Ad-Hoc Task Force on Big Data
of the
NASA Advisory Council Science Committee

Meeting Minutes

March 6-7, 2017

Charles P. Holmes

Charles P. Holmes, Chair

Gerald S. Smith, Executive Secretary

Report prepared by Joan M. Zimmermann
Ingenicomm, Inc.

Table of Contents

March 6, 2017

Opening remarks, Introduction

Mr. Gerald Smith, Executive Secretary of the NASA Advisory Council (NAC) Ad-Hoc Task Force on Big Data (BDTF), called the fourth meeting of the Task Force to order and made general announcements.

Mr. Smith introduced Dr. Charles Holmes, Chair of the BDTF. Members introduced themselves around the table. Dr. Holmes welcomed newly pending BDTF member, Dr. Chris Mentzel, reviewed the agenda, and introduced Dr. Thomas Zurbuchen, Associate Administrator (AA) of the Science Mission Directorate.

SMD AA comments

Dr. Zurbuchen offered remarks to the BDTF, and described his previous academic foci and technology programs at the University of Michigan. He presented two anecdotes that illustrated the utility of advanced data analytic methods, one of which concerned the treatment of medical data, the growth of which outperforms Moore's law and is straining the current capabilities of computing. Three-dimensional image cubes are data-intensive, and large hospitals have begun to apply more up-to-date methods to deal with the increased data volume. The other anecdote concerned the "microscopic" view in genomics. He related the story of a researcher who became famous for discovering the "fat gene," whose base-pair sequence was teased out through the use high-end computing (HEC) resources, the use of which ultimately cost less than .1% of the cost of conventional computing.

Data and data needs are going up exponentially, and trades must be made between processes and safety, and speed. Dr. Zurbuchen pointed out that the gene research innovator's HEC system was much less protected than the university's system, a potential drawback for a Federal agency. Dr. Zurbuchen expressed great interest in the BDTF's work, as NASA will need more sophisticated analytic processes, and advances in deep learning, pattern recognition, etc. The most exciting statistic in the last two years is that 40% of the papers from the Hubble Space Telescope (HST) were originating in novel analyses of old data. How do we stand on this wave as it pushes us forward, balancing the variety of values we have? NASA needs data security but also wants to benefit from the commercial pushes in computing. He encouraged BDTF to look closely at the problem, recognizing that one can't solve the problem through the efforts of one full-time employee (FTE). How does the community want to approach data, modeling, simulation, etc.? He recommended that BDTF recruit the efforts of an individual at the University of California at Berkeley, who is an economist by training.

Dr. Holmes commented, saying that since the October meeting of the BDTF he had had the opportunity to attend several workshops on the general topic of data science. He noticed that NASA was underrepresented, even at the Space Science Telescope Institute (STScI) in Baltimore. NASA needs to get funding out to researchers to work these data problems. Dr. Zurbuchen responded by noting the importance of defining the status quo

very quantitatively: what is the state of affairs in data and modeling, what are the standards, and what is the desired state and the value of it? While anecdotes imply that there are technologies NASA is not using, it is still necessary to nail down the problem beyond a few anecdotes. "Whys" are usually better than the "hows." Where is funding being wasted? What are the tools we can use? ROSES calls? The Earth Science Division (ESD) did an experiment with some of these tools and found a lot of pent-up demand for new technologies. The real needs must be rooted in reality. Dr. Holmes said he had heard excellent talks on data analysis techniques for the Palomar Transient Survey, and another astrophysics (AP) project. Dr Holmes noted that the researchers indicated they had been repeatedly rejected for NASA grants because their data science proposals didn't fit neatly in a category. ROSES should be adjusted to reflect this need for new disciplinary categories. Dr. Raymond Walker noted that while NASA was a leader in promoting an open data policy, the modeling and simulation community remains greatly reluctant to deliver results out to the broader community of users; NASA needs to educate community on opening this up. Dr. Zurbuchen indicated he had already asked the National Academies to begin a study on disbursing results from models and simulations, and offered a positive view on the cross-utility of data (AP, Planetary) and how it has opened up new science areas. That said, the science should be verifiable, independently. If a Voyager data set is not open, it is not verifiable. Modeling is subject to the reasoning: it needs to be transparent so that everyone can offer solutions. He agreed completely with Dr. Walker's assessment.

Dr. Eric Feigelson felt that NASA needs more expertise in the sociological issues surrounding data, such as the link between instrument and science; how do we pull scientific insights from data? Where is the expertise in statistics and informatics? The issue is not just about crunching numbers; ROSES must emphasize methodology more explicitly, aside from the obvious hardware and science requirements. Dr. Zurbuchen noted that ROSES stands and falls on review, and relates to the knowledge the field has in its area. In the end, the selections depend on the group of peer reviewers. The most conservative group at NASA is that of the science peer reviewer. Taking it to the next level: how is NASA addressing new learning in statistics and informatics? Dr. Zurbuchen relating having had to stop a review to teach a methodology class to the reviewers; it may be worthwhile to talk to reviewers about this very issue. Dr. Reta Beebe cited the ease with which a panel can be selected and steered to produce a desired result. She felt that people at Headquarters tend to get isolated, and often it is necessary to assist the person who is organizing the panel. Access to data is a problem. She felt that Headquarters should put more stress on documentation. On Cassini, for instance, Dr. Beebe described obtaining funds for team members to write user guides. Without these efforts, many instruments typically never become fully mature. She cited a complex ultraviolet spectrometer on the Mars Atmosphere and Volatile Evolution (MAVEN) orbiter that is still being calibrated, three years into the mission. Management should pay attention to see how they support this sort of thing. Dr. Zurbuchen agreed that missions should hire statisticians, as well as data user groups, to help on instrumentation, and to document the results.

Dr. Holmes felt that some top-down pressure to enforce improvement of data production and documentation would be useful. Dr. James Kinter commented on Mission to Planet Earth, wherein a sizable budget was put into the data information system (DIS), and which resulted in a large and unwieldy DIS. Dr. Beebe reported having had the same issue with the Cassini mission, which was affected by budget cuts, resulting primarily in the delivery of raw data. Some of the teams can deliver reduced data and derived products, but not all of the instruments are delivering calibrated data. Every time NASA cuts the budget, it also has to ask about what's happening to the data. Dr. Zurbuchen noted that stable data archives lock in technology, both good and bad. There is a lot of evidence that in AP the archives are flexible enough to lock in, but there also have to be iterations to accommodate technology evolution. NASA has to wrestle with the issues. He looked forward to receiving BDTF's final report.

<u>DOE Exascale Computing Project (ECP)</u>

Dr. Paul Messina, Project Director of the Department of Energy's (DOE) Exascale Computing Project (ECP), provided an update, beginning with some project history. DOE and other Federal agencies started looking at exascale computing in 2007. President Obama issued an Executive Order in 2015 to establish a capable exascale computing system that integrates hardware and software capability to deliver 50 times more performance than today's 20 petaflops (PF) machines, for mission-critical applications. The ECP is a 7-year project that will aim to accelerate a capable system, led by DOE's national laboratories, working with academia and industry, and looking for sustained performance in applications. Project goals are to develop applications; support national security through stockpile stewardship; deal with cybersecurity issues; contribute to economic competitiveness; collaborate with vendors to develop a software stack and partner with vendors to develop architectures to support exascale applications; and train the next-generation workforce. The ECP is a national mandate intended to serve agencies well beyond the DOE national labs. ECP is a day-to-day collaboration among six labs: Sandia, Argonne, Los Alamos, Lawrence Berkeley, Oak Ridge, and Lawrence Livermore National Laboratories (SNL, ANL, LANL, LBNL. ORNL, LLNL). ECP is a distributed effort: the project office is located at ORNL, while the deputy project director is at LANL. An initial exascale system is scheduled to be delivered in 2021, followed by the delivery capable exascale systems in 2022, and deployment in 2023. Acquisition of the systems will be carried out by the same base facility programs that are in place for DOE today.

Four key challenges for the development of an exascale system are: parallelism; memory and storage; reliability; and energy consumption. A capable exascale system must deliver 50 times the operations of today's petaflop systems; stay within a power envelope of 20-30MW; be sufficiently resilient (i.e. a perceived fault rate of less than one per week); and include a software stack that supports broad applications and workloads. The holistic co-design approach combines application development, a scalable and productive software stack, hardware technology designs, and the integration of exascale supercomputers. ECP had its first meeting in February 2017, with a total of 450 participants. The teams met in 10 to 15 concurrent sessions and identified milestones for joint production.

Exascale applications will address 25 national challenges, including nuclear energy, climate, chemical science (biofuel catalysts), wind energy, and combustion. Three applications are in weapons production and are classified. Materials science, nuclear physics, nuclear materials, accelerator physics, magnetic fusion energy, cosmology, geoscience, and precision medicine for cancer are particular challenges, which involve large data volumes and deep learning techniques. Seismology, carbon capture and storage, urban systems science (crime rates, air pollution), metagenomics, astrophysics, and power grid applications round out the remaining applications.

ECP will build a comprehensive and coherent software stack, not one-size-fits-all, that will enable developers to write highly parallel applications that can portably target diverse exascale architectures. ECP is based on a conceptual software stack that ranges from resilience settings through to workflows, flowing from left to right based on hardware interfaces The current set of ST (software technology) projects is mapped to the software stack, with current ECP data-intensive applications that include machine learning for improving cancer detection.

Next steps in developing the software stack will include a gap analysis. Based on the results of the analysis, DOE will issue requests for information and proposals (RFIs and RFPs). For the hardware technology overview, DOE issued an RFP to hardware vendors and received 14 proposals. DOE has since funded 6 well-known companies to do research and development (R&D) on new and better architectures. $280M for a 3-year contract was distributed among the 6 companies. Vendors will be working on new hardware that will be part of their new product lines, and will feature better energy consumption, more efficient I/O, and multistack memories. DOE will then be purchasing small testbed systems, and follow on with nonrecurring engineering contracts for both hardware and software. The anticipated timeline will see the first few years taken up in R&D, with targeted development starting in 2019; and systems delivered, deployed and accepted by 2022 and 2023. ECP has a total of 22 agency partners, including the national labs, National Institute of Standards and Technology (NIST) and the National Institutes of Health (NIH), 9 private sector partners, 39 university research partners, and 18 industry council members. End users include companies such as United Technologies, Eli Lilly, General Motors, General Electric, AMSYS, and Cascade Industries.

