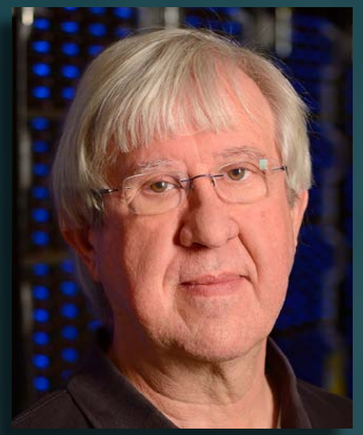
2023 Annual Review

OCTOBER 1, 2022
– SEPTEMBER 30, 2023



MESSAGE FROM
IDIES Director Alex Szalay

Welcome New IDIES Members

The following members have joined since **October 1, 2022– and before September 30, 2023**:

James Schmidt —Faculty
Ethan Vishniac—Faculty
Thi Vo—Faculty
Ali Eshragh—Faculty
James Schmidt—Faculty
Mitra Taheri—Faculty
Christopher Carroll—Faculty
Kenneth Pienta—Faculty
George Butler—Affiliate
Alan Lujan—Affiliate
Shuolin Xiao—Affiliate

ON THE COVER: *The newest dataset offered through IDIES's SciServer science platform: the history of Major League Baseball. The cover shows a visualization of a baseball field, with dots showing the typical positions of the fielding team. The field is divided into sectors, with the goal of recording where the ball ends up. SciServer now includes an example notebook to make such a plot; see page 16 for information.* **CREDIT:** *Jordan Raddick*

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 decade, 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.

With the recent announcement of a huge data science and AI initiative at Johns Hopkins, building on the momentum of the AI-X Foundry, our efforts are becoming even more relevant. In recognition of these events we have started a gradual migration of IDIES to become integral part of AI-X and the bigger initiative.

Recently, already under the AI-X umbrella, we have been awarded a 5 year, $12.5M grant from the Schmidt Futures Foundation, for the JHU SSEC, Scientific Software Engineering Center, aimed at helping project to develop software applications. Our Center is one of 4 awarded over the world. The Center is managed by our new Technical Director, Edward Hunter, and Associate Director Gerard Lemson.

Today Artificial Intelligence and Machine Learning is everywhere. To be successful in this area... ...several components are required. We need deep expertise in modern AI techniques, we have to have computational capabilities, we need translational capabilities, i.e. personnel to help research projects to get started with the right network design, and last but not least we have to make it easy to turn the research data "AI-ready". Today's experiences show that we spend a disproportionate amount of time to convert our data sets into a format that the popular AI frameworks can work with. Large scale studies need to aggregate data from many different sources, and transform them into a common system.

Making decisions based on facts, relying on sound data foundations is more important than ever. To create and grow diverse datasets, we have active partnerships with six divisions and several Institutes and large projects within JHU. Quantitative social science, using advanced computer science and statistics in tackling large scale social problems is at the heart of modern sociology.

We are also looked beyond JHU to establish new partnerships with outside organizations. We have many active collaborations all over the world. We started a new project with the National Institute of Standards and Technology (NIST) to build a SciServer-based data aggregator for data from Puerto Rico about the effects of hurricane Maria. This collaboration is now also expanding into Materials Science. Our SciServer is also running in several places over the world, like Brookhaven National Laboratory, as an AI workbench; providing the data system for a European X-ray satellite, eRosita.

IDIES, combined with the SSEC and IA-X 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.

2023 IDIES
Annual Review
TABLE of CONTENTS

2023 IDIES
Annual Symposium Schedule

Friday, October 13th, 2023
9:00 am–5:00 pm
Glass Pavilion, Homewood Campus

TIME	SESSION	SPEAKER	LENGTH
9:15 AM	Opening Remarks	Alex Szalay	15
9:30 AM	Keynote Speaker (1 of 3)	Alexis Battle	30
10:00 AM	SSEC Update	Edward Hunter	20
10:20 AM	Sheridan/Data Services & LoveDataWeek Update	Peter Lawson	15
10:35 AM	Break		40
11:15 AM	Seed Awardee Update (1 of 4)	Ed Chen* presented by Joseph Angelo	15
11:30 AM	ARCH Update	Jaime Combariza	15
11:45 AM	Summer Student Fellow Update (1 of 3)	Sampath Rapuri	10
11:[illegible]AM?M	Seed Awardee Update (2 of 4)	Caleb Alexader* presented by Venus Van Ness	15
12:10 PM	Summer Student Fellow Update (2 of 3)	Jay Kim	10
12:20 PM	Poster Madness	Two minutes per poster	20
12:50 PM	Lunch + Poster Gallery		
2:00 PM	SciServer Update + Poster Winner Announcement	Gerard Lemson	20
2:20 PM	Seed Awardee Update (3 of 4)	Corey Oses* presented by Keith Clark	15
2:35 PM	Keynote Speaker (2 of 3)	Thomas Hartung	30
3:05 PM	Summer Student Fellow Update (3 of 3)	William Shiber	10
3:15 PM	Break		30
3:45 PM	Seed Awardee Update (4 of 4)	Angelo Mele* presenting with Co-Pierre Georg	15
4:00 PM	Keynote Speaker (3 of 3)	Julia Lane	30
4:30 PM	Closing Remarks		15
4:45 PM	Reception		

2023 IDIES
Annual Symposium Keynote Speakers

Alexis Battle, PhD

**Associate Professor, Biomedical Engineering;
Associate Professor, Computer Science**

SPEAKING ON THE TOPIC OF:

Genomics and machine learning

Alexis Battle is a 2016 Searle Scholar and an Associate Professor of Biomedical Engineering and Computer Science. She specializes in unlocking secrets of the human genome by analyzing large-scale genomic sequencing data to understand the impact of genetic variation on the human body.

Battle's research is concentrated on the development of computational biology tools and machine-learning strategies to examine genetic differences on gene regulation and disease. A leading member of the National Institutes of Health's (NIH) Genotype-Tissue Expression (GTEx) Consortium, she focuses on predicting the effects of variation in noncoding DNA sequences. Findings of her work on a GTEx project, which studied how genetic patterns lead to molecular changes within specific tissues, were published in 2017 in the journal Nature.

At JHU's "Battle Lab" her research also includes development of new methods to evaluate and predict the impact of personal genomics, and rare genetic variants that may significantly impact an individual's health. These methods are helping to improve understanding of how genes work together and how their interconnected pathways may influence complex traits. Additionally, Battle is involved in a host of ongoing research initiatives, from building integrative networks for genomic analysis of autism, supported by NIH, to predicting rare Mendelian disease variants using genomic data, funded by her Searle award and 2017 JHU Catalyst Award. She is also the recipient of a 2019 Johns Hopkins Discovery Award for studying the genetics of atherosclerotic cardiovascular disease.

Battle received her formal education from Stanford University, where she earned a BS in 2003, an MS in 2013, and a PhD in computer science in 2013. She worked as a postdoctoral researcher at Stanford before joining Hopkins in 2014. Prior to her career in academia, Battle was a staff software engineer and manager for Google.

Dr. Thomas Hartung, MD, PhD

**Director of the Johns Hopkins Center for Alternatives to Animal Testing, Doerenkamp-Zbinden Chair,
Professor of Environmental Health and Engineering,
Molecular Microbiology and Immunology**

SPEAKING ON THE TOPIC OF:

Organoid cultures and the use of artificial intelligence

Dr. Thomas Hartung is the director of the Johns Hopkins Center for Alternatives to Animal Testing, which was founded in 1981 as part of the Johns Hopkins University Bloomberg School of Public Health, with a European branch (CAAT-Europe) located at the University of Kostanz, Germany.

Dr. Hartung steers the revolution in toxicology to move away from 50+ year-old animal tests to organoid cultures and the use of artificial intelligence. while also endeavoring toward a paradigm shift in toxicity testing to improve public health.

His background as head of the European Center for the Validation of Alternative Methods of the European Commission (2002-2008) has led to his involvement in the implementation of the 2007 NRC vision document "Toxicity Testing in the 21st Century – a vision and a strategy" as he endeavors toward a paradigm shift in toxicity testing to improve public health.

Dr. Hartung has furthered the translation of concepts of evidence-based medicine to toxicology (evidence-based toxicology) with the goal of systematic assessment of the quality of all tools for regulatory toxicology and the development of new approaches based on annotated pathways of toxicity (the Human Toxome).

More than 550 publications serve to document Dr. Hartung's broad background in clinical and experimental pharmacology and toxicology. His previous work centered on the immune recognition of bacteria, including pyrogen testing, and the induced inflammatory response. In experimental and clinical approaches, he studied the pharmacological modulation of these responses.

2023 IDIES

Annual Symposium Keynote Speakers

Julia Lane, PhD

Provostial Fellow for Innovation Analytics, Professor
NYU Wagner Graduate School of Public Service

SPEAKING ON THE TOPIC OF:

Democratizing Democracy

Julia is a Professor at the NYU Wagner Graduate School of Public Service. She was a senior advisor in the Office of the Federal CIO at the White House, supporting the implementation of the Federal Data Strategy. She recently served on the Advisory Committee on Data for Evidence Building and the National AI Research Resources Task Force

Julia is an elected fellow of the American Association for the Advancement of Science, the International Statistical Institute and the American Statistical Association. She is the recipient of the 2014 Julius Shiskin award and the 2014 Roger Herriot award. She is also the recipient of the 2017 Warren E. Miller Award and the 2019 Distinguished Fellow award from the New Zealand Association of Economists. She holds a PhD in Economics and an MA in Statistics.

