October 1, 2014 - September 30,

2015

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.

Today IDIES involves more than 97 faculty and more than a hundred graduate students. Over the last 12 months we have awarded 7 seed grants in a broad spectrum of topics, connecting researchers from different fields, but sharing a common interest in Big Data.

In two 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.

October 1, 2014 - September 30, 2015

The Institute for Data Intensive Engineering and Science Annual Review

SYMPOSIUM
- Agenda 1
- Keynote Speakers 2
- Speakers 3

NEWS & ANNOUNCEMENTS 5

IDIES
- Mission Statement 11
- Seed Funding Awards 12
- IDIES in Numbers 16

On the cover: Lines tangent to the vorticity vector field in isotropic hydrodynamic turbulence. Data generated from direct numerical simulation, and made available as part of the Johns Hopkins Turbulence Databases, an open numerical laboratory for turbulence research. Visualization by Kai Schneider.

AGENDA

2015 Institute for Data Intensive Engineering and Science (IDIES)

Annual Symposium
October 16, 2015 8:00 - 5:00
Mudd Hall Auditorium, Room 26

8:00 AM	Continental Breakfast - Check In
9:00 AM	*Opening Remarks* **S. Alexander Szalay, PhD**, Director of IDIES, Professor of Astrophysics & Computer Science, Krieger School of Arts & Sciences, Johns Hopkins University
9:20 AM	*MARCC: Cutting-Edge Technology for Data-Intensive Research* **Jaime E. Combariza, PhD**, Director of MARCC, Associate Research Scientist, Department of Chemistry, Krieger School of Arts & Sciences, Johns Hopkins University
9:40 AM	*Urban Planning in Baltimore City* **Tamás Budavári, PhD**, Assistant Professor, Department of Applied Mathematics & Statistics, Whiting School of Engineering, Johns Hopkins University
10:00 AM	Break
10:20 AM	KEYNOTE SPEAKER *The Materials Genome - Science in the Information Age* **Kristin Persson, PhD**, Assistant Professor, University of California at Berkeley
11:10 AM	*A Modeling Enabled Database for Aneurysm Hemodynamics and Risk Stratification - An Automated Method for Computational Modeling of Aneurysm Hemodynamics* **Jung Hee Seo, PhD**, Associate Research Scientist, Department of Mechanical Engineering, Whiting School of Engineering, Johns Hopkins University
11:30 AM	Poster Session, Mudd Hall Commons - Upper Level
12:00 PM	Lunch
1:00 PM	*Big Data Regression for Predicting Genome-wide Regulatory Element Activities* **Hong Kai Ji, PhD**, Associate Professor, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University
1:20 PM	*An interactive System to Optimize Statistical Seasonal Forecasts and Climate Change Projections* **Benjamin Zaitchik, PhD**, Assistant Professor, Department of Earth & Planetary Sciences, Krieger School of Arts & Sciences, Johns Hopkins University
1:40 PM	*Statistical Methods for Real-Time Monitoring of Physical Disability in Multiple Sclerosis* **Vadim Zipunnikov, PhD**, Assistant Professor, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University
2:00 PM	*Cost-Sensitive Prediction: Applications in Healthcare* **Daniel Robinson, PhD**, Assistant Professor, Department of Applied Mathematics & Statistics, Whiting School of Engineering, Johns Hopkins University
2:20 PM	Break
2:40 PM	KEYNOTE SPEAKER Big Data – an NIH Perspective **Philip Bourne, PhD**, Associate Director for Data Science, National Institutes of Health
3:30 PM	Closing Remarks **S. Alexander Szalay, PhD**, Director of IDIES, Professor of Astrophysics & Computer Science, Krieger School of Arts & Sciences, Johns Hopkins University
3:40 PM	Poster Session (Continued) & Cocktail Hour, Mudd Hall Commons - Upper Level

KEYNOTE SPEAKERS

KRISTIN PERSSON, PhD — Assistant Professor, University of California at Berkeley

Kristin Persson leads the Materials Project at the Lawrence Berkeley National Laboratory. She is director of the 2012 BES-funded "Materials Project Center for Functional Electronic Materials Design". Professor Persson is a PI of the Crosscutting Thrust of the recently launched JCESR hub, as well as the Batteries for Advanced Transportation Technologies (BATT) program; She is co-founder of the clean-energy start-up Pellion Technologies Inc.

PHILIP BOURNE, PhD — Associate Director for Data Science, National Institutes of Health

Philip E. Bourne of the National Institutes of Health: Associate Director for Data Science (ADDS) at the NIH; former Associate Vice Chancellor for Innovation and Industry Alliances; Professor in the Department of Pharmacology and Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California San Diego; Associate Director of the RCSB Protein Data Bank; Adjunct Professor at the Sanford Burnham Institute.

SPEAKERS

TAMÁS BUDAVÁRI, PhD

Dr. Budavári's research interests include computational statistics,Bayesian inference, low-dimensional embeddings, and streaming algorithms. He employs parallel processing on GPUs to improve processing times.

JAIME COMBARIZA, PhD

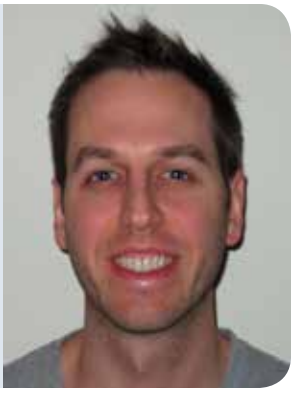
Director of the new Maryland Advanced Research Computing Center, a shared computing facility located on the Bayview Campus of Johns Hopkins University and funded by a State of Maryland grant to Johns Hopkins University through IDIES. MARCC is jointly managed by Johns Hopkins University and the University of Maryland College Park.

HONGKAI JI, PhD

Dr. Ji is interested in developing statistical and computational methods for analyzing high-throughput genomic data. He applies these tools to study gene regulatory programs in development and diseases.

DANIEL ROBINSON, PhD

Daniel designs, analyzes, and implements algorithms for large-scale optimization and complementarity problems. Applications of current interest include real-time optimization in energy systems, subspace clustering in computer vision, and predictive modeling in healthcare.



JUNG-HEE SEO, PhD

Dr. Seo's research expertise is in the areas of computational fluid dynamics and acoustics. He is currently conducting research on multi-physics computational modeling and analysis of cardiovascular flows.

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.

BENJAMIN ZAITCHIK, PhD

Dr. Zaitchik's research addresses problems of regional climate variability, water resource monitoring, disease early warning, and climate change adaptation. Prior to joining Johns Hopkins he worked in the NASA Hydrological Sciences Branch and as a Foreign Affairs Officer in the U.S. State Department.

VADIM ZIPUNNIKOV, PhD

Vadim Zipunnikov is interested in real-world data applications, especially in brain imaging and wearable computing. As a member of the Statistical Methodology and Applications for Research in Technology Working Group he is constantly involved in collaborating on new problems with scientific teams.

Maryland Advanced Research Computing Center Offers Bigger Home for Big Data Projects

New Center Gives Scholars from Johns Hopkins, U. of Maryland Far Greater Digital Power

Whether they're studying distant galaxies, deadly diseases deep within human cells or trying to understand brain functionality, Big Data researchers increasingly need more computational power and more digital storage space. To address this demand, two Maryland universities have built one of the nation's largest academic computing centers, located at the edge of the Johns Hopkins Bayview Medical Center campus in Baltimore.

Supported by $30 million in state funding, the Maryland Advanced Research Computing Center (MARCC, pronounced "MAR-see") provides state-of-the-art digital processing power to a wide array of researchers at Johns Hopkins University and the University of Maryland, College Park.