Dr. Holmes asked if there were anything BDTF should advise NASA to do with regard to ECP. Dr. Messina welcomed the pursuit of collaboration and partnership with NASA on applications; he reported having visited Langley Research Center, and was intending to consult with Headquarters and explore how to partner with the agency. For mission-critical NASA applications that require exascale computing, NASA could partner with DOE, test preliminary software, etc., in mutually beneficial ways. Secondly, DOE should understand more about NASA's data-oriented needs. Dr. Mentzel commented that the data intensive applications require hardware with different design parameters meant to handle data-intensive problem; are these designs already set? Dr. Messina said there are currently no contracts for systems in place, and there will not be for several years. Within each contract for the 6 recently selected vendors, there are several tasks that are

oriented to data issues. The task is to figure out what new designs will help with data-intensive applications, deep learning, etc., and is just getting started. Hardware designs will need to be based on real requirements, which will require more information to determine. DOE's learning about NASA's use cases will be helpful here in the early stages. NASA's HEC chief, Dr. Tsengdar Lee, is already scheduled to be at ANL, to be present at these discussions and to get a look at the research DOE is starting to fund.

Dr. Messina cautioned that funding for ECP under the new administration is currently not known, thus plans are to cast the net as widely as practical and determine needs. DOE hopes to issue an RFP for 5-6 specific applications. Dr. Clayton Tino commented that missions at NASA are typically long-lived, and tend to get hit by funding cuts and technology obsolescence. He was curious as to how DOE will approach exascale in the long term. Dr. Messina replied that nuclear weapons and science applications at DOE have long histories, but issues are typically not as severe as NASA's issues with computers out in spaceflight. Performance portability should be more feasible, DOE recognizes. DOE generally addresses it project by project, but the facilities are usually funded at an adequate level to support these portable approaches. Quantum computing, and other topics are being studied at testbeds at some of the labs. Some data needs to be kept indefinitely, however, and national labs are not particularly oriented to this. Dr. Tino said he had observed that at NASA, hardware changes over time, and was curious about whether DOE is addressing updated and changing applications. Dr. Messina agreed that it's essential to be able to update applications over time. As to funding at DOE, the cost range for the 7-year ECP project is $3.5-5B, excluding the procurement of the exascale system.

Dr. Walker noted that some big data is really big computation, such as for modeling for predicting space weather, which is particularly suitable for exascale. In testing of high-speed networks on the West coast, researchers have discovered that the bottlenecks are getting the data into and out of the supercomputers. Dr. Messina agreed completely, commented that this is why workflow is important. Dr. Kinter asked if there were any efforts to co-design; e.g., by manipulating chip design at the same time as software design. Dr. Messina felt that by starting now, vendors can explore changes to chip designs that are favorable to exascale, which they are likely to put into a product. The goal is not to have one-off results; DOE wants to have these products generally available. One example is IBM's Blue Gene family of computers, originally developed for specialized biological and electrodynamic applications; that design is now more generally applicable. ECP is taking the same approach. Dr. Lee added a few comments on his planned observations at ANL; community members in Earth Science are trying to influence the co-design process, having identified a number of bottlenecks that can be ameliorated by design. Dr. Holmes thought ECP would be a likely topic for a recommendation.

<u>NASA Big Data Working Group</u>

Mr. John Sprague, Deputy for the Technology & Innovation (T&I) Division, updated the activities of NASA"s Big Data Working Group, which he co-chairs with Dr. Lee. Mr. Sprague described T&I's purview as running the information architecture, and focused

on such issues as technology infusion (Internet of Things). The BDWG holds meetings every month and has about 85 members, and is growing continually. Notes for the group are kept on an internal website. BDWG has one or two Big Data face-to-face work sessions (Big Think meetings) per year, at rotating NASA centers. The next meeting will be in April at the Jet Propulsion Laboratory (JPL). Mr. Sprague suggested BDTF examine some relevant online training courses, such as MIT's *Tackling the Challenges of Big Data*, considered as the gold standard.

A typical BDWG agenda includes information on upcoming NASA, national, and world Big Data events, the latest hot projects and demonstrations, resource-sharing, "around the horn" reports from centers and mission directorates, and discussions of future topics. Rarely, the group will bring in guest speakers. Mr. Sprague announced new roles in Big Data at the Office of the Chief Information Office (OCIO), including a new Networking and Information Technology Research and Development (NITRD) subcommittee member, deputy CIO Terry Jackson. NASA now has a Data Scientist in the OCIO, and others have been hired at the agency. A data scientist is typically hired to solve specific issues, such as mapping the NASA data universe, and other *ad hoc* projects. There are now two contractors at Johnson Space Center in support of such efforts. Big data administration is being carried out throughout NASA. Other roles, such as Chief Data Officer (CDO) and Chief Knowledge Officer (CKO), are being filled. NASA currently has a CKO and is looking at CDO, the closest approximation of which is Mr. Sprague's current position. There are now Big Data stewards (points of contact for data), and Data Evangelists who work on projects with data scientists. There is a Data Analytics Lab and an Internet of Things (IoT) Lab located at the Johnson Space Center. BDWG is also working on data security. A Phase 2 white paper on IoT will be coming out the second week of March and will be presented to the CIOs of all the centers. A Phase 1 paper was created, but it was shared only with other Federal agencies, as it dealt with security issues. Mr. Sprague promised to circulate the Phase 2 paper to BDTF.

Dr. Lee commented briefly on the High Performance Computing Act, amended this year by Congress, which has broadened HEC definition to include large-scale data analytics, which allows NASA to embrace it more on a budgetary level.

Mr. Sprague described a just-completed phase 2 project on extravehicular activity (EVA) data integration, which has produced a state-of-the-art implementation tool that is being expanded for use with the Orion capsule. A Quantum Artificial Intelligence Lab (QuAIL) at Ames Research Center is working on special topics. The Langley Research Center has been doing some excellent work as well in areas such as space radiation and carbon nanotubes; NASA groups are also synthesizing data using Watson Content Analytics Technology. Dr. Kinter asked if there was any intention to broaden BDWG to bring in NASA grantees, as there is a large intramural group available at the universities willing to participate. Mr. Sprague responded that BDWG has been talking about bringing in communities of interest, and that he would bring Dr. Kinter's comments to the attention of the group. Dr. Feigelson asked if Mr. Sprague had any thoughts on legacy information systems. Mr. Sprague felt there was no reason to get rid of these systems, as they undergo frequent upgrades and modifications. Dr. Lee thought the

current archives were well in hand. He added that another issue in the Big Data era is that we don't know what kind of questions we are going to ask in the future. Astrophysics data, e.g., is organized in wavelengths— where do you extract the information? That is a bit of a struggle. Every discipline has similar challenges. Dr. Lee felt that moving forward, the science community must consider how it wants to organize its data sets.

Update on NASA High End Computing (HEC) and Survey of Federal HEC Capabilities
Dr. Holmes prefaced Dr. Lee's briefing by describing the origin of the survey, which was precipitated by a Heliophysics Subcommittee concern, expressed to Science Committee, that researchers are not getting the supercomputing cycles they need at NASA and HEC centers. Subsequently, a limited survey of HEC capabilities across the federal government was carried out. Dr. Jill Dahlburg, Chair of that committee, called in to ask questions.

Dr. Tsengdar Lee presented first on the upgrade of HEC capabilities at NASA. The Modular Supercomputing Facility (MSF) at Ames Research Center is the current solution to expanding HEC space at NASA. The MSF uses adiabatic computing modules that don't require an active cooling system (a scheme that is effective in dry climates). The first MSF system, Electra, was operational as of 1 January. It is fenced, fully protected and monitored. Electra has a 1.2PF capability, and provides 4652 system building units (SBUs) per hour. Ames will build a second unit on the same pad this year, pending Congressional approval.

MSF's initial test results (LINPACK- 1.096 PF; HPCG- 25.188 TF) are considered very good for such a small system. Power use efficiency is 1.03, greatly exceeding the original 1.06 target. HEC growth, based on measurements from the last calendar year, show a growth of cores by 32%, increases in TF by 58%, and increase in SBUs by 42%. NASA continues to deal with capacity oversubscription in SMD. This year, the Astrophysics division was given an additional capacity of 2M SBUs from the agency reserve, a move meant to "spread the pain" among divisions. Dr. Lee noted that this is just a temporary solution, as SMD demand is still high. During the Federal survey of HEC capacity, it was found that overall at NASA, demand vs. supply is about 2:1 in aggregate. The National Science Foundation (NSF) has a 3:1 problem, and DOE is 4:1. It must be noted that the other agencies have a review process for HEC use, while NASA does not.

In support of applications. NASA's HEC group maintains a 12-person team that specializes in modeling codes. The team has picked an Astrophysics project, ATHENA, in which it is testing magnetohydrodynamics (MHD) code to experiment with increased performance; this investigation is being carried out on an Intel Xeon Phi core.

The Federal Agency Survey on HEC included sites at the Department of Defense (DoD) (including the Air Force Research Laboratory), DOE, and NSF. Not all DoD systems reported. The survey compares the number of nodes, cores and aggregated TFLOPS. NASA supports 800-1000 projects. By comparison, DoD supports about 1800. The take-home message from the survey is that other agencies have heavily used systems as well,

and it's hard to obtain their resources. It is possible to collaborate with these agencies, but one must not expect an exchange of resource, unless it's an emergency, or a demonstration project. If NASA wants a sustained research activity, it needs to build its own HEC capacity.