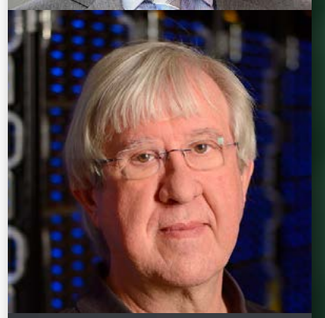
Moderators & Updates Speakers

BY ORDER OF APPEARANCE

Moderator

Tinglong Dai, PhD, is a professor of Operations Management & Business Analytics at the Carey Business School. He serves as a member of the Leadership Team of the Hopkins Business of Health Initiative (HBHI)—where he co-chairs the Hopkins Workgroup on AI and Healthcare—and on the IDIES executive committee. Tinglong holds joint faculty appointments at the Hopkins School of Nursing and Center for Digital Health and Artificial Intelligence (CDHAI). In 2021 he was named one of the world's best 40 business school professors under 40 by Poets & Quants.

IDIES



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 works on the use of big data in advancing scientists understanding of astronomy, physical sciences, and life sciences.

SSEC

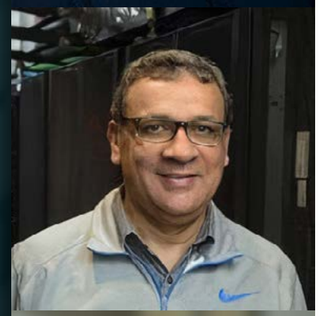
Edward Hunter holds a PhD in Electrical and Computer Engineering from UC San Diego and is the Director of Engineering at the Schmidt-funded Scientific Software Engineering Center at Johns Hopkins.

2023 IDIES
Annual Symposium

Moderators & Updates Speakers
BY ORDER OF APPEARANCE

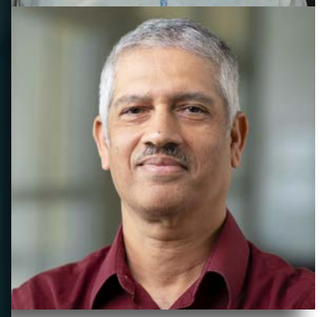
Data Services, Sheridan Libraries & Museums

Peter Lawson, PhD, the Data and Visualization Librarian at Sheridan Libraries' Data Service and the newest member of the IDIES executive board. He holds a PhD in Bioinnovation from Tulane University. As a computational scientist and librarian, Lawson leverages his research background to help students and researchers analyze and visualize their data.



Advanced Research Computing at Hopkins & Maryland Advanced Research Computing Center (ARCH & MARCC)

Jaime Combariza, PhD, is the director of the Maryland Advanced Research Computing Center (MARCC). a shared high-performance computing facility for John Hopkins University and the University of Maryland, as well as the director of the Advanced Research Computing @ Hopkins HPC group.



Moderator

Ani Thakar, PhD is a principal research scientist in the Department of Physics & Astronomy (PHA) and at the Center for Astrophysical Sciences (CAS). Thakar serves as the associate director of operations for IDIES.

SciServer

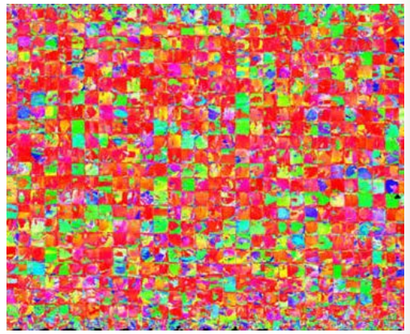
Gerard Lemson holds a PhD in theoretical cosmology and is currently a research scientist at IDIES and the associate director of the new SSEC. He is also associate director for science coordination in the NSF -funded SciServer project (www.sciserver.org) and assists in code development of the platform.

Moderator

David Elbert, PhD is the Director of the Environmental Science and Policy MS Associate Program in Advanced Academic Programs at JHU. A Hopkins alumnus himself, David is an associate research scientist in Earth & Planetary Sciences and long-time IDIES member.

IMAGE CAPTION (right) : Local crystallographic orientations within a single leg of an as-built AM Bench 2022 test artifact CREDIT: NIST

IMAGE CAPTION (far right): SMUDGE demo notebook CREDIT: : Carrie FIllion



SciServer:
New Datasets Available

AM Bench

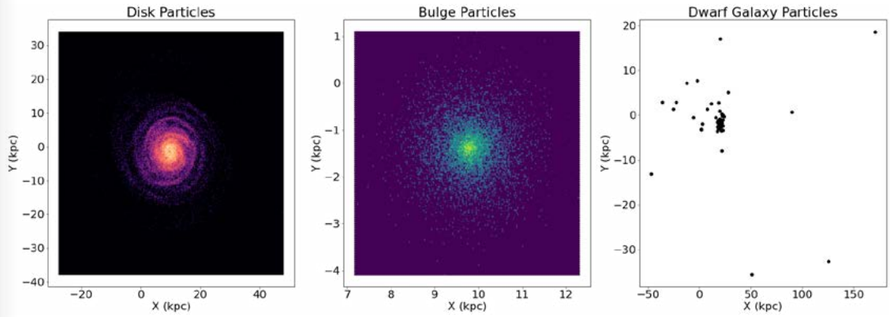
By Chandler Becker (NIST)

AM Bench is a NIST-led organization that provides a continuing series of AM benchmark measurements, challenge problems, and conferences with the primary goal of enabling modelers to test their simulations against rigorous, highly controlled additive manufacturing benchmark measurement data. All AM Bench data are permanently archived for public use using comprehensive, custom data management systems.

The first round of benchmark measurements, challenge problems, and conference (AM Bench 2018) were completed in 2018 and the second round was completed in 2022, following a one-year delay caused by the COVID-19 pandemic. AM Bench follows a nominal 3-year cycle with the next set of measurements, challenge problems, and conference scheduled for completion in 2025. Additional asynchronous benchmarks are also being planned.

The AM Bench 2022 Measurements and Challenge Problems include both metals AM and polymers AM. For metals, 3 sets of benchmarks focused on laser powder bed fusion (LPBF) of the nickel-based superalloy 718, and two sets of benchmarks included follow-on mechanical performance and microstructure measurements for the 2018 LPBF studies using the nickel-based superalloy 625. Taken together, these benchmarks cover the full processing-structure-properties range, including feedstock characterization, in situ measurements during the build process, heat treatments, 2D and 3D microstructure characterization, measurements of residual strains and part deflection, and mechanical behavior measurements.

For polymers, one set of benchmarks focused on material extrusion (MatEx) of polycarbonate test objects, and another set of benchmarks focused on vat photopolymerization measurements. Finally, an asynchronous set of benchmark measurements was completed in early 2022 and focused on melt pool and laser coupling dynamics during high-power laser interactions with Ti-alloy and Al-alloy bare metal surfaces. Links to descriptions of the 2022 benchmark measurements, challenge problems, and results are provided through the NIST AM Bench website.



SMUDGE

By Carrie FIllion

The Satellite Mergers Usher Disc Galaxy Evolution (SMUDGE) simulations are a suite of four high-resolution, dynamical N-body simulations. This suite includes two different models of isolated, disk galaxies and two versions in which these same disk galaxies experience a minor merger with a satellite galaxy. These simulations were designed to be laboratories for the study of galactic dynamics rather than tailored models of the Milky Way, and the four available models enable the study of both the dynamics in isolated disks and the dynamics in merging, disequilibrium systems.

Each host disk galaxy consists of a dark matter halo (~8×108 particles), a stellar bulge (2×107 particles), and a stellar disk (2×108 particles), with parameters for the initial conditions taken from Widrow & Dubinski (2005) (WD05). The satellite is the L2 model from Laporte et al. (2018). The initial conditions for the host were generated using Galactics (Kuijken & Dubinski 1995), and the simulations were run with the GPU accelerated N-body tree code Bonsai (Bédorf et al. 2012).

This data volume contains four simulations described below. See Hunt et al. (2021) for descriptions.

D1: A highly stable disc galaxy (MWb from WD05) included for comparison with the merger model M1. Model D1 has lower time resolution (100 Myr) snapshots as it undergoes little evolution over 5 Gyr.

D2: A more Milky Way-like isolated disc Galaxy (MWa from WD05) which forms a bar and spiral structure. Included for comparison with model M2.

M1: Relatively calm merger of the satellite into host model D1. Model M1 is not tailored specifically to the Milky Way-Sagittarius interaction, but is instead intended to be a laboratory for examining merger dynamics and host galaxy response at high resolution in an otherwise stable disc galaxy.

M2: Violent merger of the same satellite into model D2. The disc is significantly disrupted, and experiences a large degree of heating.

The provided example Jupyter Notebook (stored in SMUDGE/data01_01 and available in the SMUDGE GitHub repository) shows how to access both the individual timestep snapshots of a simulation and trajectories (orbits) for individual 'star' particles over a given time period.

