"... we must run as fast as we can, just to stay in place. And if you wish to go anywhere you must run twice as fast as that."

--Lewis Carroll, Alice in Wonderland

This year we have reached the fifth year of IDIES. This is a good time to look at what we have accomplished. IDIES is now partnering with six different Schools, the Krieger School of Arts and Sciences, the Whiting School of Engineering, the Sheridan Libraries, the School of Medicine, the Bloomberg School of Public Health and the Carey Business School.

Besides serving in a leadership role for Big Data initiatives for the University, IDIES has become responsible for the research computing efforts at JHU. The Maryland Advanced Research Computing Center (MARCC), a world class research computing facility on the Johns Hopkins Bayview campus, was built in partnership with the University of Maryland College Park. The project was supported by a $30M grant from the State of Maryland. MARCC has now been in steady operation for several years, with close to 100% of its capacity utilized, enabling cutting edge computations for JHU researchers.

Our campus infrastructure is among the best. IDIES has spearheaded an NSF proposal to build a 100G connectivity to the Internet2, which also connects to our internal 100G backbone for research data (HORNET), with 40G uplinks to individual buildings. JHU is unique among our peers in having more than 40PB of storage for research data.

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

We now 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.

The quote from Lewis Carroll provided a good summary of where we are and where we need to go. The agility, the ability to start from an idea, take it to a prototype and then to implement it is accelerated in the industry and the outside world. In academia, we still measure time in units of semesters. Machine learning, in particular Deep Learning, has emerged over the last few years revolutionizing how industry handles Big Data. JHU has created MINDS, a new institute to study the mathematical foundation of Big Data. This is clearly one area where universities can and should compete with industry leaders. We need to figure out a better way to efficiently seed the use of machine learning in individual science projects and turn these into success stories.

This is the focus of this year’s meeting, and it is no accident that all three of our keynote speakers have AI in the title of their talks. This translational effort is the focus of our engagement for the next years as well: we need to figure out how IDIES can help most efficiently, as many projects as possible, in deploying modern machine learning techniques in their science.

As we enter our next five years, we would like to accelerate, grow and become more relevant across the whole University. We would like to provide more intensive help in launching and sustaining data intensive projects in all disciplines. We seek new ideas and new directions, but we cannot do this alone: we need your help and initiative. Please send us your ideas, big or small, how we can improve our engagement with your research community.
# AGENDA

8:00 - 5:00 P.M. FRIDAY, OCTOBER 19, 2018
MUD Hall Auditorium, Homewood Campus, JHU

<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 a.m.</td>
<td>Continental breakfast &amp; Check-in – Mudd Hall UTL Commons</td>
</tr>
</tbody>
</table>
| 9:00 a.m. | OPENING REMARKS  
S. Alexander Szalay, PhD, Director of IDIES, Bloomberg Distinguished Professor, Department of Physics & Astronomy, Krieger School of Arts & Sciences, JHU |
| 9:10 a.m. | KEYNOTE SPEAKER  
Artificial Intelligence at the Edge meets Big Data and HPC in the Cloud  
Peter Beckman, PhD, Co-Director, Northwestern University/Argonne Institute of Science and Engineering, Computing, Environment & Life Science at Argonne National Laboratory and Northwestern University/Argonne Institute for Science & Engineering |
| 9:55 a.m. | SEED FUND UPDATE  
Using epidemiological and simulation data to inform the testing of autonomous vehicles  
Johnathon Ehsani, PhD, Asst. Professor, Dept. of Health Policy & Management, Bloomberg School of Public Health, JHU |
| 10:10 a.m. | SEED FUND UPDATE  
Harnessing Big Data for Population Health: Advancing Natural Language Processing Techniques to Extract Social-Behavioral Risk Factors from Free Text within Large Electronic Health Record Systems  
Elham Hatf, MD, PhD, Asst. Scientst, Dept. of Health Policy & Management, Bloomberg School of Public Health, JHU |
| 10:25 a.m. | Break – Mudd Auditorium Lobby |
| 10:45 a.m. | Transformative Research at MARCC  
Jaime Combariza, PhD, Assoc. Research Professor, Dept. of Chemistry, Krieger School of Arts & Sciences, JHU |
| 11:00 a.m. | SEED FUND UPDATE  
Is Geo-Location Information Helpful in Tracking the Spread Between Under Armour and Nike?  
Jim Kyung-Soo Liew, PhD, Asst. Professor, Dept. of Finance, Carey Business School, JHU |
| 11:15 a.m. | SEED FUND UPDATE  
A Data-Intensive Approach to Large-scale Social Network Structural Analysis: Identifying Hidden Communities  
Angelo Mele, PhD, Assoc. Professor, Dept. of Economics, Carey Business School, JHU |
| 11:30 a.m. | Poster Madness: 1 slide and 1 minute to pitch your poster |
| 11:45 a.m. | Lunch & Poster Session – Mudd Hall UTL Commons |
| 12:45 p.m. | KEYNOTE SPEAKER  
The Materials Genome Initiative and AI  
Jim Warren, PhD, Director of the Materials Genome Program, Material Measurement Laboratory, NIST |
| 1:30 p.m. | SEED FUND UPDATE  
Data driven Prediction of Risk of Sudden Cardiac Death  
Dan M. Popescu, Graduate Student, Dept. of Applied Mathematics & Statistics, Whiting School of Engineering, JHU |
| 1:45 p.m. | Toward a National Research Storage Substrate: The Open Storage Network  
Alannah White, Assistant Director, Technology, Institute for Data Intensive Engineering & Science, JHU |
| 2:00 p.m. | SciServer: Supporting Science Projects  
Gerrard Lemson, PhD, SciServer Assoc. Director of Science Outreach, Institute for Data Intensive Engineering & Science, JHU |
| 2:10 p.m. | Break – Mudd Hall Auditorium Lobby |
| 2:25 p.m. | SciServer in the Materials Science Domain: From Data to Discovery  
David Elbert, Assoc. Research Scientist, Dept. of Earth & Planetary Sciences, Krieger School of Arts & Sciences, JHU |
| 2:35 p.m. | SEED FUND UPDATE  
Developing an Analytics Software Tool to Identify Patients Diagnosed with Heart Failure and at High Risk of 30-day Hospital Readmissions  
Ali Afshar, PhD, Postdoctoral Fellow, Dept. of Anesthesiology & Critical Care Medicine, School of Medicine, JHU |
| 2:50 p.m. | SEED FUND UPDATE  
A big-data engine for large-scale soliciting screens  
Ben Langmead, PhD, Asst. Professor, Dept. of Computer Science, Whiting School of Engineering, JHU |
| 3:05 p.m. | KEYNOTE SPEAKER  
Data-driven discovery in the next generation sky survey  
Tony Tyson, PhD, DSc, Distinguished Professor, Physics Department, University of California at Davis |
| 3:50 p.m. | CLOSING REMARKS  
S. Alexander Szalay, PhD, Director of IDIES, Bloomberg Distinguished Professor, Department of Physics & Astronomy, Krieger School of Arts & Sciences, JHU |
| 4:00 p.m. | Cocktail Reception – Mudd Auditorium Lobby |
SPEAKERS

