2020 Annual Symposium

October 23, 2020

A JOINT EVENT HOSTED BY:

MiNDS
Mathematical Institute for Data Science

idies
The Institute for Data Intensive Engineering and Science

Johns Hopkins University

MESSAGE FROM THE DIRECTORS

Alex Szalay, PhD

Bloomberg Distinguished Professor
Professor, Computer Science, JHU
Director, Institute for Data Intensive Science

René Vidal, PhD

Herschel Seder Professor of Biomedical Engineering
Director, Mathematical Institute for Data Science
Professor of Computer Science,
Mechanical Engineering,
and Electrical and Computer Engineering

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, and giving JHU a strategic advantage in making new discoveries.

Today AI and Machine Learning are everywhere. To be successful, we need expertise in: modern AI techniques, computational capabilities, translational capabilities, and enabling an easy transition of data to "AI-ready". We spend a disproportionate amount of time converting our data sets into a format that AI frameworks can use. Large scale studies need to aggregate data from many different sources and transform them into a common system.

Making decisions based on facts and relying on sound data foundations is essential. To create and grow diverse datasets, we have active partnerships across the JHU community. Emerging collaborations include SAIS, School of Education, MINDS, the multi-institutional Paradim project in materials science (HEMI), 21CC, cancer immunotherapy projects with the Bloomberg Kimmel Institute, and various large-scale studies in genomics.

This year we aim to make these synergies even closer. We are holding the annual meetings of MINDS and IDIES together to demonstrate how we can work together, and make JHU a center for modern AI research in the country and the world. It is imperative to come up with innovative strategies on how to create unique data sets that can drive large-scale machine learning, like digitizing millions of physical specimens at JHMI. We are using AI in a large astronomical survey to decide which of the hundreds of millions of celestial objects should be observed through a spectrograph. We are trying to find new patterns in large numerical laboratories with a few petabytes of data in turbulence, ocean circulation and galaxy clustering, using AI techniques.

Quantitative social science, using advanced computer science and statistics in tackling large scale social problems is at the heart of modern sociology. We are working with 21CC, the SNF Agora Institute, and the Carey Business School to establish critical mass locally. We, also, have a new project with SAIS on the nation's energy infrastructure.

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

MINDS, faculty and students work together to decipher the mathematical, statistical and computational foundations of data science. Over the past year, MINDS received a $10M award from the Simons Foundation and the National Science Foundation to develop the mathematical foundations of deep learning and a $1.5M award from the National Science Foundation to create a Transdisciplinary Research in Principles of Data Science (TRIPODS) Institute on the foundations of graph and deep learning.

MINDS also recruited Dr. Rama Chellappa as a Bloomberg Distinguished Professor in Biomedical Engineering and Electrical and Computer Engineering and Dr. Soledad Villar as an Assistant Professor in Applied Mathematics and Statistics. In addition, MINDS faculty received numerous awards for their work, including the IEEE Jack S. Kirby Signal Processing Medal (Rama Chellappa), the Institute for Mathematical Statistics Fellowship (Carey Priebe), the American Institute for Medical and Biological Engineering's College Fellowship (René Vidal), the Simons Fellowship in Mathematics (Mauro Maggioni) and the NSF CAREER Award (Raman Arora and Ilya Schpitser).

Further, MINDS faculty worked with the Departments of Applied Mathematics and Statistics and Computer Science to establish a new Master program in Data Science, which is directed by Prof. Laurent Younes, and recruited its first cohort of students this spring. Moreover, MINDS awarded eight Data Science Fellowships and two MINDS Doctoral Dissertation Awards. MINDS also organized a series of collaborative events, including two symposia and twenty four seminars on the foundations of data science.

This year, we are pleased to join forces with the Institute for Data Intensive Science and Engineering (IDIES) to organize a joint symposium in both the foundations of data science and its applications in data-intensive problems. ■

CONTENTS

ON THE COVER: The cover image is a velocity magnitude contour plot on a plane across one of the world's largest numerical simulations of turbulent flow, on 1/2 trillion grid-points. Simulation by PK Yeung (IDIES collaborator at Georgia Tech); data publicly available at JHTDB (Johns Hopkins Turbulence Databases, http://turbulence.pha.jhu.edu).

START	DUR.	SESSION	SPEAKER	PRESENTATION
			MORNING SESSION A	
9:00 am	15	Opening Remarks	Alex Szalay (JHU) & Rene Vidal (JHU)	
9:15 am	15	MINDS Awards Ceremony	Rene Vidal (JHU)	
9:30 am	40	Keynote	Lauren Gardner (JHU)	Tracking COVID-19 in Real-time: Challenges Faced and Lessons Learned
10:10 am	15	IDIES Seed Awardee	Marc Stein (JHU)	Development of Tools to Automate and Harmonize Spatial Open Source Urban Data
10:25 am	30	Break		
		* Breakout Session	Gerard Lemson (JHU)	SciServer
			MORNING SESSION B	
10:55 am	40	MINDS Plenary	Francis Bach (INRIA)	The convergence of gradient descent for wide two-layer neural networks
11:35 am	20	IDIES Plenary	Brice Menard (JHU)	The Sequencer – how to reveal the main trend in your dataset?
11:55 am	15	IDIES Seed Awardee	Nik Paliwal (JHU)	An artificial intelligence approach towards predicting recurrence of atrial fibrillation in patients undergoing pulmonary vein isolation
12:10 pm	20	Poster Madness		
12:30 pm	60	Lunch		
		* Breakout Session	Poster Presentation Discussion	
			Lunch Conversation & Networking Rooms	
1:00 pm	30	* Francis Bach Q&A	Francis Bach (INRIA)	
			AFTERNOON SESSION A	
1:30 pm	40	MINDS Plenary	Sham Kakade (UW)	What are the Statistical Limits of Offline reinforcement Learning?
2:10 pm	40	IDIES Plenary	Gianluca Iaccarino (Stanford)	Engineering Simulations in the Age of Data
2:50 pm	15	IDIES Seed Awardee	Ben Haeffele (JHU) & Matthew Ippolito (JHU)	Machine Learning and Computer Vision for Malaria: Disentangling the *in vivo* Effects of Antimalarial Drugs using an Automated Malaria Microscopy Algorithm
3:05 pm	30	Break		
		* Breakout Session	Jaime Combariza (JHU)	MARCC
		* Sham Kakade Q&A	Sham Kakade (UW)	
			AFTERNOON SESSION B	
3:35 pm	40	MINDS Plenary	Deanna Needell (UCLA)	Online nonnegative matrix factorization for Markovian and other real data
4:15 pm	15	IDIES Seed Awardee	Thomas Haine (JHU)	Towards the Development of Scale-Dependent Turbulent Closures in Rotating Stratified Flows
4:30 pm	15	MINDS Thesis Award	Rene Vidal (JHU)	
4:45 pm	15	Closing Remarks	Alex Szalay (JHU) & Rene Vidal (JHU)	
5:00 pm	30	* Deanna Needell Q&A	Deanna Needell (UCLA)	

** these sessions are being held in a seperate Zoom room. For a links to these rooms please visit*

Lauren Gardner, PhD

Associate Professor, Department of Civil and Systems Engineering, JHU
Co-director, Center for Systems Science and Engineering
Affiliated Faculty, Infectious Disease Dynamics Group - Bloomberg School of Public Health
CSIRO Visiting Scientist

Source: Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time.