Open SciServer Grant

By Gerard Lemson

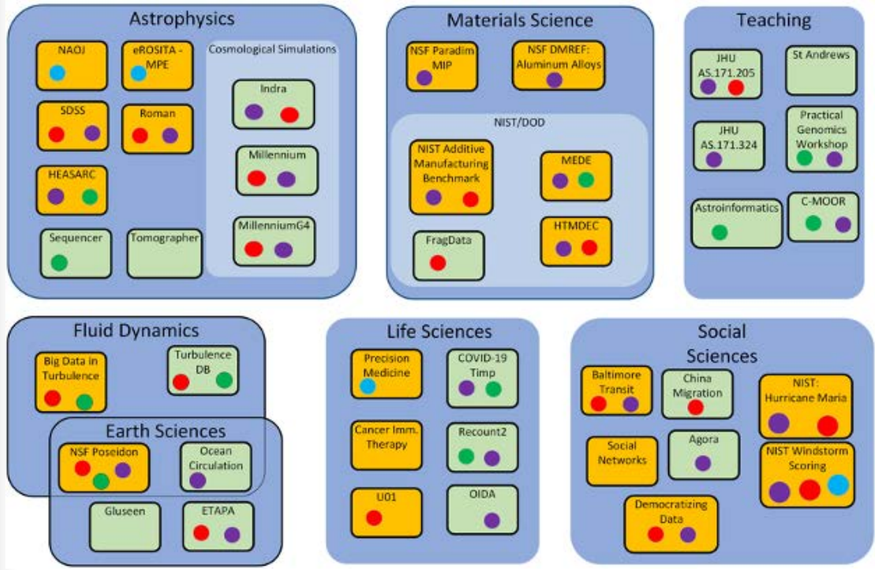
The SciServer team received an NSF The SciServer team received an NSF grant in the CSSI program. Starting Oct 2023, the Open SciServer project will implement a transition to sustainability for the SciServer collaborative data-driven science platform. SciServer is a high-impact, highly successful DIBB (data infrastructure building block) with a well-developed existing code base; an established user community; and demonstrated impact on scientific discovery, research, and education. See the figure for a schematic view of various disciplines and projects supported by SciServer. SciServer has grown from a platform for transformational impact in astronomy to one that creates significant impact in multiple science domains including - but not limited to - materials science and engineering; turbulence; oceanography; and precision medicine and genomics. To facilitate the transition to sustainability, we will work with the NSF Science Gateways Community Institute and deepen our interaction with the community.

The transition plan will migrate the SciServer code base from closed to open source enabling expansion of the user community to provide continued evolution of the platform as technologies and science drivers grow; this expansion will leverage community input to refine existing architecture to facilitate sustainable development and flexibility. The second major thrust of the transition plan is the creation of simple, rapid cloud deployment and management options to suit the broad range of current and future SciServer users. The central goal of this thrust is to facilitate democratization of access to the power of compute for Big Data research. Pre-configured options will provide tiers spanning cost, expected duration of projects, and computing power. Data store and user management will be opened to work with existing community resources. The SciServer transition to sustainability will also generalize its educational tools to allow all users to integrate SciServer with curricular efforts to accelerate translation of education to research and workforce development in data-intensive science domains.

SciServer's focus on transparent, reproducible data-intensive research aligns with science needs and provides infrastructure to facilitate applications of FAIR data. During the two years of this transition plan, the project will particularly focus on instantiations and priorities resonant in many fields, but particularly clearly articulated in the Materials Genome Initiative strategic plan. This focus will leverage SciServer experience in existing projects (e.g. NSF PARADIM MIP, two NSF DMREF, the Army Research Lab HTMDEC project) to develop an open and interoperable infrastructure with multiple options for sustainability from many fields and funding sources.

The proposed research will provide a free and openly available science platform cyberinfrastructure system for data-driven science that can be used, installed, managed, and extended by research projects from community colleges to R1 institutions. The project will extend best practices for transitioning...major science infrastructure to sustainability. The project will identify new research problems to address interoperability of major data infrastructure. SciServer is a resource that thousands of science users and several science communities and projects depend upon worldwide, so ensuring its sustainability is crucial.

The impact of this project will be far reaching and extend to multiple science communities. It will provide training on Open Source principles and practices for a diverse group of developers and students working to extend an infrastructure for their research needs. It takes a highly successful cyberinfrastructure that continues to fi ll a critical, community identified gap and creates a sustainable path to meet current and future needs. The creation of generalized components and a sustainable model will provide a basis for a range of important cyberinfrastructure projects that must survive past the development to meet the needs of science research in the US. Creation of tiered deployments will provide a model for a data-intensive science ecosystem that can be leveraged to meet priority goals of data democratization and a widely available platform to realize the promise of FAIR data and Open Science principles. SciServer already has an excellent track record of broader impact within the educational community, including underserved institutions.



Take Me Out to Big Data:
The History of Major League Baseball Comes to SciServer

By Jordan Raddick

DATA USAGE MESSAGE: Recipients of Retrosheet data are free to make any desired use of the information, including (but not limited to) selling it, giving it away, or producing a commercial product based upon the data. Retrosheet has one requirement for any such transfer of data or product development, which is that the following statement must appear prominently:

The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at www.retrosheet.org.

Retrosheet makes no guarantees of accuracy for the information that is supplied. Much effort is expended to make our website as correct as possible, but Retrosheet shall not be held responsible for any consequences arising from the use of the material presented here. All information is subject to corrections as additional data are received. We are grateful to anyone who discovers discrepancies and we appreciate learning of the details.

Baseball is America's Pastime, and the favorite sport of data scientists. Fans obsessively track what happens in every game, and use those records to calculate statistics, from simple metrics like batting average to complex parameters like OPS+, VORP, WAR, and many other "sabermetrics" that they use to evaluate players and teams - and professionals do the same. These efforts mean that baseball has brought the techniques and habits of mind of data scientists into mainstream culture like few other activities ever have.

The mainstream appeal of baseball statistics means that baseball data can provide an intuitive and engaging introduction to data science for students. Many baseball datasets exist on the Internet for use in learning activities, but most contain summary statistics only, making it difficult to learn the most important skill of a data scientist: proposing and answering research questions.

To address this need, IDIES now offers a new dataset in SciServer's Getting Started public data repository, called baseball. The dataset is available in SciServer Compute in the directory getting_started/bigdata/baseball, and is documented on the Datasets page of www.sciserver.org.

This dataset comes from the Major League Baseball events dataset compiled over more than 20 years by online volunteers at www.retrosheet.org, containing complete records of all Major League Baseball games since 1974, with events for some games as far back as 1915. "Events" means everything that happens in a game: pitches, strikes, hits, fielding events, unusual occurrences, and many other pieces of information.

Crucially, the SciServer baseball dataset builds on the Retrosheet data by adding context. Not only does each record show who was pitching and batting, but also the game information at that moment: the score, the number of outs, and any runners on base. This opens up a broad range of questions that can be asked - for example, investigating which hitters perform the best with two outs and runners on base. Full events data is available in the seasons directory, one CSV file per season.

What can learners do with this new dataset? The Getting Started data volume includes a working example, in the /getting_started/Example-Notebooks/baseball/ directory. The example looks at where balls in play ended up in the 2022 season - when a hitter hits into an out, into which section of the field did the ball travel?

Results are shown in Figure 1 for all fielded outs in the 2022 Major League Baseball season (n = 82,804). The results are not a surprise - hitters hit across their bodies into the infield, resulting most often in balls fielded by the shortstop for right-handed hitters, and between first and second base for left-handed hitters. This simple example offers a powerful glimpse into the types of questions that can be answered with our new dataset.

We hope you find this new dataset and example helpful in getting started with SciServer. Please let us know what you do with it!

Racial Disparities in Business Lending in Baltimore City

By Jordan Raddick (IDIES) & Mac McComas (JHU 21st Century Cities Initiative)

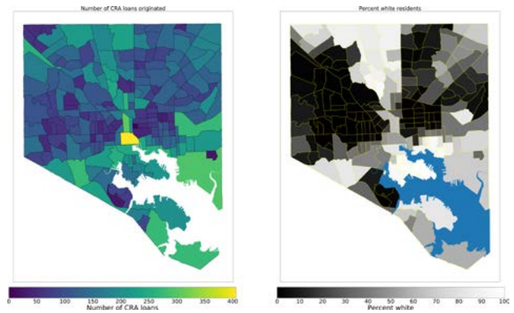


Figure 1. Maps of census tracts in Baltimore City. The left map shows the total number of loans made under the Community Reinvestment Act (2012-2017); the right map shows the percent of white residents.

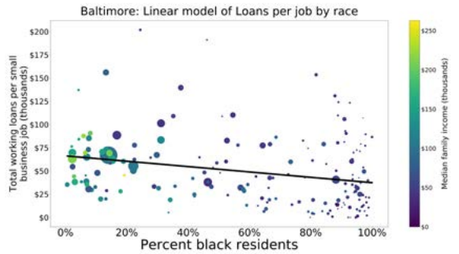


Figure 2. Scatterplot of Baltimore City census tracts, showing percentage of black residents (x-axis) and total 2012-2017 CRA loans (y-axis). Dots are scaled by total tract population and colored by median family income (color scale on the right).

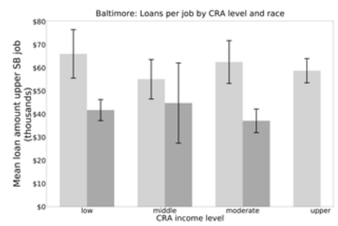


Figure 3. Mean CRA loan amount disaggregated by income level and race for Baltimore City census tracts (after Theodos et al. 2019). Error bars show standard errors.

Nearly seventy years after segregation officially ended in the United States, the legacy of racial disparity is sadly alive and well in Baltimore. Our city is divided into clear majority-white and majority-black neighborhoods, with major differences in poverty and opportunity.

