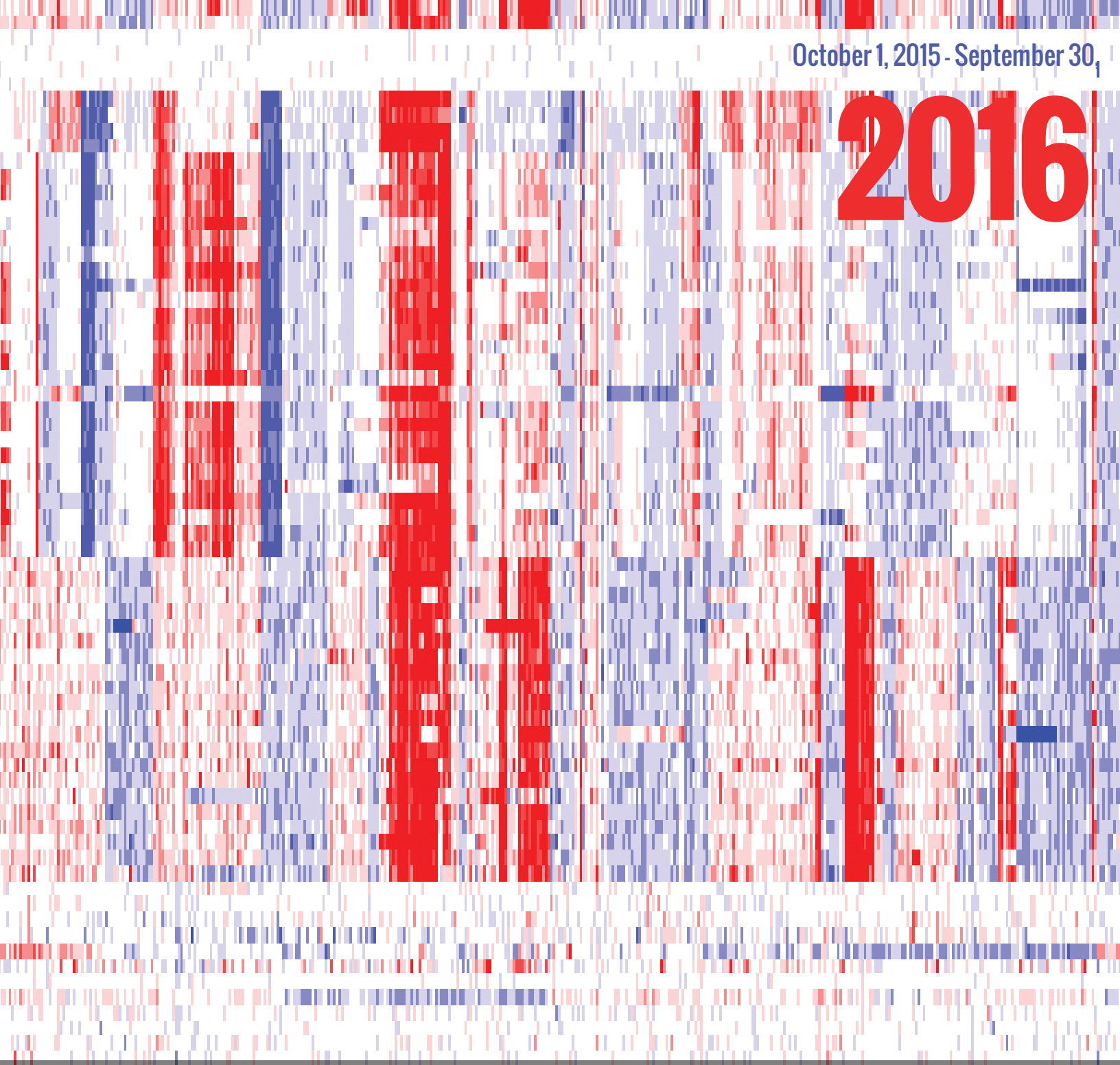
October 1, 2015 - September 30, 2016



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 includes 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.

Besides serving in a leadership role for Big Data initiatives for the University, IDIES has become responsible for the research computing efforts at JHU. In Sep 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. The project was supported by a $30M grant from the State of Maryland. The system has a world-class, 100G connectivity to Internet2, with a similar campus backbone and 10-40G uplinks into the individual buildings. The facility has now been in steady operation for more than a year, and has enabled cutting edge computations for researchers.

Today IDIES involves more than 100 faculty and a similar number of graduate students. Over the last 12 months we have awarded 4 seed grants in a broad spectrum of topics, connecting researchers from different fields, but sharing a common interest in Big Data.

In three 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.

This year we decided to host the All Hands Meeting in East Baltimore, to make it easier for our colleagues at the School of Medicine and Public Health to engage with our activities. Our speakers also reflect this trend. We hope that this will help to grow our engagement in health related areas further. I would like to encourage our colleagues in these fields to follow our announcements, join our ranks as members of IDIES, and help us to advance Big Data at Johns Hopkins.

October 1, 2015 - September 30, 2016

The Institute for Data Intensive Engineering and Science Annual Review

On the cover: A representation of analyzing hundreds of individual cells in a breast cancer tumor to recognize the specific changes to the DNA in each cell. This visualization indicates how the tumor progresses, and how the primary tumor is related to the metastases. See *IDIES Researchers Reveal Intratumor Heterogeneity in Metastatic Breast Cancer* on page 8.

AGENDA

2016 Institute for Data Intensive Engineering and Science (IDIES)
Annual Symposium, October 21, 2016 8:00 – 5:30
East Baltimore Campus – Sheldon Hall

8:00 a.m.	***Continental Breakfast – Check In***
9:00 a.m.	**Opening Remarks** **S. Alexander Szalay, PhD**, Director of IDIES, Professor of Astrophysics & Computer Science, Johns Hopkins University
9:15 a.m.	KEYNOTE SPEAKER **How much Cancer can we Cure with the Immune System?** **Andrew Pardoll, MD, PhD**, Director, Bloomberg-Kimmel Institute for Cancer Immunotherapy, School of Medicine, Johns Hopkins University
10:05 a.m.	**Using Causal Inference to Make Sense Of Messy Data** **Ilya Shpitser, PhD,** John C. Malone Assistant Professor, Department of Computer Science, Whiting School of Engineering, Johns Hopkins University
10:25 a.m.	**SEED FUND UPDATE** **In Pursuit of DNA and RNA Amalgamations** **Sarah Wheelan, MD, PhD**, Associate Professor, Institute for Genetic Medicine, School of Medicine, Johns Hopkins University
10:45 a.m.	***Break***
11:00 a.m.	**Data Subject to Restrictions at MARCC** **Jaime E. Combariza, PhD**, Associate Research Professor, Department of Chemistry, Krieger School of Arts and Sciences, Johns Hopkins University; Director, Maryland Advanced Research Computing Center (MARCC)
11:20 a.m.	**SEED FUND UPDATE** **Scalable Framework for Statistical Inference on Big Multimodal Data via Sketching and Concentration** **Vladimir Braverman, PhD**, Assistant Professor, Department of Computer Science, Whiting School of Engineering, Johns Hopkins University
11:40 a.m.	**Poster Madness:** 1 slide and 1 minute to Pitch Your Poster!
11:50 p.m.	***Lunch and Poster Session - Feinstone Hall, Bloomberg School of Public Health – E2030***
1:20 p.m.	**SEED FUND UPDATE** **Towards the Johns Hopkins Ocean Circulation Database: Method Development and Prototype** **Thomas Haine, PhD**, Morton K. Blaustein Chair and Professor of the Earth & Planetary Sciences Department, Krieger School of Arts and Sciences, Johns Hopkins University
1:40 p.m.	**SEED FUND UPDATE** **Metabolic Compass: A Mobile Health Platform for Understanding the Impact of Circadian Sleeping, Eating and Exercise Behaviors on Metabolic Syndrome, and Obesity** **Jeanne Clark, MD, MPH**, School of Medicine, **Yanif Ahmad, PhD**, Whiting School of Engineering, Johns Hopkins University
2:00 p.m.	**Optimization Challenges for Cost-Sensitive Model Prediction with Applications in Healthcare** **Daniel Robinson, PhD**, Assistant Professor, Department of Applied Math and Statistics, Whiting School of Engineering, Johns Hopkins University
2:20 p.m.	**SciServer Compute: Bringing Analysis Close to the Data** **Mike Rippin, PhD**, Associate Research Scientist, Department of Physics and Astronomy, Krieger School of Arts and Sciences, Johns Hopkins University
2:40 p.m.	***Break***
2:55 p.m.	KEYNOTE SPEAKER **Analyzing Big Data from Big Brains** **Terrence Sejnowski, PhD**, Francis Crick Professor, Computational Neurobiology Laboratory, Salk Institute for Biological Studies
3:45 p.m.	**The 'Sixth' Factor -- Social Media Factor Derived Directly from Tweet Sentiments** **Jim Kyung-Soo Liew, PhD**, Assistant Professor, Chairman of the Carey Business School Innovation Factory, Carey Business School, Johns Hopkins University
4:05 p.m.	**Navigating tens of thousands of RNA-seq datasets with recount, SciServer and Jupyter** **Benjamin Langmead, PhD,** Assistant Professor, Department of Computer Science, Whiting School of Engineering, Johns Hopkins University
4:25 p.m.	**Closing Remarks** **S. Alexander Szalay, PhD,** Director IDIES, Professor Astrophysics & Computer Science, Johns Hopkins University
4:30 p.m.	***Cocktail Reception – Feinstone Hall, Bloomberg School of Public Health – E2030***