Tracking COVID-19 in Real-time: Challenges Faced and Lessons Learned

The dashboard, first shared publicly on January 22, illustrates the location and number of confirmed COVID-19 cases, deaths and recoveries for all affected countries. It was developed to provide researchers, public health authorities and the general public with a user-friendly tool to track the outbreak as it unfolds.

Lauren Gardner, associate professor in the Department of Civil and Systems Engineering at Johns Hopkins Whiting School of Engineering, is the creator of the interactive web-based dashboard being used by public health authorities, researchers, and the general public around the globe to track the outbreak of the novel coronavirus. The dashboard, which debuted on January 22, became the authoritative source of global COVID-19 epidemiological data for public health policy makers and many major news outlets worldwide. Because of her expertise, Gardner was one of six Johns Hopkins experts who briefed congressional staff about the outbreak during a Capitol Hill event in early March 2020.

Gardner is co-director of the Center for Systems Science and Engineering and affiliated faculty in the Johns Hopkins Bloomberg School of Public Health. Prior to joining JHU in 2019, Gardner was a senior lecturer in civil engineering at the University of New South Wales (UNSW) Sydney, in Australia. Her research expertise is in integrated transport and epidemiological modeling. Gardner has previously led related interdisciplinary research projects which utilize network optimization and mathematical modeling to progress the state of the art in global epidemiological risk assessment. Beyond mobility, her work focuses more holistically on virus diffusion as a function of climate, land use, mobility, and other contributing risk factors. On these topics Gardner has received research funding from organizations including NIH, NSF, NASA, the Australian Government National Health and Medical Research Council (NHMRC), Australian Research Council (ARC), Commonwealth Scientific and Industrial Research Organization (CSIRO), Queensland Health Department, Transport for New South Wales and GoGet Car Share. Outcomes from her research projects have led to publications in leading interdisciplinary and infectious disease journals, presentations at international academic conferences, as well an invited seminarsandkeynotetalksatUniversitiesand various events. Gardner is also an invited member of multiple international professional committees, reviewer for top-tier journals and grant funding organizations, and invited participant of various Scientific Advisory Committees. She has also supervised more than 30 students and post-docs, and teaches undergraduate and graduate level courses on network modeling and transport systems.

Research

My research focuses on advancing the state-of-art in data-driven epidemic planning and decision making in order to provide outbreak assessment and control recommendations based on best available evidence. Most notably, I lead the efforts behind the interactive web-based dashboard being used by public health authorities, researchers, and the general public around the globe to track the outbreak of the novel coronavirus in 2020.

Research Interests:

- Public health
- Bio-secure mobility
- Transport systems
- Infectious disease spread
- Network modeling

Marc L. Stein, PhD

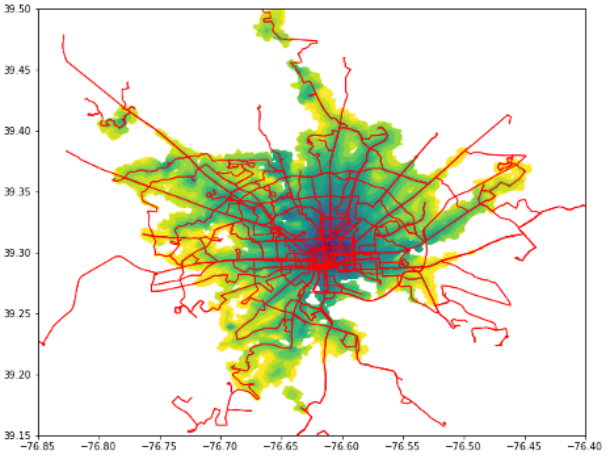
Associate Professor in the School of Education, JHU
Research Co-Director of the Baltimore Education Research Consortium (BERC)
Faculty Associate with the Hopkins Population Center (HPC), Faculty Affiliate with the Center for Social Organization of Schools (CSOS).

Development of Tools to Automate and Harmonize Spatial Open Source Urban Data

Getting to school in a city with high crime rates, concentrations of poverty, crumbling physical infrastructure, and no designated "yellow bus" student transportation can be time-consuming and stressful. However, little is known about how students get to school using public transportation and the kinds of social and environmental stressors they experience along the way. Our recent work has begun to explore the relationships between commuting difficulty and student disengagement from school in the form of increased absenteeism has been shown in the transition from middle to high school in Baltimore City, early high school transfer and decreased on-time arrival at school. Further, we have shown absenteeism to be related to commuting difficulty in the form of increased exposure to violent crime along the route to school.

In this seed grant we are developing a system for ingesting real-time transit vehicle location data and weekly crime incidence data that will allow for the examination of the daily dynamics of attendance in relationship to daily differences in environmental exposure to violence and transit unreliability in ways that provide more concrete behavioral mechanisms than aggregate outcomes and static student predictors ever could. To our knowledge, there are no existing studies that examine attendance with this level of temporal specificity.

The key aim of this project is develop the strategies and methods for ingesting real-time vehicle location feeds which will be converted into daily general transit feed specification (GTFS) data files. These daily GTFS files will then be used to generate daily routes to school using a routing engine that we have developed using the OpenTripPlanner (OTP) software. This routing engine takes as inputs OpenStreetMap (OSM) data, school locations and a regular mesh grid of origin points that have a spatial resolution of 300 x 300 feet and cover the entirety of Baltimore City. Together these methods will afford the capacity to estimate daily routes to school from actual arrival, departure and transit times of vehicles which could then be spatially and temporally merged to violent crime incident data and linked to students' daily attendance and administrative data.



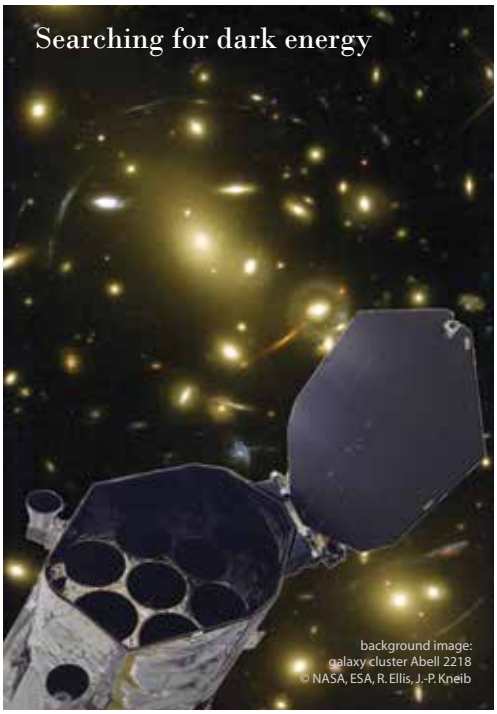
Isochrones for reaching a certain location using bus routes in Baltimore. Bus lines are superimposed.

To date we have received a storage space allocation for the project in a Microsoft SQL Server Big Data Cluster (BDC) instance at IDIES. This system combines relational database technology with a Spark/HDFS big data framework. It is a perfect system for implementing ETL pipelines such as required for this project. We have written scripts for automating loading of GTFS streams every seven seconds and store the resulting JSON files in the HDS system. The script runs in the same Kubernetes system supporting the BDC itself, which makes it robust for failures. The results are transformed daily to a parquet format that allows very efficient querying using Spark SQL. We also have loaded metadata about the routes of the MTA public transport system. This step can be fully performed from python code running in a Jupyter notebook within SciServer. It uses pyspark to connect to the BDC from within a container using a special purpose built SciServer/Compute Docker Image.