Dr. Dahlburg asked if Dr. Lee had talked to the program offices. Dr. Lee affirmed this, and added that the communication had taken place through the interagency working groups, and by sending out physical surveys. DoD has 1800 users, DOE has 10,000 users, and NSF has 3500 users. NASA is low on aggregated PF, and has less compute capability for the number of users. Dr. Dahlburg asked about next steps for NASA. Dr. Lee felt the community should advocate at advisory meetings and with NASA management with knowledge of programmatic priorities. Dr. Dahlburg noted that DoD has priorities but still manages to put out 20 PF for 1800 users. Dr. Lee thought the issue is not just about hardware; it's also application software. He agreed to brief HPS on the survey when possible. Dr. Feigelson asked about the needs and allocations for producing data products, and modeling; most university allocations are modeling. Are all data processing needs being met? Dr. Lee explained that data processing has its own separate funding, but many users are coming to HEC to do processing. He guessed the proportion to be 10-20%. Most HEC (at NASA) is devoted to modeling. Kepler is only one pipeline. Dr. Tino asked how work like MHD code is being funded. Dr. Lee replied that this comes out of the HEC funds, to maintain the 12-person workforce that awaits assignment. Tasks are assigned based on impact, speed and savings. Dr. Neal Hurlburt commented that missions will need more support as modeling becomes more integrated into mission operations. Dr. Lee noted that NASA is already in discussion for the Wide Field Infrared Survey Telescope (WFIRST), to understand the overall requirement, and future requirements are being to the budget planning process.

<u>Comments from the BDTF Chair</u>
Dr. Holmes reported on his activities since the last meeting. He delivered the latest BDTF conclusions to the Science Committee and to the leadership of SMD, attended 3 data science workshops, and visited David Spergel at the Simons Foundation. The BDTF report was well received by the Science Committee, which passed two findings to the NAC. The recommendation on an exoplanet cloud computing demonstration of pipeline processing will be revisited.

Dr. Holmes visited NASA HQ at the end of November and delivered to SMD officers the Interim Report of the BDTF, which summarized the accomplishments of its first year and a look forward to the TF's goals for the second year. The report was well received and he took away some good feedback from the HQ officers. Some trends and common themes reported were: SMD needs to formally adopt data science approaches, hire a Data Scientist Program Officer, along with a research program; incorporate data science factors into Announcements of Opportunity (AOs), selections, etc.; employ Cloud Computing services for data processing, and possibly for storage; start leveraging on national science data communications programs; and evolve to "data analytics" architectures. Dr. Holmes felt that NASA management is receptive to these messages. The goals of the present meeting are to revisit Headquarters data and computing

program officers, hear about the DOE exascale program, finish the review of NSF's Big Data Hubs project; discuss study topics; and generate findings and recommendations. Given that changes are expected in key areas of the NASA budget, Dr. Holmes felt that any findings and recommendations should be low-key and general.

Space Telescope Science Institute (STScI)

Dr. Joshua Peek gave a briefing on a conference entitled "Detecting the Unexpected," which focused on discovery in the era of astronomically Big Data, and which was held at the Space Science Telescope Institute (STScI) in Baltimore. The institute's first Big Data conference had 150 applications for 65 spots, and employed an interactive approach. Big Data, as defined at STScI, is a term that signifies data whose raw form is so large that users must qualitatively change the way it is reduced, stored and accessed; data whose reduced form is so large that users must qualitatively change the way in which they interact with and explore it; and data whose structure is so complex that current tools cannot efficiently extract the scientific information that is sought.

The conference was not concerned with data volume but with the scientific method. Technology growth is outpacing astronomical imaging, while exploring data is getting more difficult. There is great value of serendipity in astronomy, as well as in large area surveys. Is the serendipitous value secure in the Big Data area? The conference explored methodological themes such as machine learning (computer does the looking); citizen science (have many people looking at it); visualization, especially data integrated visualization (look at it better). Machine learning, or non-parametric classification of high-dimensional data, included discussions of outlier detection methods. Many recent astronomical finds have included galaxies and stars with unexplained spectra. Often outliers are "close," not extreme. Classifiers are becoming portable from theory to data, and from data to data. Models and theory can quantify the expected, and look for unexpected in residuals.

Deep learning (data classification directly from pixel data) is seeing a fast expansion in astronomy, especially in the use of Convolutional Neural Nets (CNNs). CNNs are useful for classification, and have been used effectively in galaxy morphology identification. Citizen Science has resulted in many big unexpected discoveries to date, and can be used in synergy with machines. Dr. Holmes added that almost all the citizen scientist groups have "superusers" who feed information back to science investigators. Citizen Science requires a new skill set for astronomers to use it effectively, and will benefit from community management.

Fast visualization of data is an engineering challenge; there will be a need to federate big data systems to deal with this. Philosophy and culture are also important considerations. The community needs better ways to protect junior people who are doing high-risk exploration research. The conference held a data and methods bazaar, with both machines and software. A "hack day" was also held, and included exercises such as exploiting variation in survey filters to find extreme emission lines, and the use of new technologies for serving histograms of big data over networks.

Dr. Feigelson asked how one could feasibly do distributed visualization, such as for LSST data. Dr. Peek replied that for a narrow application, there is an architecture for that. The Blue tool brings different data sets to the user, and can be attached to a remote server. Dr. Kinter was interested in the conference focus on outliers. In Earth Science, researchers are also interested in climatology and expected values, and how to characterize the main part of a data distribution. Dr. Peek added that other considerations are how to properly compare simulations with real data, or "Mocking the Universe." Forward modeling is another approach. Dr. Peek felt NASA should bundle forward modelers with telescope projects, which Dr. Kinter likened to the OSSEs (Observing System Simulation Experiments) in Earth Science. Dr. Mentzel commented on how software makers vs. instrumentation makers should be acknowledged in the community. Dr. Peek felt the tension lay in how to give people credit; one way is to consider software as being publishable in the science journals. This would require an alteration in journal structure.

Public comment
No comments were noted.

Science Mission Directorate Program Officer Panel
BDTF heard reports from NASA program officers on the state of the archives. Dr. Holmes felt that the projects had definite ideas of where they were headed, but that the programs didn't have a specific vision. BDTF has been concerned that SMD's data and computing abilities keep pace with the larger community.

Dr. Kevin Murphy of the Earth Science Division (ESD) spoke first. Dr. Murphy runs the Earth Science Data System (ESDS), included distributed data archive system, and the Earth Observing System Data Information System (EOSDIS), managed out of Goddard Space Flight Center. EOSDIS has been around since the mid-1990s, and has moved from non-Internet to almost solely Internet-based process, serving more than 3 million users per year. Over the past 2 years, NASA has substantially reviewed EOSDIS architecture, including looking at processing in place, and expanding access, to see how it will accommodate the Surface Water Ocean Topography (SWOT) and NASA-ISRO Synthetic Aperture Radar (NISAR) missions, each representing 10s of PB of data per year. In Spring of 2016, EOSDIS had a presentation on the evolution of this activity, and presented a plan to expand to the commercial cloud; this plan took about 3 years worth of work. ESD has also begun to consider how to operate archives in the Cloud environment, and how to optimize storage and access. The other part is the reduction of storage at NASA facilities; users will need data close to the processing. The plan is to have the data processed by the expert, and have the Data Analysis Center (DACs) manage it. NASA is currently going DAC by DAC to develop business processes and cost models. There are some issues with egress from the Cloud outward (cost) and some log-in regulations from Department of Homeland Security. If archives can be managed in Cloud Computing (CC) environments, then anyone who has a NASA business or research grant can access those products. NASA is also working with CC vendors on how they can access NASA archives from anywhere.

Dr. Jeffrey Hayes, who runs the Space Physics Data Facility and Solar Data Analysis Centers, spoke primarily of *in situ* and remote sensing data. He reported that the astrophysics (AP) community has been evolving over the last 6-7 years, from proprietary data to open data and acceptance of data standards. Solar is a little different; hence Dr. Hayes was looking for someone like Bob McGuire, who had actively lobbied the AP community to become more open. Progress in moving to Cloud Computing is partly constrained by the budget in Heliophysics, and partly by the anticipated change in government policy. Another hurdle to "full and free" data access is the waffling on security; NASA needs a crisp definition in order to move forward. The good news is that all operating missions are required to have archiving plans, with standard formats, while tentative steps are being taken toward Cloud Computing. The next goal is to get the Community Coordinated Modeling Center (CCMC) to think about what it wants to do about modeling data. Is there a global standard for metadata? Standards will be necessary to drive the research-to-operations movement. Dr. Hayes felt that Astrophysics and Heliophysics are not quite far along as Earth Science, but they're getting there.

Dr. Tom Morgan sat in for Dr. Michael New, to give an update on the Planetary Data System (PDS), which contains about 1.4 PB. As part of a Roadmap activity, the Planetary Science Division set out to estimate requirements for two new Discovery missions. The PDS will grow by about a factor of 13 over the next 10 years, and will be hold about 20 PB. The Roadmap has given PDS a different focus, as Planetary started hearing that customers have "small data" problems such retrieving data from PDS, and finding the tools to work on data. Small users are referred to as "retail" data. By contrast, the planetary missions are generally pretty good on data management and archiving plans. PDS has had the continuing issue of the unexpected, but that is a happy, however costly, issue. The real problems are in the "retail" side of the house; developing higher-order products and putting them back into the archive, especially for the small user who needs tools and needs to know how to archive that data into PDS. They don't like doing this. PDS is going to start seeing users going across all the archives, reaching outside of PDS for planetary data; that's a big data problem. Most countries belong to the international Planetary Data Archive (IPDA), which will be helpful. PDS hopes to have a draft roadmap by the time of the Lunar and Planetary Science Conference (LPSC) meeting in Spring 2017.

Dr. Hashima Hasan summarized activities at the Astrophysics archives, Mikulski Archive for Space Telescopes (MAST), High Energy Astrophysics Science Archive Research Center (HEASARC), Infrared Science Archive (IRSA) and NASA Extragalactic Database (NED), which go back many years and hold standardized data, mainly categorized by wavelength. The Astrophysics Data System (ADS) holds an archive of the literature, which also includes planetary and solar data. NASA is continually working to keep the archives up to date with new technology. Responding to the BDTF finding on ADS that implementation of the proposed modernization of its database engine, user and visualization interfaces may not be feasible at current funding levels, Dr. Hasan informed the Committee that the Astrophysics Division had augmented the ADS budget

after the Archives Programmatic Review held in 2015. With this enhanced level of funding, the delay in implementation has been caused by difficulty in hiring the right talent. Since the Astrophysics archives last spoke with BDTF, all of them had their user groups meet at least once. The NASA Astrophysics archives took over the management of the infrastructure created within NSF's Virtual Astronomical Observatory (VAO), in FY2015. The result of this joint archival effort, the NASA Astronomical Virtual Observatories (NAVO), led by HEASARC, enhances public access to the archived data, responding to the community's and the general public's expectations to find these data on the Web. This year, the Astrophysics Division will have the NAVO project reviewed to determine whether it is providing value.

