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

Meeting Minutes

June 22-23, 2017

Charles P. Holmes

Charles P. Holmes, Chair

Gerald S. Smith, Executive Secretary

Table of Contents

Report prepared by Elizabeth Sheley
Ingenicomm, Inc.

June 22, 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 teleconference to order and reviewed the Federal Advisory Committee Act (FACA) rules. He then introduced Dr. Charles Holmes, Chair of the BDTF.

Dr. Holmes explained that the purpose of the teleconference was to review draft documents the Task Force had developed. These documents fell into two categories: draft recommendations, and study topics with program reports. A goal of the meeting was to finalize and approve all the draft recommendations.

FY18 Budget Update - SMD's Data Archives and HEC

Mr. Craig Tupper of NASA's Science Mission Directorate (SMD) presented program highlights and budget information. The Fiscal Year 2017 (FY18) President's Budget Request (PBR) for SMD supports full funding for the Europa Clipper mission, as well as an SMD-wide cubesat/smallsat initiative, formulation for the Wide Field InfraRed Space Telescope (WFIRST), and a number of mission launches. In addition, operating missions continue to be supported with the exception of the NASA portion of DSCOVR. Space weather and Science, Technology, Engineering, and Math (STEM) activities are funded as well. There are some proposed cuts within the Earth Science Division (ESD), but the Planetary Science Division (PSD) will receive an increased budget, while funding for the Heliophysics Division (HPD), the Astrophysics Division (APD), and the James Webb Space Telescope (JWST) are flat.

Mr. Tupper next reviewed each division's data funding. PSD's Planetary Data System (PDS) and Science Data and Computing budgets are essentially flat, as is APD's Astrophysics Data Curation and Archival budget. HPD's Space Physics Data Archive and Solar Data Center are flat as well. ESD's budget covers SMD's High End Computing (HEC) capability and Scientific Computing. The HEC budget appears to drop in the FY18 PBR. However, the lower figure results from a transfer of $15 million to support the modular super-computing construction at Ames Research Center (ARC). While this allocation is outside of the science budget, it funds construction in support of science.

Turning back to the ESD budget, the Multi-Mission Operations line includes some computing. A decrease in the Earth Observing System Data and Information System (EOSDIS) line and a new line for Making Earth Science Data Records for Use in Research Environments (MEaSUREs) gave the appearance that MEaSUREs was being broken out from EOSDIS, which was to be confirmed later in the meeting. Mr. Tupper pointed out that all of the reductions in the ESD budget came from flight programs and the Research and Analysis (R&A) program. The Congressional mark-up was ongoing at the time of the meeting, so it was unclear when there might be an appropriations bill, though July was a possibility.

Member Reports

Dr. Holmes reported that he had visited a commercial cloud specialist at the Google offices in New York, where they discussed commercial cloud services. Commercial operators are evolving their fee structures, which may make the commercial sector more attractive for archiving

NASA's science data. Dr. Holmes also received some good questions and comments at the April meeting of the NAC Science Committee (SC), which accepted the BDTF finding on the upgrades to NASA's HEC capabilities. The SC discussed Dr. Tsengdar Lee's survey of Federal high-end computing (HEC) capabilities, and the Heliophysics Advisory Committee (HPAC) will follow up once it is able to meet. Dr. Holmes would like the four science advisory committees to discuss the issue of sufficient capacity for NASA's HEC capabilities, as he does not see that as falling to BDTF and considers the capacity topic as more of a policy issue. Whereas the BDTF can follow the technical approaches for achieving the desired capacities.

The fall American Geophysical Union (AGU) meeting will be in mid-December, and BDTF has been approved to present a session, "Big Data Approaches for NASA Science Missions." The deadline for submission of abstracts is August 2. Dr. Holmes listed some of the complementary AGU sessions. Related to this is the AGU Space Physics and Aeronomy e-newsletter of May 2, 2017, which included a call placed by Dr. Lika Guhathakurta, currently on detail at ARC, for applicants to the NASA Frontier Development Lab (FDL). The lab will bring physical scientists together with specialists in data science and machine learning. The FDL was conducting a space weather workshop for the following week.

Dr. Chris Mentzel reported next, explaining that his program is doing an evaluation that will produce relevant findings in regard to the education and training of software engineers for the research sciences. These are researchers who produce working software. He also had done some work with Google's Earth Engine, a platform for analyzing Earth observation system data and computing with scalable data. The system provides access to publicly available data sets, and has a capacity to broaden the scale-out. Dr. Reta Beebe reported that, since the last meeting, she has been dealing with the close-out and archiving of metadata. At the same time, NASA is pressuring teams to use new archiving methods. This has been a full-time commitment and has not left her much time to think about to think about big data.