The further development of this system is still in progress. We have started some analysis using notebooks with the same SciServer compute image. For example, one important step we must make is to compare an actual route to the one that was planned thus we need to infer the likely stop times that a bus made at the planned stops. We do this by finding the nearest location of the trip to each of the stops using the coordinates of the bus and stops which is straightforward with Spark SQL. Figure 1 shows a plot of the planned vs the expected stop times for the same trip for a given route on a number of days. We have also run an OTP router inside of a SciServer container and have made requests to its REST API (see Figure 2), showing the ease by which this tool can be incorporated in the analysis pipeline. Next steps in the project include developing daily GTFS schedules, ingesting weekly crime data and demonstrating variability in transit reliability and crime exposure along routes over time. ■

Gerard Lemson, PhD

Research Scientist, JHU
IDIES Director of Science



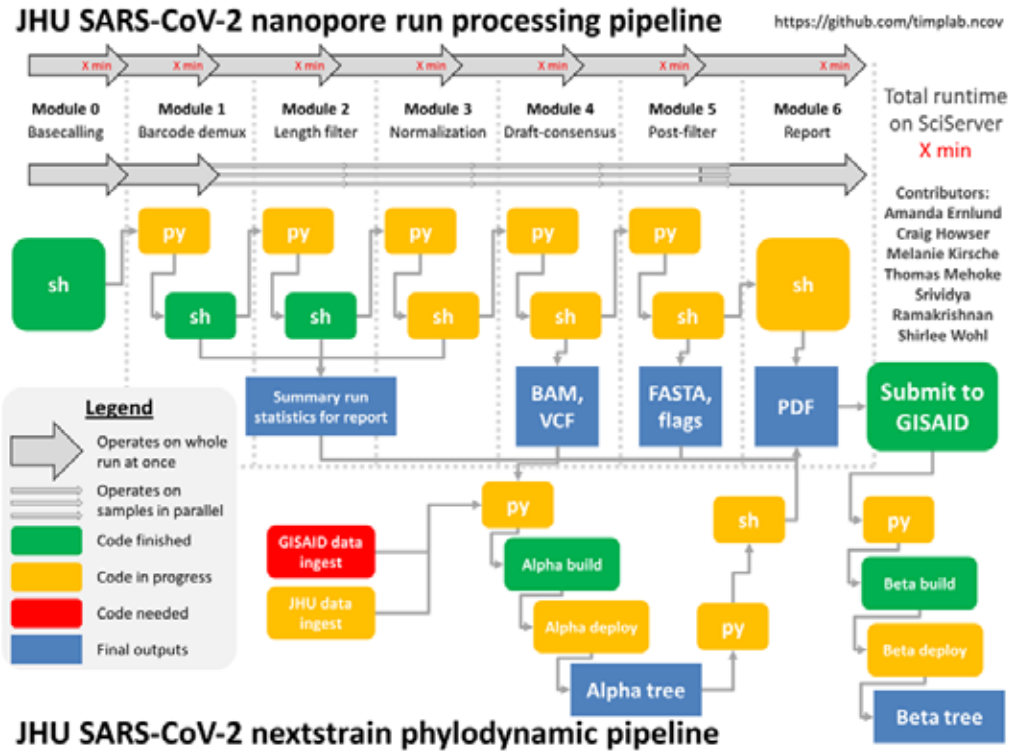
SciServer 2020

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

In the past year SciServer has been used to support various new scientific projects from groups at JHU. A special case was our support for combined efforts across JHU that established a pipeline to create a broad genetic characterization of SARS-COV-2, the virus which causes COVID-19. Sciserver/IDIES helps to house, organize and analyze the data, making it accessible on the back-end to computational analysis and on the front-end through web servers to their group and other researchers throughout the JHU family (JHU, SPH, APL). We have also started a project to host all public data sets from the HEASARC high-energy astronomy archive at Goddard Space Flight Center. We host their hardware with initially some 125TB of data, which is expected to increase by some 25TB/yr, and make it available to SciServer users.

The SciServer platform has also had its first full deployments outside of JHU. At the Max-Planck institute for extraterrestrial Physics in Munich, Germany, a SciServer instance was deployed for hosting and disseminating results of the eROSITA X-Ray satellite. At NIST an instance was deployed in a protected AWS environment and various use cases are in progress to test the system. This is supported further by a second PREP agreement between NIST and IDIES.

If anyone is interested in using the SciServer platform, please contact Gerard Lemson (glemson1@jhu.edu). ■



Francis Bach, PhD

Research Faculty, INRIA – National Institute for Research in Digital Science and Technology, Paris, France
Scientific Leader, SIERRA – Machine learning research laboratory

Francis Bach is a researcher at Inria, leading since 2011 the machine learning team which is part of the Computer Science department at École Normale Supérieure. He graduated from Ecole Polytechnique in 1997 and completed his Ph.D. in Computer Science at U.C. Berkeley in 2005, working with Professor Michael Jordan. He spent two years in the Mathematical Morphology group at Ecole des Mines de Paris, then he joined the computer vision project-team at Inria/ École Normale Supérieure from 2007 to 2010. Francis Bach is primarily interested in machine learning, and especially in sparse methods, kernel-based learning, large-scale optimization, computer vision and signal processing. He obtained in 2009 a Starting Grant and in 2016 a Consolidator Grant from the European Research Council, and received the Inria young researcher prize in 2012, the ICML test-of-time award in 2014, as well as the Lagrange prize in continuous optimization in 2018, and the Jean-Jacques Moreau prize in 2019. He was elected in 2020 at the French Academy of Sciences. In 2015, he was program co-chair of the International Conference in Machine learning (ICML), and general chair in 2018; he is now co-editor-in-chief of the Journal of Machine Learning Research.

On the convergence of gradient descent for wide two-layer neural networks

Abstract

Many supervised learning methods are naturally cast as optimization problems. For prediction models which are linear in their parameters, this often leads to convex problems for which many guarantees exist. Models which are non-linear in their parameters such as neural networks lead to non-convex optimization problems for which guarantees are harder to obtain. In this talk, I will consider two-layer neural networks with homogeneous activation functions where the number of hidden neurons tends to infinity, and show how qualitative convergence guarantees may be derived. I will also highlight open problems related to the quantitative behavior of gradient descent for such models. (Joint work with Lénaïc Chizat) ■

Brice Ménard, PhD

Associate Professor, Department of Physics & Astronomy

Brice Ménard joined the faculty at Johns Hopkins in 2010. He received his Ph. D. from both the Institut d'Astrophysique de Paris and the Max Planck Institute for Astrophysics in Germany. He was a postdoctoral member of the Institute for Advanced Study in Princeton and a senior research associate at the Canadian Institute for Theoretical Astrophysics in Toronto. His research combines astrophysics and statistics. His work has led to the detection of gravitational magnification by dark matter around galaxies, the discovery of tiny grains of dust in the intergalactic space, a new technique to estimate the redshift (or distance) of extragalactic objects and a new way to automatically find trends in complex datasets.

