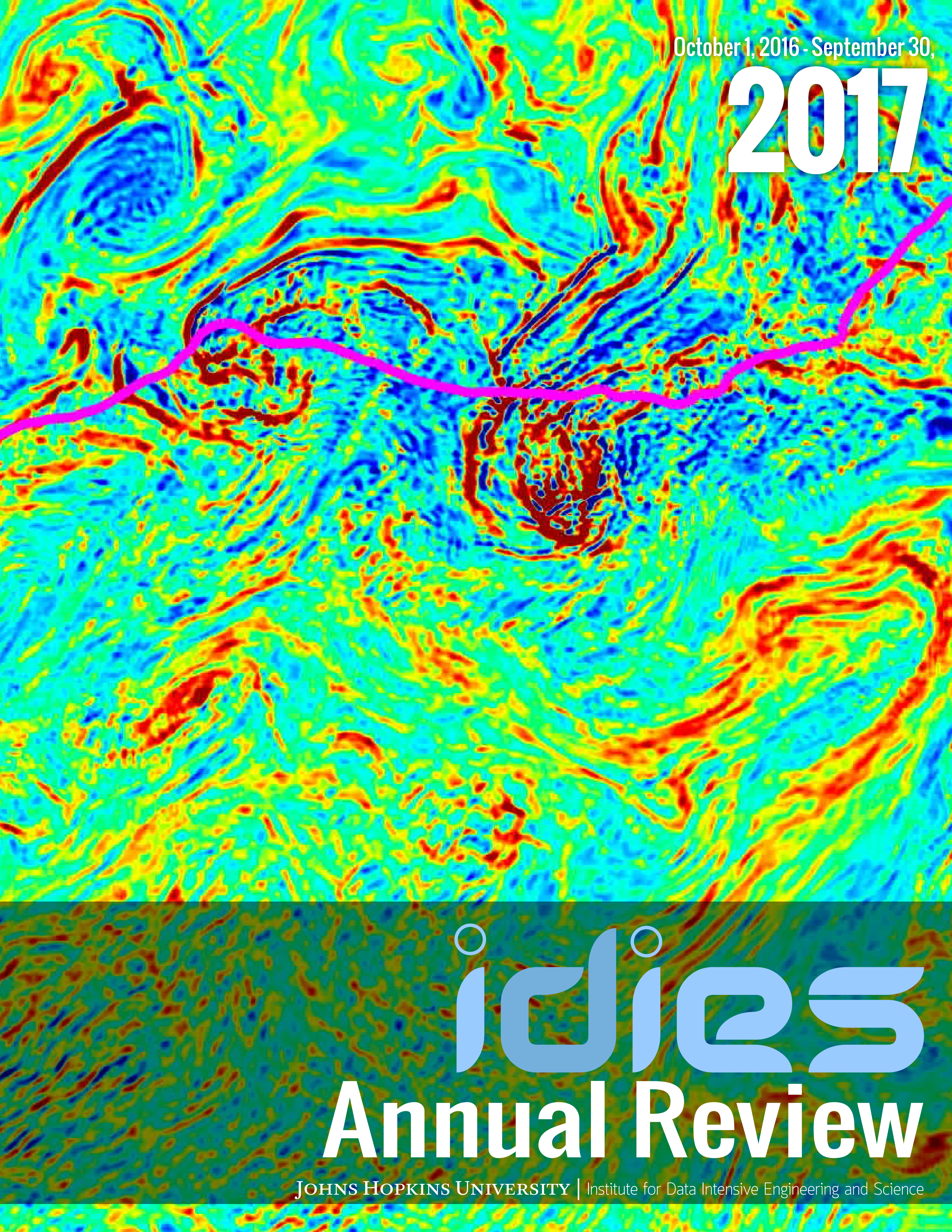


October 1, 2016 - September 30,

2017

idies

Annual Review

Johns Hopkins University | Institute for Data Intensive Engineering and Science

MESSAGE FROM THE DIRECTOR

Recognizing the strategic importance of computing and data across the whole university, in 2013 President Daniels decided to substantially broaden the scope of IDIES. The Institute included five schools, the Krieger School of Arts and Sciences, the Whiting School of Engineering, the Sheridan Libraries, the School of Medicine and the Bloomberg School of Public Health. Now I am happy to report that the Carey Business School has also decided to join IDIES.

In Sept 2015, after several years of hard work, we have opened the Maryland Advanced Research Computing Center (MARCC). The MARCC is a new, world class research computing facility on the Johns Hopkins Bayview campus, partnering with the University of Maryland College Park, supported by a $30M grant from the State of Maryland. The system has 22,000 cores, a 100G connectivity to Internet2, with a similar campus backbone and 10-40G uplinks into the individual buildings. MARCC is now almost saturated and we need to start looking for further opportunities for new enhancements and additions.

Today IDIES involves more than 100 faculty and a similar number of graduate students. As every year, we have awarded four seed grants in a broad spectrum of topics, connecting researchers from different fields, but sharing a common interest in Big Data. Also, over the last twelve months, we have dramatically increased our involvement in projects at JHMI, and SPH. Our involvement in material science is also growing, we are working with HEMI and projects like Paradigm to increase the ability of researchers to take data and analyze them much easier than before.

Our collaboration with the Bloomberg-Kimmel Center for Cancer Immunotherapy has been one of the more recent additions to our collaborative research. We are trying to scale up the amount of data collected about cancer cells by a factor of a 1,000!

In four years we have come a long way, now we have a major interdisciplinary program, a large, diverse effort, where faculty and students work together to solve amazing data-intensive problems, from genes to galaxies, starting new projects in materials science and urban planning, in collaboration with the City of Baltimore. Our members have successfully collaborated on many proposals related to Big Data, and we have hired several new faculty members, all working on different aspects of data-driven discoveries.

Our next challenge is to put the Institute on a path that leads to long-term sustainability.

Over the last few years IDIES had a much broader impact on JHU as a whole than we ever hoped for. It has turned into one of the great examples of what it means to be One University!

October 1, 2016 - September 30, 2017

The Institute for Data Intensive Engineering and Science ANNUAL REVIEW

SYMPOSIUM

NEWS

IDIES

On the cover: Using the HHPCv2 and MARCC, IDIES affiliate Thomas Haine and colleague Marcello Magaldi generated this simulation of the surface vertical vorticity field in the Denmark Strait. Circulation is driven in part by formation of dense waters on the Southeast Greenland Continental shelf (upper left of back cover). See the article on page 8 for the full story. From Magaldi & Haine, 2015.

AGENDA

IDIES Annual Symposium
October 20, 2017

8:00 a.m.	***Continental breakfast & Check-in***
9:00 a.m.	OPENING REMARKS **S. Alexander Szalay, PhD**, Director of IDIES, Bloomberg Distinguished Professor, Professor of Astrophysics & Computer Science, Johns Hopkins University
9:15 a.m.	KEYNOTE SPEAKER AWS Research Initiatives **Sanjay Padhi, PhD**, Amazon Web Services Research Initiatives
10:05 a.m.	SEED FUND UPDATE Enter the matrix: Interpreting unsupervised feature learning with matrix decomposition to answer unasked questions in high-throughput data **Elana Fertig, PhD**, Assistant Professor, Department of Oncology, School of Medicine, Johns Hopkins University
10:25 a.m.	SEED FUND UPDATE New Tools for an Old Problem: Building a Global and Historical Data Set of Social Unrest **Sahan Karatasli, PhD**, Assistant Research Scientist, Department of Sociology, Arrighi Center for Global Studies, Krieger School or Arts and Sciences, Johns Hopkins University
10:45 a.m.	***Break – Mudd Auditorium Lobby***
11:00 a.m.	MARCC Update **Jaime Combariza, PhD**, Associate Research Professor, Department of Chemistry, Krieger School of Arts & Sciences, Johns Hopkins University
11:20 a.m.	IDIES Proposal Submission Service **Margie Gier**, Senior Administrative Manager, Institute for Data Intensive Engineering and Science
11:25 a.m.	The MEDE Data Science Cloud: Building a Bridge to Materials Science **David Elbert,** Associate Research Scientist, Department of Earth and Planetary Sciences, Krieger School of Arts & Sciences, Johns Hopkins University
11:45 a.m.	Poster Madness
12:00 p.m.	***Lunch & Poster Session – Mudd Hall UTL Commons***
1:15 p.m.	KEYNOTE SPEAKER When Big Data is Not Enough **Alan Yuille, PhD**, Bloomberg Distinguished Professor, Department of Cognitive Science, Krieger School of Arts & Sciences, Johns Hopkins University
2:05 p.m.	SEED FUND UPDATE EchoSIM: Multiscale Acoustic Simulations Integrated with Free-Flight Experiments for Echo Scene Analysis of an Echolocating Bat **Rajat Mittal, PhD**, Professor, Department of Mechanical Engineering, Whiting School of Engineering, Johns Hopkins University
2:25 p.m.	***Break – Mudd Hall Auditorium Lobby***
2:40 p.m.	KEYNOTE SPEAKER Big Scientific Data and Data Science **Tony Hey, PhD**, Chief Data Scientist, Science & Technology Facilities, UK
3:30 p.m.	SEED FUND UPDATE Integrating environmental genomics into biogeochemical models **Sarah Preheim, PhD**, Department of Environmental Health & Engineering, Whiting School of Engineering, Johns Hopkins University
3:50 p.m.	CLOSING REMARKS **S. Alexander Szalay, PhD**, Director of IDIES
4:00 p.m.	***Cocktail Reception – Mudd Auditorium Lobby***