Dr. James Kinter said that he had attended several meetings that had some relevance to BDTF. (1) There is a multi-agency effort involving NASA, the National Oceanic and Atmospheric Administration (NOAA), and the National Center for Atmospheric Research (NCAR) that will generate a lot more data from climate models. This initiative is part of an international program called the Coupled Model Intercomparison Project, which is in its 6th generation (CMIP6). The CMIP6 international coordinating body has defined a set of protocols that all modeling groups are expected to apply to their models, and the output of those numerical experiments will be shared in data repositories around the world, coordinated by the Earth System Grid Federation (ESGF). He noted that planning for CMIP6 began 2 years ago, and the protocols appeared in a collection of peer-reviewed journal articles (*Geoscientific Model Development*). The groups are now running experiments, which they expect to take a year, at which point the data will be available through distributed archives. (2) The U.S. National Weather Service (NWS) is reaching out to partner agencies like NASA, including Goddard Space Flight Center (GSFC) that operates the Joint Center for Satellite Data Assimilation with NWS and NOAA. The NWS effort, called the Next Generation Global Prediction System project, aims to increase high-performance computing capabilities and modeling at all time scales, driving global prediction to 10-kilometer grid spacing or less. This will greatly increase the corresponding data volumes. The NWS initiative also engages working groups from various agencies and universities, including an

infrastructure working group on sharing code and data, and a software architecture group thinking about coupling technologies to make the Earth observing components talk to each other more efficiently.

A third meeting that Dr. Kinter attended in Nanjing also has implications for Big Data. China has an aggressive program to build information sciences, weather prediction and climate simulation capabilities. For example, China will have seven models participating in the CMIP6 climate projection experiments, which is the same number of models as the United States. With the increases in number of models, number of experiments and resolution of models, archives will balloon as much as fivefold (from generation 5 to generation 6 of CMIP, for example) and the data sets will be too big to move. Dr. Kinter has been gathering information about modeling work flows, and noted that a lot of people are thinking about the issues to be discussed at BDTF's AGU session.

Mr. Smith reported that NASA was moving to a risk management format in order to respond to a White House initiative involving IT security. The Agency is assessing its status in this regard, while also looking at how to enable access with a type of guest credential. The Office of Management and Budget (OMB) has made this a priority.

Dr. Neal Hurlburt said that he submitted a notice for the BDTF's AGU session to Solar News, to appear in the next issue. He looked at a design review for solar physics and ground data regarding calibration and other concerns, and developed a concept of a private cloud that could be pushed out to the larger cloud. The Frontiers Development Lab will be supporting this concept, and the group he is involved with is addressing migration of data stores.

Dr. Raymond Walker explained that he had discussed BDTF at an April meeting of NASA's internal big data group, receiving lots of comments. Those who run large data stores were interested in the commercial cloud but had misgivings about costing and where that might lead. There was a very interesting presentation on data access by a group at the Jet Propulsion Lab (JPL) regarding new ways to access data. In May, there was a meeting on data modeling and simulations with the goal of finding how to address the needs of the next generation of simulations, and what to do with the output. Another issue is that those outside the Agency who work to provide NASA systems are being affected by a security issue that requires resources, essentially an unfunded mandate that also involves the Department of Homeland Security (DHS). It is increasingly difficult for those members who are not U.S. citizens to access data. A similar situation exists in Japan. Dr. Holmes suggested that he, Mr. Smith, and Drs. Walker and Beebe discuss this offline.

Dr. Eric Feigelson was on sabbatical at the National Science Foundation's (NSF's) Statistical and Applied Mathematical Sciences Institute (SAMSI), which brings mathematicians and statisticians together with scientists. A recent workshop included astrophysicists and engineers. The goals seem to be to delve into petabytes of noise and pull out faint signals and other data, and to classify a billion variable items from poorly sampled survey data. The focus was on methodology.

Dr. Holmes reported the communications that had already occurred for the AGU session and asked if anyone could reach out to the planetary communities; Dr. Beebe volunteered. Dr. Holmes emphasized that the abstracts were due in 5 weeks. Once the community was alerted, he wanted BDTF to submit at least four papers, one for each of the task force topics. Each paper must have at least one co-author who is an AGU member, which he did not expect to be a problem since most BDTF members belonged to the organization. Dr. Holmes also sought a presentation of at least 20 minutes on astrophysics. He hoped to invite a speaker on another topic and asked BDTF members to send in suggestions.

Dr. Beebe expressed concern about presenting BDTF products before they are published. Dr. Holmes explained that the materials would be on the Task Force website in November, before the AGU meeting. In addition, there is information about the white papers' contents in the minutes, which are public and which will be available before the meeting. He would still like to publish beforehand, but the information will be available regardless. BDTF will select the abstracts to be included. AGU will put the session on the schedule once that is done.

List of Draft Task Force Findings Recommendations for the Science Committee

The discussion of findings began with Dr. Holmes' draft on participation in NSF's Big Data Regional Innovation Hubs and Spokes. This is an important network in addressing science problems, and participation would benefit NASA's research community. The recommendation was that: 1. NASA/SMD use the NSPIRES messaging list to alert and inform its research community about this program and the benefits of working with participants in discussing data analysis problems and, 2. NASA investigate the possibilities of a joint research solicitation with NSF to address problems in analyzing the large and complex data sets that result from NASA's science missions. SMD should appoint a program scientist to contact the NSF project officer, Dr. Fen Zhao, to help develop messaging and to initiate joint solicitation discussions.