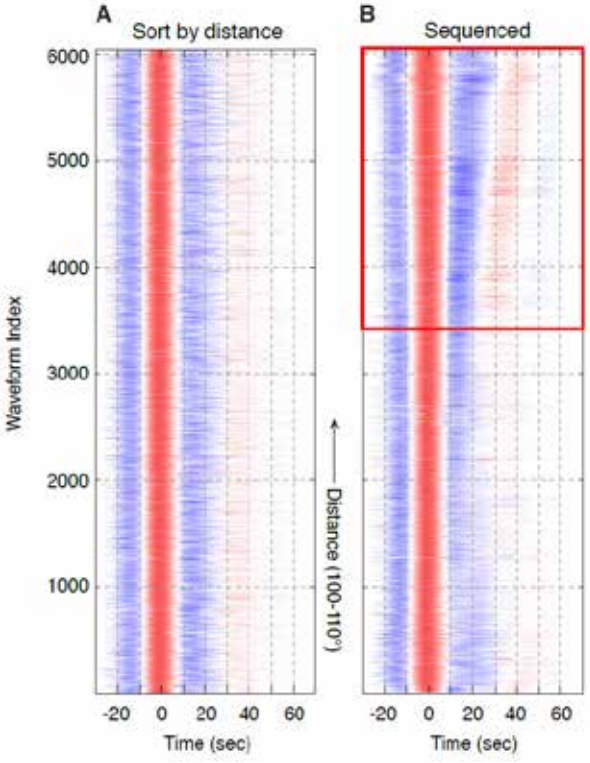
Ménard is a joint member of the Kavli Institute for Physics and Mathematics at Tokyo University. He received the Johns Hopkins President Frontier award (2019), the Packard fellowship for Science and Engineering (2014), the Sloan Research fellowship (2012), was named the 2012 Outstanding Young Scientist of Maryland, and was awarded the 2011 Henri Chrétien grant award by the American Astronomical Society.

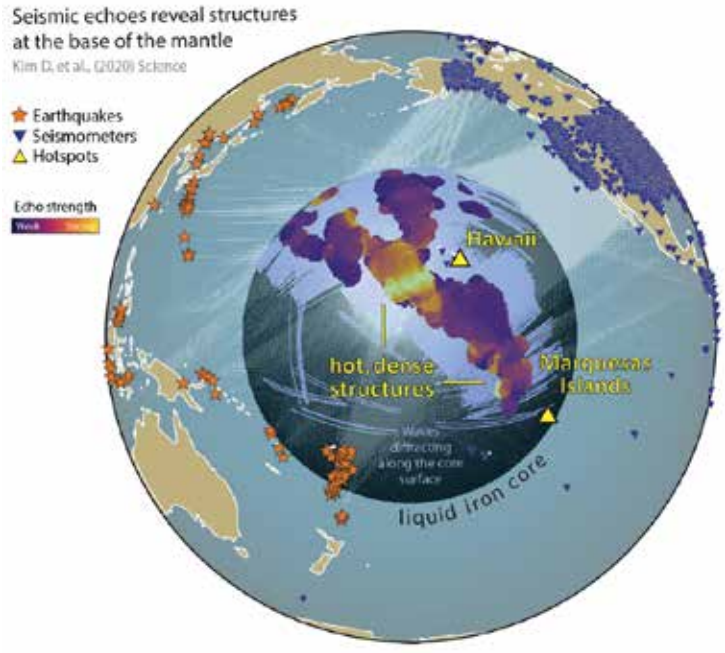
The Sequencer – how to reveal the main trend in your dataset?

Abstract

Scientists aim to extract simplicity from observations of the complex world. With data at hand, they look for trends but this process typically is more an art than a science. I will show how to think about this problem generically and present a new tool, the Sequencer (http://sequencer.org), which is able to automatically find trends in arbitrary datasets, without any training. I will then present some of its recent discoveries in astronomy and geology. ■

A: a set of 6,000 source-deconvolved seismograms from deep Earthquakes. When sorted by distance they do not show any trend.
B: same data re-ordered by the Sequencer and showing a population with delayed echoes. They reveal hot/dense structures at the core-mantle boundary, as shown in the adjacent figure. earth





Nikhil Paliwal, PhD

Postdoctoral Fellow in the Alliance for Cardiovascular Diagnostic and Treatment Innovation (ADVANCE) at JHU

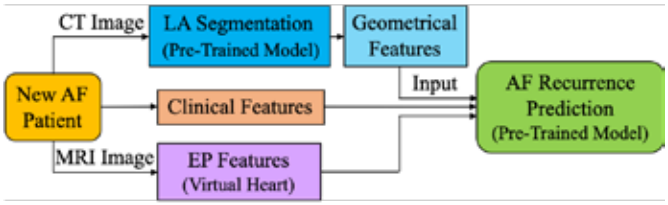
Nik is a postdoctoral fellow in the Alliance for Cardiovascular Diagnostic and Treatment Innovation (ADVANCE) at the Johns Hopkins University. He completed his PhD in Mechanical Engineering from the State University of New York at Buffalo (2019) and B. Tech in Bioengineering from the Indian Institute of Technology Kanpur, India (2011). His research interests include use of computational modeling and artificial intelligence to explore and predict clinical outcomes of cardiovascular diseases. His current work involves predictions of stroke risk and recurrence in patients with atrial fibrillation after catheter ablation, using electrophysiological and hemodynamic simulations along with machine learning.

SEED Update

An artificial intelligence approach towards predicting recurrence of atrial fibrillation in patients undergoing pulmonary vein isolation

Atrial fibrillation (AF) is the most common cardiac rhythm disorder, in which the atrium beats in an abnormal manner. Pulmonary vein isolation (PVI) of the left atrium (LA) has emerged as a radical alternative technology over traditional antiarrhythmic drug therapy for the treatment of AF patients. However, the long-term success of PVI is modest at best, with recurrence rates of ~50-60% after the first catheter ablation (CA) procedure. We proposed a data-driven technology to enable a priori prediction of the success of PVI for AF patients, which will maximize the benefit of PVI while minimizing the financial costs and procedural risks of unnecessary ablation procedures. As shown in Figure 1, our data-driven approach takes leverage from AF patient clinical data, and additional features like LA geometry and results from virtual heart simulation of AF induction to develop machine learning (ML) based algorithm for predicting the success of PVI after CA. To achieve this goal, we have progressed on two parallel fronts: (1) using ML and mathematical methodologies to develop tools for efficient and objective LA geometrical feature extraction, and (2) understand if clinical features alone can provide accurate outcome predictions, and whether simulation-based features provide additional insights and higher predictive accuracy for AF recurrence prediction after PVI.

On the tool development front, we have worked on efficient and accurate image segmentation and objective LA geometrical feature-extraction tools. For automated segmentation of cardiac images, we trained a convolutional neural network on pre-segmented cardiac CT and LGE-MRI images and tested its performance on a different cohort. Our results show that it accurately predicts the LA segmentation within 2 minutes, which takes hours when done manually. Furthermore, to remove the variabilities in the LA geometrical measurement, we are developing an objective LA geometrical assessment tool that takes in the automatically segmented LA geometry and outputs LA geometrical features like LA size, volume, and pulmonary veins and LA appendage diameter.



