within mixed infections. Our project expands these approaches to address the distinct demands of clinical recurrence classification.

During the IDIES award period to date, we have identified internal and external datasets for use in training and testing, cleaned experimental artefacts, and interrogated the assumption of independence between loci

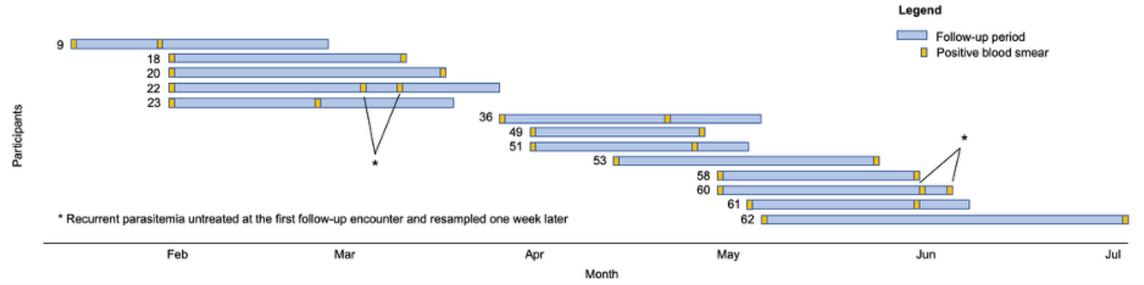


Figure 1: . Time series of participants who experienced recurrent Plasmodium falciparum infection. Two participants (22 and 60) had recurrent parasitemia that was untreated at the first instance of detection and resampled one week later. The paired recurrences of known identical infections serve as positive controls for genotyping comparisons.

often made by existing models for polyclonal infection analysis. To ensure clinical relevance, we are focusing on molecular inversion probe (MIP) panels. MIP panels are increasingly used in clinical trials and therapeutic efficacy studies because they measure alleles across the P. falciparum genome without the expense or alignment ambiguity of whole genome sequencing. While alleles at MIP panels are not readily phased into haplotypes, the same would be true of data from short-read next-gen sequencing.

Longitudinal MIP data from therapeutic efficacy studies in Zambia, Rwanda, and Uganda are the primary internal datasets for this study. Our current dataset comprises fourteen patients with recurrent infections, including two patients with recurrent parasitemia who went untreated off-protocol between the time of first and second follow-up sampling and thus approximate positive known controls for recrudescent infection (indicated by asterisks in Fig. 1).

Public data from the MalariaGEN consortium will address the assumption of locus independence common to polyclonal analysis models. The MalariaGEN Pf8 database contains 33,325 P. falciparum samples annotated with an FWS statistic, which measures the expected clonality of the sample on a scale from 0 to 1 (FWS = 1 – HW/

CONTINUED ON NEXT PAGE

SEED FUNDING PROJECT UPDATE

SOLVING THE PHASING PROBLEM OF INFECTION COMPLEXITY IN POLYGENOMIC MALARIA: CLASSIFYING RECURRENT INFECTION

CONTINUED FROM PREVIOUS PAGE

HS, where HW = individual sample heterozygosity and HS = local parasite population heterozygosity). Higher FWS indicates lower genomic complexity. We subset African samples that pass quality control with an FWS ≥ 0.95 (7,763 samples), a proxy for infection by a single clone, and are using these to build custom linkage disequilibrium (LD) matrices relevant to targets of our MIP panel. This quantitative measure of locus independence will inform development of our model. MalariaGEN data will also be used to develop synthetic "MIP" panels mimicking the structure of our data for use in training.

We also developed an independent locus model for predicting recrudescent from new recurrent infection in MIP data. Without accounting for data missingness or linkage disequilibrium, maximum likelihood could not distinguish between recrudescence versus new infection, highlighting the necessity to account for linkage disequilibrium by exploiting data resources from MalariaGEN. In parallel, we have made progress in error correction in microhaplotype calling, addressing artefacts that arise from trailing-base misalignment ("ragged-end" or "overhang" sequences)—cases where near-identical microhaplotypes are incorrectly classified as distinct.

By developing reliable and practical measures of recrudescent infection by phasing complex infections, we will extract more information from longitudinal malaria studies and improve patient care. The models we hope to develop could be deployed in ongoing clinical studies to classify recurrences, improve treatment policy, identify drug resistance, and develop more effective therapeutics.

ABOUT THE PI: Joel Bader

Joel S. Bader, a professor in the Department of Biomedical Engineering, is a leading innovator in genomics and biotechnology. His research explores the connection between genotype and phenotype, relating an organism's genetic material to its observable characteristics.

Bader develops new computational methods to define how inborn genetic variants and acquired mutations lead to disease, primarily complex genetic disorders and cancer, with the goal of identifying new therapeutic targets. He also leads synthetic biology projects that design and build entire chromosomes and genomes to create new forms of life. He also leads an effort to create cells that rely on RNA (ribonucleic acid) rather than DNA for their genetic information. In related efforts, Bader's lab develops technologies for biosafety and biosecurity, preventing escape of engineered life and identifying evidence of genetic engineering in DNA samples.

ABOUT THE SYMPOSIUM PRESENTER: Megan Clement

Megan Gelement is a Ph.D. candidate in Biomedical Engineering at the Johns Hopkins University School of Medicine. She received her M.S. in Computer Science from Tufts University in 2023, and her B.S. in Biology from the University of Alabama in 2016. In the interim, she worked in the healthcare and publishing industries. Her current research focuses on the computational deconvolution of polyclonal Plasmodium falciparum infections, with an aim to distinguish recrudescent from new recurrent infections using longitudinal genomic parasite data. Her thesis advisor is Dr. Joel Bader, Professor of Biomedical Engineering and Director of the Institute for Computational Medicine, and she is additionally advised on malaria genomics projects by Dr. Matthew Ippolito, Associate Professor of Medicine and Director of Clinical Epidemiology for the Southern and Central Africa International Centers of Excellence for Malaria Research.

SEED FUNDING PROJECT UPDATE

eSTOMACH – A MULTIPHYSICS IN-SILICO MODEL OF THE HUMAN STOMACH

RAJAT MITTAL
MICHAEL SCHWEITZER
JUNG HEE SEO

The stomach plays a pivotal role in digestion. It receives and stores ingested food, mixes, softens and breaks it down, and then delivers digesta to the duodenum in a regulated manner. This process called “gastric digestion is accomplished by the complex interactions of gastric motility (stomach wall motion), fluid dynamics inside the stomach, and flow mediated biochemical reactions between the digesta and enzymes.

Although the science of food digestion has been studied for over 300 years, our understanding of the underlying biomechanics and chemo-fluid dynamics principles of the gastric digestion process is extremely limited due to a complex multi-physics nature. Conducting studies of gastric biomechanics and chemo-fluid dynamics is challenging. In-vivo studies are difficult since imaging of stomach contents during the time-course of digestion is complex, costly and involves risks. In-vitro models to date have been unsuccessful in recapitulating both the complex contractile movements as well as the secretory behavior of the stomach with sufficient fidelity. Computational “in-silico” models hold great promise in this regard.