SPEAKERS

ALAN YUILLE, PhD KEYNOTE

Bloomberg Distinguished Professor, Department of Cognitive Science, Krieger School of Arts and Sciences

Dr. Yuille is a mathematician and computer scientist interested in the biology of vision. His research spans several disciplines including computer vision, vision science, and neuroscience. Alan Yuille received a BA degree in mathematics from the University of Cambridge in 1976. His PhD on theoretical physics, supervised by Prof. S.W. Hawking, was approved in 1981. He was a research scientist in the Artificial Intelligence Laboratory at MIT and the Division of Applied Sciences at Harvard University from 1982 to 1988. He served as an assistant and associate professor at Harvard until 1996. He was a senior research scientist at the Smith-Kettlewell Eye Research Institute from 1996 to 2002. He joined the University of California, Los Angeles, as a full professor with joint appointments in computer science, psychiatry, and psychology. He moved to Johns Hopkins University in January 2016 where he was appointed a Bloomberg Distinguished Professor. He holds joint appointments in the Departments of Cognitive Science and Computer Science. His research interests include computational models of vision, mathematical models of cognition, medical image analysis, and artificial intelligence and neural networks.

He directs the research group on Computational Cognition, Vision, and Learning (CCVL). He is affiliated with the Center for Brains, Minds and Machines, and the NSF Expedition in Computing, Visual Cortex on Silicon.

SANJAY PADHI, PhD KEYNOTE

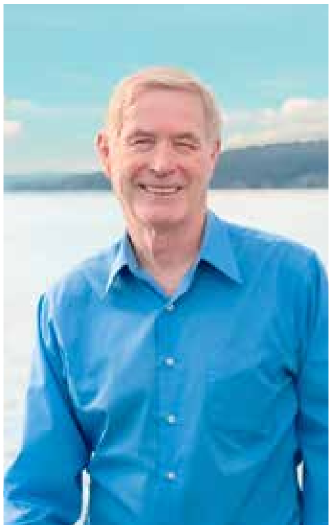
Amazon Web Services Research Initiative

Dr. Sanjay Padhi, leads the AWS Research Initiatives including AWS's federal initiatives with the National Science Foundation. Dr. Padhi has more than 15 years of experience in large-scale distributed computing, Data Analytics and Machine Learning. He is the co-creator of the Workload Management System, currently used for all the data processing and simulations activities by CMS, one of the largest experiments in the world at CERN, consisting of more than 180 institutions across 40 countries. He also co-founded the ZEUS Computing Grid project at Deutsches Elektronen-Synchrotron (DESY), Germany before joining CERN. Dr Padhi obtained his PhD from McGill University in High Energy Physics and is also currently appointed by the Dean of Faculty as an Adjunct Associate Professor of Physics at Brown University.

TONY HEY, PhD

KEYNOTE

Chief Data Scientist, Science and Technologies Facilities, UK



Tony Hey began his career as a theoretical physicist with a doctorate in particle physics from the University of Oxford in the UK. After a career in physics that included research positions at Caltech and CERN, and a professorship at the University of Southampton in England, he became interested in parallel computing and moved into computer science. In the 1980's he was one of the pioneers of distributed memory message-passing computing and co-wrote the first draft of the successful MPI message-passing standard.

After being both Head of Department and Dean of Engineering at Southampton, Tony Hey escaped to lead the U.K.'s ground-breaking 'eScience' initiative in 2001. He recognized the importance of Big Data for science and wrote one of the first papers on the 'Data Deluge' in 2003. He joined Microsoft in 2005 as a Vice President and was responsible for Microsoft's global university research engagements. He worked with Jim Gray and his multidisciplinary eScience research group and edited a tribute to Jim called 'The Fourth Paradigm: Data-Intensive Scientific Discovery.' Hey left Microsoft in 2014 and spent a year as a Senior Data Science Fellow at the eScience Institute at the University of Washington. He returned to the UK in November 2015 and is now Chief Data Scientist at the Science and Technology Facilities Council.

In 1987 Tony Hey was asked by Caltech Nobel physicist Richard Feynman to write up his 'Lectures on Computation'. This covered such unconventional topics as the thermodynamics of computing as well as an outline for a quantum computer. Feynman's introduction to the workings of a computer in terms of the actions of a 'dumb file clerk' was the inspiration for Tony Hey's attempt to write a popular book about computer science: 'The Computing Universe: A Journey through a Revolution'.

Tony Hey is a fellow of the ACM and of the AAAS in the US and also of the UK's Royal Academy of Engineering. In 2005, he was awarded a CBE by Prince Charles for his "services to science."

✦ ✦ ✦

ALEX SZALAY, PhD

Professor Szalay is the founding director of IDIES, a Bloomberg Distinguished Professor, Alumni Centennial Professor of Astronomy, 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.

JAIME COMBARIZA, PhD

Jaime Combariza is the director of the Maryland Advanced Research Computing Center, a shared computing facility of the University of Maryland and Johns Hopkins University. MARCC is funded by a State of Maryland grant to Johns Hopkins through IDIES.

DAVID ELBERT

David is an environmental geochemist and mineralogist with broad expertise. He's recently worked on methodology development for electron microscopy, synchrotron X-ray, and neutron scattering applications to toxic-metal, crystal chemistry.

ELANA FERTIG, PhD

The Fertig Lab pursues research in the systems biology of cancer and therapeutic response. Her group develops computational methods for pattern detection from genomics data. Elana and her group develop bioinformatics tools that integrate pathway structure in pattern detection algorithms for gene expression data.

SAHAN SAVAS KARATASLI, PhD

Sahan Savas Karatasli is an Assistant Research Scientist and Lecturer at the Department of Sociology and the Arrighi Center for Global Studies at the Johns Hopkins University. His work examines dynamics of social movements, nationalism and historical capitalism from a long historical and global perspective.

RAJAT MITTAL, PhD

Professor Mittal's research interests include computational fluid dynamics, low Reynolds number aerodynamics, biomedical flows, active flow control, LES/DNS, Immersed Boundary Methods, fluid dynamics of locomotion (swimming and flying), biomimetics and bioinspired engineering, and turbomachinery flows.

SARAH PREHEIM, PhD

Dr. Preheim researches the ecology of microorganisms impacting water quality in lakes, estuaries and coastal oceans to better inform remediation strategies. She uses a combination of field sampling, laboratory experiments and computational analysis to improve our understanding of the microbial processes that impact water quality.

NEWS

JHU Researchers Launch "recount," the Largest Summarized Collection of Human RNA Sequencing Data to Date

The **recount2** project, the largest summarized collection of human RNA sequencing data to date, was launched near the end of 2016. The project was primarily the work of Leonardo Collado Torres and Abhinav Nellore. The recount2 team spans several departments, with most of the work taking place in the labs of Ben Langmead (Computer Science), Jeff Leek (Biostatistics), Kasper Hansen (Institute for Genetic Medicine & Biostatistics) and Andrew Jaffe (The Lieber Institute for Brain Development). A summary paper for the project, entitled "Reproducible RNA-seq analysis using recount2" was released in Nature Biotechnology on April 11th, 2017.

Public sequencing data archives are growing by petabases — millions of billions of DNA letters — per year. But for typical researchers, using these data requires huge downloads and laborious re-analysis. Most researchers aren't equipped for this, so valuable data go unused. The JHU team, led by IDIES affiliate Ben Langmead (Computer Science) and Jeffrey Leek (Biostatistics), is working to make public data easier to use by choosing valuable subsets of the archive — human RNA sequencing data in this case — analyzing it, and summarizing it using a carefully crafted and uniform bioinformatics pipeline. The team then makes resources like **recount2** available to the community in a form that allows researchers to ask and answer sophisticated questions.