Thanks to speedy fiber-optic cable connections to the participating campuses, Big Data university researchers in their labs or offices tap into the new computing center. "Everyone is going to be able to access the new facility on a remote basis," said Jaime Combariza, a Johns Hopkins computational chemist who became director of MARCC in June of last year. "MARCC allows all of Johns Hopkins and the University of Maryland to centralize and more effectively share their computing power."

For participating researchers, he said, the arrangement should lead to significant cost savings and greater efficiency. Instead of requiring individual research groups to use time, money and space to create their own small data centers, all participants will share the costs of space, cooling, networking and operation of a the single, more capable, data center.

The shared equipment within the nondescript 3,786-square new building is capable of delivering a hefty digital punch. The initial configuration includes 19,000+ processors and 17 petabytes of storage capacity–that's 17 million gigabytes. The combined theoretical computing capability is about 900 TFLOPs (900x1012 Floating-point Operations Per Second). MARCC provides different computing environments to match a wide variety of needs: traditional CPU time-intensive computing, high throughput computing, large memory needs, use of accelerators, and large amounts of space for data intensive analytics.

Currently over 115 research groups have requested allocations and close to 400 individual accounts have been created, spanning researchers from the Bloomberg School of Public Health, Krieger School of Arts and Sciences, School of Medicine and Whiting School of Engineering at Johns Hopkins University, and scholars from the University of Maryland, College Park. In the near future we expect many other researchers in the State of Maryland to be able to access MARCC.

The users include astrophysicists who grapple with vast amounts of celestial data from powerful telescopes. Scholars from biophysics, material science, chemistry, engineering, brain science, genomics and many other diverse disciplines are using MARCC to analyze their large datasets.

Alex Szalay, a professor in the Krieger School's Department of Physics and Astronomy who pioneered the use of Big Data in sky-mapping projects, has also begun to apply his expertise to biomedical research. One new project, slated to run on MARCC computers, involves newly designed software, written in collaboration with scientists from the McKusick-Nathans Institute of Genetic Medicine and the Department of Computer Science. The software is designed to perform demanding genetics tasks. With MARCC on board, computations that used to take a day to complete will finish in less than an hour, enabling Szalay's team to crunch several hundred genomes' worth of data in a matter of days.

Many other Big Data projects in biology and medicine have become popular, and they also require significant computing resources. For example, just one experiment comparing gene activity in two types of tissue generates 30 to 40 gigabytes of data. A simulation of the workings of the heart generates one terabyte of data. A single MRI or CT scan creates one-to-two terabytes. MARCC will speed up the completion of studies involving such information.

Natalia Trayanova, a Johns Hopkins professor of biomedical engineering, leads a team that creates complex simulations of the heart, using everything from MRIs to the latest information on heart-specific proteins. Her team currently uses several HPC resources at Johns Hopkins Homewood campus and often must wait for enough processors to become available. If Trayanova's team needs 10 nodes and only nine are available, they have to wait. Now, with thousands of processors shared in a central location, the turnaround time will be drastically reduced.

As a shared resource, MARCC provides a local flexible environment for computational scientists allowing them to successfully advance their research agendas, foster new collaborations and maintain a competitive edge over most other academic institutions.

This article was adapted from a story by Phil Sneiderman and Cathy Kolf, published in the JHU HUB 07/06/2015.

SciServer

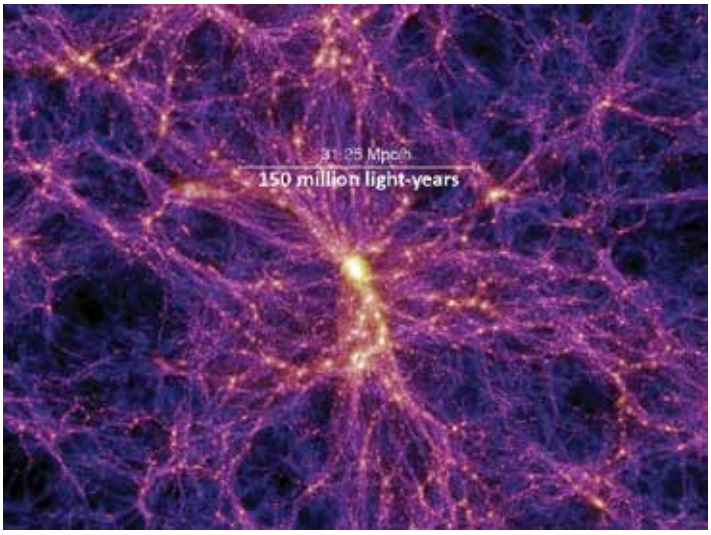
A Scientific Ecosystem Based on Collaborative, Data-Driven Science

With support from the National Science Foundation, IDIES is meeting the challenge of Big Data through an ambitious project called SciServer. Built around several large datasets for the broader science community, this 5 year program will enable a new scientific approach in which data, analysis tools, and compute capabilities are linked into one framework.

Now nearing the end of its second year, SciServer is developing a framework of reusable tools in a sustainable collaborative ecosystem. SciServer grows out of our Institute's expertise with the Sloan Digital Sky Survey (SDSS), an ongoing project to make a digital map of the Universe in unprecedented detail. Building on our tools for SDSS, we are working with communities in Turbulence, Cosmology, Genomics, Earth Science, and Oceanography to develop, deploy and serve scientific data in new ways.

Our strategy is to create scalable big data "building blocks" with portability, generality, and economies of scale. SciServer also helps other institutions and communities deploy services and applications to rapidly test new ideas.

So far, we have completed SciServer's key initial objectives – to rewrite and extend existing SDSS online data access tools (SkyServer and CasJobs), and to integrate SkyQuery, a powerful system for cross-matching data across multiple astronomical surveys in multiple wavelengths. As we expand our efforts to support the global astronomy community, we are also developing new tools to benefit all our science collaborations.

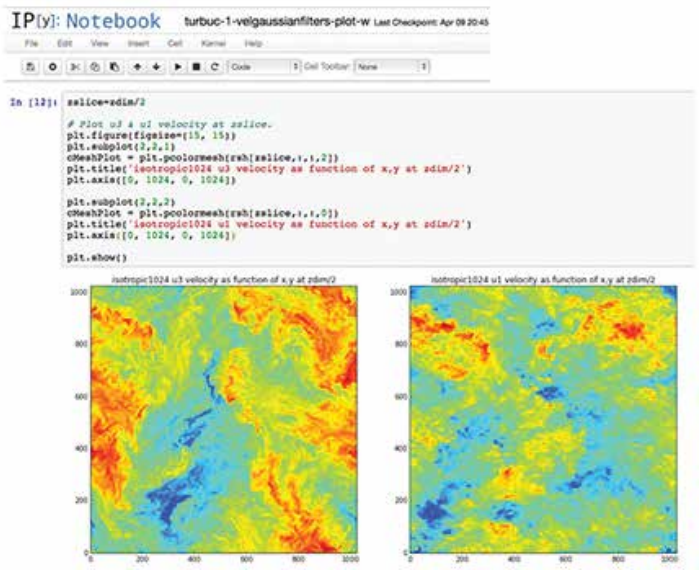


Cosmology: Dark matter distribution from the Millennium Simulation

We have integrated a new distributed data storage and sharing application called SciDrive, and prototyped the integration of Turbulence, Cosmology, Genomics and Earth Science data. Most excitingly, we are developing "Open Numerical Laboratories" to support innovative new approaches to analysis. We have developed SciScript, a set of tools based on Python and Matlab. SciScript, combined with the large temporary working space SciServer provides, allows researchers to perform both queries and processing on datasets far larger than ever possible before. All these services are available through a Keystone-based single sign-on login portal.

The SciServer system is now available for testing, and will continue to grow over the next three years.