One of the most effective ways to build wealth in neighborhoods is to develop locally owned small businesses like restaurants, hardware stores, and doctors' offices. Starting such businesses requires access to capital, most often in the form of loans from financial institutions – but those institutions are often reluctant to lend to businesses in low-income areas. To address these challenges, in 1977 the U.S. Federal Government passed the Community Reinvestment Act (CRA; Macey & Miller 1993), which encourages financial institutions to provide loans to low-income areas. Unfortunately, research has shown that businesses owners in black-majority neighborhoods face unconscious discrimination in lending (e.g. Lederer & Oros 2020), and the CRA does not explicitly address race.

How well does the CRA do in reducing racial disparities in lending in our city? To find out, we looked at reported data from the Federal Financial Institutions Examination Council (FFIEC), the federal agency that evaluates banks on meeting their CRA lending goals and many other federal financial programs; we retrieved aggregated data by year and by Baltimore City census tract.[1] We normalized the aggregated loans by business activity in each tract, measured by the number of jobs reported in the U.S. Census Bureau's Longitudinal...

...Employer-Household Dynamics (LODES) dataset.[2] Finally, we got demographic data for each census tract, including the percentage of people reporting their race as "White" and as "Black / African American," from the 5-year average data in the... Census Bureau's American Community Survey (ACS).[3] We retrieved all this data programmatically and stored it in a shared data volume in IDIES's own SciServer science platform (www.sciserver.org). Our data and analysis notebooks are available in a SciServer research group; sign up for a free account and contact the authors for an invitation.

Figure 1 shows our results; the left panel shows the number of loans per small business from 2012 to 2017 in each census tract; the right panel shows the average racial makeup of each tract over the same period. Both maps show the same pattern, with white-majority neighborhoods in the center and black-majority neighborhoods to the east and west. The maps are clear, but can we quantify the relationship between race and small business lending?

Figure 2 is a scatterplot in which each point represents a census tract. The x-axis shows the percentage of black residents in the tract, while the y-axis shows the total amount of small business loans between 2012 and 2017, labeled in thousands. The dots are scaled by total population and colored by median family income. The figure also shows the best-fitting linear model for total loan amount as a function of percentage of black residents. For every additional one percent black population, a neighborhood receives on average $295 less in CRA loans per job ($r^2 = .071$). The race effect explains 7.8 percent of the variance in the data, higher than any other variable in our models.

To compare directly with previous research (Theodos, Hangen, Meixell, & Rajasekaran 2019, p. 3), we also looked at the combined effect of race and poverty. Figure 3 shows the results in terms of mean loan amount per small business job as a function of both majority race and poverty level, as defined in the Theodos reference. Error bars show standard errors within each category. An ANOVA shows no statistically significant difference between tracts by poverty level alone, but does show a statistically significant difference in both race and poverty together and race alone.

Together, these results provide more evidence that unconscious racism is still here, and still impacting Baltimore for the worse. We are now adding more years into our analysis, and looking at the same data for other cities to see how poverty and race affect people around the country.

REFERENCES

Lederer, A., & Oros, S. (2020). Lending Discrimination within the Paycheck Protection Program. 18. https://www.ncrc.org/lending-discrimination-within-the-paycheck-protection-program/ . Accessed October 6, 2023.

Macey, J. R., & Miller, G. P. (1993). The community reinvestment act: An economic analysis. Va. L. Rev., 79, 291.

Theodos, B., Hangen, E., Meixell, B., & Rajasekaran, P. (2019). Neighborhood Disparities in Investment Flows in Chicago. Washington, DC: Urban Institute.

[1] Annual aggregated data are available online at https://www.ffiec.gov/craadweb/aggregate.aspx

[2] https://lehd.ces.census.gov/data/lodes/LODES7/

[3] https://www.census.gov/programs-surveys/acs

Democratizing our Data

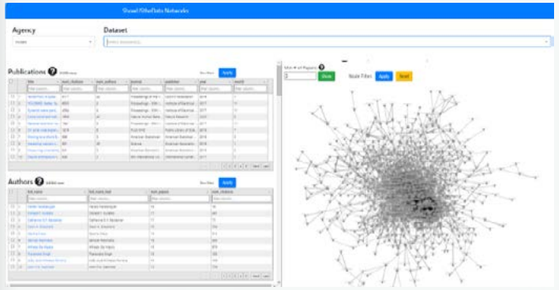
By Gerard Lemson

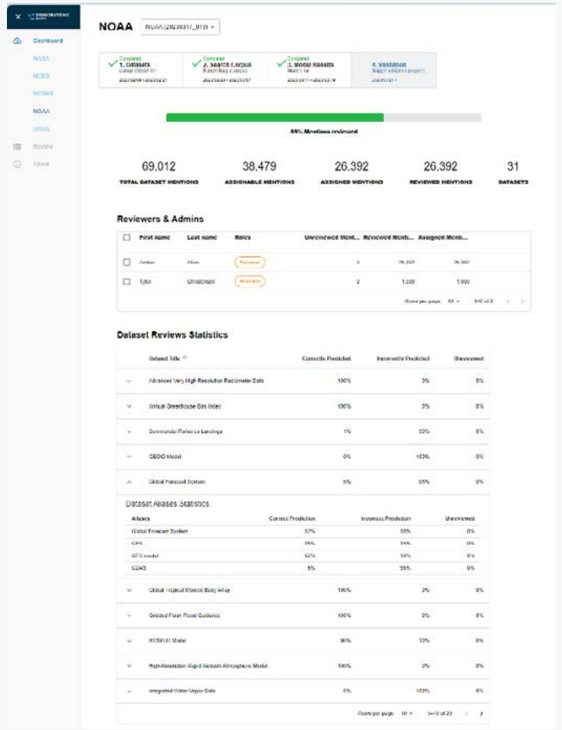
IDIES personnel have worked with Prof. Julia Lane and co-workers from NYU, Elsevier and TACC on a project to investigate the usage of datasets produced by (mainly) federal agencies. Here we give a summary of the background to the project and of the tasks performed by IDIES personnel in the course of the past year.

Background (by Julia Lane)

The call for better data and evidence for decision-making has become very real in the US as evidenced by the passage of both the Foundations of Evidence-based Policymaking Act (Evidence Act) and the CHIPS+ Act, establishing a National Secure Data Service. The challenge to be addressed is finding out not just what data are produced but how they are used – in essence, to build an Amazon.com for data -so that both governments and researchers can quickly find the data and evidence they need.

IDIES is participating in a massive effort that started five years ago which has been focused on finding out how data are being used, to answer what questions, and find out who are the experts, by mining text documents that are hidden in plain sight – in the text of scientific publications, government reports and public documents. The core idea has been to use Artificial Intelligence tools – Natural Language Processing and Machine Learning (ML) – to search and discover how datasets are referenced by authors in scientific text. A major Kaggle competition, a follow on conference, and a piece in the Harvard Data





CAPTION (left): Screen shot of admin dashboard

CAPTION (right): Screen shot of analysis dashboard

The current work—which is sponsored by agencies such as NSF's National Center for Science and Engineering Statistics (NCSES), NSF's Technology, Innovation, and Partnerships (TIP) Directorate, the Department of Education's National Center for Education Statistics (NCES), the US Department of Agriculture (both the National Agricultural Statistical Service and Economic Research Service) and the National Institutes of Health, and private foundations such as Schmidt Futures, the Alfred P. Sloan Foundation, and the McGovern Foundation—has generated a prototype API and dashboards that can be used – so that, for example, agencies can document dataset use for Congress and the public, program managers can identify investment opportunities rapidly and researchers can more easily build on existing knowledge rather than redoing things from scratch. More information is available at democratizingdata.ai

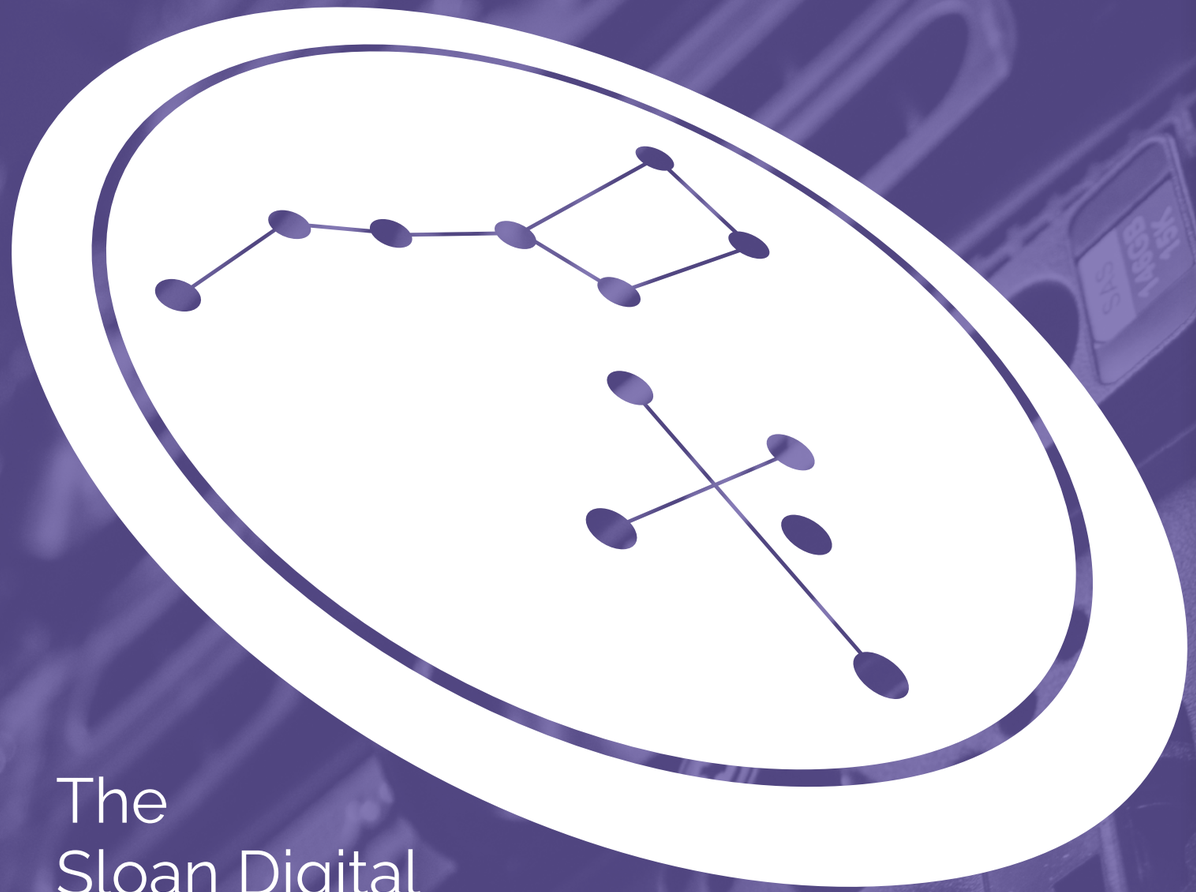
We expect the results to make a huge difference in understanding the public good produced by government dataset – to paraphrase Lee Platt's aphorism about HP – "If government knew what government knows, it would be three times more productive".