Dr. Murphy commented that he anticipated the first round of the Earth Science Cloud Computing project to take place in the Fall, after which ESD can evaluate performance and cost. There will also be a test of a full-resolution browse service, imported to the Cloud environment; e.g., 10 years of high-resolution MODIS data. ESD is committed to moving the metadata repository to Amazon by the end of Summer 2017. Earth Science has moved some search programs to the Cloud already, where cost has proven beneficial. It is also experimenting on how to manage software development in this environment. SAUCE, an open-source, metadata formatting tool, is in use as well. In ROSES 2017, ESD has added language that permits open-source planning to be used as an evaluation criterion. Dr. Walker asked about the timeline from beginning to the implementation of migrating to Cloud Computing. Dr. Murphy noted that security is always a moving target, and thought it would end up being more of a hybrid situation. ESD studied building its own Cloud management software system, but it was faster to use what is commercially available. NASA procures hardware in increments. Commercial cloud buys are similar in nature. If security interferes, the service will have to be reduced.

Dr. Hayes, addressing the concept of retail data, commented that for some things like Grand Challenges, the government will have to store data. This is an unfunded mandate on the proposer. PDS still does peer review on data before it goes to the archives. The question is what to do with the little guys. What's the right level of detail? As an example, the NuStar Astrophysics mission provided data that was relevant to Solar science, raising the issue of how the two archives communicate with different software. There needs to be an intelligent way to associate the data between the two disciplines. Another example is the Trappist-1 exoplanet system, which has applications for all four disciplines. Interoperability needs to be based on well-defined standards common to all the disciplines.

Dr. Feigelson also commented on retail data. He said the AAAS journals are encouraging the definition of data behind figures, and are hiring staff to deal with it. Commonly, this translates to associating DOI (digital object identifier, a unique persistent identifier) links with articles. SMD archives might want to talk with the journals to help close the gaps. Dr. Murphy noted that there is a lot of consideration in Earth Science on how to

support DOIs for its products, and how to work with journals in making it available. Dr. Walker said AGU has a policy that data must be publicly available; and although DOI is the obvious way, not everyone uses it. Dr. Mentzel commented that Cloud Computing has had a secondary effect around the talent question; it makes the data more available to the new generation of scientists, as well as more ways to make data accessible, enhancing the ability of others to use the data, or build new tools. Dr. Murphy said that Earth Science (ES) does use Docker for workflows, as it wants people to transparently understand how ES generates its products. It is still an open question about whether it is more affordable. Ideally, NASA wants to help users bring their expertise to NOAA or USGS, and also wants to facilitate work with other international groups on Earth-observing satellites. The agency has an interoperability catalogue to make data discoverable.

Dr. Kinter asked if Earth Science (ES) could do the data analytics part on the Cloud; e.g., can a graduate student do it and publish it to the cloud? Dr. Murphy said ES would like to put up URLs for those objects; Cloud access doesn't cost NASA, the Cloud conveys the costs to the user. Users want to be able to operate on data that is *in situ* or on the Cloud. Dr. Murphy said NASA wants to enable people to bring algorithms to the data, but at their own cost. Dr. Tino, reading a press release from a NOAA project, related that while Amazon Web Services (AWS) bore the cost of hosting data on the Cloud, customers run algorithms on it at their own individual cost. Dr. Murphy related that NASA and NOAA are interacting, while Earth Science is testing how to more effectively manage archives on the Cloud; the answer is not yet clear. Dr. Tino noted that the question goes back to how to assign value to data sets; NOAA seems to have found one way: is there a similar value that can be attributed to NASA data sets? Dr. Murphy noted that ES has worked with Google Earth, and AWS, and just doesn't know yet if the commercial value is there. ES met with senior Amazon management last June and learned that they see value in structured data sets for machine learning, and that they're willing to work with NASA on this. Dr. Kinter cautioned that NASA could ultimately end up with all the low-value data in this scenario. Dr. Hayes reminded BDTF that NASA must hold all the data, period, as an inherently governmental function, as a permanent archive and curation organization. Dr. Kinter noted that sometimes unexpected things can be discovered in low-value data. Another consideration is the reprocessing of old observational data; there is a process in Earth science referred to as "reanalysis" which uses modern techniques to quality-assure and assimilate past observations, and that requires well-documented, digitized versions of archival data. Once data is stored in a low-cost medium, there is the risk of high cost associated with making it usable again: How do you determine what type of data is worth resurrecting?

Dr. Holmes asked if SMD was addressing data analytics. Dr. Murphy said that many in Earth Science are doing so, as there needs to be a foundation of basic data and structures that can help users generate their own data analytics tools. Dr. Hayes replied in the affirmative as well, and added that he thought the age of agency-developed tools is gone, and that discovery and enabling are key. Let the community figure it out for themselves. Dr. Feigelson felt that NASA emphasizes data reduction software in its AOs,

which are treated very seriously. He thought the AO peer reviews were lacking the equivalent of a science review of the algorithms and methodologies used for reducing data. He asked if the need for mathematicians in the peer reviews resonated with the panel. Dr. Morgan said he had heard Dr. Zurbuchen echo this same point. Dr. Hayes commented that data reduction runs into some intellectual property issues, as well as data quality, and data flags; 40% of the project is the back end. He felt that was something NASA still doesn't consider. Proposers tend to low-ball the AOs because they are conditioned to do so. An extra or supplemental (algorithmic) review is going to cost time and money. Dr. Holmes raised the issue of stovepiping. Dr. Hayes said that exoplanet interdisciplinary research has been going well, but he was not sure it could be generally expanded. Dr. Mentzel suggested there could be lessons learned from how NSF rolled out their data plans, adding that there can be problems if you bake specific instructions into the AO; he recommended starting out with what's been working in exoplanets and work it outward, and look for areas that benefit the community—that will move the peers. Dr. Murphy noted that Earth Science has a history of requiring data management plans, and communities have sprung from this requirement, resulting in more efforts in how to improve data services and standards. NASA actively participates in this process. There is also the ESDS Working Group, which looks at best practices, how to do a mission close-out, etc. Dr. Murphy felt that ES already has a rigorous process on determining data processing methodologies up front, and the AOs are updated regularly. Dr. Hayes noted that as cubesats become more pervasive, with different requirements, and less rigorous parameters: Do we want to apply the same standards here?

<u>Big Data And Data Science at JPL</u>
Mr. Daniel Crichton gave a briefing on big data and data science at JPL. Dr. Richard Doyle joined the briefing as well. The presentation dealt with looking at data from point of collection all the way out to data analytics, a Big Data life cycle that focuses on computing technology needs, from end to end.

Dr. Crichton described work at UC Riverside, a recipient of a 5-year grant on developing data science, and which just created a Masters degree in the field. Work is also ongoing Caltech, NIH, the Defense Advanced Research Projects Agency (DARPA), and NSF, on extracting insights from increasing data volumes, and questions of data infusion. A challenge that goes beyond science, e.g. in NASA's Deep Space Network (DSN) is monitoring anomalies. JPL and Caltech have formed a joint initiative in Data Science and Technology, with 10 projects under way. Big Data for JPL is a statement of the problem, and data science is a response to the big data challenge. Data science includes scalable architectural approaches, techniques, software and algorithms. Big Data methods are required when data collection processing and management exceed the capacity of available methods and software systems. JPL is also looking to optimize choices for NASA-unique problems.

In 2010, NASA/JPL began work with the National Research Council's development of a report, *Frontiers of Massive Data Analysis* (published in 2013), which discussed various needs for data analytic methods. As data increases, there is an opportunity to

systematize the analysis of data. The report called out the need for an end-to-end data life cycle, from point of capture to analysis, and integration of multiple discipline experts. Dr. Crichton emphasized the latter point. Overall, the report called for an application of novel statistical and machine learning approaches for data discovery.

Today, NASA science and Big Data starts with mission operations, moves to science data processing, science data management archive and distribution, and how they all connect to Big Data infrastructure (data, algorithms, machines; or, the data ecosystem). The JPL Data Science Working Group was established in 2014, which started an effort in onboard agile science, Earth science distributed data analytics, and machine learning for transit detection in astronomy. A Data Science WG was chartered in 2016 to report to JPL's Leadership Management Council, covering all aspects of the Lab's operations. It is now working on identifying pilots to help transform the way NASA is executing missions as an institution.

Data science strategy guiding principles include data life cycle, data ecosystem, and data sciences as cutting across capabilities at NASA. As an example in Planetary, there is sifting through data to find dust devils on Mars. For ground-based systems for AP, there is streaming and extracting information from real-time data feeds, with data distributed in massive archives. Interoperability of data archives will enable international data archive and sharing architectures. Apache in Space is an open-source big data stack that has been brought in to build data systems. Google, Amazon and Microsoft are starting to make their libraries available. Common data elements and information models (about 4000 different data products) need to be described in standard ways, to bring them into visualization techniques.

The future of data science in the Big Data research environment will benefit from a reduction in "data wrangling." JPL is exploring opportunities and use cases across the ground environment, such as intelligent ground stations; intelligent archives and knowledge bases; intelligent MOS-GDS; and data analytics and decision support. A sample case is analyzing water cycle patterns in the Western US. In 2015-16 an Advanced Information Systems Technology (AIST) Big Data Study was undertaken by JPL, which mapped technology and data needs against the mission data life cycle; results were used as key input for the 2016 ROSES call in AIST. The AIST Big Data Study identified 10 Year Capability Needs for NASA in observational programs, ground-based mission systems, massive data archives, and distributed data analytics.

Dr. Crichton discussed aspects of PDS, whose purpose it is to collect archival data and to make digital data and documentation accessible. Challenges for PDS include the wide variety of planetary data, data volume, and the federation of disciplines. These factors affect computation, data consistency, data storage, discovery, etc. PDS4 has been adopted by international partners, in support of an open Planetary approach. PDS4 was endorsed by the IPDA in 2012. Its tools include a Lunar Mapping and Modeling Portal, which helps to explore data visually. Another use case is the Western States Water Mission, which seeks to understand water availability in aquifers and snowpack. Models and data are beginning to be compared, to better understand the cycle.