KEYNOTE SPEAKERS

DREW M. PARDOLL, PhD, MD

Abeloff Professor of Oncology, Director, Cancer Immunology, Director, Bloomberg-Kimmel Institute for Cancer Immunotherapy

Dr. Pardoll is an Abeloff Professor of Oncology, Medicine, Pathology and Molecular Biology and Genetics at the Johns Hopkins University, School of Medicine. He is the Director of the Bloomberg~Kimmel Institute for Cancer Immunotherapy and Co-Director of the Cancer Immunology Program at the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins.

TERRY SEJNOWSKI, PhD

Professor & Laboratory Head Computational Neurobiology Laboratory Francis Crick Chair, Salk Institute for Biological Studies

Terrence Sejnowski is a pioneer in neural computation and his goal is to understand the principles that link brain to behavior. He is a member of the Institute of Medicine, National Academy of Sciences and the National Academy of Engineering, one of only 10 living persons to be a member of all three national academies.

SPEAKERS

VLADIMIR BRAVERMAN, PhD

Vladimir Braverman is an Assistant Professor in the Department of Computer Science in the Whiting School of Engineering at the Johns Hopkins University. His research interests include data stream, sub-linear and randomized algorithms.

JEANNE CLARK, MD, MPH

Jeanne Clark is the Director of the Division of General Internal Medicine as well as a Professor of Medicine at Johns Hopkins University. Dr. Clark is a practicing general internist and is an expert in obesity, diabetes, and nonalcoholic fatty liver disease.

JAIME COMBARIZA, PhD

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

THOMAS HAINE, PhD

Thomas Haine is the Morton K. Blaustein Chair and Professor of the Earth & Planetary Sciences Department at Johns Hopkins University. Dr. Haine's research interests specialize in ocean circulation and dynamics, and the ocean's role in climate.

BENJAMIN LANGMEAD, PhD

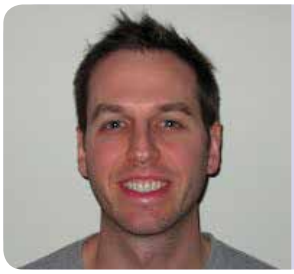
Benjamin Langmead is an Assistant Professor in the Computer Science Department at Johns Hopkins university, His research interests include genomics, sequence alignment, text indexing, high performance computing, and big data.

JIM LIEW, PhD

Jim Liew, is an Assistant Professor in the Carey Business School at Johns Hopkins University. He is an expert in multiple areas of finance, including the topics of derivatives, fixed income, hedge fund strategies and wealth management.

MIKE RIPPIN, PhD

Mike Rippin is the project manager for the SciServer project. He has 20 years of experience working in IT and software services for commercial scientific research and development organizations, and for 5 years was VP of Operations at Tessella Inc.



DANIEL ROBINSON, PhD

Daniel Robinson is an Assistant Professor in the Department of Applied Mathematical & Statistics at Johns Hopkins University. His research interests include real-time optimization in energy systems and predictive modeling in healthcare.

ILYA SHPITSER, PhD

Ilya Shpitser is a John C. Malone Assistant Professor in the Department of Computer Science at Johns Hopkins University. His research includes all areas of causal inference and missing data, particularly using graphical models.

ALEX SZALAY, PhD

Professor Szalay is the founding Director of IDIES, a Bloomberg Distinguished Professor, Alumni Centennial Professor of Astronomy, and Computer Science Department Professor. He is a cosmologist, working on the use of big data in advancing scientists' understanding of astronomy, physical science, and life sciences.

SARAH WHEELAN, MD, PhD

Sarah Wheelan is an Associate Professor in the School of Medicine at Johns Hopkins University. Her research interests include genomics, high throughput sequencing, next generation sequencing, sequence analysis, and transposon.

Maryland Advanced Research Computing Center

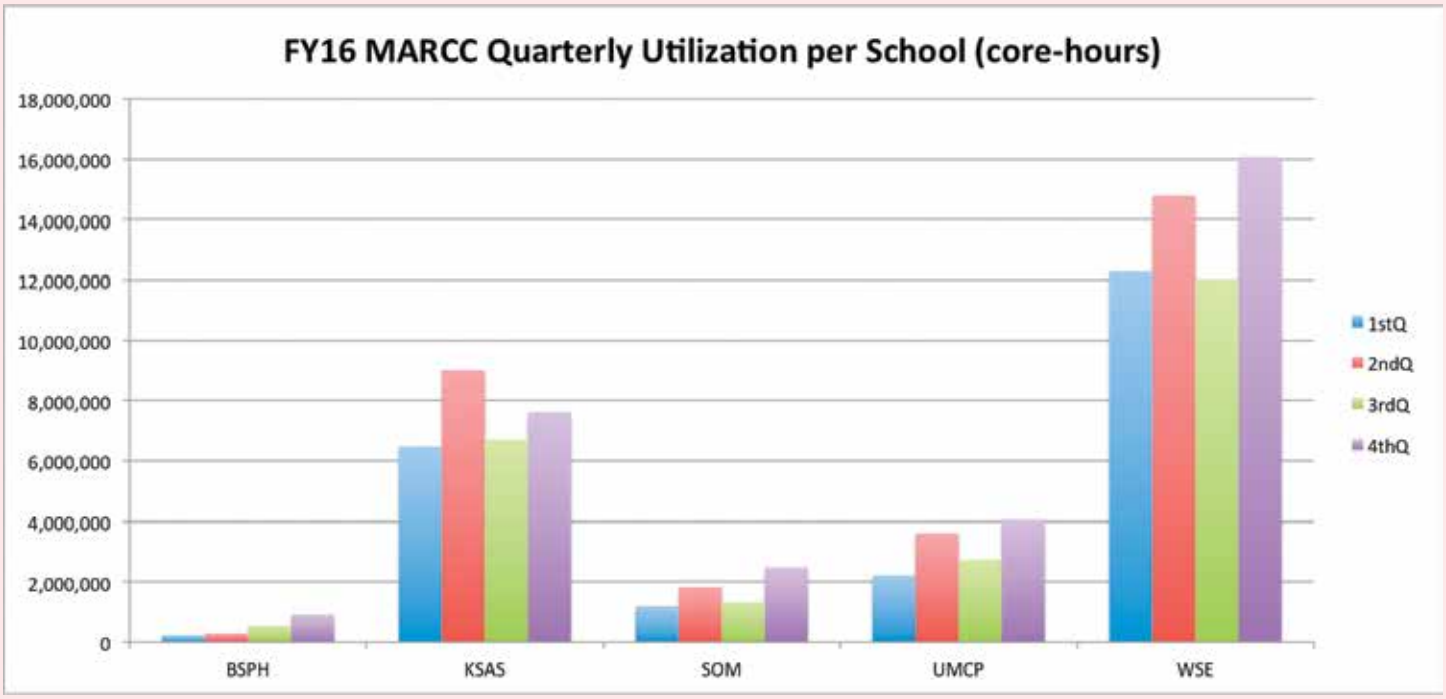
First Year Impact, Transforming Research Computing

In its first year of production, the Maryland Advanced Research Computing Center (MARCC) has contributed to transform research computing in Maryland. MARCC has granted 200+ requests for allocation from the schools participating in this project: Bloomberg School of Public Health, Krieger School of Arts and Sciences, School of Medicine, the Whiting School of Engineering, and the University of Maryland at College Park.