TONY TYSON, PhD
Distinguished Professor, Physics Department
University of California at Davis

Tony Tyson is a Distinguished Professor of Physics at University of California, Davis. Before that, he worked 35 years for Bell Labs in the physics division. While applying CCDs to astronomy in the early 1980s he discovered a population of faint blue galaxies, and then pioneered the field of weak gravitational lensing using these distant galaxies as sources. His current research is in cosmology: dark matter distribution, gravitational lens effects, cosmic shear, and the nature of dark energy. He is currently Chief Scientist on the Large Synoptic Survey Telescope project.

Tyson received his B.S. in Physics from Stanford in 1962 and PhD in condensed matter physics from University of Wisconsin in 1967. He is a Fellow of the American Physical Society and the American Academy of Arts and Sciences, and a member of the National Academy of Sciences and the American Philosophical Society.

JIM WARREN, PhD
Director of the Materials Genome Program, National Institute of Standards and Technology

Dr. James A. Warren is the Director of the Materials Genome Program in the Material Measurement Laboratory of the National Institute of Standards and Technology (NIST). After receiving his PhD in Theoretical Physics at the University of California, Santa Barbara, which was preceded by an A.B. (also in Physics) from Dartmouth College, in 1992 he took a position as a National Research Council post-doc in the Metallurgy Division at NIST. In 1995, with three other junior NIST staff members, he co-founded the NIST Center for Theoretical and Computational Materials Science, which he has directed since 2001. From 2005-2013 he was the Leader of the Thermodynamics and Kinetics Group. His research has been broadly concerned with developing both models of materials phenomena, and the tools to enable the solution of these models. Specific foci over the years have included solidification, pattern formation, grain structures, creep, diffusion, wetting, and spreading in metals. In 2010-11, Dr. Warren was part of the ad hoc committee within the Office of Science and Technology Policy’s National Science and Technology Council (NSTC) that crafted the founding white paper on the Obama Administration’s Materials Genome Initiative (MGI). Since 2012, Dr. Warren has served as the Executive Secretary of the NSTC MGI Subcommittee, coordinating inter-agency efforts to achieve the goals laid out in the MGI.
Keynote

PETE BECKMAN, PhD
Co-Director,
Northwestern University &
Argonne Institute of Science and Engineering

Pete Beckman is a recognized global expert in high-end computing systems. During the past 25 years, his research has been focused on software and architectures for large-scale parallel and distributed computing systems. Pete enjoys the challenging step of moving novel research software into production on extreme-scale systems. After receiving his PhD degree in computer science from Indiana University, he helped found the university’s Extreme Computing Laboratory, which focused on parallel languages, portable run-time systems, and collaboration technology. In 1997 Pete joined the Advanced Computing Laboratory at Los Alamos National Laboratory, where he founded the ACL’s Linux cluster team and launched the Extreme Linux series of workshops and activities that helped catalyze the high-performance Linux computing cluster community.

Pete also has been a leader within industry. In 2000 he founded a Turbolinux-sponsored research laboratory in Santa Fe that developed the world’s first dynamic provisioning system for cloud computing and HPC clusters. The following year, Pete became Vice President of Turbolinux’s worldwide engineering efforts, managing development offices in the US, Japan, China, Korea, and Slovenia.

Ali Afshar received his PhD in Electrical & Computer Engineering, Computational Medicine at JHU. He has led translational research projects and built interdisciplinary teams of clinicians, engineers and designers at Johns Hopkins Medicine in collaboration with industry partners including Apple, Nokia Health and Google.

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

Jonathan Ehsani, PhD, researches the role of policy and technology on driver behavior and safety. He has ongoing studies examining the effectiveness of autonomous vehicle testing policies and is collaborating on a number of instrumented vehicle studies. He is the current Leon S. Robertson Faculty Development Chair in Injury Prevention.

David Elbert is an environmental geochemist and mineralogist with broad expertise. He’s recently worked on methodology development for electron microscopy, synchrotron X-ray, and neutron scattering applications to toxic-metal, crystal chemistry.
Elham Hatef, MD, MPH is a preventive medicine physician and clinical informatician. She is on the faculty of the Center for Population Health IT at JHSPH and the academic director of the General Preventive Medicine Residency Program. Dr. Hatef focuses on population health and health information technology. Her main field of work is the impact of social determinants of health on health-related outcomes using health IT and Big Data.

Ben Langmead, PhD, and his group apply ideas from sequence alignment, text indexing, statistics and parallel programming. Ben has released high-impact software tools (e.g. Bowtie, Bowtie 2) addressing common genomics research questions. He has also created scalable software tools that use commercial cloud computing to analyze large collections of archived sequencing data.

Gerard Lemson received his PhD in theoretical cosmology and is currently a research scientist at Johns Hopkins University. He is Associate Director for Science Coordination in the NSF-funded SciServer project (www.sciserver.org) and assists in code development of that platform.

Jim Kyung-Soo Liew, PhD, is an Assistant Professor of Finance at Johns Hopkins Carey Business School and revels in pushing the boundaries of financial knowledge and product development both as an academic and FinTech Entrepreneur. He has published pioneering research in the intersection of social media big data, crypto-currencies, and financial markets.

Angelo Mele, PhD, is an applied economist interested in social and economic networks and their effect on economic performance. His research combines theoretical modeling, structural and reduced-form econometrics with state-of-the-art computational methods.