The SciServer Framework



iPython: A user-created filtering script and its results

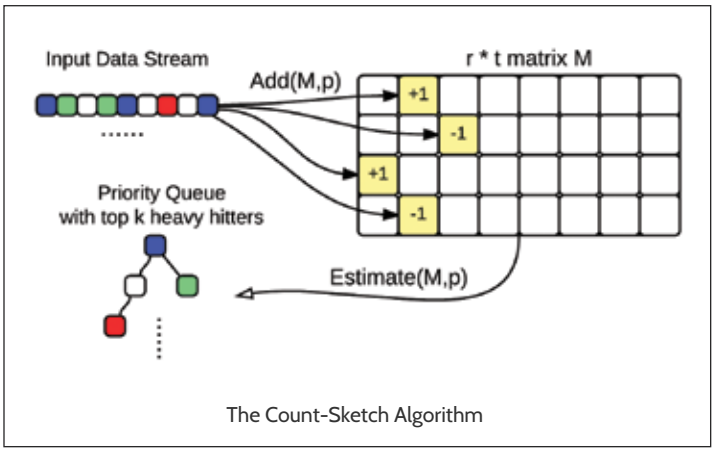
SciServer investigators include IDIES Director Alexander Szalay, IDIES Associate Directors Steven Salzberg, Ani Thakar, and Charles Meneveau, and IDIES affiliates Randal Burns and Michael Rippin.

SciServer is supported by National Science Foundation award # ACI-1261715.

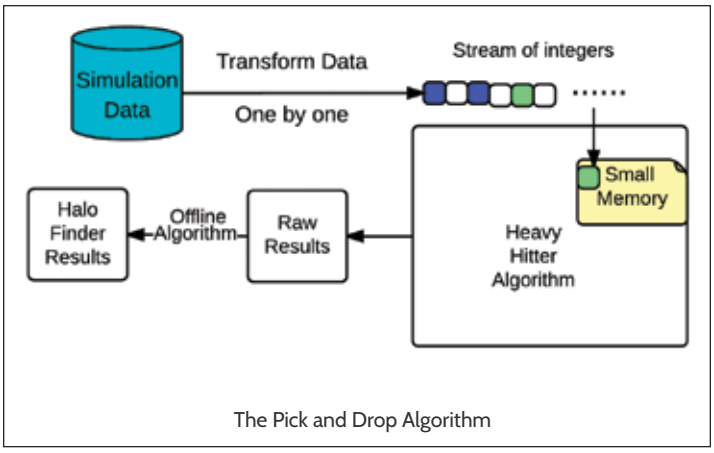
Clustering Algorithms for Data Streams and Related Applications

Cosmologists and Computer Scientists Collaborate on Scalable Algorithms for Big Data

Cosmological simulations are among the largest computer experiments currently run in science, the largest of them currently using over a trillion particles to represent the matter content in synthetic universes that model significant subsets of the observable universe. The traditional way of analyzing their results generally requires machines with memory sufficient to hold complete snapshots of the data. At 10s of Terabytes this can only be achieved on machines similar to the parallel super-computers capable of producing the simulations in the first place and is open to only the handful of specialist programmers of the simulation codes. This, together with the fact that I/O is poses a severe performance bottle neck at these data volumes has meant that the analysis of recent simulations has been lagging behind their production more and more.

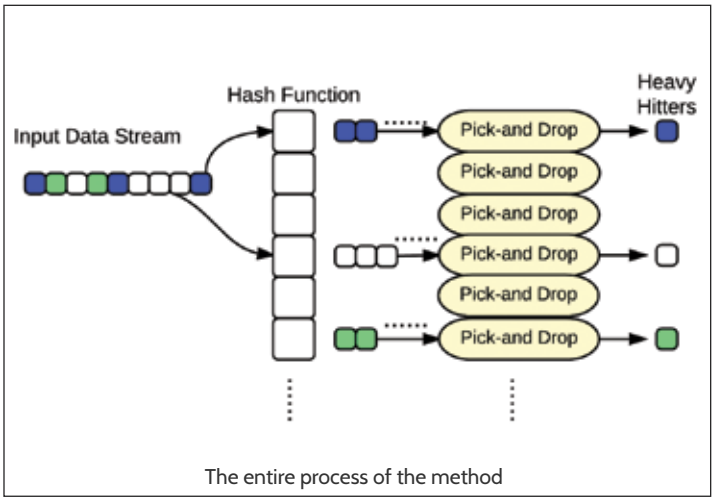


The Count-Sketch Algorithm

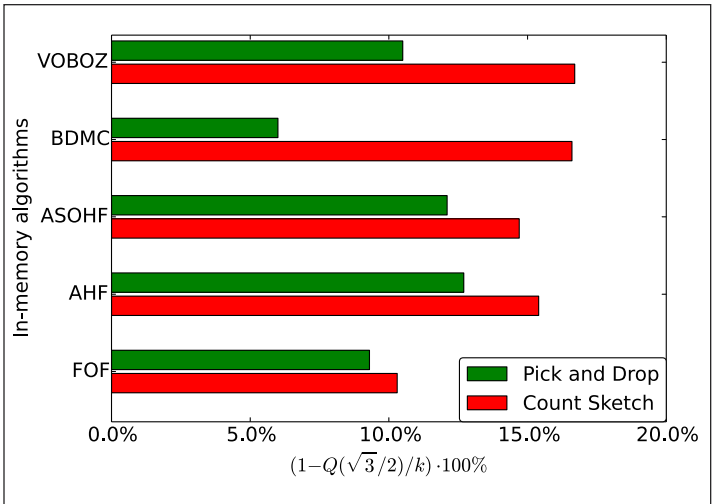


The Pick and Drop Algorithm

One way to address this problem is to try to find algorithms that put lower demands on resources whilst still producing scientifically meaningful results. We have started a collaboration between the CS and Cosmology groups at JHU with the specific aim of analyzing streaming and randomized algorithms. In particular we want to apply these to clustering problems. An important first step in much of the analysis of cosmological simulations is to identify clusters of particles and derive properties of these. Analyzing such cluster catalogues, which generally are 2-3 orders of magnitudes smaller in size than the raw data, is more feasible and on themselves they provide valuable starting points for further analysis and model comparison.



The entire process of the method



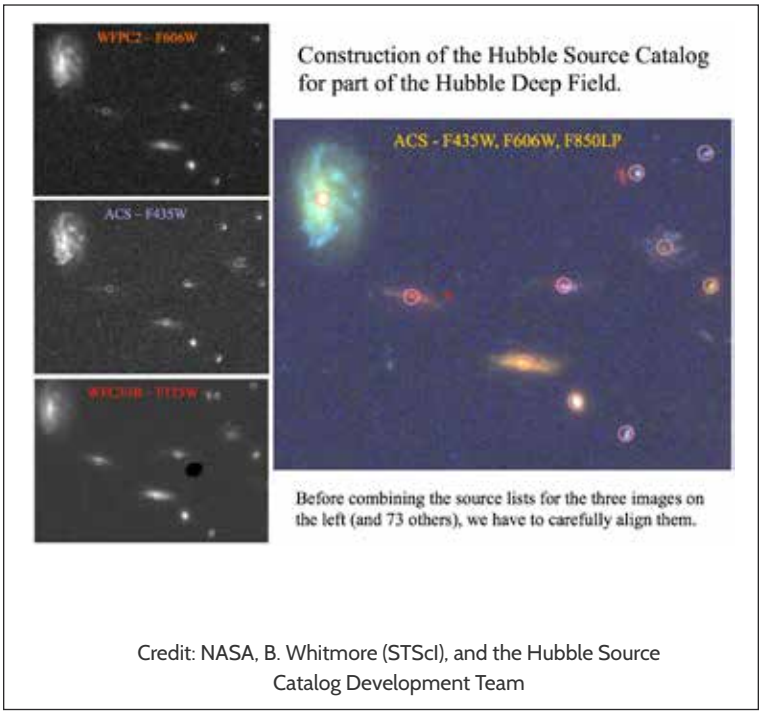
Error rate comparison between different algorithms

In a first attempt we have used two algorithms, known as Count-Sketch (Figure 1) and Pick-and-Drop (Figure 2) to find the 1000 most massive clusters. Comparisons to traditional "offline" implementations shows that our randomized streaming approach finds more than 90 percent of all these 1000 most massive clusters, using a small fraction of the memory. This is significant step forward bringing scalable methods of computations to cosmological simulations.