Observations for modeling from visualization to research, in a project called WaterTrek, are providing greater insight into the data.

JPL is also looking at methodology transfer into other domains, as in a partnership with the National Cancer Institute (NCI), which is capturing data on cancer biomarkers, proteomics, genomics, and imaging. The collaboration was the subject of a Congressional briefing in October 2015. The Early Detection Research Network, another project in this effort, is a comprehensive infrastructure to support biomarker data management across centers. JPL is running programs normally used for transient detection to pick up anomalies in cellular images, and is also working on identifying and classifying cancer's molecular signatures. Other partnerships are being carried in searching the dark web, SPAWAR (information warfare), the Earth System Grid Federation, and next-generation data infrastructures (NSF).

Caltech now has a Center for Data-Driven Discovery on campus, and a Center for Data Science. Caltech/JPL established a 9-day virtual summer school in 2014, which has drawn 25,000 students. The UC Riverside data science program placed 19 students at JPL for training, supported by a $4.5M grant from NASA. Dr. Crichton summarized with recommendations:

- Use the mission science data lifecycle to organize Big Data at NASA.
- Enable use and data analytics for the community (data ecosystem)
- Explore opportunities for methodology transfer
- Establish multi-disciplinary teams between science/discipline experts, computer science/data science.

He concluded by mentioning the Planetary Science Informatics and Data Analytics Conference, to be held in April 2018.

Dr. Holmes asked how JPL gets the code sitting next to the data archives so that remote analysts can use it. Dr. Crichton answered that JPL is trying to model one aspect of the problem by using a simulation engine (modeling computational workflows). Dr. Doyle noted that simulation can reveal whether the solution is speed, or Cloud Computing. The Earth System Grid is doing work on model output, which is going to generate about 30 PB. Dr. Tino said he resonated with the concept of agile science, which is similar to modern software development approaches, as it will be essential to be able to quickly able to feed back new data into systems.

<u>Discussion</u>

Dr. Holmes reviewed the next day's agenda, and floated the idea of a face-to-face meeting at JPL.

Dr. Walker was happy to hear Dr. Zurbuchen's answers on archiving simulation outputs, to enable comparisons of models and codes. He noted there would be a CCMC modeling

and simulation workshop the first week in April at Cape Canaveral; one topic will be archiving simulation and results.

Dr. Holmes briefly touched on his activities, first on giving the BDTF interim report to Dr. Zurbuchen. He reported also having spent time at the Simons Foundation with David Spergel, formerly of the Science Committee and Space Studies Board. Dr. Spergel agreed that BDTF is congruent with Dr. Zurbuchen's SMD philosophy. Dr. Tino brought up the issue of the proper process for bring software refreshment into budget process, through the Senior Review. Based on Dr. Beebe's comments on Cassini, he wondered if it were still the case that NASA is approaching hardware as platform outlays. He hoped the Cassini issue would raise awareness. Dr. Holmes commented that people will need to optimize codes for efficient use on supercomputers, and he felt that Dr. Zurbuchen spoke to this issue earlier. Dr. Tino called for a more holistic approach to platforms, to consider them as involving both hardware and software. Dr. Kinter commented that NASA will need to throw big dollars at vendors for co-design tasks; DOE is getting interest from vendors by offering billions. Dr. Feigelson thought that NASA has been blurring the distinction between pipeline processing and science processing, and that this was good. But there are branches of NASA providing sophisticated science products, as in EOSDIS, and in AP, that are intended for science analysis. It must be recognizes that it is not a mechanical process, it requires years of advanced methodology performed with advanced software. The problem is the lack of quality control; some researchers don't know how to do basic statistics. As a result, NASA-provided software is inconsistent in quality. Dr. Feigelson felt that if you embed the requirement in the AO of the mission, it's simple. If it's created by the center, he didn't know how to do this. Dr. Tino noted that it's getting to the point where the software is the science; NASA can't ignore this.

Dr. Holmes, referencing the morning's briefing by Mr. Sprague, noted that his agency working group was internal to NASA. Dr. Mentzel asked whether SMD would want the same structure, and in addition a more bottoms-up community effort. He felt a monthly meeting could encourage cross-pollination. At present, awareness of data science and the application of it across SMD is spotty; this issue might be remedied by shining a spotlight on it. Dr. Kinter suggested reaching out beyond the labs at NASA and to Dr. Murphy's Earth Science Information Partners (ESIP) federation. Dr. Holmes said he would add this thought to his note to Mr. Sprague. Dr. Feigelson suggested encouraging NASA to increase its HEC capability 42% per year.

Dr. Holmes wondered if anything similar to the data science workshop at STScI was happening in other disciplines. Dr. Feigelson noted that in AP, the workshops are populated by the younger generation, and that NASA would do well to do more in this area. Dr. Beebe confirmed that Planetary is spending funds to train users in data sciences. Dr. Walker noted that in the case of the Magnetospheric Multiscale (MMS) mission, they're running hard to keep up with state of the art instruments. The mission has had to stop and re-invent methods in order to accommodate the instruments. Dr. Feigelson felt that if BDTF does recommend an Office of Data Science, it should also

recommend that the funds come from Education. Methods-implemented software can improve the sensitivity of instruments.

In regard to the SMD archive panel, Dr. Holmes felt encouraged by the demonstrations in EOSDIS. Dr. Walker remembered an old NASA program, Applied Information Technology Research, which might be resurrected in some form to accommodate data science. Dr. Tino mentioned that in tests with AWS, the actual metadata was able to be hosted on AWS object storage, and added that it would be worthwhile to find out how they made progress on the storage issue. He also felt Dr. Kinter's previous point on the low-value data/Obamacare analogy is spot on, and that NASA needs to figure out when moving data to cold storage is worth the cost, from a future technology standpoint.

As to the JPL discussion, Dr. Beebe felt it was worthwhile hearing about cross-discipline work.

March 7, 2017

Member Reports
BDTF members presented reports on their individual activities.

Dr. Kinter described recent thoughts on Earth Science modeling for coupled Earth-system models. In the past, the components were modeled independently, while now more attention is being paid to interactions, e.g. between atmosphere and ocean. Modelers are starting to think about it in a more rigorous way, with the addition of variables like aerosols, ocean waves, and nonlinear effects. These changes will have impacts on software, data flows and workflows. There is a multiagency effort under way (14 agencies, mandated by Congress in the 1990 Global Change Research Act), that is building software infrastructure across agencies. The next National Climate Assessment is coming out in 2018, a process the agencies are now committed to every 4 years. The bold new thing coming out here is the science of attribution of extreme events. While the statistics are not easy and the physics not clear, there is nonetheless an effort to try to attribute events to human interference in the climate system, which is looking to use satellite data and trace back processes. The annual meeting of American Meteorological Society was centered around the theme: Observations Lead the Way. A special issue of the Bulletin of the AMS is forthcoming on this concern. There is also some community concern about the dilution of science rigor.

The NASA Modeling Analysis and Prediction program had its quadrennial review panel, with proposals to the AO that include tropical interseasonal oscillations in global climate models. There is also keen interest in gravity waves, as in vertically propagating gravity waves, which can break and deposit momentum in the Earth's upper atmosphere. Researchers are keen to observe them better. Larger Earth Science models that incorporate vegetation dynamics, and connections to human health (black carbon and respiratory disease) are also in development.

The Earth Science modeling community is concerned about coarse spatial resolution of models; processes such as cloud formation and dissipation must be parameterized. Parameterizations don't work well enough. There have been attempts to simulate these small-scale processes directly, e.g. with the use of cloud-resolving models. This means more data and process studies, thus the community is expecting an explosion of data requirements to solve the parameterization problem.

Another ongoing activity in the Earth science community is coupled model inter-comparison, which is conducted internationally about every 7 years. The modeling groups in many countries, including 7 in the US, run the same prescribed experiments with their Earth system models and share the output. The 6th generation of this effort (CMIP-6) is underway now, and the voluminous and complex output data will be available in distributed archives worldwide. The thought of 35 PB of data distributed around the world is giving everyone a headache.

Dr. Walker reported on his efforts on testing high-speed networks on the West Coast, thinking about running 50 PB per run in modeling, and how to handle it. He began a study in cooperation with Ames and UCLA, and like other universities in the West, set up Pacific Rim, a high-speed network, and began a series of tests to optimize transmission. Results of the tests found that the limiting factor is getting the data off the system. At Ames, he got it to 6Gb/s, and could pull the data across the network as fast as it could be read. The numbers are now 12 Gb/s on a solid-state system. Dr. Walker was very encouraged by the results. The other part of the testing will involve doing some of the processing on the supercomputer, which is to be decided.

Modeling and simulation: how do we meet the requirements to make them publicly available and archive the results? Both in Heliophysics and Planetary, there's much cooperation between ESA and NASA. ESA is taking lead on the problem, and has come up with metadata standards for archiving. At the upcoming CCMC workshop at the Cape, one of the working groups is on archiving and making available simulation outputs. A number of other groups on science and modeling issues will be represented there. Dr. Holmes asked how one might extend the work of the CCMC to the broader community. Dr. Walker said the CCMC had been set up by plasma physicists, very much based on their discipline; the metadata is unique to them. They usually run simplified calculations. CCMC came up from the science community; Dr. Walker would hope that other disciplines would organize similarly from their grass roots communities. He noted that CCMC is funded by both NASA and NSF. Dr. Kinter noted that Earth Science has an Earth System Grid Federation (ESGF), initially funded by DOE, which now has centers in the UK, Japan, and China. The centers receive modeling results data, and they require the data to come in a standard format and by established conventions. The data require postprocessing by the modeling groups, and they deliver to the Earth System nodes. Some countries require credentials (for use), some don't. NOAA is required to archive atmospheric monitoring data. Dr. Kinter recommended contacting Lawrence Livermore National Laboratory, as they run the ESGF. In Astrophysics, Dr. Feigelson noted there is a famous code called Gadget, which has been around for 15 years and is

widely used. The use of it is independent, but there is no organized repository. Dr. Holmes asked Dr. Kinter to include archiving results information in his paper.