Advances in technology continue their expected path. In 2015, MARCC started with 19,000+ cores spanning three different types of compute nodes: large memory, standard compute, and compute nodes with accelerators (2 Nvidia K80 GPUs). GPU computing, in particular, is in high demand as applications are optimized to use accelerators. MARCC is adding more compute nodes by the end of this year: 24 GPU nodes plus 48 standard compute nodes with the latest Broadwell processors. This addition increases MARCC's theoretical computing capability to about 1.1 Peta-FLOPs, making MARCC one of the most powerful computational resources in academia in the US.

Perhaps the most important metric is the impact that MARCC has had on research agendas:

- The ready availability of cutting edge computational resources at MARCC creates an efficient, secure, dynamic environment that drives transformative research in many disciplines and it is becoming a crucial tool that enables researchers to conduct desired research.
- In the past few years, big data analytics has transformed many research agendas. MARCC's capability to quickly transfer, store and analyze large amounts of data is reflected in an increase in productivity and allows researchers to tackle more ambitious problems.
- Often timely results are critical to compete with researchers at peer institutions and high volume



Over 800 user accounts have been generated with about half of them active in a particular month. Researchers ran data analysis that utilized 105,880,609 core-hours for about 75% effective utilization. In FY16, with only a few PIs reporting, we received information that over 70 publications and presentations used MARCC resources for data analytics.

Continued on next page.

simulations are necessary to produce statistically meaningful results.

MARCC provides tools to facilitate the interaction between computation and physical experiments, which are now very common in many disciplines. Software applications have been developed, installed and optimized to provide a practical environment so that researchers can concentrate on their investigations without operational distractions.

MARCC Data Center at the Johns Hopkins Bayview Campus in Baltimore, MD

A few projects that take advantage of MARCC's resources are highlighted. Professor Julian Krolik in the Physics and Astronomy Department (JHU) is conducting research on the effects of matter in the vicinity of black holes. They investigate the tidal disruption of ordinary starts that happen to pass close to a large black hole that torn them apart. They calculated radiation-matter interactions to support observational diagnostics of matter orbiting in the outside of a black hole.

In another Physics study, Professor D. Richardson (UMCP) is conducting simulations of dense regions of Saturn's rings, which feature continual collisions between small icy particles in close proximity as they orbit the giant planet. The major finding is that certain dynamical instabilities can co-exist in the densest part of the rings, giving rise to specific wave-like behavior that may explain features observed by spacecraft, including the ongoing Cassini mission at Saturn.

Professor Natalia Trayanova's group (BME, JHU) is developing a highly innovative patient-specific MRI-based heart modeling environment that represents cardiac functions from molecular processes to electrophysiological and electromechanical interactions at the organ level. They term this environment ``virtual electrophysiology lab'', and propose to translate it into the clinic and apply it to the non-invasive diagnosis and treatment of heart rhythm and contractile disorders in patients with structural heart disease.

Professor T. Mueller in Materials Science Engineering (JHU) has demonstrated that the cluster expansion method is able to accurately model the structure and properties of metal alloy surfaces, making it a valuable tool in the design of new catalysts. However, this method is limited to studying the surface of bulk materials with fixed lattice parameter. An extension of the cluster expansion method allows predicting the properties of surfaces as a function of the underlying bulk lattice parameter. The group is using MARCC to generate DFT data to develop and validate their new method.

The group of professor Xin-Zhong Liam (Atmospheric Sciences at UMCP) is coupling the state-of-the-art mesoscale regional Climate-Weather Research and Forecasting model (CWRF) with the most-comprehensive crop growth models to study climate-crop interactions. They are developing an advanced model infrastructure to quantify the impact of key environmental stress factors, including temperature, moisture, nutrient, UV radiation, CO2 concentration, aerosols and other air pollutants, on agricultural crop quality and yields.

For these and many more projects MARCC has been an invaluable resource. The Center fulfills the need for advanced computational facilities and support in the region, as foreseen by those that envisioned it.

SciServer Compute Bringing Analysis Close to the Data

Dmitry Medvedev, Mike Rippin, Gerard Lemson, Jordan Raddick, Bonnie Souter (Institute for Data Intensive Engineering and Science)

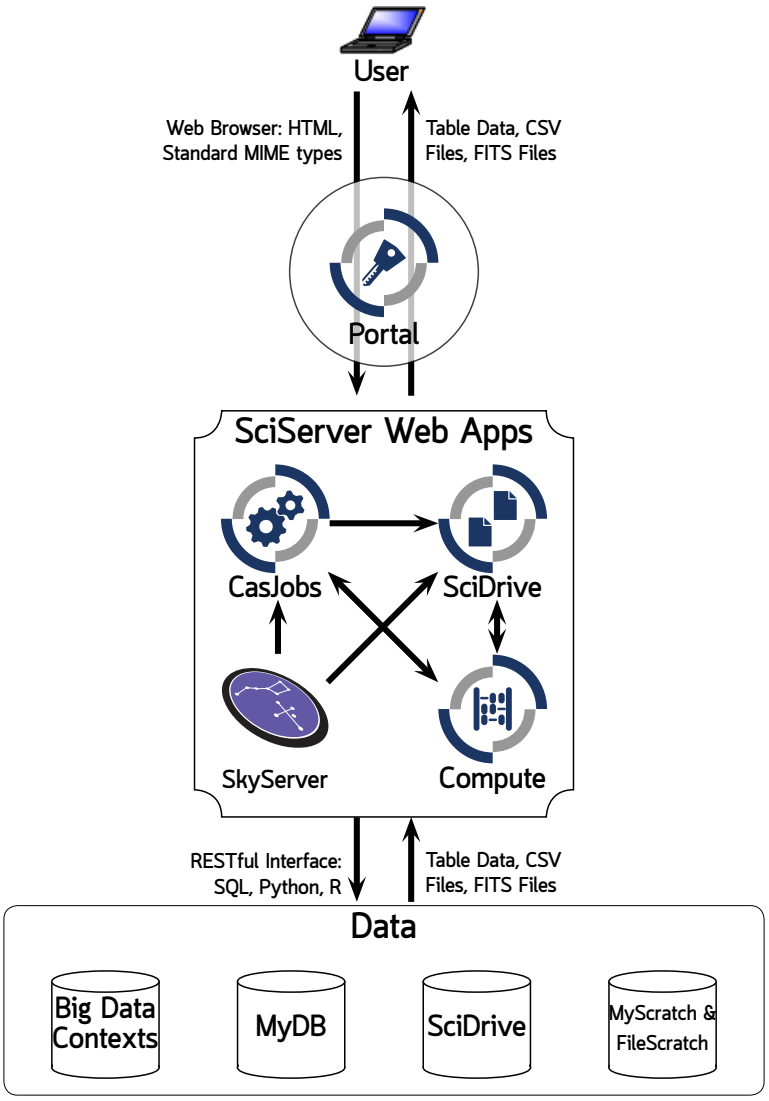
SciServer Compute features Jupyter notebooks running in server-side Docker containers attached to large relational databases and file storage to bring advanced analysis capabilities close to the data. SciServer Compute is a component of SciServer, a big-data infrastructure project developed at Johns Hopkins University that will provide a common environment for computational research.

With SciServer, launched in May, 2016, users use CasJobs and SkyServer to create freeform SQL queries on SciServer-hosted large relational databases, such as the Sloan Digital Sky Survey's astronomy dataset. In addition to downloading query results, users save their results in the cloud in several ways: in their personal MyDB, in the shared MyScratch Database and File temporary storage, or in the file storage service SciDrive. Each of these SciServer-provided storage options minimizes data-movement, keeping the data close to SQL-based analysis tools.

SciServer Compute, added in June, expands users' capabilities, enabling them to perform scientific analysis with scripting languages such as Python, R, and Matlab with Jupyter Notebooks deployed in Docker containers on a dedicated, scalable, cluster of servers. SciServer Compute integrates with other SciServer tools and data stores, allows users to query SciServer-hosted and other databases and file systems, and read and write their results to MyDB, MyScratchDB and FileScratch, and SciDrive. Compute also provides direct access to large data archives and additional libraries and tools through shared, read-only data volumes.