Dr. Holmes reported that Dr. Max Bernstein, who publishes SMD's research solicitations (ROSES), agreed that this would be a good idea. The messaging would require some text, and a joint SMD-NSF solicitation has precedent and is not uncommon, though a bit more involved. The consequences of no actions are that both communities would miss opportunities to have modern data science techniques applied to finding new science results in NASA's large and complex science data holdings.

Dr. Feigelson advised checking with NSF to ensure that the plan is consistent with the wording of their announcements and budgetary restrictions. His specific concern was that NSF rules for this project prohibit the funding of academic salaries. While he understood that NASA and NSF have collaborated in the past, he wanted to note that NSF prohibits some expenditures that NASA views as standard. He sought confirmation of what was and was not possible.

Public Comment

Members of the public were given the opportunity to speak, but no one came forward to do so.

List of Draft Task Force Findings Recommendations for the Science Committee continued

Dr. Feigelson reiterated his concerns, noting that NASA's Research Opportunities in Space and Earth Sciences (ROSES) solicitations only allow Principal Investigators (PIs) to ask for

informatics on archived missions. This seemed to limit flexibility for proposals involving math, statistics, and computer science, and Dr. Feigelson wondered if this draft recommendation could make a clear statement that data scientists should have funds for extra- and intramural programs with few restrictions. He worried that BDTF might make two recommendations for two different grant programs – a highly restricted one involving NSF, and an open-ended one under SMD. Dr. Holmes decided to reserve the recommendation for the next day.

The next recommendation addressed the Department of Energy (DOE) Exascale Computing Project (ECP), advising SMD to continue its participation permitting NASA inputs to the ECP leaders. Increased NASA involvement would be helpful, especially if the HEC programs grow as anticipated with modeling and simulations. It was noted that the ECP team is still wondering how to deal with the volume of data, so perhaps this recommendation should be tied to big data directly. There are new techniques for capturing simulation data, which should also be mentioned. In addition, there is growing use of a technique that captures intermediate data from simulations, which is particularly relevant for exascales. Dr. Mentzel volunteered to write some additional text to circulate, and Dr. Kinter offered to provide an Earth science example, while another member suggested one for solar.

Next, BDTF discussed the recommendation to join the Science Data Super Highway. Dr. Holmes provided some background on NASA's relatively limited involvement in this area despite a number of previous recommendations, from multiple sources, that the Agency start participating in the superhighway with some urgency. The result is that NASA has not kept up, and BDTF specifically noted a lack of focused activity or awareness on the part of SMD. Because NASA has not kept track of what the science community needs, the Task Force was recommending that SMD establish a 2- to 3-year temporary position to focus on big data networks. Having a dedicated manager would allow NASA to leverage activities with other Federal agencies, allied services, etc., while also determining what is needed and available for extending the nation's science data superhighway to many of NASA's sponsored research groups. The officer would sponsor a program involving a ROSES solicitation so that research groups can propose setting up connections to larger campus and national science data networks. There is also a need to reserve funds for the Agency's Communications Services Office under the OCIO to provide necessary links to bridge to junction points in the high-speed national networks. Other parts of the Federal government have invested heavily in upgrading their science data transmission capabilities.

Dr. Walker observed that this is only part of the answer, but it is necessary. Dr. Kinter expressed concern about the "haves" and "have-nots," noting that some groups will continue to be left out, and it was not clear how to address that problem. For example, he was not sure how he would differentiate his campus from another campus with the same needs. This is unlike science proposals that can be judged on their merit. Dr. Holmes cited a successful grants program during his NASA tenure that supported the costs of refreshing hardware and software. Dr. Beebe pointed out that multiple BDTF recommendations called for an additional person. She wanted to be clear that this could be one or two people instead of a new manager for each recommendation.

The group moved on to discuss Dr. Hurlburt's assessment of approaches of the SMD data programs, projects, and centers to future challenges. The BDTF had posed three questions to managers of NASA's major data center activities. Specifically, the questions asked what

processes they had in place for planning for future capabilities, what services they would like to stop, and what steps they are taking toward interoperability both inside and outside of NASA.

When preparing for the future, much of SMD relies on the results from Senior Reviews. The level of effort also scales to the size of communities. Dr. Hurlburt described the interests in data centers for cloud computing, the rapidity of the changes, and the question of what to do with old data. There is a disparity among the disciplines regarding the means to discover and use the data appropriately. For example, ESD and PSD have distributed data management, while HPD and APD have different needs that are more focused. The latter two divisions have more papers coming out of new data and need more uniform data that can be accessed more easily. Dr. Hurlburt was still crafting a recommendation.

Dr. Holmes said that he was seeing the details but not the big picture here, and sought a clearer message. Dr. Feigelson suggested that the methodology white paper had material that could be used in a summary. However, he wondered if BDTF should be so straightforwardly negative about NASA's past performance in this area. Should the executive summary be explicitly negative in order to wake up the readers and tell them that this is an area in which NASA has not kept up? Finally, while this piece provided no conclusion, the title implied one. He agreed to hold this comment so that Dr. Hurlburt could revise the title.