Dr. Holmes reported having attended data science workshops at NYU, Columbia University, and the Space Telescope Science Institute (STScI). He several people at Columbia University, one of which was an applied physicist at Google. The other is a key figure in Internet 2 activity and is involved with DOE's Energy Sciences Net and the Pacific Research Platform. He felt it critical to get SMD more involved in these national science community networks. The concept of the Data Scientist really hit home; it is clear SMD needs someone working on digging information out of data sets. The BDTF would be better termed a Data Science Task Force, as data scientists are attacking real problems in our society.

Dr. Holmes also met with David Spergel at the Center for Computational Astrophysics, a division of the Simons Foundation. Dr. Spergel felt the Space Studies Board would be receptive to BDTF report results, particularly regarding workflow activity. He suggested looking at what happened with the revamping of models for stellar evolution which significantly changed the efficiency of getting results out to the community.

Dr. Beebe reported that the planetary atmospheres community is starting to deal with archiving modeling results for the Martian atmosphere, and that she had been interacting with Ames about this. It is currently possible to provide links to web pages of groups involved in atmospheric modeling, but NASA must still develop standards for archiving model results in this discipline. She noted that the Ames group is in the process of structuring modeling questions. The PDS, in conjunction with ESA/PSA has led the development of the International Planetary Data Alliance (IPDA). Membership includes national agencies that are involved in planetary missions. The IPDA has accepted PSD4 (an XML based system developed by the PDS) as the archiving standard. This has been motivated by existing situations. When a country comes on line with a mission, they frequently find there's no money for archiving. The PDS4 model is available to them in such cases. ESA has worked closely with NASA, and is using PDS4 for BepiColombo and their Mars 2020 mission. The United Arab Emirates (UAE) is planning a Mars orbiter, and has consulted closely with the US to help them develop the mission. Their orbiter will be filling a niche for 24-hour atmospheric observations of Mars, and is also using PDS4. The next candidate is a South Korean lunar mission. This mission will present at the Town Hall at LPSC. They are committed to using PDS4 and possibly will join IPDA officially.

Addressing Dr. Morgan's "retail" problem, Dr. Beebe said the PDS Atmospheric Discipline Node (ATMOS) has been canvassing recent grantees to get them started on setting up their data bundles and labeling their data; ATMOS is nearly ready to demonstrate beta-testing (in June, approximately) on software to simplify this process. One argument to use in this effort is to demand that users learn how to set up their own dataset, and as an extra benefit it will help them learn how to use PDS4. Beebe is hoping to reduce the grumbling of the data providers. Encouraging the providers to create a file

and dump information in it as they go will help the data supplier and NASA manage the retail system.

Dr. Hurlburt noted that small Explorer missions (SMEXs) are moving to K_a band, which will increase data volume in small missions. He assumed this growth would apply to other missions. The Daniel K. Inouye (DKI) solar telescope will generate 10TB/day, for instance. The telescope's data center goes to Preliminary Design Review (PDR) next year, and to operations in late 2019. The European Flarecast project is making progress in data science, using machine learning and feature extraction techniques for an automated pipeline. In the US, Dr. Hurlburt noted progress being made by young researchers at Stanford and Berkeley, applying machine learning techniques for coronal mass ejection (CME) prediction, including addition coronal observations from the Solar Dynamics Observatory's Atmospheric Imaging Assembly (AIA). He also detailed his experience at the Astronomical Data Analysis Software and Systems (ADASS) meeting in Italy, where he found efforts to bring code to data, focusing on VOspace, a uniform authentication scheme for logging into systems and running code locally. He felt it similar to the DOE DMZ's Asterisk project, which is experimenting with a 3G pipeline from Europe to Africa. He also related that Lockheed Martin now has a data scientist.

Dr. Feigelson noted that Big Data currently plays a relatively small role in AP, mostly from ground-based instruments and new radio interferometers. The Large Synoptic Survey Telescope (LSST) is expected to generate 5-10 TB per day, and the radio LOFAR interferometer requires 30 TFLOPS continuously. WFIRST, in comparison to ground-based projects is relatively easily. Kepler's use of the Pleaides supercomputer has been praiseworthy. In Big Theory (computational cosmology, computational astrophysics, planetary accretion, etc.), progress has been good and steady. The archives are effective and stable. For source detection, and spatial analysis for extragalactic astronomy, the science is complex but the methodology is fairly simple.
American journals are modernizing in their efforts to edit statistical techniques, and improving the quality of ADS, which contains links to data and software, some of which is done via DOI. Journals are encouraging links to instruments as well. ADS is recognized as an excellent portal into both NASA-produced literature and data, and it is being filled in in a mature fashion, aided by the energies of young scientists like Joshua Peek. Dr. Feigelson noted that IEEE has been doing well in getting engineers to pay attention to astronomy, and felt that SMD should talk to mathematicians in a more systematic way.

WFIRST is currently in phase A, and had its first data challenge based on simulation codes; Dr. Feigelson felt that NASA overall is doing well in Big Data and Big Theory. Dr. Beebe added that ADS also provides abstracts from meetings, which is a good resource for students looking for dissertation topics. Dr. Feigelson noted that ADS is also at the cutting edge of library science, in not just data but knowledge.

Dr. Mentzel shared details of a three–day symposium at the Moore Foundation, which brought together 50 post-docs across the natural sciences, computer science, and statistics for. These were early-career researchers, all data-driven in some way, at the

intersection of many domains. The methodologists have a generally positive outlook, while many natural scientists are pursuing science outside academia. It is important for NASA to look for recruits in data science as an opportunity to foster a different type of research, perhaps by tweaking incentives for what it means to be a researcher. Include the notion of software as instrumentation and technology development. Instrument builders are essential, and one must think of the data scientist as being equally essential to the mission scientist. Dr. Mentzel felt that the best way to get the eclectic group together was over data types; scientists from different disciplines can make huge progress in sharing best practices, and uncovering problems and solutions. Such an approach might help NASA break down silos. He concluded with a last thought, that data science is still not a well-defined term, and probably many data scientists at NASA would not self-identify as such. NASA might also want to alter titles to reflect this change.

Dr. Tino commented on discussions he'd been having in San Jose on the subject of "technical depth," noting that it is very common in industry to take legacy codes and leave them running, which becomes expensive. Coupling engineering to science can help address the issue to minimize long-term costs. He thanked Dr. Feigelson for helping him think about this.

Having shifted into management, Dr. Tino had taken to thinking more about his responsibility for general software as a whole, and also getting a sense of how business is reacting to Big Data. This is leading to a different feel about how to monetize programs. Data analysis is essentially trying to discover questions, which is a hard sell in business. Data analysis requires organizing and crossing silos even when it is not known what the results will be. How do we identify the value of data analysis. Financially, there are now interesting questions around scale, where industry is starting to see inflection points for minimal scale; this applies to both NASA and commercial enterprise. From a technology standpoint, more users are going to container-based platforms, and DOCKER. He mentioned that Dr. Beebe had inspired him to use Jupyter notebooks for documentation, both for archives and active projects. He had also begun taking implementation to business logic for financial staff, as it does have a place in how we talk about metadata, and how to store analysis code. Dr. Tino also investigated some Google projects, and deferred that discussion to a later one on research results.

<u>Progress on BDTF research topics</u>

Modeling workflows
Dr. Kinter commented that the problem in a lot of modeling for the NASA science disciplines is that there hasn't been much progress since the 1970s; workflows have remained essentially unchanged. Models are ingesting and generating 10-100s of TB. Dr. Kinter sent questions to a list of experts at NASA Goddard, GISS, NCAR, and one researcher in Belgium, and received some initial thoughts.

Dr. Mentzel interjected that Earth Science can be broken down into data assimilation (prediction); long-term climate modeling, (simulation); and climate interdisciplinary

modeling, (coupling climate models to socioeconomic models). The three categories vary: some are volume constrained, some are variety-constrained. The data needs to be more understandable, accessible, and incorporable into models.

Dr. Kinter related that NCAR's Dr. Richard Loft, a CTO at NCAR for a few decades, answered his query. Dr. Loft designs and manages data systems for NCAR, and had remarked that modeling data production rates track with sustained performance. Sensors are growing exponentially, with 50B sensors estimated to be connected to the Internet by 2020. Latency bandwidth is not keeping pace with data, or data budget. The government mandate to store and make available data is a challenge for researchers. Earth System modeling workflows are unmanageable, and are reaching a tipping point in terms of workflows. The critical path to streamlining workflow includes modularization, parallelization, lossy data comprehension, and other tools for managing hierarchies and workflows. Dr. Loft felt it essential to invest in fundamental statistical machine learning, and also noted there is a penchant for bit-for-bit reproducibility, but that reproducibility is ephemeral at best if the order of operations is not preserved. He felt bit-by-bit is a bygone expectation in Earth Science, and that the community needs to get over it. The bottom line is that everything is a probability distribution. Also, slavishly reproducing metadata in every file is unnecessary; NASA needs to go back to a linked-object approach to metadata documentation.

Dr. Walker took an action item to ping Giovanni LaPenta for his insights, and said he strongly agreed with Dr. Loft's comments on reproducibility. Dr. Feigelson suggested names off line pertaining to the workflow effort, and agreed to send contact information to Dr. Kinter. He also asked for further context from the Planetary and Heliophysics community models. He also mentioned a new way to view the uncertainty of a computational mode; its nickname is Uncertainty Quantification, and is a hot topic. There is a UQ toolbox and 2004 Data Challenge at Langley. Dr. Mentzel noted that there is also a study in engineering management science on SIP, stochastic information packets.

<u>NSF Big Data Innovation Hubs</u>

Dr. Fen Zhao presented a briefing on NSF's Big Data Regional Innovation Program, Hubs and Spokes. She noted that she came from a computational astrophysics background and was well acquainted with BDTF challenges.

The ultimate goal of the Big Data program is to bring together domain scientists, computer scientists and end users to use data to solve challenges. The Hubs and Spokes program is really a mechanism to reach out to stakeholders.