IDIES Activities

IDIES personnel have been involved in design and implementation of the database storing the result of the ML algorithms and has automated the loading pipeline leading from Elsevier through IDIES to TACC where results are stored for the public APIs. IDIES has upgraded a validation tool that allows members from the agencies to review the results of the ML algorithms. We have created an administration dashboard that allows agency members to track the progress of the pipeline and the validation process. And in collaboration with data scientists from Elsevier we have worked on improving the machine learning algorithms themselves. Through the SciServer platform we have provided agency users with advanced dashboards and custom Jupyter notebooks for deeper analysis of the results, for example of the author networks related to specific data sets.

This work was initially funded by Schmidt Futures through the Scientific Software Engineering Center and later through grants from the McGovern foundation, from various federal agencies and from the NSF.

CAPTION (right): *A time-lapse photo of the LVM instrument at Las Campanas Observatory in the Atacama Desert of Chile. The four black boxes contain the telescopes that make up the instrument; fiber optic cables behind the telescopes carry light into the integrated field unit (IFU) spectrographs. Left to right: Tom Herbst, Nick Konidaris, Florian Briegel, Markus Kuhlberg Image Credit: Maximilian Häberle (Max-Planck-Institute for Astronomy) and the SDSS collaboration.*



The Sloan Digital **Sky Survey-V**

SDSS-V is the first facility providing multi-epoch optical & IR spectroscopy across the entire sky, as well as offering contiguous integral-field spectroscopic coverage of the Milky Way and Local Volume galaxies. This panoptic spectroscopic survey continues the strong SDSS legacy of innovative data and collaboration infrastructure.

SDSS DR18

By Ani Thakar

IDIES continues to be a critical part of the Sloan Digital Sky Survey (SDSS, **https://www.sdss.org/**), the "Cosmic Genome Project" that launched the Big Data revolution in Astronomy at the turn of the millennium and is now into its fifth phase (SDSS-V). A grant from the Moore Foundation funds the SDSS-V Catalog Archive Server (CAS) that hosts and serves the multi-terabyte catalog data to the world from a dedicated cluster in the IDIES Data Center. The stalwart web portals - SkyServer and CASjobs - that were re-engineered as part of the SciServer collaborative science platform development by IDIES data scientists, are still going strong after 20+ years and continue to provide the bulk of data access and visualization for the SDSS catalog data.

The first public data release for SDSS-V, Data Release 18 or DR18, was announced to the world in January 2023. DR18 includes spectroscopic targeting catalogs prepared for the Black Hole Mapper and Milky Way Mapper science programs, as well as spectra for other specialized science programs and a Value Added Catalog of redshifts and classifications for a significant fraction of DR18 spectroscopic targets.

These data products comprised a major new subschema of data tables added to the CAS database, and in the tradition of cumulative SDSS data releases since the beginning of the project, DR18 incorporated all the prior data to date (DR17) along with the newly released data for the maximum science and discovery potential. New capabilities were added to SkyServer to support access to the new data products.

IDIES Seed Funding Initiative

Every fall, The Institute for Data Intensive Engineering and Science (IDIES) offers its members the chance to receive up to $25,000 in grant funding for eligible data-intensive research projects.

The goal of the Seed Funding initiative is to provide pilot funding for data-intensive computing projects that:

(a) will involve areas relevant to IDIES and JHU institutional research priorities;

(b) are multidisciplinary; and

(c) build ideas and teams with good prospects for successful proposals to attract external research support by leveraging IDIES intellectual and physical infrastructure.

Dr. Caleb Alexander, MD, MS
Bloomberg School of Public Health

Harnessing Image Detection To Help Address The U.S. Opioid Epidemic: An Analysis Of The Opioid Industry Documents Archive

Dr. Edward Chen, MD, MS
School of Medicine

Expanding the Clinical Capability and Scalability of Truly Remote Vital Sign Monitoring

Angelo Mele
Carey Business School

Systemic Risk and Externalities in Software Dependency Networks

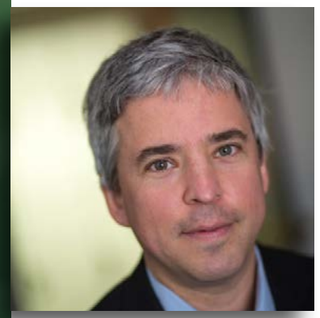
Corey Oses
Whiting School of Engineering

High-Entropy Anchors for High-Performance Lithium-Sulfur Batteries

2023 IDIES

Seed Funding Awards

Harnessing Image Detection To Help Address The U.S. Opioid Epidemic: An Analysis Of The Opioid Industry Documents Archive



SEED AWARDEE

Dr. Caleb Alexander, MD, MS is a Professor of Epidemiology and Medicine at Johns Hopkins Bloomberg School of Public Health, where he serves as a founding co-Director of the Center for Drug Safety and Effectiveness and Principal Investigator of the Johns Hopkins Center of Excellence in Regulatory Science and Innovation (CERSI). He is a practicing general internist and pharmacoepidemiologist and is internationally recognized for his research examining prescription drug utilization, safety and effectiveness.

PRESENTED BY

Venus Van Ness , JD, MS is a collections development Archivist currently working on the Opioid Industry Documents Archive at the Bloomberg School of Public Health.

Expanding the Clinical Capability and Scalability of Truly Remote Vital Sign Monitoring

SEED AWARDEE

Dr. Edward Chen, MD is the medical director of respiratory care services at Johns Hopkins Bayview and chair of the Hopkins Epic development project (electronic medical record) critical care workgroup. His dual roles facilitated implementation of many necessary changes to patient care practice to maintain safety during the COVID crisis. One key interest is to leverage technology to improve health care outcomes, particularly for underserved patient populations. The current grant application reflects the intersection of his clinical and research interests, recognizing the distinct value that a multi-disciplinary team provides for successful development and implementation of novel approaches to patient care.

PRESENTED BY

Dr. Joseph Angelo, PhD is a member of the Senior Professional Staff at the Johns Hopkins University Applied Physics Laboratory (JHU/APL), where he works in the Experimental and Computational Physics Group. Dr. Angelo received his B.S. in Physics from Drexel University in 2010. After a year of service, he attended graduate school and received his Ph.D. in Biomedical Engineering from Boston University in 2017. After receiving his Ph.D., Dr. Angelo was awarded the National Research Council Postdoctoral Fellowship to work at the National Institute of Standards and Technology for two years. Afterward, he dedicated a year to diffuse optics research with a team at Strasbourg University in France. In 2020, he joined the research staff at JHU/APL where he has been actively researching problems in experimental optics for health applications.

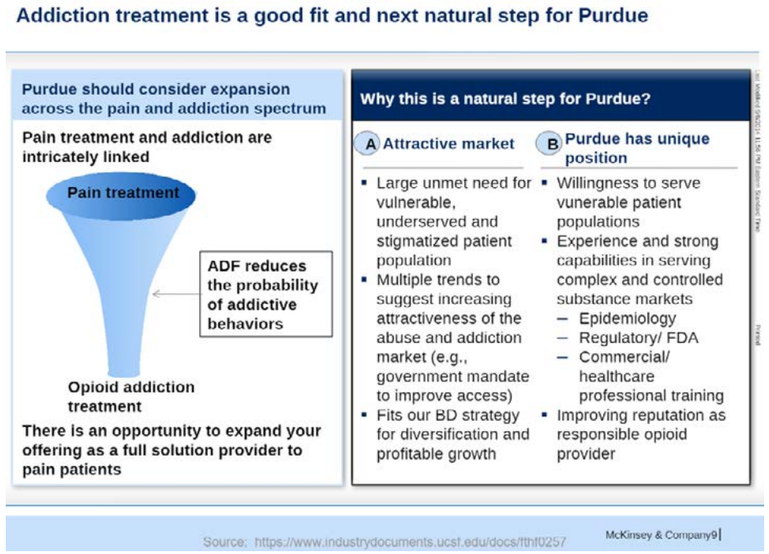


Figure 1: Addiction treatment is a good fit and next natural step for Purdue.