The objective of the present study is the development of an in-silico, virtual human stomach model called “eStomach” which simulates all the key biomechanics/biophysics in gastric digestion: gastric motility, fluid dynamics with flow-structure interaction, and flow mediated biochemical reactions for digestive process. eStomach will not only allow us to investigate the details of gastric biomechanics but also enable “virtual” gastric surgeries and prediction of expected outcomes. This could revolutionize bariatric surgery by providing more predictable and potentially better outcomes for patients.

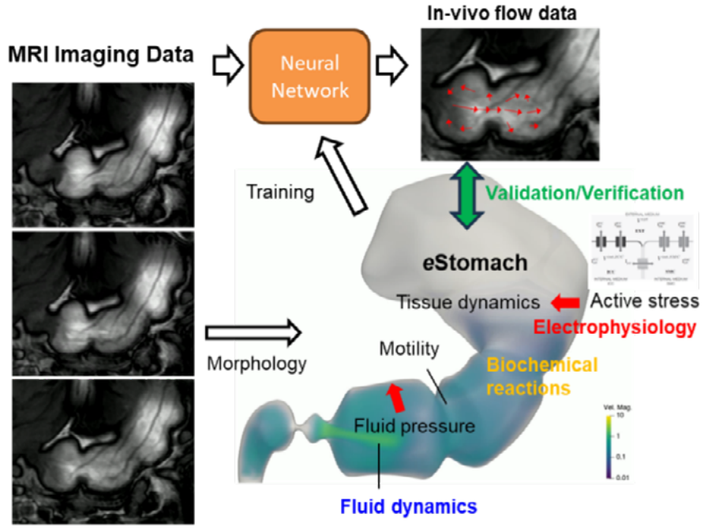


Figure 1: Schematic diagram of eStomach. Patient-specific, predictive virtual stomach model is developed from MRI imaging data

We have successfully developed a method to generate a patient-specific stomach model by using the in-vivo MRI imaging data. The gastric mixing and emptying are then simulated by a high-fidelity, immersed boundary based, multi-physics flow solver developed by our research team. The key physiologies of the stomach: fundic accommodation, terminal antral contraction, and the phasic opening of the pyloric sphincter are all incorporated into the model. The pathological conditions such as the enlargement and disruption of pyloric sphincter due to pyloric intervention procedures or pyloric incompetence are also implemented.

The developed eStomach model has been used to investigate the influence of food properties (density and viscosity), motility disorders, and pyloric intervention procedures on gastric mixing, duodenogastric reflux, and solid-food trituration. The gastric emptying rate and the transpyloric pressure gradient predicted by the eStomach were in good agreement with existing in-vivo measurements in the literature.

The eStomach model developed in this study allows the exploration of a wide set of parameters related to

CONTINUED ON NEXT PAGE

SEED FUNDING PROJECT UPDATE

CONTINUED FROM PREVIOUS PAGE

gastric function. eStomach offers a novel and potentially effective modality for examining the effect of gastric conditions and gastric surgical procedures on gastric function. We will continue the development of eStomach for application to virtual bariatric surgery.

The results obtained from this study were used for the NIH grant proposal "A Magnetic Resonance Imaging-Based Digital Twin of the Stomach for Improved Evaluation and Phenotyping of Gastroparesis" and the proposal is under review now.

> *eStomach will not only allow us to investigate the details of gastric biomechanics but also enable "virtual" gastric surgeries and prediction of expected outcomes.*
>
> *This could revolutionize bariatric surgery by providing more predictable and potentially better outcomes for patients.*

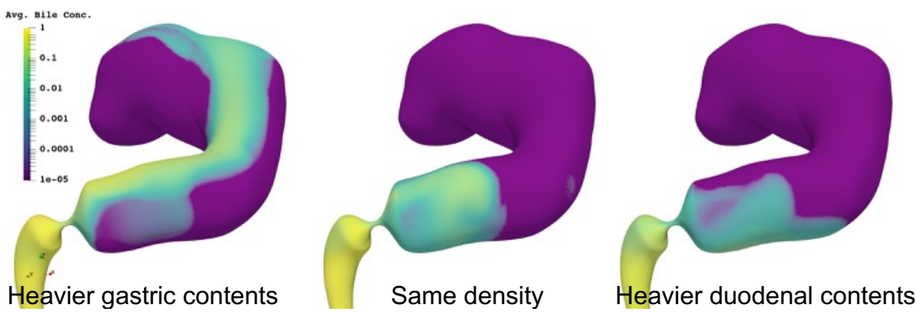


Figure 2: Bile reflux simulated by eStomach. Averaged concentration of bile on the stomach surface if the food is heavier, of same density, or is lighter than the duodenal contents.

ABOUT THE PI:

Rajat Mittal, a professor of mechanical engineering, is an expert in computational fluid dynamics.

Mittal's research group, the Flow Physics and Computation Lab, develops and employs computational methods to model a variety of flows, with a special focus on immersed boundary methods, vortex dominated flows, biomedical fluid dynamics, biological and bioinspired locomotion (swimming and flying), bioacoustics, active flow control, fluid-structure interaction, and high-performance computing. Due to the cross-disciplinary nature of his work, he collaborates extensively with researchers from other disciplines, such as zoology, cardiology, robotics, biomechanics, and dynamics.

Mittal's work has had a significant impact on the field of computational fluid mechanics and computational biomechanics. His recent research has focused on active and passive control of flows, swimming and flying in animals, multiphysics modeling of heart murmurs, blood clots and heart valves, and the biomechanics of digestion. His research has received funding from the National Institutes of Health, National Science Foundation, U.S. Air Force, Office of Naval Research, Army Research Office, Defense Advanced Research Projects Agency (DARPA), and NASA.

ABOUT THE SYMPOSIUM PRESENTER: Jung-Hee Soo

Jung-Hee is an associate research professor at the Johns Hopkins University, department of mechanical engineering. His research is focused on the biological and biomedical fluid flow and his expertise is computational modeling of multiphysics problems including fluid-structure interaction, acoustics, and chemical transport. He received his PhD at the Korea University in 2006 and was a Post-doctoral fellow at the Stanford University, department of aeronautics and astronautics. He joined the Johns Hopkins in 2009 and performed the research projects about human phonation, cardiovascular flow, heart sound, and flying and swimming animals.

SEED FUNDING PROJECT UPDATE

ARTIFICIAL INTELLIGENCE-AIDED VIDEO-BASED ASSESSMENT OF MASTOIDECTOMY SURGICAL SKILLS

FRANCIS X CREIGHTON
GEORGE LIU

Traditional surgical skills assessments are often subjective, relying on the attending surgeon's personal evaluation. To support the shift toward competency-based medical education, there's a critical need for objective, quantifiable tools to ensure fair and accurate assessment of technical skill. Video-based assessment shows significant promise as one such tool. In otolaryngology—head and neck surgery, mastoidectomy is a fundamental yet complex procedure. It involves using a high-speed drill to sculpt the mastoid bone, navigating within millimeters of critical structures such as the brain and facial nerve. Although video recording is standard practice via surgical microscopes, these videos are rarely used for skills assessment because conventional analysis methods require time- and labor-intensive manual annotations.

