



DIBBs17

Final Report: 1st NSF Data Infrastructure Building Blocks PI Workshop



Chair, DIBBs17

David Lifka, Cornell University

Program Committee

Duncan Brown, Syracuse University

Stephen Ficklin, Washington State

Ken Koedinger, Carnegie Mellon

Kristin Persson, UC Berkeley

Linda Schadler, RPI

Carol Song, Purdue University

Editor

Paul Redfern, Cornell University

Contents

1.0 Introduction.....	5
2.0 DIBBs17 Welcome - David Lifka, Chair	5
3.0 Keynote: DIBBs Successes and Future Challenges - Irene Qualters, NSF	6
3.1 NSF Overview	6
3.2 OAC Staff and Unique Contribution	6
3.3 DIBBs Categories: Right for the Future?	7
3.4 Other Data Awards to Stimulate Thought (recent non-DIBBs)	8
3.5 Challenges/Next Steps for NSF and the DIBBs Community.....	9
3.6 Comments following the Keynote	10
4.0 Significant/Innovative DIBBs Results	10
4.1 Panelists	10
4.1.1 Carol Song, Purdue University Open Source, Self-Service for Geospatial Data Exploration, Computation, and Sharing.....	10
4.1.2 Feifei Li, University of Utah STORM: Towards Building a Spatial Temporal Online Reasoning and Management System	10
4.1.3 Kenton McHenry, NCSA DIBBs Brown Dog: A Science Driven Data Transformation Service.....	11
4.1.4 Catherine Larson, University of Arizona DIBBs for Intelligence and Security Informatics Research and Community	12
4.2 Roundtable Discussions/Report Outs	13
4.2.1 DIBBs Project Successes Report Outs	13
4.2.2 How to Create a Successful DIBBs Project	16
4.2.3 Posters from Recently Awarded Projects	16
4.3 Project Posters.....	17
5.0 Most Significant DIBBs Challenges/Solutions	18
5.1 Panelists	18
5.1.1 Thomas Furlani, University at Buffalo Aristotle Cloud Federation: Building a Federated Cloud Model.....	19
5.1.2 Geoffrey Fox, Indiana University Middleware and High Performance Analytics Libraries for Scalable Data Science.....	19
5.1.3 Klara Nahrstedt, University of Illinois at Urbana-Champaign 4CeeD DIBBs Challenges and Solutions	20

5.1.4 Alex Szalay, Johns Hopkins University SciServer - Long Term Access to Large Scientific Data Sets: SkyServer and Beyond.....	20
5.2 Roundtable Discussions/Report Outs.....	21
5.2.1 Report Outs on Shared DIBBs Challenges and Solutions.....	21
5.3 White Papers on Project-Specific Challenges and Solutions	23
6.0 Day 1 Takeaways.....	25
6.1 Panelists.....	26
6.1.1 Stephen Ficklin, Washington State University	26
6.1.2 Linda Schadler, Rensselaer Polytechnic Institute.....	26
6.1.3 Manish Parashar, Rutgers University.....	26
6.2 Comments following Panel.....	26
7.0 Remaining DIBBs Challenges and Future Directions	27
7.1 Panelists.....	27
7.1.1 Bonnie Hurwitz, University of Arizona Accelerating Comparative Genomics through an Ocean Cloud Commons.....	27
7.1.2 Santosh Kumar, University of Memphis Provenance-based Data Analytics CI for High-frequency Mobile Sensor Data (mProv)	28
7.1.3 Jerome Reiter, Duke University An Integrated System for Providing Access to Large-Scale, Confidential Social Science Data.....	29
7.1.4 Ken Koedinger, Carnegie Mellon University Infrastructure for Data-Driven Innovation in Education	29
7.2 Security-Related Comments following Panel.....	30
7.3 Roundtable Discussions/Report Outs: Remaining Challenges and Future Directions	30
8.0 Wrap Up and Summary Discussion.....	32
8.1 Panelists.....	32
8.1.1 Duncan Brown, Syracuse University	32
8.1.2 Victor Pankratius, MIT	32
8.1.3 Camille Crittenden, UC Berkeley	32
8.2 Summary Discussion following Panel	33
9.0 Closing Comments – Amy Walton, NSF	33

10.0 Appendices	34
10.1 NSF Dear Colleague Letter: DIBBs program PI/coPI Meeting	34
10.2 Workshop Proposal	36
10.3 Agenda	41
10.4 Attendees	44
10.5 Workshop Evaluation/Suggestions for Future DIBBs Workshops.....	48
11.0 References	56

1.0 Introduction

On August 29, 2016, a Dear Colleague Letter from the National Science Foundation's Division of Advanced Cyberinfrastructure (now Office of Advanced Cyberinfrastructure) in the Directorate for Computer & Information Science & Engineering (CISE) announced the organization of the first workshop for PIs and co-PIs funded by active awards under the Data Infrastructure Building Blocks (DIBBs) program [1].

More than 40 awards have been funded or co-funded by the DIBBs program since 2013 through a collaborative approach with all seven NSF research and education directorates, including CISE's three research divisions.

YEAR	TYPE	N	VALUE	CO-FUNDING DIRECTORATES
2013	Implementation	4	\$ 27,521,583	
2013	Conceptualization	4	\$ 429,392	
2014	Early Implementation	2	\$ 9,830,819	EdHR; SBE
2014	Pilot Demonstrations	16	\$ 21,340,996	BIO; CISE; ENG; GEO; MPS; SBE
2015	Multi-campus; Multi-institutional	5	\$ 23,685,304	*co-located with CC*
2016	Pilot Demonstrations	5	\$ 1,946,064	BIO; EdHR; ENG; GEO; MPS
2016	Early Implementations	8	\$ 28,115,008	BIO; CISE; ENG; MPS; SBE
		44	\$ 112,869,166	

Table 1: Summary of type, number, and value of DIBBs awards (2013-2016) [2].

The purpose of this workshop, as stated in the letter signed by CISE Assistant Director James Kurose, was to exchange results and lessons learned from the DIBBs projects and to outline next steps based on research advances in data. Other stated goals were to enable better communications among funded investigators, reduce programmatic redundancies, and foster team-building within and across institutional boundaries.

Subsequently, the Cornell University Center for Advanced Computing (CAC) proposed to organize, manage, and chair the workshop under the leadership of David Lifka. The purpose set forth in the Cornell proposal (pgs. 36-40) was to provide an opportunity for PIs, co-PIs, and NSF program directors to consider DIBBs project results, identify and recognize achievements, understand current challenges (technical, financial, and social), and discuss future challenges and models to address them, with the goal of informing a future vision for data cyberinfrastructure and the science and engineering disciplines it enables. A workshop agenda was established [3], [workshop website](#) launched [4], and the 1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17) scheduled.

2.0 DIBBs17 Welcome - David Lifka, Chair

David Lifka, Vice President and CIO and Director, Center for Advanced Computing, Cornell University welcomed 72 participants (pgs. 44-47) to the 1st NSF Data Infrastructure Building Blocks PI Workshop on January 11-12, 2017 at the Westin Arlington Gateway, Arlington, VA. Lifka encouraged the PIs/co-PIs to explore synergies with other DIBBs projects and to think outside the box about future infrastructures: what they will look like and how they will integrate with national cyberinfrastructure. He explained that the workshop was designed to build community amongst the participants, cross cultivate ideas, and create new ideas for the future. Lifka closed by thanking the Office of Advanced Cyberinfrastructure (OAC) for funding the event and for the outstanding venue for the 5 DIBBs workshop panels, 3 roundtable topic discussions, and poster reception. He also expressed appreciation to the participants for submitting [37 posters](#) on DIBBs successes [5] and [37 white papers](#) outlining project-specific challenges and solutions [6].

3.0 Keynote: DIBBs Successes and Future Challenges - Irene Qualters, NSF

Irene Qualters, Director of the Office of Advanced Cyberinfrastructure (OAC), gave the DIBBs17 keynote address beginning with an overview of NSF and OAC's focus, and concluding with challenges and next steps for NSF and the DIBBs community ([slides](#)) [7].

3.1 NSF Overview

- NSF by the numbers:
 - \$7.72 billion (FY16 budget request),
 - 11,000/50,000 proposals funded,
 - 38,000 reviewers,
 - 300,000 researchers at 2,000 institutions supported.
- New leadership is expected at NSF GEO and ENG, and at the National Science Board (NSB) where the first two female NSB Chairs were named last year.
- 24% of science and engineering research and 82% of computer science research is supported by NSF (percent of total federal support).
- NSF is uniquely positioned to address infrastructure needs because it is at the frontiers of basic research in multiple disciplines.
- National priorities are the focus:
 - supporting fundamental research in [food/energy/water](#) [8];
 - [understanding the brain](#) [9];
 - examining/analyzing data that's never been analyzed together before (e.g., water data with decision-support and simulation data);
 - data is varied, multi-scale, and image-based, and requires domain expertise such as physics and genomics.
- [10 Big Ideas for Future NSF Investments](#) [10] was launched by NSF director France A. Córdova. Each idea is multi-disciplinary and will require a highly capable, highly interoperable research infrastructure driven by research needs.

For example, the LIGO's (Laser Interferometer Gravitational-Wave Observatory) stunning results in the detection of gravitational waves required researcher sustained access to diverse and interoperable cyberinfrastructure (e.g., critical U.S. and international network upgrades, HPC services and resources including Open Science Grid, SDSC Comet, UIUC Blue Waters and XSEDE, and computational science advances in software infrastructure (i.e., simulations, visualization, and workflows/dataflows) [11].



Figure 1: Waves from merger of black holes envisioned in a computer simulation (SXS).

3.2 OAC Staff and Unique Contribution

- The Office of Advanced Cyberinfrastructure is a small, highly interdisciplinary staff that works with divisions and directors across the Foundation:
 - Office Director: Irene Qualters
 - Office Deputy Director: Amy Friedlander
 - Public Access: Patricia Knezek
 - Cooperative Agreements: Alejandro Suarez

- Focus areas are:
 - Data: Bob Chadduck, Amy Walton
 - HPC: Bob Chadduck, Rudolf Eigenmann, Edward Walker
 - Networking/Cybersecurity: Anita Nikolich, Kevin Thompson
 - Software: Rajiv Ramnath, Vipin Chaudhary
- Science Advisor Cross Cutting-CI: Bill Miller
- Learning/Workforce Development: Sushil Prasad.
- OAC supports research cyberinfrastructure to uniquely enable collaboration and discovery frontiers at all scales.

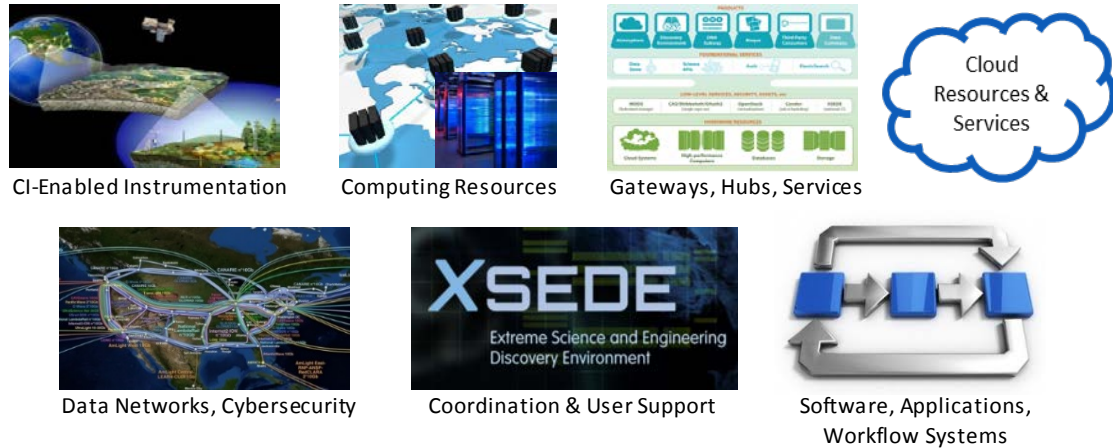


Figure 2: Shared resources, capabilities and services across the scientific workflow.

- Community input is essential to NSF planning, such as this meeting and the *Dear Colleague Letter: Request for Information on Future Needs for Advanced CI to Support Science and Engineering Research (NSF CI 2030)* initiated by OAC Science Advisor Bill Miller [12].

3.3 DIBBs Categories: Right for the Future?

- DIBBs was launched as a result of the *Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21)* report [13]. The program is crosscutting and NSF-wide with co-funding from other NSF Directorates in most solicitations. There was a strong science emphasis in 2015 focused on getting institutions to work together. A total of 43 awards have been distributed across three DIBBs categories: Generation/Acquisition/Discovery, Curation/Storage/Management, and Analysis/Modeling/Visualization. Is this the right way to think about DIBBs for the future?

DIBBs CATEGORY	N	EXAMPLE TOPICS
Generation/Acquisition/Discovery	9	Access, New Data Types, Instrument Data
Curation/Storage/Management	18	Provenance, Storage, Cloud Resources, Automatic Curation, Repository Indices
Analysis/Modeling/Visualization	16	Tools, Data Integrity and Security, Spatial Data Analysis, Collaborative Data Analysis
	43	

Table 2: DIBBs categories, number of awards (N), and example topics.

3.4 Other Data Awards to Stimulate Thought (recent non-DIBBs)

- In support of the recent Innovations at the Nexus of Food, Energy, and Water Systems (INFEWS) solicitation, OAC funded the Northern Arizona University award (“INFEWS/T1: Mesoscale Data Fusion to Map and Model the U.S. Food, Energy, and Water (FEW) System”) [14], a project with strong infrastructure requirements. The goal is to create the first national fused water map of the U.S. using geo data and a robust decision support system capable of guiding federal policy making [15].
- The Johns Hopkins University BrainLab CI award [16] is building a prototype system to aggregate multiple CI components (e.g., Jupyter Notebook, HPC cluster, cloud computing) into a workflow that will integrate thousands of MRI and neurophysiology data sets to increase experiment reproducibility and community sharing.

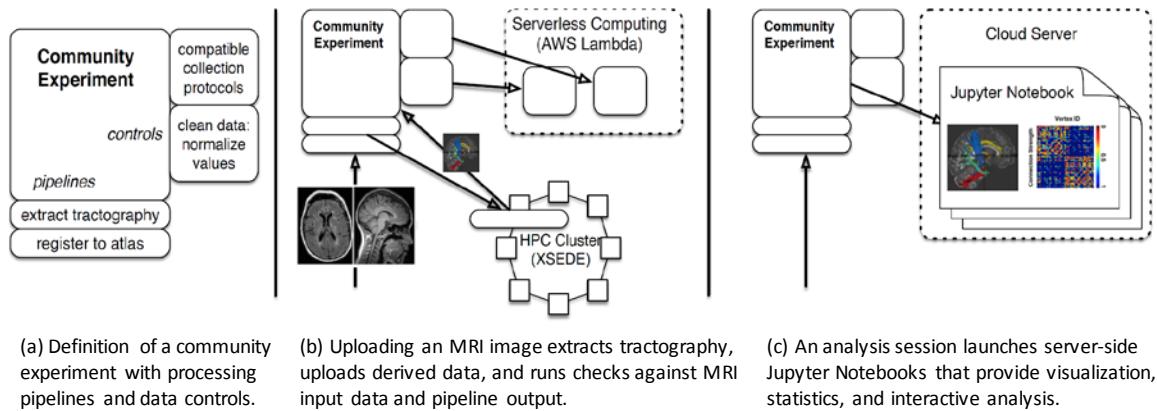


Figure 3: BrainLab CI prototypes a cloud-based experimental-management system for reproducible science.

- A Major Research Instrumentation (MRI) award to the University of Chicago (“MRI: Development of an Urban-Scale Instrument for Interdisciplinary Research”) [17] is a collaboration with the Argonne National Laboratory and the City of Chicago. ["Array of Things"](#) (AoT) instrument nodes are being mounted around the city to measure temperature, barometric pressure, vibration, carbon monoxide, ozone, ambient sound intensity, pedestrian and vehicle traffic, etc. and will make the resulting data available to the public and scientists for research [18]. This will be the first instance of a general-purpose research infrastructure that allows researchers to rapidly deploy networks of sensors, embedded systems, computing, and communications systems at scale in an urban environment.



Figure 4: 500 Array of Things sensors will be installed in the City of Chicago in 2017-18.

A second University of Chicago award, “MRI: Acquisition of a Data Lifecycle Instrument (DaLI) for Management and Sharing of Data from Instruments and Observations” [19], will enable researchers to acquire, transfer, process, store, manage and share, in a unified workflow telescopes, microscopes, parallel sequencers, and other instruments and feature scalable resources (e.g., HPC, storage, etc.).

- The final example presented was HPC/Big Data as an Enabler of NSF Big Ideas, namely, navigating the new Arctic. [ArcticDEM](#) is an NGA-ESRI-NSF public-private collaboration (University of Minnesota and 4 other universities) to produce a high-resolution, digital surface

model of the entire Arctic using 3D imagery and the Blue Waters supercomputer with the resulting data set being publicly available on Amazon Web Services [20].

3.5 Challenges/Next Steps for NSF and the DIBBs Community

An architectural vision for cyberinfrastructure for the next 5-10 years is needed. OAC displayed a preliminary "work in progress" architecture (see Figure 5 below) at the DIBBs17 poster reception. Each PI attached a DIBBs project sticker to this poster so that NSF and the DIBBs community could see where investments have been made to date.

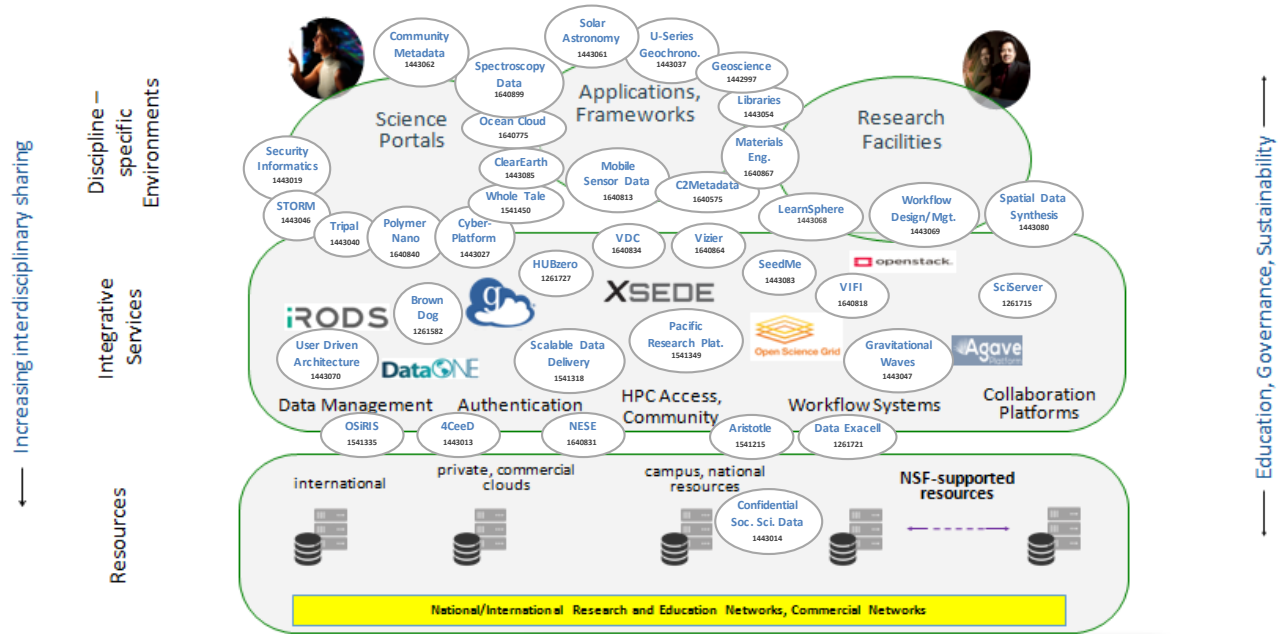


Figure 5: "Investment in Context of NSF Architectural Vision for Research CI" poster (with DIBBs projects added by DIBBs PIs).

This architectural vision for research cyberinfrastructure was also presented during the keynote presentation and workshop participants were asked to think about:

- How well would the architecture work in practice?
- How do we leverage your DIBBs work and experience?
- What are the alternatives?
- How do we optimize NSF investments to benefit cyberinfrastructure as a whole for the long term and still support near term project needs?

When considering the five data Vs: Volume, Velocity, Variety, Veracity and Value, Qualters said one area that she would like to see explored more is the 5th "V"—the sustainability and value of data over time [21]. Other questions posed were:

- How do commercial products and commodity infrastructure play (not just cloud, but software)?
- How much emphasis should be placed on research dataflows and workflows in support of "multi-cloud" and multi-institutional systems, facilities, and instruments?
- How do we qualify the degree of reproducibility required? Can I do what you do better?
- Could a statistical approach be used to verify the credibility of analysis?
- How do we create incentives for research cyberinfrastructure staff and data scientists so that they have a clear and compelling career path?

3.6 Comments following the Keynote

- In cases when reproducibility cannot be achieved, transparency should be the goal; transparency is a prerequisite for credibility.
- Computing is ephemeral, but data wants to persist (especially if it's "my" data). Guides on what data to save might be useful, keeping in mind that the data life cycle is very different for different kinds of data (e.g., almost all of detector data is tossed).
- Early access to data may be more important than long-term storage.
- Internet of Things (IoT), streaming data, distributed sensor networks, and security as it relates to data at scale are worthy research areas, perhaps with industry or interagency collaborators.