Workflow of the AF recurrence prediction tool

Concurrently, to identify which features are critical for accurate outcome predictions, we have performed two separate AF recurrence prediction studies including different sets of input features: (1) image and simulation-based features and (2) clinical features alone. In the first proof-of-concept study on 32 AF patients, we have demonstrated that prediction model that includes simulation-based AF induction features have better accuracy as compared to image-based features, and the combination of both image-based and simulation-based features had the highest prediction for success of PVI for AF patients (AUC=0.89). To test whether clinical features alone can predict the success of PVI, we performed another study on 300 AF patients that underwent PVI. ML-based classifiers using 50 clinical features, including clinically measured LA geometrical features, had modest AF recurrence prediction accuracy. Results from both studies suggest that even for a small cohort of 32 patients, ML-based models trained using simulation-based features were able to achieve high predictive accuracy, whereas preliminary ML model trained using only clinical features for 300 patients did not achieve high accuracy.

In the next part of our project, we plan to use our novel image analysis tools to efficiently and accurately process the cardiac images of the 300 AF patients, and extract their LA geometrical features. For patients with LGE-MRI imaging, we will perform computational simulations on LA models to obtain AF induction-based features. Then, we will train ML classifier algorithms using different combinations of the features (clinical, image, LA geometrical and AF induction) to identify which features are important for accurate recurrence prediction after PVI in AF patients. Finally, we will develop an optimized predictive model that gives the highest performance using the best combination of the input features. ■

Sham Kakade, PhD

Professor, Department of Computer Science and the Department of Statistics, University of Washington

Sham Kakade is a Washington Research Foundation Data Science Chair, with a joint appointment in both the Allen School and Department of Statistics at the University of Washington. He works on the theoretical foundations of machine learning, focusing on designing (and implementing) statistically and computationally efficient algorithms. Amongst his contributions with a diverse set of collaborators are: establishing principled approaches in reinforcement (including the natural policy gradient, conservative policy iteration, and the PAC-MDP framework); optimal algorithms in stochastic and non-stochastic multi-armed bandit problems (including the linear bandit and the Gaussian process bandit); computationally and statistically efficient spectral algorithms for estimation of latent variable models (including estimation of mixture of Gaussians, latent Dirichlet allocation, hidden Markov models, and overlapping communities in social networks); faster algorithms for large scale convex and nonconvex optimization. He is the recipient of the IBM Goldberg best paper award (in 2007) for contributions to fast nearest neighbor search and the best paper, INFORMS Revenue Management and Pricing Section Prize (2014). He has been program chair for COLT 2011.

Sham completed his Ph.D. at the Gatsby Computational Neuroscience Unit at University College London, under the supervision of Peter Dayan, and he was a postdoc at the University of Pennsylvania, under the supervision of Michael Kearns. Sham was an undergraduate at Caltech, studying in physics under the supervision of John Preskill. Sham has been a Principal Researcher at Microsoft Research, New England; an associate professor at the Department of Statistics, Wharton, UPenn; and an assistant professor at the Toyota Technological Institute at Chicago.

What are the statistical limits of offline reinforcement learning with function approximation?

Abstract

The area of offline reinforcement learning seeks to utilize offline (observational) data to guide the learning of (causal) sequential decision making strategies. It is becoming increasingly important to numerous areas of science, engineering, and technology. Here, the hope is that function approximation methods (to deal with the curse of dimensionality) coupled with offline reinforcement learning strategies can provide a means to help alleviate the excessive sample complexity burden in modern sequential decision making problems. However, the extent to which this broader approach can be effective is not well understood, where the literature largely consists of sufficient conditions.

This talk will focus on the basic question of what are necessary representational and distributional conditions that permit provable sample-efficient offline reinforcement learning? Perhaps surprisingly, our main result shows even if: i) we have realizability in that the true value function of _all_ target policies are linear in a given set of features and 2) our off-policy data has good coverage over all these features (in a precisely defined and strong sense), any algorithm information-theoretically still requires an exponential number of offline samples to non-trivially estimate the value of the _any_ target policy. Our results highlight that sample-efficient, offline RL is simply not possible unless significantly stronger conditions hold: such as either having extremely low distribution shift (where the offline data distribution is close to the distribution of the target) or far stronger representational conditions must hold (beyond realizability). ■

Gianluca Iaccarino, PhD

Professor of Mechanical Engineering
Director of the Institute for Computational Mathematics, Stanford University
Co-Founder of Cascade Technologies Inc.

Gianluca Iaccarino is a professor in Mechanical Engineering and the Director of the Institute for Computational Mathematical Engineering (https://icme.stanford.edu).

He received his PhD in Italy from the Politecnico di Bari in 2005 and worked for several years at the Center for Turbulence Research (NASA Ames & Stanford) before joining the faculty at Stanford in 2007.

Since 2014 he has been the Director of the PSAAP Center at Stanford, funded by the US Department of Energy: a $20M research Center focused on multiphysics simulations, uncertainty quantification and exascale computing (http://exascale.stanford.edu).

In 2010, he received the Presidential Early Career Award for Scientists and Engineers (PECASE) award from President Obama. In the last couple of years, he has received best paper awards from AIAA, ASME IMECE and Turbo Expo Conferences.

Over the years, his interests in research and teaching have touched many topics, but always revolved around the use of computing and data to solve problems in energy, biomedicine, aerodynamics, propulsion, design.

Engineering Simulations in the Age of Data

Abstract

95% of today's data has been created in the last three years!

Object recognition, health monitoring, language translation are only three examples of how society is

being transformed in the age of data. The exponential growth of interest in data is fueled by three elements (1) the introductions of new, inexpensive sensors and diagnostic technology, (2) the availability of high-quality, general-purpose and open source machine learning tools and (3) the development of novel computer architectures tailored for handling data tasks.

How science and technology innovation will change in the age of the data is still somewhat unclear and the current impact is more subtle, partly because of existing practices based on well-established physical and mathematical principles. The talk will provide a historical perspective on the use of data in engineering simulations for validation, calibration and inference. This will be contrasted with new emerging trends in data science and specifically in merging mathematical models with data.

Many of the examples described emerge from the use of data science tools in simulations of turbulence. ■

Ben Haeffele, PhD

Research Faculty Member, MINDS, JHU
Research Faculty Member, Center for Imaging Science (CIS), JHU

Matthew Ippolito, PhD

Assistant Professor, School of Medicine, JHU

Machine Learning and Computer Vision for Malaria: Disentangling the in vivo Effects of Antimalarial Drugs using an Automated Malaria Microscopy Algorithm

This project is a collaboration between the School of Engineering and School of Medicine to develop machine learning-based computer vision algorithms to innovate new methods for antimalarial drug development. Malaria is due to the protozoan parasite *Plasmodium* spp. which causes potentially lethal infection through the invasion and remodeling of circulating red blood cells. The malaria parasite undergoes complex transformations in the human host within a 48-hour life cycle. Each transformation is associated with varying susceptibility to treatment, with some life-stages being impervious to the most commonly used antimalarial drugs.