Dr. Holmes had an issue with disjointed thoughts, and Dr. Beebe said that an outline of the whole report should show how everything fits together, which was not evident in these drafts. Dr. Mentzel volunteered to do some revisions to this piece overnight, and Dr. Holmes decided that BDTF would revisit the document the next day.

Next was the SMD Data Science Research Program. This document pointed out that SMD controls a lot of data, and there are activities in industry and academia that develop new techniques that could be relevant. However, the NASA research community has not embraced these. There should be a permanent SMD position of data program scientist to address many functions that are not currently covered. This person would direct a data science applications program, convene workshops on the application of data science methodologies and analysis, participate in NASA's mission development process, and interact with others at NASA to implement data science methodologies.

Dr. Holmes noted that most of a mission's funds are spent in implementation phases. The data scientist will help advise the mission scientists. This person will review draft AOs, assist mission program scientists in nominating prominent data scientists for the science review panels, and work with program scientists to ensure that appropriate data science methodologies are incorporated and reviewed. Dr. Holmes noted that the data scientist could be recruited through the Intergovernmental Personnel Act (IPA) for the first year or two. He did not see this as an add-on responsibility for another manager. Dr. Feigelson thought this was BDTF's most important recommendation. Dr. Beebe advised including a concise summary of expected accomplishments. Dr. Holmes said he would review the piece overnight.

<u>BDTF Study Topics – Progress Reports and Drafts</u>

Dr. Beebe presented the first white paper, which addressed data availability and accessibility. She described the paper as something to critique, and sought input from the Task Force members. The white paper described each of the SMD divisions in terms of their data access. While the material for HPD seemed complete, the other three divisions could use more examples and approaches. Dr. Beebe was also concerned that the paper was overly broad. Regarding division-level data policies, Dr. Holmes offered to help her and Dr. Walker find them.

Drs. Feigelson, Hurlburt, and Mentzel wrote the methodology white paper, which also needed some work. Dr. Feigelson thought it might be too detailed. The paper was to recommend ways to improve methodologies for across all science domains; develop software architectures and tools for data analyses on distributed or cloud-based systems; and formulate strategies to make better use of advances in data science. The overall conclusion was that NASA science operations are not uniformly integrating data science advances, and NASA no longer leads in statistical and computational methodologies for space science.

Dr. Holmes thought some of this material should move to another recommendation, which he would handle. Dr. Beebe advised making sure that the paper's recommendations fit NASA's charter. Dr. Feigelson pointed out that the issue of which computations are needed goes beyond just volume and into the variety. NASA's culture is oriented to improvements in instrumentation, but data needs are less well recognized, leading to brilliant instruments that employ old-fashioned data approaches. Dr. Holmes liked that point and suggested emphasis on the need for improved analysis techniques.

Dr. Feigelson reviewed the background section, which addressed the fourth paradigm, in which the computer becomes the instrument rather than the camera or telescope. The paper cited a range of sources regarding this need for informatics. For example, NSF has emphasized bringing computer science into other sciences and has opened up entire programs in this area. NASA has not. Dr. Beebe suggested using the phrase "training the NASA cadre," which resonated in a prior effort. Dr. Hurlburt noted five bullets that detailed what data science encompasses. He welcomed additional description. The "statement of the problem" section pointed out that while enormous strides are being made, NASA has not kept up, resulting in uneven quality in data and scientific analysis products from NASA science missions. The Agency should have high standards, but instead there are no standards for mission analysis pipelines, etc. He has seen brilliant work and terrible work as a result.

Dr. Feigelson added that in the course of mission development, there is no evaluation of the quality of the data analysis. The scientific goals do not include data analysis methodology. Dr. Mentzel said that everyone should be concerned with making science efficient in this area. Efficiency and effectiveness could be a compelling argument, especially given the funding situation. They could include areas that verge on data engineering and that can provide much greater efficiencies in the data analysis pipeline. Dr. Feigelson agreed, noting that software is written inefficiently in many cases and a software engineer brought in early could help tremendously. As for CPU usage, that can vary a great deal. Some of these jobs could entail highly efficient algorithms that the scientists do not know about. This improvement could save money and enhance efficiency. Space scientists are often poorly trained for informatics, so that they are smart but insufficiently informed, leading to inefficiency.

The recommended approaches began with training workshops, which NASA does not fund very much despite the potential benefits. Another recommended approach was NSF's Spokes and Hubs program. There was also considerable discussion of software engineering, maintenance, and related techniques that warrant focus. Approaches that had yet to be developed were on algorithms, deep learning, and optimization, and visualization.

The authors generated several overarching conclusions:

- NASA should promote professional development in statistics and informatics for scientists and engineers.
- Software development for space sciences should be professionalized so that it is not left to the astronomers; this should be specified in calls for proposals.
- Software engineering should be a cross-directorate effort and employed at the early stages of design.