4.0 Significant/Innovative DIBBs Results

Four panelists—Carol Song, Feifei Li, Kenton McHenry, and Catherine Larson— moderated by Ashit Talukder, UNC Charlotte, discussed their most significant and innovative DIBBs results, including advances in science, innovative technologies, and new capabilities for the community. A question/answer period followed.

4.1 Panelists

4.1.1 Carol Song, Purdue University—Open Source, Self-Service for Geospatial Data Exploration, Computation, and Sharing ([slides](#)) [22]

- Modeling a domain science is difficult: project teams must be managed; data shared; metadata tracked; data browsed, searched, and viewed; tools created and supported; and, results published.
- Creating a software stack for spatial data is not trivial. What formats can you handle? Can you transform data into a format I can display? How do I serve the data? There are many choices to make for a wide range of demands (e.g., agricultural economics, climate science, remote sensing, meteorology, hydrology, crowdsourcing, training, etc.).
- Geospatial Analysis Building Blocks (GABBs) is integrating geospatial modeling data and analysis toolkits into the HUBzero software stack so users can create their own workflows and deploy, analyze, explore and share interactive geospatial apps on the web. There were 7,100 GABBs users in 2016. A data service API is available for 3rd party tools such as the iPhone. Plans are also underway to make GABBs available as a one-click start-up on AWS (test VMs are now available) and as a Linux package.
- Visit [MyGeoHub](#) [23] to request a demo or to sign up to try GABBs. See the project [poster](#) [24] and [white paper](#) [25] for additional information.

4.1.2 Feifei Li, University of Utah—STORM: Towards Building a Spatial Temporal Online Reasoning and Management System ([slides](#)) [26]

- Distributed spatial and temporal data is massive (e.g., 10TB/day NASA robotic mission data, 6PB/day WDCC weather data, 323TB AT&T phone call information, 42TB Amazon.com consumer and product data).
- The STORM platform is an automatic query engine for large, heterogeneous data that reduces analysis time from hours to minutes by looking at each data source (e.g., social media, sensors, and transaction data logs) and merging summaries of the data rather than integrating the raw data at scale. Queries, combined with machine learning techniques, are sent over the data summaries

to enable effective approximate analysis and data integration. In some cases, in-memory computation over a cluster (rather than hard disks) can enable 100TB analyses on 1000 machines in 1-5 minutes rather than ½ to 1 hour.

- Application examples are: (1) [MesoWest Weather Data](#) [27] - 40,000 weather stations manage many sensors with different formats that have produced 10 billion readings to date. STORM allows them to quickly interact and query such data, (2) [Interactive Health Indicators by County](#) [28] - STORM is delivering interactive summaries of map layers based on percent of tweets about exercise, fast food, healthy food, etc. and health indicators such as average caloric density of food, age adjusted mortality, premature mortality rate, diabetes, obesity, and physical inactivity.

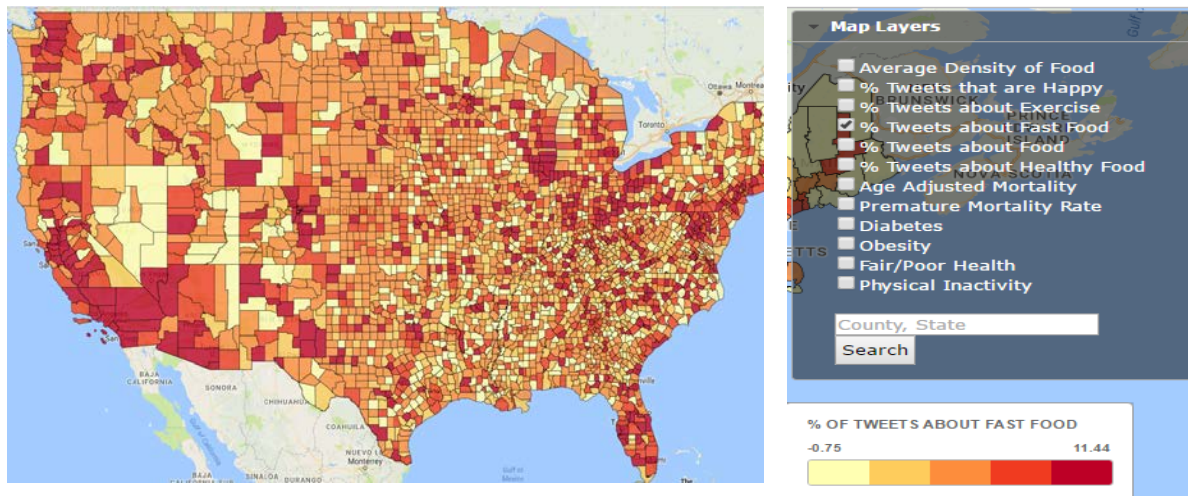


Figure 6: Interactive Health Indicators by County based on percent of tweets about exercise, fast food, healthy food, etc.

- For more information on STORM, see the project [poster](#) [29] and [white paper](#) [30].

4.1.3 Kenton McHenry, NCSA—DIBBs Brown Dog: A Science Driven Data Transformation Service ([slides](#)) [31]

- Vast amounts of digital data (e.g., collections of images, video, audio, documents, spreadsheets) are difficult to search, use, and reproduce because they are trapped in difficult-to-read formats and because the metadata is nonexistent.
- Brown Dog provides “data wrangling” services for data format conversion, metadata extraction from existing data, and the indexing of uncuration collections of data. It provides the framework for an extensible suite of new and existing tools and an API for others to build on. Over 200,000 files/datasets have been transformed to date.
- Brown Dog supports the data conversion and extraction/analysis needs for a broad range of communities. Eighty-two tools are in beta release spanning Ecology, Biology, Hydrology, Civil Engineering, Social Science, as well as general users. Examples are: (1) SEAD (Sustainable Environment Actionable Data) - a smart drop box interface that overlays floodplains extracted from LIDAR data and historic river locations from digitized maps to improve flood basin hydrology, (2) [PEcAn](#) (Predictive Ecosystem Analyzer) [32] – transforms weather and vegetation data to 30 different models to enable climate prediction forecasting and metadata analysis.

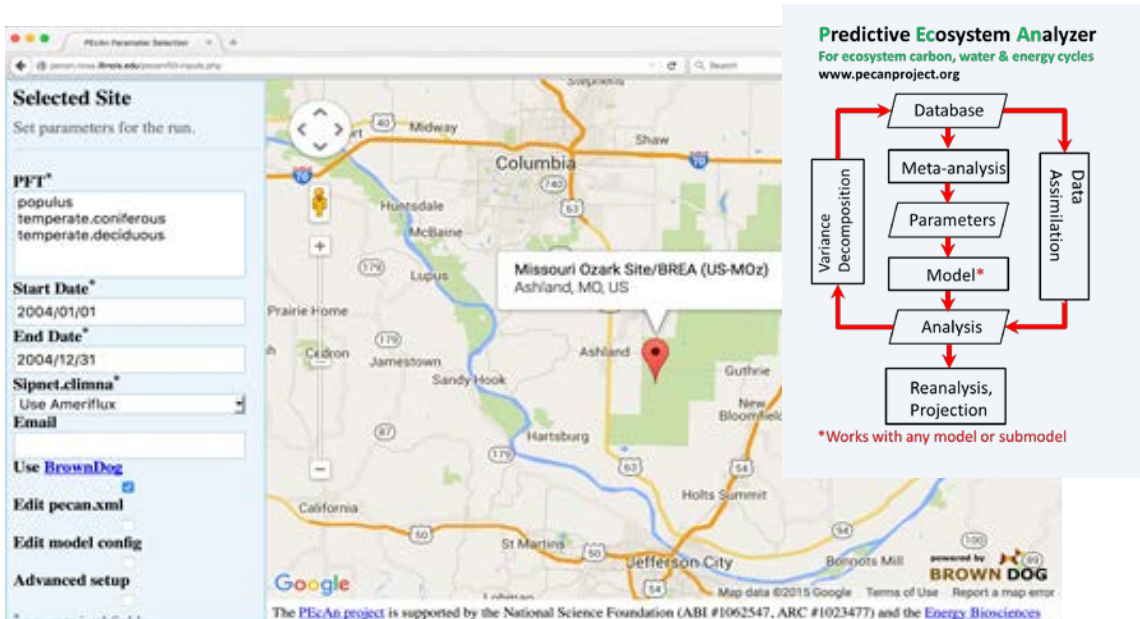


Figure 7: Predictive Ecosystem Analyzer modules used for climate forecasting run within the Brown Dog framework.

- Another important Brown Dog innovation is the development of the Distributed Data Transformation Services which is analogous to a DNS for applications to build on.
- Visit [Brown Dog](#) [33] to learn more or see the project [poster](#) [34] and [white paper](#) [35].

4.1.4 Catherine Larson, University of Arizona—DIBBs for Intelligence and Security Informatics Research and Community ([slides](#)) [36]

- Since 2000, over 73,000 terrorist attacks have killed more than 170,000 people globally. The economic impact of terrorism reached \$90 billion in 2015 [37]. Cyber crime costs are estimated to be \$400 billion a year, with estimates going as high as \$2 trillion by 2019 [38].
- DIBBs for ISI (Intelligence and Security Informatics) develops research infrastructure (e.g., online archives and analysis tools) that integrate divergent data sources allowing an estimated 4000+ scientists, students and others in the security community to collaborate, track down malicious actors, and identify weaknesses in cyberinfrastructure networks.
- COPLINK (the “Google for cops” used by over 6,000 law enforcement agencies) was developed by Hsinchun Chen and University of Arizona researchers under NSF and NIH funding and was subsequently commercialized by IBM [39]. It led to the [dark web and geopolitical web research](#) [40] which generates massive amounts of data (over 39 million extremist and social movement postings in multiple languages) to study extremism and other state and non-state movements.
- DIBBs for ISI developed [PhishMonger](#) [41], a new DIBBs tool that collects live phishing websites for study and analysis (171,360 sites and 129 targeted brands as of 9/2016). For information on this and other ISI data sets, visit [AZSecure-data.org](#) [42], a data science testbed for security researchers.
- Future plans include the NSF-funded Hacker Web project that will integrate 185+ million records from 79 platforms to study the international hacker community and behavior, including darknet marketplaces, hacker forums, and bitcoin transactions.
- See the Intelligence and Security Informatics [poster](#) [43] and [white paper](#) [44] for more information.

4.2 Roundtable Discussions/Report Outs

Nine roundtables discussed their most significant/innovative DIBBs results to date and provided report outs to the group. In addition, [37 project posters](#) (pgs. 17-18) were on display during the workshop and poster reception [45]. Information from these posters supplemented the “DIBBs Project Successes Report Outs” below.

4.2.1 DIBBs Project Successes Report Outs

- [Aristotle Cloud Federation](#) [46] – deployed DIBBs storage assets at Cornell University, University at Buffalo, and UC Santa Barbara. Created a schema for federated allocations and accounting. Redesigned the XDMoD data warehouse to work with a new federated framework that when completed will report on cloud resource usage and performance and integrate a prediction tool for optimal resource selection. Used IoT sensors, an Edge Cloud, Aristotle, and a machine learning (ML) model to improve food production and security at the Sedgwick Reserve.
- [CI for Interpreting and Archiving U-series Geochronologic Data](#) [47] – developed a data dictionary and schema, and compiled data from many different labs into a large database published recently by PALSEA (PALEo constraints on SEA level rise). Identified both a minimum set of data needed to reproduce U-series dates, and a larger set of recommended data for quality control and assessment. A recently submitted paper argues that these should become data reporting standards for the larger U-series community [48].
- [ClearEarth](#) [49] – documented, schematized, made the software more robust, and published a systematic approach to seed machine learning and NLP processes with annotated texts. Produced the first set of annotation guidelines: "Sea Ice Based on Reference Ontologies" [50]. Completed double annotation and adjudication of the Semantic Role Labels for 41300 instances in sea-ice, verifying that the annotation faithfully captures the content of "gold standard" ontologies.
- [Computer-Aided Discovery in Geoscience](#) [51] – used a novel computer-aided methodology and infrastructure, and incorporated models of volcanology physics to analyze volcanic deformations. Discovered two previously unnoticed transient inflation events at Alaskan volcanoes which was published in the *Journal of Volcanology and Geothermal Research* [52].
- [Confidential Social Science Data Access](#) [53] – created a fully synthetic federal Office of Personnel Management (OPM) database, including synthetic careers. Verification measures for Level 1 users all satisfy differential privacy. Successfully tested remote data access by approved researchers from multiple universities.
- [DataCenterHub](#) [54] – developed an innovative approach and features to systematically collect, curate, and preserve data. These include (1) a discipline-neutral organization with datasets organized as experiments with common top-level metadata, (2) a simple tabular interface that is intuitive and scalable, (3) annotative file collections in which hierarchies are extracted automatically and used as annotation with all metadata becoming searchable fields, (4) parameters for custom interaction by researchers.
- [Data Exacell](#) [55] – SLASH2 wide area file system and other DIBBs were extended, tested, and made production quality. Applications such as the Pittsburgh Genome Resource Repository and the Casual Web were built on them and then transitioned to production on Bridges.

- [4CeeD](#) [56] – developed the 4CeeD Uploader to securely upload files; the 4CeeD Curator to perform adaptive data collection from instruments by wrapping data with metadata in real-time; and, the 4CeeD Cloud Coordinator to filter data, extract metadata from microscopic images, find correlations, and identify new dependencies between materials and semiconductor device fabrication processes. Labs testing 4CeeD 1.0 report significant time and cost reductions.
- [Gravitational-Wave Science](#) [57] – made a significant impact on LIGO's search for black holes [58] and the reliable use of the Open Science Grid by hardening Pegasus' existing data re-use capabilities; improving the Pegasus Stampede Dashboard for visualization of workflow status and progress; providing tools to integrate the Dashboard into workflows and reuse; and, developing an initial metadata model for gravitational-wave science.
- [LearnSphere](#) [59] – doubled to 1,300 the learning data sets available for sharing and analysis at LearnSphere. Developed DiscourseDB to enable research and practice on collaborative and discussion-based learning. Developed new MOOCdb capabilities. Distributed the DataShop data analysis service version for the learning science community. Released LearnSphere's workflow authoring tool contributed to by many, and published results with analytics.
- [Local Spectroscopy Data Infrastructure \(LSDI\)](#) [60] – developed a fully automated workflow for NMR data with the results stored in a database for distribution and further analysis. Created fully functional and tested software tools for the high throughput generation and storage of X-Ray Absorption (XAS) data. Currently, under production on K-edge spectra. Over 10,000 spectra for 1,500 compounds from the Materials Project [61] database have been successfully computed.
- [Middleware and High Performance Analytics Libraries for Scalable Data Science](#) [62] – delivered and extended the HPC-ABDS (Apache Big Data Stack) as Cloud-HPC interoperable software with the performance of HPC. MIDAS integrating middleware now has architecture for Big Data analytics, and an integration of HPC in communications and scheduling on ABDS. SPIDAL (Scalable Parallel Interoperable Data Analytics Library) [63] now has 20 members with domain-specific and core algorithms.
- [MI-OSiRIS \(Multi-Institutional Open Storage Research InfraStructure\)](#) [64] – deployed a single, software-defined storage infrastructure to support computation "in place" at the University of Michigan, Michigan State, and Wayne State University. MI-OSiRIS uses Ceph storage building blocks for seamless expansion as well as SDN, network-topology mapping, and perfSONAR-based components for network optimization. A fourth site was successfully deployed at SC16. The team is currently working with ATLAS to scale interfaces to our service to meet their needs.
- [NanoMine](#) [65] – 85 papers were included in NanoMine under a consistent schema. Characterization tools and multi-scale models were made available to the polymer nanocomposites community. A preliminary XML schema was developed, as well as an ontology called Nanopublications. It uses existing standards to encode uploaded data and results as manageable units of experimentally or computationally supported knowledge.
- [Pacific Research Platform \(PRP\)](#) [66] – integrated campus Science DMZs into a high-capacity CA regional network so large amounts of data can be moved between scientist's labs, and their collaborators' sites, supercomputer centers or data repositories. In tests of new 100G hardware at SC16, PRP partners achieved 74Gb/s memory-to-memory between Stanford and Salt Lake City.
- [Scalable Capabilities for Spatial Data Synthesis](#) [67] – created novel, scalable capabilities for spatial data synthesis enabled by cloud computing and cyber GIS, and innovative data models and

transformation processes that take into account quality and uncertainty across diverse and massive data sources. Multiple collaborations solving emergency management and urban problems are working actively to exploit the capabilities. Over 300 participants attended a CyberGIS [68] Summer School and training workshops.

- [SeedMe](#) [69] – developed a small-data API for visualization of transfer, one of several modular building blocks that will plug into the open source Drupal web content management system to support scientific data sharing and lightweight visualization within a web browser.
- [SciServer](#) [70] – built a 20-year open data archive for the Sloan Digital Sky Survey (4 million users). Created sustainability by modularizing the code for reuse. Currently, enabling interactive science on petascale data across many disciplines using commonly shared DIBBs.
- [Shared Services for Community Metadata Improvement](#) [71] – developed tools for evaluating metadata records and collections in multiple dialects. An Excel dashboard compares metadata collections and the Recommendation Analysis Dashboard presents four views of the results: Dialect Suitability, Signature Score, Results Summaries, and Concept Guidance. An evaluation engine is being developed. The first live implementation is available at the NSF Arctic Data Center [72].
- [Solar Astronomy](#) [73] – developed innovative measures for highly dynamic and variable types of spatio-temporal events, implemented scalable algorithms that can analyze solar astronomy big data, and discovered novel and actionable/useful knowledge.
- [Syndicate](#) [74] – made technology advancements in wide area read/write volumes and programmable storage fabric. Advancements include: (1) user defined-properties (consistency, confidentiality, authenticity, I/O side-effects), (2) sustainable compatibility (write once, use anywhere, composable drivers, live driver updates, automatic deployment), (3) minimized operations (declarative configuration, auto reconfiguration, self-healing fabric).
- [Tool Supporting Collaborative Data Analytics Workflow Design and Management](#) [75] – developed a collaborative provenance data model with a graph-level provenance querying formalism; hypergraph theory-based algorithms for provenance management and mining; and, a novel software tool that supports (a)synchronous collaborative workflow design, composition, reduction, and visualization. Also, extended the workflow tool VisTrails as a proof of concept.
- [Tripal Gateway](#) [76] – published 3 journal articles and 5 conference papers and presentations. Issued 4 software releases (Tripal v3 alpha; blend4php; Big Data Smart Socket v1.0; and, 1b1 and 1b2). Created 6 GitHub repositories. Developed a Galaxy instance for analytics and created a Tripal v3 demo site to improve the access and use of systems-genetics datasets.
- [User Driven Architecture for Data Discovery](#) [77] – developed a scalable recommendation system that can produce personalized recommendations for millions of datasets. Designed a tile-based user interface that balances graphics and text and a mechanism to convert user opinion about validity, etc. into a numerical rating scale. Created options to switch between different recommendation algorithms. Provided a pilot recommendation system for DOI-based datasets.
- [Whole Tale](#) [78] – began integrating technologies to examine, transform and seamlessly republish research data used in articles. Completed an initial system architecture and mockup of a user dashboard. Designed a data model and initial REST API. Deployed a federated storage system between TACC and NCSA. Currently, extending Globus authentication with support for ORCID.

4.2.2 How to Create a Successful DIBBs Project

During the roundtable session, PIs/co-PIs shared lessons learned on how to create a successful project:

- Line up your consortia ahead of time and pick partners who work well together and produce.
- Develop a strong relationship with your user community ahead of time.
- Create the right mix of producers and consumers.
- Build a team with the skills to integrate different, evolving technologies.
- Take exciting tools and make them more accessible (e.g., in Jupyter Notebooks, Docker containers, VMs). Create a tool sustainability plan.
- "Eat your own dog food."
- Envision how your project will integrate with national cyberinfrastructure.
- How will you enable workflow re-use?
- Should you develop a point solution for a specific problem or a general solution that has broad value but may have less value for a specific problem? What is the right balance for DIBBs projects (and for NSF funding of DIBBs?).
- If developing a point solution, how will you move it to wider distribution/community practice?
- Can you port your data science tool or technology to another scientific domain? If so, what value would it provide?
- What is the role of commercial software and tools?
- Can an industry partner enhance your proposal by reducing development time and/or costs?
- Have you considered the latest technologies and what their roles might be (e.g., Machine Learning, IoT, etc.)
- How will you know at the end of the day that your results were good?
- Remember the lessons from *The Cathedral and the Bazaar* [79].
 - "Good programmers know what to write. Great ones know what to rewrite."
 - "Treating your users as co-developers is your least-hassle route to rapid code improvement and effective debugging."
 - "Release early. Release often. And listen to your customers."
 - "The next best thing to having good ideas is recognizing good ideas from your users. Sometimes the latter is better."

4.2.3 Posters from Recently Awarded Projects

Posters from recently awarded projects (August 2016 to present) highlight project goals rather than successes. Links to these posters are located at the bottom of Table 3 in Section 4.3.

- Alter/Lyle: C²Metadata: Continuous Capture of Metadata (#1640575)
- Hurwitz: The Ocean Cloud Commons (#1640775)
- Kumar/Ives: Provenance-based Data Analytics Cyberinfrastructure for High-frequency Mobile Sensor Data (#1640813)
- Talukder/Crichton: VIFI: Virtual Information Fabric-Infrastructure for Data-Driven Decisions from Distributed Data (#1640818)
- Cuff/Youssef: NESE: The North East Storage Exchange (#1640831)
- Parashar: Virtual Data Collaboratory (VDC): A Regional Cyberinfrastructure for Collaborative Data Intensive Science (#1640834)
- Kennedy/Freire: Streamlining and Understanding Curation with Vizier (#1640864)
- Govindaraju/Rajan: Data Laboratory for Materials Engineering (#1640867).