We are developing an AI-aided approach for video-based assessment of mastoidectomy, with the goal of objectively evaluating surgical technical skills. Our initial focus is on creating computer vision tools to automatically extract quantitative features of surgical instrument movements from video. To do this, we've built a comprehensive dataset of mastoidectomy videos from over a dozen surgical residents and are adding videos from experienced surgeons with

over a decade of dedicated ear surgical experience. These videos focus on two critical procedural steps: opening the mastoid antrum and exposing the incus. Using this dataset, we have developed a deep learning computer vision framework that can accurately track the drill and suction instruments. This approach uses one-shot memory that requires manual annotation of only a single video frame. Preliminary data demonstrates that this approach allows for fast and accurate segmentation of these instruments (Figure 1). We are now expanding this model to track the instrument tip location and the drill's on/off state (Figure 2, next page).

Our second objective is to correlate instrument movements with technical skill. We are collecting expert ratings by having experienced ear surgery attendings perform blinded assessments of our videos using a modified

CONTINUED ON NEXT PAGE

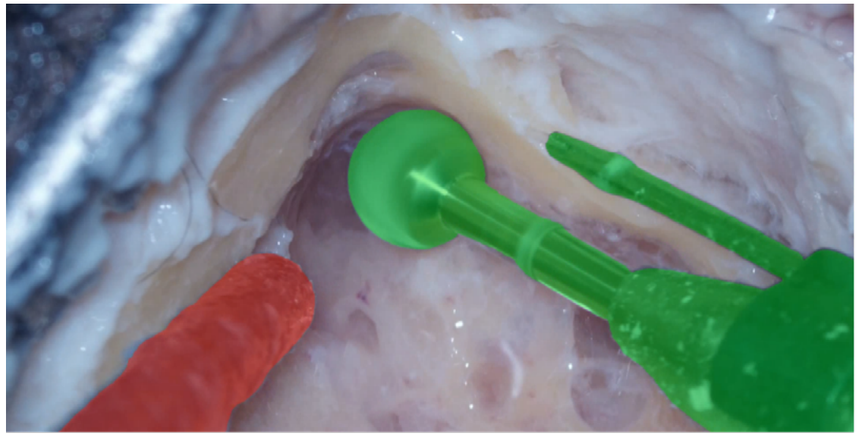


Figure 1: Deep learning, one-shot segmentation examples for the drill (green) and suction (red) instruments in an example frame from a mastoidectomy surgery video. Here, a large 6 mm drill burr is removing bone from the superior-posterior aspect of the bony external auditory canal to open the mastoid antrum.

SEED FUNDING PROJECT UPDATE

ARTIFICIAL INTELLIGENCE-AIDED VIDEO-BASED ASSESSMENT OF MASTOIDECTOMY SURGICAL SKILLS

CONTINUED FROM PREVIOUS PAGE

mastoidectomy OSATS (Objective Structured Assessment of Technical Skill) survey. By pairing these expert ratings with our AI-extracted data on instrument movements, we will identify specific motion patterns that indicate surgical proficiency. We anticipate that this research will serve as a foundational study for the future implementation and validation of an AI-aided video-based assessment tool. If successful, our approach could be adapted to assess surgical skills across other specialties and contribute to the adoption of fair, accurate, and objective tools for evaluating technical skill in surgical residency.

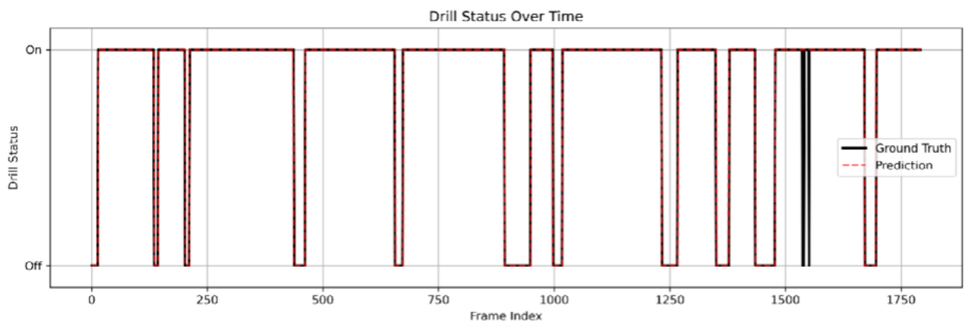


Figure 2: Deep learning predictions of the drill's on/off state. Predictions are shown for a video used for validation of the model.

ABOUT THE PI: Francis Creighton

Dr. Creighton is an ear and lateral skull base surgeon. Dr. Creighton graduated from the Georgia Institute of Technology with a Bachelor of Science in biomedical engineering and earned his M.D. from Emory University. He completed a general surgery internship at the Mass General Hospital and residency in otolaryngology-head and neck surgery at Harvard. Following this he completed a fellowship in neurotology and lateral skull base surgery at Johns Hopkins.

His clinical practice specializes in surgical and medical management of middle ear, inner ear, skull base and facial nerve disorders. These include skull base tumors, vestibular schwannomas (acoustic neuromas), hearing loss, cholesteatoma, cochlear implantation, stapedectomy, CSF leaks and ear drum perforations. He is trained in minimally invasive and endoscopic approaches to the ear for cholesteatoma and ear drum perforations, which reduces the need for visible incisions, and performed the first in office tympanoplsty repairs at Johns Hopkins.

ABOUT THE SYMPOSIUM PRESENTER: George S. Liu

George S. Liu, MD is a 2nd year fellow in Neurotology and Skull Base Surgery at Johns Hopkins University. His research focuses on developing precision approaches to treating hearing and balance disorders through the translational application of artificial intelligence and computer vision methods to clinical datasets. As a surgeon-scientist, he combines his clinical experience which serves as a direct source of research questions with his technical expertise in applied computation and machine learning. He collaborates with scientists, clinicians, and engineers at the Johns Hopkins School of Medicine and Whiting School of Engineering. Prior to coming to Johns Hopkins, he completed his undergraduate degree in physics at Princeton University and medical school and otolaryngology—head and neck surgery residency at Stanford University.

SEED FUNDING PROJECT UPDATE

CORRECTING FOR BIAS AND UNCERTAINTY IN AI ALGORITHMIC OUTPUTS USED IN GLOBAL CHILD MORTALITY ESTIMATES

ABHI DATTA

Estimates of population-level cause-specific mortality fractions (CSMFs) from verbal autopsy (VA) data in low-and-middle-income countries (LMICs) often rely on AI-algorithm predicted causes of death, which are subject to both bias and uncertainty. Calibration methods can adjust for algorithmic bias but typically fail to propagate this uncertainty into downstream analysis. Cut-Bayes or cutting-feedback method provides a principled framework to propagate upstream uncertainty from AI model generated output when used in downstream analysis. However, MCMC-based implementations of cut Bayes are computationally infeasible, and existing variational inference methods suffer from statistical limitations, which often rely on crude approximations of upstream distributions or restrictive assumptions on downstream variational families, limiting accuracy. The proposal aims to (1) develop a general deep-learning based framework for unidirectional propagation of uncertainty in AI-generated outputs when used in downstream analysis; and (2) apply this framework to produce improved estimates of neonatal and child CSMFs using data from VA studies in LMICs.