Dan Popescu is a PhD student at Johns Hopkins University in the Department of Applied Mathematics and Statistics. He completed his BS in 2012 at University of Richmond, with a double major in Mathematics and Economics. Dan's research interests can be described broadly as using deep learning techniques on extracting information from clinical cardiac images.

Alex Szalay, PhD, is the founding director of IDIES, a Bloomberg Distinguished Professor, Alumni Centennial Professor of Astronomy, and a professor of Computer Science. As a cosmologist, he works on the use of big data in advancing scientists’ understanding of astronomy, physical sciences, and life sciences.

Alainna White is the Director of Technology for IDIES. She has over a decade of experience in scientific computing and has architected computational and storage systems of various scale for educational, government and commercial institutions. Alainna manages the technical resources and vision of the Institute and is currently spearheading the Open Storage Network effort with our NSF and NDSC partners.
Transformative Research at MARCC
Jaime Combariza (Director of MARCC)

MARCC has become a high demand core facility at Johns Hopkins University and the University of Maryland at College Park, with 220+ operating research groups and 750+ dynamic users computing at MARCC. The powerful computing resources at MARCC (23,000+ cores, 150 GPUs; about 1.4PFLOPs) are used to advance and transform research across many disciplines from genomics, deep learning, computational mechanics, molecular dynamics, quantum chemistry, astronomy and cosmology, study of black holes effects, to computational fluid dynamics and many more fields. MARCC has not only enabled data-driven computations but also has fostered significant local and external collaborations to teach different techniques to new and promising users. During FY18, MARCC was used as a teaching tool for 15 different classes and workshops, which trained over 200 new users. Additionally, simulations conducted at MARCC produced results that contributed to 400+ publications and MARCC resources supported over $120M in grant projects. As an example, the genomics research group of Professor Liliana Florea (SOM), working in developing computational methods for ‘big data’, has the ability to use MARCC resources along with the hundreds of genomics applications and libraries already available is critical, allowing the group to develop and evaluate powerful and highly scalable tools for analyzing next generation sequencing data, design new experiments, and maximize productivity.

The Bluecrab cluster continues to grow; research groups have added 8 condos totaling 60 nodes, 1500 cores and 4 P100 GPUs, supplemented with the final 20 nodes purchased using the remaining of the state grant, totaling over 3300 new cores. The overall compute power enables high quality simulations that contribute to JHU researchers’ position at the forefront of data-driven research and closes the gap between our local computing infrastructure and capabilities at peer institutions. As a result, researchers can propose ambitious, computationally intensive projects that lead them in different directions, stimulating their ability to procure funding for novel areas of exploration. Many projects have benefited from the readily available resources at MARCC and researchers point out that these important results could not have been achieved without access to local resources.

MARCC added power backup (generator power) to the facility to protect critical servers and valuable data from power outages, improve reliability and availability, and perform maintenance tasks with limited downtimes. The power backup is an incentive for research groups who want to use our colocation facility.

Another item of importance is the capability of storing, transferring and sharing data simply and easily. MARCC and IDIES provide access to Globus+ (a data transfer service from globus.org) that manages all data transfers. Researchers using Globus can easily transfer files within the Hopkins domain. For example, using a test sample (500 GB worth of large files) we were able to move data at about 8 gigabits per second (gbps) between Homewood and MARCC. That is, 8-10 minutes to move half a terabyte of data. Likewise, using Globus, researchers are able to share files with other collaborators at Hopkins or any other institution. Globus also allows for the creation of APIs that can transparently manage data transfers from, for example, instruments at East Baltimore to MARCC.

The HIPAA system (MARCC Secure Environment, MSE) continues to run as a prototype. We have about 10 groups who are actively using this resource. MARCC has partnered with the data trust, IRB group and the legal office to provide a secure environment for research groups who need to analyze data that contains Protected Health Information (PHI). For this particular service that involves high security, MARCC needs to be added as a collaborator to the project before it is approved by the IRB.

(Continued on page 9)

2018 IDIES ANNUAL REVIEW | NEWS
Toward Exascale Community Ocean Circulation Modeling

Thomas Haine, Department of Earth and Planetary Sciences, Johns Hopkins University

A new $2.8M project will push the envelope on ocean simulations and build a data server for public access. Principal Investigator Thomas Haine, Professor in the Department of Earth & Planetary Sciences, leads a team including IDIES researchers in the Departments of Physics & Astronomy and Computer Science, plus collaborators at Columbia University and MIT.

Professor Haine’s project is built on three premises regarding cyberinfrastructure in Earth & Planetary Sciences:

First, Computational Oceanography has come of age. Computational Oceanography is the study of ocean processes by numerical simulation, and is a new branch of marine science. Numerical simulation of the circulation has grown increasingly realistic with exponentially greater compute power in the last few decades.

In comparison, the observational database has grown more slowly. Thus, oceanographers are approaching the point when ocean circulation models can exactly match the data in an infinite number of ways. Such ocean circulation simulations must be treated as seriously as real observations.

Second, interactive analyses of these petascale – and in the future, exascale – ocean simulations must be widely available. Seamless access to ocean simulations has not kept pace with the increase in model resolution. Although the simulation output is in principle available to anyone, severe barriers exist to using the data in practice. To exploit simulation output it’s imperative that the model data are accessed and disseminated as widely as possible, including to casual, non-expert, and non-professional users. The simulation output must be “democratized” by making it (at least) as easy to use as oceanographic observations.

Third, Exascale Computational Science is on the horizon and ocean/atmosphere/climate dynamics constitute a vanguard use-case for computational innovation. Ocean model simulations will migrate to exascale compute resources in the foreseeable future. Moreover, this use-case is built on a generic vision to develop cyberinfrastructure...
Over the past year, SciServer has made significant progress towards its goal: build an accessible and easy-to-use collaborative platform for scientists at JHU and around the world to store, share and analyze their scientific data sets.

The overall capabilities of SciServer have been extended and improved, and the system moved from its previous "alpha" state to full production status in July 2018. SciServer provides collaborative data sharing capabilities through its new dashboard (see figure), database query services, and both interactive and asynchronous compute capabilities. We still provide, and have extended the support of, users' own storage space, both in databases (through MyDBs) and in file systems (through "persistent" storage). Our user base is growing as SciServer is becoming more widely known, and our hardware resources are growing to meet this need.