The vision of the program, launched in 2012, came out of the Networking and Information Technology Research and Development (NITRD), which recognized that data science is happening everywhere, and wanted to bring data scientists into the fold. The Hubs program goal is to create a community-driven partnership. NSF launched four Hubs, or consortia in four geographic regions, which are represented by industry, academia and government. The Hubs awards were made in November 2015, and the

Spokes awards were funded (10 in 2016, one additional in 2017). NSF also funded ten planning grants ($100K@), ahead of the second solicitation. Examples of grants are data science for criminal justice, and an energy project in upstate NY. Why pick geographically? Wanted F2F interactions, have participants work together work to work. NSF already funds topic-based research (RCN mechanism).
Each Hub has a series of Spokes, or local priority areas, often with specific projects (Nodes). The Spokes are free-standing project ideas that are not necessarily being funded by NSF. Over the long term, Spokes will have to figure out how to sustain themselves operationally.

NSF has also have initiated a Big Ideas project; one of which is Harnessing the Data Revolution, and includes computer human frontier (robotics); multimodal physics for astronomy; genomics. Harnessing the Data Revolution is thinking about funding programs beyond disciplines to include data science workflows, and challenge centers for targeted problems. The TRIPODS program brings together computer scientists, statisticians, and mathematicians to work on data science. NSF has started funding "centerlets" in this area. NSF is also developing ideas in infrastructure (DIBBS, which produced EarthCube); and Education (data science workforce; National Research Traineeship). The new Big Idea framework is in work, and TRIPODS is part of it. Big Ideas is now a five-part program: domain science and applications; data systems; data science foundations; cyberinfrastructure; and workforce and education.

The Hubs themselves over the past year have been working with technology vendors such as Microsoft, Amazon and Google in helping the community. Microsoft, Amazon and Google have awarded Hubs $3M in Cloud Computing credits. There have also been massive regional all-hands meetings with hundreds of attendees interested in data science; and the establishment of early career researcher programs with the Computing Community Consortium (CCC). There are also two PIs working on a three-year sociotechnical study on the ecosystem of data science within the Hubs.

Spokes are designed to be mission-driven. A typical Spoke project would be developing tools for a community, a little broader than an R&D project. Spokes have three major themes: Grand Challenges (water in the West, food safety in Midwest); data sharing; and automation (Artificial Intelligence; automation of part of the data pipeline; thinking about the entire data life cycle). On top of themes are areas of emphasis: neuroscience; replicability and reproducibility in data science; data privacy; education; data-intensive research in social behavioral and economic sciences; smart and connected communities. The distribution of funding per region is roughly equivalent. The total Spokes outlay is about $12M in the first round. Dr. Zhao highlighted one notable Spoke project based on data sharing within legal frameworks: automating the production of a legal document. This application is tied to a distributed database developed by MIT, and it is hoped that the application can change the way data sharing is done, such as in health documentation.

Other Spokes projects involve IBM's Watson and the Encyclopedia of Life (EOL; the largest database of biological species and biodiversity, held by the Smithsonian), from

the Automation theme. Georgia Tech is working with IBM Watson to make EOL easier to access and use. One tool is a semantic processing app to allow middle school-level students to ask a question and have it answered at their level. There are two tracks in this Spoke, one for scientists and the other for the lay user.

Smart Grid Data Sharing is creating an organization that brings together disciplines across industry, academia, and government. There are many international partners in this effort. Its particular challenges are privacy rules for sharing utility data.

The Digital Agriculture project matches unmanned aerial vehicle (UAV) data with plant phenomic data to understand the state of crops. Currently, there is a problem with harmonizing agricultural drone data with plant phenomics data. This project is building tools that can make conclusive statements about what the drone sees. Webinars are used to teach students how to use this data, which helps to harmonize community around the method. Asked if there was potential here for interoperability with NASA data, Dr. Zhao said she could easily see this extension. Spokes are three-year grants; the plan is to fund Hubs for another cycle and allow people to re-apply and renew. Spokes are also eligible for supplemental funding.

Other projects are: Metro Insights (weather and satellite data); human health (merging environmental data sets with human health records); and Smart Grid (weather data). There is now a Department of Commerce mechanism for engaging the Hub; any major data science project can go through it to engage the Spokes. Dr. Holmes saw two potential recommendations: that NASA hold a coupled research solicitation with NSF, which SMD can broadcast to the research community, and that NASA send out an FYI about the existence of the Hub and Spoke infrastructure. Dr. Zhao said each of the Hubs has a point of contact, and that NSF is also trying to build out a virtual organization, which is probably about a year from being online. Each of the Hubs has its own website. Dr. Feigelson asked if the Hubs were limited to societal challenges. Dr. Zhao said there will be more of an emphasis on scientific topics in the future.

Dr. Holmes felt satisfied that NSF has organized an infrastructure of expertise that the NASA research community can leverage. Drs. Hurlburt and Tino affirmed Holmes' thoughts as well. Dr. Mentzel asked what the vision was for increasing participation in the Hubs. Dr. Zhao said that NSF is trying to figure out how to scale up slowly; the creation of a virtual organization will help. Some of the Hubs are also trying hard to get outside funding to get more people onboard. NSF needs to be selective for now. She mentioned there would be a Hubs PI meeting coming up 15-17 March, held jointly with the Big Data Research Program.

Discussion of Research Topics

Server side analytics

BDTF discussed the topic of server-side analytics, or data analytics, which involves moving the code to the data and requires network proximity to data. It is very

cumbersome to transfer large data sets, buffer them locally, etc., and the time has come to do something different. At the STScI data science workshop, there were discussions of the pressing need to do data reduction in order to accommodate the big Astrophysics missions of the future (WFIRST, JWST). The community is looking at CyServer, developed from the Sloan Data Survey, which does some processing at the data site. There are also European data reduction tools for Gaia, called GAVIP. At Goddard, John Schnase at the Climate Analytics Service uses the Modern-Era Retrospective analysis for Research and Applications (MERRA) database, an amalgamation of Earth Science data sets useful for climate research, produced on a regular basis. CDSLive, based on a Python library, looked at MERRA data and how it can describe the impact of precipitation on irrigated land. In this scenario, the metric is large; it takes 8.4TB to transport data for this one problem set. A significant reduction in data volume and time can be achieved by using a two-step process. Dr. Tino commented that this is similar to what Google is providing with TimeLapse. TimeLapse is a collection of all of Google's catalogue for standard Earth Science data sets. It contains a 40-year history of LandSat data, as well as Sentinel and MODIS data, and high-resolution imagery. A functional platform is on top of it. There is a separate program called Google Cloud, which provides raw compute on top of the data, and provides reduced function sets on top of the data set itself. Users have access to convolutions and functions over a portion of the data set over a period of time, and can get reduced data sets as a result. It is very similar to the Schnase approach. CyServer appears to be somewhere in between TimeLapse and Schnase's approach. Dr. Tino felt that both approaches were valid, and that it would be in NASA's interest to classify access to data more specifically. Dr. Holmes commented that users have a spectrum of expectations at NASA for access to data; e.g., modelers need to work with the real code, and need to come up with a justification for the resources to deal with it. Dr. Kinter noted that the Global Modeling and Assimilation Office runs a data assimilation code, which merges all available observagtions of the atmosphere with a state-of-the-art model. The output data set is MERRA, which goes back to 1979. Giovanni sits on top of the DAAC at Goddard, which puts tools on top of the data sets. There are two paths from this approach. The user can decide data reduction, like zonal averages, time averages, or staggered epoch averages in time, chosen from a menu. The other way is to provide a data analytics engine to the user, who can then specify an algebraic expression; this is more flexible to the user and more complex, taking longer to process.

Dr. Holmes said he was hearing caution about how to set up the group that sets up the analytic processing. Sandboxes in institutions must avoid overexpense and underperformance. Dr. Hurlburt cited the experience of a Joint Science Operations Center (JSOC), which houses the SDO data set; the total volume distributed to any user is limited and scriptable. They also support users running their code on the JSOC machines, thereby bringing code to data. This has been common practice in the Heliophysics seismology community for a long time. Dr Hurlburt's group set up a Jupytr hub to experiment with server-side processing using Python and IDL for the International Space Science School in November, but it didn't work due to a poor Internet connection. They are now try again on a local project.

Data discovery methods

Dr. Beebe said she was still planning to visit the four SMD disciplines and do a fair survey on the state of the need in each subdiscipline. She planned to two questions: does the data exist, and can you get it all?

Need for improved data/science analysis methodology (and technology)

Dr. Hurlburt discussed the theme of the data science triad presented at the last BDTF meeting by Dr Mentzel and how it fits in history, while trying to update the triad to reflect the need for data science in missions; to lay out and review the rules of the road for open data, e.g. standard formats, communications protocols, etc. Within the mission concept, there are recs for project data plans and maps. Missions are getting more tightly integrated with models, which need to incorporate the data. A data scientist can play a role in revitalizing advanced software development to include data science; embedding data science teams in missions would be advantageous. Dr. Feigelson commented that data or science methodology, and software, needs to be a reviewable component of NASA missions, and reviewable at CDR and PDR. He further requested support by anecdote (from the 3 non-AP disciplines) of improvement of algorithms that lead to a significantly increased sensitivity of an instrument. Ideally, he wanted to have examples from all the SMD divisions. Dr. Holmes asked Drs. Feigelson and Hurlburt to investigate SIPs and DACs as well.

Dr. Kinter provided an anecdote on MERRA-like functions. The Weather Service did a reanalysis of variables using a computing resource that was available only briefly, so they did four 10-year segments with 2-year overlaps, and later discovered hysteresis effects in these parallel-in-time runs. The results contained jump discontinuities as an unintended consequence. Dr. Tito asked about the method development process at NASA: is the science method developed and handed off to software groups to implement? Dr. Hurlburt said the work typically gets sent off to a young scientist (in Heliophysics). Dr. Feigelson commented that the concept of algorithm development in many NASA environments is absent, except in Earth Science. It's an organic and sloppy process in the other disciplines. Dr. Kinter noted that climate models are tuned, and tuning means the data available constrains the system, which introduces a difficulty in determining what to calibrate against. Typically, the models are tuned to reproduce a time series. Models have varying sensitivity to aerosol concentrations based on data in the 20th Century record; the tuning steps are not documented in peer-reviewed papers. Now people are talking about provenance, to document what steps the modeling versions have taken. NCAR will be doing this. Dr. Walker noted that PDS keeps all versions of the data and documents the changes. Dr. Tito felt that the documentation of the software engineering process along with science process should be consolidated together. Dr. Feigelson commented that in the AP proposal process at Goddard, when finalizing mission design, proposers begin with science requirements, which flow down to engineering requirements, which flow down to budget/schedule in a well-organized fashion. Afterward, writing software is done to implement the specifications, and this

tends to be done internally, within the mission team, and with little to no oversight. It is not done as well as the proposal process.