This project is funded by NSF Award # 1447639 - Principal Investigator - Vladimir Braverman; Co-Principal Investigators - Alexander Szalay, Randal Burns, Tamás Budavári, Benjamin Van Durme; Sr. Personnel - Gerard Lemson, Mark Neyrinck.

Hubble Source Catalog

One-Stop Shopping for Astronomers



Credit: NASA, B. Whitmore (STScI), and the Hubble Source Catalog Development Team

Researchers at IDIES, together with astronomers at the Space Telescope Science Institute, have created a new "master catalog" for astronomy called the Hubble Source Catalog. The new catalog provides one-stop shopping for measurements of objects observed with NASA's Hubble Space Telescope.

"The Hubble Source Catalog is arguably the Hubble Space Telescope's ultimate legacy," said Tamas Budavari of the Johns Hopkins University, IDIES affiliate and member of the catalog's development team. "Not only is it a one-stop shop, but it's the first place to go. It's the table of contents for most Hubble observations. If a zillion investigators pointed Hubble in the same direction and took images in different wavelengths, we have now put all those observations together in one place."

The Hubble Space Telescope has amassed a rich legacy of images and other scientific data over its 25 years of exploring the universe. All its observations are available freely online in an archive containing more than a million images showing about 100 million objects, from distant galaxies to compact star clusters to individual stars. For astronomers, however, just sifting through this gold mine of data for their objects poses a major challenge. The Hubble Source Catalog now allows astronomers to perform a computer search for characteristics of these objects, receiving the information they need within seconds or minutes.

This graphic shows an example of how the Hubble Source Catalog was constructed for a small part of the Hubble Deep Field. The catalog includes data from 76 separate images, three of which are shown on the left side of the graphic. The three images come from three different instruments on the Hubble Space Telescope: the Wide Field Planetary Camera 2 (top), the Advanced Camera for Surveys (middle), and the Wide Field Camera 3 (bottom), which uses infrared light. The larger image on the right is the combined image created for the Hubble Source Catalog. Note that the objects in the original images are not perfectly aligned. Specially-developed software was created to align sources before making the final version of the catalog.

Salzberg and Colleagues Find Genome Databases Contain Cross-Species Contamination

Samier Merchant and co-authors Derrick Wood and IDIES Associate Director Steven Salzberg have found evidence in public genome databases of unnoticed and unexpected errors. In their November 2014 paper, Unexpected cross-species contamination in genome sequencing projects, published in PeerJ, they show that many putatively 'complete' genomes, including Neisseria gonorrhoeae TCDC-NG08107, contain cross-species contamination. Their work indicates that genomes in databases may contain sequences from other species, even other kingdoms, such as mammalian DNA in bacterial sequences.

The errors Salzberg's group found were on the order of a few base pairs per million - a small number with big potential to cause problems for researchers. The errors may have been introduced during sampling, preparation, or during the computational process of sequence assembly. Within the Neisseria genome that Salzberg's group focused on, the importance of the errors grew due to the incorrect submission of the genome as a 'complete' genome rather than a draft genome. Such mislabeling of the genome's status can give researchers a greater than justified confidence in the genome's correctness. These findings underscore a danger researchers unknowingly face when comparing their own results against contaminated genomes: treating complete, or even partial, genome sequences from databases as validated. Against this flawed standard, researchers unwittingly ascribe deviations from the norm as originating in their own samples, leading to confusing and erroneous results. While the results of this work led to the removal of the incorrect Neisseria genome submission from the public GenBank database, that database still has tens of thousands of other genomes that have not undergone the same scrutiny by independent reviewers. The paper reminds us of the importance of validating genome sequences and metadata obtained from and deposited into genome databases. Although the art and science of genome processing has made tremendous strides, it is still far from perfected.

Sloan Digital Sky Survey Releases the Largest and Most Detailed Digital Map of the Universe

IDIES plays key role in Data Release 12 (DR12) of the Sloan Digital Sky Survey (SDSS)

In January 2015, IDIES played a key role in Data Release 12 (DR12) of the Sloan Digital Sky Survey (SDSS), by far the largest and most detailed digital map of the Universe and one of most successful international projects in the history of astronomy. The new release continues the IDIES tradition of making the full SDSS catalog dataset freely available to the world.

Along with images of nearly half a billion unique astronomical objects from a billion unique detections, DR12 brings the total number of galaxies with observed spectra – and thus with reliable distance measurements – to more than 3.8 million. The 12 Terabyte DR12 catalog archive includes new observations of more than 160,000 stars in our own Milky Way Galaxy, with never-before-released estimates of each star's chemical composition in 15 different elements. DR12 also includes an entirely new type of data – high-quality measurements of the radial velocities of more than 10,000 stars, taken with the goal of finding potential extrasolar planets.

As has been the case from the beginning, the DR12 catalog data is served through a commercial database management system (Microsoft's SQL Server) and available freely online, through a variety of web-based IDIES tools, including the SkyServer (http://skyserver.sdss.org) and CasJobs (http://skyserver.sdss.org/casjobs/) websites. SkyServer includes several easy-to-use tools for browsing and searching SDSS data, as well as educational activities teachers can use to help students learn science using the same data that professional scientists are using. CasJobs allows researchers to submit complex search queries to DR12 data, and to save the results of those queries to personal database space for later analysis.

IDIES's involvement in the SciServer project has led to the realization of another important goal – all phases of the SDSS have now been unified into a single web site going forward (http://www.sdss.org). The new site will make it much easier for astronomers and the public to find the information they need, whether it be a learning activity from the earliest days of the SDSS in 2001 or the latest galaxy spectra from DR12.

The Sloan Digital Sky Survey is supported by the Astronomical Research Consortium, award # SSP430.

Credit: Dana Berry / SkyWorks Digital, Inc. and Jonathan Bird (Vanderbilt University)