SciServer currently supports and collaborates with researchers in astronomy, cosmology, genomics, turbulence, environmental science, oceanography, and materials science, and plans to expand to more scientific domains to support researchers at Johns Hopkins and beyond.

SciServer is a collaborative research environment for large-scale data-driven science. It is being developed at, and administered by the Institute for Data Intensive Engineering and Science at Johns Hopkins University. SciServer is funded by the National Science Foundation Award ACI-1261715. For more information about SciServer, please visit http://www.sciserver.org.



Major components and data sources for the SciServer System.

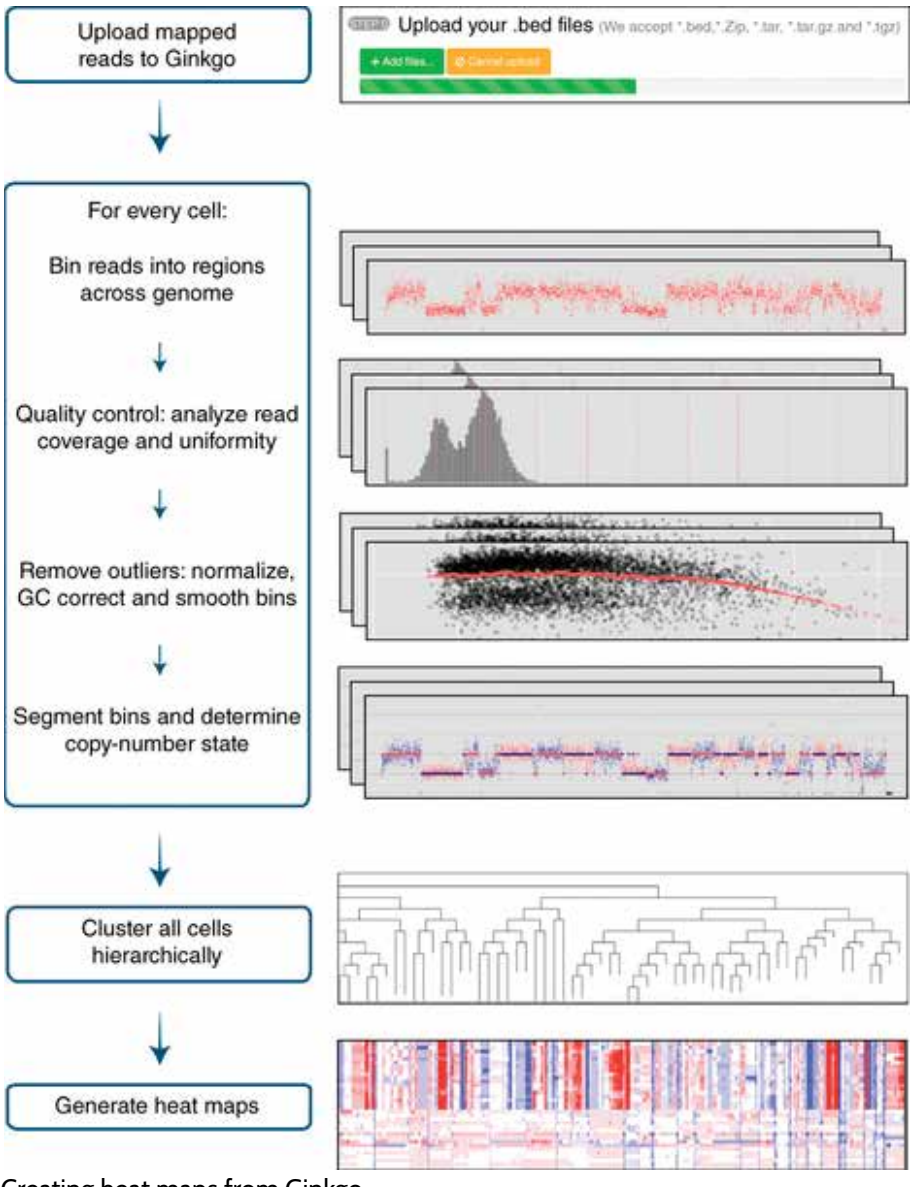
IDIES Researchers Reveal Intratumor Heterogeneity in Metastatic Breast Cancer

Michael Schatz (Department of Computer Science)

Cancer is a disease of the genome caused by the accumulation of genetic and epigenetic changes that leads to abnormal cell growth and proliferation. One common genetic mutation in cancer is called a copy number variation. While cells in normal healthy tissues have two copies of each chromosome, cancerous cells often acquire extra or loose copies of different portions of their chromosomes. It is critically important to cancer patients to know when cancer-promoting genes called oncogenes are amplified or when cancer-fighting genes called tumor suppressors are lost. New single cell biotechnology allows researchers to start to address this question by isolating and collecting genomic information from individual tumor cells, although this creates vast quantities of data far too complex to understand without sophisticated analysis software.

IDIES member Schatz recently published Ginkgo in Nature (http://rdcu.be/kOdD), as the first automated analysis system for single cell copy number profiling. One of the main applications of the software is to profile individual cells from a tumor to understand how it is progressing at the genetic level. The figure shows the estimated copy number created by Ginkgo for 100 individual cells isolated from a patient with metastatic breast cancer. Each row represents the profile of a single cell across the 3 billion base pairs of her genome. Values near 1.0 indicate the chromosomes are healthy, with two copies per cell. Values nearer to 0 means one or two copies of the chromosome are lost. Values above 1.0 indicate extra chromosome copies. This profile shows that approximately one half of the cells in the tumor have abnormal copy number profiles (top half with bright red and blue bands), while the other half (bottom) have "quiet" genomes with limited copy number changes. Interestingly, the cells with copy number changes can be further divided into two main groups, suggesting that her tumor contains 2 main populations called clones that will need to be individually treated to combat her cancer.

Central to this analysis is the Bowtie program, which is used by Ginkgo to map DNA sequencing data from each tumor cell to the reference human genome. Bowtie was created and is maintained by IDIES members Ben Langmead and Steven Salzberg. See "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome" (https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25).



Creating heat maps from Ginkgo.

Cost-Sensitive Prediction: Applications in Healthcare

Daniel Robinson (Department of Applied Mathematics & Statistics), Suchi Saria (Department of Computer Science)

Advances in model prediction for problems that have a non-trivial cost structure are needed. In healthcare, the financial, nurse time, and wait time costs share a complicated dependency with the clinical measurements needed and medical tests performed. In 2014, the healthcare budget in the United States came to 17% of the GDP with a total annual expenditure of $3.1 trillion dollars. It is estimated that between one-fourth and one-third of this amount was unnecessary, with most attributed to avoidable testing and diagnostic costs. Therefore, the design of new cost-sensitive models that faithfully reflect the preferences of a user is paramount.

In the first part of this project we represented the cost graph as finite layer boolean circuit, and then used properties of boolean circuits to formulate a new class of overlapping group-sparse regularizers that incorporate a patient's preferences. The use of the new regularizer has lead to better predictions at lower costs, which has the potential to influence personalized healthcare. This work has been published in the International Joint Conferences on Artificial Intelligence (IJCAI), and a chapter publication in The Institution of Engineering and Technology (IET).

The second part of this project is focusing on the design of an efficient optimization solver for ℓ_1-regularized large-scale convex problems. This serves as a first step toward the development of a method that efficiently minimizes group-sparse regularized functions. A manuscript detailing the new algorithm and its convergence theory has been submitted to the SIAM Journal on Optimization. A second paper that details the local convergence analysis and C++ implementation is in preparation for submission to the SIAM Journal on Optimization. Our preliminary implementation is outperforming other state-of-the-art solvers across a diverse set of model prediction problems, and hopefully will be available to the public within the next six months.

Genomics Data Explosion will Require Proactive Data Management Strategy

IDIES Affiliate Michael Schatz extrapolates current Genomics data growth rates and asserts that genomicists must create strategies now to avoid becoming inundated with data.