We have completed much of Aim 1. This includes designing a Cut-Bayes framework that directly incorporates uncertainty-quantified upstream results, for example, MCMC samples of algorithmic bias estimates, into downstream analysis, ensuring uncertainty is propagated without reverse feedback. Our proposed method, Neural Network–based Variational Inference for Cut-Bayes (NeVI-Cut, Fig. 1), conditions on each upstream particle (one sample) and employs normalizing flows to approximate conditional variational distributions for downstream parameters. Normalizing flows are neural network–parameterized expressive families of probability distributions. We then implemented on a triply stochastic, scalable algorithm. We have also established theoretical justification for universal approximation capability of NeVI-Cut. Finally, we have conducted extensive numerical analysis using simulated and real data to demonstrate that NeVI-Cut achieves accuracy comparable to the gold-standard nested MCMC sampling, while offering substantial computational gains. Its success stems from two key design features: (i) incorporating upstream samples to preserve the uncertainty in downstream analysis, and (ii) employing neural network transformations within each flow layer to enable expressive and flexible posterior approximations.

We have also started work on Aim 2. We have applied NeVI-Cut to obtain estimates of child and neonatal mortality in Mozambique. We use publicly available posterior samples of confusion matrices of cause-of-death predicting AI algorithms, pre-computed from a previous (upstream) analysis. These samples are incorporated into the downstream analysis to calibrate cause-specific mortality fractions (CSMFs) from COMSA verbal autopsy data in Mozambique. We benchmarked NeVI-Cut against multiple imputation (the current gold standard for cut-Bayes models), parametric variational inference approaches, and full Bayesian models. The CSMF distributions were often multimodal and skewed (Fig. 2). While full Bayesian analysis produced biased results, NeVI-Cut matched the accuracy of multiple imputation while

CONTINUED ON NEXT PAGE

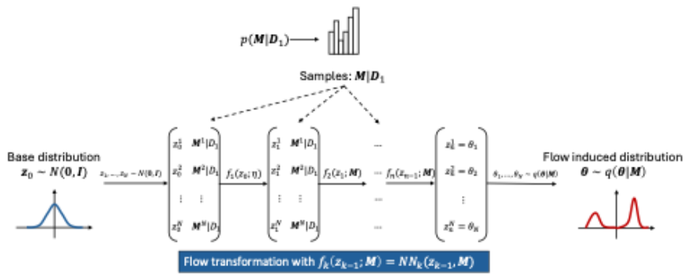


Figure 1. NeVI-Cut framework. The upstream analysis uses the CHAMPS dataset (D1), which contains paired information on causes of death (COD) from both verbal autopsy (VA, processed by an AI algorithm) and gold-standard COD from minimally invasive tissue sampling (MITS). Comparing VA COD against MITS COD yields estimates of algorithmic bias, summarized in a confusion matrix (M). MCMC samples of confusion matrix are then passed as inputs into the downstream analysis. NeVI-Cut transforms a base distribution through a sequence of invertible and differentiable functions f (parameterized by neural networks) to obtain a conditional variational

CORRECTING FOR BIAS AND UNCERTAINTY IN AI ALGORITHMIC OUTPUTS USED IN

CONTINUED FROM PREVIOUS PAGE

substantially reducing computation time. This analysis has served as a proof-of-concept demonstrating the success of NeVI-Cut in an LMIC.

For Aim 1, we plan to finalize the manuscript and release the software as a public open-source package by the end of 2025. For Aim 2, we aim to complete a manuscript on the methods development and benchmarking of NeVI-Cut. We also plan to expand the application of NeVI-Cut to propagate uncertainty in the CA-CODE analysis, a large scale analysis to produce global estimates of child and neonatal cause-specific mortality using verbal autopsy data from over 300 studies from 80+ high mortality countries. Beyond VA applications, we also aim to extend this framework to other scientific domains where AI-generated estimates or samples are used in downstream analyses without allowing reverse feedback. For example, uncertainty-quantified environmental exposure predictions (e.g., pollutant levels) derived from AI models are frequently used in health association studies, where accurate uncertainty propagation is equally critical but often computationally challenging to implement.

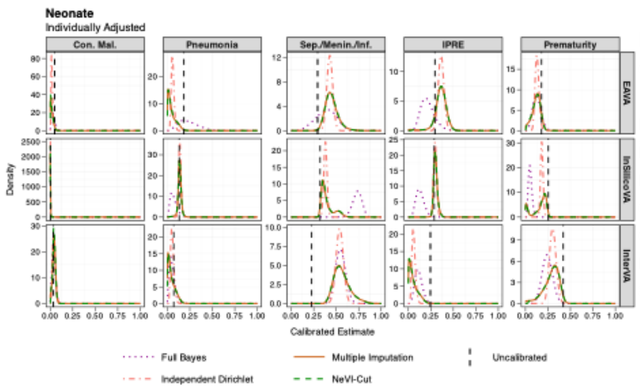


Figure 2: Posterior distributions of cause-specific mortality fractions (CSMFs) among neonates population using individually corrected confusion matrix. The NeVI-Cut approach effectively captures complex distributional features, including multimodality and skewness, achieving similar performance to multiple imputation.

ABOUT THE PI: Abhi Datta

Dr. Abhirup (Abhi) Datta is a Professor in the Department of Biostatistics at Johns Hopkins Bloomberg School of Public Health.

Dr. Datta's research focuses on developing statistical and machine learning methods for geospatial data with applications to air pollution, forestry, ecology, infectious disease modeling, and on Bayesian hierarchical models for multi-source data with applications to global health. Geospatial AI methods and Environmental Health research: Dr. Datta has been leading highly innovative and impactful research in geospatial AI methods (statistical modeling and machine learning algorithms) for analysis of large and complex environmental data. His work on Nearest Neighbor Gaussian Processes (NNGP) has become one of the most widely used methods for scalable analysis of massive geospatial data. The original NNGP paper led by Dr. Datta is one of the top-5 most cited papers published in the Journal of the American Statistical Association from 2016-2020.

ABOUT THE SYMPOSIUM PRESENTER: Jiafang Song

Jiafang Song is a fifth-year PhD student in Biostatistics at the Johns Hopkins Bloomberg School of Public Health, advised by Dr. Abhirup Datta. Her research focuses on developing Bayesian inference and machine learning methods for large and complex datasets.

She has developed cut-Bayes methods for multi-source data, designing approaches that incorporate upstream results into downstream models while properly propagating uncertainty without reverse feedback. To address challenges in posterior approximation, she developed neural network–based transformations that efficiently capture complex posterior distributions, improving both computational scalability and the accuracy of inference. This work has been applied to enhance estimates of cause-specific mortality fractions (CSMF) and to better understand the burden of disease in low- and middle-income countries.

Jiafang has also contributed computational innovations for large-scale environmental data analysis. She designed a fast variational inference algorithm based on the Nearest Neighbor Gaussian Process (NNGP), achieving significant speedups over existing methods. This framework has been released as an open-source R package, making the tools broadly accessible to the research community.