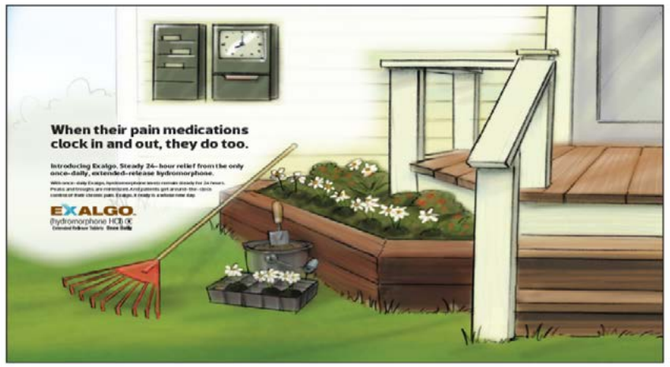


Figure 2: When their pain medications clock in and out, they do too.

Harnessing Image Detection To Help Address The U.S. Opioid Epidemic: An Analysis Of The Opioid Industry Documents Archive

Dr. G Caleb Alexander In collaboration with: Anqi Liu

JHU researchers from the Department of Epidemiology and the Department of Computer Science are rapidly and efficiently identifying compelling visual artifacts in the UCSF-JHU Opioid Industry Documents Archive, a highly innovative document collection focused on the U.S. opioid epidemic. The team is using complementary techniques to detect images in different file formats in the archive: rule-based techniques (image size and entropy of images) for images found within for PowerPoint and Excel documents and computer-vision techniques (distinguishing images from text in raster page images) for detecting images within raster page images. Figures 1-3 are examples of these images, which succinctly reveal some of the ways that opioid litigation defendants sought to increase sales of drugs they knew to be addictive and deadly.

The team is also building a pipeline for filtering and sorting a subset of the extracted images in order to construct training sets to use machine-learning for automated culling of the extracted images, yielding a refined set of content-rich images from the archive that can be made available on a public website. This OIDA Image Collection will include automatically and manually created metadata about the images and the documents from which they come. As a compelling additional entry point to the archive, the OIDA Image Collection will broaden the OIDA user audience and deepen the level of interaction with, and exploration of, archive documents, facilitating a broader understanding of the U.S. opioid epidemic.

Team members grateful for the support of the SciServer team in helping make the data available in the IDIES storage environment and are experimenting with running code on a copy of the data in AWS cloud storage, which allows for simple integration of third-party web-based tools like Label Studio.

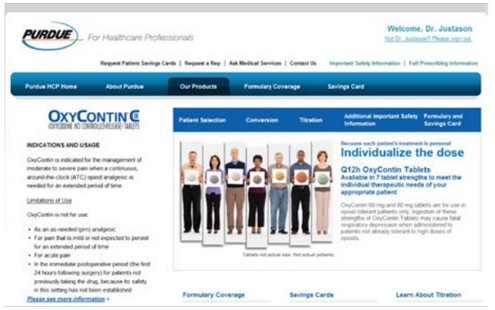


Figure 3: Individualize the dose

Expanding the Clinical Capability and Scalability of Truly Remote Vital Sign Monitoring

Edward Chen in collaboration with Joseph P. Angelo, Anissa Elayadi, and Robert L. Wilson

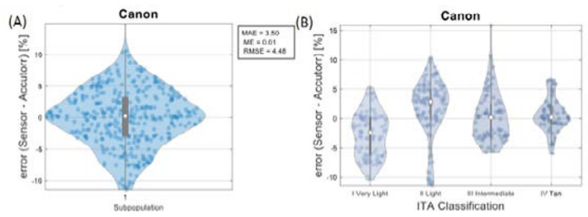


***Figure 1**: Oxygenation performance of the calibrated-device approach for 20 participants against the clinic-grade Accutorr gold standard system. (A) All participant errors are compiled for an overall mean-absolute error (MAE) of 3.5 and a mean error (ME) of 0.01. (B) The participant errors are analyzed by skin tone at grouped by ITA, showing a fairly balanced performance across skin tone.*

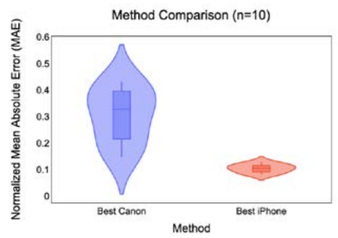


***Figure 2**: Comparing heartrate estimation from mobile device selfie-video with algorithms optimized for (blue) Canon DSLR from approximately 12 feet vs (red) parameters tuned for iPhone selfie-video.*

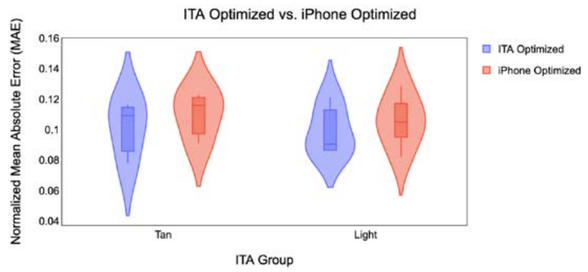


***Figure 3**: Comparing heartrate estimation performance for algorithms/parameters tune for (red) all participants vs (blue) skin-tone specific tuning.*

Access to healthcare remains a challenge due to factors such as sociodemographic inequities and limited capacity for health care delivery. The COVID-19 pandemic immediately revealed the severity of inequities in health care delivery and introduced new safety challenges to both healthcare workers and patients. There is a clear unmet need to develop novel approaches to extend the scope of healthcare delivery and limit the risks associated with highly transmissible infections. Specifically, developing strategies that can be adapted to commonly available equipment currently in use for telemedicine (smart phones, tablets, and webcams) would be the most efficient approach. In this update, we present work that attempts to helps bridge the gap between modern non-contact vital sign measurement methods and deployable telemedicine techniques for underserved communities.

This work stems from previously collected human subject data. Internal JHU Applied Physics Lab and JHU Discovery Program funding supported our effort to amass a comprehensive data set encompassing 91 healthy participants for testbed development (JHM IRB #00252571 Assessment of Remote Vital Signs for Telemedicine Applications; www.clinicaltrials.gov NCT04820387). While instrumented with a suite of clinical and research grade ground truth contact sensors (e.g., EKG, PPG, BP), subjects completed manual tasks and light physical exercise over a period of roughly 10 minutes. Concurrently, we recorded the subjects with standoff sensors to obtain non-contact imaging including thermal, hyperspectral, and RGB cameras of varying capability (i.e., 4K UHD, DSLR, smartphone), with all data (i.e., standoff sensors, clinical, and research) being time synchronized through a custom ArUco tag implementation. We built a data pipeline and presented two baseline results for estimating heart rate (HR) through remote-photoplethysmography (rPPG) and blood oxygenation (SpO2) via rPPG and a clinical system calibration. Two results were clear: both HR and SpO2 results depended on a participant's skin tone, and a data processing algorithm for one camera would yield non-optimized results if applied to another. These are major gaps for deployable telemedicine applications and our IDIES work pushes the needle forward.

For an updated SpO2 approach, we avoided techniques that require a ground-truth heartrate measurement to train a model. Instead, a "live calibration" approach requires a simple calibration target in the camera's view, largely following the work of Nishidate (Nishidate et al., 2020). We used a standard color chart to characterize the illumination and to white balance the scene for every frame of participant footage to convert the RGB measurements into a calibrated CIEXYZ color space that is device-independent. To interpret the calibrated images, a 7-layer Monte Carlo model of human skin was simulated for varying hemodynamics (blood volume and oxygenation) and skin tone (percent melanin) (Zherebtsov et al., 2019) using a GPU-accelerated photon transport model (Yan & Fang, 2020). Figure 1 demonstrates the error distribution in the approach's performance, showing a balanced performance across skin tone, though future work will apply active correction for melanin content.

To address the sensor-based variability of HR estimation, we applied a testing framework developed by Boccignone et al (2022) called pyVHR (https://github.com/phuselab/pyVHR). This enabled the testing of 2 skin extraction methods, 2 skin sampling methods, 3 patch sizes, 4 skin locations, 3 temporal window sizes, 3 blood volume pulse extraction methods, and 2 signal filters for a total parameter space of 450 permutations for each sensor for all participants. We then compared the HR estimation performance of the mobile device camera when running the DSLR-optimized algorithm versus the updated mobile-device-optimized algorithm. Results shows a clear benefit in overall accuracy of approximately 20-points in normalized mean-absolute error (MAE) (Figure 2), clearly showing that algorithm optimization should be scene and sensor specific.

2023 IDIES

Seed Funding Awards

Systemic Risk and Externalities in Software Dependency Networks

SEED AWARDEE

Angelo Mele, PhD is an Associate Professor of Economics at Johns Hopkins University – Carey Business School. His research analyses how social and strategic interactions affect individual and aggregate socioeconomic outcomes. His work has been published in Econometrica, American Economic Journal: Economic Policy, Journal of Business and Economic Statistics and The Review of Economics and Statistics. He has a PhD in Economics from University of Illinois at Urbana-Champaign.

PRESENTED WITH

Co-Pierre Georg is an Associate Professor at EDHEC Business School in France.

High-Entropy Anchors for High-Performance Lithium-Sulfur Batteries

SEED AWARDEE

Corey Oses, PhD Corey Oses is an assistant professor in the Department of Materials Science and Engineering. He leads the Entropy for Energy (S4E) Laboratory focusing on the discovery of materials for clean and renewable energy using computational and data-driven approaches. More information can be found at https://entropy4energy.ai.