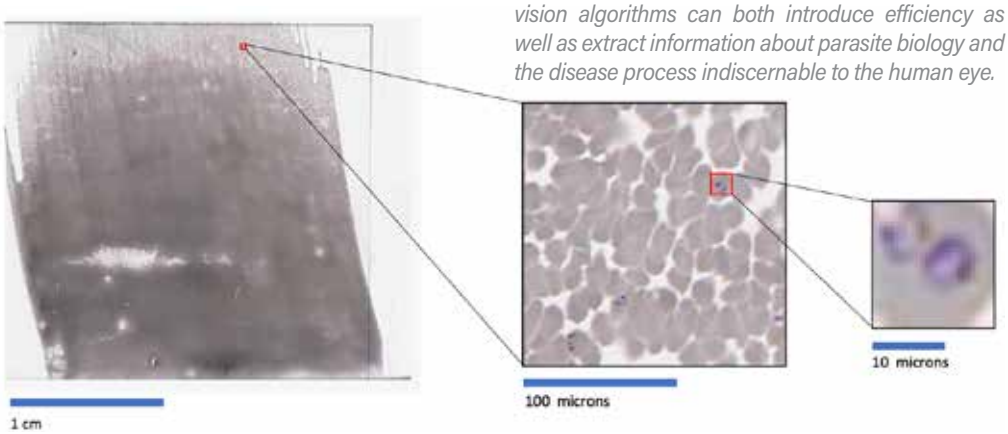
Malaria is notable among infectious diseases because we can witness its lifecycle unfold within a drop of blood. In contrast to HIV or TB, which conceal latent reservoirs of infection in parenchymal tissue observable only with ultrasensitive molecular methods, malaria in its various forms can be assessed directly in the peripheral blood. Routine blood smears can provide information about severity, prognosis, and response to therapy. However, manual readings are time consuming, error prone, and the human eye is insufficient to extract all of the clinically relevant information contained in a blood slide (Figure 1).

Our aim is therefore to develop a computer vision and machine learning-based platform for malaria microscopy to facilitate preclinical and clinical research into malaria treatment. With IDIES support, and in collaboration with the Johns Hopkins Malaria Research Institute, we have begun to develop and validate algorithms for the detection of malaria parasites in the early life phase, which is the least drug-susceptible parasite stage.

One of the key challenges of biomedical applications of computer vision is the lack of heavily annotated training data. For example, in this project one would like ground truth labels of the location of each parasite in a given image along with the corresponding life-stage of each parasite. To overcome these challenges we have begun by leveraging over 10,000 archival images from the NIH to train our computer vision-based system to identify parasites from blood slides prepared under simulated field conditions with ~85% sensitivity.

As we acquire additional data we anticipate the performance of our initial algorithm to improve. To construct ground-truth information for the parasite life-stage we have curated a novel initial training set of synchronized parasites cultured under laboratory conditions so that all of the parasites within a given image are known to be at roughly the same life-stage, which will inform ongoing algorithm development. For further validation and practical application, we will import malaria blood slides from a recently completed clinical trial of antimalarial drugs to cross-validate and augment manual evaluations of our algorithm to delineate drug effects beyond what is currently possible with manual assessments. The deliverables from this IDIES-supported project will provide critical support for external funding to continue development and application. ■

Figure 1. *A single drop of blood from a typical patient with malaria harbors on average 125,000 individual parasites. Conventional methods rely on manual enumeration by trained laboratory technicians examining blood smears, such as the one depicted above, at 1000x magnification. Machine learning-based computer vision algorithms can both introduce efficiency as well as extract information about parasite biology and the disease process indiscernable to the human eye.*



Jaime Combariza, PhD

Associate Research Scientist
Director, Maryland Advanced Research Computing Center (MARCC)

Rockfish: the Latest in High Performance Computing

Advanced computing, integrating traditional High Performance Computing (HPC), Data Intensive (DI), and Artificial Intelligence (AI, machine and deep leaning), continues to grow rapidly due to increased data-driven research. Given ever escalating research demands, academic institutions need to stay at the forefront of new developments to provide adequate resources to advance computational research and remain competitive in this essential endeavor.

At MARCC, the original Bluecrab cluster (now 5.5 years old) is aging and lacks more recently developed features crucial for support of innovative research. Furthermore, critical components are beginning to fail, thus introducing anxiety and frustration to the user community. The Bluecrab will continue in production until June of 2022 within clear guidelines established regarding its reliability. As must all HPC centers, MARCC is implementing new models for sustainability to provide and maintain a state of the art facility. For an essential core facility of this magnitude, no single entity's financial infrastructure can be relied upon to support refresh cycles every three to five years. Thus arose the need for a communal support system.

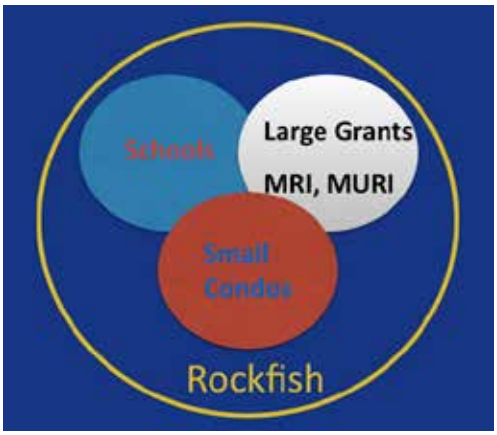


Figure 1. Proposed Rockfish sustainability Model.

A three pronged plan (as described in figure 1) relying on the contributions of separate yet interdependent support groups was proposed. The core facility is supported by the University through each of its schools, by research groups that join forces and apply for large grants (MRI, MURI, DURIP), and by individual researchers that add small condos to better serve their computational needs. The results are community-shared, guarantee sustainability and create a core facility with enough capacity to enable desired research.

With this in mind, MARCC , in collaboration with faculty members, is dedicated to several projects that will bring this plan to fruition. In 2019, MARCC was given an opportunity from the University to submit a Major Research Instrumentation (MRI) proposal to the National Science Foundation (NSF). JHU received notice in October 2019 that the grant would be funded and would be used to provide the foundation of a new cluster. Later in 2020, a large grant was submitted to the Defense University Research Instrumentation Program (DURIP). Notification of its successful funding is expected in the near future.

The institution has demonstrated its commitment to support faculty's HPC research needs. The Deans of KSAS and WSE planned contributions for a large condo that would provide additional resources for these schools. However, due to COVID-19 these plans are postponed but there are indications that a reduced contribution will result in smaller condos added in the near future. In addition, several individual PIs form the third group contributing to the new instrument. Thanks to their funding, 30 nodes (5 condos) will be added in the next few weeks. At this time there are also several PIs expecting grant confirmation to add more condos. If these plans continue as expected, we should have a very large system by early 2021 with over 25,000 cores.

The new system will be called *rockfish.jhu.edu* and will have three sets of compute nodes. A large number of standard compute nodes with 48 cores per node, 192GB RAM and a local NVMe SSD with 1 TB capacity. A small set of nodes will have 1.5 TB of memory and finally a small set of GPU nodes is planned, featuring the newest Nvidia technology (Ampera-100). Each node will have 4 GPUs. All nodes and storage are connected via 100gbps Infiniband. Rockfish will have a parallel file system with 13 PB of storage and will begin its mission to support advanced computing at Johns Hopkins University in December 2020. Rockfish will also support two other important groups: Morgan State University as a partner on the MRI, and the national community through the distribution of 20% of the computing resources via XSEDE.