For this last step, **recount2** depends on the SciServer system and borrows from its philosophy. SciServer provides storage, hosting, and a sophisticated interface for computing and visualizing the data "on-site" at JHU, without having to download either raw or summarized data. This allows biological researchers to use powerful, flexible systems (Jupyter, R and Python) to interact with concise summaries of many public datasets.

✦ ✦ ✦

HIPPA-Capable Research Infrastructure at MARCC

Jaime Combariza (Director of MARCC)

The Maryland Advanced Research Computing Center I(MARCC) is deploying additional services that will make MARCC resources more attractive to researchers and at the same time facilitate research endeavors. Many researches have been asking for secure resources that are in compliance with the Health Insurance Portability and Accountability Act (HIPAA).

MARCC has implemented data security measures to protect privacy of research data according to JHU's policies and standards, applicable legal requirements and expected applicable safeguards under the HIPAA Privacy and Security Rules. The policies in place at the MARCC Secure Environment (MSE) strictly apply to all personnel who are involved in research endeavors dealing with Protected Health Information (PHI) data. MARCC also complies with JHU's minimum-security standards for systems with PHI, namely physical security, encryption of data at rest, and facility monitoring. It is critical that both parties, PIs and MARCC, understand and agree that handling sensitive data is a shared responsibility and that policies and procedures must be strictly followed to avoid potential damages and minimize risk. This MSE will be available at the end of October 2017 after the data trust committee grants authority to operate.

Tamas Budavari uses Big Data Astronomy Research Methods Aiming to Solve Entrenched Problems in Baltimore

Hopkins initiative aims to solve entrenched problems in Baltimore, other cities

Back in January, IDIES affiliate and JHU mathematician Tamas Budavari was featured in the Atlantic's CityLab blog detailing his research to help predict housing abandonment in the city of Baltimore. From CityLab, "Budavari and Phil Garboden, a doctoral student in sociology and applied math, are working on a statistical tool to predict abandonment. They're combining publicly available data with GIS technology to create a database of the city's housing stock. This will serve as a base to do high-level statistical analyses that can help officials make better, data-driven evaluations of current and future interventions."

Some of Baltimore's deserted row houses. *Image credit: Chuck Robinson, chuckrobinsonphoto.com.*

"Just like how galaxies cluster in the universe, houses also cluster in the city," Budavari told CityLab. "So if you have a vacant house in a given place, there's a higher probability of finding other ones next to it." Professor Budavari and Garboden are also looking to expand to other cities as well, already submitting grants for funding to apply their research methods to New Orleans and Kansas City as well."

But that's not all, later in July Dr. Budavari was once again recognized in the Baltimore Sun for the use of his research in helping the city's Department of Housing and Community Development identify dangerous vacant houses. From the Sun article:

"Tamas Budavari, an expert in cosmology and galaxy evolution, joined the city housing department to map abandoned buildings and predict future blight to help officials figure out where to intervene before a property became vacant.

When a house fell and killed a man in West Baltimore last year and four others crumbled amid powerful wind gusts, Budavari immediately responded to help the city identify which of Baltimore's thousands of vacants were in imminent risk of collapsing. He and researchers mined city databases to assess risk factors: the height of a building, whether the roof was missing and the number of years vacant. The data helped the city identify more than 300 houses that needed to be torn down right away."

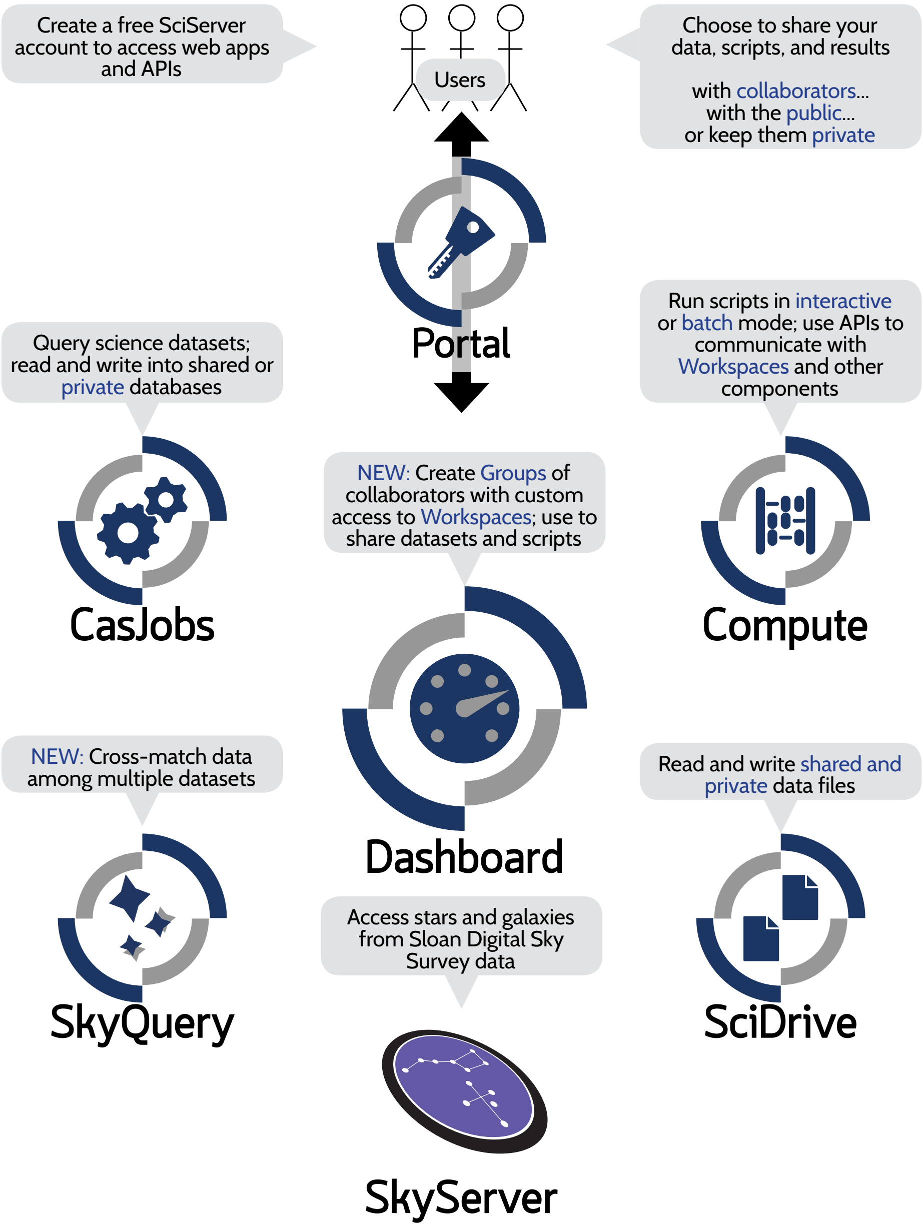
The article also goes on to mention how Dr. Budavari and the rest of the initiative's work is being shared with officials from cities such as Kansas City and New Orleans to fight similar problems. The initiative's director, Ben Seigel, says the program "is designed to bring together Hopkins students and faculty, government officials, policymakers, philanthropists, activists and others to talk about the problems, solutions and the potential to harness big data analytics." A symposium is being planned for December 2017.

The New SciServer: Bringing Collaboration Close to the Data

The next few months will bring several exciting changes to SciServer, our online environment for science research and education with big data. SciServer (www.sciserver.org) offers free access to big datasets online, alongside tools to visualize and analyze those datasets in synchronous or asynchronous mode using browsers or APIs. Our upcoming new features will enable new ways to share datasets, analyses, and publications with colleagues all over the world.

The heart of the new system will be the SciServer Dashboard, a single interface offering quick access to all your prior work across all components. From your dashboard, you will be able to work with all datasets you have access to – whether you own data or Petabyte-scale online datasets, whether stored as databases or as collections of files. You will also have easy access to the entire history of queries you have submitted to any of those datasets, and to any visualization or analysis scripts you have run through the SciServer Compute interface.

A second major feature is the inclusion of data access controls and “Workspaces,” which will provide an easy way to share datasets and scripts in a self-contained, collaborative environment. Mounting a Compute volume using a Workspace will allow you to share many heterogenous types of data in exactly the same way; your mounted volumes may be either read-only or writable. Our data access controls will make it easy for data providers to control who sees what data and who can perform what operations.