PRESENTED BY

Keith Clark Keith is a second-year graduate student in The Johns Hopkins University's Department of Materials Science & Engineering. He graduated from George Mason University with a Bachelor of Science in Mathematics, concentrating in Applied Mathematics, and Chemistry, concentrating in Materials Chemistry. Keith's research career began with the NSF Chemistry REU program at the University of Utah where he studied the bond dissociation energies of diatomic transition-metal compounds via gas-phase laser spectroscopy under Prof. Michael Morse. Upon returning to George Mason he joined the Nanosynthesis, Optics, and Energy Laboratory led by Prof. Hao Jing where he studied the optical and catalytic effects of the site-selective deposition of palladium onto gold nanotriangles. Now he develops the aflow++ framework to discover high-entropy materials for clean and renewable energy in the Entropy for Energy Laboratory at Johns Hopkins, led by Prof. Corey Oses.

Systemic Risk and Externalities in Software Dependency Networks

Angelo Mele in collaboration with: Co-Pierre Georg

“Modern software development involves collaborative efforts and re-use of existing software packages and libraries, to reduce the cost of developing new software. However, package dependencies expose developers to the risk of contagion from bugs or other vulnerabilities that may cost billions of dollars. This project will model the maintainers’ decisions to create dependencies among software libraries in an equilibrium strategic network formation game. After estimating the parameters of such model using data from https://libraries.io, we can quantify and understand the externality imposed by such dependencies in terms of contagion risk from bugs or other vulnerabilities. This analysis will provide a measure of systemic risk for a software ecosystem.”

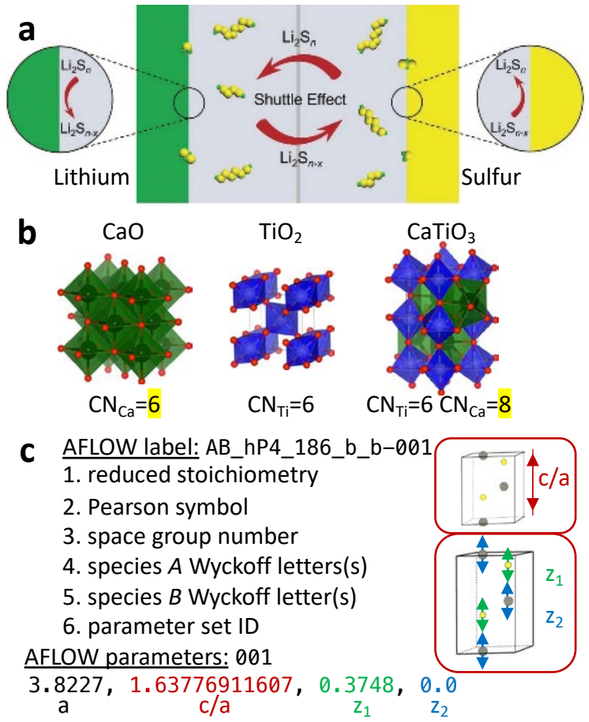


Figure 1. Database construction for HE-MOF stability analysis. (a) schematic of the polysulfide shuttle effect [4]. (b) coordination change; crystal structures of CaO, TiO2, and CaTiO3: Ca is black, Ti is grey, and O is red. Notice Ca has six-fold coordination with oxygen in CaO and eight-fold in CaTiO3 [5]. (c) sample AFLOW prototype designation for Wurtzite.

High-Entropy Anchors for High-Performance Lithium-Sulfur Batteries

Corey Oses in collaboration with: Sara Thoi

Lithium-sulfur (Li-S) batteries pose game-changing advantages over state-of-the-art Lithium-ion systems: an order of magnitude higher theoretical charge-storage capacity, up to 3 times higher gravimetric energy density, low cost, light weight, and fewer safety concerns [1–3]. However, Sulfur cathodes degrade in common electrolytes, forming lithium polysulfide species (LiPSs) that insulate the anode, diminishing cyclability and preventing commercialization (i.e. the "polysulfide shuttle" effect (Fig. 1a)). This project aims to discover the first high-entropy metal-organic frameworks (HE-MOFs) for LiPS immobilization by employing the high-throughput ab-initio aflow++ framework to screen promising candidates followed by experimental verification. This report provides an update on the first project objective: constructing a database of binary and ternary sulfide systems for high-throughput stability analysis of candidate HE-MOF materials, generated by prototyping experimental structures from the Inorganic Crystal Structure Database (ICSD).

Two issues inhibit database construction: (i) known errors in density functional theory's predicted formation enthalpy values for polar compounds (i.e. sulfides) (ii) absence of sulfide prototypes in the aflow++ codebase. AFLOW's coordination corrected enthalpies scheme (CCE) addresses the first issue [5, 6]. CCE offers corrections on a per-bond basis, enabling users to predict the relative stability of polymorphs with high accuracy (Fig. 1b) [5]. CCE corrections are defined as the difference between experimental and predicted formation enthalpies; sparse empirical sulfide data requires an abinitio approach. We aim to expand CCE by machine learning corrections, the first step of which is to construct a training set based on existing experimental data.

Prototyping the ICSD sulfides serves two purposes: (i) populating a structural database of experimental binary and ternary sulfides, and (ii) enabling the generation of a training dataset to machine learn sulfide CCE corrections. Pairing structural data with thermodynamic quantities derived from the learned CCE corrections will establish a database suitable for high-throughput sulfide stability analysis.

This seed funding has enabled the Entropy for Energy team to develop an automated workflow for high-throughput prototype injection into the aflow++ codebase.

2023 IDIES

Summer Student Fellowship (SSF) Awards

The Summer Student Fellowship program (SSF) offers awards of $6,000 to support summer research projects led by undergraduate students with the guidance of an IDIES faculty member mentor. These projects are meant to provide an opportunity for students to participate in a 10-week (June–August) full-time data-science focused research project.

This year, IDIES congratulates the following three Summer Student Fellowship recipients on the successful completion of their data-intensive research.

Jay Kim
Mentored by Mitra Taheri

Intelligent Microscopy Methods for Characterizing Intermediate States Using EELS

Sampath Rapuri
Mentored by Dr. Robert D. Stevens

Development of a Machine Learning Model for Pulmonary Embolism Prediction in Intensive Care

William Shiber
Mentored by Corey Oses

Creating a Federated Materials Data Infrastructure: Completing the AFLOW Data Pipeline

SUMMER STUDENT FELLOWSHIP AWARDEE

Jay Kim

Mentored by Mitra Taheri

Jay Kim is a sophomore attending Johns Hopkins University double majoring in Materials Science & Engineering (MSE) and Computer Science. Originally from New York, He has experience in high school in TEM research and is currently continuing work in the same branch with the Taheri Group in the department of MSE.

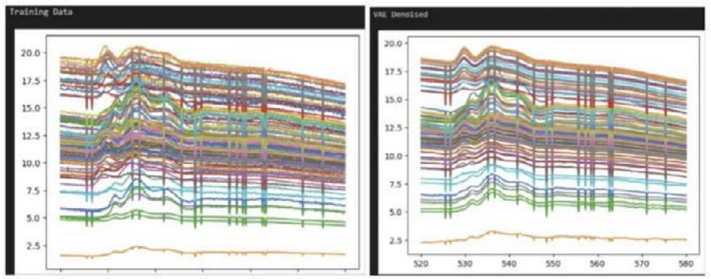


Figure 1. (above) *Original training data gathered from TEM EELS (left); deionized cleaned spectra from our VAE classifier*

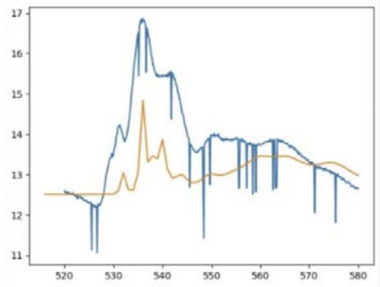


Figure 2. (right) *VAE denoised spectra (blue) vs. FDMNES simulated spectra (orange)*

Intelligent microscopy methods for characterizing intermediate states using EELS

Jay Kim* in collaboration with PI Mitra Taheri & Jonathan Hollenbach

Transmission Electron Microscopy (TEM) is a powerful tool for visualizing nanoscale samples, able to magnify specimens up to two million times. A technique called Electron Energy-Loss Spectroscopy (EELS) can give additional information, such as bonding/valence states, near-neighbor atomic structures, and free-electron density for any element on the periodic table. However, these techniques require extremely high data rates (up to 4.5 TB per hour), and return incredibly complex information that is hard to interpret. These limitations have created a need for improved imaging quality and efficiency in TEM methods – needs that can be partially met through machine learning (ML) techniques.

The ML application we chose use was a Variational Autoencoder (VAE). Like traditional autoencoders, VAEs work by compressing data from a higher dimensional space to a lower, and then uncompressing back. Though autoencoders are mainly used for image compression and reconstruction, we chose one to focus on the latent space, which serves as a representation for input data in a compressed format that can be processed through computation. We constructed our ML model using Keras[8], a deep learning Application Programming Interface, built on the Tensorflow open-source software.

The target dataset for this study was spectra gathered from an EELS TEM process of a $SrFeO_3$ perovskite structure. The model system was the reduction of a SrFeO3-δ thin film. After a process of in-situ heating on the sample, a focused ion beam (FIB) was used to perform a conventional FIB lift-out. The sample was then removed and allowed to anneal. This film was then reduced to δ = 0.5, producing a SrFeO2.5 brownmillerite structure. These TEM-EELS measurements were recorded using a JEOL 2100F Microscope.

We first gathered our experimental spectra by running our SFO data through our VAE code. We then created simulated spectra for each structure by plugging CIF files for each structure (obtained from the Open Quantum Materials Database) into a program called the Finite Difference Method for Near Edge Spectra (FDMNES).