Alex Szalay and Steven Salzberg named Bloomberg Distinguished Professors

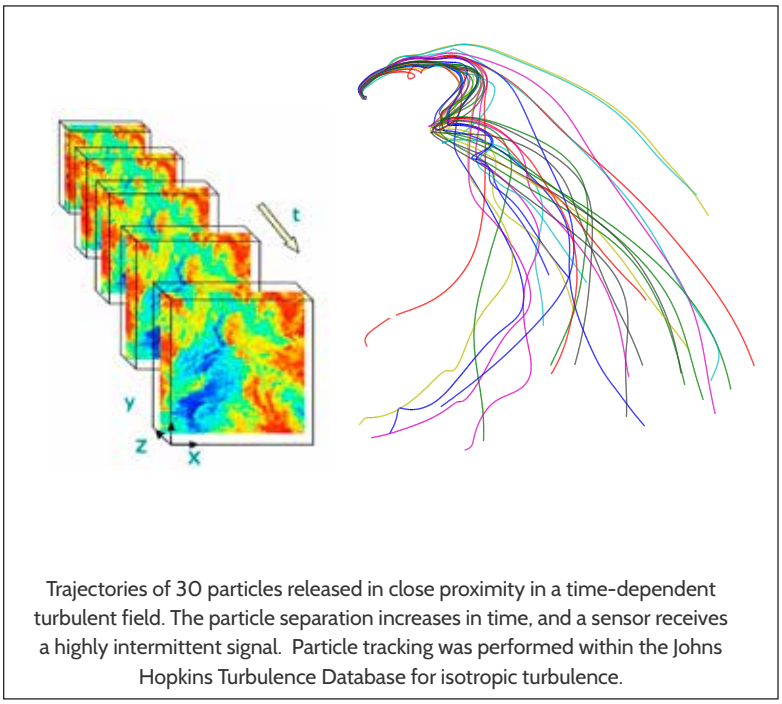
The Institute for Data Intensive Engineering and Science (IDIES) Director Alex Szalay and IDIES Associate Director Steven Salzberg are two of four Johns Hopkins Professors awarded the prestigious endowed Bloomberg Distinguished Professorship. The BDPs were established with a landmark gift from alumnus Michael R. Bloomberg. The goal of this initiative is to bridge the university's schools and divisions, conduct and stimulate innovative research that crosses traditional disciplinary boundaries, and train a new generation of native "inter-disciplinarians."

Alexander Szalay has broken new ground in the use of big data to advance scientists' understanding of astronomy, physical science, and life sciences. As the founding director of the Institute for Data Intensive Engineering and Science, he and his team are developing methods which are helping Johns Hopkins become a leader in the world of high-performance computing. Alex's collaboration on a scientific database for astronomy led to similar collaborations, and his technologies have provided the platform for big data analysis across disciplines, including radiation oncology and genetics. He is on the faculties of the Krieger School of Arts and Sciences and the Whiting School of Engineering.

Steven Salzberg's work at the convergence of computer science and genomics will help physicians customize medical care by giving them tools to connect and translate huge sets of data. As director of the university's Center for Computational Biology, he leads a group whose research focuses on developing new methods and software for analyzing DNA and RNA sequences across a vast range of species. Steven started out as a professor of computer science at Johns Hopkins, but his intellectual interests led him to work on some of the first genomes to be sequenced, including the human genome and many of the early bacterial genomes. He has appointments on the faculties of the School of Medicine, Bloomberg School of Public Health and the Whiting School of Engineering.

Professors Zaki and Meneveau receive NSF Award to Study Contaminant Dispersion by Turbulence

IDIES Seed Fund Grant Leads to NSF-Funded Project



Trajectories of 30 particles released in close proximity in a time-dependent turbulent field. The particle separation increases in time, and a sensor receives a highly intermittent signal. Particle tracking was performed within the Johns Hopkins Turbulence Database for isotropic turbulence.

A team led by Professors Tamer Zaki and Charles Meneveau of the Department of Mechanical Engineering has received funding from the National Science Foundation for A Big Data Computational Laboratory for the Optimization of Olfactory Search Algorithms in Turbulent Environments.

The researchers will use the unique Johns Hopkins Turbulence Databases to predict the source of contaminant release in turbulent environments. A detector placed away from an unknown source in a turbulent field receives a highly intermittent signal of contaminant concentration. The intermittency of the signal is encoded by the stochasticity of turbulence and is very challenging to decode. In this project, we devise strategies to decode that information. Starting from the detected contaminant and some knowledge of the flow conditions, we attempt to fully characterize the source. Our approach can determine the source location, its intensity and release history in the presence of measurement and flow uncertainties. We also examine the ideal placement of multiple sensors and the optimal search path for moving sensors. For moving sensors, we compare our algorithm to the instructive examples provided by nature where insects and animals have integrated olfactory sensors in their search strategy for nutrition and mating.

This effort utilizes the Johns Hopkins Turbulence Databases as a Numerical Turbulence Laboratory where virtual sensors can be placed and forward- and backward-particle tracking are performed. The diversity of flow configurations within the database ensures that our search strategies are robust. Finally, in collaboration with our project partners at the University of Tokyo, we will perform an experimental demonstration of our olfactory search algorithm using a state-of-art autonomous underwater vehicle.

The proposal was submitted under IDIES to a joint US-Japan (NSF-Japan Science and Technology Agency) program on use of Big Data for disaster mitigation. It was one of six awarded grants ($300K, 01/04/15–31/03/18) featured in NSF press release 15-029.

Professor René Vidal receives BIGDATA Award from NSF

Dr. René Vidal, Professor in the Center for Imaging Science, Department of Biomedical Engineering, and newest IDIES affiliate, recently received funding from the National Science Foundation's Critical Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) Program, in support of "developing or studying fundamental techniques, theories, methodologies, and technologies of broad applicability to Big Data problems."

Professor Vidal's $600K Award, Learning a Union of Subspaces from Big and Corrupted Data, is a collaboration with Professor Daniel Robinson of the Department of Applied Mathematics and Statistics, which supports the development of theory and algorithms for automatically discovering multiple low-dimensional structures in high-dimensional data. In the case of uncorrupted data, the research team studies conditions on the data under which a perfect clustering is possible. In the case of data corrupted by outliers, the research team studies conditions under which perfect clustering and outlier rejection are possible. In the case of data with missing entries, the research team studies conditions under which perfect clustering and data completion are possible. The project also develops efficient and scalable algorithms that benefit from distributed and high-performance computing for solving subspace clustering problems. These algorithms will be evaluated in large-scale problems in computer vision, including image clustering.

The techniques to be developed will enhance our ability to handle big data problems from multiple sources and modalities, and advance the knowledge on how to interpret massive amounts of complex high-dimensional data. Such techniques will significantly broaden the applicability of existing results in sparse representation theory to subspace clustering problems, which have found widespread applications in image processing (e.g., image denoising, compression, representation, and segmentation), computer vision (e.g., motion segmentation and face clustering) and dynamical systems (e.g., hybrid system identification).

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!

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.

Fall, 2014

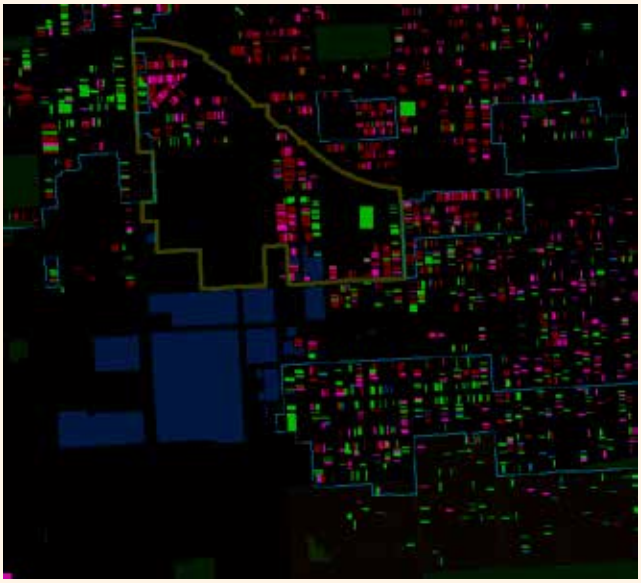
Urban Planning in Baltimore City

Tamás Budavári (Dept. of Applied Mathematics & Statistics), Kathryn Edin (Dept. of Sociology), and Michael Braverman (Dept. of Housing & Community Development, Housing Authority of Baltimore City)

IDIES Affiliate Professor Budavari's research into vacant housing dynamics dovetails with Johns Hopkins 21st Century Cities (21CC) Initiative's goal: to initiate innovative solutions for creating wealth, expanding opportunities, transforming education, promoting well-being and health, strengthening infrastructure, and cultivating the arts in our cities.