4.3 Project Posters

Thirty-seven DIBBs projects and OAC participated in a poster reception. The table below provides links to the posters, attendees, and associated NSF award abstracts. The posters are listed in order of award, from earliest to latest. [Project posters](#) [80] are also available online at [DIBBs17.org](https://dibbs17.org).

DIBBS17 ATTENDEES	POSTER TITLE	AWARD NO.
McHenry/Dietze	Brown Dog – A Science Driven Data Transformation Service [81]	1261582
Szalay/Rippin	SciServer: Bringing Analysis Close to the Data [82]	1261715
Scott/Nystrom	Data Exacell [83]	1261721
Song/Zhao	GABBs: Geospatial Data Analysis Building Blocks [84]	1261727
Pankratius	An Infrastructure for Computer Aided Discovery in Geoscience [85]	1442997
Nahrstedt/Gupta	4CeeD: Real-Time Data Acquisition and Analysis Framework for Material-related Cyber-Physical Environments [86]	1443013
Reiter	An Integrated System for Public/Private Access to Large-scale, Confidential Social Science Data [87]	1443014
Chen/Larson	Data Infrastructure Building Blocks for Intelligence and Security Informatics Research and Community [88]	1443019
Catlin	CIF21 DIBBs: Building a Modular Cyber-Platform for Systematic Collection, Curation, & Preservation of Large Engineering & Science Data [89]	1443027
Bowring	Collaborative Research: Cyberinfrastructure for Interpreting and Archiving U-series Geochronologic Data [90]	1443037
Ficklin/Feltus	CIF21 DIBBs: Tripal Gateway – a Platform for Next-Generation Data Analysis & Sharing Platform [91]	1443040
Li	STORM: Spatio-Temporal Online Reasoning and Management of Large Data [92]	1443046
Brown	CIF21 DIBBs: Domain-Aware Management of Heterogeneous Workflows: Active Data Management for Gravitational-Wave Science [93]	1443047
Fox/Jha	Middleware and High Performance Analytics Libraries for Scalable Data Science [94]	1443054
Angryk/Martens	CIF21 DIBBs: Systematic Data-Driven Analysis and Tools for Spatio-temporal Solar Astronomy Data [95]	1443061
Habermann	Shared Services for Community Metadata Improvement [96]	1443062
Koedinger/Veeramachaneni	LearnSphere: Data-Driven Discovery and Innovation in Education [97]	1443068


Zhang	A Tool for Collaborative Data Analytics Workflow Design and Management [98]	1443069
Manepalli/Powell	User Driven Architecture for Data Discovery [99]	1443070
Wang/Keahey	Scalable Capabilities for Spatial Data Synthesis [100]	1443080
Qualters/Walton	Investment in Context of NSF Architectural Vision for Research Cyberinfrastructure [101]	
Chourasia	CIF21 DIBBs: Ubiquitous Access to Transient Data Preliminary Results via the SeedMe Platform [102]	1443083
Jenkins/Martin	ClearEarth: Preparing a Science Domain for NLP/ML, drawing on Biomedical Semantic Technologies [103]	1443085
Lifka/Furlani	Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation [104]	1541215
Peterson	Give Your Data the Edge: A Scalable Data Delivery Platform [105]	1541318
Merz/Meekhof	OSiRIS: Ceph and SDN for Multi-Institutional Research [106]	1541335
Crittenden/Papadopoulos	Pacific Research Platform [107]	1541349
Ludaescher/Gaffney	CC*DNI DIBBs: Merging Science and Cyberinfrastructure Pathways: The Whole Tale [108]	1541450
Alter/Lyle	C²Metadata: Continuous Capture of Metadata [109]	1640575
Hurwitz	The Ocean Cloud Commons [110]	1640775
Kumar/Ives	Provenance-based Data Analytics Cyberinfrastructure for High-frequency Mobile Sensor Data [111]	1640813
Talukder/Crichton	VIFI: Virtual Information-Fabric Infrastructure for Data-Driven Decisions from Distributed Data [112]	1640818
Cuff/Youssef	NESE: The North East Storage Exchange [113]	1640831
Parashar	Virtual Data Collaboratory (VDC): A Regional CI for Collaborative Data Intensive Science [114]	1640834
Schadler/McGuinness	Ontology-enabled Polymer Nanocomposite Open Community Data Resource [115]	1640840
Kennedy/Freire	Streamlining and Understanding Curation with Vizier [116]	1640864
Govindaraju/Rajan	Data Laboratory for Materials Engineering [117]	1640867
Persson	The Local Spectroscopy Data Infrastructure (LSDI) [118]	1640899

Table 3: Links to project posters, attendees, and associated NSF award abstracts (from earliest to latest).

5.0 Most Significant DIBBs Challenges/Solutions

Four panelists—Thomas Furlani, Geoffrey Fox, Alex Szalay, and Klara Nahrstedt—moderated by Kate Keahey, Argonne National Laboratory/University of Chicago, discussed their most significant DIBBs challenges and solutions, followed by a question/answer period.

5.1 Panelists

5.1.1 Thomas Furlani, University at Buffalo—Aristotle Cloud Federation: Building a Federated Cloud Model ([slides](#)) [119]

- The Aristotle Cloud Federation (Cornell University, University at Buffalo, and UC Santa Barbara) deployed first-year DIBBs storage assets and is working with 7 strategic science use cases to demonstrate the potential of federated cloud as a campus bridging paradigm. The goal is to optimize time to science by providing researchers with a portal that makes it easy to identify available cloud resources in the federation and, additionally, NSF resources such as Jetstream and public resources such as AWS.
- Two significant challenges are predicting the availability of cloud resources, and tracking and insuring the efficient use of federated resources with a new federated version of Open XDMoD that is being developed to collect usage and performance data not only from HPC centers, but from clouds and other emerging platforms. UC Santa Barbara developed a new methodology called [DrAFTS](#) (Durability Agreements from Time Series) [120] that utilizes QBETS statistics to make online forecasts of the availability of resources. In addition, DrAFTS [predicts the bid price](#) [121] necessary to guarantee a specific duration of execution in the AWS Spot Tier, decreasing the cost of bursting by a factor of 4 to 10. The challenge will be seamlessly integrating DrAFTS into the federated version of Open XDMoD.
- Other project challenges include implementing a single sign-on for federated cloud resources (solved by partnering with HPE to add OAuth 2.0 support to the Eucalyptus console) and creating a federated allocation and accounting system that tracks usage on any federated resource so allocation units can be traded before having to pay a public cloud provider.
- Visit the [Aristotle portal](#) [122] to learn more or see the project [poster](#) [123] and [white paper](#) [124].

5.1.2 Geoffrey Fox, Indiana University—Middleware and High Performance Analytics Libraries for Scalable Data Science ([slides](#)) [125]

- A collaboration of teams from Arizona, Emory, Indiana (lead), Kansas, Rutgers, Virginia Tech, and Utah, is building DIBBs to overcome CI limitations in 7 communities: Biomolecular Simulations, Network and Computational Social Science, Epidemiology, Computer Vision, Spatial GIS, Information Systems, Remote Sensing for Polar Science, and Pathology Informatics.
- The main challenge is developing a scalable library by using HPC-ABDS (High Performance Computing enhanced Apache Big Data Stack) which layers ABDS software on top of HPC technologies to improve performance. Several representative Big Data capabilities from 350+ separate HPC-ABDS software modules were selected by the team and delivered and extended.
- MIDAS (Middleware for Data Intensive Analytics and Science) is the actual software that links HPC and ABDS. It has plug-ins to Hadoop which make Hadoop and Spark run faster, and ways to make jobs run as fast as C++. As MIDAS matures, the challenge will be reworking MIDAS with the HPC-ABDS infrastructure.
- HPCCloud is used as the primary development platform because the team needs features from both. SPIDAL (Scalable Parallel Interoperable Data Analytics) is being built on top of HPCCloud with domain specific (general) and core algorithms. Twenty members have joined SPIDAL.
- Significant progress has been made in all aspects of the project. The challenge going forward is to pull all this together with good software engineering, and package and test SPIDAL and MIDAS. Good APIs are also needed (Apache libraries have poor APIs) so that application developers and XSEDE resource providers can download SPIDAL libraries and MIDAS middleware.
- For more information, see the project [poster](#) [126] and [white paper](#) [127] or read the 21-month [Progress Report](#) [128] and [Kaleidoscope of ABDS and HPC Technologies](#) [129].

5.1.3 Klara Nahrstedt, University of Illinois at Urbana-Champaign—4CeeD DIBBs Challenges and Solutions ([slides](#)) [130]

- It typically takes 20 years to go from the discovery of new materials to the semiconductor fabrication of next-generation devices based on those materials. The challenge is accelerating this cycle by speeding-up the processes of collecting data about materials and making the data available to computational tools used to develop new materials and fabricate new devices.
- Current data capture and storage in materials and semiconductor fabrication is impeded by: (1) “sneaker-net” data transfer (96% use flash drives), (2) local hard drive or cloud-based storage that does not provide any assistance in organizing the data, (3) lack of digital connections between material sciences and semiconductor fabrication areas that prohibits data access and sharing. Other challenges include highly diverse users, materials, and instrumentation (e.g., clean rooms with different types of chemicals, different types of microscopes reconfigured every hour, etc.).
- University of Illinois DIBBs collaborators developed a framework called [4CeeD](#) [131] to allow microscopes to connect to a private cloud with real-time, trustworthy upload and curation capabilities. The 4CeeD curator service collects data workloads from instruments with metadata and enables data sharing from the cloud. A coordinator service filters data, extracts metadata from microscope images, analyzes and finds correlations among the data, and identifies new dependency relations between materials and device fabrication processes.
- Feedback from test users at two major research labs indicates 4CeeD has helped them to significantly reduce time and cost, and prevent unnecessary repetitions of experiments.
- This solution integrates and enhances NCSA DIBBs Clowder/Brown Dog for smart data management. 4CeeD version 1.0 is available on GitHub. See the [poster](#) [132] and [white paper](#) [133] for more information.

5.1.4 Alex Szalay, Johns Hopkins University—SciServer - Long Term Access to Large Scientific Data Sets: SkyServer and Beyond ([slides](#)) [134]

- SciServer is in the third year of a five-year DIBBs project to develop online tools to enable researchers to cope with scientific big data. SciServer builds upon a decade long effort on the Sloan Digital Sky Survey (SDSS), SkyServer, and its ad hoc spinoffs, whose components are being re-engineered in portable, generalized building blocks. The data have been expanded to include turbulence simulations (500TB), oceanography (~50TB), and genomics data (~200TB open data, 600TB HIPAA data).
- The goal of SciServer is to provide instant, interactive access to very large, rich data. DIBBs challenges and solutions underway include: (1) changing technologies – adopted Jupyter/iPython as the main scripting platform, created large data containers (e.g., the SDSS container provides access to 150TB of raw images and spectra), and used high-speed data movement (disproportionate time is still spent on case-by-case tuning), (2) scalability and robustness – consolidated the security framework with Keystone sign-on, collecting 50 different counters of every drive every 30 seconds in a database with plans to correlate these logs with actual failures,

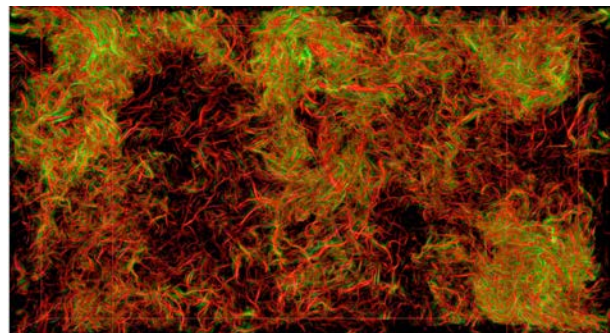


Figure 8: Tangle of vortices in a snapshot of numerically simulated 3-D isotropic fluid turbulence (Visualization by Kai Buerger based on data from Johns Hopkins database).

(3) robust environment – using at least 2 “hot” copies for most databases, heavily partitioned the data so any missing partitions can be quickly replaced, and optimized a 2.8PB backup server.

- To learn more about building a collaborative research environment for large-scale data-driven science, visit [SciServer](#) [135] or see the project [poster](#) [136] and [white paper](#) [137].

5.2 Roundtable Discussions/Reports

Nine roundtables discussed their most significant DIBBs challenges and solutions and provided report outs.

5.2.1 Report Outs on Shared DIBBs Challenges and Solutions

The roundtable report outs coalesced around the types of challenges encountered when implementing a DIBBs project and possible solutions.

CHALLENGE	SOLUTION
Communications between computer scientists and domain scientists is difficult.	Integrate computer scientists and domain scientists as a team. Meet face to face early, then regularly, by phone or in person. Embed data science students with scientists to provide a truly interdisciplinary experience. Use tools like Slack.
Scientists have difficulty appreciating the importance of CI.	Show them an initial proof of concept and communicate the "power of your approach." Get a go/no-go decision.
Scientists are not always sure what they are looking for in their data.	Develop systems for interactive prototyping back and forth to ensure regular feedback. Keep asking: Is it interesting? Is it benefiting the scientists?
Software and systems requirements often do not exist for science domains.	Consider developing a set of software and systems engineering requirements for a science domain which projects can use, extend, or make unique contributions to. Avoid reinventing the wheel.
Data owners are too busy and preoccupied to share their data or consider data to be a competitive advantage. Data may be incompatible, heterogeneous, or have privacy issues.	Short-term: make intensive contact over a period of time with selected, targeted researchers via multiple means. Long-term: effect a culture change for researchers to incorporate the future sharing of their data into a conscious part of their research process. Alternatively, consider letting researchers keep their data and move your solution to them.
Data is often stored in multiple, separate data repositories without ontology; no one else can use the data.	Fund projects that make people work together and share data with ontology.
Encourage repository administrators across institutions to participate in a federated approach to data discovery.	Rely on dataset usage rather than just on metadata (the quality and quantity of which vary substantially across communities). Access usage details from existing analytics software (e.g., Google Analytics, Kissmetrics).

Establishing a repository can have a steep learning curve.	Select a repository platform that meets users' needs for access, retrieval, and other functions, and hire an experienced developer.
Data users want data on-demand.	Make sure the user can find the data quickly, and know where the methods and algorithms are. Try to build standards with them over time.
Querying relational databases is difficult.	Future solicitations should address making interactive queries easier. Siri for SQL queries, anyone? Siri is now integrated with the WolframAlpha mathematical equation solver and there is some research underway, e.g., Natural Language Interface for Relational databases (NaLIR).
Confidentiality of data can impede scientific discovery and public policy decision-making.	Problem may be more sociological, than technical (although scrub too much data and it's less accurate). Solicitations should include how sociological issues will be overcome.
High degree of variability in programming skills among scientists and students across various scientific domains.	Include plug-ins for data manipulation, libraries within languages such as MATLAB, R, and Python, etc. Science degree programs need to recognize and emphasize the importance of programming skills.
Interoperability	Interoperability is a moving target and very hard to achieve. Try to think longer term; otherwise, in the end, it will be very hard to interoperate.
Getting people to use our software through proper design and packaging.	It takes a lot of effort to package software and tools and make them easy for a community to use. There usually isn't a budget for it. Try to design with one click to install and use. One click to log data. Design and implement suitable, effective QA/QC filters before releasing the information product. Develop a "cookbook" of protocols and virtual community via protocols.io for using your resources or cyberinfrastructure. Provide ample training and proactive outreach to users and service providers.
Getting people to know our software or tool exists.	Engage the user community early and often. Convene domain-specific user groups. Release for an extended phase of wider community testing and input. Ask for help in integrating the software. Run workshops and hackathons to launch the software and engage the community, in addition to conferences and papers. Engage campus communications professionals.
Hiring and retaining personnel with the right cross-disciplinary skills and reducing turnover.	Personnel with development, operation, and scientific use skills are hard to find and difficult to retain. New job profiles, salary scales, and career paths are needed. CI staff needs to be recognized as contributing team members and public awareness improved. Ample training/cross-training opportunities can mitigate the impact of staff turnover.

Providing DIBBs students with professional training or development opportunities.	Work with project co-PIs and partners to identify all possible avenues for training and education.
Ensuring our data will be available long term and that our work is sustainable.	Try to find a balance between specificity required by a domain science and general applicability. Define what long term success looks like early. Work with students who need data for their thesis and colleges will come up with an economic model. Expand outreach to foster community interest and strategize early on how to secure funds for long-term support. Create interoperability with other major CI efforts. Since DIBBs teams are rarely equipped with entrepreneurship skills or interests, engage business schools to build a sustainable business model that users are confident will last.

Table 4: Roundtable report outs on shared DIBBs challenges and solutions.

5.3 White Papers on Project-Specific Challenges and Solutions

The table below provides links to 37 white papers on project-specific DIBBs challenges/solutions in order from the earliest to the latest NSF award.

AUTHORS	WHITE PAPER TITLE	AWARD NO.
K. McHenry, S. Bradley, M. Dietze, P. Kumar, J. Lee, R. Marciano, L. Marini, J. McDonough, B. Minsker, A. Schmidt, B. Sullivan	DIBBs Brown Dog – The Need for and Challenges of a Science Driven Data Transformation Service [138]	1261582
A. Szalay	Long Term Access to Large Scientific Data Sets: The SkyServer and Beyond [139]	1261715
J.R. Scott, N. Nystrom, R. Roskies	The Data Exacell [140]	1261721
C. Song	CIF21 DIBBs: Integrating Geospatial Capabilities into HUBzero [141]	1261727
V. Pankratius, P. Erickson, F. Lind, M. Gowanlock, C. Rude, J. Li, G. Rongier	An Infrastructure for Computer Aided Discovery in Geoscience [142]	1442997
K. Nahrstedt, S. Konstanty, T. Nicholson, P. Ngyuen, T. Spila, T. O'Brien, M. Chan, A. Schwartz-Duval, N. Aluru, P. Braun, R. Campbell, B. Cunningham, I. Gupta, K. McHenry, J. Rogers	4CeeD: Real-Time Data Acquisition and Analysis Framework for Material-related Cyber-Physical Environments [143]	1443013

J. Reiter	An Integrated System for Providing Access to Large-scale, Confidential Social Science Data [144]	1443014
H. Chen	Data Infrastructure Building Blocks for Intelligence and Security Informatics (DIBBs for ISI) for Research and Community [145]	1443019
S. Pujol, A.C. Catlin, M. McLennan, C. Sim, L. Zilinski	CIF21 DIBBs: Building a Modular Cyber-Platform for Systematic Collection, Curation, and Preservation of Large Engineering and Science Data – A Pilot Demonstration Project [146]	1443027
J. Bowring, A. Dutton, N. McLean, K. Rubin	CIF21 DIBBS #1443037 Collaborative Research: Cyberinfrastructure for Interpreting and Archiving U-series Geochronologic Data [147]	1443037
S. Ficklin, J. Wegrzyn, F. Feltus, M. Staton, D. Main, S. Jung, K. Wang	Challenges/Future Directions for CIF21 DIBBs: Tripal Gateway, a Platform for Next Generation Data Analysis & Sharing [148]	1443040
F. Li	CIF21 DIBBs: STORM: Spatio-Temporal Online Reasoning and Management of Large Data [149]	1443046
D. Brown, E. Deelman, J. Qin	CIF21 DIBBs: Domain-Aware Management of Heterogeneous Workflows: Active Data Management for Gravitational-Wave Science [150]	1443047
G. Fox	CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science [151]	1443054
R. Angryk, P. Martens, K. Reeves	CIF21 DIBBS: Systematic Data-Driven Analysis and Tools for Spatiotemporal Solar Astronomy Data [152]	1443061
R. Habermann, M. Jones	CIF21 DIBBs: Beyond Data Discovery: Shared Services for Community Metadata Improvement [153]	1443062
K. Koedinger, K. Veeramachaneni, C. Thille, P. Pavlik, U-M O'Reilly, C. Rose, J. Stamper	Infrastructure for Data-Driven Discovery and Innovation in Education: Challenges and Future Directions [154]	1443068
J. Zhang, S. Lu	Collaborative Scientific Workflow Composition as a Service [155]	1443069
G. Manepalli, A. Powell	User Driven Architecture for Data Discovery [156]	1443070
S. Wang, K. Keahey, A. Padmanabhan	CIF21 DIBBs: Scalable Capabilities for Spatial Data Synthesis [157]	1443080
A. Chourasia, M. Norman	CIF21 DIBBs: Ubiquitous Access to Transient Data and Preliminary Results via the SeedMe Platform [158]	1443083
C. Jenkins, J. Martin, M. Palmer, S. Myers, R. Duerr, S. Ramdeen, A. Thessen, J. Preciado	CIF21 DIBBs: Porting Practical Natural Language Processing and Machine Learning Semantics from Biomedicine to the Earth, Ice and Life Sciences [159]	1443085
D. Lifka, T. Furlani, R. Wolski	CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation [160]	1541215

L. Peterson	Give Your Data the Edge: A Scalable Data Delivery Platform [161]	1541318
S. McKee	CC*DNI DIBBs: Multi-Institutional Open Storage Research InfraStructure (MI-OSIRIS) [162]	1541335
L. Smarr, C. Crittenden, T. DeFanti, P. Papadopoulos, F. Wuerthwein	Pacific Research Platform (PRP) [163]	1541349
B. Ludascher, K. Chard, N. Gaffney, M. Jones, J. Nabrzyski, V. Stodden, M. Turk, K. Turner	CC*DNI DIBBs: Merging Science and Cyberinfrastructure Pathways: The Whole Tale [164]	1541450
G. Alter	Continuous Capture of Metadata for Statistical Data [165]	1640575
B. Hurwitz, J. Hartman	CIF21 DIBBs: PD: Accelerating Comparative Metagenomics through an Ocean Cloud Commons [166]	1640775
S. Kumar, Z. Ives, I. Sim, M. Srivastava	mProv: Provenance-based Data Analytics Cyberinfrastructure for High-frequency Mobile Sensor Data – Challenges and Risk Mitigation Strategy [167]	1640813
A. Talukder	CIF21 DIBBs: EI: VIFI: Virtual Information-Fabric Infrastructure for Data-Driven Decisions from Distributed Data [168]	1640818
J. Cuff, J. Goodhue, S. Youssef, R. Shridhar, R. Zottola, C. Hill, G. Bresnahan	NESE: The North East Storage Exchange [169]	1640831
M. Parashar, V. Honavar, and the VDC Team	Virtual Data Collaboratory (VDC): A Regional Cyberinfrastructure for Collaborative Data Intensive Science [170]	1640834
L. Schadler, D. McGuinness, C. Brinson, W. Chen	CIF21 DIBBs: Ontology-enabled Polymer Nanocomposite Open Community Data Resource [171]	1640840
O. Kennedy, B. Glavic, J. Freire	Streamlining and Understanding Curation with Vizier [172]	1640864
V. Govindaraju, K. Rajan, T. Furlani, S. Setlur, S. Broderick	CIF21 DIBBs: EI: Data Laboratory for Materials Engineering [173]	1640867
K. Persson	The Local Spectroscopy Data Infrastructure [174]	1640899

Table 5: Authors and links to DIBBs17 white papers on project-specific challenges and solutions in order of NSF award (earliest to latest).