Genomics, a science that didn't exist 15 years ago, is on course to join Astronomy, Twitter, and YouTube as frontrunners of Big Data. In the July 7, 2015 issue of PLOS Biology, IDIES affiliate Michael Schatz and his co-authors document how Genomics data and the demands of working with it have increased at an astounding rate. According to Professor Schatz and his colleagues, knowing that Genomics is growing at this unprecedented pace will enable genomicists to develop a strategies to capture, store, process, and interpret genomics data.

Genomics may soon generate more electronic bytes per year than any other field. Even by conservative estimates, genomicists predict that by the end of the decade more than one billion billion (1 exabase) base pairs will be sequenced. With the global trend to sequence large populations of genomes, as well as the strides in DNA sequencing technology, it is likely that this estimate is conservative. Professor Schatz, who joined Johns Hopkins in January 2016 as an Associate Research Professor of Computer Science, is a co-author of the PLOS paper. The authors estimate that genomics data may experience growth of four to five orders of magnitude by 2025, making it one of the fastest-growing Big Data sciences.

Honors and Achievements for our Affiliates

IDIES Associate Director Charles Meneveau received an honorary doctorate from the Danish Technical University (DTU) for "Outstanding and highly innovative scientific achievements in fluid dynamics, particularly for his work on turbulence and atmospheric physics and its applications to wind energy."

...

IDIES Director Alex Szalay received the IEEE Computer Society 2015 Sidney Fernbach Award for his outstanding contributions to the development of data-intensive computing systems and on the application of such systems in many scientific areas including astrophysics, turbulence, and genomics.
Alex Szalay also received the Microsoft Outstanding Collaborator Award for his significant contribution to computational research in a variety of scientific fields.

...

The Very Large Database of Lipids (VLDL) received a charitable gift of $300,000 from the Trone Family Foundation to support the VLDL big data project. The VLDL Project, led by IDIES Affiliate Steven Jones is an ongoing project to leverage industry-derived data from a leading clinical laboratory for independent use by academic investigators.

...

Members of Bioinformatics.org voted Ben Langmead as the laureate of the 2016 Benjamin Franklin Award in Life Sciences for promoting free and open access to the materials and methods used in the life science.

...

Steven L. Salzberg and Alex Szalay were among the 2015 Thompson Reuters Highly Cited Researchers.

...

Professor Sarah Wheelan and co-principal investigator Carol Greider received $100,000 from the Allegheny Health Network Cancer Institute for their project "Distinct mechanism of telomere maintenance: toward individualized treatment."

...

Michael Schatz was named as the 21st Bloomberg Distinguished Professor, an honor he shares with Alex Szalay and Stephen Salzberg.

...

Ben Langmead received an NIH Grant. Professor Langmead and his team used IDIES seed funding to develop Rail-RNA, a software tool that enables scientists, including those with limited computational resources and expertise, to execute a uniform, splicing-aware, annotation-agnostic analysis of many RNA sequencing datasets. The NIH/NIGMS R01 grant will fund future development of the software

...

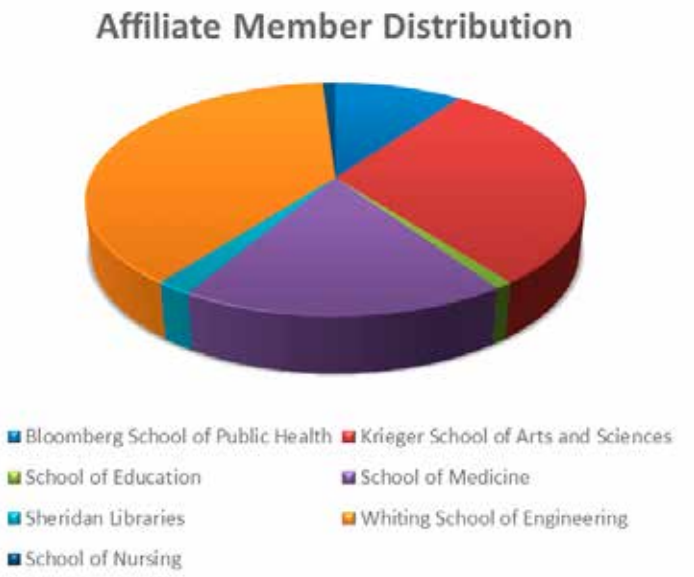
SciServer, an NSF funded data-infrastructure project sponsored and administered by IDIES went live in May, 2016, providing a collaborative environment for data-intensive research.

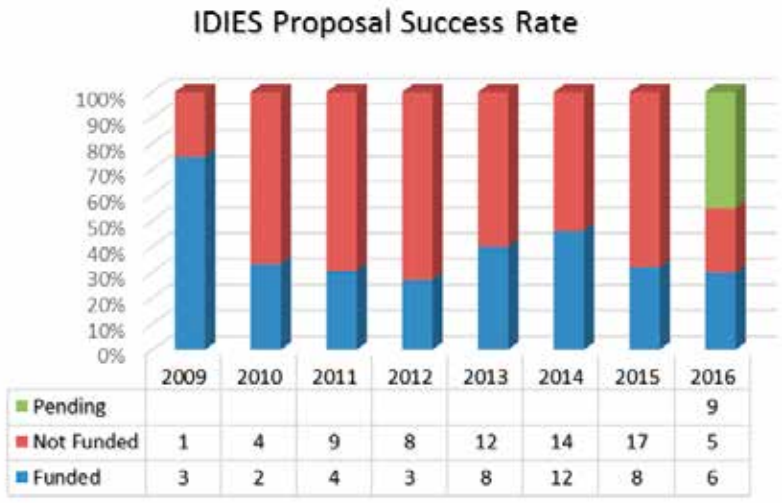
...

IDIES Director Alex Szalay gave a Keynote at the #DataDriven Microsoft Event, which featured the Sloan Digital Sky Survey database.

IDIES IN NUMBERS

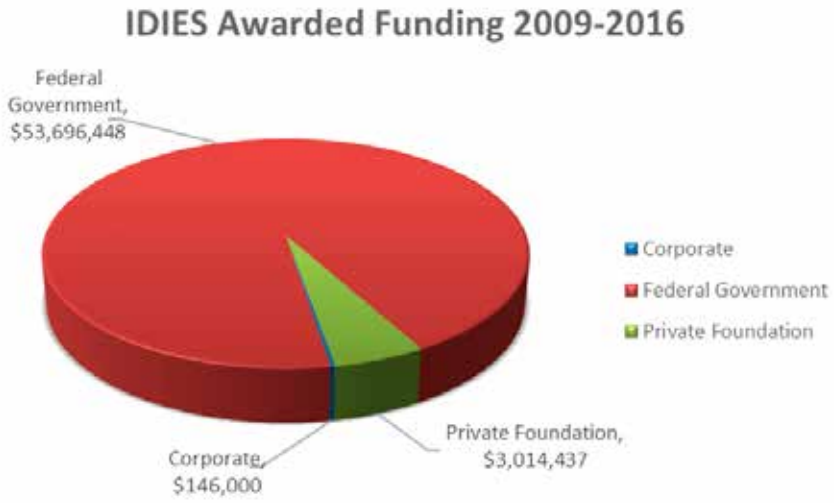
In fiscal year 2016 (July 2015 through June 2016), IDIES experienced continued growth while supporting the IDIES vision of facilitating high performance and data intensive computing across all of JHU. We saw a 7% growth in affiliate members, and 20% of 2016 proposal submissions came from new IDIES affiliates. IDIES FY16 proposal submissions kept up with our positive trend, totaling 20 submissions for the year. FY16 SEED funding was awarded to two new and two existing IDIES affiliates. SEED awardees have been submitting their SEED research ideas to external agencies for additional support and we continue to see positive results.





Government agencies continue to be the primary source of sponsored funding for IDIES, followed by awards from charitable foundations and industry. Our primary federal support is through grants from the National Science Foundation. As of the close of FY16, of the 125 sponsored application submitted through IDIES, 40% have been awarded. As big data continues to rise with increasing importance to the forefront across all research disciplines, IDIES hopes to expand and diversify its collaborations and partnerships with additional government agencies, private foundations, and industry partners.

As the prevalence of Big Data grows in all areas of research, the focus is growing as well to include training and additional outreach for the upcoming generation of researchers. In addition to applying to new funding opportunities from NIH and NSF targeting the development of training programs, IDIES is expanding its outreach activities this year with several new workshops and seminar series. Make sure you visit our website to check them out!



SEED FUNDING AWARDS

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.