Create a free SciServer account to access web apps and APIs

Choose to share your data, scripts, and results

with collaborators...
with the public...
or keep them private

Query science datasets; read and write into shared or private databases

Run scripts in interactive or batch mode; use APIs to communicate with Workspaces and other components

NEW: Create Groups of collaborators with custom access to Workspaces; use to share datasets and scripts

NEW: Cross-match data among multiple datasets

Read and write shared and private data files

Access stars and galaxies from Sloan Digital Sky Survey data

A third major new feature is the ability to submit asynchronous Compute Jobs in “batch” mode. You can develop a Jupyter Notebook interactively, then submit it as a batch job to take advantage of significantly more hardware resources to support compute jobs that require lots of processing power and/or memory. This new feature also comes with extended support for API integration with other SciServer components.

In addition to this system’s obvious capabilities for collaborative science, it has the potential to revolutionize science education, particularly at the introductory college level. Students can learn science with real, modern scientific data – and they can also get hands-on experience with programming, a critical skill in today’s information economy. And they can do this all without needing to go

Continued on page 11

Non-Hydrostatic simulation of dense waters cascading off the East Greenland continental shelf

Marcello G. Magaldi[1] and Thomas W.N. Haine[2] (Istituto di Scienze Marine[1]; Department of Earth and Planetary Sciences, Johns Hopkins University[2])

Using the HHPCv2 and MARCC, IDIES affiliate Thomas Haine and colleague Marcello Magaldi generated this simulation of ocean circulation driven by formation of dense waters on the Southeast Greenland Continental shelf. The cover image shows part of the calculated surface vertical vorticity field from a Denmark Strait simulations run on HHPCv2 and MARCC, analyzed on DataScope and published in Magaldi & Haine (Deep Sea Res. I, 2015). On the cover figure, positive vorticity features (red) are Denmark Strait cyclones rotating counter-clockwise and moving southwestward. Negative vorticity features (blue) are anticyclones rotating clockwise and occurring mainly on the inner shelf. The high-resolution non-hydrostatic simulations show that Denmark Strait cyclones are made up of different vorticity filaments intertwined in strands. Anticyclones exhibit an internal structure with positive elongated filaments at their edges.

Figure 1 on the right shows the full simulation, including (c) sea surface temperature,(f) vorticity and (i) strain fields on 19 July, 1800 UTC. The yellow arrows in (c) show five cyclones moving along the shelf break. The 500 m isobath is in magenta in (f) and (i) for reference.

Marcello G. Magaldi, Thomas W.N. Haine, Hydrostatic and non-hydrostatic simulations of dense waters cascading off a shelf: The East Greenland case, Deep Sea Research Part I: Oceanographic Research Papers, Volume 96, February 2015, Pages 89-104, ISSN 0967-0637, https://doi.org/10.1016/j.dsr.2014.10.008. *http://www.sciencedirect.com/science/article/pii/S0967063714001915*

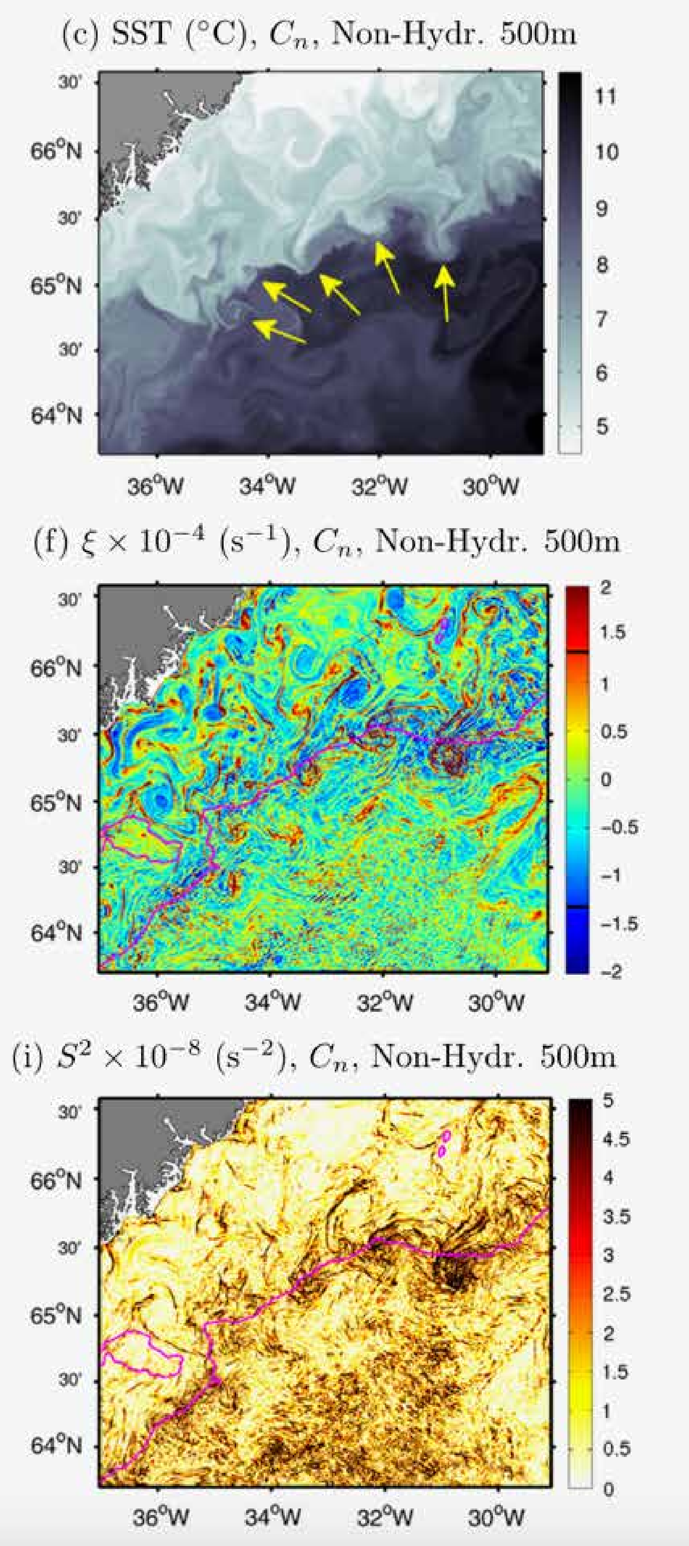

(c) SST (°C), C_n, Non-Hydr. 500m
30'
66°N
30'
65°N
30'
64°N
36°W
34°W
32°W
30°W
11
10
9
8
7
6
5
(f) $\xi \times 10^{-4}$ (s^{-1}), C_n, Non-Hydr. 500m
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
(i) $S^2 \times 10^{-8}$ (s^{-2}), C_n, Non-Hydr. 500m
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0

Figure 1

Simulating Gravitational Lensing with CUDA and OpenGL

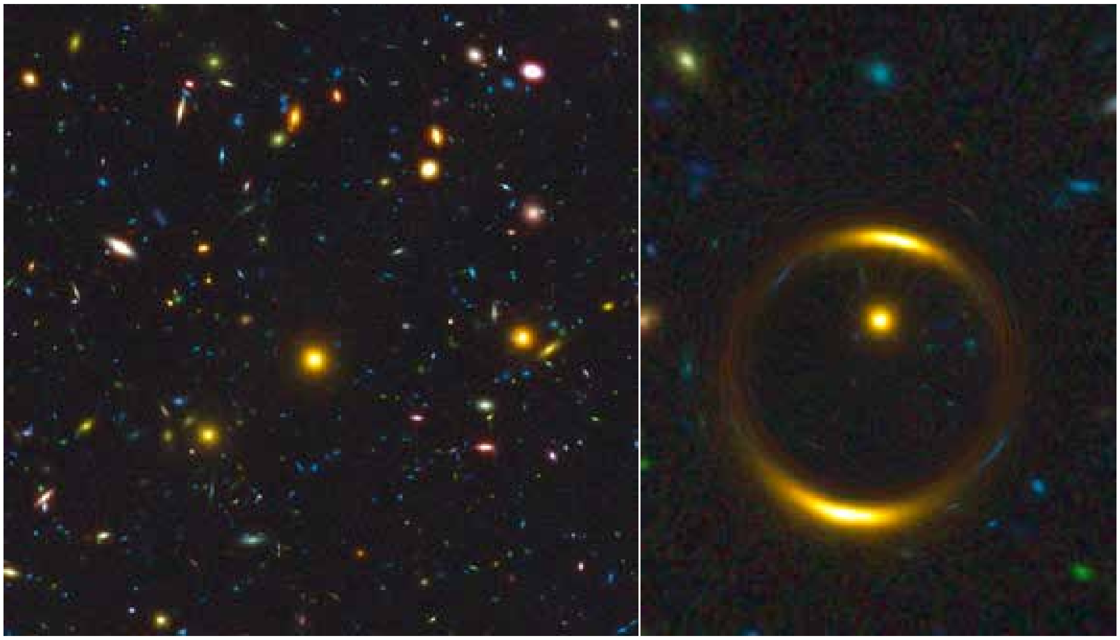
Adharsh Babu[1], Alex Szalay[2], and Gerard Lemson[2]
(University of Maryland, College Park[1], Department of Physics and Astronomy[2])

Knowing that gravitational lensing must exist is not the same as knowing how to recognize it. Calculations required to predict the effects of gravitational lensing are prohibitively time-consuming using traditional techniques. We have developed and applied a technique to accurately and rapidly calculate the gravitational potential effect using CUDA (Compute Unified Device Architecture) on Nvidia GPUs. Our simulation generates hundreds of frames per second, a vast improvement over traditional calculations which take hours per frame.