6.0 Day 1 Takeaways

Three panelists—Stephen Ficklin, Linda Schadler, and Manish Parashar— moderated by Kristin Persson, University of California, Berkeley, discussed the main takeaways from Day 1 which primarily focused on two themes: most significant/innovative DIBBs results and DIBBs challenges and solutions.

6.1 Panelists

6.1.1 Stephen Ficklin, Washington State University

As a domain scientist in genomics who “dabbles in cyberinfrastructure,” Stephen Ficklin said he took quite a few notes on software development, data management, security, id management, and the social problem of how to get people to interact with scientists and interact across institutions and architectures. He was surprised to find the depth of problems in some emerging sciences and very large infrastructure. Ficklin wondered where we will be in 2030 and expects data science will rise up and fill the niche and that the current struggles with emerging technologies will be overcome and, hopefully, democratized so that scientists will be able to afford to collect the data they need. He closed by asking if we can create components that an operating system could fire up and perform the analysis.

6.1.2 Linda Schadler, Rensselaer Polytechnic Institute

Linda Schadler, a materials science researcher, said one of her main takeaways from Day 1 was the degree to which data is siloed and stored in formats that scientists can neither reach nor share. She suggested that domain users and professional societies should consider establishing standards for what constitutes quality data and best practices for data management.

6.1.3 Manish Parashar, Rutgers University

Computer scientist Manish Parashar said that the DIBBs17 workshop posters and panel presentations were very exciting and posed three questions for the DIBBs community: (1) how do we create more outreach and communications between DIBBs projects so we can learn from one another and build a coherent cyberinfrastructure? (2) how do we transition our ideas from C.S. sandboxes to real production environments that scientists want and are able to use? (3) how do we build trust with the user community so they know our solution will be there when they need it?

6.2 Comments following Panel

Comments following the panel included:

- Scientists and computer scientists “must be married” to build successful, collaborative software; more often than not, they’re living in a parallel universe.
- The development and use of software is not given proper credit in publications. More needs to be done (e.g., Kate Keahey is Editor-in-Chief of a new journal called *SoftwareX* [175] whose goal is to acknowledge the impact of software on research).
- Software developers do not have a clear career path in academe and are often on soft money; we need to push institutions to recognize this, otherwise hiring and retaining talent will continue to plague us.
- If our model is “I’m a domain scientist and you’re a computer scientist,” that is not a transformative pathway. We need to blend the conventional domain scientist with the digital scientist, and engineer a new frontier that compels universities to respond to this convergence.
- A lot of people mistake Open Source for sustainability. Making something Open Source or even putting it on GitHub does not make it sustainable.
- Some people are innovators, others are sustainers. We need more sustainers to maintain the software and data.
- By funding XSEDE’s Extended Collaborative Support Service (ECSS) and DIBBs, NSF is changing the cyberinfrastructure and data science model in a good way.

7.0 Remaining DIBBs Challenges and Future Directions

Four panelists—Bonnie Horwitz, Santosh Kumar, Jerome Reiter and Ken Koedinger—moderated by Amit Chourasia, University of California, San Diego, discussed remaining DIBBs challenges and future directions. After Q&A, roundtables convened on the same subject, followed by report outs.

7.1 Panelists

7.1.1 Bonnie Hurwitz, University of Arizona—Accelerating Comparative Genomics through an Ocean Cloud Commons ([slides](#)) [176]

- The 3-year Tara Oceans Expedition created an unprecedented planetary scale dataset of sequencing, microscopy, and physical/chemical metadata to explore ocean biodiversity. Scientists are eager to use this data to understand the distribution of organisms across the sea and how they affect climate and ecosystems function. They have a trillion bases of metagenomic data and associated metadata (~10TB raw data).
- Only a few weeks old, this DIBBS project is developing a scalable data platform called the Ocean Cloud Commons (OCC). The cloud-based service is designed to perform big data analysis on the Tara Oceans Expedition data. It will be deployed as a large-scale prototype in OpenCloud and use an algorithm that computes all-vs.-all sequence analyses in a Hadoop framework.
- The goal is to facilitate the systematic study of the entire spectrum of microbial life—viruses, bacteria, archaea, protists, and metazoans—in oceans all around the world and enhance reproducibility. The challenge is to: (1) integrate large-scale -omics datasets and additional data from optics and satellites, (2) interlink physiochemical and environmental content that originates in the ocean, (3) examine spatial scales across diverse ecosystems, (4) cooperate among disciplines to harmonize data and put it into frameworks, (5) maintain all data and sampling and processing protocols. Processing that took 6 months is expected to be reduced to 1 day.
- OCC partners are CyVerse Cyberinfrastructure, Agave Platform, OpenCloud, XSEDE, and the Texas Advanced Computing Center. To learn more, see the [poster](#) [177] and [white paper](#) [178].

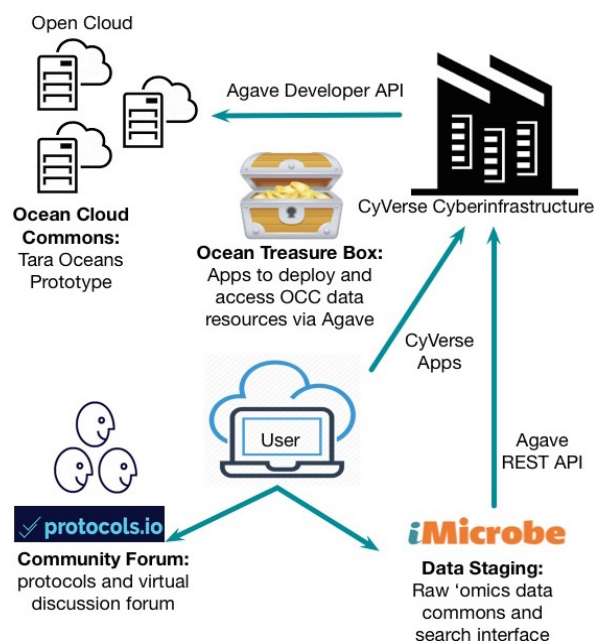


Figure 9: A CI to deploy and analyze 'omics data in the Ocean Cloud Commons through: (1) staging raw data in an iMicrobe and transferring to CyVerse via the Agave REST API, (2) developing Apps that convert raw data into persistent comparative metagenomic data clouds via the Agave Developer API (Ocean Cloud Commons and Ocean Treasure Box), (3) developing Apps that analyze data in the Ocean Cloud Commons via the Agave Developer API (Ocean Treasure Box), (4) developing a "cookbook" of protocols and virtual community via Protocols.io for using these resources.



Figure 10: Tara Oceans Schooner

7.1.2 Santosh Kumar, University of Memphis—Provenance-based Data Analytics Cyberinfrastructure for High-Frequency Mobile Sensor Data (mProv) ([slides](#)) [179]

- Smartphones, wearables such as fitness trackers, and other mobile sensor devices are continuously streaming data which have a potential to advance science and improve human health and wellness.
- The NIH Mobile Sensor Data-to-Knowledge (MD2K) Center of Excellence is conducting ongoing health studies, building tools and predicting behavior based on wearable sensor data [180]. High data rate sensors present many data challenges: velocity (hundreds of samples per second per sensor), variety (tens of sensors per sensor), volume (GBs per day per person), variability (variations in attachment, placement, signal quality), veracity (multiple biomarkers from the same sensor), and validation (sources of validation for specific biomarkers such as stress, smoking, eating, heart motion, and drug use).

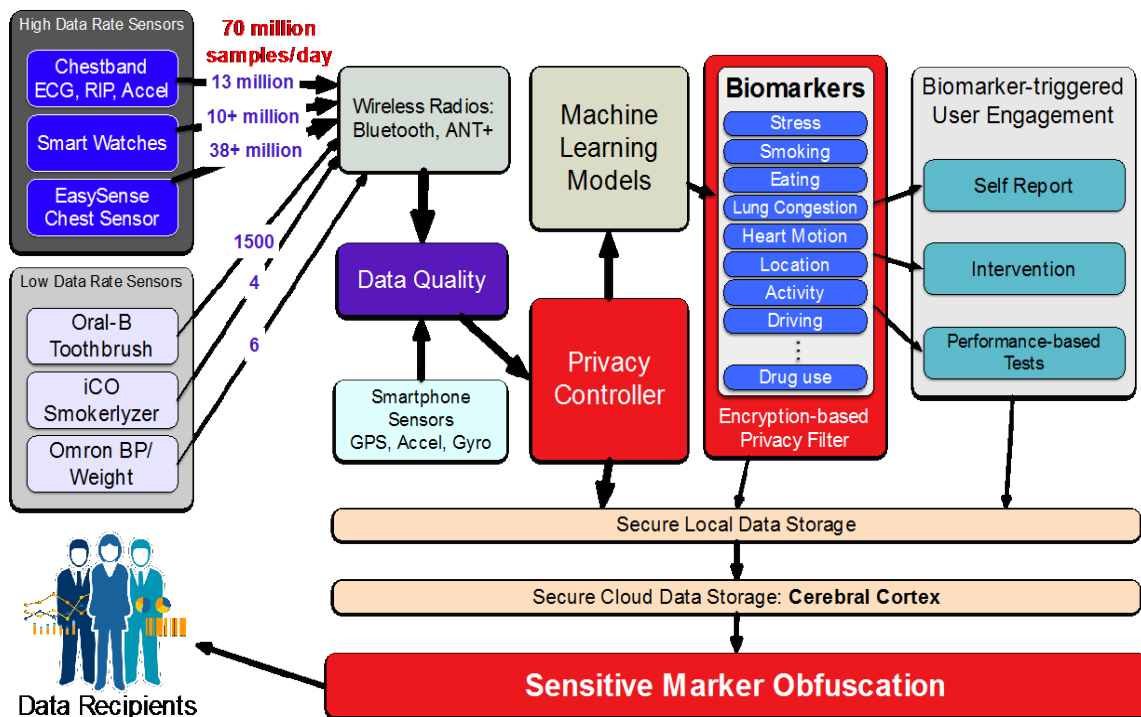


Figure 11: Ongoing real-time data collection from sensors in devices such as smartphones and wearables is enabled by MD2K's mCerebrum mobile software platform. It can capture hundreds of data points per second and uses analytics to convert streaming sensor data into biomarkers of health, behavior, and risks. ~150TBs of sensor data is at 8 sites.

- The sharing of raw mobile sensor data can accelerate research, but provenance infrastructure is needed to enable reproducibility and comparative analysis. The mProv team led by DIBBs PIs/co-PIs at the University of Memphis, U. Penn, UCSF, and UCLA and their collaborators (Ohio State, GA Tech, Open mHealth and others) is developing data models, metadata standards, API's, and runtime support for annotating MD2K data streams with source, semantics, provenance, validity, and privacy.
- The mProv provenance infrastructure will enable replay, interpretability, comparative analysis, and reproducibility for a wider research community, including industry.
- For additional information, visit [mProv](#) [181] or see the [poster](#) [182] and [white paper](#) [183].

7.1.3 Jerome Reiter, Duke University—An Integrated System for Providing Access to Large-Scale, Confidential Social Science Data ([slides](#)) [184]

- Decision makers need government agency data to inform public policies, and scientists and students need it to advance basic social, behavioral, and economic research; the question is how and among whom to share it.
- The Duke DIBBs team is developing/piloting software and methods for government agencies and data stewards to share large-scale confidential data. The initial use case is record-level data on the work histories of 3.5 million people, provided by the U.S. Office of Personnel Management (OPM). The types of analyses are: do salaries differ by gender or race, holding all else constant? What do typical career trajectories look like? What happens to government after elections?
- The integrated system has three parts: (1) unrestricted access to the fully synthetic OPM data (simulated data generated from statistical methods), (2) means for approved researchers to access confidential data with remote access solutions, glued together with, (3) verification servers that use innovative methods to allow users to access the quality of their inferences with the synthetic data so as to be more efficient with their use (if necessary) of the confidential data.
- Currently, verification measures for Level 1 (low trust) users have been completed that satisfy differential privacy, and include plots of residuals vs. predicted values for linear regression, Receiver Operating Characteristic (ROC) curves in logistic regression, and sign of and significance of regression coefficient in any model. Data is available to approved researchers via a secure server at Duke and remote access has been tested.
- Remaining challenges are seeing how the system works in practice, scaling up the infrastructure to expand to users outside the trusted group, understanding when to shut the system off after so many queries (determining the “privacy budgets”), and satisfying the university and OPM so they will want to sustain the system. See the Duke [poster](#) [185] and [white paper](#) [186] for details.

7.1.4 Ken Koedinger, Carnegie Mellon University (CMU)—Infrastructure for Data-Driven Innovation in Education ([slides](#)) [187]

- Led by CMU and collaborators at MIT, Stanford, and the University of Memphis, the LearnSphere team sees a big opportunity in the area of educational applications and more broadly in advancing learning science, but progress is hampered because data is kept in separate silos.

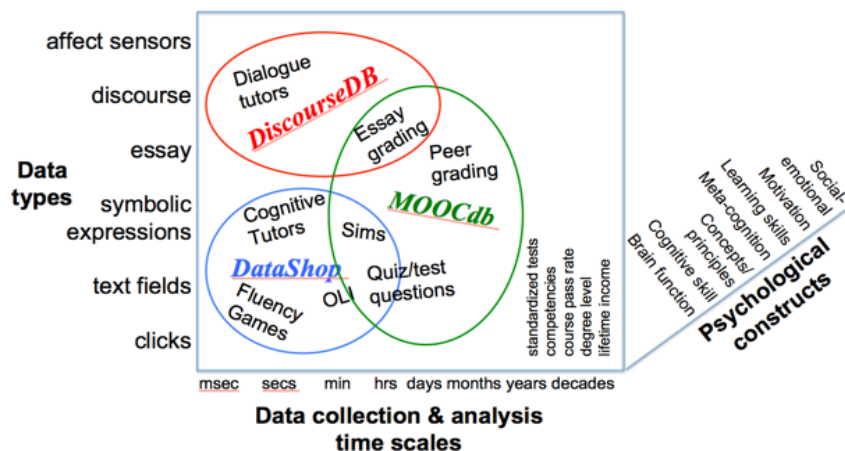


Figure 12: Thousands of education data types, data and collections with a large variation in time scales and psychological concepts are siloed, including click stream data in CMU's DataShop, MOOC analytics in MIT's MOOCdb, MOOC data in Stanford's DataStage, and language and discourse data in CMU's new DiscourseDB.

- LearnSphere integrates existing data across silos at CMU, Stanford and MIT with a web-based portal that points to these and other learning analytic resources and provides a new workflow authoring and sharing tool called Tigris.
- The goal of this DIBBS project is to make it easier for researchers, course developers, and instructors to use learning analytics to improve learning online, even without programming skills. A recent analysis across four online courses involving 5 million interactions (tutor, MOOC, and discussion board) of over 12,000 students discovered that active learning activities (e.g., answering questions with feedback) are associated with 6x more learning outcomes than are passive learning activities.
- Remaining challenges include: (1) in a community with no single data schema, how do you balance uniformity (toward maximum reusability) with flexibility in representations (to adapt to user needs)? Progress on the uniformity-flexibility challenge is being made with Tigris. (2) How do you maximize the sharing of human data without sacrificing student privacy? (3) How do you balance advancing the sophistication and variety of analytic DIBBs available to users versus having a smaller set of well-documented analytics that users can understand and trust? (4) How do you meet the occasional "need for speed"? Should this challenge be met by integrating cloud services into the workflow?
- Learn more by visiting [LearnSphere](#) [188] or by accessing their [poster](#) [189] and [paper](#) [190].

7.2 Security-Related Comments following Panel

- University administrative IT personnel are well versed in security and consider it a priority; a higher degree of security concern and practice may be necessary on the part of faculty and student researchers who are increasingly analyzing data from sensor, social media, and mobile devices. Researchers may be the weakest link that bad actors target to gain access to administrative systems.
- IT professionals understand security, faculty do not. Best practices for research data security need to be clarified and communicated, with clear delineations of responsibilities. Ensuring faculty, staff, and students receive IT security education and training was recently identified as the top strategic information security issue by the Higher Education Information Security Council [191].
- A protected research data network where sensitive data lives and network IT personnel manage the risk may be necessary for social science research that has privacy issues [192].
- Re-identification of individual genomes, inference of sensitive phenotype information from DNA sequences, and correlations between regions of the genome are threats to privacy and should be addressed in future architectures [193].
- There are best practices for security that we as a community must embrace, but we must also remember that innovation involves risk. A balance between acceptable use and privacy needs to be established for different data types.

7.3 Roundtable Discussions/Report Outs: Remaining Challenges and Future Directions

Nine roundtables discussed remaining challenges and future directions and provided report outs to the larger group.

- **Intelligent Data Collection:** For many researchers, the current approach to data collection is collect everything, then figure out later how to analyze it. We need to be more intelligent in how we collect data. CERN data collection, for example, is done in a selective, statistically valid way, using algorithms to process a subset of 600 million events per second. How will we monitor data

streams in the future in order to only collect "interesting" phenomena? Intelligent data collection ideas need further exploration.

- **Data Publishing:** Best practices for research data publishing are needed as well as an understanding of how repositories should support data publishing and "democratize" access, particularly for the "long tail" of science.
- **Internet of Things:** IoT technologies and streaming data from smartphones and wearable technologies will push data science innovations and require a better understanding of (1) how to use Edge capabilities effectively, (2) how to protect confidentiality, (3) how to choose the right technologies and methodologies to deal with dynamic data. Applying this complex digital space to science with reproducibility and transparency will be challenging.
- **Machine Learning:** ML (computer programs that change when exposed to new data) will play a greater role in intelligent data collection and analysis. For example, UC Santa Barbara's "Where's the Bear?" project [194] uses IoT sensors (cameras), an Edge cloud, and the Aristotle back-end cloud to classify over 200,000 animal images per month. Researchers used the Aristotle cloud to train a Google TensorFlow model on stock images from Google's image catalogue. The ML solution allows only images of interest to be catalogued. Machine learning algorithms may also be helpful in monitoring data corruption and flagging anomalies as data is being loaded into a system.
- **Virtualization:** Virtualizing instrumentation (e.g., mass spectrometers) so that they deliver data directly into the cloud would eliminate a lot of redundant, local infrastructure. Working with samples from the real dataset would be invaluable.
- **Containers:** Containers (e.g., Docker Swarm, Kubernetes, Mesos, Singularity) are the new "module" and their use is maturing rapidly. Many scientists, however, use containers as a "tar file" which takes extra work to integrate with other systems. Container training is needed. The ability to leverage micro-service architectures and scaling containers will grow in importance in data science.
- **Cloud Support:** The availability of better cloud support for the average scientist would be a big help when analyzing Big Data (e.g., knowing how to automatically distribute data, knowing how to build representative datasets with certain guarantees as to sampling quality, etc.).
- **Secure Use of Commercial Clouds:** Universities and public agencies will increase their use of commercial clouds in scientific research because of data sharing, bursting, and the rapid introduction and availability of new analytic and machine learning tools. While commercial clouds may be more secure than university data centers due to fewer points of entry, the risk of data breaches, compromised credentials, hacked APIs, advanced persistent threats, etc. remain the responsibility of the data steward or the university [195]. Best practices for mitigating these risks need to be researched, disseminated, and employed, with particular attention paid to issues such as liability for data breaches, terms of use, cross border transfers of data, data location, ownership of data, and authentication policies.
- **Common Research Platforms:** This workshop was valuable in providing 67 PIs/co-PIs with an opportunity to learn about other DIBBs projects and explore how they might share and leverage each other's components. Recognizing potential connections between projects, the question was asked, should more DIBBs funding be focused in the future on building common research platforms that serve multiple disciplines? Such an approach might foster more consensus and interoperability, and provide the DIBBs community with standards to adopt and/or build upon, and reduce duplication of effort.
- **Crosscutting Tools:** Opportunities to develop crosscutting tools and infrastructures should be identified and encouraged (e.g., 'omics data have disparate tools in ecological and medical

applications that may be crosscutting; gravitational wave detection patterns are very noisy and may apply to, for example, social media patterns; CyVerse [196] infrastructure is designed to work equally well on data from plants, animals, or microbes, etc.). When forced to work together, apparently disparate groups may find commonalities which can result in similar technologies being used across different domains. This approach could lead to new found efficiencies for the next generation of DIBBs. NASA has been very successful in discovering commonalities and identifying technologies that cross multiple technology areas.