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

Deanna Needell, PhD

Professor, Dunn Family Endowed Chair in Data Theory
Department of Mathematics, UCLA

Deanna Needell earned her PhD from UC Davis before working as a postdoctoral fellow at Stanford University. She is currently a full professor of mathematics at UCLA. She has earned many awards including the IEEE Best Young Author award, the Hottest paper in Applied and Computational Harmonic Analysis award, the Alfred P. Sloan fellowship, an NSF CAREER and NSF BIGDATA award, and the prestigious IMA prize in Applied Mathematics. She has been a research professor fellow at several top research institutes including the Mathematical Sciences Research Institute and Simons Institute in Berkeley. She also serves as associate editor for IEEE Signal Processing Letters, Linear Algebra and its Applications, the SIAM Journal on Imaging Sciences, and Transactions in Mathematics and its Applications as well as on the organizing committee for SIAM sessions and the Association for Women in Mathematics.

Online nonnegative matrix factorization for Markovian and other real data

Abstract

Online Matrix Factorization (OMF) is a fundamental tool for dictionary learning problems, giving an approximate representation of complex data sets in terms of a reduced number of extracted features. Convergence guarantees for most of the OMF algorithms in the literature assume independence between data matrices, and the case of dependent data streams remains largely unexplored. In this talk, we present results showing that a non-convex generalization of the well-known OMF algorithm for i.i.d. data converges almost surely to the set of critical points of the expected loss function, even when the data matrices are functions of some underlying Markov chain satisfying a mild mixing condition. As the main application, by combining online non-negative matrix factorization and a recent MCMC algorithm for sampling motifs from networks, we propose a novel framework of Network Dictionary Learning that extracts ` network dictionary patches' from a given network in an online manner that encodes main features of the network. We demonstrate this technique on real-world data and discuss recent extensions and variations. ■

Thomas Haine, PhD

Professor, Earth & Planetary Science, JHU

Thomas Haine's research interests are in ocean circulation and dynamics, and the ocean's role in climate. He is involved in improving estimates of the geophysical state of the ocean circulation through analysis of field data and circulation model results. He is particularly interested in the high latitude oceans, including the subpolar North Atlantic, Arctic, and Southern Oceans. He studies water-mass ventilation processes (rates, pathways, variability, and mechanisms), three-dimensional circulation, and geophysical fluid dynamics. He also investigates key physical processes that maintain the state of the extra-tropical upper ocean focusing on fluid dynamics and thermodynamics and their role in controlling sea surface temperature variability over years to decades.

Current research in Prof. Haine's group addresses: Arctic/North Atlantic dynamics, geophysical fluid dynamics, computational oceanography, passive tracer and transport diagnostics, and pedagogical oceanography.

He joined Johns Hopkins in 2000.

Towards the Development of Scale-Dependent, Non-Local, Turbulent Closures in Rotating Stratified Flows

SEED Update

The ocean circulation is characterized by large scale (> 100km) currents with a rich, unsteady, turbulent eddy field. This ocean "weather" disperses tracers at shorter horizontal scales, like salinity and plankton. The resulting complex spatial patterns dramatically affect the ocean circulation and ecosystems [Figure 1]. Models used to project global anthropogenic climate change cannot resolve these effects in order to integrate over years and decades. Thus, the interplay between mean currents, eddies, and small-scale turbulence must be represented by artificial parameterizations in climate models. A grand challenge in climate modeling is to accurately represent tracer dispersion (turbulent tracer transport) in the ocean. There are enormous implications for projections of anthropogenic climate change of ocean ecosystems. Early results from our seed fund project suggest that reconsidering a classic fluid dispersion problem can illuminate the way ahead.

Figure 1. Surface ocean "weather" in the Tasman Sea shown by a satellite image of chlorophyll. Colors reveal mesoscale eddies (> 10 km diameter), surrounded by submesoscale fronts, filaments, and vortices (< 10km). Taken from NASA Goddard Space Flight Center Gallery.

One class of artificial parameterization represents the effects of these unresolved currents as irreversible mixing. Most commonly, it is treated as isotropic down-gradient diffusion following the 19th century law due to Adolf Fick. Realistic and ideal simulations of ocean circulation show that this diffusion is inhomogeneous, meaning that the diffusion magnitude varies from place to place, and anisotropic, meaning that a preferred diffusion direction exists. A canonical example is a sheared current in a straight channel with the flow direction alternating across the channel [Figure 2]. Such shear flows are ubiquitous in the ocean, for example at the edges of the mesoscale eddies in Figure 1. In typical cases, the tracer averaged across the channel disperses much more rapidly along the channel than it would without the alternating currents, a phenomenon known as shear dispersion. Shear dispersion was characterized by G. I. Taylor in the 1950s and exemplifies anisotropic tracer transport because the spread of tracer along the channel is enhanced by the currents.

Evidence exists that the Fickian diffusion paradigm sometimes fails. For example, the parametrization of shear dispersion as anisotropic diffusive mixing can be qualitatively wrong. Specifically, the diffu-

continued on p.14

cont.

Towards the Development of Scale-Dependent, Non-Local, Turbulent Closures in Rotating Stratified Flows

sive parametrization exaggerates the speed at which a tracer blob initially spreads, although it gives an accurate estimate for the eventual spread. This initial dispersion bias derives from the mathematical form of Fick's law, which only considers information that is local in space. To fix the initial dispersion bias a non-local mathematical form is needed. This non-local form is a hybrid mathematical operator that acts like irreversible advection at short times and reversible mixing at long times.

Our seed fund project is exploring and applying these ideas. Work to date has focused on the theory of the shear dispersion problem. Preliminary results by postdoc Miguel Jimenez-Urias suggest that an exact analytical solution is available, which advances Taylor's work of seventy years ago. The analytical solution exposes the novel hybrid mathematical dispersion operator, which is important because it quantifies the transition between short-time advective and long-time diffusive shear dispersion. It also provides a theoretical path to studying related problems, for example that replace Taylor's cross-channel average with more flexible space-time filters needed for ocean climate models.

Our future work seeks to develop and apply computational methods to diagnose the novel mathematical operator in the shear dispersion problem. We intend to use a promising new development in turbulence modeling called the Macroscopic Forcing Method (MFM, Mani & Park, 2019). The MFM has the advantage that the dispersion operator is directly obtained from the governing equation, as opposed to inferring it from empirical fits. No simplifying assumptions of isotropy or separation of scales are required. The MFM quantifies the degree of non-locality and anisotropy of the operator and it remains uniformly valid over a wide range of scales, unlike other approaches.