> "Our close collaboration with City Leadership guarantees that we focus on the most relevant high-level questions."
>
> -- Tamas Budavari

With the advent of sensor networks and computerized record keeping of public data, a new branch of data-driven science aimed at improving the quality of city life is emerging, and being integrated into redevelopment policy and administration in government. Macroeconomic and demographic trends have left Baltimore with 300,000 fewer residents than 60 years ago. This depopulation has resulted in more than 16,000 vacant, uninhabitable buildings. These buildings pose significant challenges to the City Leadership, from maintenance and crime to negative perceptions hampering reinvestment. While the preferred outcome for a vacant building is rehabilitation, in many cases demolition of an entire row is the only viable option. Addressing the vacancy crisis is essential to attracting and retaining people in Baltimore.



Visualizations such as this blacklight map help the Housing Department set policies. Hopkins Medical Campus: blue; vacant buildings: red and pink, formerly vacant building: green.
Illustration by John David Evans | Code Stat Director | Baltimore City Housing

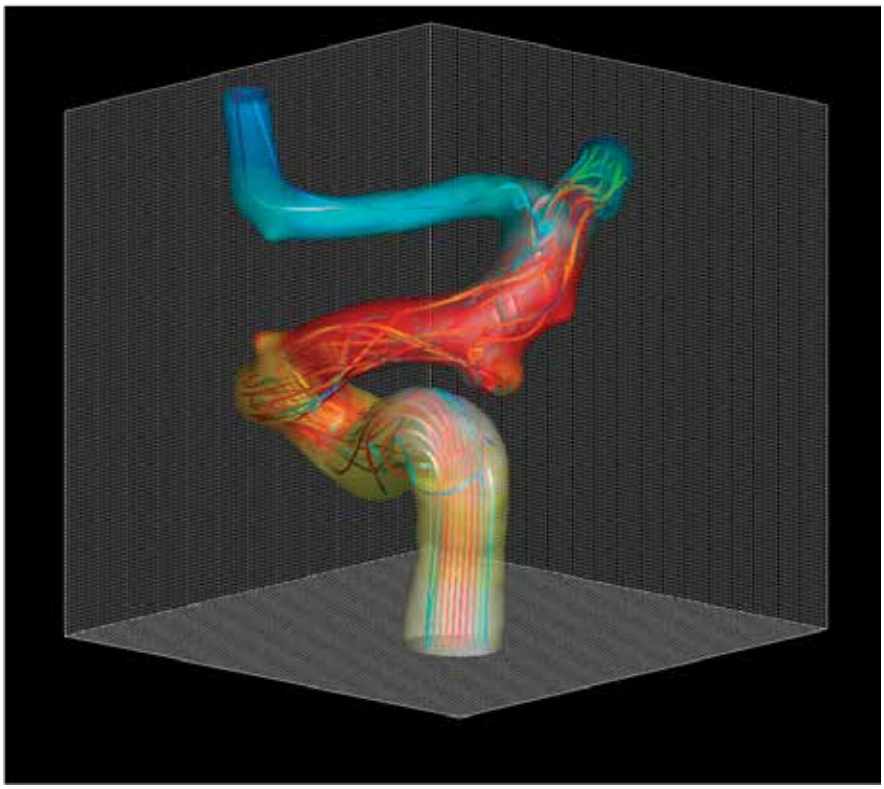
The interdisciplinary team has created a novel framework to address the challenges posed by the complex datasets about the city of Baltimore. The team's approach is to build a unique database of the geometries of all housing lots in the city, then add in layers of pertinent information to study the vacant housing dynamics. The auxiliary data has to include everything from construction permits and violation notices to water usage and calls to emergency services. They use the available information to track each lot's history from occupied to vacant, rehabilitated, or demolished. With this new system the researchers aim to measure the effects of city interventions and to predict possible outcomes in hypothetical scenarios.

The PIs of this Seed Fund Grant play key roles in the current development of the recently established 21CC Initiative: Tamas Budavari is member of Steering Committee of 21CC, Kathy Edin is the Director. Michael Braverman is Deputy Commissioner of Baltimore Housing.

A Modeling Enabled Database for Aneurysm Hemodynamics and Risk Stratification

Jung Hee Seo, (Dept. of Mechanical Engineering), Rajat Mittal (Dept. of Mechanical Engineering), and Rafael Tamargo, (Dept. of Neurosurgery & Otolaryngology), and Justin Caplan, (Dept. of Neurosurgery)

An aneurysm is a pathological, localized, balloon-like bulge in the wall of blood vessel. While ruptured intracranial aneurysms are associated with high mortality and morbidity, treatment of aneurysms considered to be at risk of rupture also brings serious risks. Thus, prompt and accurate stratification of rupture risk is the "holy-grail" in treating this pathology. Physics-based computational models of aneurysm biomechanics including the simulation of blood flow holds great promise in developing a reliable risk stratification method. However, current approaches are not designed to deal with large sample sizes that are essential for developing insights and reliable statistical correlations/metrics. The ultimate goal of the research proposed here is to lay the foundation for a modeling-enabled, crowd-sourced approach to the analysis of aneurysm hemodynamics and risk; the Aneurysm Hemodynamics and Risk Data Hub (AHRDH) to be developed is a unique WEB-based, crowd-sourcing Data Hub to generate a large (and continuously growing) database of simulation based, patient-specific aneurysm hemodynamics. The near-term goal of this seed funding program is to develop the automated patient-specific image processing and flow simulation methods to generate computational hemodynamic results, that will allow us to deal with large number of patient data.



Towards a Global, Streaming Data Exploration Testbed in Astrophysics

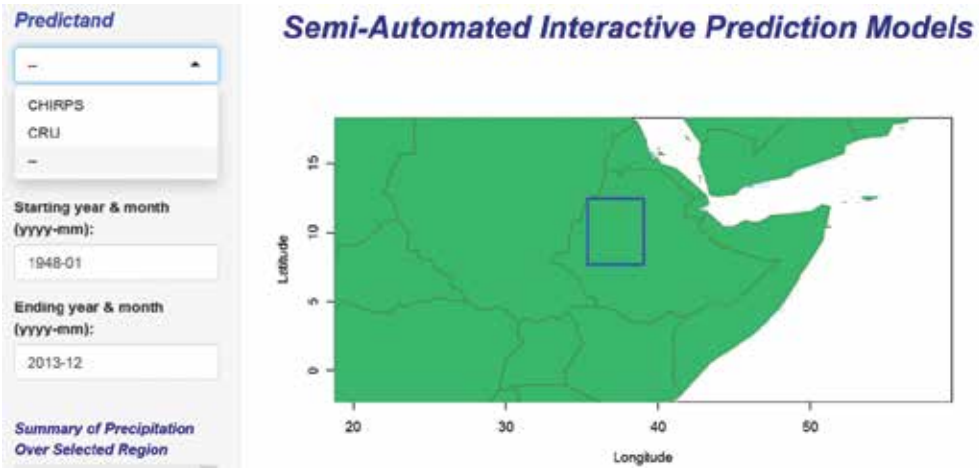
Brice Menard, (Dept. of Physics & Astronomy), Yanif Ahmad (Dept. of Computer Science), and Raman Arora, (Dept. of Computer Science)

The astronomical data space has dramatically increased over the past fifteen years, thanks to detector technology and space-based observations opening up new wavelengths channels. Surprisingly, attempts to characterize and represent the data globally have been rather limited. With this project, we propose to: (i) identify a standard set of operations to look globally at datasets; (ii) explore the potential of various techniques used in statistics and Machine Learning; (iii) define and build efficient tools to conducting global data exploration given one dataset or a combination of them. The goal of this project is to develop a preliminary package allowing a user to perform global data exploration and gain knowledge on the content of the data space.