Dr. Walker pointed out that a lot of the software development techniques have been used widely in heliophysics and planetary science. He volunteered to write something on this for the paper. Dr. Holmes assigned various BDTF members to work on specific sections.

Dr. Kinter presented the workflow white paper, which was also incomplete and for which he and his co-authors sought input. It began with an introduction that served as a primer on modeling, especially in earth science, and described large modeling projects as examples. The paper needed more on modeling issues in other areas, however. There was a discussion of challenges, such as computing and hardware not keeping pace, etc. The third section stated the problem, which was that the existing situation results in workflows that are inefficient and insufficient. Features of the problem include the inability to bring data into models; the massive data volume; reproducibility, which is a growing issue because HPC runs no longer produce the same results due to parallelism and potential for uncorrectable errors in a small subset of the O(million) chips; data compression; and inherent performance limitations of standard data storage formats. Dr. Feigelson mentioned compressed sensing, a new solution that had not yet been promulgated in astronomy. Dr. Hurlburt planned to contribute language on this.

Dr. Kinter presented the next section, which provided examples of approaches, such as prototypes and demonstrations. As recommended approaches, the document cited three types of investments: hardware, software, and human resources (includes institutional cultural change). There was also a point about infrastructure design issues. The document then noted that having a specific project objective and timeframe can help set requirements and drive development of workflow improvements. There was also a list of longer-range investments for innovative capabilities. Data accessibility is an issue that affects a number of areas, including modeling.

Adjournment for Day 1

The meeting adjourned for the day at 5:55 p.m.

Friday, June 23, 2017

Finalize Task Force Findings and Recommendations for the Science Committee

Mr. Smith called the meeting to order and reviewed the FACA rules. Dr. Clayton Tino was present for the teleconference that day. Dr. Holmes reviewed the agenda and the plan for the day, which began with finalizing the draft findings and recommendations, moved on to program reports, and ended with the white papers.

Dr. Holmes reviewed the paper on the data superhighway, which had not changed much since the previous day. BDTF found that there is little knowledge within SMD of the recommendations authored by the NAC's IT Infrastructure Committee (ITIC). The recommendation under discussion was for a temporary position to focus on the data superhighway and build on existing infrastructure. NSF has provided grants to hundreds of universities nationwide to upgrade their internal campus systems in an ongoing Federal investment, and the groups already in have benefited tremendously. Many NASA-affiliated research groups are not yet involved, however, so SMD needs to provide guidance to ensure this activity occurs. Failure to do so will result in NASA falling even further behind despite the growing fleet of missions, satellites, and other operations.

Dr. Tino questioned the use of the DMZ acronym, which connotes battlegrounds. Dr. Holmes said that it has become a term of art used by DOE and others. He agreed that a brief definition in a footnote would be helpful. Dr. Feigelson thought the recommendation was quite general, and Dr. Holmes clarified that the recommendation addresses research groups rather than the NASA centers. Dr. Tino wondered if the concern might be more that NASA is not being proactive, because research groups could participate through other avenues. He thought the paper should distinguish between NASA being left out and NASA science suffering due to lack of engagement. Dr. Holmes agreed to address this, then send the corrections around via email.

Dr. Hurlburt explained that on the assessment of SMD data centers, only the last paragraph had changed. Dr. Feigelson observed that APD seems to emphasize this area more than the other divisions. It can take a long time because some of the work is agreed upon at the international level. However, once it is done, a very powerful archival science environment emerges that can be quite valuable scientifically. Dr. Walker noted that within the silos, there is uniform metadata and it is positive. The real question is going perpendicular to the silos. He thought NASA was better off than the paper implied. Dr. Holmes suggested a modification to say that NASA needs to continue its evolution of the metadata standards, with future goals in mind.

Dr. Holmes thought the recommendation was soft, and suggested that BDTF defer further discussion to the next meeting while continuing to work on the paper. The assessment is that NASA is doing a good job with its resources, and there will be challenges ahead. The paper then offers some ideas. He asked Dr. Tino to help Dr. Hurlburt. Dr. Kinter made two points regarding Earth science. First, there needs to be a purpose to cross-silo interoperability, and ESD has data archives and old datasets that might not warrant the amount of work required to make them interoperable. On the other hand, there is evidence that cross-silo interoperability can be successful in areas like four-dimensional output of climate modeling.

The next paper discussed SMD's data science research program, which Dr. Kinter had tweaked. Dr. Holmes noted that the recommendation points out that NASA data are continuing to increase

in volume and complexity. NASA research programs should blend science expertise with data approaches, and SMD should have a program officer in this area. The data program scientist will direct an annual research program that is solicited via ROSES, convene workshops on applying data science to NASA science mission data analysis problems, participate in the mission development process, and interface with NASA offices regarding data science methodologies.

In answer to a question, Dr. Holmes explained that the solicitation funding would go to those eligible to apply to ROSES. It would not go to Phase A of a science mission. The data archives would be a policy issue for the program officer to work out with management. Dr. Feigelson saw this as the most important recommendation, so he was concerned that it was less detailed than the others, and that much of the background was in the methodology white paper. Dr. Holmes thought it would be appropriate to add a footnote to that effect.