Gravitational lensing is a rare, natural phenomenon which causes light to unnaturally warp around massive celestial clusters, resulting in an observer to see distorted images of distant galaxies. The discovery of the gravitational lens lead to the validation of Einstein's theory of relativity, and is fundamental for understanding the universe at a massive scale. This occurs when a cluster of mass lies between an observer and a distant galaxy. Normally, a cluster of mass will block the observer's view of the distant galaxy. However, if the cluster of mass is dense enough to cause a gravitational trough across the fabric of space, light will be pulled by the cluster to travel along the trough. As a result, light emitted from a distant galaxy will be pulled around the cluster of mass, causing an observer to see the hidden, distant galaxy. The dense cluster of mass acts as a lens to view galaxies that are much farther away, usually on the scale of ten billion light years away. Traditional gravitational lensing simulations require an intricate ray-tracing algorithm to simulate each ray of light as it travels across space. Such scientifically accurate simulations take hours to compute a single frame. Throughout cinema and photography, black hole lensing effects are commonly used and rapidly computed, but they lack the scientific accuracy required for astronomical calculations.

Using CUDA, a parallel computing platform released by Nvidia, the traditional lensing simulation can be optimized to hundreds of frames per second. OpenGL is used to project the graphics to the screen. A spatial density is either calculated or provided by the user, and a Fast Fourier Transform is applied to obtain the spatial gravitational potential, which compares favorably to observed gravitational lenses (see figure). In the future, we hope to expand on this technique to map out the spatial density of the dark matter present in our universe.

Adharsh Babu completed this project while he was a student at Centennial High School, Baltimore, MD with guidance and advice from Alex Szalay and Gerard Lemson.



Simulated image (left) and lensed version of a small part of it (right).

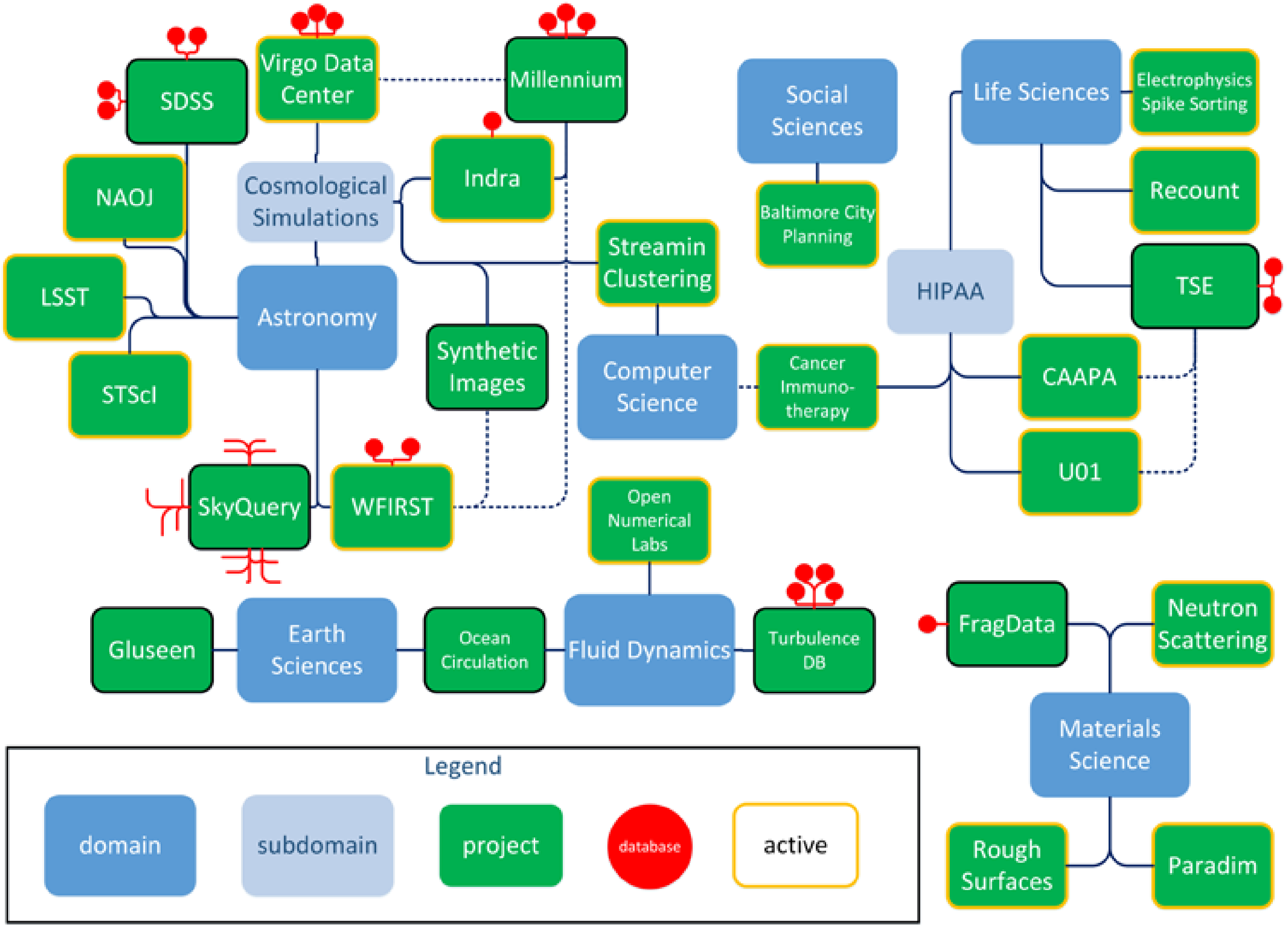
SciServer's build-out generates expertise along with petabytes

Since the initial release of SciServer in June 2016, researchers and educators world-wide have increasingly used the system to host, analyze, and visualize their data for research and teaching. The SciServer team continues to support such collaborations across diverse fields of science and engineering, with an emphasis on assisting these groups in the dissemination of the sometimes very large and complex data products they produce.

For each project, the SciServer team works with researchers to incorporate their datasets and computational requirements into the system. Building cross-domain collaborations generates expertise on the SciServer team: by importing and optimizing new types of data into SciServer contexts, by adding new analysis and visualization tools and techniques, and by understanding the constraints specific to each field. The team hopes to leverage this expertise to tackle new types of data in new science domains.

These projects incorporate SciServer tools in many ways. Some groups, such as the Turbulence Research Group and FragData project of the Hopkins Extreme Materials Institute, use the SciServer team's data modeling expertise to optimize and provide access to relational databases. Other projects, such as the JHU Ocean Circulation Database group and the **recount2** genomics project, make datasets available via SciServer Compute containers, where they can be accessed from Compute's Jupyter Notebooks.

SciServer's experienced team is eager to continue their build-out. They want to identify and work with new research groups across the scientific spectrum. If you want to know more about adding your data to SciServer, please contact sciserver-helpdesk@jhu.edu.



SciServer's science outreach projects. Major research domains are shown as blue nodes, subdomains as gray, and projects as green. Active projects are highlighted in yellow.

ANNOUNCEMENTS

IDIES initiated our Bi-Monthly Seminar Series, hosting seminars on data- and computationally-intensive topics by distinguished speakers from around the globe. See http://idies.jhu.edu/news-events/bimonthly/ for more information.

Alex Szalay and Steven Salzberg were among the 2016 Thompson Reuters Highly Cited Researchers.

Steven Salzberg was named to American Institute for Medical and Biological Engineering's 2017 College of Fellows.

recount2, *an IDIES sponsored project, published Reproducible RNA-seq analysis using recount2 in Nature Biotechnology (http://go.nature.com/2y4Rfxc).*

IDIES Affiliate Thomas Haine was featured in an article in the Washington Post (http://wapo.st/2yNBFU3).