Spring, 2016

Scalable Framework for Statistical Inference on Big Multimodal Data via Sketching and Concentration

Vladimir Braverman, PhD, (Department of Computer Science) and Carey Priebe, Professor, (Applied Mathematics and Statistics)

The goal of this project is to develop scalable frameworks for statistical inference of multimodal high dimensional, large and complex data. For example, community detection for large graphs can often be approximated on a sketched version of the given graph, in which the inherent dimension is much smaller than the number of vertices.

Developing comprehensive machinery for multi-sample multimodal inference problems involving graphs is of both theoretical and practical importance. For instance, testing for similarity across brain graphs is an area of active research at the intersection of neuroscience and machine learning, and practitioners often use classical parametric two-sample tests, such as edgewise t-tests on correlations or Mantel tests, or permutation tests on subgraphs, as approaches to graph comparison. Related examples include multimodal inference problems of regression or testing for independence between a subject attributes, such as the Creativity index score [2], or a time series of stimulation and neuronal activities, against the subject's brain graph. Many existing inference procedures either do not take into account the special topological structure inherent in the graphs, or incorporate only simplistic features and use oft-irrelevant graph differences such as edge density.

Co-PI Priebe and his coauthors recently provided the first principled approach to an important multi-sample multimodal inference problem for random graphs, namely the problem of determining whether a collection of random graphs, not necessarily having the same vertex set or even the same number of vertices, are from equivalent distributions [3]. In contrast to two-sample or multi-sample testing in classical statistics, the number of observations (graphs) are usually small as small as two in many applications of interest but the complexity of each observation (graph) is large. For example, a large-scale brain graph generated from a MRI pipeline could have on the order of 10^5 vertices and 10^8 edges.

We propose to develop efficient algorithmic tools that, collectively, we call Sketching and Measure Concentration for Data Analysis (SMCDA). Informally, SMCDA presents solutions that scale polylogarithmically in data size by discarding the vast majority of data while approximately preserving the information of interest. The correctness of SMCDA stems from deep and well-established mathematical theories such as Local Theory of Banach Spaces. The connection between these theories and analysis of massive data has become clear only recently [1].

References

[1] V. Braverman, S. R. Chestnut, R. Krauthgamer, and L. F. Yang. Streaming symmetric norms via measure concentration. CoRR, abs/1511.01111, 2015.

[2] R. Jung, J. H. Segall, H. J. Bockholt, R. A. Flores, S. M. Smith, R. S. Chavez, and R. J. Haier. Neuroantomy of creativity. Human Brain Mapping, 31:398{409, 2010.

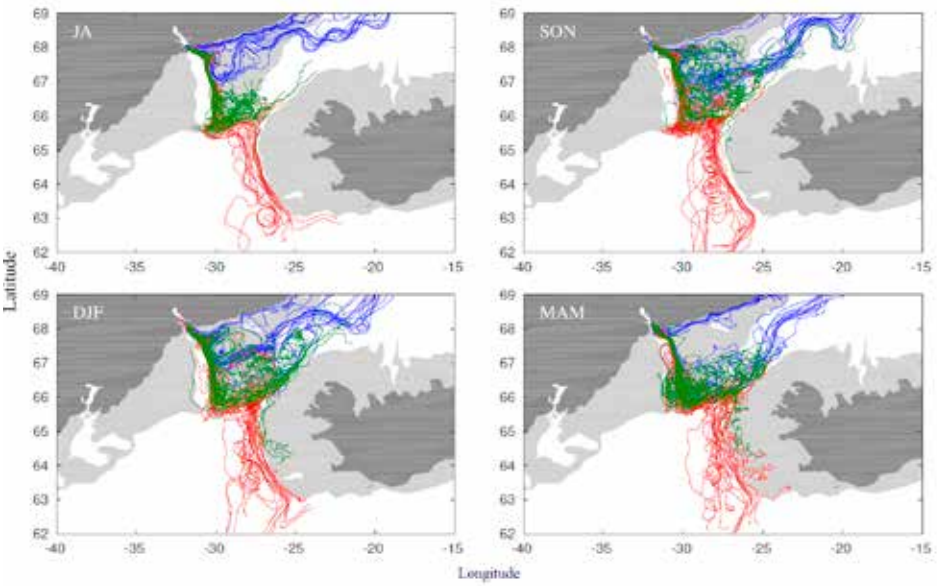
[3] M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe. A nonparametric two-sample hypothesis testing problem for random dot product graphs. Bernoulli, 2016.

Towards the Johns Hopkins Ocean Circulation Database: Method Development and Prototype

Thomas Haine and Renske Gelderloos (Department of Earth & Planetary Sciences), Alex Szalay, and Gerard Lemson (Department of Physics & Astronomy)

Numerical ocean circulation models are getting more and more realistic, but tools to analyze the model output are still primitive. We are using this seed fund award to develop a prototype environment that will make ocean-model output analysis easy, fast, and (eventually) accessible to the general public. We envision tools to enable on-the-fly analysis of ocean-model output on a computer cluster. They will advance the post-processing of ocean model output, which will in turn increase its scientific value.

Over the past few months we have improved the implementation of our in-house numerical particle tracking algorithm, which enables a Lagrangian (drifting in space) analysis of the Eulerian (fixed in space) model output. Substantial speedup has been achieved with these improvements, as reported by Gelderloos et al. (2016a). The improvements will enable particle tracking in global high-resolution model solutions over several years, whereas the studies are now limited to regional configurations and a time span of weeks to months, like the one displayed in the Figure.



Apart from Lagrangian methods, we also propose to develop tools that improve the analysis capabilities of the Eulerian model output. As ocean model output is often highly structured, database methods can be used to optimize the retrieval of subspaces of the four-dimensional (space-time) output fields. The improved field retrieval will accelerate the subsequent calculation of diagnostic quantities, such as time series of heat content in a specific subspace of the model ocean domain, or energy fluxes across a temperature isosurface. The tools will be made available via an ePortal, with different levels of complexity for general and expert users.

In the figure above you see the Ocean Lagrangian particles trajectories for flow between Greenland and Iceland. The particle trajectories are integrated backwards in time from the Kangerlugssuaq Fjord, which drains a large part of the southeast Greenland Ice Sheet and whose discharge flux is accelerating. Each panel is a different season (JA = July–August, SON = September–November, DJF = December–February, MAM = March–May). The shading indicates the fraction (%) of particle trajectories approaching Kangerdlugssuaq Fjord that originate from the Arctic (blue; cold and fresh water), Atlantic (red; warm and salty water), and the sill in the shallow Greenland–Iceland ridge (green; cool and salty water). Only values exceeding 1% are shaded. Dots show two representative 90-day trajectories in each case. The pathways are 3-dimensional, but are projected onto the horizontal plane in this figure. From Gelderloos et al. (2016b).

References:

R. Gelderloos, A. S. Szalay, T. W. N. Haine, and G. Lemson. A fast algorithm for neutrally-buoyant Lagrangian particles in numerical ocean modeling. IEEE 12th International Conference on eScience, 2016a.

R. Gelderloos, T. W. N. Haine, I. M. Koszalka, and M. G. Magaldi. Seasonal variability in warm-water inflow towards Kangerdlugssuaq Fjord. J. Phys. Oceanogr., in review, 2016b.