- **Workflow Plans:** Workflows are essential for replication. Data management plans should include a workflow plan. How the framework will be used to conduct the data analysis should be clear.
- **Institutional Review Board (IRB):** Better approaches are needed to streamline IRB reviews and approvals for multi-site research collaborations. Oversight of these projects can be time-consuming and burdensome.
- **Sustainability:** Users have to have confidence that production systems are going to be reliable and stable for them after the project funding ends. The challenge is building a budget or sustainability model to fund future maintenance. Building trust with the user that tools and their data will be available in the future is also essential. Reuse of tools by other projects or domains may enhance sustainability.
- **Outreach:** Making researchers aware that databases are available for analyses can be challenging. Technologies such as a recommendation system (like Amazon product recommendations) that upon downloading a data set suggests additional data sets that one might be interested in, would be helpful.

8.0 Wrap Up and Summary Discussion

8.1 Panelists

Three panelists—Duncan Brown, Victor Pankratius, and Camille Crittenden—moderated by Sue Fratkin, Fratkin Associates, discussed the main takeaways from the workshop. Q&A and a summary discussion followed.

8.1.1 Duncan Brown, Syracuse University

Astrophysicist Duncan Brown said that he saw many commonalities in the 37 project posters and he thinks that many of the building blocks could fit together and be effectively applied within specific domains. He believes that the NSF can help identify commonalities and encourage proposals that deepen the collaborations between domain scientists and computer scientists.

8.1.2 Victor Pankratius, MIT

MIT computer scientist Victor Pankratius said that what works for him is to embed science use cases with the infrastructure development. Use cases ensure that infrastructure is built at the right level, neither too purpose-built nor too general-purpose. Use case teams regularly review design decisions and after installation, provide early user feedback. Multiple use case scientists are essential to good infrastructure design.

8.1.3 Camille Crittenden, UC Berkeley

Camille Crittenden, Deputy Director of the Center for Information Technology Research in the Interest of Society (CITRIS) and a speaker on human rights, technology, and new media said that future

achievements in CI are in part dependent upon developing a special academic track for data scientists so they have a solid career path. Crittenden emphasized that we all need to advocate for recognition of these intellectual achievements in tenure promotions and pressure review committees to recognize these accomplishments as an essential contribution to discovery.

8.2 Summary Discussion following Panel

- A common theme discussed on the second day of the workshop was the need for sustainability and interoperability. Project scale has a clear impact on sustainability (i.e., sustainability requires critical mass).
- Facilitating the transition from small-scale projects that are known to satisfy a critical need to large-scale, self-sustaining projects is challenging. One possibility might be to model project evolution based on NASA Technology Readiness Levels (TRL) [197]. NASA starts projects with an exploratory phase. If a project is successful, then it moves to a readiness phase with, for example, more formal interfaces, and so on. This approach enables NASA to focus on innovation and only grows projects to handle a large user base when appropriate.
- NSF might consider creating a program similar to CI-SUSTAIN to "keep the lights on" existing projects, but prioritizing efforts that link multiple projects funded by existing DIBBs awards.
- If you go from 10 users to a 1000 users who are not paying, that is a sustainability problem, not a sustainability solution.
- A broader vision for software development is needed; otherwise, we're developing components without an architecture to plug into.
- Creating a common cloud infrastructure that includes a Jupyter Notebook with plug-in tools might be helpful. Whenever a project creates a new tool, it would be a plug into that environment.
- The reward structure for domain scientists participating in computer science or cyberinfrastructure projects is weak. Being a co-author in a computer science publication is not the same recognition as being a co-author in a domain publication. Domain scientists need to see clearly that if they can get more work done by participating in data science projects, they will be able to write more papers. Domain scientists should recognize the contributions of data scientists and their platforms in domain publications.
- New DIBBS proposals should first mine existing DIBBs awards for technologies that might be re-used in the new proposal.

9.0 Closing Comments – Amy Walton, NSF

Amy Walton, a Program Director in the Office of Advanced Cyberinfrastructure, thanked the participants for being such an articulate group and for participating with exceptional energy and enthusiasm which resulted in a highly productive workshop. She explained that at this juncture, the NSF is considering what's next for DIBBs and said that the accomplishments, current challenges and future directions discussed at DIBBs17 will help answer that question.

Walton also observed that on day 2 of the workshop, all attendees were still present and engaged, so something went very right with this workshop. She encouraged the PIs and co-PIs to complete a post workshop survey conducted by Cornell and to provide comments on future workshop themes and formats (see the survey results on pgs. 48-55). Walton closed by saying she looked forward to working with this new community and thanked the participants for taking the time to attend and make new connections.

10.0 Appendices

10.1 NSF Dear Colleague Letter: DIBBs program PI/coPI Meeting

NSF 16-126

Dear Colleague Letter: Data Infrastructure Building Blocks (DIBBs) program PI/coPI Meeting

August 29, 2016

Dear Colleagues:

With this Dear Colleague Letter (DCL), the National Science Foundation's (NSF) Division of Advanced Cyberinfrastructure (ACI) in the Directorate for Computer & Information Science & Engineering (CISE) announces the organization of the first workshop for Principal Investigators (PIs) and co-PIs funded by active awards under the Data Infrastructure Building Blocks (DIBBs) program. This invitation-only workshop will take place in early FY 2017, contingent upon available funding.

BACKGROUND AND CONTEXT

Over the past decade, ACI has supported a series of programs and activities that provision cyberinfrastructure (CI) and treat CI — composed of interoperable, heterogeneous physical resources and systems, software, and data — as itself an object of research. Of critical importance is meeting the needs of the scientific and engineering community. Science and engineering use cases drive CI development, and successful CI systems strike a balance reflective of both the underlying technology and disciplinary research needs.

Since its launch in July, 2012, as part of the CIF21 Initiation,¹ the Data Infrastructure Building Blocks (DIBBs)² program has funded or co-funded more than 40 awards through a collaborative approach that involves representatives from all seven NSF research and education directorates, including CISE's three research divisions. More generally, ACI has emphasized the need to seek to situate these data infrastructure technologies and systems in contexts that also address high-performance computing (HPC) and research instrumentation, networking and security, and software together with the learning and workforce development that will enable next-generation science and engineering.

The collaborative approach that distinguishes ACI's programs in data speaks to the central role that data play in current and future scientific and engineering research, as reflected in a recent discussion about NSF's Big Ideas at the May 2016 meeting of the National Science Board (NSB).³ Many threads intersect to create the vision that was articulated at that NSB meeting: fundamental research in mathematics, statistics, and computational science; fundamental research on data topics; engagement of the research domains; embodiment of these innovations in a comprehensive data cyberinfrastructure ecosystem that enables and accelerates data-intensive research; and development and evaluation of innovative learning opportunities and educational pathways.

NEXT STEPS

To identify ways to build on prior and concurrent successes, reinforce and complement companion efforts in service of the vision outlined at the May 2016 NSB meeting, and evolve ACI's strategic programmatic needs, ACI envisions ongoing structured dialog with the data infrastructure research community, similar to PI workshops in the HPC, software, networking and cybersecurity programs. The initial PI workshop will be organized to allow PIs and co-PIs with active DIBBs awards at the time of the invitation to meet, exchange results and lessons learned, and outline next steps based on their research advances. Such workshops have historically enabled better communication among funded investigators, reduced unnecessary programmatic redundancies, and fostered team-building within and across institutional boundaries. ACI expects the PI workshop to continue on an annual basis and may eventually expand it to include wider participation from industry as well as other Federal and State agencies. The shape and form of such future opportunities are partially contingent upon the outcomes of the initial PI workshop. Next steps, including future workshops, will be announced in a future DCL later in FY 2017 or in FY 2018.

Questions about this DCL may be directed to Amy Walton at awalton@nsf.gov and Robert Chadduck at rchadduc@nsf.gov.

Sincerely,

James Kurose
Assistant Director, CISE

¹ Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF 21), https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504730.

² [NSF 16-530](#), Data Infrastructure Building Blocks (DIBBS): http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504776.

³ See especially "Harnessing Data for 21st Century Science and Engineering," https://www.nsf.gov/about/congress/reports/nsf_big_ideas.pdf.

10.2 Workshop Proposal

Summary

A NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17) will be convened in Arlington, VA on January 11-12, 2017 to exchange results and lessons learned from the projects, and to consider the implications of project results for advances in the vision and goals for data cyberinfrastructure. Prior to the workshop, each PI will submit a PDF of a poster on their DIBBs successes and a short white paper describing current and future challenges. A Program Committee comprised of a representative set of PIs will use these project-specific materials to organize the workshop panels and small discussion groups. Panels will discuss significant and innovative DIBBs results, current DIBBs challenges and solutions, and future DIBBs challenges, including sustainability issues. Each panel will be immediately followed by small group discussions and report-outs to increase PI/Co-PI participation and facilitate DIBBs community building. A 2017 DIBBs PI Workshop report will summarize the progress and challenges of the DIBBs projects and describe potential gaps and future opportunities that were discussed during the one and one half days workshop. Links to PDFs of all PI white papers and posters will be included in the report as well as the workshop website.

1.0 Overall Organization

David Lifka, Vice President & CIO and Director Center for Advanced Computing, Cornell University will serve as the workshop Chair [1]. Lifka is PI of CC*DNI DIBBs award #1541215 [2]. He and his team will be responsible for workshop planning and logistics. Lifka will work with NSF Advanced Cyberinfrastructure (ACI) Program Directors to assemble a Program Committee of representative DIBBs awardees covering the time span of the program, and representing the types of challenges addressed or future challenges anticipated. A maximum of 2 PIs/Co-PIs per DIBBs project will attend the invitation-only event for a maximum workshop attendance of 80 (40 projects x 2 attendees per project). Invitations will be distributed by NSF to DIBBs project PIs and will point to the workshop website as the source for all workshop information, registration, and requirements.

2.0 Project-Specific Materials

Each project will be asked to provide two inputs: (1) a poster describing the most significant successes to date, and (2) a short white paper identifying remaining challenges and future directions.

- **Posters:** PIs are asked to summarize progress on their awards by preparing a required poster addressing the question: “What are you most proud of about your DIBBs project?” Responses can be advances in science, innovative technologies, new capabilities for the community, major findings, or similar topics. A PDF of the poster is due December 9, 2016; PDFs will be submitted via the workshop website. The project title, PI name, and NSF award number is required at the top of each poster. Posters will be 36”H x 48”W (required size) and be displayed by the PI just prior to the Day 1 breakfast so that workshop attendees can (1) visit the posters during workshop breaks, and (2) PIs and Co-PIs can mingle and share their project successes with NSF program officers and their DIBBs colleagues at the Poster Reception scheduled at the end of Day 1.
- **White Papers:** PIs are also asked to submit a short (maximum 1-page) required white paper that addresses two questions: (1) What is the most significant challenge(s) encountered in your DIBBs project and how did you overcome it? (2) What future DIBBs challenges do you envision and are there sustainability issues or other barriers to success? PIs who just received a DIBBs award may write about anticipated challenges and how they might overcome them, in addition to whatever future challenges they envision. A PDF of the white paper is due Dec. 9, 2016; PDFs will be submitted via the workshop website. The project title, PI name, and NSF award number is required

at the top of each white paper. Papers may not exceed 1-page in length; the format is 11pt New Times Roman.

3.0 Panel & Small Group Organization and Process

The workshop Program Committee will review the white papers and posters to identify commonalities in results and unique innovations, commonalities in challenges and solutions, and insightful visions of the future of DIBBs, and invite the suitable PIs/Co-PIs to moderate or participate in three panel discussions. Each panel discussion will be followed by small group discussions (round tables of 8 or less) that will afford an opportunity for all attendees to contribute useful observations.

4.0. Proposed Dates and Location

The one and one half days workshop will occur Wednesday, January 11 and Thursday morning, January 12, 2017 in Arlington, VA, near NSF headquarters, at the Westin Arlington Gateway Hotel [3].

5.0. Agenda

The workshop will feature a series of panel discussions organized by the Program Committee around the white papers and posters submitted by PIs from all active DIBBs awards at the time of the workshop. After each panel discussion, individuals at each table will share their project experiences with each other. A volunteer at each table will summarize that table's discussion. We will assign attendees to different tables for each of the 3 small group discussions to mix up the discussion groups. This will provide attendees the opportunity to meet more people, share project experiences, and, potentially, spark future collaborations.

Day 0 - Tuesday, Jan. 10, 2017

Day/Evening Attendees Arrive at Westin Arlington Gateway Hotel (no scheduled activities)

Day 1- Wednesday, Jan. 11, 2017

7:30-9:00	Badges, Poster Setup & Continental Breakfast F. Scott Fitzgerald Ballroom A/B/C (all events)
9:00-9:15	DIBBs17 Welcome David Lifka, Workshop Chair; Vice President & CIO and Director Center for Advanced Computing, Cornell University
9:15-10:00	Keynote: DIBBs Successes & Future Challenges Irene Qualters, Division Director, ACI/CISE, National Science Foundation
10:00-10:45	Panel 1: Most Significant/Innovative DIBBs Results
10:45-11:15	Coffee Break
11:15-11:45	Small Group 1 Discussions: Most Significant/Innovative DIBBs Results
11:45-12:15	Small Group 1 Report-Outs
12:15-1:30	Lunch Buffet
1:30-2:15	Panel 2: Most Significant DIBBs Challenges/Solutions

2:15-2:45	Coffee Break
2:45-3:15	Small Group 2 Discussions: Most Significant DIBBs Challenges/Solutions
3:15-3:45	Small Group 2 Report-Outs
3:45-4:15	Discussion of Main Takeaways from Day 1 David Lifka & Program Committee
4:15-5:00	Break
5:00-6:30	Poster Reception for All DIBBs Attendees - hors d'oeuvres and cash bar
6:30	Dinner on Your Own

Day 2 - Thursday, Jan. 12, 2017

8:00-9:00	Continental Breakfast
9:00-9:45	Panel 3: Future DIBBs Challenges/Sustainability
9:45-10:15	Small Group 3 Discussions: Future DIBBs Challenges/Sustainability
10:15-10:45	Small Group 3 Report-Outs
10:45-11:00	Coffee Break
11:00-11:30	Summary Discussion & Wrap Up David Lifka & Program Committee
11:30-11:45	Closing Comments Amy Walton, Program Director, NSF
11:45	Box Lunches to go

6.0 Other Details

6.1 Hotel Meeting Room & Food Provided

The workshop is expected to take place in the F. Scott Fitzgerald Ballroom (sections A/B/C) of the Arlington Westin Gateway Hotel. Food provided will include 2 breakfasts, a luncheon buffet, hors d'oeuvres at the Poster Reception, coffee breaks, and a box lunch at the end of the workshop.

6.2 Workshop Website

There will be a workshop website hosted on Red Cloud at the Cornell Center for Advanced Computing (DIBBs17.org). Costs to host and maintain this site for five years are included as staff logistical support. The website will include workshop purpose, theme, PI paper and poster requirements/submittal forms, workshop agenda, registration, PDFs of white papers and posters, and the final workshop report.

7.0. Workshop Outcomes

The workshop outcomes and products will include:

1. White papers and posters from each PI covering topics such as successes (including new and innovative methods), challenges encountered and strategies to address them, and challenges and opportunities for future research to continue the progress made thus far.
2. A concise workshop report will summarize current project progress and challenges as well as gaps that represent future opportunities. The report will help the community and NSF better understand how DIBBs project developments advance a vision and goals for data cyberinfrastructure and advance science and engineering across disciplines.

A draft of the workshop report will be shared as a Google Doc for all to comment on for 4-weeks. A final report will be submitted to the NSF and shared with attendees and the research community within 3 months of the completion of the workshop.

Justification

The National Science Foundation *Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21)* considers an integrated, scalable, and sustainable cyberinfrastructure to be crucial for the advancement of new research practices and transformative advances across all fields of science and engineering [4]. A *Data Vision for CIF21* outlines data-specific strategies to provide a national data infrastructure [5]. The Data Infrastructure Building Blocks (DIBBs) program supports that vision by encouraging the development of robust and shared data-centric cyberinfrastructure capabilities [6]. Four DIBBs program solicitations (2012, 2014, 2015, and 2016) have resulted in 40 active awards ranging in funding from approximately \$100,000-\$10,000,000.

Dr. Lifka will serve as workshop chair. He is the PI for one of the 40 DIBBs awards (#1541215), developing advanced data-driven capabilities supporting seven major scientific user communities. This workshop provides an opportunity for the current set of DIBBs investigators, which are addressing similar data-driven issues, to discuss project results and identify best practices and gaps. The workshop website will be hosted on Red Cloud at the Cornell Center for Advanced Computing, providing access to all PI papers and posters developed for the workshop, as well as the final workshop report.

Intellectual Merit

This workshop provides an opportunity for PIs, Co-PIs, and NSF program directors to consider DIBBs project results, identify and recognize achievements, understand current challenges (technical, financial, and social), and discuss future challenges and models to address them, with the goal of informing a future vision for data cyberinfrastructure and the science and engineering disciplines it enables. This will be the first ever collective assembly of DIBBs PIs and Co-PIs awardees and, as such, it has the potential to advance the knowledge of the DIBBs program and to generate new ideas to ensure that discovery stimulated by data is properly supported by an integrated and sustainable data cyberinfrastructure. The community will gain insights into the successes and challenges faced by DIBBs projects as they address the needs of both existing and emerging research domains.

Broader Impacts

A workshop website consisting of DIBBs project posters, white papers, and a final workshop report will inform the community of data cyberinfrastructure innovations being undertaken by DIBBs projects and convey future data infrastructure needs and potential strategies to meet them. The workshop itself will help to develop a DIBBs community in which potential synergies and future partnerships may be explored. The workshop will also help NSF program officers in this cross-cutting program envision future data cyberinfrastructure program needs. The continual evolution of the DIBBs program is essential to the achievement of CIF21 goals. This workshop will further clarify the role of DIBBs in enabling data-focused services and capabilities to broadly impact and strengthen NSF's research portfolio.

Workshop Proposal References

- [1] Cornell University Office of CIO and Vice President for Information Technologies: About the CIO. Available from: <http://cio.cornell.edu/community/office-cio-and-vice-president-information-technologies>.
- [2] NSF Award Abstract #1541215 – CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation. Available from: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1541215&HistoricalAwards=false.
- [3] The Westin Arlington Gateway Hotel. Available from: <http://www.westinarlingtongateway.com/>.
- [4] Cyberinfrastructure Framework for 21st Century Science and Engineering: Vision. May 2012 [PDF]. Available from: <http://www.nsf.gov/cise/aci/cif21/CIF21Vision2012current.pdf>.
- [5] Data Vision for CIF2: Cyberinfrastructure Framework for 21st Century Science and Engineering – A Vision and Strategy for Data in Science, Engineering, and Education. April 2012 [PDF]. Available from: <https://www.nsf.gov/cise/aci/cif21/DataVision2012.pdf>.
- [6] National Science Foundation: Data Infrastructure Building Blocks (DIBBs). Available from: https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504776.