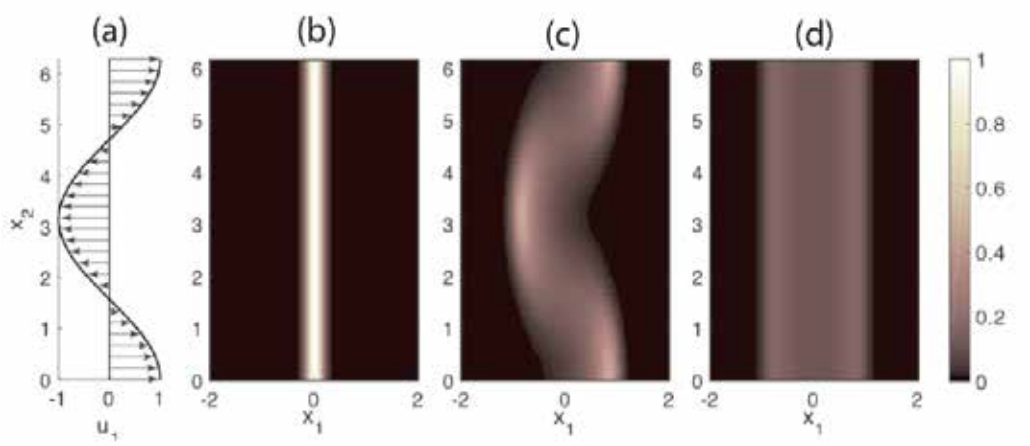


Figure 2. Shear dispersion of tracer in alternating flow in a long straight channel. (a) Flow speed along the channel. (b) Tracer blob at the initial time and (c) after one characteristic time unit. (d) Across-channel average tracer field corresponding to (c). Shear dispersion enhances tracer spread along the channel by the alternating flow. The challenge is to quantify the time evolution of the cross-channel average tracer field. From Mani & Park (2019).

The MFM method requires a set of Direct Numerical Simulations (DNS) of the flow under study, which resolve the space and time scales up to the dissipative scales. We will apply the MFM method to DNS of shear dispersion. The exact analytical solution from the first stage of research will be used to check the computational results and calibrate the MFM. We will then apply the MFM to a canonical ocean dynamics problem that spans the mesoscale to submesoscale transition seen in Figure 1. Performing the DNS for this case is computationally expensive so we will perform the calculations on the new MRI Rockfish MARCC cluster. We will analyze the simulations on the SciServer using OceanSpy tools, which were developed by an earlier IDIES Seed Grant (Almansi et al., 2019).

The long term goal is to expand these prototypical approaches to other flows. We envisage a sequence of canonical ocean circulation solutions of increasing complexity and a set of theoretical and computational tools to diagnose their tracer dispersion properties. ■

Winston Timp, PhD

Assistant Professor, Biomedical Engineering, JHU

Overview:

Understanding the genetic variability of virus genomes informs many public health and basic research priorities, such as the impact on social distancing and insight into vaccine development. Genetic characterization of viral pathogens requires specialty expertise from many disciplines, including molecular biology, virology, computational biology, and epidemiology. By combining efforts across JHU, we are establishing a pipeline to sequence virus using the ARTIC method for broad genetic characterization of SARS-COV-2, the virus which causes COVID-19. Specifically, we are: 1) Working to identify the the genetic diversity of the virus in Baltimore/DC, what viruses are part of our initial introduction, and how does this change over time and with distancing measures. 2) Measuring strains in individuals and how this correlates to disease state/virulence, an individual's immune response, and pharmacological resistance and 3) How strains move among the population - i.e. what is the flow from broad to fine scale: (nation, state, city, neighborhood) using sequence evolution to identify and track community transmission if possible.

Sciserver:

IDIES is helping to house, organize and analyze the data, making it accessible on the back-end to computational analysis and on the front-end through web servers to our group and other researchers throughout the JHU family (JHU, SPH, APL).

Our preprint on this is here:

Genomic Diversity of SARS-CoV-2 During Early Introduction into the United States National Capital Region

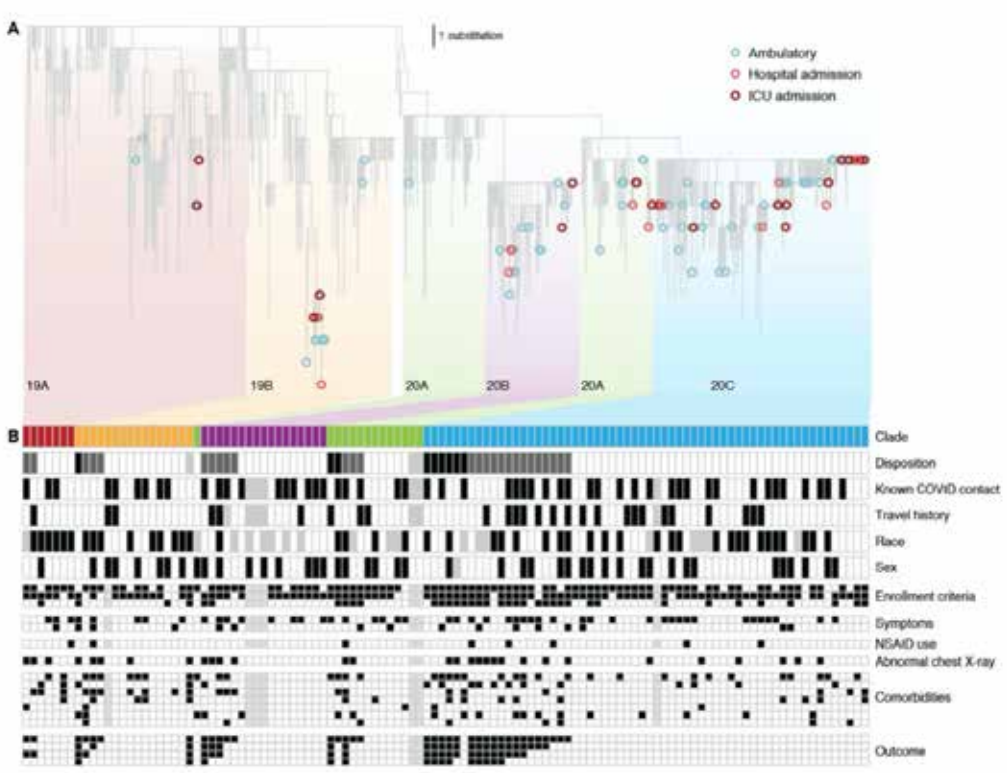


Figure 3: JHHS sequences and patient outcome. (A) Maximum likelihood tree of subsampled SARS-CoV-2 global dataset and all 114 sequences generated in this study. Ambulatory (blue) includes all patients with no known admission to the hospital. Hospital admission (light red) includes admitted patients with no known admission to the intensive care unit (ICU), including patients administered oxygen. (B) Clinical metadata and virus clade. Each column is one of the 114 patients with virus sequenced in this study, and columns are grouped by disposition within each clade. Unless otherwise specified: black = yes, white = no, gray = unknown. Disposition: black = still in hospital or deceased as of 15-May-2020, dark gray = discharged, white = never admitted. Race: black = Black, white = White, gray = other. Sex: black = female, white = male. Enrollment criteria (top down): Fever, cough, shortness of breath. Symptoms (top down): body ache, GI. Comorbidities (top down): Cardiac disease, Lung disease, diabetes, obese, alcohol, history of smoking (current and former smokers), immunocompromised. Outcome (top down): hospital admission, supplementary oxygen, ICU admission, ventilator administration.

This years poster session was open to the greater Johns Hopkins community, including the University, Hospital, Healthcare, Space Telescope, APL, and Peabody student community.

We would like to thank the following participants:

Laszlo Dobos
Carrie Filion
Zherui Xuan
Ambar Pal
Nick Carey
Coco Cai
Vedant Chandra
Gina El Nesr
Peter Gu
Brian Broll

A copy of their posters will be available on the IDIES and MINDS websites after the Symposium has concluded.

There will be breakout sessions during the lunch break to discuss poster content and ask any questions of the poster session participants.

JOHNS HOPKINS
UNIVERSITY

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

idies.jhu.edu minds.jhu.edu