Additionally, SciServer has supported summer schools on Astroinformatics and Turbulence, and is used in university classes in the U.S. and abroad (e.g. the University of St Andrews in Scotland). We are creating tools to facilitate creating a virtual classroom inside SciServer, where teachers can share resources like assignments, data, and demo notebooks, as well as lecture notes and other materials. SciServer can create isolated storage spaces where students can work on assignments, and share them with their instructors and TAs. There is rapidly growing demand for such a straightforward, low-cost, and flexible approach to data driven classroom activities. SciServer makes creating and managing such activities easy.

A major focus going forward is to extend our engagement with the scientific communities where SciServer can provide real value. Over the past year, SciServer has continued to work with scientists in an ever-growing set of

(Continued on page 9)
The Open Storage Network: A National Research Storage Substrate:
Alainna White, IDIES

The amount of available research data continues to grow; however, researchers’ ability to access these data-sets has historically been, and continues to be, an ongoing challenge. The Open Storage Network (OSN) will provide a national storage substrate for the research community, enabling simple, reliable and fast sharing of data.

The deployment of low-cost, purpose-built petabyte-scale storage blocks at institutions across the country is the foundation of the OSN. More than two hundred research institutions in the United States possess research-dedicated high-speed Internet connections. The OSN endeavors to deploy a petabyte (minimum) of usable storage at each of these research endpoints, creating the largest research data storage system – in excess of 200PB of usable storage – in the world.

Distributed storage deployment provides the system with geography-based redundancy, improved data locality, and a greater number of high-speed connections than more centralized deployments. Furthermore, the cumulative encumbrance of the physical deployment (footprint, power and cooling) is dispersed across a number of institutions. Although the system is physically distributed, it is centrally managed from the OCC (OSN Command Center) in an effort to reduce the system’s administrative burden on individual institutions. This holistic management philosophy minimizes the need for local systems support and allows researchers to spend more time on science and less on logistics.

The OSN is partnered with Globus for authentication and data transfer functionality. The Globus Connect software provides scientists a familiar interface for sharing and accessing data on the OSN. Additional functionality and accessibility for community-based tools is provided using the industry-standard S3 API, allowing established communities and tools to maintain their existing workflow.

Originating at IDIES, the project has rapidly expanded beyond the institute. Over the next twelve months, IDIES and its partners at the NSF Big Data Hubs and the National Data Service Consortium will deploy four additional prototype storage blocks across the country. These prototypes will support several real-world scientific use cases, demonstrating the value of a national storage substrate for the research community. The prototype phase will lay the groundwork for large-scale deployment in 2020 and beyond.

Come join the conversation in Dallas, TX at the 2018 Supercomputing Conference!
Learn more at https://www.openstoragenetwork.org/sc18-bof.
Supported by NSF Award 1747493.

(Continued from page 6, Toward Exascale...)
diverse disciplines. Within the past year, we have started projects with teams in Sociology, Education, and Business. We have also started new collaborations with Johns Hopkins Medicine and the Bloomberg School of Public Health. The breadth of services that SciServer can offer to science projects is evident from our current capabilities:

- **Big Data dissemination:** We support publishing large scientific data sets to worldwide research communities. This often requires special-purpose hardware resources, development of loading pipelines, data modeling, documentation, etc.

- **Relational database usage:** We have experience with applying relational database technology in science, and we have shown the power of this technology as an alternative to traditional use of spreadsheets. We assist teams in data modeling, transforming data sets for ingestion into databases, database tuning, developing query patterns, etc.

- **Collaboration-wide resource sharing:** Relational database technology is not suitable for all data sets. We support special data volumes for storing collections up to 10-20 TB. Access to these collections can be restricted to within a collaboration, or can be opened to everyone, with SciServer’s access control system.

- **Customized compute support:** We can create customized compute environments for analyzing project-specific data sets, which greatly facilitates the user experience of working with data.

- **Web service development:** SciServer who know the contents of the data provide their end users with a services allow us to wrap the user interfaces, enabling counters and single button clicks.

- **Consulting:** In most, if not personnel help researchers analyzing problem areas providing startup implementations or, if desired, projects mandate compliance with the Health Insurance Portability and Accountability Act (HIPAA). We are working with the Applied Physics Laboratory (APL) to set up a separate instance of SciServer only accessible from a HIPAA-compliant desktop environment.

- **HIPAA:** Several medical projects with HIPAA or similar requirements are welcome to discuss with us whether this type of system might work for them.

If you are a Project, Group, or Department lead and would like to learn more about how SciServer can help your research, please visit our website at http://www.sciserver.org, email the helpdesk at sciserver-helpdesk@jhu.edu, or create an account.

See for yourself at https://apps.sciserver.org!

(Continued from page 5, *Transformative*)

In conclusion, after 3+ years in production mode MARCC has become an integral component of scientific computing. For the most part, MARCC’s resources have been free of charge to PIs. However, we need the help and support of the scientific computing community so we can establish a sustainable environment that continues enabling transformative research.
ANNOUNCEMENTS

Eleven IDIES Members Receive 2018 JHU Discovery Awards
IDIES member recipients include Tamas Budavari, Jonathan Ehsani, Greg Hager, Lingxin Hao, Tak Igusa, Steven Jones, Mauro Maggioni, Michael Schatz, Natalia Trayanova, Alex Szalay, Peter Zandi.

Congratulations to Jaafar El-Awady on winning a Johns Hopkins Catalyst Award
Professor El-Awady’s research interests include multiscale materials modeling, damage and fracture mechanisms of materials in mechanical design, material degradation in extreme environments, nano-materials and structures, and impact dynamics and wave propagation.

Quest Diagnostics & Hopkins Health System Labs Adopt VLDL Algorithm
The Very Large Database of Lipids (VLDL) project of IDIES Members Seth Martin and Steven Jones has reached a new milestone with the national adoption by Quest Diagnostics of VLDL’s new algorithm to determine LDL cholesterol. The Johns Hopkins Health System Laboratories has also adopted the new cholesterol algorithm.

Three IDIES Affiliates win Schmidt Sciences Nascent Innovation Grants
Professors Brice Ménard, Sean Sun, and Joshua Vogelstein each won a $250K grant from Schmidt Sciences. The Schmidt Grants support innovative data science and technology projects. Professor Ménard is an astrophysicist, Professor Sun is a mechanical engineer; Professor Vogelstein is a biomedical engineer.