DEMYSTIFYING THE DEEP IMAGE PRIOR: AN EXPLORATORY ANALYSIS OF WEIGHT STRUCTURE

JAMES D'ALLEVA-BOCHAIN
MENTORED BY IDIES MEMBER PROFESSOR
BRICE MENARD

The Deep Image Prior (DIP), introduced in Ulyanov et al. (2017), is a deep learning based algorithm for image inverse tasks such as in-painting, denoising, and super-resolution. In contrast to most deep learning approaches to these problems, the DIP does not train on large datasets. Instead, it maps a fixed random tensor into image space with a neural network and updates the network's parameters only using the corrupted image. Despite this minimal setup, the DIP achieves competitive performance and produces outputs on the image manifold. This raises questions that confront foundational assumptions about neural networks. For example, how can a network generalize so well with so little data?

Figure 1. Example of DIP in-painting from original paper.

Following Guth et al. (2023), weights are analyzed by viewing neurons as a finite sampling from a distribution defined in the infinite-width limit. Thus, understanding the weights amounts to understanding this distribution. In this framework, the covariance of the weights reveals the network's features. Covariance eigenvectors act as the features and the eigenvalues capture their importance. This leads to notions of effective dimensionality comparisons of what is learned across networks.

The original DIP paper found an encoder-decoder convolutional network to perform best. However, analysis of the weights showed that the encoder was not essential. The output remained invariant under strong perturbations of its weights, and decoder-only networks reached equivalent performance. Moreover, the dimensionality of learned features increased multiplicatively across layers. Together, these results suggested that a decoder-only network where capacity increases with depth. This refined architecture maintains performance but is more resonant with internal feature structure and allows for a more principled analysis of the encodings.

Investigations of the covariance eigenvalue spectra elucidated properties of the encoding mechanism of the network (see Figure 2). Across layers, the spectra belong to the same family of curves. These spectra admit a characteristic scale at increasing rank deeper into the network. The low-rank power law region can be shown to correspond to the learned features, while the remainder are random features. This was reasoned by comparing the covariance eigenbasis of weights across networks and random

CONTINUED ON NEXT PAGE

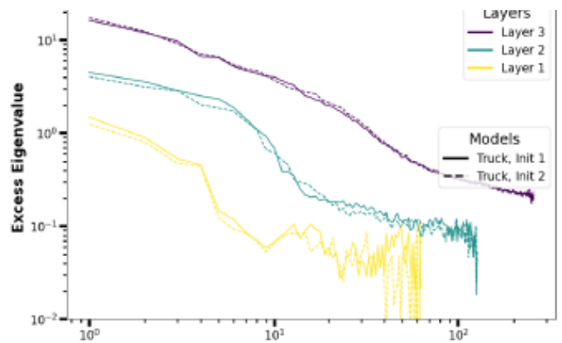


Figure 2. Excess covariance eigenvalue spectrum by layer for two networks both trained to reconstruct an image of a truck starting from two different random initializations. Excess covariance eigenvalue is just a measure of eigenvalue movement from initialization, meant to clarify what is learned versus random in the geometry of the weights. Note the characteristic scale emerging at ranks 4, 10, and 25 across the layers. These correspond to transitions between learned and random features.

DEMYSTIFYING THE DEEP IMAGE PRIOR: AN EXPLORATORY ANALYSIS OF WEIGHT STRUCTURE

CONTINUED FROM PREVIOUS PAGE

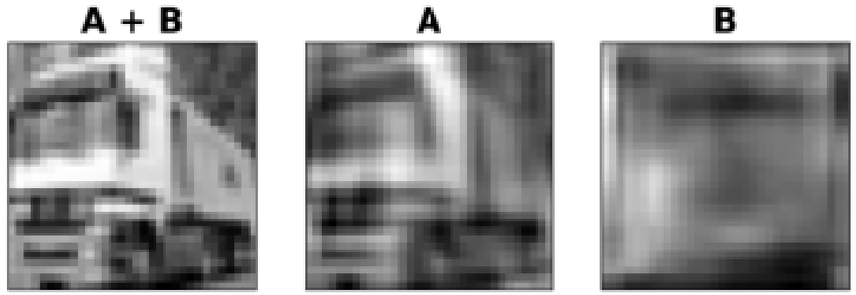
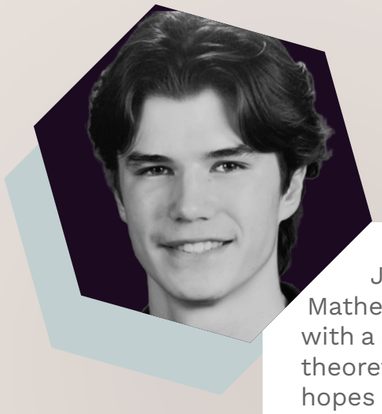


Figure 3: Truncation of DIP weights to regime A (learned) vs. regime B (random) across all layers. The truncation to regime A visually corresponds to a smoothed version of the image. Note that the truncation to regime B fits the fact that neural networks form Gaussian processes at initialization—a topic with a rich literature on its own.

initializations. Truncating the weights to only learned components preserved large scale structure of the image but lost the small scale fluctuations of the image.

This and other experiments proposed that the random features are important to the network and encode small scale information in the image. In contrast, learned features can be shown to encode localized waves in image space. This partition of weight space provides a consistent picture across layers and a preliminary model of the encoding mechanism of DIP networks.

Among other things, future work will model more precisely how random features encode small scale fluctutions in image space and determine the spatial scale at which the encoding strategy transitions from learned to random features. Answering these questions will advance a weights-based understanding of how the DIP achieves effective generalization in this minimal data regime.



James D'Alleva Bochain is a junior majoring in Mathematics and Applied Mathematics & Statistics, with a minor in Physics. He is interested in developing theoretical understandings of deep learning and hopes to pursue research in the field in graduate school and beyond. In his free time, James enjoys playing basketball and visiting art museums.

> "Through studying the weight structure of DIP networks, the goal is to demystify this algorithm and understand its implications on our fundamental understanding of neural networks.

References

[1] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep Image Prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 9446–9454, 2018. arXiv:1711.10925.

[2] F. Guth, B. Ménard, J. Rochette, and S. Mallat. A Rainbow in Deep Network Black Boxes. Proceedings of the National Academy of Sciences (PNAS), 120(31): e2302071120, 2023. arXiv:2206.05264.

QUANTIFYING ICP/ABP DYNAMICS DURING ATRIAL FIBRILLATION EPISODES

ALEX LARSON
MENTORED BY IDIES MEMBER PROFESSOR
TAMAS BUDAVARI

We examine how features of the Intracranial Pressure and Arterial Blood Pressure waveforms are affected before, during, and after AFib segments are detected on ECG. We hypothesize that AFib increases beat-to-beat and intrabeat variability and shifts morphology in both ICP and ABP.