Dr. Beebe cautioned that the last time NASA had funding in this area, the major complaint was that the research did not feed back into the methodology. She did not want to see that come up again. Dr. Feigelson suggested that the current environment is very different, with specialized repositories. He wanted to write a paragraph about that, and Dr. Beebe agreed it was important. There will need to be a structure, and it must be properly presented in the AO. Dr. Holmes said that he would add a responsibility to procure a feedback loop of products and grants. Drs. Mentzer and Beebe advocated for a mention of standards and reviews. SMD does not have a general standard for usability. Dr. Holmes said he would edit the piece and send it out for additional feedback.

For discussion of the Hub and Spokes recommendation, Dr. Fen Zhao of NSF joined the teleconference. Dr. Holmes explained that in 2015, NSF created a series of awards targeting big data innovations. Members of the BDTF visited and interviewed many of the participants. The BDTF has concluded that this is a new and important resource. The recommendation had two parts. First was that SMD inform its research community via the NSPIRES messaging service that this NSF program is operating and that NASA PIs may benefit from making contact with the participants to discuss data analysis problems they face. The second part was that SMD should consider establishing a joint research solicitation with NSF to bring NASA research PIs together with regionals hubs and spokes to attack some of the daunting problems in analyzing large and complex data sets. The NASA PIs will benefit from learning and applying new and emerging methodologies, while NSF participants will be able to apply their craft to Earth and space science problems.

Dr. Feigelson raised his concern about funding. Specifically, his impression was that NASA has space scientists who could benefit from more contact with informatics and data science professionals, but it was not clear that the cross-disciplinary experts could be funded through the hub and spokes mechanism. Dr. Zhao said that a computer scientist can indeed receive funds from NSF for this type of work. The funding restrictions were meant to bring in working groups rather than the traditional “PI with a graduate student” arrangement. In addition, NSF cannot fund others from Federal agencies. An addendum would be required for a situation in which NASA funds the space scientists and NSF funds the computer personnel and applied mathematicians, but NSF has done that for other agencies. Dr. Zhao added that she was excited

to see the hubs engage with the NASA community, and she saw a lot of potential. NSF is ready to take this on.

NASA Data Science Program Updates

Dr. Holmes explained that BDTF had invited the SMD division program officers for the data archives to provide updates. Mr. Kevin Murphy of ESD began by explaining that the usual work of archiving data from field experiments has continued, the volume of which will increase significantly over the next 5 years. To mitigate the risks of the archive volume growth, ESD has begun some commercial prototyping activities. There are also preparatory activities for the NASA-Indian Space Agency (ISRO) Synthetic Aperture Radar (NISAR), including archive distribution activities from the Amazon Web Services (AWS). There is also an ongoing Cumulous project. The Division continue exercising the plan to evaluate Amazon's cloud services, having implemented a common metadata within AWS. This has optimized costs and shows improvements. In the fall, ESD hopes to see what other systems can be moved. The Earth Observing System Data and Information System (EOSDIS) Cloud Evolution (ExCEL) Project. serves as the umbrella and includes six or seven prototype efforts. ESD entered into a prototyping effort with AWS to demonstrate a 5-petabyte archive. The demonstration is going well and includes public distribution. Regarding Google's Earth Engine effort, ESD provides support as needed in a partnership. However, ESD provides services to all users. Last fall, the Division published a white paper with Amazon, Google, and Microsoft on how to move science archives into their systems. The metadata standards that ESD uses are in ISO 19155-2, and the Division works with a range of external organizations to define the standards further. To a certain extent, these are used internationally. One of the most effective ways to adopt these standards is through the efforts of a data working group. MEaSUREs is a competitive processing activity, as opposed to data systems. Mr. Murphy explained that the core systems are covered under the Multi Mission Operations budget line, while MEaSUREs concerns long-term climate data records and EOSDIS encompasses the access program, citizen science, and other activities.

Dr. Jeffrey Hayes provided the HPD update. The two budget lines are small and have had little change. There was not much else to report other than some personnel changes. He now has a deputy, Dr. Heather Futrell, who will take over the work on the archives. Dr. Joe Gurman is retiring, and there was a candidate to fill his position. Dr. Gurman is working with GSFC and Stanford to create a permanent archive for Solar Dynamics Observatory (SDO) data.

Dr. Tom Morgan of PSD said that that Division's greatest challenge at the moment was a modest funding shortfall. PSD completed its roadmap, though it was not yet on web pending approval of PSD. The report will have 19 findings regarding the PDS. The original plan to migrate data from PDS3 to PDS4 was to move conservatively, but the roadmap recommends speeding up the process. He wanted to defer other discussion until the roadmap's release.