Congratulations to Steven Salzberg on being elected to the American Academy of Arts and Sciences.

Kevin Schlaufman Proposes New Upper Limit on the Size of Planets

Congratulations to Charles Meneveau on being elected to the National Academy of Engineering.

IDIES Member Hongkai Ji and Team Publish Paper in Nature
Hongkai Ji published “Genome-wide prediction of DNase I hypersensitivity using gene expression” in Nature. Professor Ji’s research into a method to predict samples’ gene regulatory landscape was funded in part by the IDIES Seed Fund program.

MARCC praised in The Hub for Boosting research at Johns Hopkins
Researchers around the Johns Hopkins Community describe the advanced research computing center as “absolutely essential” and “a godsend.”

Congratulations to Mark Robbins on being elected Fellow of the American Association for the Advancement of Science.

Congratulations to Steven Salzberg and Alexander Szalay Credited as Most Cited Researchers by Clarivate Analytics in their 2017 Highly Cited Researchers List.

3D Visualization Hackathon, 2018
On June 9, 2018, 30 faculty, staff, and students from ten departments across Johns Hopkins converged on IDIES for the inaugural IDIES Visualization Hackathon. Big data visualization is one of the biggest challenges facing researchers in data-intensive and computationally-intensive fields. Participants from fields as diverse as particle physics, social sciences, turbulence, and earth sciences brought their expertise and techniques to the table to learn and share how to create virtual reality scenes, stereoscopic images and movies, and more.

Building on the success of our first Hackathon, IDIES plans to host another Big Data Hackathon in January, 2019. Stay tuned to the IDIES website, http://idies.jhu.edu, for more information!
Five years ago, the scope of IDIES expanded in recognition of the strategic importance of big data and data science. From the launch of the “new” IDIES in 2013, we now include six schools: Krieger School of Arts & Sciences, Whiting School of Engineering, School of Medicine, Bloomberg School of Public Health, Sheridan Libraries, and the Carey Business School. IDIES has grown to 140 faculty affiliate members from across the University, and, in 2017, opened memberships to include non-tenure track research faculty & scientists, postdocs, students, and non-JHU faculty and external researchers. We look forward to increasing our member base and strengthening our opportunities for collaboration and interdisciplinary research through continued support of our Seed Funding Initiative, seminar series, hackathons, and new outreach events we hope to announce in the coming year.

The IDIES Seed Fund Initiative received 72 submissions since its inception in 2014, and was able to provide awards for 27 projects. As a result of the awarded Seed projects, there have been 23 submissions to external sponsors, to date, as a direct result of these awards and nine of those submissions were awarded. IDIES proposal submissions continue to see benefits of IDIES collaborations and resources. Proposals submitted through IDIES show a success rate of 38% versus 26% for JHU big data proposals not associated with IDIES. IDIES aims to expand its faculty memberships so that all JHU proposals utilizing big data and data science can benefit from IDIES resources.

IDIES award funding continues to be primarily from government agencies. As we move forward, IDIES hopes to balance out this support with additional funding from private foundations and donors, and industry partners. We will seek to leverage current IDIES expertise and resources to spur new project ideas, team collaborations, and training opportunities. Make sure to keep an eye on your inbox and the IDIES website for all our new and upcoming initiatives!
Seed Fund Updates, Spring 2018

Each spring, the IDIES Seed Funding Program invites proposal submissions for $25,000 awards for Big Data pilot or seed projects. 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. Traditionally, IDIES awards four to five projects annually; however in 2018 the JHU School of Medicine generously funded two additional awards. We would like to thank the JHU School of Medicine, and all of our sponsors, for their support of this worthwhile program.

Using Epidemiological and Simulation Data to Inform the Testing of Autonomous Vehicles

Johnathon Ehsani (Center for Injury Research and Policy, Department of Health Policy and Management, Department of Health, Behavior and Society, Bloomberg School of Public Health), Tak Igusa (Center for Systems Science and Engineering, Department of Civil Engineering, Whiting School of Engineering), Hadi Kharrazi (Center for Population Health Information Technology, Department of Health Policy and Management, Johns School of Public Health)

Abstract: Autonomous vehicles (AVs) have the potential to transform mobility and reduce the burden of motor vehicle crashes. Before this promising future can become reality, however, there is a need for extensive testing of AVs. As industries aggressively roll out testing plans, they have found that the most challenging questions are on the location and timing of AV testing. While AV engineers are mastering factors such as motion control, path planning, localization, perception and mapping, they have not yet considered in suitable depth the epidemiology of crash risk, particularly within urban settings. This collaboration between public health and systems engineering will develop Geo-spatial distribution of high-severity motor vehicle crashes in Maryland between January-March 2018.

(Continued on page 20)
Harnessing Big Data for Population Health: Advancing Natural Language Processing Techniques to Extract Social-Behavioral Risk Factors from Free Text within Large Electronic Health Record Systems

Jonathan Weiner, Hadi Kharrazi, Elham Hatef (Center for Population Health Information Technology, Health Policy and Management, Bloomberg School of Public Health), Mark Dredze (Center for Language and Speech Processing & Malone Center for Engineering in Healthcare, Whiting School of Engineering), Christopher Chute (School of Medicine & Chief Research Information Officer, Johns Hopkins Health System)

Social determinants of health (SDOH) – a combination of behavioral, social, economic, environmental, and occupational factors – are powerful drivers of morbidity, mortality, and future well-being of individuals and communities, yet they mostly lie outside domain of the conventional medical care delivery system.

Despite the importance and significant impact of SDOH, medical care providers often rely on administrative claims for assessment of these factors, which mostly lack in-depth information on important SDOH affecting health. Another challenge for healthcare systems is access to data on SDOH through their electronic health records’ (EHRs) structured and unstructured data (clinical free-text notes). There is no standardized format for documenting SDOH in EHRs’ structured data and extracting SDOH data from unstructured EHRs requires time-consuming and subjective methods such as chart review for identification of patients with high social risk, which is not a feasible approach to screen a large population of patients.

We explored the use of natural language processing (NLP) and text mining techniques such as pattern-based methods to determine SDOH from Johns Hopkins Health System EHR’s unstructured data (free-text clinical notes) from 23 million clinic notes (of 1,188,202 unique patients and 9,066,508 unique encounters) from July 1, 2016 through May 31, 2018. We pre-processed the notes to mark the line boundaries and normalized character encodings for faster processing. To identify notes containing SDOH, we used hand-crafted linguistic patterns that a team of experts developed using ICD-10, CPT, and LOINC codes, SNOMED terminology, as well as SDOH description in public health surveys and Instruments.