Dr. Hurlburt noted that a NASA mission's primary task, typically, is to get the hardware working, while data is a secondary consideration. Dr. Mentzel thought that "software as instrument" was a good metaphor, but not perfect; agile software development allows for a lighter weight and more interconnected product between science and engineering. There are individuals who can do both, but in any case, ideally a mission should have engineers embedded in the science group. Dr. Hurlburt referenced a recent presentation at ADASS arguing that hardware design is detailed but most of the effort is in the build, while in the software arena, design is everything and it is slow and hard. Dr. Tito cautioned that poor implementation of agile software methodology process can often used as an excuse for poor performance. A mission still needs a robust picture of what the science is trying to accomplish. Dr. Feigelson noted that mission proposals often describe the critical science, and then hone in on capabilities to achieve science goals. Data analysis focuses more on the pipeline, and Level 1 and requirements, and is usually competent. Science analysis, on other hand (creating the science, Level 3 and 4 requirements) is more problematic; NASA is involved in both analyses in terms of algorithms and software. Dr. Mentzel felt that deep learning belongs under this study topic, and was not hearing it in the discussion.

Dr. Holmes asked that the BDTF report include a look at data policies for the four divisions. Dr. Beebe remarked on Dr. Mentzel's prior comment on the lack of consideration of science reduction in mission proposal reviews. She felt that in Planetary, science reduction was done very carefully. Dr. Feigelson seconded this observation, but felt, however, that phase PDR/CDR tends to be less rigorously reviewed in terms of data science. Dr. Holmes agreed. Dr. Hurlburt commented that a recommendation to strengthen the phase B milestone review will help to push the "wet noodle."

Public comment period
No comments were noted.

Findings and recommendations
Dr. Holmes asked everyone to send him bullets on the progress of their respective research projects. He felt that recommendations were not quite ready to present, and asked that BDTF plan to hammer them out at the next meeting. He felt that the overall assessment of the briefings on NASA's computing programs could be summed up as: "doing a great job, but." Drs. Walker and Tino agreed to flesh out this assessment. Dr. Holmes also supported developing a recommendation on the data science program, and wanted to submit text pages to the Science Committee on the data program. He thought that a simple recommendation based on Paul Messina's program briefing could pave way for NASA to start materially participating in project.

Commenting on Dr. Lee's discussion on a survey of federal HEC capabilities, he proposed sending it to Dr. Dahlburg to present to the Science Committee as simple narrative. Dr. Feigelson suggested that a recommendation might lead to an increase in HEC. Dr. Holmes felt that HEC access is a policy issue, but that BDTF could support a case by providing the technical background. Dr. Tito proposed that BDTF request more information from Dr. Lee on how to improve capacity/utilization, and compute efficiency. Dr. Beebe felt the Task Force should commend Dr. Lee on his awareness of and proactive stance on HEC.

Dr. Hurlburt volunteered to work up an assessment of NASA's data program. Dr. Holmes agreed to draft a write-up on NSF's Big Data Hubs program, first as a recommendation to broadcast its existence to the NASA research community, and a second recommendation that NASA consider joint solicitation with NSF. Dr. Holmes also agreed to develop a finding on NASA's data science program.
Dr. Tito suggested a finding on technical staffing for the HEC program.

Outlines of white papers
BDTF members briefly summarized the outlines of white paper topics. In general, the papers were to be four pages, containing an executive summary, introduction (statement of problem), facts and figures/summary of examples and demonstrations, recommended approach/list of dos and don'ts, list of recommendations, and conclusion.

Mr. Smith and Dr. Holmes made tentative plans for the next meeting, tentatively to be a teleconference in June/July, and a final meeting at JPL in October. Dr. Holmes suggested a wrap-up session at the AGU meeting during the second week of December. Drs. Hurlburt and Kinter took an action item to submit a proposal for an AGU session on the results of the BDTF effort, to highlight the subject of Big Data.

BDTF reviewed Dr. Tito's finding on Lee's HEC program and rearranged some wording.

Dr. Holmes adjourned the meeting at 3:18 pm.

Appendix A
Attendees

Ad Hoc Big Data Task Force Members
Charles P. Holmes, ***Chair***, Big Data Task Force
Reta Beebe, New Mexico State University
Eric Feigelson, Pennsylvania State University
Neal Hurlburt, Lockheed Martin
James Kinter, George Mason University
Chris Mentzel, The Moore Foundation (pending membership)
Clayton Tino, Virtustream, Inc. (Webex)
Raymond Walker, University of California at Los Angeles
Gerald Smith, ***Executive Secretary***, NASA HQ

NASA Attendees
Joe Bredekamp, NASA HQ Retired
T. Jens Feeley, NASA HQ
Hashima Hasan, NASA HQ
Jeffrey Hayes, NASA HQ
Tsengdar Lee, NASA HQ
Thomas Morgan, NASA HQ
Kevin Murphy, NASA HQ
John Sprague, NASA OCIO, Big Data Working Group

Non-NASA Attendees
Judith Greco, SAE
Logen Johnson, SAE
Amy Reis, Ingenicomm
Pamela Tomski, SAS
Fen Zhao, NSF
Joan Zimmermann, Ingenicomm

Webex Attendees
Alberto Accomazzi, Harvard University
Brent Auble, LMI
Ryan Cobb, SAS
Melissa Cragin, Midwest Big Data Hub
Daniel Crichton, JPL NASA
Jill Dahlburg, NRL, NAC Heliophysics Advisory Committee
Lane Brent Forsythe, Mutatio, Inc.
David A. Imel, Caltech
James Lochner, USRA

Thomas McGlynn, NASA
Paul Messina, Argonne National Laboratory
Joshua Peek, STScI
David Peterson, LMI
Marc Postman, STScI
A. Riojas
Laura Viveck, ADA

Appendix B
Membership

Charles P. Holmes, Chair
Retired
Formerly at NASA HQ

Gerald Smith, Executive Secretary
Science Mission Directorate
NASA HQ

Reta F. Beebe
Professor, Department of Astronomy
New Mexico State University

Ashok Srivastava
Chief Data Scientist, Verizon Wireless
Verizon Development Labs

Eric D. Feigelson
Department of Astronomy and Physics
Pennsylvania State University

Clayton P. Tino
Software Architect, Virtustream Atlanta
Virtustream Incorporated

Neal E. Hurlburt
Research Science Manager,
Solar and Astrophysics Laboratory
Lockheed Martin Space Systems Company

Raymond J. Walker
Professor, Institute of Geophysics and
Planetary Physics
University of California, Los Angeles

James L. Kinter
Director, Center for Ocean-Land
Atmosphere Studies
George Mason University

Appendix C
Presentations

1. Department of Energy Exascale Computing Project; *Paul Messina*
2. Update of NASA HEC Capabilities, and Federal Survey of HEC Capabilities; *Tsengdar Lee*
3. NASA Big Data Working Group; *John Sprague*
4. Space Telescope Science Institute: Detecting the Unexpected; *Joshua Peek*
5. Big Data and Science at the Jet Propulsion Laboratory; *Daniel Crichton, Richard Doyle*
6. Big Data Regional Innovation Hubs and Spokes; *Fen Zhao*

Appendix D
Agenda

Ad Hoc Big Data Task Force
of the
NASA Advisory Council Science Committee

March 6-7, 2017

NASA Headquarters
300 E Street, SW Washington DC
Room 5H41-A

Agenda
(Eastern Standard Time)

Monday, March 6

9:00 – 9:45	Opening Remarks / Introduction	Dr. Charles Holmes Mr. Gerald Smith
9:45 – 10:45	DOE Exascale Computing Project (ECP)	Dr. Paul Messina
10:45 – 11:00	***BREAK***	
11:00 – 11:30	NASA Big Data Working Group	Mr. John Sprague Dr. Tsengdar Lee
11:30 – 12:00	Update NASA High End Computing (HEC) & Survey of Federal HEC capabilities	Dr. Tsengdar Lee
12:00 – 13:00	***LUNCH*** ***12:30-12:50 Highlights of the "Detecting the Unexpected" Workshop at Space Telescope Science Institute***	Dr. Josh Peek
13:00 – 13:05	Public Comment	
13:05 – 14:45	Science Mission Directorate (SMD) Program Office Panel	Dr. Hashima Hasan Dr. Jeffrey Hayes Dr. Thomas Morgan Mr. Kevin Murphy
14:45 – 15:00	***BREAK***	
15:00 – 16:00	Data Science program at the Jet Propulsion	Mr. Daniel Crichton

	Laboratory (JPL)	
16:00 – 17:00	Discussion	Membership
17:00	***ADJOURN FOR DAY 1***	

Tuesday, March 7

9:00 – 9:45	Member Reports	Membership
9:45 – 10:45	Progress on research on the Big Data Task Force (BDTF) study topics – Part 1	Membership
10:45 – 11:00	***BREAK***	
11:00 – 12:00	National Science Foundation (NSF) Big Data Innovation Hubs and Spokes Project	Dr. Fen Zhao
12:00 – 13:00	***LUNCH***	
13:00 – 13:05	Public Comment	
13:05 – 14:45	Develop findings and recommendations for the Science Committee to include: • Overall assessment of SMD's data and computing programs. • Statement on NSF's Big Data Innovation Hubs and Spokes Project.	Membership
14:45 – 15:00	***BREAK***	
15:00 – 16:00	BDTF study topics. Present outlines of the white papers – Part 2	Membership
16:00 – 17:00	Final Discussion/Next Meeting/Conclusion	Membership
17:00	***ADJOURN FOR DAY 2***	