Dr. Hashima Hasan of APD reported on a number of initiatives, some of which involve the Infrared Processing and Analysis Center (IPAC). IPAC would like to build on plans to use high-performance network connectivity, explore the application of machine-learning techniques and data integration to facilitate science discovery, investigate complex queries and data packaging at the Archive for Data Analysis and Image Processing, and initiate pilot projects to move portions of the archives to the cloud to determine the feasibility of an entirely cloud-hosted archive. The

funding is uncertain at this point, however, so it is not yet clear which of the work can be done in the near future. IPAC has already established connectivity to the Pacific Rim high-speed network. A user panel report indicated concern that availability of connectivity might limit the ability of researchers to use big data as needed. Dr. Hasan described a number of maps and catalogues generated by the Herschel mission. The Mikulski Archive for Space Telescopes (MAST) now has an input catalogue for the Transiting Exoplanet Survey Satellite (TESS) discoveries

The Neutron star Interior Composition Explorer (NICER) recently launched, and Dr. Hasan noted the expected High Energy Astrophysics Science Archive Research Center (HEASARC) milestones in support of the mission, beginning with the first software release of NICER tools in August, the first throughput of data by the end of August, another software release in September, and opening of the public archive 6 months after successful calibration. Regarding NASA Astronomical Virtual Observatories (NAVO), four members of the team gave presentations at an international meeting in Shanghai. Dr. Hasan reported on a number of efforts to update standards. The NASA/IPAC Extragalactic Database (NED) has begun updating its DBMS to improve performance and support enhanced spatial processing. Finally, the Astrophysics Data System (ADS) is moving ahead as planned and making good progress.

Dr. Holmes asked the four program officers to send their division's high-level metadata standards to Mr. Smith. Specifically, he was looking for the policies on setting requirements and standards, so that information could go into the white paper on data availability.

Dr. Lee provided an update on HEC activities. At the March BDTF meeting, Dr. Paul Messina of the DOE discussed the ECP. DOE has now awarded the next phase of the project to six vendors. SMD involvement has been hampered by budget issues but the goal is to participate. Dr. Lee noted the $15 million shift in funds from ESD to the new HEC facility. The modular computing center is being built as planned. The 2017 modules will be the same size as the FY16 modules but with three times the capacity. There is an aggressive over guide plan to integrate and operate the modules, which will be located at ARC, where the staff is sufficient and prepared to bring up the new modules. Distributing the computing center to multiple centers is not practical. The next phase of procurement will cover the next 10 years and enable new system acquisition. The drop in R&A for ESD in FY18 PBR will be a factor, and SMD is still working on that.

Public Comment

Members of the public were given the opportunity to speak, but no one came forward to do so.

Finalize Task Force Findings and Recommendations for the Science Committee

BDTF resumed discussion of the recommendation for continued participation in ECP. Dr. Walker had added an example for Earth science. The recommendation noted that as the project proceeds, NASA will need to continue participating and support modeling development activities to move into the next generation. In addition, there are serious problems that this increased capability will allow NASA to address. Lack of involvement would impede the Agency's progress on some significant science projects. Dr. Lee thought this was good, and suggested more specific language regarding what they hope to accomplish. Dr. Tino thought it was insufficiently detailed, and Dr. Holmes encouraged him to contribute to the piece. Dr. Lee

explained that a group he had organized wrote a position paper stating that the current system might not meet their needs in the future and calling for more bandwidth.

Dr. Holmes asked if SMD might be taking a path with ECP that could be hard to shift if, for example, a commercial group were to come up with something preferable. Dr. Lee said that ECP is working with all of the major vendors, so he did not foresee a surprise from the commercial side. The competition is from other countries, such as China. ECP is organized around different families of technologies. He did find the legacy requirements of the current vendors to be a potential concern, but he did not foresee a radical new design from elsewhere.

The BDTF members voted to approve the draft, and Dr. Holmes asked Dr. Walker to smooth out the language by the end of the next week so that he could circulate it to the NASA program officers and to Dr. Bradley Peterson of the NAC Science Committee.

BDTF Study Topics – Progress Reports and Drafts continued

Dr. Kinter received some feedback on the workflow white paper from Drs. Feigelson and Hurlburt, adding material on compressive sensing. Dr. Holmes also provided feedback and would pursue a contact. Dr. Kinter hoped to work more closely with his co-authors and wanted to add references while also improving the organization, particularly in the "recommended approaches" section. Dr. Hurlburt offered to provide a discussion of conversion of data from physics models in solar research. Dr. Kinter planned to add something similar from Earth science, in which a model outputs data as if it is an observation. Dr. Walker had a blurb on methodology. He suggested that there be some generalized problems that cross disciplines. With Dr. Kinter's approval, Dr. Holmes said he would send a subsequent draft to some modeling groups for feedback. He noted that most of other white papers address internal NASA processes. Dr. Kinter said that the server side analytics paper included relevant information; he and Dr. Holmes agreed to integrate and cross-reference some of the materials in the two papers.

Dr. Holmes said that he still had work to do on the server-side analytics paper. He then discussed the history of NASA data storage, mentioning physical media transfers. Currently, there is a lot of interest in changing the architecture to allow researchers to go through an analytics processor that will go to a database and process in a way that reduces the volume through the use of some elementary processes, known as "pre-processing." The idea is to allow a tailored processing capability appropriate to the type and use of the data served. This would be controlled by the user and result in a lower level of data transfer and storage by user. Examples of approaches included Google Earth and SciServer from Johns Hopkins and the Space Telescope Science Institute (STScI), but Dr. Holmes asked other Task Force members to provide more examples and case studies. He also sought help on recommended approaches.