Optimized Empirical-statistical Downscaling of Global Climate Model Ensembles for Climate Change Impacts Analysis

Benjamin Zaitchik, (Dept. of Earth & Planetary Sciences), Seth Guikema (Dept. of Geography & Engineering), and Dr. Sharon Gourdji, (International Center for Tropical Agriculture (CIAT) Cali, Colombia)

Under IDIES Seed Grant funding we are developing an automated climate forecast and downscaling system. The system uses an empirical-statistical approach in which remote variables (e.g., Pacific Ocean temperatures) are employed to predict local change (e.g., Baltimore rainfall). Our primary innovations are to implement the system in a manner that allows users to select predictors using automated or interactive methods and then to apply a combination of optimization and statistical learning theory to generate skillful predictive models. The system is capable of searching massive 4-D (x,y,z,t) climate databases to find optimal combinations of predictors. Models can be applied in seasonal forecast mode using real-time analyses or in climate change mode using Global Climate Model projections.



The image shown is a screen capture from the interactive online version of the semi-automated prediction tool, developed using Shiny by Rstudio. The user selects a region and time series variable of interest. This selection serves as the basis for diagnostic plots that help the user to identify promising predictors drawn from climate reanalyses and satellite-derived datasets. After the user selects predictors the modeling system applies an iterative model generation procedure in which a genetic algorithm is employed to optimize predictors in a suite of parametric and nonparametric statistical models. The product of the algorithm is a set of candidate models ranked by predictive skill that the user can apply to study climate process or generate operational predictions.

Spring, 2015

Genome-wide Prediction of DNase I Hypersensitivity and Transcription Factor Binding Sites Based on Gene Expression

Hongkai Ji, (Biostatistics), Ted Dawson (Neurology (SOM), Neuroscience (SOM)), and Valina Dawson, (Neurology (SOM), Neuroscience (SOM), Physiology (SOM))

The objective of this proposal is to develop a data science approach for studying global gene regulation. The proposal tackles a problem at the interface of data science, statistics, machine learning, and biology. We propose to use massive amounts of publicly available functional genomic data to build computational models to predict genome-wide DNase I hypersensitivity (DHS) and transcription factor binding sites (TFBSs) based on gene expression data. We will develop new big data regression methods for this purpose, and we will test the feasibility of using gene expression data generated from small number of cells to predict DHS and TFBSs.

The project has two specific aims. Our first aim is to develop a big data regression method for predicting genome-wide DHS and TFBSs using RNA-seq or exon array gene expression data. In this prediction prob-

lem, both predictors and responses are ultra-high-dimensional. We will develop a solution that is both statistically and computationally efficient. Data from the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects will be used to train and test the prediction models. We will evaluate the feasibility of applying the trained models to gene expression data in the Gene Expression Omnibus (GEO) database to predict DHS.

Our second aim is to test the feasibility of using small-cell-number gene expression data to predict DHS and TFBSs. We will generate RNA-seq or exon array gene expression data using different cell numbers (ranging from ~10^2 to ~10^6 cells) for 1~2 ENCODE cell lines. We will then predict DHS using these data and evaluate how prediction accuracy decreases as a function of cell number. Through this analysis, we will identify the minimal cell number required for making practically useful DHS predictions, which will provide important guides for future applications of the prediction approach to decoding gene regulatory programs in small-cell-number samples.

Statistical Methods for Real-Time Monitoring of Physical Disability in Multiple Sclerosis

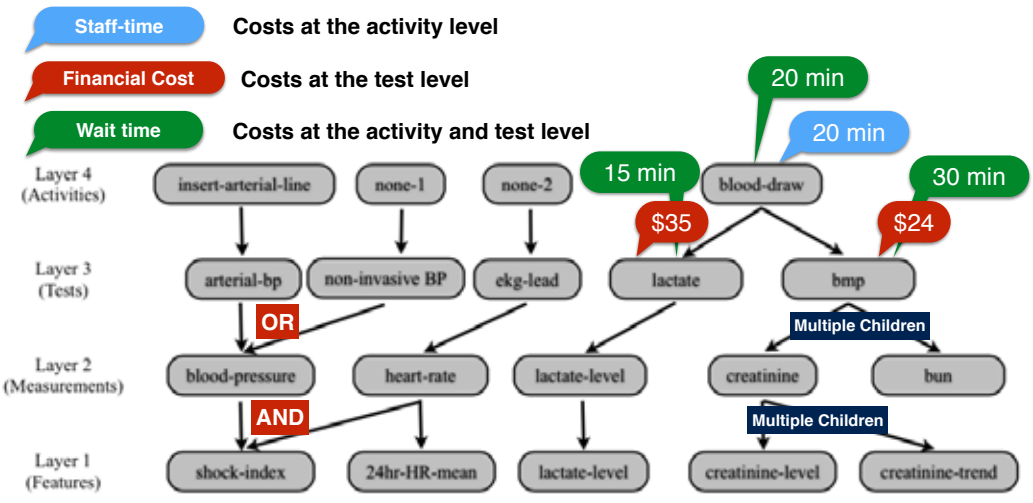
Vadim Zipunnikov (Biostatistics) and Kathleen Zackowski (Motion Analysis Lab)

The lack of sensitive outcomes capable of detecting progression of Multiple Sclerosis (MS) is a primary limitation to the development of newer therapies. Wearables provide real-time objective measurement of physical activity of MS patients in a real-world context. We put forward a novel statistical framework that simultaneously characterizes multiple features of physical activity profiles over the course of a day as well as their day-to-day dynamics. The proposed framework will allow MS researchers to identify physical activity signatures that will distinguish between individuals with different MS types and will help to understand physical activity differences in disability progression.

Cost-Sensitive Prediction: Applications in Healthcare

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

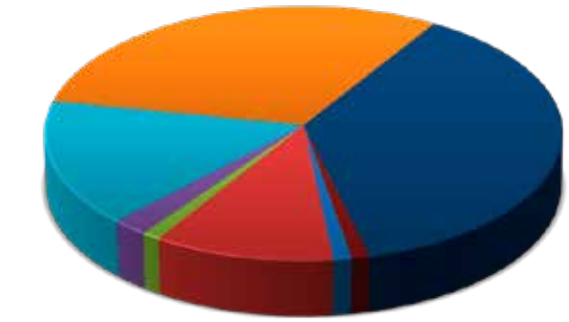
Advances in model prediction for problems that have a non-trivial cost structure are needed. In healthcare, the financial, nurse time, and wait time costs share a complicated dependency with the clinical measurements needed and medical tests performed. In 2014, the healthcare budget in the United States came to 17% of 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. We will develop such models and new optimization algorithms to solve them that give better predictions at lower costs, incorporate a patient's preferences, and assist in personalized healthcare.



IDIES IN NUMBERS

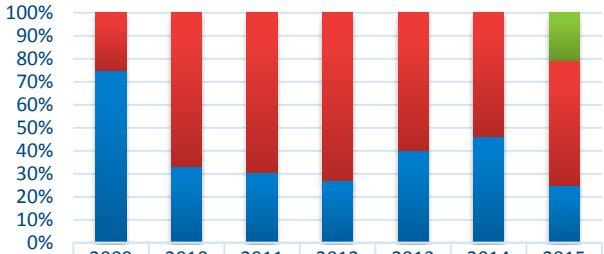
In fiscal year 2015 (July 2014 through June 2015), IDIES experienced continued growth while supporting the IDIES vision of facilitating high performance and data intensive computing across all of JHU. We saw a 12% growth in affiliate members, and 17% of 2015 proposal submissions came from first-time IDIES proposers. IDIES FY15 proposal submissions kept up with our positive trend, totaling 24 successful submissions for the year. FY15 SEED funding was awarded to five new and two existing IDIES affiliates. SEED awardees have begun to submit their SEED research ideas to external agencies for additional support and we are already seeing positive results.