10.3 Agenda



1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17)

Final Agenda (*with links to slides*)

Wednesday, January 11, 2017

- | | |
|--------------------------|--|
| 730 am – 900 am | Continental Breakfast - Westin Arlington Fitzgerald Ballroom A/B/C
Pick up your DIBBs17 badge, setup your poster (by 900 am), and enjoy continental breakfast with your DIBBs colleagues. Sit where you like. |
| 900 am – 915 am | DIBBs17 Welcome
David Lifka, Workshop Chair
Vice President & CIO and Director, Center for Advanced Computing,
Cornell University |
| 915 am – 1000 am | Keynote: DIBBs Successes & Future Challenges
Irene Qualters, Division Director, OAC/CISE, National Science Foundation |
| 1000 am – 1045 am | Panel 1: Most Significant/Innovative DIBBs Results
Moderator: Ashit Talukder, University of North Carolina, Charlotte

Panelists:
Carol Song, Purdue University
Open Source, Self-Service for Geospatial Data Exploration, Computation, Sharing

Feifei Li, University of Utah
STORM: Spatio-Temporal Online Reasoning and Management of Large Data

Kenton McHenry, NCSA
Brown Dog: A Science Driven Data Transformation Service

Catherine Larson, University of Arizona
DIBBs for Intelligence and Security Informatics Research and Community |
| 1045 am - 1115 am | Coffee Break |
| 1115 am – 1145 am | Small Group 1 (SG1) Discussions:
Most Significant/Innovative DIBBs Results
Discuss your most significant DIBBs results (advances in science, innovative technologies, or new capabilities for the community) and, if you have one or two you'd like to share, pick a rep from your table to give a rapid report-out. |
| 1145 am – 1215 pm | Small Group 1 Report Outs:
Most Significant/Innovative DIBBs Results
Moderator: David Lifka |
| 1215 pm – 130 pm | Lunch Buffet (switch tables ~120pm, see back of badge for table number) |

130 pm – 215 pm	Panel 2: Most Significant DIBBs Challenges/Solutions: Moderator: Kate Keahey, Argonne National Laboratory/University of Chicago Panelists: Thomas Furlani, University at Buffalo Aristotle Cloud Federation: Building a Federated Cloud Model Geoffrey Fox, Indiana University Middleware and High Performance Analytics Libraries for Scalable Data Science Klara Nahrstedt, University of Illinois at Urbana-Champaign 4CeeD DIBBs Challenges and Solutions Alexander Szalay, Johns Hopkins University Long Term Access to Large Scientific Data Sets: SkyServer and Beyond
215 pm – 245 pm	Coffee Break
245 pm – 315 pm	Small Group 2 (SG2) Discussions: Most Significant DIBBs Challenges/Solutions Discuss your most significant DIBBs challenge and solution and, if you have one or two you'd like to share, pick a rep from your table to give a rapid report-out.
315 pm – 345 pm	Small Group 2 Report-Outs: Most Significant DIBBs Challenges/Solutions Moderator: David Lifka
345 pm – 415 pm	Takeaways: Discussion of Main Takeaways from Day 1 Moderator: Kristin Persson, University of California, Berkeley Panelists: Stephen Ficklin, Washington State University Linda Schadler, Rensselaer Polytechnic Institute Manish Parashar, Rutgers University
415 pm – 500 pm	Break
500 pm – 630 pm	Poster Reception
630 pm	Dinner on Your Own

Thursday, January 12, 2017

- 800 am – 900 am** **Continental Breakfast - F. Scott Fitzgerald Ballroom A/B/C**
Enjoy continental breakfast and sit at your assigned Thursday morning table prior to today's 900 am start (see the back of your badge for your table number)
- 900 am – 945 am** **Panel 3: Remaining Challenges and Future Directions**
Moderator: Amit Chourasia, University of California, San Diego
- Panelists:
Bonnie Hurwitz, University of Arizona
[The Ocean Cloud Commons](#)
- Santosh Kumar, University of Memphis
[Provenance-based Data Analytics CI for High-Frequency Mobile Sensor Data \(mProv\)](#)
- Jerome Reiter, Duke University
[An Integrated System for Providing Access to Large-Scale, Confidential Social Science Data](#)
- Ken Koedinger, Carnegie Mellon University
[LearnSphere: Infrastructure for Data-Driven Innovation in Education](#)
- 945 am – 1015 am** **Small Group 3 (SG3) Discussions:**
Remaining Challenges and Future Directions
Discuss your remaining DIBBs challenges and future directions and, if you have one or two you'd like to share, pick a rep from your table to give a rapid report-out.
- 1015 am – 1030 am** **Small Group 3 Report Outs:**
Remaining Challenges and Future Directions:
Moderator: David Lifka
- 1045 am - 1100 am** **Coffee Break**
- 1100 am – 1130 am** **Summary Discussion & Wrap Up**
Moderator: Sue Fratkin, Fratkin Associates
- Panelists:
Duncan Brown, Syracuse University
Victor Pankratius, MIT
Camille Crittenden, University of California, Berkeley
- 1130 am – 1145 am** **Closing Comments**
Amy Walton, Program Director, OAC/CISE, National Science Foundation
- 1145 am** **Box Lunches**

10.4 Attendees



1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17)

George Alter

Research Professor
University of Michigan
altergc@umich.edu

Rafal Angryk

Associate Professor
Georgia State University
angryk@cs.gsu.edu

James Bowring

Associate Professor
College of Charleston
bowringj@cofc.edu

Duncan Brown

Charles Brightman Professor of Physics
Syracuse University
dabrown@syr.edu

Ann Christine Catlin

Senior Research Scientist
Purdue University
acc@purdue.edu

Bob Chadduck

Program Director
National Science Foundation
rchadduc@nsf.gov

Vipin Chaudhary

Program Director
National Science Foundation
vipchaud@nsf.gov

Hsinchun Chen

Regents' Professor
University of Arizona
ailab@eller.arizona.edu

John Cherniavsky

Senior Advisor for Research
National Science Foundation
jchernia@nsf.gov

Amit Chourasia

Senior Visualization Specialist
University of California, San Diego
amit@sdsc.edu

Daniel Crichton

Director, CDST
JPL
daniel.j.crichton@jpl.nasa.gov

Camille Crittenden

Deputy Director
UC Berkeley
ccrittenden@berkeley.edu

James Cuff

Assistant Dean for Research Computing
Harvard University
james_cuff@harvard.edu

Michael Dietze

Associate Professor
Boston University
dietze@bu.edu

Alex Feltus

Associate Professor
Clemson University
ffeltus@clemson.edu

Stephen Ficklin

Assistant Professor
Washington State University
stephen.ficklin@wsu.edu

Geoffrey Fox

Professor
Indiana University
gcf@indiana.edu

Susan Fratkin

Fratkin Associates
sfratkin8@gmail.com

Amy Friedlander

Deputy Division Director
National Science Foundation
afriedla@nsf.gov

Juliana Freire

Professor Computer Science and Data Science
New York University
juliana.freire@nyu.edu

Thomas Furlani

Director, Center for Computational Research
University at Buffalo, SUNY
furlani@buffalo.edu

Niall Gaffney

Director of Data Intensive Computing
The Texas Advanced Computing Center
ngaffney@tacc.utexas.edu

Venu Govindaraju

SUNY Distinguished Professor and VP for
Research and Economic Development
University at Buffalo
govind@buffalo.edu

Indranil Gupta

Associate Professor
University of Illinois at Urbana-Champaign
indy@illinois.edu

Ted Habermann

Director of Earth Science
The HDF Group
thabermann@hdfgroup.org

Daryl Hess

Program Director
National Science Foundation
dhess@nsf.gov

Bonnie Hurwitz

Assistant Professor
University of Arizona
bhurwitz@email.arizona.edu

Zachary Ives

Professor
University of Pennsylvania
zives@cis.upenn.edu

Christopher Jenkins

Associate Researcher
University of Colorado, Boulder
chris.jenkins@colorado.edu

Shantenu Jha

Associate Professor
Rutgers University
shantenu.jha@rutgers.edu

Kate Keahey

Scientist
University of Chicago
keahey@anl.gov

Oliver Kennedy

Assistant Professor
University at Buffalo
okennedy@buffalo.edu

Ken Koedinger

Professor
Carnegie Mellon University
koedinger@cmu.edu

Santosh Kumar

Professor and Chair of Excellence in Computer
Science
University of Memphis
santosh.kumar@memphis.edu

Catherine Larson

Associate Director, AI Lab
University of Arizona

Feifei Li

Associate Professor
University of Utah
lifeifei@cs.utah.edu

David Lifka

Vice President and CIO
Director, Ctr. for Advanced Computing (CAC)
Cornell University
lifka@cornell.edu

Bertram Ludaescher

Professor
University of Illinois at Urbana-Champaign
ludaesch@illinois.edu

Jared Lyle

Archivist
University of Michigan
lyle@umich.edu

Giridhar Manepalli

Director of Information Management
Technology
Corporation for National Research Initiatives
gmanepalli@cnri.reston.va.us

Petrus Martens

Professor
Georgia State University
martens@astro.gsu.edu

James Martin

Professor
University of Colorado, Boulder
james.martin@colorado.edu

Peter McCartney

Program Director
National Science Foundation
pmccartn@nsf.gov

Deborah McGuinness

Professor
Rensselaer Polytechnic Institute
dlm@cs.rpi.edu

Kenton McHenry

Senior Research Scientist
NCSA
mchenry@illinois.edu

Ben Meekhof

OSiRIS Lead Engineer
University of Michigan
bmeekhof@umich.edu

Kenneth Merz

Professor
Michigan State University
merzjrke@msu.edu

Klara Nahrstedt

Director CSL/Professor CS
University of Illinois at Urbana-Champaign
klara@illinois.edu

Nick Nystrom

Senior Director of Research
Pittsburgh Supercomputing Center
nystrom@psc.edu

Victor Pankratius

Head of Astro-&Geo-Informatics Group
MIT Haystack
pankrat@mit.edu

Philip Papadopoulos

Program Director
University of California, San Diego (SDSC)
ppapadopoulos@ucsd.edu

Manish Parashar

Distinguished Professor
Rutgers University
parashar@rutgers.edu

Kristin Persson

Assistant Professor
UC Berkeley
kapersson@lbl.gov

Larry Peterson

Professor
University of Arizona
llp@cs.arizona.edu

Allison Powell

Senior Research Manager
Corporation for National Research Initiatives
apowell@cnri.reston.va.us

Irene Qualters

Division Director
National Science Foundation
iqualter@nsf.gov

Krishna Rajan

Eric Bloch Chair
University at Buffalo, SUNY
krajan3@buffalo.edu

Rajiv Ramnath

Program Director
National Science Foundation
rramnath@nsf.gov

Paul Redfern

Assistant Director, Strategic Partnerships
Cornell University Center for Advanced
Computing (CAC)
red@cac.cornell.edu

Jerome Reiter

Professor of Statistical Science
Duke University
jerry@stat.duke.edu

Michael Rippin

Project Manager
Johns Hopkins University
mike.rippin@jhu.edu

Linda Schadler

Vice Provost and Dean of Undergraduate
Education
Rensselaer Polytechnic Institute
schadl@rpi.edu

J. Ray Scott

Senior Director, Facilities Technology
Pittsburgh Supercomputing Center
scott@psc.edu

Carol Song

Senior Scientist
Purdue University

Alexander Szalay

Professor
Johns Hopkins University
szalay@jhu.edu

Ashit Talukder

Director, UNC Data Visualization Center
UNC Charlotte
atalukde@uncc.edu

Kalyan Veeramachaneni

Principle Research Scientist
MIT
kalyan@csail.mit.edu

Amy Walton

Program Director
National Science Foundation
awalton@nsf.gov

Shaowen Wang

Professor
University of Illinois at Urbana-Champaign
shaowen@illinois.edu

Saul Youssef

Research Associate Professor
Boston University
youssef@bu.edu

Jia Zhang

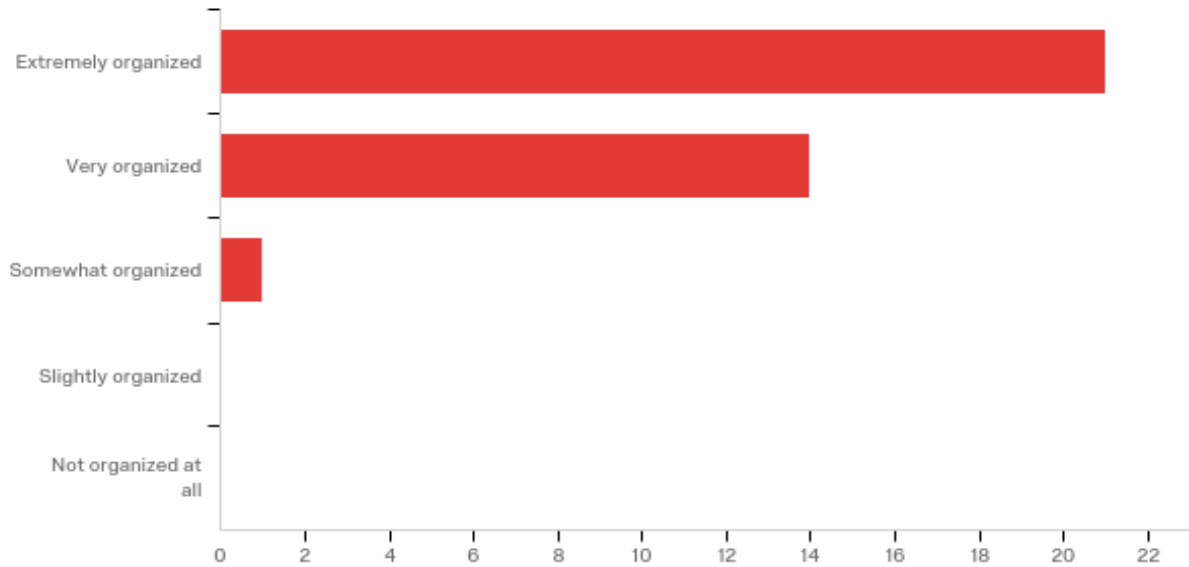
Associate Professor
Carnegie Mellon University
jia.zhang@sv.cmu.edu

Lan Zhao

Research Scientist
Purdue University
zhao4@purdue.edu

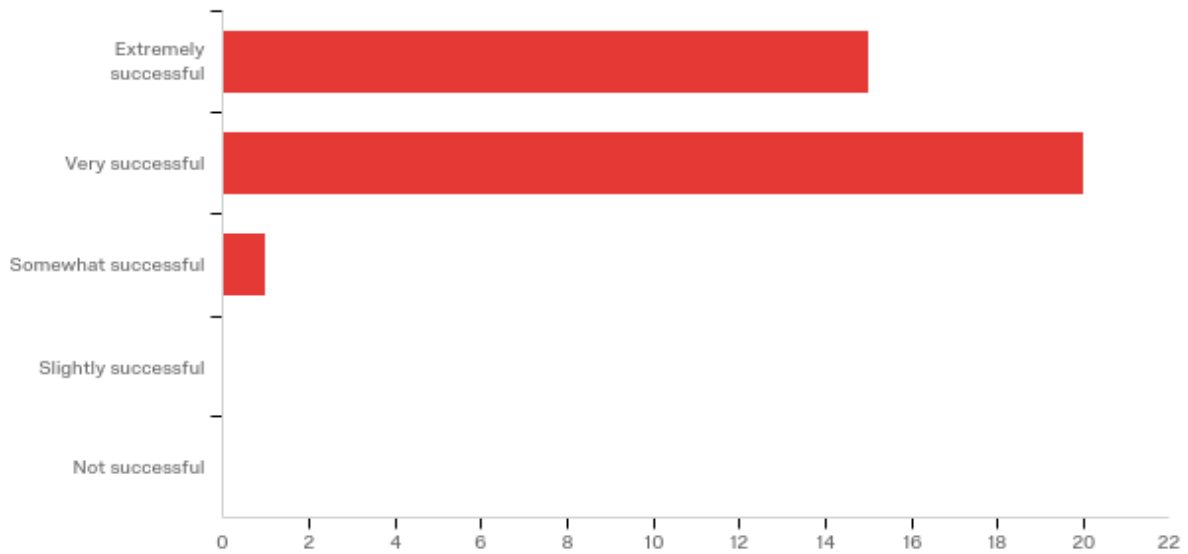
10.5 Workshop Evaluation/Suggestions for Future DIBBs Workshops

Q1 - How organized was the DIBBs17 workshop?



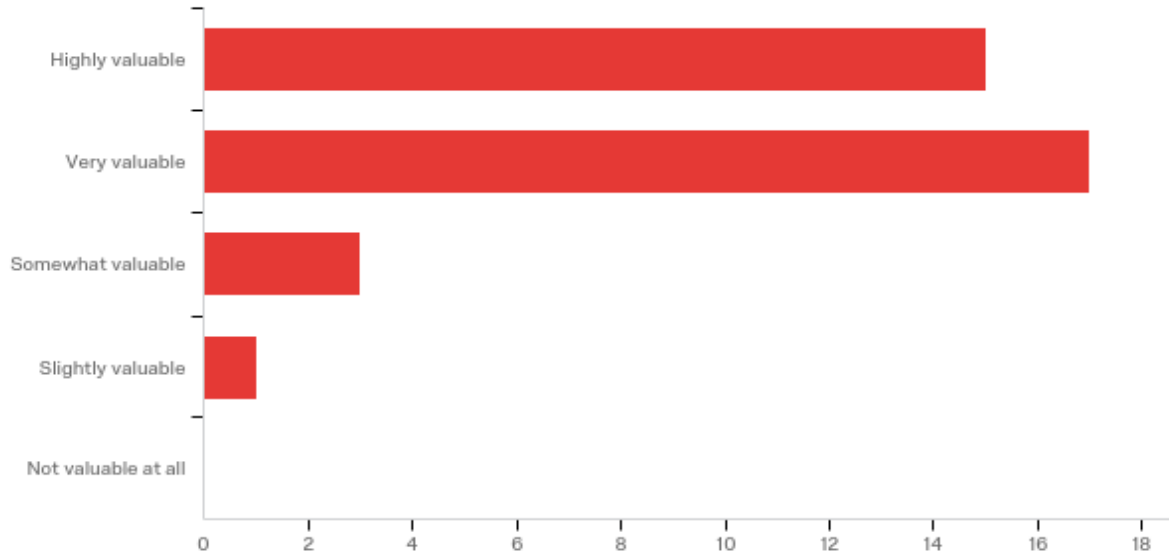
#	Answer	%	Count
1	Extremely organized	58.33%	21
2	Very organized	38.89%	14
3	Somewhat organized	2.78%	1
4	Slightly organized	0.00%	0
5	Not organized at all	0.00%	0
	Total	100%	36

Q2 - How successful was the workshop in providing you with opportunities to interact with your peers, share your views, and develop community?



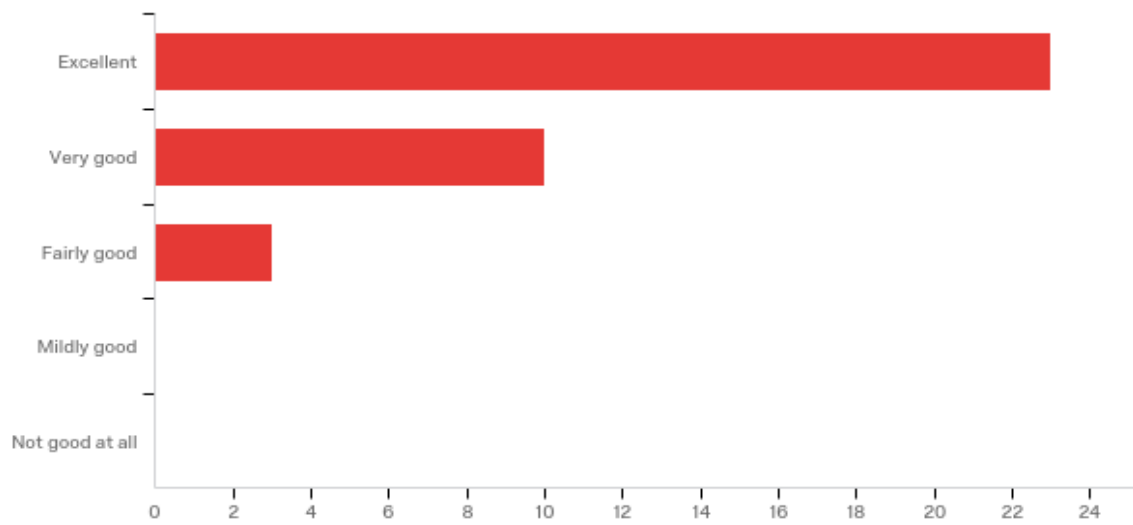
#	Answer	%	Count
1	Extremely successful	41.67%	15
2	Very successful	55.56%	20
3	Somewhat successful	2.78%	1
4	Slightly successful	0.00%	0
5	Not successful	0.00%	0
	Total	100%	36

Q3 - Did you derive value from the workshop (e.g., knowledge gained, synergies explored, etc.)?



#	Answer	%	Count
1	Highly valuable	41.67%	15
2	Very valuable	47.22%	17
3	Somewhat valuable	8.33%	3
4	Slightly valuable	2.78%	1
5	Not valuable at all	0.00%	0
	Total	100%	36

Q4 - Overall, how would you rate the DIBBs17 workshop?



#	Answer	%	Count
1	Excellent	63.89%	23
2	Very good	27.78%	10
3	Fairly good	8.33%	3
4	Mildly good	0.00%	0
5	Not good at all	0.00%	0
	Total	100%	36

Q5 - If you had to say one thing about the workshop, what would you say?

Big challenge, well met.

Highly interactive.

Only that I wish we had more time.

More of these workshops may be needed to facilitate the organization of the building block investments.

Focused.

Good discussions.

I appreciated that there was sufficient time to interact with other DIBBs groups.

Opportunity to identify challenges, solutions, future directions ACROSS projects, i.e., classifying important areas common to projects.

Great chance to meet other DIBBs PI's.

Very well organized, very informative.

Excellent for networking.

I liked the mixture of panels, followed by the discussion around the tables and their summaries.

A very good mix of size, structure, and topics that lead to very good collaborative discussions.

Strong leadership, extremely well executed meeting.

The beneficial part of the workshop was the opportunity for networking. The group discussions were diffuse and unproductive.

Excellent opportunity to learn what others are doing and how it might be leveraged for my institution.

I was shocked how our building block fit with others. It was like a LEGO set. Kudos to the DIBBS Program and Workshop.

Kept a tight schedule to ensure maximum participation and managed to keep focus on scientific objectives.

Posters and table discussions good. Talks less successful.

The requirement to send a poster PDF in December served no clear purpose.

Good opportunity to see where our project fits in infrastructure and how we might be inspired by or leverage others.

Great combination of presentations, facilitated discussion and open time.

It was good to bring together folks from different disciplines to learn from one another - that was what was keeping my attention.

Useful.

Educational.

Fun and informative.

It was good to see the critical mass emerging.

Well-done.

Nice to see what others are doing but slides were hard to follow -- too detailed for 6 minutes.

Superb organization team.

Q6 - Is there a suggestion you'd like to share about future workshops?

Keep having them.