He had been advised that the implementers should be on the science side, not the archives side. The rationale was that the standing groups are not as hungry as the science side. Dr. Hurlburt agreed that the scientists should implement, noting that one of the big issues is security. Dr. Mentzel described Google Earth Imaging as a platform for computing on the NASA data, noting that it is relatively safe with a secure access format that relates to a recommended approach. Dr. Kinter suggested encapsulating the idea of putting an Application Programming Interface (API) between the data access and the generator. There should also be a balance between data center

personnel and external sources in provision of the tools and applications; there must be a collaboration, which can be messy but which is necessary.

Final Discussion/Next Meeting/Conclusion

Dr. Holmes noted that almost everyone would be available for the planned November 1-3 meeting, with Dr. Feigelson participating by phone if he could not attend. There will be updates, a tour at JPL, and continued work on the papers and recommendations. The final Task Force meeting will be in New Orleans at the same time as the AGU conference. BDTF will have a session at AGU, followed by a half-day BDTF meeting to close out any unfinished business and officially shut down the group.

Mr. Smith noted that SMD is concerned about presenting the results before the minutes are approved and on the website. Dr. Holmes agreed to ensure that this timeline is met. He then asked each member for last thoughts regarding this meeting.

Dr. Feigelson thought they were doing well on the details and needed better integration to clarify the big picture. He advised making the documents similar in the name of coherence. Dr. Beebe said that she and Dr. Walker would work on their sections, seek input, and look at the other white papers to match the structure. Dr. Hurlburt thought they had done a good job and looked forward to having an overall structure. Dr. Mentzel agreed. Dr. Walker said that he would make his work on the project a priority and get it in quickly. Mr. Smith asked that the authors coordinate their documents.

Dr. Holmes explained that each author would be responsible for version control, and that this also applied to edits. The draft managers were to collect and distribute the most recent versions.

Adjourn for Day 2

The meeting was adjourned at 3:07 p.m.

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.
Raymond Walker, University of California at Los Angeles
Gerald Smith, ***Executive Secretary***, NASA HQ

NASA Attendees
Tsengdar Lee, NASA HQ
Thomas Morgan, NASA HQ
Craig Tupper, NASA HQ

Non-NASA Attendees
Elizabeth Sheley, Ingenicomm

Webex Attendees
Elaine Denning, NASA HQ
Heather Futrell, NASA HQ
Hashima Hasan, NASA HQ
Jeffrey Hayes, NASA HQ
Tsengdar Lee, NASA HQ
James Lochner, USRA
Brendan Marozas, Cornell University
Thomas Morgan, NASA HQ
Kevin Murphy, NASA HQ
Larry Roelofs, NASA Goddard
John Sprague, NASA OCIO
Fen Zhao, NSF

Appendix B
Membership

Charles P. Holmes, Chair
Retired
Formerly at NASA HQ

Gerald Smith, Executive Secretary
Science Mission Directorate
NASA Headquarters

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

Clayton P. Tino
Software Architect, Virtustream Atlanta
Virtustream Incorporated

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

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

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

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

Appendix C
Presentations

1. SMD FY2018 Budget Estimates; *Craig Tupper*
2. Chair's Report: 5th Meeting of the Big Data Task Force; *Charles Holmes*

Appendix D
Agenda

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

June 22-23, 2017

Teleconference/WebEx Meeting

Agenda
(Eastern Daylight Time)

Thursday, June 22nd

11:00 – 11:30	Opening Remarks/Introduction	Dr. Charles Holmes Mr. Gerald Smith
11:30 – 12:00	FY18 Budget Update – SMD's Data Archives and HEC	Mr. Craig Tupper
12:00 – 13:30	Member Reports	Membership
13:30 – 14:30	***Lunch***	
14:30 – 14:50	Summary of Task Force's Report to Science Committee in April 2017	Dr. Charles Holmes
14:55 – 15:00	Public Comment	
15:00 – 15:30	List of Draft Task Force Findings Recommendations for the Science Committee discussion (online collaboration)	Membership
15:30 – 16:00	***BREAK***	
16:00 – 18:00	BD Task Force Study Topics – progress reports and drafts	Membership

18:00 ***ADJOURN FOR DAY 1***

Friday, June 23rd

11:00 – 12:45	Finalize Task Force Findings and Recommendations for the Science Committee discussion (online collaboration)	Membership

12:45 – 13:00	***BREAK***	
13:00 – 14:00	NASA Data Science Program Updates	Dr. Michael New Dr. Hashima Hasan Dr. Jeffrey Hayes Mr. Kevin Murphy Dr. Tsengdar Lee
14:00 – 14:05	Public Comment	
14:05 – 14:30	Final Discussion/Next Meeting/Conclusion	Membership
14:30	***ADJOURN FOR DAY 2***	