Data and Methods

We analyzed patient ECG, ICP, and ABP data from the Neurosciences Critical Care Unit (NCCU) at Johns Hopkins Hospital. This dataset consisted of 52 patients who were designated with AFib via ICD-10 codes, and over 500 days of readings overall. AFib and non-AFib labels were determined via a pretrained resnet1d deep neural network model [1], and these labels were assigned to 10-second ECG segments. For each window, we computed ABP (mean and std ABP, min/max ABP, mean and std upstroke slope, mean and std decay rate, mean and std notch prominence, pulse amplitude, systolic/diastolic per beat values) and ICP features (mean and std ICP, min/max ICP, P2/P1 ratio mean and std, mean and std interpulse time, mean and std time-to-peak, spectral entropy) . We aligned signals to AFib→Normal transitions and Normal→AFib transitions, and summarized two regions: the 10 seconds before, and the 10 seconds after. Cohort comparisons used paired t-tests and effect sizes using Cohen's distance; a randomly selected patient's changes are shown in Figure 2.

Preliminary Results

Across current subjects (N = 5), we find no statistically significant difference at the 5% level in any of the ICP features between the 10 seconds of AFib and non-AFib at each transition. However, we do find some statistically significant differences in several of the ABP features, though the particular features vary between patients. Patient 427 Block 1 recorded a higher average mean diastolic blood pressure (p = .039) during these AFib segments (Figure 2). Patient 355 block 1 experienced several changes during AFib, including higher diastolic std (.003) and lower upstroke slope (.008). These patterns are consistent with the hypothesis that AFib perturbs ABP, although not ICP thus far, though it is important to note that these results are preliminary and sample size-limited.

Limitations and Next Steps

Limitations include small N, potential mislabels by the neural network during episodes and at boundaries, and patient-level heterogeneity. Regions that were classified as "Other" by the neural network were inconclusive on if they were AFib, so such segments were not taken into account when locating episode transitions. Future plans for this project include running the same t-tests

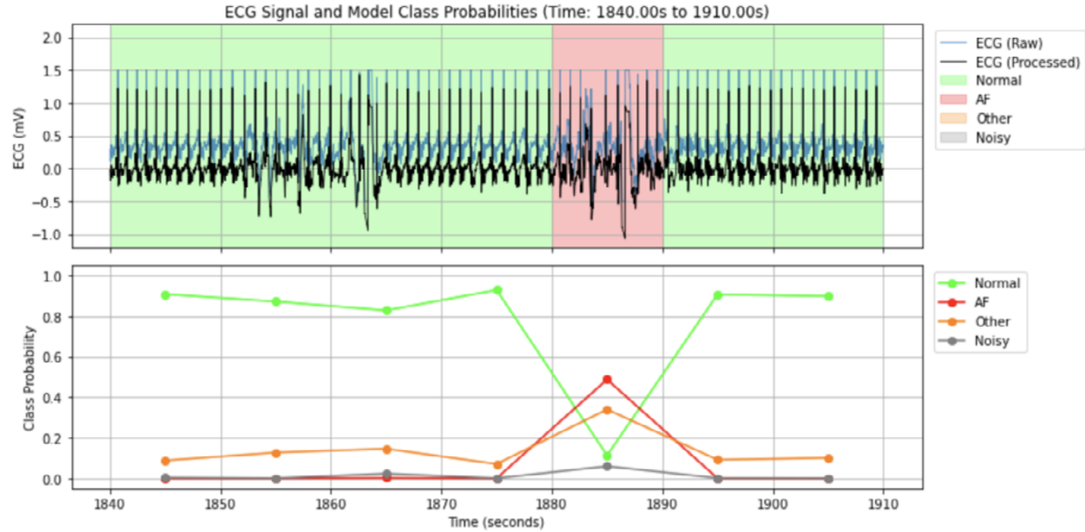


Figure 1: ECG segment plotted with predicted classes and predicted class probabilities. The upper graph depicts the raw signal with the background shaded by predicted class, and the lower graph plots the probabilities for each 10-second segment.

ENHANCING POST-OPERATIVE EFFICIENCY IN THE NEUROSCIENCES CRITICAL CARE UNIT USING MACHINE LEARNING AND DATA SCIENCE

CONTINUED FROM PREVIOUS PAGE

on the rest of the cohort and evaluating the results as a whole to see if there are consistent outcomes, and creating new plots that can further visualize the effects we are analyzing. We will also aim to add patient normalization to reduce between-patient variance. A potential offshoot of this project is using the ICP/ABP values during AFib to predict what a subject's normal/non-AFib ICP/ABP waveform should look like.

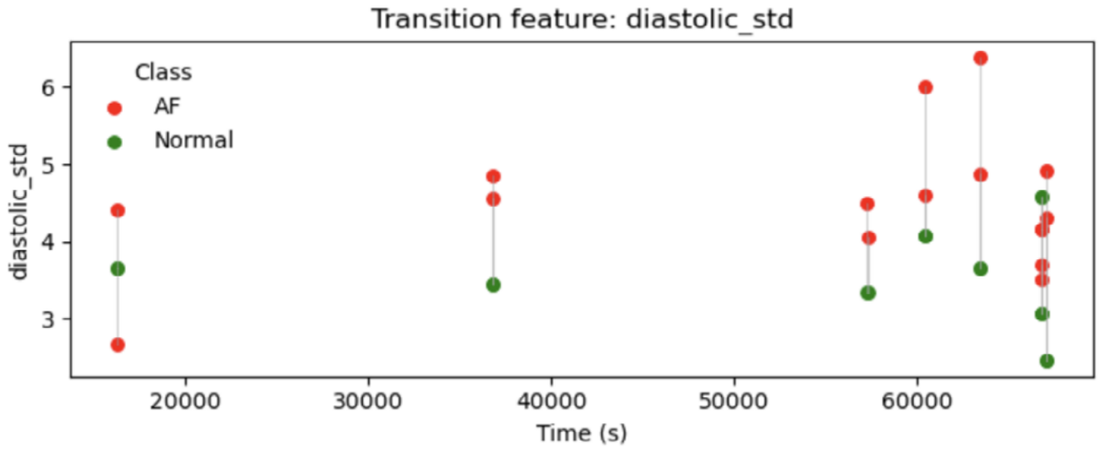


Figure 2: Paired line plots for ABP amplitude standard deviation at AF/Normal transition for a randomly selected NCCU patient (427 block 1).

Alex Larson is a senior at Johns Hopkins University majoring in Applied Mathematics & Statistics with a minor in Physics. He is from New York, also enjoys playing baseball and plans to pursue a PhD in astrophysics, where he hopes to apply data science and statistical modeling to uncover patterns in large-scale datasets.

References

1. Hong, S., Xu, Y., Khare, A., Priambada, S., Maher, K., Aljiffry, A., Sun, J., & Tumanov, A. (2020). HOLMES: Health OnLine Model Ensemble Serving for Deep Learning Models in Intensive Care Units. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20), 1614–1624.

A NON-PARAMETRIC APPROACH FOR LEARNING INTERACTION LAWS IN AGENT-BASED SYSTEMS

SRISHA MURTHY NIPPANI
MENTORED BY IDIES MEMBER PROFESSOR
MAURO MAGGIONI

Systems of interacting agents arise in various academic disciplines, where agents may represent particles, cells, animals, people, rational agents, etc... We seek to answer the following questions:

Given observations of trajectories of a system of interacting agents, how can the interaction laws be estimated from data?