More strongly address the interoperability aspect of the "building blocks," specify which others and how you will interoperate (e.g. protocols, interfaces, standards). How are you flexible so that others can broadly utilize your component (e.g. outside of the project funded use cases)? Be prepared for the re-occurring conversation of "avoiding the building of cyber infrastructure without community engagement and assuming the users will come." While absolutely important, this is a point that is accepted and understood by most if not all. The conversation can be moved on to the next level, e.g. how is each effort engaging with their communities while simultaneously addressing broader needs towards serving larger groups of users and working towards the sustainability of the developed infrastructure.

Make the awardees present 2 -4 min. lightening talk about their project in addition to and (well) before the poster session.

Maybe an evening reception to encourage more discussion

Since the meeting, I've been thinking a bit about the roundtable discussion groups. I really liked the chance to discuss with colleagues during the actual discussion, and was trying to figure out how the report-out could have been more interactive. Perhaps the report-out served its purpose by passing information up to NSF, but I was wondering if a brief discussion period after the report-out might prompt some useful reactions to points raised at other tables.

Some projects did not have a chance to present as a panel member.

More time for full group interaction.

Felt a bit rushed during the small group discussions. A slightly longer session might help.

It will be useful to organize a demo session to see how projects work in live action.

Distinguishing the roles of each roundtable breakout session would have helped to reduce overlap, especially between the first and second.

Give a max number of slides (3) for panelists to allow for more discussion.

It might be good to encourage possible demonstrations during the poster session if possible, or videos of the systems.

Not sure how to keep this size and structure as DIBBs grows but it will be very important to keep this to facilitate the integration between these projects that is envisioned by NSF.

Great format, longer time for talks, posters took significant time to construct.

Bring in a keynote speaker.

Focus discussions on common problems rather than generalities (e.g. "challenges"). A number of predictable themes emerged at the meeting, e.g. community engagement, moving software from prototype to production, sustainability. These issues deserve more attention than a 30 minute discussion followed by a 3 minute report.

Keep it small. Keep it real.

Make them annual events!

Maybe that we have some of the more general questions discussed perhaps the next one would focus expertise on more specific questions.

1 day.

Workshop format worked well for diverse group.

Wasn't entirely clear what you wanted out of the group discussions. Seemed like we focused on each project reporting 2-3 highlights independently (which would have been easier to have done via a simple survey) rather than really synergizing the emerging challenges and opportunities. This got a bit better as we went, but it wasn't always clear what the goal/objective of a specific set of talks or discussions actually was.

A second (or just a longer) poster session?

Give all projects opportunity to participate in panels

Have posters first. Have short research presentations around themes: how to make software, how to get people to use methods, how to deal with privacy.

Perhaps more birds-of-a-feather activities?

The activities encouraged direct participation by all attendees.

11.0 References (all websites accessed 3/24/2017)

- [1] Kurose, James. "NSF 16-126 Dear Colleague Letter: Data Infrastructure Building Blocks (DIBBs) program PI/coPI Meeting." National Science Foundation, August 29, 2017.
<https://www.nsf.gov/pubs/2016/nsf16126/nsf16126.jsp>.
- [2] Qualters, Irene. "DIBBs: Successes and Future Challenges – Summary of DIBBs funding, 2013-2016 (slide 15)." Keynote presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017. <https://dibbs17.org/report/Presentations/KeynoteQualters.pdf>.
- [3] Cornell University Center for Advanced Computing. "1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17): Final Agenda (with links to slides)." <https://dibbs17.org/report/DIBBs17Agenda.pdf>.
- [4] Cornell University Center for Advanced Computing. "DIBBs17: 1st NSF DIBBs PI Workshop." <https://dibbs17.org>.
- [5] Cornell University Center for Advanced Computing. "Posters: 1st NSF Data Infrastructure Building Blocks PI Workshop." <https://dibbs17.org/report/dibbs17posters.pdf>.
- [6] Cornell University Center for Advanced Computing. "White Papers: 1st NSF Data Infrastructure Building Blocks PI Workshop." <https://dibbs17.org/report/dibbs17whitepapers.pdf>.
- [7] Qualters, Irene. "DIBBs: Successes and Future Challenges." Keynote presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017.
<https://dibbs17.org/report/Presentations/KeynoteQualters.pdf>.
- [8] National Science Foundation Blog. "Food/Energy/Water. NSF Innovations at the Nexus of Food + Energy + Water Systems." <https://foodenergywater.wordpress.com/>.
- [9] National Science Foundation. "Understanding the Brain." https://www.nsf.gov/news/special_reports/brain/.
- [10] National Science Foundation. "10 Big Ideas for Future NSF Investments." (2016).
https://www.nsf.gov/about/congress/reports/nsf_big_ideas.pdf.
- [11] National Science Foundation Press Release 16-015. "Gravitational Waves Detected 100 Years After Einstein's Prediction," February 11, 2016. https://www.nsf.gov/news/news_summ.jsp?cntn_id=137628.
- [12] Kurose, James, James Olds, Joan Ferrini-Mundy, Barry Johnson, Rebecca Lynn Keiser, Roger Wakimoto, F. Fleming Crim, Fay Cook, and Suzanne C. Iacono. "NSF 17-031 Dear Colleague Letter: Request for Information on Future Needs for Advanced Cyberinfrastructure to Support Science and Engineering Research (NSF CI 2030)." National Science Foundation, January 5, 2017.
<https://www.nsf.gov/pubs/2017/nsf17031/nsf17031.jsp>.
- [13] "Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21)." National Science Foundation (2016). https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504730.
- [14] National Science Foundation. "Award Abstract #1639529 - INFEWS/T1: Mesoscale Data Fusion to Map and Model the U.S. Food, Energy, and Water (FEW) System." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1639529&HistoricalAwards=false.
- [15] "\$3 million grant to support first detailed map of the nation's food, energy and water systems." *Northern Arizona University News*, September 2, 2016. <http://news.nau.edu/nau-researcher-awarded-3-million-grant-build-first-detailed-map-nations-food-energy-water-systems/>.
- [16] National Science Foundation. "Award Abstract #1649880 - Computational Infrastructure for Brain Research: EAGER: BrainLab CI: Collaborative, Community Experiments with Data-Quality Controls through Continuous Integration." https://nsf.gov/awardsearch/showAward?AWD_ID=1649880&HistoricalAwards=false.
- [17] National Science Foundation. "Award Abstract #1532133 - MRI: Development of an Urban-Scale Instrument for Interdisciplinary Research." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1532133&HistoricalAwards=false.
- [18] Mitchum, Robert. "Chicago becomes first city to launch Array of Things: Urban sensing project will measure air quality, traffic, climate, and more." *UChicagoNews*, August 29, 2016.
<https://news.uchicago.edu/article/2016/08/29/chicago-becomes-first-city-launch-array-things>.

- [19] National Science Foundation. "Award Abstract #1626552 - MRI: Acquisition of a Data Lifecycle Instrument (DaLI) for Management and Sharing of Data from Instruments and Observations." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1626552.
- [20] University of Minnesota. "Polar Geospatial Center, ArcticDEM." <http://pgc.umn.edu/arcticdem>.
- [21] Marr, Bernard. "Why only one of the 5 Vs of big data really matters." *IBM Big Data Hub*, March 19, 2015. <http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters>.
- [22] Song, Carol. "GABBS: Open Source, Self-Service for Geospatial Data Exploration, Computation and Sharing." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017. <https://dibbs17.org/report/Presentations/Panel2Song.pdf>.
- [23] Purdue University. "GABBS - mygeohub." <https://mygeohub.org/groups/gabbs>.
- [24] Song, Carol, and Lan Zhao. "Geospatial Data Analysis Building Blocks (GABBS)." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1261727poster.PDF>.
- [25] Song, Carol. "CIF21 DIBBs: Integrating Geospatial Capabilities into HUBzero." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1261727paper.PDF>.
- [26] Li, Feifei. "STORM: Towards Building Spatial Temporal Online Reasoning and Management Systems." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Presentations/Panel1Li.pdf>.
- [27] University of Utah. "MesoWest Data." <http://mesowest.utah.edu/>.
- [28] University of Utah. "Hashtag Health Map." <https://hashtaghealth.github.io/countymap/map.html>.
- [29] Li, Feifei, Jeff Phillips, John Horel, and Paul Rosen. "STORM: Spatio-Temporal Online Reasoning and Management of Large Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443046poster.PDF>.
- [30] Li, Feifei. "CIF 21 DIBBs: STORM: Spatio-Temporal Online Reasoning and Management of Large Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443046paper.PDF>.
- [31] McHenry, Kenton. "DIBBS Brown Dog: A Science Driven Data Transformation Service." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017. <https://dibbs17.org/report/Presentations/Panel1McHenry.pdf>.
- [32] "PEcAN: Ecosystem science, policy, and management informed by the best available data and models." <http://pecanproject.github.io/>.
- [33] University of Illinois at Urbana-Champaign. NCSA Brown Dog. <http://browndog.ncsa.illinois.edu/>.
- [34] McHenry, Kenton, Shannon Bradley, Mike Dietze, Praveen Kumar, Jong Lee, Richard Marciano, Luigi Marini, Jerome McDonough, Barbara Minsker, Art Schmidt, and Bill Sullivan. "Brown Dog – A Science Driven Data Transformation Service." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1261582poster.PDF>.
- [35] McHenry, Kenton, Shannon Bradley, Michael Dietze, Praveen Kumar, Jong Lee, Richard Marciano, Luigi Marini, Jerome McDonough, Barbara Minsker, Art Schmidt, and Bill Sullivan. "DIBBs Brown Dog – The Need for and Challenges of a Science Driven Data Transformation Service." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1261582paper.PDF>.
- [36] Larson, Catherine. "Data Infrastructure Building Blocks (DIBBs) for Intelligence and Security Informatics Research and Community." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017. (Project team: Hsinchun Chen, Mark Patton, Ahmed Abbasi, Paul Hu, Bhavania Thurasingham, and Chris Yang). <https://dibbs17.org/report/Presentations/Panel1Larson.pdf>.
- [37] The Institute for Economics and Peace. "2015 Global Terrorism Index." <http://economicsandpeace.org/wp-content/uploads/2015/11/Global-Terrorism-Index-2015.pdf>.

- [38] Morgan, Steve. "Cyber Crime Costs Projected to Reach \$2 Trillion by 2019." *Forbes*, January 17. 2106. <http://www.forbes.com/sites/stevemorgan/2016/01/17/cyber-crime-costs-projected-to-reach-2-trillion-by-2019/#5476f4943bb0>.
- [39] IBM. "IBM i2 COPLINK." <http://www-03.ibm.com/software/products/en/coplink>.
- [40] University of Arizona. "Dark Web and GeoPolitical Web Research." <https://ai.arizona.edu/research/dark-web-geo-web>.
- [41] Dobolyi, David, and Ahmed Abbasi. "PhishMonger: A Free and Open Source Public Archive of Real-World Phishing Websites." *IEEE ISI 2016*. https://s3-us-west-2.amazonaws.com/azsecure-phishingwebsites-3/PhishMonger_Dobolyi_Abbasi_ISI-2016_preprint.pdf.
- [42] University of Arizona. "AZSecure-data.org: Intelligence and Security Informatics Data Sets." <http://www.azsecure-data.org/>.
- [43] Chen, Hsinchun, Mark Patton, Cathy Larson, Ahmed Abbasi, Paul Hu, Bhavani Thurasingham, and Chris Yang. "DIBBs for ISI." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443019poster.PDF>.
- [44] Chen, Hsinchun. "Data Infrastructure Building Blocks for Intelligence and Security Informatics (DIBBs for ISI) for Research and Community: Challenges." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443019paper.PDF>.
- [45] Cornell University Center for Advanced Computing. "Posters: 1st NSF Data Infrastructure Building Blocks PI Workshop." <https://dibbs17.org/report/dibbs17posters.pdf>.
- [46] Cornell University Center for Advanced Computing. Aristotle Cloud Federation. <https://federatedcloud.org/>.
- [47] National Science Foundation. "Award Abstract #1443037 - CIF21 DIBBs: Collaborative Research: Cyberinfrastructure for Interpreting and Archiving U-series Geochronologic Data." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1443037.
- [48] Arctur, David, W. Christopher Lenhardt, Denise Hills, Reyna Jenkyns, Kelly Stroker, Nancy Stella Todd, Emilie Dassie, and James Bowring. "Metadata, Identifiers, and Physical Samples." American Geophysical Union Fall Meeting, San Francisco, December 2016. https://www.researchgate.net/publication/311588003_PA51B-2257_Metadata_Identifiers_and_Physical_Samples.
- [49] National Science Foundation. "Award Abstract #1443085 - CIF 21 DIBBs: Porting Practical Natural Language Processing (NLP) and Machine Learning (ML) Semantics from Biomedicine to the Earth, Ice and Life Sciences." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1443085&HistoricalAwards=false.
- [50] "Clear Earth Annotation Guidelines: Sea Ice Based on Reference Ontologies (Version 4)." <http://bit.ly/2eaSVrG>.
- [51] National Science Foundation. "Abstract #1442997 - CIF21 DIBBs: An Infrastructure for Computer Aided Discovery in Geoscience." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1442997.
- [52] Li, Justin D., Cody M. Rude, David M. Blair, Michael G. Gowanlock, Thomas A. Herring, and Victor Pankratius. "Computer aided detection of transient inflation events at Alaskan volcanoes using GPS measurements from 2005-2015." *Journal of Volcanology and Geothermal Research*. Volume 327, 15 (2016): 634-642. <http://www.sciencedirect.com/science/article/pii/S0377027316303808>.
- [53] National Science Foundation. "Award Abstract #1443014 - CIF21 DIBBs: An Integrated Systems for Public/Private Access to Large-Scale, Confidential Social Science Data." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1443014.
- [54] "DATAHUB: Your data preserved and discovered." <http://datacenterhub.org/>.
- [55] Pittsburgh Supercomputing Center. "Data Exacell." <https://www.psc.edu/index.php/research/data-handling-analytics/data-exacell>.
- [56] "4Ceed: Capture, Curate, Coordinate, Correlate, and Distribute Your Data." <https://4ceed.github.io/>.
- [57] Syracuse University. "Duncan Brown: NSF DIBBS ACI 1443047." <https://dbrown10.expressions.syr.edu/?portfolio=nsf-aci-1443047>.

- [58] Enslin, Rob. "Gravitational Waves Detected 100 Years After Einstein." *Syracuse University News*, February 10, 2106. <https://news.syr.edu/2016/02/gravitational-waves-detected-100-years-after-einsteins-prediction-38878/>.
- [59] "LearnSphere: A community data infrastructure to support learning improvement online." <http://learnsphere.org/>.
- [60] National Science Foundation. "Award Abstract #1640899 - CIF21 DIBBs EI: The Local Spectroscopy Data Infrastructure (LSDI)." https://nsf.gov/awardsearch/showAward?AWD_ID=1640899&HistoricalAwards=false.
- [61] "The Materials Project." <https://materialsproject.org/>.
- [62] Fox, Geoffrey, Judy Qui, David Crandall, Gregor von Laszewski, Shantenu Jha, Madhav Marathe, John Paden, Fusheng Wang, Oliver Beckstein, and Thomas Cheatham. "Datanet: CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science. NSF14-43054 Progress Report, August 2016." http://dsc.soic.indiana.edu/publications/SPIDAL-DIBBSreport_July2016.pdf.
- [63] "SPIDAL Project." <http://www.spidal.org/index.html>.
- [64] "OSIRIS: Open Storage Research Infrastructure." <http://www.osris.org/>.
- [65] Northwestern University. "NanoMine: an Online Platform of Material Genome Nanocomposites." <http://brinson.mech.northwestern.edu/research/Nanomine.html>.
- [66] "PRP: Pacific Research Platform." <http://prp.ucsd.edu/>.
- [67] National Science Foundation. "Award Abstract #1443080 - CIF21 DIBBs: Scalable Capabilities for Spatial Data Synthesis." https://nsf.gov/awardsearch/showAward?AWD_ID=1443080.
- [68] NCSA – University of Illinois at Urbana-Champaign. "CyberGIS Center for Advanced Digital and Spatial Studies." <http://cybergis.illinois.edu/>.
- [69] "SeedMe." <https://www.seedme.org>.
- [70] Johns Hopkins University. "SciServer." <http://www.sciserver.org/>.
- [71] National Science Foundation. "Award Abstract #1443062 - Beyond Data Discovery: Shared Services for Community Metadata Improvement." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1443062.
- [72] "NSF Arctic Data Center: Data and software from NSF Arctic research." <https://arcticdata.io/>.
- [73] National Science Foundation. "Award Abstract #1443061 - CIF21 DIBBs: Systematic Data-Driven Analysis and Tools for Spatiotemporal Solar Astronomy Data." https://nsf.gov/awardsearch/showAward?AWD_ID=1443061&HistoricalAwards=false.
- [74] "Syndicate." <http://syndicatedrive.com/>.
- [75] National Science Foundation. "Abstract #1443069 - CIF21 DIBBs: An Infrastructure Supporting Collaborative Data Analytics Workflow Design and Management." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1443069.
- [76] "Tripal." <http://tripal.info/>.
- [77] National Science Foundation. "Award Abstract #1443070 - CIF21 DIBBs: User Driven Architecture for Data Discovery." https://nsf.gov/awardsearch/showAward?AWD_ID=1443070&HistoricalAwards=false.
- [78] "Whole Tale." <http://wholetale.org/>.
- [79] Raymond, Eric Steven. *The Cathedral and the Bazaar*, Version 3.0. <http://www.catb.org/esr/writings/cathedral-bazaar/cathedral-bazaar/>.
- [80] Cornell University Center for Advanced Computing. "Posters: 1st NSF Data Infrastructure Building Blocks PI Workshop." <https://dibbs17.org/report/dibbs17posters.pdf>.
- [81] McHenry, Kenton, Shannon Bradley, Mike Dietze, Praveen Kumar, Jong Lee, Richard Marciano, Luigi Marini, Jerome McDonough, Barbara Minsker, Art Schmidt, and Bill Sullivan. "Brown Dog – A Science Driven Data Transformation Service." Poster presented at presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1261582poster.PDF>.

- [82] Szalay, Alex. "SciServer: Bringing Analysis Close to the Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1261715poster.PDF>.
- [83] Scott, J. Ray, Nick Nystom, and Ralph Roskies. "Data Exacell." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1261721poster.pdf>.
- [84] Song, Carol. "Geospatial Data Analysis Building Blocks (GABBS)." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1261727poster.PDF>.
- [85] Pankratius, Victor, Philip J. Erickson, Frank D. Lind, Michael Gowanlock, Cody M. Rude, Justin D. Li, and Guillaume Rongier. "An Infrastructure for Computer Aided Discovery in Geoscience." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1442997poster.PDF>.
- [86] Nahrstedt, K., S. Konstanty, T. Nicholson, P. Nguyen, T. Spila, T. O'Brien, M. Chan, A. Schwartz-Duval, N. Aluru, P. Braun, R. Campbell, B. Cunningham, I. Gupta, K. McHenry, and J. Rogers. "4Ceed: Real-Time Data Acquisition and Analysis Framework for Material-related Cyber-Physical Environments." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443013poster.PDF>.
- [87] Reiter, Jerome. "An Integrated System for Public/Private Access to Large-scale, Confidential Social Science Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443014poster.PDF>.
- [88] Chen, Hsinchun, Mark Patton, Cathy Larson, Ahmed Abbasi, Paul Hu, Bhavani Thurasingham, and Chris Yang. "DIBBs for ISI." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443019poster.PDF>.
- [89] Pujol, Santiago, Michael McLennan, Ann Christine Catlin, Chungwook Sim, and Lisa Zilinski. "datacenterhub.org: Your data preserved and discovered." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443027poster.PDF>.
- [90] Bowring, James F., Andrea Dutton, Noah M. McLean, and Kenneth Rubin. "CIF21 DIBBs #1443037: Collaborative Research: Cyberinfrastructure for Interpreting and Archiving U-series Geochronologic Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443037poster.PDF>.
- [91] Ficklin, Stephen P., Alex Feltus, Dorrie Main, Meg Staton, Jill Wegerzyn, Sook Jung, Kuangching Wang, Ming Chen, Nate Henry, Chun-Huai Cheng, Brian Soto, Connor Wytke, Mark Clytus, Nick Mills, Nick Watts, Emily Grau, and Nic Herndon. "CIF21 DIBBs: Tripal Gateway, a Platform for Next-Generation Data Analysis and Sharing." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443040poster.PDF>.
- [92] Li, Feifei, Jeff Phillips, John Horel, and Paul Rosen. "STORM: Spatio-Temporal Online Reasoning and Management of Large Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443046poster.PDF>.
- [93] Brown, Duncan, Ewa Deelman and Jian Qin. "CIF21 DIBBs: Domain-Aware Management of Heterogeneous Workflow Active Data Management for Gravitational-Wave Science." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443047poster.PDF>.
- [94] Fox, Geoffrey. "NSF 1443054: CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443054poster.PDF>.
- [95] Angryk, Rafal, Petrus Martens, Katherine Reeves, M. Schuh, S. Hazra, B. Aydin, D. Kempton, T. Gholston, and W.G. Johnson. "CIF21 DIBBs: Systematic Data-Driven Analysis and Tools for Spatio-

- temporal Solar Astronomy Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443061poster.PDF>.
- [96] Habermann, Ted, Matthew B. Jones, Sean Gordon, Ben Leinfelder, Bryce Mecum, Peter Slaughter, and Lindsay A. Powers. "Shared Services for Community Metadata Improvement." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443062poster.PDF>.
- [97] Koedinger, Ken, John C. Stamper, Carolyn Rose, Kalyan Veeramachaneni, Una-May O'Reilly, Candace Thille, and Phil Pavlik. "LearnSphere: Data-Driven Discovery and Innovation in Education." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443068poster.PDF>.
- [98] Zhang, Jia, and Shiyong Lu. "NSF ACI-1443069: A Tool Supporting Collaborative Data Analytics Workflow Design and Management." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443069poster.PDF>.
- [99] Manepalli, Giridhar, and Allison Powell. "User Driven Architecture for Data Discovery." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443070poster.PDF>.
- [100] Wang, Shaowen, Kate Keahey and Anand Padmanabhan. "CIF21 DIBBs: Scalable Capabilities for Spatial Data Synthesis (NSF 1443080)." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443080poster.PDF>.
- [101] Qualters, Irene, and Amy Walton. "NSF architectural vision for research cyberinfrastructure." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/NSFCIarchitectureposter.pdf>.
- [102] Chourasia, Amit, and Michael Norman. "CIF21 DIBBs: Ubiquitous Access to Transient Data and Preliminary Results via the SeedMe Platform." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-1, 2017. <https://dibbs17.org/report/Posters/1443083poster.PDF>.
- [103] Jenkins, Chris, Jim Martin, Martha Palmer, Ruth Duerr, Anne Thessen, Skatje Myers, Jenette, Preciado, and Sarah Ramdeen. "ClearEarth: Preparing a Science Domain for NLP/ML, drawing on Biomedical Semantic Technologies." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443085poster.PDF>.
- [104] Lifka, David, Thomas Furlani, and Rich Wolski. "CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1541215poster.PDF>.
- [105] Peterson, Larry. "Give your Data the Edge: A Scalable Data Delivery Platform." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1541318poster.PDF>.
- [106] McKee, Shawn, Douglas Swany, Patrick Grossman, and Kenneth Merz. "OSiRIS: Distributed Ceph and Software Defined Networking for Multi-Institutional Research." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1541335poster.PDF>.
- [107] Smarr, Larry, Camille Crittenden, Tom DeFanti, Philip Papadopoulos, and Frank Wuerthwein. "PRP: Pacific Research Platform." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1541349poster.PDF>.
- [108] Ludascher, Bertram, Kyle Chard, Niall Gaffney, Matthew B. Jones, Jaroslaw Nabrzyski, Victoria Stodden, Matthew Turk, and Kandace Turner. "#1541450: CC*DNI DIBBs: Merging Science and Cyberinfrastructure Pathways: The Whole Tale." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1541450poster.PDF>.