Continued from page 7

through a complicated install and setup process on their machines.

The new SciServer system is currently in an alpha testing phase with "Early Adopters" from our science collaborations. Astronomers are using our new system to access and share raw sky images underlying SQL databases; fluid mechanics researchers and oceanographers are using it to run particle tracking codes on large-scale simulations; genomics researchers are using it to organize thousands of studies about millions of RNA sequences; and soil scientists are using it to manage access to data collected at field sites worldwide.

With the new SciServer, you can:

- Run Python, R, or Matlab scripts as asynchronous jobs
- Share datasets and scripts with colleagues
- Control who has read and/or write access to your shared data

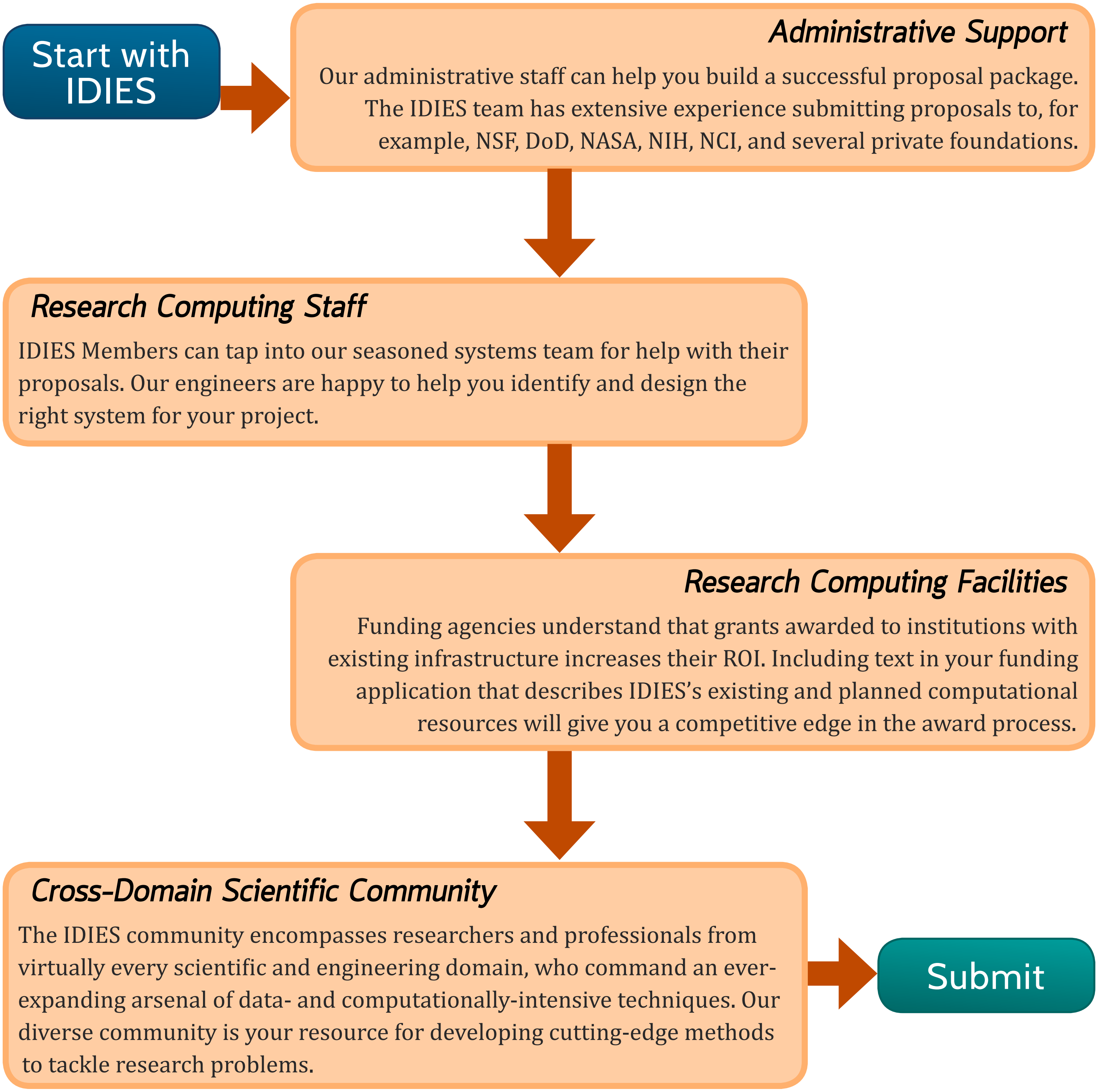
The system is being hosted on production hardware so we can monitor performance – but the real reason for our alpha testing is to get feedback on bugs, feature requests, and usability. The new SciServer will go through several cycles of feedback this fall before gaining wide public release.

We are actively seeking new Early Adopters to test our system. We would particularly love to hear from scientists with datasets that they want to make available to their colleagues and to the public. To learn more about our Early Adopters program, or to apply for an Early Adopter invite, please email the SciServer helpdesk at sciserver-helpdesk@jhu.edu.

The SciServer team is Evelin Bányai, Joseph Booker, Tamás Budavári, Camy Chhetri, László Dobos, Lance Joseph, Jai Won Kim, Jordan Raddick, Gerard Lemson, Dmitry Medvedev, Victor Paul, Mike Rippin, Bonnie Souter, Alex Szalay, Manuchehr Taghizadeh-Popp, Ani Thakar, Jan Vandenberg, Sue Werner, and Alainna White. SciServer is a collaborative research environment for large-scale data-driven science, developed at and administered by IDIES. SciServer is funded by National Science Foundation award ACI-1261715. For more information, please visit www.sciserver.org.

IDIES PROPOSAL SERVICES

IDIES provides a **Proposal Submission Service** tailored for "Big Data", one of the Signature Initiatives of Johns Hopkins University. Coming through IDIES, your proposal will be recognized as "Big Data". If you do not submit through IDIES, be sure to have your administrator check the "Big Data" box on the COEUS Conditional Compliance Questionnaire so that your proposal is properly identified.



Start with IDIES

Administrative Support

Our administrative staff can help you build a successful proposal package. The IDIES team has extensive experience submitting proposals to, for example, NSF, DoD, NASA, NIH, NCI, and several private foundations.

Research Computing Staff

IDIES Members can tap into our seasoned systems team for help with their proposals. Our engineers are happy to help you identify and design the right system for your project.

Research Computing Facilities

Funding agencies understand that grants awarded to institutions with existing infrastructure increases their ROI. Including text in your funding application that describes IDIES's existing and planned computational resources will give you a competitive edge in the award process.

Cross-Domain Scientific Community

The IDIES community encompasses researchers and professionals from virtually every scientific and engineering domain, who command an ever-expanding arsenal of data- and computationally-intensive techniques. Our diverse community is your resource for developing cutting-edge methods to tackle research problems.

Submit

For more information, visit http://idies.jhu.edu/proposal-submission-service/

SEED FUNDING AWARDS SPRING 2017

The IDIES Seed Funding Program RFP was issued for competitive awards of $25,000. The goal of the Seed Funding initiative is to provide 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.

Variational Bayes Gene Activity in Pattern Sets (VB-GAPS) bioinformatics algorithm for efficient precision medicine in oncology

Elana J. Fertig (Department of Oncology, School of Medicine), Raman Arora (Department of Computer Science, Whiting School of Engineering)

Technological advances in molecular profiling have made high-dimensional data standard for biological inquiry. The empirical knowledge that biological processes are interrelated systems has lead to the realization that gene regulatory mechanisms and gene-gene interactions responsible for phenotypes are low dimensional structures. Matrix factorization, also known as compressed sensing techniques, learn such low-dimensional mathematical representations from high-dimensional data. These factorization techniques can embed assumptions about pleiotropy, epistasis, inter-relationships between complex traits, and context-dependent interactions to accurately model biology. They have been broadly applied to uncover new biology in a breadth topics ranging from pathway discovery to time course analysis, across diverse high-throughput omics technologies in bulk and single-cell. Although relying on the same underlying algorithmic structure, the visualization and biological interpretations of matrix factorization are diverse and lack standardization. We review the interpretation of the systems level analyses obtained from matrix factorization. Codifying the mechanisms to decipher biologically relevant features with matrix factorization enables their broad application to discovery beyond the limits of current biological knowledge—answering questions from high-dimensional data that we have not yet thought to ask.