To craft the linguistic patterns, the expert team focused on the three selected domains namely housing issues, social connection/isolation, and income/financial resource strain. Listing all the available codes for each selected SDOH domain and cross matching them across different coding systems and terminology the expert team developed a comprehensive list of all available codes and specific content areas, which guided them to develop a number of phrases for each SDOH domain.

Figure 1. Characteristics of EHR Unstructured Data (Free-Text Clinical Notes) Containing SDOH Domains. a) percentage of notes related to selected SDOH domains completed by each provider type. b) percentage of note types including information related to selected SDOH domains.

(Continued on page 18)
Data-driven Prediction of Risk of Sudden Cardiac Death

Natalia A. Trayanova (Department of Biomedical Engineering and Medicine, School of Medicine), Katherine C. Wu (Department of Medicine, Division of Cardiology, School of Medicine), Dan M. Popescu (Department of Applied Mathematics and Statistics, Whiting School of Engineering)

The overall goal of our research project is to develop targeted strategies for predicting risk of sudden cardiac death (SCD) from arrhythmias and use them in clinical practice. SCD is the largest cause of natural death in the U.S., a large proportion of deaths resulting from ventricular arrhythmias, in particular among patients with prior heart disease. Current clinical cardiology practice relies on “one-size-fits-all” measures to ascertain SCD risk, often leading to unnecessary expensive treatments and failures to identify a majority of SCD victims. Our proposed methodology marries two patient-personalized approaches: the Virtual Heart approach – based on cardiac imaging and computational modeling – and machine learning. While the Virtual Heart approach relies on constructing a 3D computer model of patients’ individual hearts from a variety of clinical imaging modalities and disease-specific electrophysiological properties, the machine learning approach supplements this by evaluating the contribution of other factors that may lead to electrical instabilities. The results of the two approaches will be integrated to yield a risk assessment score (Fig. 1).

In the first part of the project, we focused on the machine learning data-driven approach. Initially, we analyzed 109 patients with ischemic cardiomyopathy (infarction) enrolled in the PROSE ICD clinical trial with MRI scans acquired before April 2009. In light of the success deep learning – a new way of learning from data representations by separating the task into successively meaningful layers often modeled through neural networks – has had in computer vision in recent years, we developed a deep convolutional neural network algorithm to classify patients according to risk of SCD. To enhance the network’s learning abilities, we created a pre-processing algorithm shown schematically in Fig. 2. For each patient, a stack of 16 late-gadolinium enhanced (LGE)-MRIs with pre-segmented left ventricle myocardial wall and fibrotic tissue are cropped and one-hot encoded into 4 areas: outside region of interest (black), non-fibrotic tissue (green), infarct border zone (yellow) and core scar tissue (red). After re-scaling all patient images to ensure constant voxel dimension across the entire dataset, we apply a problem-specific data-augmentation algorithm and adjust weights.

Figure 1: Architecture of the SCD risk prediction system combining the Virtual Heart and Deep Learning approaches and how it is trained.

Figure 2: Pre-processing algorithm for the LGE-MRI dataset for a given patient.
Developing an Analytics Software Tool to Identify Patients Diagnosed with Heart Failure and at High Risk of 30-day Hospital Readmissions based on Physiologic and Clinical Data Obtained from Electronic Health Records

Nauder Faraday (Anesthesiology and Critical Care Medicine, School of Medicine), Alexis Battle (Department of Biomedical Engineering, Whiting School of Engineering), Kasper Hansen (Department of Biostatistics, Bloomberg School of Public Health), Ali Afshar (Department of Biomedical Engineering, Whiting School of Engineering)

The goal of our project is to develop an analytics software tool to identify patients diagnosed with Heart Failure (HF) and at high risk of 30-day hospital readmissions based on physiologic and clinical data obtained from Electronic Health Records (EHRs). Cardiovascular disease remains the leading cause of death in the United States, with HF specifically affecting nearly 6 million US adults. HF is the most common cause for hospital admission among people over 65 years of age, and it accounts for a high percentage of hospital readmissions. The estimated cost for treatment of HF in Medicare recipients is $31 billion and is expected to increase to $53 billion by 2030.

To conduct this project, our team obtained access to the MIMIC-III (‘Medical Information Mart for Intensive Care’) clinical database from Beth Israel Deaconess Medical Center of Harvard Medical School. The MIMIC-III dataset was derived from the Electronic Health Records of ~54,000 unique hospital admissions to critical care units at Beth Israel. The dataset includes patient demographics, medical diagnoses, procedures, laboratory values, and physiologic data from patients admitted to the critical care units between 2001 and 2012.

To characterize the patterns and features of time series physiologic data that could be related to 30-day readmission among HF patients, our team worked on cleaning the large-scale physiologic data from MIMIC-III database by removing the artifacts specific to physiologic signals of interest, which include Blood Pressure and Heart Rate, among others. Table 1 represents the number of physiologic records in the database and the corresponding number of analyzable records that we identified after applying our artifact correction algorithm. Figure 1 represents an instance of a set of raw (i.e. unprocessed) Arterial Blood Pressure signals (Systolic, Diastolic and Mean) obtained directly from the database, and its processed version after applying our artifact correction algorithm.

<table>
<thead>
<tr>
<th>Item</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of Physiologic Records</td>
<td>22246</td>
</tr>
<tr>
<td>Minute Records (Sampling Frequency: 0.01667 Hz)</td>
<td>13310</td>
</tr>
<tr>
<td>Second Records (Sampling Frequency: 1.0 Hz)</td>
<td>8936</td>
</tr>
<tr>
<td>Analyzable Minute Records</td>
<td>5924</td>
</tr>
<tr>
<td>Analyzable Second Records</td>
<td>3232</td>
</tr>
<tr>
<td>Total # of Analyzable Records</td>
<td>9156</td>
</tr>
<tr>
<td>Patients Represented</td>
<td>5430</td>
</tr>
</tbody>
</table>

Figure 1. (A) Part of original ABP signal set (Red: Systolic, Green: Mean, Blue: Diastolic) obtained from the database. Black arrows point towards specific instances of artifacts present in the raw signal. (B) The processed version of this signal set obtained by our artifact correction algorithm.
Is Geo-Location Information Helpful in Trading the Spread between Under Armour and Nike?