- [109] Alter, George. "C²Metadata: Continuous Capture of Metadata." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640575poster.PDF>.
- [110] Hurwitz, Bonnie, Illyoung Choi, and John Hartman. "Ocean Cloud Commons: A Cyberinfrastructure for Microbial Ecology." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640775poster.PDF>.
- [111] Kumar, Santosh, Zachary Ives, Ida Sim, Mani Srivastava, and Timothy Hnat. "CIF21 DIBBs: EI mProv: Provenance-based Data Analytics Cyberinfrastructure for High-frequency Mobile Sensor Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640813poster.PDF>.
- [112] Talukder, Ashit, W. Dou, Y. Tao, Y. Zhu, G. Djorgovski, A. Mahabal, D. Crichton, W. Tolone, M. Hadzikadic, E. El-Shaer, Y. Wang, and W. Zadrozny. "VIFI: Virtual Information-Fabric Infrastructure for Data-Driven Decisions from Distributed Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640818poster.PDF>.
- [113] Cuff, James, Scott Yockel, John Goodhue, Saul Youssef, Rajiv Shridhar, Ralph Zottola, Chris Hill, and Glenn Bresnahan. "NESE: The North East Storage Exchange." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640831poster.PDF>.
- [114] Parashar, Manish. "Virtual Data Collaboratory (VDC): A Regional Cyberinfrastructure for Collaborative Data Intensive Science." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640834poster.PDF>.
- [115] Schadler, Linda, L. Catherine Brinson, Wei Chen, and Deborah L. McGuinness. "Ontology-enabled Polymer Nanocomposite Open Community Data Resource." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640840poster.PDF>.
- [116] Freire, Juliana, Oliver Kennedy, and Boris Glavic. "Streamlining and Understanding Curation with Vizier." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640864poster.PDF>.
- [117] Govindaraju, Venu, Krishna Rajan, Thomas Furlani, Srirangaraj Setlur, and Scott Broderick. "CIF21 DIBBs: EI: Materials Data Engineering (MaDE) Laboratory." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640867poster.PDF>.
- [118] Persson, Kristin. "The Local Spectroscopy Data Infrastructure." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.
<https://dibbs17.org/report/Posters/1640899poster.PDF>.
- [119] Furlani, Thomas, "Aristotle Cloud Federation: Building a Federated Cloud Model." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017.
<https://dibbs17.org/report/Presentations/Panel2Furlani.pdf>.
- [120] Aristotle Cloud Federation. "Emerging Technologies: DrAFTS."
<https://federatedcloud.org/using/drafts.php>.
- [121] Wolski, Rich, John Brevik, Ryan Chard, and Kyle Chard. "Probabilistic Guarantees of Execution Duration for Amazon Spot Instances." University of California Technical Report Number 2016-05, July 1, 2016.
<https://www.cs.ucsb.edu/sites/cs.ucsb.edu/files/docs/reports/master.pdf>.
- [122] Cornell University Center for Advanced Computing. Aristotle Cloud Federation.
<https://federatedcloud.org>.
- [123] Lifka, David, Thomas Furlani, and Rich Wolski. "CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation." Poster presented at

the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Posters/1541215poster.PDF>

[124] Lifka, David, Thomas Furlani, and Rich Wolski. "CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1541215paper.PDF>.

[125] Fox, Geoffrey C. "NSF 1443054: CIF21 DIBBS: Middleware and High Performance Analytics Libraries for Scalable Data Science." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017.

<https://dibbs17.org/report/Presentations/Panel2Fox.pdf>.

[126] Fox, Geoffrey C. "NSF 1443054: CIF21 DIBBS: Middleware and High Performance Analytics Libraries for Scalable Data Science." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443054poster.PDF>.

[127] Fox, Geoffrey C. "NSF 1443054: CIF21 DIBBS: Middleware and High Performance Analytics Libraries for Scalable Data Science." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443054paper.PDF>.

[128] Fox, Geoffrey, Judy Qui, David Crandall, Gregor von Laszewski, Shantenu Jha, Madhav Marathe, John Paden, Fusheng Wang, Oliver Beckstein, and Thomas Cheatham. "Datanet: CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science. NSF14-43054 Progress Report, August 2016. http://dsc.soic.indiana.edu/publications/SPIDAL-DIBBSreport_July2016.pdf.

[129] NIST. Big Data Public Working Group Use Case Survey. "High Performance Computing Enhanced Apache Big Data Stack." <http://hpc-abds.org/kaleidoscope/>.

[130] Nahrstedt, Klara. "T2C2: 4CeeD DIBBs Challenges and Solutions." Panel presentation at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017.

<https://dibbs17.org/report/Presentations/Panel2Nahrstedt.pdf>.

[131] "4CeeD: Capture, Curate, Coordinate, Correlate, and Distributed Your Data."

<https://4ceed.github.io/index.html>.

[132] Nahrstedt, K., S. Konstanty, T. Nicholson, P. Nguyen, T. Spila, T. O'Brien, M. Chan., A. Schwartz-Duval, N. Aluru, P. Braun, R. Campbell, B. Cunningham, I. Gupta, K. McHenry, and J. Rogers. "4CeeD: Real-Time Data Acquisition and Analysis Framework for Material-related Cyber-Physical Environments." Poster presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443013poster.PDF>.

[133] Nahrstedt, K., S. Konstanty, T. Nicholson, P. Nguyen, T. Spila, T. O'Brien, M. Chan., A. Schwartz-Duval, N. Aluru, P. Braun, R. Campbell, B. Cunningham, I. Gupta, K. McHenry, and J. Rogers. "4CeeD: Real-Time Data Acquisition and Analysis Framework for Material-related Cyber-Physical Environments." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443013paper.PDF>.

[134] Szalay, Alex. "SciServer - Long Term Access to Large Scientific Data Sets: the SkyServer and Beyond." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11, 2017. <https://dibbs17.org/report/Presentations/Panel2Szalay.pdf>.

[135] "SciServer." <http://www.sciserver.org/>.

[136] Szalay, Alexander S. "SciServer: Bringing Analysis Close to the Data." Poster presented at the 1st NSF Data Infrastructure Building Block PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Posters/1261715poster.PDF>.

[137] Szalay, Alexander S. "ACI-1261715: Long Term Access to Large Scale Scientific Data Sets: The SkyServer and Beyond." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1261715paper.PDF>.

[138] McHenry, Kenton, Shannon Bradley, Michael Dietze, Praveen Kumar, Jong Lee, Richard Marciano, Luigi Marini, Jerome McDonough, Barbara Minsker, Art Schmidt, and Bill Sullivan. "DIBBs Brown Dog – The Need for and Challenges of a Science Driven Data Transformation Service." Paper

presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1261582paper.PDF>.

[139] Szalay, Alexander S. "ACI-1261715: Long Term Access to Large Scientific Data Sets: The SkyServer and Beyond." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1261715paper.PDF>.

[140] Scott, J. Ray, Nick Nystrom and Ralph Roskies. "The Data Exacell." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1261721paper.PDF>.

[141] Song, Carol. "CIF DIBBs: Integrating Geospatial Capabilities into HUBzero." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1261727paper.PDF>.

[142] Pankratius, Victor, Philip J. Erickson, Frank D. Lund, Michael Gowanlock, Code M. Rude, Justin D. Li, and Guillaume Rongier. "An Infrastructure for Computer Aided Discovery in Geoscience." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1442997paper.PDF>

[143] Nahstedt, K., S. Konstanty, T. Nicholson, P. Ngyuen, T. Spila, T. O'Brien, M. Chan, A. Schwartz-Duval, N. Aluru, P. Braun, R. Campbell, B. Cunningham, I. Gupta, K. McHenry, and J. Rogers. "4CeeD: Real-Time Data Acquisition and Analysis Framework for Material-related Cyber-Physical Environments." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443013paper.PDF>.

[144] Reiter, Jerry. "ACI-14-43014: An Integrated System for Providing Access to Large-Scale, Confidential Social Science Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443014paper.PDF>.

[145] Chen, Hsinchun. "Data Infrastructure Building Blocks for Intelligence and Security Informatics (DIBBs for ISI) for Research and Community: Challenges." Paper presented at 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1443019paper.PDF>.

[146] Pujol, Santiago, Ann Christine Catlin, Michael McLennan, Chungwook Sim, and Lisa Zilinski. "CIF21 DIBBs: Building a Modular Cyber-Platform for Systematic Collection, Curation, and Preservation of Large Engineering and Science Data – A Pilot Demonstration Project." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1443027paper.PDF>.

[147] Bowring, James, Andrea Dutton, Noah M. McLean, and Kenneth Rubin. "CIF21 DIBBs #1443037 Collaborative Research: Cyberinfrastructure for Interpreting and Archiving U-series Geochronologic Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443037paper.PDF>.

[148] Ficklin, Stephen, Jill Wegrzyn, Frank Feltus, Margaret Staton, Doreen Main, Sook Jung, and Kuangching Wang. "Challenges and Future Directions for CIF21 DIBBs: Tripal Gateway, a Platform for Next-Generation Data Analysis and Sharing (Award #1443040)." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1443040paper.PDF>.

[149] Li, Feifei. "CIF21 DIBBs: STORM: Spatio-Temporal Online Reasoning and Management of Large Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443046paper.PDF>.

[150] Brown, Duncan, Ewa Deelman, and Jian Qin. "CIF21 DIBBs: Domain-Aware Management of Heterogeneous Workflows Active Data Management for Gravitational-Wave Science." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1443047paper.PDF>.

[151] Fox, Geoffrey C. "NSF 1443054: CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443054paper.PDF>.

- [152] Angryk, Rafal, Petrus Martens, and Katherine Reeves. "CIF21 DIBBs: Systematic Data-Driven Analysis and Tools for Spatiotemporal Solar Astronomy Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443061paper.PDF>.
- [153] Habermann, Ray, and Matt Jones. "CIF21 DIBBs: Beyond Data Discovery: Shared Services for Community Metadata Improvement." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443062paper.PDF>.
- [154] Koedinger, Ken, Kalyan Veeramachaneni, Candace Thille, Phil Pavlik, Una-May O'Reilly, Carolyn Rose, and John Stamper. "Infrastructure for Data-Driven Innovation in Education: Challenges and Future Directions." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443068paper.PDF>.
- [155] Zhang, Jia, and Shiyong Lu. "Collaborative Scientific Workflow Composition as a Service." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443069paper.PDF>.
- [156] Manepalli, Giridhar, and Allison Powell. "User Driven Architecture for Data Discovery." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443070paper.PDF>.
- [157] Wang, Shaowen, Kate Keahey, and Anand Padmanabhan. "CIF21 DIBBs: Scalable Capabilities for Spatial Data Synthesis." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443080paper.PDF>.
- [158] Chourasia, Amit, and Michael Norman. "CIF21 DIBBs: Ubiquitous Access to Transient Data and Preliminary Results via the SeedMe Platform." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443083paper.PDF>.
- [159] Jenkins, Chris, Jim Martin, Martha Palmer, Skatje Myers, Ruth Duerr, Sarah Ramdeen, Anne Thessen, and Jenette Preciado. "CIF21 DIBBs: Porting Practical Natural Language Processing (NLP) and Machine Learning (ML) Semantics from Biomedicine to the Earth, Ice and Life Sciences." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443085paper.PDF>.
- [160] Lifka, David, Thomas Furlani, and Rich Wolski. "CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1541215paper.PDF>.
- [161] Peterson, Larry. "Give Your Data the Edge: A Scalable Data Delivery Platform." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1541318paper.PDF>.
- [162] McKee, Shawn. "CC*DNI DIBBs: Multi-Institutional Open Storage Research Infrastructure (MI-OSIRIS)." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1541335paper.PDF>.
- [163] Smarr, Larry, Camille Crittenden, Tom DeFanti, Phil Papadopoulos, and Frank Wuerthwein. "Pacific Research Platform (PRP)." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1541349paper.PDF>.
- [164] Ludascher, B., K. Chard, N. Gaffney, M. Jones, J. Nabrzyski, V. Stodden, M. Turk, and K. Turner. 2017. "#151450: CC*DNI DIBBs: Merging Science and Cyberinfrastructure Pathways: The Whole Tale." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1541450paper.PDF>.
- [165] Alter, George. "Continuous Capture of Metadata for Statistical Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1640575paper.PDF>.
- [166] Hurwitz, Bonnie, and John Hartman. "CIF 21 DIBBs: PD: Accelerating Comparative Metagenomics through an Ocean Cloud Commons." Paper presented at the 1st NSF Data Infrastructure

Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1640775paper.PDF>.

[167] Kumar, Santosh, Zachary Ives, Ida Sim, and Mani Srivastava. "mProv: Provenance-based Data Analytics Cyberinfrastructure for High-frequency Mobile Sensor Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1640813paper.PDF>.

[168] Talukder, Ashit. "CIF21 DIBBS: EI; VIFI: Virtual Information-Fabric Infrastructure for Data-Driven Decisions from Distributed Data: Challenges and Risk Mitigation Strategy." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1640818paper.PDF>.

[169] Cuff, James, John Goodhue, Saul Youssef, Rajiv Shridhar, Ralph Zottola, Chris Hill, and Glenn Bresnahan. "NESE: The North East Storage Exchange." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1640831paper.PDF>.

[170] Parashar, Manish, Vasant Honavar, and the VDC Team. "Virtual Data Collaboratory (VDC): A Regional Cyberinfrastructure for Collaborative Data Intensive Science." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1640834paper.PDF>.

[171] Schadler, Linda, Deborah L. McGuinness, Cate Brinson, and Wei Chen. "CIF21 DIBBS: PD: Ontology-enabled Polymer Nanocomposite Open Community Data Resource." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1640840paper.PDF>.

[172] Freire, J., O. Kennedy, and B. Glavic. "Streamlining and Understanding Curation with Vizier." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1640864paper.PDF>.

[173] Govindaraju, Venu, Krishna Rajan, Thomas Furlani, Srirangaraj Setlur, and Scott Broderick. "CIF21 DIBBS: EI: Data Laboratory for Materials Engineering (Award #1640867)." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Papers/1640867paper.PDF>.

[174] Persson, Kristin. "NSF DIBBs Award 1640899: The Local Spectroscopy Data Infrastructure." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1640899paper.PDF>.

[175] ELSEVIER. "SoftwareX." <https://www.journals.elsevier.com/softwarex/>.

[176] Hurwitz, Bonnie. "The Ocean Cloud Commons." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 12, 2017.

<https://dibbs17.org/report/Presentations/Panel3Hurwitz.pdf>.

[177] Hurwitz, Bonnie, Illyoung Choi, and John Hartman. 2017. "Ocean Cloud Commons: A Cyberinfrastructure for Microbial Ecology." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017.

<https://dibbs17.org/report/Posters/1640775poster.PDF>.

[178] Hurwitz, Bonnie, and John Hartman. "CIF21 DIBBS: PD: Accelerating Comparative Metagenomics through an Ocean Cloud Commons." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1640775paper.PDF>.

[179] Kumar, Santosh. "Provenance-based Data Analytics CI for High-frequency Mobile Sensor Data (mProv)." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 12, 2017. <https://dibbs17.org/report/Presentations/Panel3Kumar.pdf>.

[180] National Institutes of Health. "MD2K: Center of Excellence for Mobile Sensor Data-to-Knowledge." <https://md2k.org/>.

[181] "mProv." <http://mprov.md2k.org/>.

[182] Kumar, Santosh, Zachary Ives, Ida Sim, Mani Srivastava, and Timothy Hnat. "CIF21 DIBBS: EI mProv: Provenance-based Data Analytics Cyberinfrastructure for High-frequency Mobile Sensor Data."

- Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1640813poster.PDF>.
- [183] Kumar, Santosh, Zachary Ives, Ida Sim, and Mani Srivastava. "mProv: Provenance-based Data Analytics Cyberinfrastructure for High-frequency Mobile Sensor Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1640813paper.PDF>.
- [184] Reiter, Jerry. "An Integrated System for Providing Access to Large-Scale, Confidential Social Science Data." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 12, 2017. <https://dibbs17.org/report/Presentations/Panel3Reiter.pdf>.
- [185] Reiter, Jerome. "An Integrated System for Providing Access to Large-Scale, Confidential Social Science Data." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Posters/1443014poster.PDF>.
- [186] Reiter, Jerry. "ACI-14-43014: An Integrated System for Providing Access to Large-scale, Confidential Social Science Data." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443014paper.PDF>.
- [187] Koedinger, Ken. "LearnSphere: Infrastructure for Data-Driven Discovery & Innovation in Education." Panel presentation at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 12, 2017. <https://dibbs17.org/report/Presentations/Panel3Koedinger.pdf>.
- [188] "LearnSphere." <http://learnsphere.org/>.
- [189] Koedinger, Ken, John C. Stamper, Carolyn Rose, Kalyan Veeramachaneni, Una-May O'Reilly, Candace Thille, and Phil Pavlik. "LearnSphere: Data-Driven Discovery and Innovation in Education." Poster presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, and January 11-12, 2017. <https://dibbs17.org/report/Posters/1443068poster.PDF>.
- [190] Koedinger, Ken, Kalyan Veeramachaneni, Candace Thille, Phil Pavlik, Una-May O'Reilly, Carolyn Rose, and John Stamper. "Infrastructure for Data-Driven Innovation in Education: Challengers and Future Directions." Paper presented at the 1st NSF Data Infrastructure Building Blocks PI Workshop, Arlington, January 11-12, 2017. <https://dibbs17.org/report/Papers/1443068paper.PDF>.
- [191] Grama, Joanna, and Valerie Vogel. "The 2016 Top 3 Strategic Information Security Issues." EDUCAUSE, January 11, 2016. <http://er.educause.edu/articles/2016/1/the-2016-top-3-strategic-information-security-issues>.
- [192] Duke University. "SSRI Protected Research Data Network." <https://ssri.duke.edu/data-it-services/protected-data-support-prdn/ssri-protected-research-data-network-prdn>.
- [193] Naveed, Muhammad, Erman Ayday, Ellen W. Clayton, Jacques Fellay, Carl A. Gunter, Jean-Pierre Hubaux, Bradley A. Malin, and XiaoFeng Wang. "Privacy in the Genomic Era." arXiv.org 1405.1891v3, June 17, 2015. <https://arxiv.org/pdf/1405.1891.pdf>.
- [194] Elias, Andy Rosales, Nevena Golubovic, Chandra Krintz, and Rich Wolski. "Where's the Bear? – Automating Wildlife Image Processing Using IoT and Edge Cloud Systems." University of California, Santa Barbara Tech Report 2016-07, October 12, 2016. <https://www.cs.ucsb.edu/sites/cs.ucsb.edu/files/docs/reports/tr.pdf>.
- [195] Rashid, Fahmida Y. "The dirty dozen: 12 cloud security threats." *InfoWorld*, March 11, 2016. <http://www.infoworld.com/article/3041078/security/the-dirty-dozen-12-cloud-security-threats.html>.
- [196] "CyVerse." <http://www.cyverse.org/>.
- [197] NASA. "Technology Readiness Level." https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html.