✦ ✦ ✦

New Tools for an Old Problem: Building a Global and Historical Data Set of Social Unrest

Beverly J. Silver (Sociology Department; Arrighi Center for Global Studies), Sahan Savas Karatasli (Sociology Department, Arrighi Center for Global Studies), Christopher Nealon (English Department)

With the seed grant we are developing methods to semi-automate the collection of data on labor

Professor Silver

and social protest from New York Times, The Guardian and Le Monde with the goal of both reducing the time and increasing the accuracy for coding event information (e.g., location, actors, actions, demands). Most existing social science research in this area automate the data collection process, but do so at the cost of including an unacceptable level of false positives and failing to take advantage of the rich detailed information provided in the newspaper articles themselves. Our current NSF-funded research on Global Social Protest uses search strings to extract relevant articles from the digitized newspaper archives and relies on a custom-built website for data coding and analysis; however, to avoid the above-mentioned pitfalls it used to rely exclusively on human coding of articles (which was time consuming). We are now developing natural language processing tools that allow for a middle path between full automation and manual coding. By comparing different randomized portions of the data coded using manual coding, full automation and semi-automation, we will assess the relative reliability, strengths and limitations of each method. In addition to English and French language newspapers, we plan to run pilots in Japanese, Korean and Spanish newspapers. The extension of the project to other languages allows us to widen and deepen ongoing international research collaborations. We will organize a workshop at Johns Hopkins University in Spring 2018, where social scientists and Big Data experts will be asked to comment and critique the methods and preliminary outcomes of the project.

✦ ✦ ✦

Integrating environmental genomics into biogeochemical models

Sarah Preheim (Department of Environmental Health and Engineering, Whiting School of Engineering), Anand Gnanadesikan (Department of Earth and Planetary Science, Krieger School of Arts and Sciences)

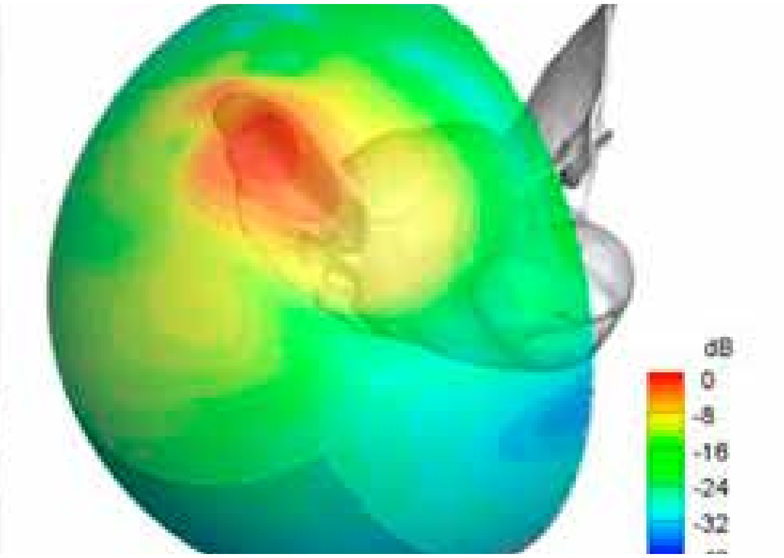
Environmental policy is increasingly based on results from computer simulations, but more integration between models and observations is needed to make sound decisions. We investigate how best to use environmental genomics to inform, constrain and improve biogeochemical models of oxygen-free (anoxic) dead-zones, such as in the Chesapeake Bay. To investigate the relationship between microbial genes and biogeochemical model predictions, we began with a simple aquatic ecosystem, Upper Mystic Lake. Samples were collected approximately every meter from the surface to 22 meters depth from spring to summer in 2013. The distribution in the water column of genes known to be involved biogeochemical processes was explained by biogeochemical process rates predicted by the model, suggesting genes may be a useful proxy for biogeochemical rates under certain circumstances. By reconstructing microbial genomes from complex assemblages of microorganisms and looking at co-occurrence patterns, we were able to determine additional biogeochemical processes that, if active, would significantly alter the

predicted biogeochemistry of the lake. We are also investigating the relationship between microbes, their genes and model predictions in a more complex ecosystem: the Chesapeake Bay. Samples were collected in the Chesapeake Bay at one site monthly for 3 months during the period of anoxia from 2015-2017. We also collected water samples from the bottom of the Bay throughout the dead-zone in 2016 and 2017. By analyzing the microbial community within the water samples, we found that many microorganisms are present throughout the water column, but that their relative abundance changes based on the presence of oxygen. We also found that genes involved in anoxic processes are more abundant in the water column where oxygen is depleted. We will analyze whether microbial activity, assessed through gene transcription, is more informative than gene abundance in complex ecosystems like the Chesapeake Bay. We will use genome reconstruction to identify biogeochemical processes of importance that are not currently included in the models of dead-zone biogeochemistry. With these observations, we hope to improve predictions of how size and duration of the dead-zone could respond to changes.

EchoSIM: Multiscale Acoustic Simulations Integrated with Free-Flight Experiments for Echo Scene Analysis of an Echolocating Bat

Rajat Mittal (Department of Mechanical Engineering), Jung Hee Seo (Department of Mechanical Engineering), Cynthia F. Moss (Psychological and Brain Sciences), Susanne J. Sterbing-D'Angelo (Psychological and Brain Sciences)

Animals that rely on active sensing provide a powerful system to investigate the neural underpinnings of natural scene representation, as they produce the very signals that inform motor actions. Echolocating bats, for example, transmit sonar signals and process auditory information carried by returning echoes to guide behavioral decisions for spatial orientation. Bats compute the direction of objects from differences in echo intensity, spectrum, and timing at the two ears; while an object's distance is measured from the time delay between sonar emission and echo return. Together, this acoustic information gives rise to a 3D representation of the world through sound, and measurements of sonar calls and echoes provide explicit data on the signals available to the bat for orienting in space.



In the present seed funding program (SFP), we are developing a first-of-its-kind computational simulation-enabled method for echo scene analysis of an echolocating bat. We refer this echo-simulator as "EchoSIM". The proposed method integrates tightly with free-flight laboratory assays of bats and takes as input, variables such as the bat's flight path, hear-ear anatomy, position and orientation as well as the sonar call wave form. The simulation results (3D echo scene and echo signal) together with the experimental measurements will provide a unique and powerful integrated dataset that enable unprecedented analysis of active sensing and adaptive flight behavior of bats in complex environments.

The key components of the "EchoSIM" has been developed this year with the support of the SFP. EchoSIM integrates experimental measurements, full 3D direct numerical simulation of ultrasound wave scattering in the near-field, and an acoustic analogy based model for the propagation and back-scattering of ultrasound waves in the far-field. A series of simulations have been performed using EchoSIM for a bat tracking finite size, moving objects, and it is observed that the spectrum of echo signal contains information related to the size and shape of the object. In addition, the distance and speed information the object can be derived from the time-delay and Doppler shift analysis. Laboratory experiments are also being conducted for the same situations, and the measured data will be compared with the simulation results to validate EchoSIM.

IDIES New Membership Categories

All IDIES members can

- Use IDIES resources, attend IDIES events, find potential collaborators, and connect with experts in diverse areas of computationally- and data-intensive research.
- Gain early access to registration for IDIES events.
- Research Big Data funding opportunities on our website.
- Join the IDIES mailing list.

IDIES Faculty

...are tenure-track and research faculty from our sponsoring divisions who can serve as a PI on a research grant.

Can:

- Submit Big Data-related proposals via the IDIES Proposal Submission Service.
- Use IDIES resources, including the Data-Scope, SciServer, etc.
- Participate in the IDIES Seed Funding Initiative.
- Participate in IDIES new initiatives.

IDIES Affiliates

...are teaching professors, lecturers, adjunct professors, research staff, and postdocs from our sponsoring divisions who work with IDIES but cannot act as a PI on a research grant.

Can:

- Submit Big Data-related proposals as a Co-I with IDIES Faculty via the IDIES Proposal Submission Service.
- Participate in the IDIES Seed Funding Initiative as a Co-I with IDIES Faculty.
- Participate in IDIES new initiatives.

IDIES Collaborators

...are JHU faculty from divisions that do not sponsor IDIES, or are non-JHU faculty and external researchers (reviewed on a case-by-base basis).

...must have an IDIES Faculty sponsor.

Can:

- Access IDIES computing resources as allowed by their Faculty or Affiliate sponsor.

IDIES Students

...are JHU doctoral, medical, masters, and undergraduate students.

...must have an IDIES Faculty or Affiliate sponsor.