Jim Kyung-Soo Liew (Department of Finance, Carey Business School), Tamas Budavari (Department of Applied Mathematics and Statistics, Whiting School of Engineering)

In this work we examine fundamental linkages between “location data” (a.k.a. “human movement data”) and financial market equity price behavior. We concentrate our study on a popular hedge fund trading strategy known as pairs-trading. In pairs-trading, one stock is purchased while another stock is sold-short, resulting in a bet on the price spread between these two stocks. We computed the price spread between Under Armour and Nike and examined this spread behavior with respect to several variables. We also investigated the relative volume of human traffic visiting each physical store location, as proxied from anonymous cell phone geolocation traffic. The geolocation positions were recorded intraday and cover the time span from January 1, 2018 to July 17, 2018. The data set is massive, consisting of over 47.2 million rows. Longitude and latitude coordinates are employed to create economically distinct, yet competitive, paired-physical store locations of Under Armour and Nike, respectively. We monitored the relative activities in these non-overlapping ring-fenced locations to use as a proxy for the consumers’ competing fundamental demands. In the midst of our monitoring, we gleaned the following fascinating results regarding relative location data: (1) relative location data is a statistically significantly feature at the daily frequency, but not at the intra-day frequency, (2) it is a contrarian indicator, with higher/lower relative location ratios predicting next day contraction/expansion in spreads, and (3) it is consistently in the top five features across the usual suspect of machine learning algorithms. Given our limited data set, we disclaim our findings accordingly, but these initial results show that geolocation data should be a significant factor when building professional-grade pairs-trading models.

Figure 1: Mobile activity around the world from the Fysical data
to July 17, 2018. The data set is massive, consisting of over 47.2 million rows. Longitude and latitude coordinates are employed to create economically distinct, yet competitive, paired-physical store locations of Under Armour and Nike, respectively. We monitored the relative activities in these non-overlapping ring-fenced locations to use as a proxy for the consumers’ competing fundamental demands. In the midst of our monitoring, we gleaned the following fascinating results regarding relative location data: (1) relative location data is a statistically significantly feature at the daily frequency, but not at the intra-day frequency, (2) it is a contrarian indicator, with higher/lower relative location ratios predicting next day contraction/expansion in spreads, and (3) it is consistently in the top five features across the usual suspect of machine learning algorithms. Given our limited data set, we disclaim our findings accordingly, but these initial results show that geolocation data should be a significant factor when building professional-grade pairs-trading models.

Figure 2 UA and Nike Store pairs (200m to 1000m)
In Figure 2, we select 52 pairs of stores whose distance are between 200 meters and 1000 meters. Each marker on the map is one pair of Under Armour and Nike stores in the US.

Figure 3: Customer count around UA and Nike with a radius of 100 meters

(Continued on page 19)
IDIES | 2018 IDIES ANNUAL REVIEW
A Data-Intensive Approach to Large-scale Social Network Structural Analysis: Identifying Hidden Communities

Angelo Mele (Department of Economics, Carey Business School), Lingxin Hao (Department of Sociology, Krieger School of Arts and Sciences), Gerard Lemson (Department of Physics and Astronomy, Krieger School of Arts and Sciences)

Social networks are important determinants of socio-economic outcomes like education, health, information diffusion, migration decisions, and consumer choices. This project attempts to estimate complex models of social network structures. The existing studies and methods focus on estimation for moderate size networks because of computational challenges. We integrate methods from structural econometrics and agent-based modeling literature with techniques developed in the physical sciences for data-intensive computation, to estimate models for large networks. The primary objective is to identify a finite number of hidden communities as the macro property of the network structure. The first approach we are implementing is an approximate estimation method for large-scale networks with 10,000 nodes and observed individual characteristics, and ties formed based on preferential attachments. We will experiment with several alternative algorithms and modeling strategies.

At the symposium, we will present our current work on stochastic block-models, where the unobserved blocks are hidden communities of the entire network. The model assumes that there are communities of agents, which are unobserved by the researcher, and are modeled as a multinomial random variable. Conditional on the unobserved communities, a pair of individuals form links, with probability that depends on their observed characteristics and their unobserved community. We show that these finite mixture models have behavioral microfoundations and we can formulate a game theoretical model of link creation and deletion that generates stochastic block-models as equilibrium outcomes.

Performing exact maximum likelihood inference for these models is too burdensome to be practical, especially if we are controlling for observable characteristics of the individuals. Therefore, instead of focusing on likelihood estimation through Monte Carlo simulation, we are implementing approximate variational inference, an algorithm of relatively fast convergence to the final approximation. The algorithm finds the best approximation of the likelihood by minimizing the Kullback-Leibler divergence between approximating distribution and likelihood, by optimizing the variational parameters. This approach provides faster convergence of the estimated parameters, while retaining a reliable method of inference. Recent results using probability theory have shown that the estimates with this algorithm are consistent and asymptotically normal. The method is applied to a simulated network of 10,000 individuals in a first step and to China’s internal migration network generated via agent-based modeling in a second step. We will extend the methods for estimating the structure of social network snapshots and detecting structural changes of social network dynamics.

Please join us for the IDIES Annual Symposium Every year on Fall Break Day

2018 IDIES ANNUAL REVIEW | IDIES
A Big-Data Engine for Large-Scale Splicing Screens

Ben Langmead (Department of Computer Science, Whiting School of Engineering), Seth Blackshaw (Department of Neuroscience, School of Medicine), Jonathan Ling (Neuroscience, School of Medicine)

We aim to build software that enables large-scale investigation of splicing patterns in hundreds of thousands of archived RNA sequencing datasets. We will build software systems that (a) answer targeted queries about splicing relationships, e.g. between transcription factors and splicing patterns in disease, and (b) perform large-scale bulk screens to find associations between metadata variables, e.g. knock-downs and disease states, and splicing patterns. We will apply these new capabilities to find associations that can help to explain and treat disease, including neurodegenerative disease and cancer.