The 45 files in our experimental TEM EELS dataset were split up into X-labeled images (the original single pixel images) and Y-labeled images (the denoised clean spectra for each of the points). Using the VAE to predict the Y spectra, we were able to denoise the training spectra. Figure 1 shows the training spectra (left) and the denoised clean spectra from our VAE classifier (right).

With dimensionality-reduced latent space representations as well as the experimental spectra we gathered, we moved into the FDMNES simulations. We searched through a variety of different $SrFeO_3$ perovskite structures to find possible matches for our experimental data. One such example is found in Figure 2. The simulated spectra that we found is the OQMD Cubic Perovskite structure. The match is not perfect, possibly due to a background subtraction error, but the fact that the peaks align is a strong indication that this particular structure was either an intermediate state or initial/annealed state found in our TEM EELS.

Based on the results, this does justify the need for from-scratch DFT with actual experimental to accurately identify intermediate states, and our FDMNES simulations are a prominent first step in that process.

* *This article has been abridged for length. The full version of the project report will be available on the IDIES website.*

SUMMER STUDENT FELLOWSHIP AWARDEE

Sampath Rapuri

Mentored by Dr. Robert D. Stevens

Sampath Rapuri is a second-year student pursuing a dual degree in Biomedical Engineering and Computer Science. He is interested in developing computational pipelines that can improve the precision, efficacy, and outcomes of care delivered to critically ill patients. After graduation, he plans to pursue an MD-PhD and establish his own research lab. In his free time, Sampath enjoys playing badminton and gardening.

Figure 1.

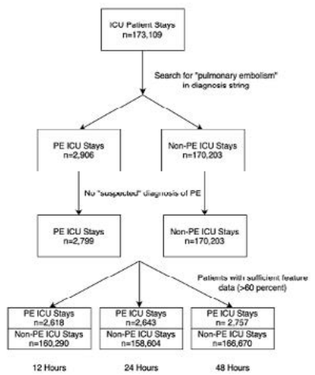
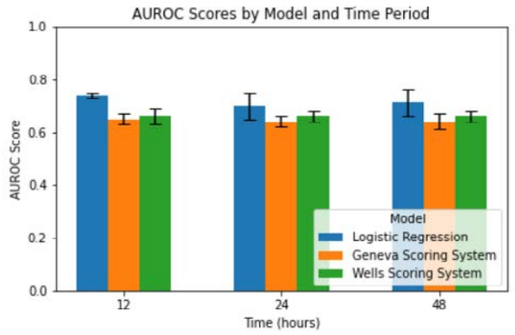


Figure 2.



Development of a Machine Learning Model for Pulmonary Embolism Prediction in Intensive Care

Sampath Rapuri in collaboration with Dr. Robert D. Stevens

Pulmonary embolism (PE) is a frequent and life-threatening complication in hospitalized patients. While statistical risk scores have been proposed, there is an unmet need for more accurate methods that can forecast the likelihood for developing PE.

Current PE scoring systems (e.g., Modified Wells Scoring System, Revised Geneva Scoring System, and the Pulmonary Embolism Rule Out Criteria) allow clinicians to estimate risk based on a discrete number of predictive factors used as inputs in multivariable logistic regression equations [1]. However, these models have limited predictive accuracy, in part because they may not consider the physiological complexity of acutely ill patients, and because the statistical modeling techniques used may not be optimized for dynamic prediction tasks [2,3].

I used a large, multicenter database (eICU) containing >200,000 ICU admissions from 208 hospitals across the US to gather and analyze 2,799 unique ICU stays where a diagnosis of PE was recorded. From this data, I created three separate 'datasets' using differing observation windows (12, 24, and 48 hrs.) of time-dependent features (e.g. lab values and vital signs) due to no clear way to get the exact time of occurrence of PE. Figure 1. Details the inclusion and exclusion criteria of ICU stays used in this study. On each observation window, I evaluated multiple different machine learning (ML) models: decision tree, random forest (RF), gradient boosting (XGBoost, CatBoost, and GBoost), generalized linear models (GLM), support vector machine (SVM), and artificial neural network (ANN). After hyperparameter tuning, I compared these models to current PE risk scoring models (Wells and Geneva risk scoring models). Figure 2. Outlines the AUROC scores for the top performing model (logistic regression) and current risk scoring models for all three observation windows.

I am currently in the process of external validation with the Precision Medicine Analytics Platform (PMAP) dataset from Johns Hopkins. Additionally, I am revisiting the hyperparameter tuning process to address issues of overfitting.

After external validation, I aim to publish these results in a peer-reviewed journal.

This project allowed me to learn so much – both technically as well as professionally. I learned about the entire scientific process from crafting a hypothesis to sharing my results with the broader scientific community.

References:

1. Doherty S. Pulmonary embolism An update. Aust Fam Physician. 2017 Nov;46(11):816-820. PMID: 29101916

2. van Doorn, W. P., Stassen, P. M., Borggreve, H. F., Schalkwijk, M. J., Stoffers, J., Bekers, O., & Meex, S. J. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. PLOS ONE, 16(1). https://doi.org/10.1371/journal.pone.0245157

3. Kamran Boka, M. D. (2022, June 29). Pulmonary embolism clinical scoring systems. Overview, Modified Wells Scoring System, Revised Geneva Scoring System. Retrieved January 16, 2023, from https://emedicine.medscape.com/article/1918940-overview

SUMMER STUDENT FELLOWSHIP AWARDEE

William Shiber

Mentored by Corey Oses

William Shiber is a third-year student pursuing a BS/MS in Applied Mathematics and Statistics with a focus on Statistics and Statistical learning. While originally from northern New Jersey, Shiber has spent much of his time in in New Hampshire.

Creating a Federated Materials Data Infrastructure: Completing the AFLOW Data Pipeline

William Shiber in collaboration with Corey Oses

Recent applications of machine learning to materials data repositories have accelerated computational materials design. Machine learning algorithms have enhanced the prediction of materials properties such as thermodynamic stability, potential energy surfaces, and electronic structures [1–3]. The Automatic Flow (aflow++) framework for materials discovery is an open-source software for high-throughput materials simulations that has generated the largest database for computationally investigated inorganic materials [4, 5]. It has successfully predicted novel experimentally verified magnets [6], super-alloys [7], high-entropy car- bides [8, 9], and phase change memory compositions [3]. The aflow++ framework [10] has created over 3.5 million entries spanning 1,100 crystallographic prototypes, all housed within the aflow.org online data repositories [11].

This summer we created AFLUX@JHU: a materials search-API for the aflow.org data repositories generated and housed at Johns Hopkins University. The original AFLUX implementation — written in PHP — was not built within the public distribution of the aflow++ framework. AFLUX@JHU is the first publicly available AFLUX implementation, written in C++ directly within aflow++. AFLUX@JHU connects the repositories generated and stored at JHU with those housed at other universities, creating a federated environment of materials data — enhancing research efficiency while maintaining data provenance.

The ability to filter materials on specific tensor components is extremely valuable. By implementing new syntax into the AFLUX language and developing new SQL queries to search the database we were able to allow for tensor-like filterability. Additionally, many materials parameters are specific to certain atoms or species within the crystal. Because of the dynamic nature of these data vectors, filtering within a relational database using SQL is almost infeasible. Instead, we created a second layer of filtering using C++ within aflow++ to allow for users to filter on these properties. Both new types of filterability make use of the '@' operator within the AFLUX language to specify the desired splicing of these matrices.

https://s4e.ai/search/API/?ael_stiffness_tensor(@(3,3)(0.2*))

Tensor like filtering example on the component in position 3,3 of the ael stiffness tensor

https://s4e.ai/search/API/?bader_net_charges(@(Fe)(0*))

Filtering on the bader charge of the Fe atoms in the structure (note that this is "if any")

Further work is needed to fully connect our data with that at other universities and complete the "plug-and-play" ability of AFLOW. This simply implies editing how AFLOW chooses where to pull the data from, a DB file on the machine or online from the JHU, Duke, or other databases.

Coming into this project I had very little coding experience, none with C++, SQL, or anything close to the size of the AFLOW framework. Throughout the summer I was able to hone my skills within database design and implementation, developing with C++, and leveraging SQL to query databases. Beyond the basic hard-skills I developed through this project, I was able to learn what it was like to work within a research group. I was able to strengthen my communication and teamwork skills as well as becoming a better at managing large, long-term projects.

IDIES Mission Statement

To further the JHU mission of “Knowledge for the world” by providing intellectual leadership in data-driven science:

- Research and Education in adaptive disruptive technologies
- Technical and domain expert guidance via collaboration and consultation
- Open and sustainable long-term access to high-value datasets

IDIES is a partnership of:

Bloomberg School of Public Health
Carey Business School
Krieger School of Arts and Sciences
School of Medicine
Sheridan Libraries
Whiting School of Engineering

Johns Hopkins Bloomberg School of Public Health

Johns Hopkins Sheridan Libraries & University Museums

Johns Hopkins Carey Business School

Johns Hopkins School of Medicine

Johns Hopkins Krieger School of Arts & Sciences

Johns Hopkins Whiting School of Engineering

THANK YOU TO OUR GENEROUS SPONSORS

The IDIES Executive Committee would like to extend our heartfelt gratitude to our affiliates, collaborators, contributors, editors, and staff, without whose continued support and cooperation IDIES would not be possible.

IDIES is always accepting members. For more information about our membership categories, please visit **idies.jhu.edu** joi