Can:

- Access IDIES computing resources as allowed by their Faculty or Affiliate sponsor.
- Volunteer at IDIES hosted conferences.

For more information, visit http://idies.jhu.edu/join/

OUR MISSION

We foster education and research in the development and application of data intensive technologies to problems of national interest in physical and biological sciences and engineering. The institute provides faculty, researchers and students with the structure and resources needed to accomplish these goals.

Leadership

Intellectual leadership in addressing research challenges related to the "Science of Big Data," establishing a group that leads the world in new discoveries enabled by next-generation data sets and analytics. Provide coordination of integrative activities, such as seminar series, and visitors.

Vision

Continue to provide vision and oversight to high performance and data intensive computing across all of JHU, in the spirit that has proven to be highly successful over the last four years (HHPC 1 and 2, GPU). Having a large shared facility enables leveraging needed for seeking further funding opportunities.

Growth

Given the emerging need of data analytics skills for the workforce of the future, IDIES will work with the departments to establish new masters, graduate, and undergraduate programs, minors, etc, that emphasize these new skills.

Management

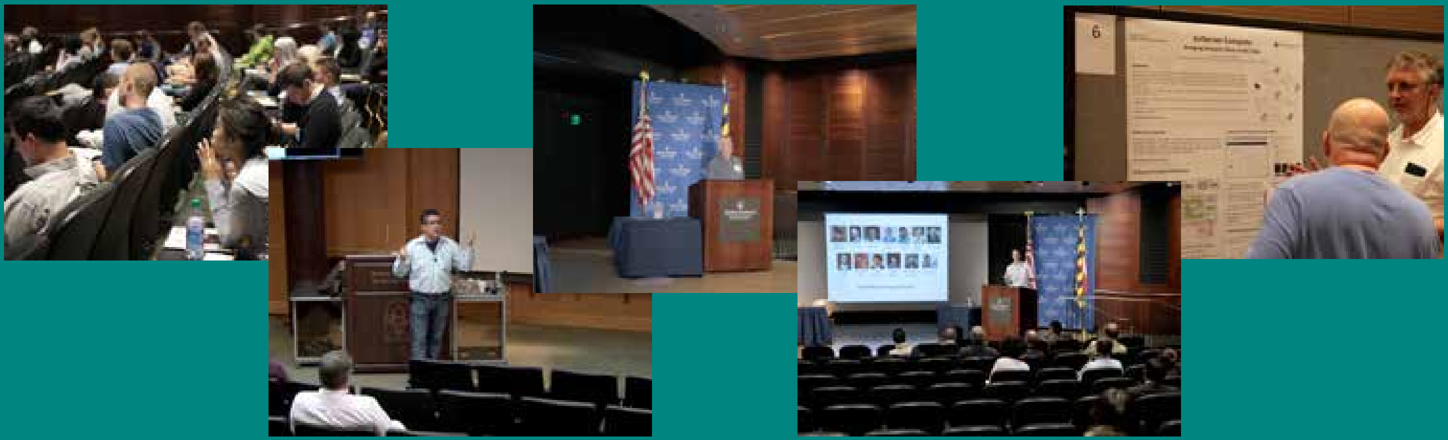
Management of a significant high-performance computing facility. IDIES needs state of the art facilities to enable its members to use data in new ways and compete for new funding. MARCC provides exciting opportunities for continuing our development of facilities that are a magnet attracting new JHU researchers to the institute.

Development

Continue to develop mutually beneficial corporate partnerships and through these affiliations transform research into sustainable, real-world applications.

Incubator

An incubator for creating/curating/publishing new data sets at JHU that could be preserved within the JHU Data Archive. This would give the group an "unfair advantage," name recognition, and additional leverage, while also motivating and focusing research around challenges and opportunities of dealing with Big Data.



IDIES is always accepting affiliates who are Faculty and Research Scientists within the Johns Hopkins community. Visit idies.jhu.edu/join for more information, and to join today!

THANK YOU

To our generous sponsors

NASA

SCHMIDT FAMILY FOUNDATION
Philanthropy for a Better World

JOHN TEMPLETON FOUNDATION

NSF

FRED L. EMERSON FOUNDATION

NIH
National Institutes of Health

ALFRED P. SLOAN FOUNDATION
1934

intel

GORDON AND BETTY MOORE

NOKIA

MRF MELANOMA RESEARCH FOUNDATION

NVIDIA

Microsoft

W.M. KECK FOUNDATION

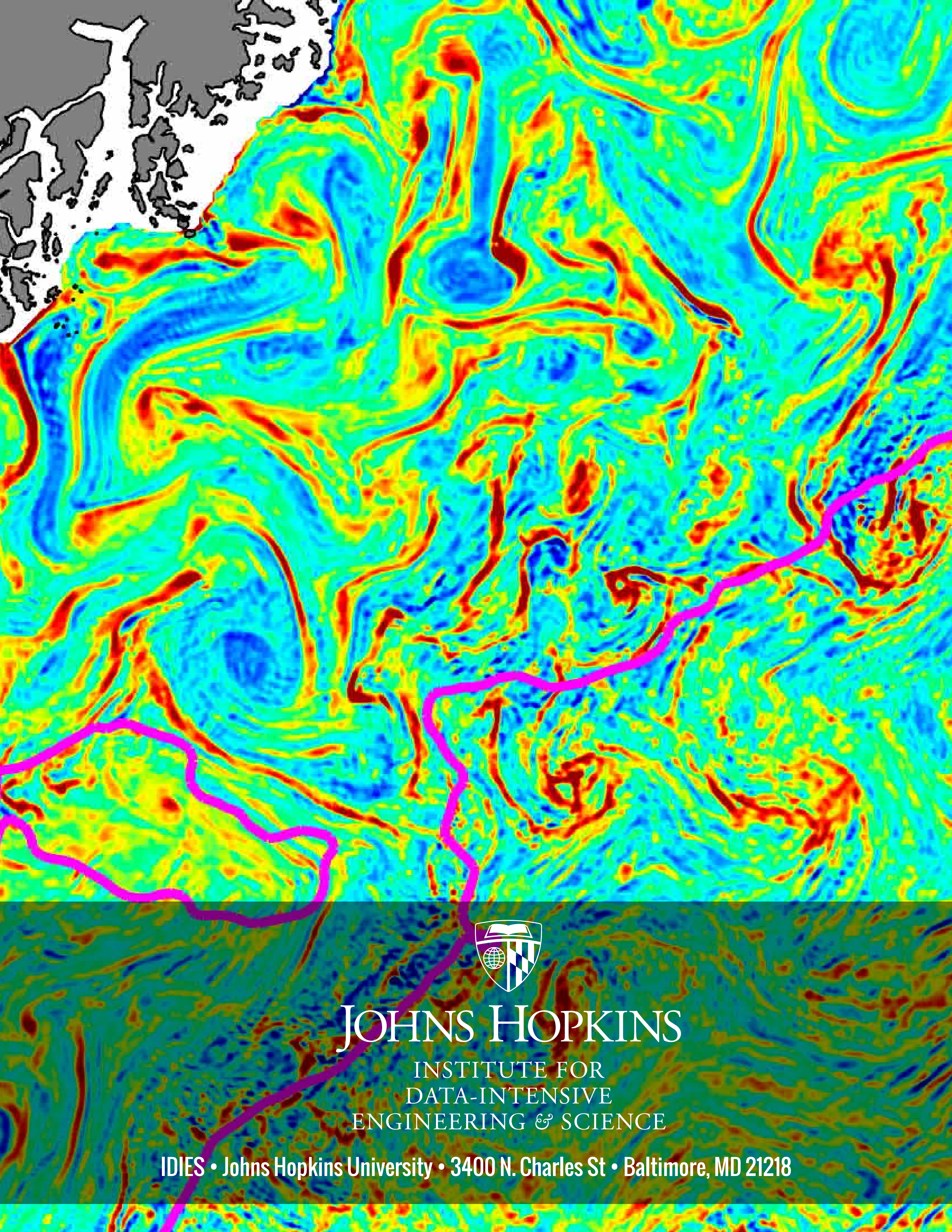
The IDIES Executive Committee would like to extend our heartfelt gratitude to our affiliates, collaborators, contributors, editors, and staff, without whose continued support and cooperation IDIES would not be possible.

—*Alex Szalay, Charles Meneveau, Stephen Salzberg, Mark Robbins , Ani Thakar, Sayeed Choudhury, Margie Gier, & Alainna White.*

Steven Salzberg

Alain C. White

Margie Gier



JOHNS HOPKINS

INSTITUTE FOR
DATA-INTENSIVE
ENGINEERING & SCIENCE

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