Progress: We are developing Snaptron 2, a redesign of the Snaptron system for enabling splicing queries that we developed in 2017. Snaptron 2 extends Snaptron’s capabilities to enable bulk queries -- the simultaneous processing of several individual queries at once. We are using these new bulk-querying capabilities to perform large-scale screens, to accelerate the quantification of a new gene annotation, and to transform splice-junction usage data into higher-level summaries of alternative splicing.

As a major case study, we are applying Snaptron 2 to study alternative splicing patterns specific to rod photoreceptors. Photoreceptors are retinal cells that sense light and transduce information for the brain. They are unique in terms of morphology, metabolism, and function — characteristics that require specialized alternative splicing patterns. We used Snaptron 2 to analyze datasets from a manually curated list of purified mouse cell types (732 run accessions) in the Sequence Read Archive (SRA), tissue datasets from the human Gene Tissue Expression Consortium (GTEx – 9,662 run accessions), shRNA-seq datasets from the ENCODE Project (1,159 run accessions), 43 single-cell studies (33,303 cells) in human and mouse including the Allen Brain Institute adult mouse primary visual cortex study. We pooled these newly-analyzed datasets with over 50,000 human RNA-seq run accessions from the Sequence Read Archive that we analyzed for previous efforts. We identified important neuron-specific splicing patterns. We also identified many exons that are highly utilized in neurons but unused in other cell types, i.e. that have a high “Percent Spliced In” (PSI) statistic in neurons but low PSI elsewhere. We can further subgroup neuron-specific exons based on their utilization in muscles and/or pancreatic islet cells. We also identified exons that were very specific to NRL+ rod photoreceptors isolated from an early developmental stage (not present in other neuronal cell types). Overall, we found many exons that were only utilized by a single cell type, despite near ubiquitous expression of the associated gene.

(Continued from page 13, Harnessing Big Data...)
Figure 3: 3D Deep Convolutional Neural Network used for predicting SCD risk

Figure 4 Time series plots of 30-minute data

Figure 5 Time series plots of daily data
(Continued from page 12, *Using Epidemiological...*)

an epidemiology-based simulation tool, operating within IDIES’ SciServer, that would enable AV R&D to generate high-resolution data of crash risk that can inform the development of AV testing programs in urban centers in the U.S.

What have we accomplished? Since March 1st, 2018 we have focused on building the database that will form the basis of simulation tool that will be hosted on IDIES’ SciServer. We have hired two research assistants who have focused on compiling and merging multiple sources of publicly available data on crashes and neighborhood characteristics to develop a crash risk profile for each section of roadway in the US.

The first step required the geo-coded crash data to be located within the road network in Maryland (Figure 1). The next step required the merging the crash data with traffic volume data to develop a measure of risk-per-road-segment. The final step will combine the crash risk data with census-level information from the block, group tract, and county level. A preliminary version of this dataset has been created, and initial analyses are underway.

The investigators have shared the concept with a number of potential future collaborators, including a major city government in the North-Eastern U.S. and an autonomous vehicle company. Both parties have expressed interest in access to the data and simulation tools that we are developing. When the tool becomes mature, it could be provided on a license or fee-for-service basis.

What will we do next? Informed by the analyses of the dataset, we will create a series of simulation models using a defined geographic area (e.g. city of Baltimore). The results of our findings will be disseminated through at least one peer-reviewed publication and a public-facing website that describes the tool and capabilities of our research team.

For future funding to support this work, we will apply to an R21 or equivalent mechanism from the National Highway Traffic Safety Authority, and the National Institute for Occupational Safety and Health. The Johns Hopkins Bloomberg School of Public Health Technology Development Accelerator may be an additional source of funding to support a commercialization pathway for this tool.

(Continued from page 14, *Data-driven Prediction...*)

generate a collection of random transformations (e.g., translations, rotations, etc.). We apply a different random transformation to each patient’s MRI stack and arrange the data into a 16x128x128 3D tensor, representing one data-point for the network.

After preprocessing, we input the data to a novel deep convolutional neural network with sequential layers (Fig. 3). The feature-extraction part of the model is made up of three 3D convolutional layers, each followed by a “max pooling” layer for subsampling. Both layer types have a stride of 1 and window sizes of 3x3x3 and 2x2x2, respectively. Finally, in the classification part, a flattening layer and two fully-connected layers (one 16-channel “ReLU” and one 2-channel “softmax”) map the data into a value in the interval 0, 1, interpreted as the probability of an SCD event. We used the Adam optimizer for the backpropagation algorithm with a learning rate of 0.001. Although we will continue to refine the network architecture, we are already seeing several encouraging results in the learning process. First, a relatively small network capacity (approx. 250,000 parameters) is sufficient for the algorithm to learn the training data, with much overfitting. Second, the small number of convolutional layers imply that the features extracted will potentially preserve interpretability, thereby revealing novel relationships between fibrosis distribution and risk of SCD. Working with the 70% of the data reserved for training and validation, k-fold validation revealed an average accuracy of over 75% for k=5 on the validation set.

Going forward, we will optimize the network hyperparameters and architecture to improve accuracy numbers. Additionally, we will further increase performance by parallelizing the current model using a second sequential branch which learns from patients’ phenotypic profile data. As we grow the dataset, we will also improve the model by using machine learning for automatically segmenting the myocardium and scar tissue, thus eliminating manual intervention in the pre-processing step. This will allow us to robustly and inexpensively assess risk of arrhythmia without the need of invasive procedures, which will result in significant improvements in patient well-being, thus having massive medical and economical societal benefits.

IDIES | 2018 IDIES ANNUAL REVIEW
OUR MISSION

Leadership
Develop intellectual leadership within IDIES to address research challenges related to the "Science of Big Data", and establish a group that leads the world in new discoveries enabled by next-generation data sets and analytics. Additionally, IDIES leadership will continue to provide coordination of integrative activities, such as seminar series and visitors.

Vision
Provide vision, leadership, and oversight for high performance and data intensive computing across all of JHU. Utilizing the same spirit proven to be highly successful over the last four years (HHPC and MARCC) to maintain a large shared facility and enable the leverage needed to seek and secure future funding opportunities.

Management
Manage and grow existing IDIES high-performance computing facilities to become a more substantial resource as we seek new funding, new partnerships, and to attract new affiliates to IDIES.

Incubator
Become an incubator for creating, curating, and 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.

Growth
Work with JHU academic departments to establish and develop new masters, graduate, and undergraduate programs that address the emerging need for a workforce with data analytic skills.

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

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.