Metabolic Compass: A Mobile Health Platform for Understanding the Impact of Circadian Sleeping, Eating and Exercise Behaviors on Metabolic Syndrome, and Obesity

Jeanne Clark (SOM – General Internal Medicine), Thomas Woolf (Physiology & Computer Science), Yanif Ahmad (Computer Science)

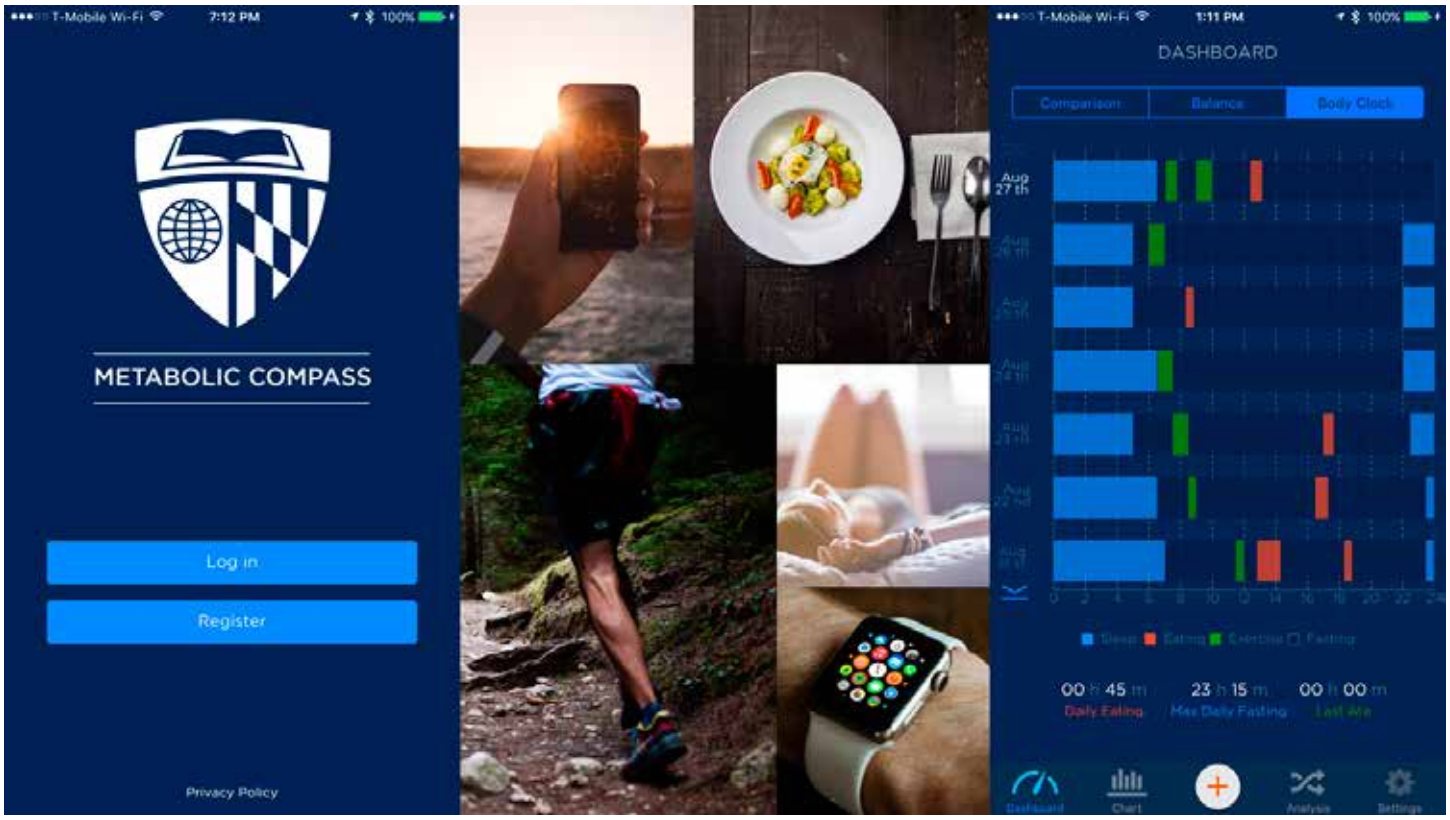
Circadian rhythms drive much of our health, but are not well understood. Our App, Metabolic Compass, is designed to collect information on when we eat, when we sleep, and when we exercise. These events are displayed for users to quickly see their daily behavior within a weekly context. Our long-term goal is an understanding of how daily timing decisions impact our health.

Our IDIES Seed award has enabled the development of our App for iOS/Android. We use Apple's frameworks, ResearchKit and HealthKit, for electronic consent and for the data sharing of more than 60 health-metrics. Our cloud-based (Amazon Web Services) backend lets us provide each user with a streaming data update for how their data compares against others in the study. We provide the ability to filter the comparison data and to see weekly and monthly statistics on study progress.

Data entry is directly within the App, and can be passively added via wearables and voice commands. The engineering of rapid data entry, to encourage participants to stay with the App, was a primary goal. In addition notifications can be adjusted to remind users to enter data.

Wearables like the Apple Watch passively record heart-rate. Voice-activated input, like Amazon Alexa (and in the future Google Home), provides in-home data entry along with reminder and analysis queries.

One proposal has been submitted to the NIH, with an American Heart Association and a revised NIH proposal being planned based on the initial support provided by IDIES. Having strong preliminary data, showing that we can mount this study, is central to our argument that we should be funded to perform these national studies. We are also pursuing the development of a wearable ketone sensor. The wearable data will be fully integrated with the iOS/Android platforms and will enable users to continuously see their ketone values. This should have utility for Type I diabetic patients in providing early warnings against diabetic ketoacidosis. We plan to have a small clinical trial on the ketone wearable underway within the next 3-6 months.

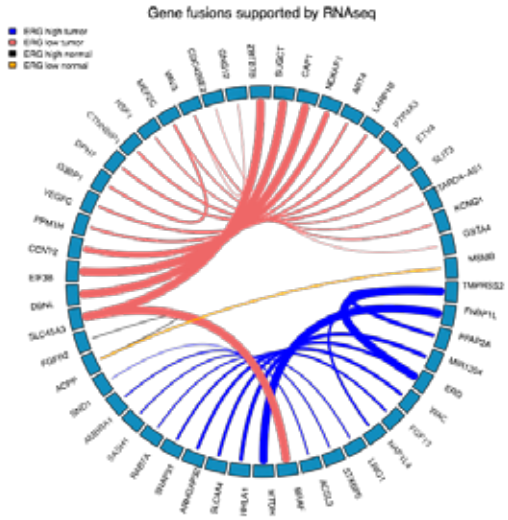


In pursuit of DNA and RNA amalgamations

Sarah J. Wheelan, MD, PhD, (Institute of Genetic Medicine) and Michael C. Schatz, PhD, (Department of Computer Science)

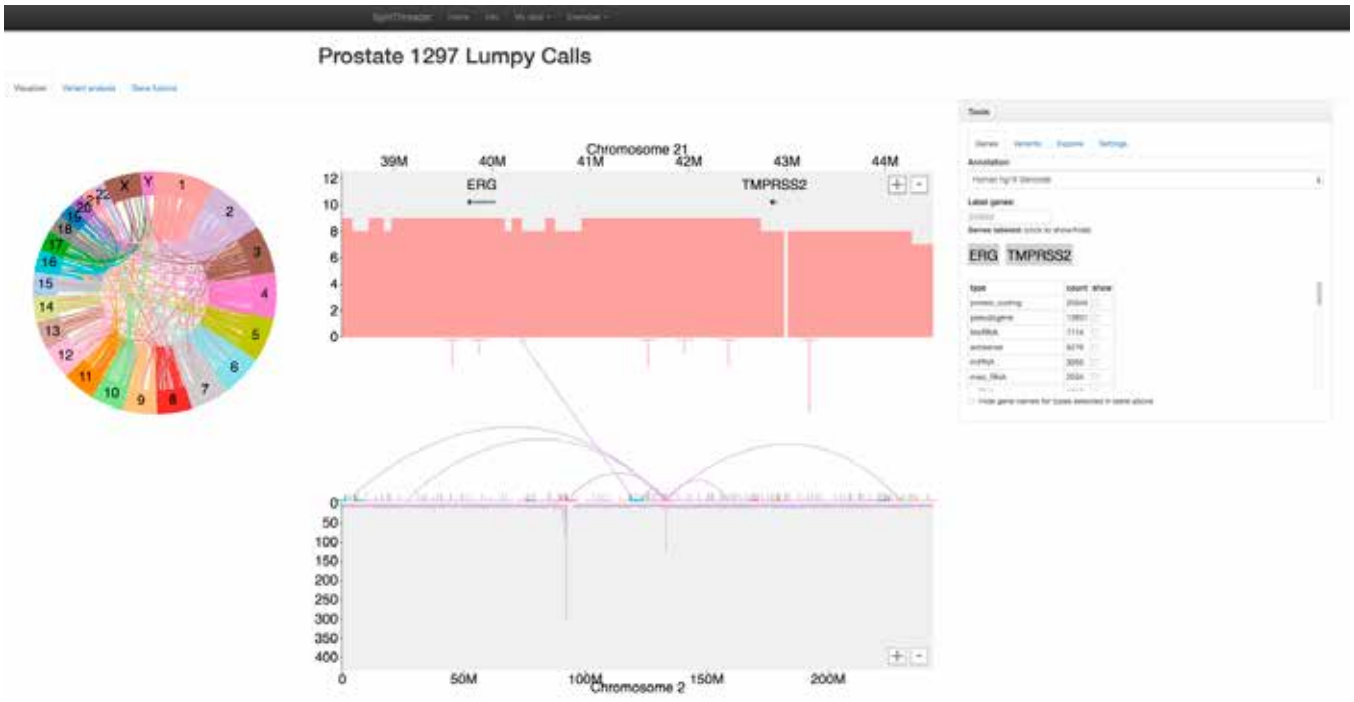
Interestingly, prostate tumors are very often characterized by gene fusions, making this a perfect set of specimens in which to begin our search. In fact, we can be nearly certain that when the ERG gene is highly expressed, there is a gene fusion, as from historical data this overexpression has always been associated with DNA rearrangements.