My advisor, Professor Mauro Maggioni, had previously answered these questions for heterogeneous systems with pairwise interactions, in the form:

$$d\boldsymbol{x}_i(t) = \frac{1}{N}\sum_{i'=1}^{N} A_{i,i'}\phi(\boldsymbol{x}_{i'}(t)-\boldsymbol{x}_i(t))(\boldsymbol{x}_{i'}(t)-\boldsymbol{x}_i(t)) + \sigma d\mathbf{B}_i(t), i = 1,\dots,N$$

where $A \in \mathbb{R}^{N\times N}$ is the weight matrix of a network with N nodes, and at each node i there is an agent with state $\boldsymbol{x}_i \in \mathbb{R}^d$; only pairs (i, i') of agents connected by an edge (corresponding to $A_{ii'} > 0$) interact, and ϕ is the interaction kernel (IK). Maggioni used an alternating least squares algorithm to do a joint parametric estimation of the network A and IK ϕ, with A being drawn from an 'admissible' set and ϕ being drawn from a finite-dimensional function space of dimension p. The system was modified to:

$$\left[\Delta\boldsymbol{x}_i^m(t_l)\right]_{l,m} / \Delta t = A_{i,.}\left[\Phi_i^m(t_l)\right]_{l,m} c, \quad \forall i \in [N]$$

where $c \in \mathbb{R}^p$ is the coefficient vector for the IK, $\left[\Delta\boldsymbol{x}_i^m(t_l)\right]_{l,m} \in \mathbb{R}^{dLM}$ is the observed data across L time steps and M trajectories (indexed by l and m respectively), and the learning matrix $\left[\Phi_i^m(t_l)\right]_{l,m} \in \mathbb{R}^{N\times dLM\times p}$ is constructed by appropriately stacking $\Delta\boldsymbol{x}_i^m(t_l) \in \mathbb{R}^d$ and $\Phi_i^m(t_l) \in \mathbb{R}^{N\times d\times p}$ over variables l and m. The ALS operator $\mathbf{ALS}_{c,i}$ for every i estimates $A_{i,.}$ while fixing c (define $A_i, \mathbf{ALS}_{c,i} := A_{i,.}\left(\left[\Phi_i^m(t_l)\right]_{l,m} c\right)$ for each i) while $\mathbf{ALS}_A$ estimates c while fixing A (define $\mathbf{ALS}_A c := \left(A_{i,} \cdot \left[\Phi_i^m(t_l)\right]_{l,m}\right) c$). The ALS algorithm appeared statistically efficient and robust.

A goal of my research was to answer the very same questions but for heterogeneous systems with up to ternary interactions; that is, a system that allows for self, pairwise, and ternary interactions. Such a system takes the form:

$$\begin{aligned} dx_i(t) = \chi(x_i(t))dt + \frac{1}{N-1}\sum_{i'\neq i} A_{ii'}\phi(x_{i'}(t), x_i(t))dt \\ + \frac{1}{(N-1)(N-2)}\sum_{i',i''\neq i} A_{ii'i''}\psi(x_{i''}(t), x_{i'}(t), x_i(t))dt \\ + \sigma dB_i(t), \quad i = 1,\dots,N \end{aligned}$$

where $A \in \mathbb{R}^{N\times N}$ is the weight matrix (for pairwise interactions) and $\mathscr{A} \in \mathbb{R}^{N\times N\times N}$ is the weight tensor (for ternary interactions), and χ, ϕ, ψ denote the self, pairwise, and ternary IK. Given the limited literature on this topic, Maggioni advised me to generalize the techniques developed for the pairwise setup, write code to implement the model, and run simulations to determine if said technique was promising. Much of my time went toward coding and running simulations.

CONTINUED ON NEXT PAGE

A NON-PARAMETRIC APPROACH FOR LEARNING INTERACTION LAWS IN AGENT-BASED SYSTEMS

CONTINUED FROM PREVIOUS PAGE

To accommodate self interactions, we added the unary term to the system:

$$[\Delta \boldsymbol{x}_i^m(t_l)]_{l,m} / \Delta t = [\mathbf{X}_i^m(t_l)]_{l,m} b + A_{i,\cdot} [\Phi_i^m(t_l)]_{l,m} c, \forall i \in [N].$$

Here, we have a different hypothesis space of dimension s for the self IK, with $b \in \mathbb{R}^s$ as the coefficient vector. Now, the ALS algorithm must be adjusted to (1) estimate both interaction kernels b, c given the weight matrix A and (2) estimate A given both interaction kernels b, c . The second task is straightforward; given b, c one has to solve

$$[\Delta \boldsymbol{x}_i^m(t_l)]_{l,m} / \Delta t - [\mathbf{X}_i^m(t_l)]_{l,m} b = A_{i,\cdot} \left(\mathbf{ALS}_{c,i}\right)$$

for each row of A, where the operator $\mathbf{ALS}_{c,i} := [\Phi_i^m(t_l)]_{l,m} c \in \mathbb{R}^{N \times dLM}$. However, the first task will require what we shall henceforth call the "stacking" technique. Given A, define the operator $\mathbf{ALS}_A := \left[[\mathbf{X}^m(t_l)]_{l,m}, \left(A_{i,} [\Phi_i^m(t_l)]_{l,m}\right)_i\right] \in \mathbb{R}^{N \times dLM \times (s+p)}$ by concatenating the two hypermatrices. Concatenate the coefficient vectors b and c. Then, solve

$$[\Delta \boldsymbol{x}_i^m(t_l)]_{l,m} / \Delta t = \left(\mathbf{ALS}_A\right) \begin{bmatrix} b \\ c \end{bmatrix}$$

for b and c simultaneously. Based on our simulations, the stacking technique seemed to work remarkably well; it appeared to be just as accurate. For instance, if there were no self interactions, the algorithm would correctly estimate $b \simeq 0$. We currently do not have any theoretical justification as to why this is the case. Typically, when one has more parameters, one would expect the estimation to be worse.

Because of this, we decided to apply the stacking technique to the ternary setup. Now, we have another hypothesis space of dimension q for the ternary IK, with $\xi \in \mathbb{R}^q$ as the coefficient vector. To estimate the IKs, define the operator:

Sri Nippani is a junior at Johns Hopkins University majoring in Mathematics with minors in Economics and Computer Science. He aims to pursue graduate studies in Economics and is enthusiastic about contributing to interdisciplinary research that leverages quantitative and computational methods.