Affliate Member Distribution



- Applied Physics Laboratory
- Bloomberg School of Pubic Health
- School of Education
- Sheridan Libraries
- School of Medicine
- Krieger School of Arts and Sciences
- Whiting School of Engineering
- School of Nursing

IDIES Proposal Success Rate

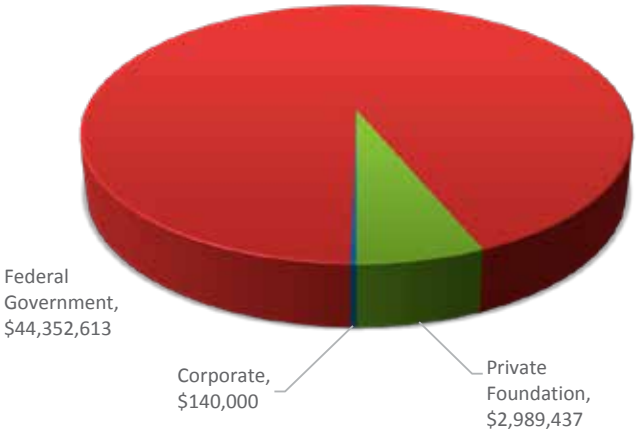


	2009	2010	2011	2012	2013	2014	2015
Pending							5
Not Funded	1	4	9	8	12	14	13
Funded	3	2	4	3	8	12	6

Government agencies continue to be the primary source of sponsored IDIES funding, with research grants from the National Science Foundation leading the way. IDIES awarded funding grew by 15% from federal sponsors and 45% from private foundations. While 2015 saw no direct funding from corporate partners to IDIES, we continue to seek additional contacts and to establish closer links. 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!

IDIES Awarded Funding 2009-2015



THANK YOU

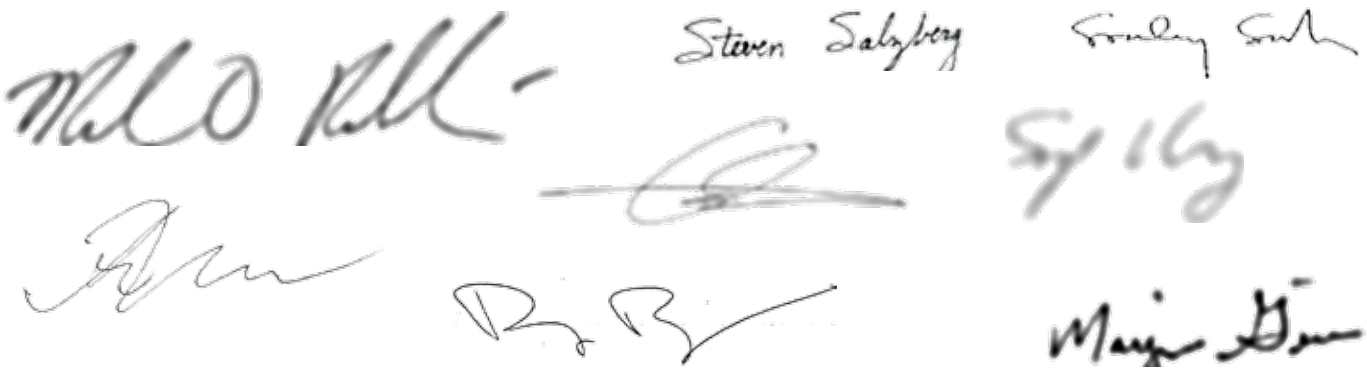
To our generous sponsors

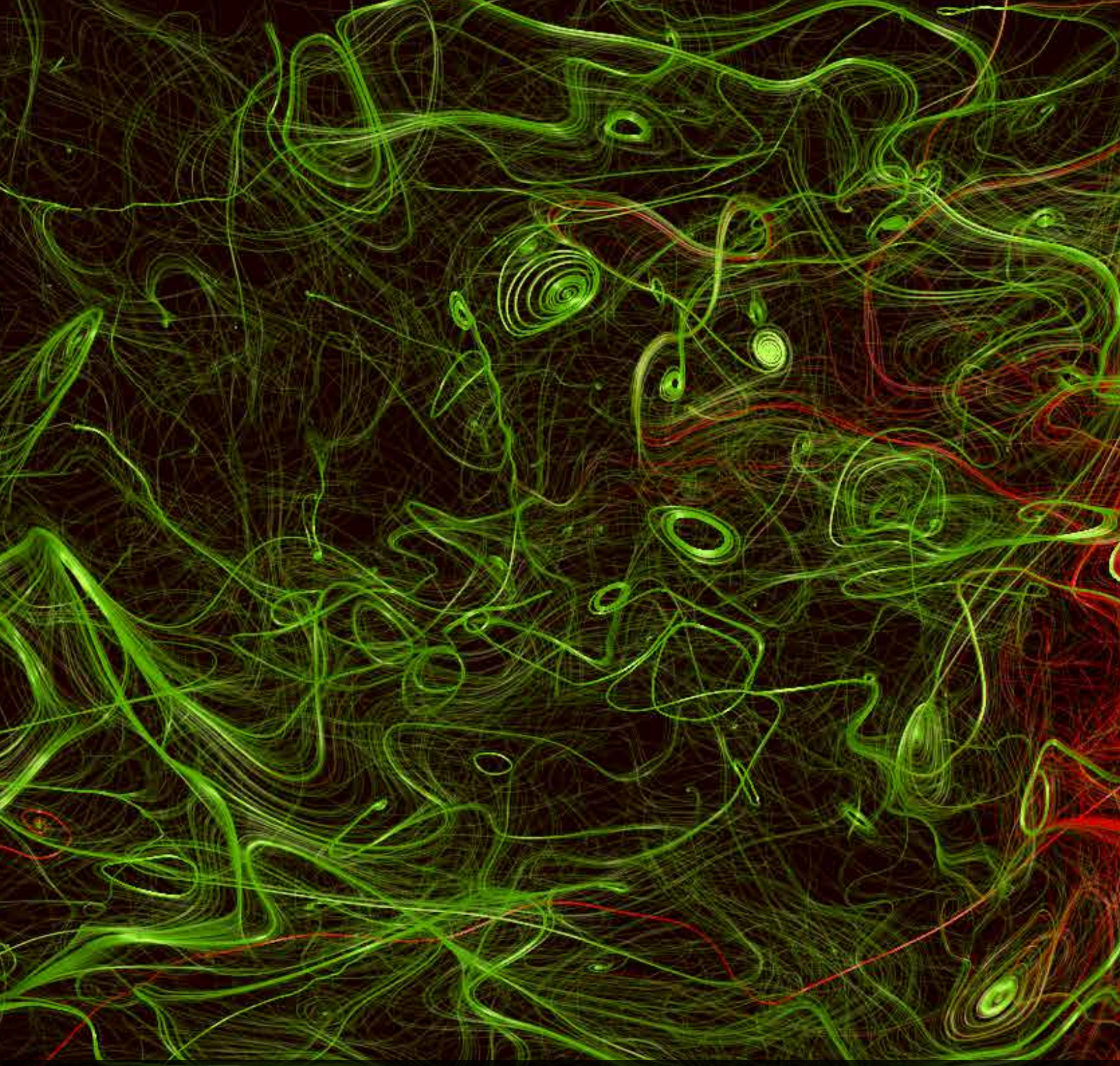
JOHN TEMPLETON FOUNDATION

GORDON AND BETTY MOORE 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