Using existing state-of-the-art methods and new methods that we have recently developed, we have analyzed the sequencing data of these samples. We are able to identify a number of RNA-level fusions in these samples (Figure 1), but, interestingly, we have not yet found obvious underlying DNA rearrangements.



Gene fusions shown by RNA-seq in the tumor and normal prostate samples

Importantly, to state that an RNA fusion does not derive from a genomic rearrangement, we need to be extraordinarily sensitive in our search for DNA rearrangements and extraordinarily specific in our search for RNA-fusions. We control for these factors by using a new ensemble approach that we call SURVIVOR in which we run multiple analysis programs at once and then use the consensus of those calls. We find that this significantly improves the specificity of the analysis while maintaining a high level of sensitivity. Nevertheless we have not found the expected DNA rearrangements for several strongly supported RNA-level fusions (Figure 2). We are currently evaluating these candidates informatically, and then a selection of the top candidates will be further validated using experimental techniques. Once optimized, we aim to apply this approach to an even larger collection of samples to analyze the processes that lead to the formation of RNA-fusion genes.



SplitThreader visualization of prostate sample 1297. The RNA-seq analysis showed a strongly supported ERG-TMPRSS2 gene fusion, but is unsupported in the genomic data except for a few small deletions on chr21 and a candidate chromosome fusion between chr21 and chr2.

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, visitors, and so on.

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

NSF

JOHN TEMPLETON FOUNDATION

NIH National Institutes of Health

intel

ALFRED P. SLOAN FOUNDATION 1934

NOKIA

GORDON AND BETTY MOORE FOUNDATION

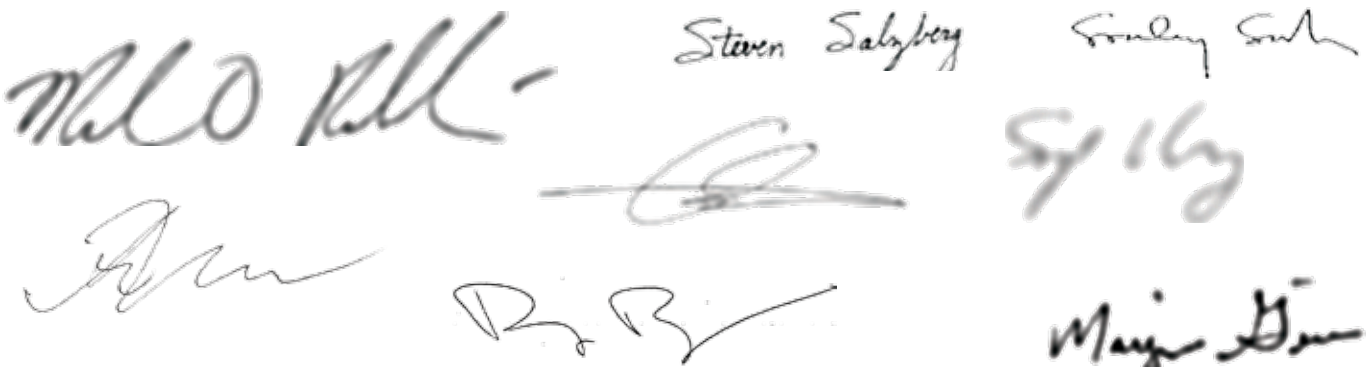
Microsoft

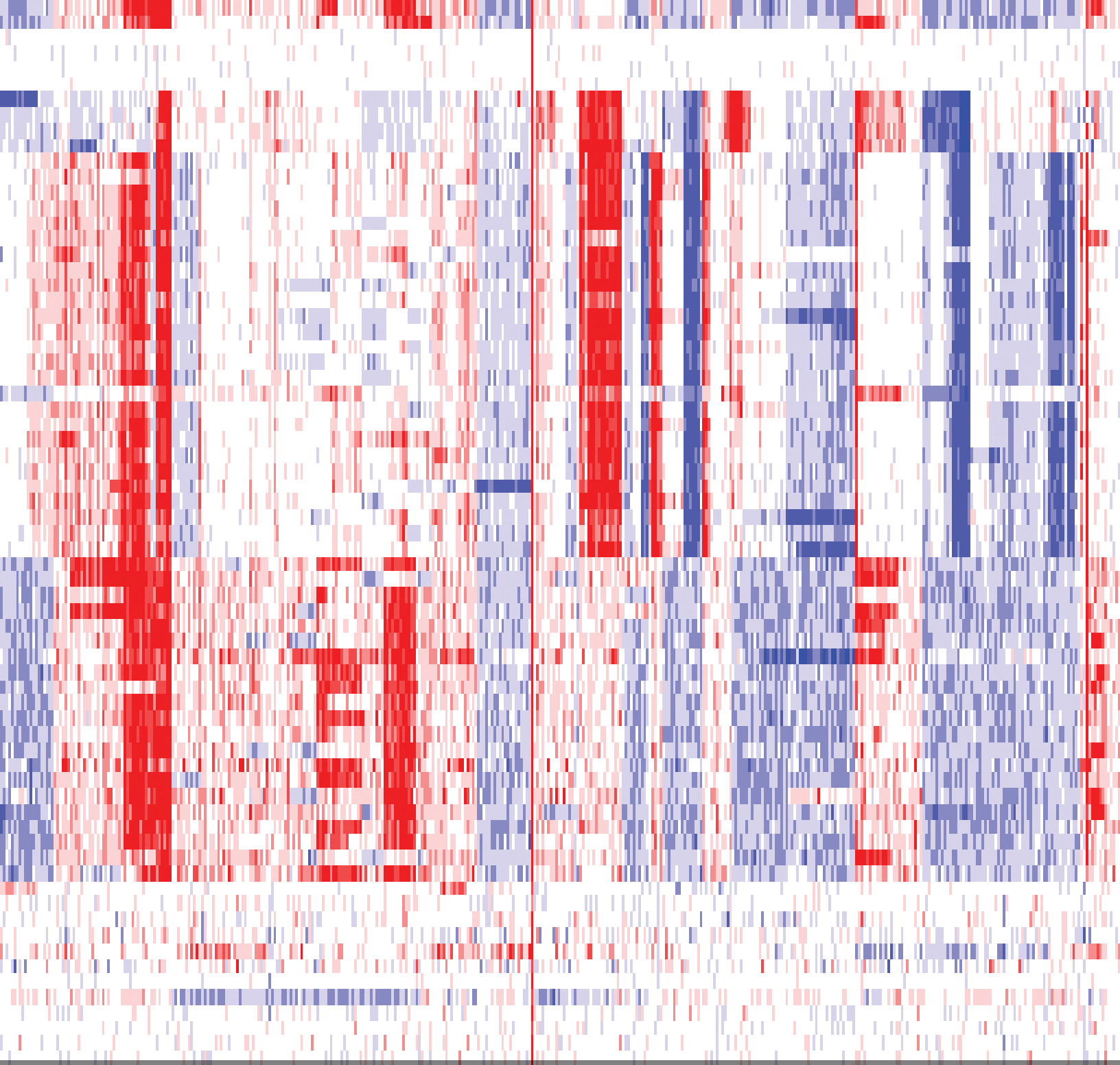
NVIDIA

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, Roger Peng, & Margie Gier





JOHNS HOPKINS

INSTITUTE FOR
DATA-INTENSIVE
ENGINEERING & SCIENCE

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