$$\mathbf{ALS}_{A,\mathscr{A}} := \left[[\mathbf{X}^m(t_l)]_{l,m}, \left(A_{i,} \cdot [\Phi_i^m(t_l)]_{l,m}\right)_i, \left(\mathrm{Vec}\left(\mathscr{A}_{i,\cdot,\cdot}\right)^\top [\Psi_i^m(t_l)]_{l,m}\right)_i\right] \in \mathbb{R}^{N \times dLM \times (s+p+q)}$$

by (horizontally) concatenating the three hypermatrices. Concatenate the vectors b, c, ξ . Then, solve

$$[\Delta \boldsymbol{x}_i^m(t_l)]_{l,m} / \Delta t = \left(\mathbf{ALS}_{A,\mathscr{A}}\right) \begin{bmatrix} b \\ c \\ \xi \end{bmatrix}$$

for b, c, and ξ simultaneously. To estimate the networks, define, $\forall i \in [N]$, the operator

$$\mathbf{ALS}_{c,\xi,i} := \left[[\Phi_i^m(t_l)]_{l,m} c; [\Psi_i^m(t_l)]_{l,m} \xi\right] \in \mathbb{R}^{(N+N^2) \times dLM}$$

by (vertically) concatenating the two matrices. Concatenate the vectors $A_{i,\cdot}$ and $\mathrm{Vec}\left(\mathscr{A}_{i,\cdot,\cdot}\right)^\top$. Then, solve

$$[\Delta \boldsymbol{x}_i^m(t_l)]_{l,m} / \Delta t - [\mathbf{X}_i^m(t_l)]_{l,m} b = \left[A_{i,\cdot}, \mathrm{Vec}\left(\mathscr{A}_{i,\cdot,\cdot}\right)^\top\right] \left(\mathbf{ALS}_{c,\xi,i}\right)$$

for $A_{i,}$ and $\mathrm{Vec}\left(\mathscr{A}_{i,\cdot,\cdot}\right)^\top$ simultaneously. We can easily recover A and $\mathscr{A}$ after the estimation.

So far, I have been writing code to implement this model and run simulations. Professor Maggioni and I plan to run further simulations to decide if this estimation technique seems promising. Only then, can we look into providing theoretical guarantees for the estimator. Moreover, one should note that this stacking method can be generalized to higher-order interactions. These are the next steps for the project.

AI-DRIVEN ECHOCARDIOGRAPHIC MODEL FOR EARLY IDENTIFICATION OF STRESS CARDIOMYOPATHY

JOOYOUNG RYU
MENTORED BY IDIES MEMBER PROFESSOR
ROBERT D. STEVENS

My funded proposal, "AI-driven Echocardiographic Model for Early Identification of Stress Cardiomyopathy," set out to refine diagnostic labels by tracking recovery of left ventricular function across sequential echocardiograms. Over the past several months, we conducted extensive analysis to evaluate the feasibility of this approach, including systematic searches for longitudinal cases within the MIMIC-IV dataset and exploratory pipelines for automated recovery tracking of heart function. Ultimately, we determined that the strategy was not viable due to the very small number of patients in MIMIC-IV with three or more follow-up echocardiograms.

Rather than abandon the central goal of improving diagnosis of Stress Cardiomyopathy (SCM), we shifted to a more feasible yet still novel direction: a classification framework to differentiate SCM from myocardial infarction (MI) using multimodal data. This preserves the original clinical motivation of the project while ensuring the research remains novel and impactful.

Background and Prior Work

Two recent studies have applied deep learning to echocardiography for SCM vs. MI classification.[1] [2]

These were restricted to standard apical views (A2C, A4C), excluded electronic health record (EHR features, and evaluated perfomance only under artificially balanced class ratios. Reported AUROC reached ~0.79 record (EHR features, and evaluated perfomance only under artificially balanced class ratios. Reported AUROC reached ~0.79 with sensitivity and specificity ≥0.70, but these findings were not tested in real-world imbalanced settings.

Methods

In contrast, this project leverages:

- Comprehensive imaging: all available echocardiographic views (not just A2C/A4C)

- Clinical data integration: 45 structured EHR features capturing demographics, comorbidities, and lab values.

- Multimodal modeling: 1024-dimensional embeddings from the EchoFM[4] foundation model, combined with XGBoost[5] classifiers under three settings—echo only, EHR only, and echo+EHR

- Robust evaluation: training on the MIMIC-IV[3] cohort (561 MI, 33 SCM patients) with exclusive external validation planned on a Johns Hopkins dataset. To address diagnostic overlap, patients with dual SCM/MI codes were conservatively retained in the SCM group

Results

The multimodal approach substantially outperformed unimodal baselines:

- Echo-only model: AUROC 0.71, AUPRC 0.14

- Multimodal (echo + EHR): AUROC 0.85, AUPRC 0.29

	SCM	IM
Patients	33	561
Echo studies	63	1,230
Dicom files (including all echo views)	2,435	51,495

Figure 1: Cohort Selection in MIMIC-IV Dataset

These results are particularly meaningful given the baseline AUPRC of only 0.05, reflecting the severe class imbalance (1:20). Sensitivity reached 0.80, though specificity remained limited at 0.55. Cross-validation revealed wide variance (AUROC 0.3–0.7), a consequence of <15 SCM cases per test fold.

CONTINUED ON NEXT PAGE

AI-DRIVEN ECHOCARDIOGRAPHIC MODEL FOR EARLY IDENTIFICATION OF STRESS CARDIOMYOPATHY

CONTINUED FROM PREVIOUS PAGE

Significance and Next Steps

Although constrained by small sample size, this work demonstrates the feasibility and potential of multimodal approaches for SCM vs. MI classification, addressing limitations of prior work by incorporating all echo views and structured EHR features.

Next steps include:

- External validation on the independent Johns Hopkins cohort

- Developing strategies to mitigate class imbalance

- Exploring interpretability methods on imaging and clinical features

Citations

1. Zaman, F., Ponnapureddy, R., Wang, Y. G., Chang, A., Cadaret, L. M., Abdelhamid, A., Roy, S. D., Makan, M., Zhou, R., Jayanna, M. B., Gnall, E., Dai, X., Singh, A., Zheng, J., Boppana, V. S., Wang, F., Singh, P., Wu, X., & Liu, K. (2021). Spatio-temporal hybrid neural networks reduce erroneous human “judgement calls” in the diagnosis of takotsubo syndrome. EClinicalMedicine, 40, 101115. https://doi.org/10.1016/j.eclinm.2021.101115

2. Laumer, F., Di Vece, D., Cammann, V. L., Würdinger, M., Petkova, V., Schönberger, M., Schönberger, A., Mercier, J. C., Niederseer, D., Seifert, B., Schwyzer, M., Burkholz, R., CorinziaL., Becker, A. S., Scherff, F., Brouwers, S., Pazhenkottil, A. P., Dougoud, S., Messerli, M., Templin, C. (2022). Assessment of Artificial Intelligence in echocardiography diagnostics in differentiating takotsubo syndrome from myocardial infarction. JAMA Cardiology, 7(5), 494. https://doi.org/10.1001/jamacardio.2022.0183

3. Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). Mimic-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1).https://doi.org/10.1038/s41597-022-01899-x

Jooyoung Ryu is a fourth-year computer science student originally from Seoul, South Korea. He is passionate about AI- and machine learning-driven precision medicine and developing computational tools that improve patient outcomes. With aspirations to attend medical school after graduation, Jooyoung hopes to combine his background in computer science with clinical training to advance healthcare innovation.

REFERENCES CONTINUED

4. Kim, S., Jin, P., Song, S., Chen, C., Li, Y., Ren, H., Li, X., Liu, T., & Li, Q. (2025). ECHOFM: Foundation model for Generalizable Echocardiogram Analysis. IEEE Transactions on Medical Imaging, 1–1. https://doi.org/10.1109/tmi.2025.3580713

5. Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

NEXT ISSUE:
NOVEMBER 2026

EMPOWERING RESEARCHERS
TO MAKE BIG BREAKTHROUGHS
THROUGH BIG DATA