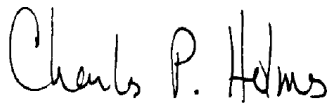


**Ad Hoc Big Data Task Force
of the
NASA Advisory Council Science Committee**

Meeting Minutes

**Inaugural Meeting
February 16, 2016
NASA Headquarters
Glennan Conference Room, 1Q39**



Charles P. Holmes, Chair



Erin C. Smith, Executive Secretary

NASA Advisory Council Ad Hoc Big Data Task Force, February 16, 2016

*Report prepared by Joan M. Zimmermann
Ingenicomm, Inc.*

Table of Contents

Introduction	3
Charter/Science Committee and Subcommittee Feedback	3
Legacy from NAC ITIC	4
Discussion	5
HPD Big Data	6
Science Committee Greetings	8
Big Data and Earth Science	9
Supercomputing and Big Data	10
APD and Big Data	11
Public comment	13
Other Federal Big Data Initiatives	13
Planetary Science Big Data	14
Discussion/wrap-up	15

Appendix A- Attendees

Appendix B- Membership roster

Appendix C- Presentations

Appendix D- Agenda

Introduction

Dr. Erin Smith, Executive Secretary of the NASA Advisory Council (NAC) Ad Hoc Big Data Task Force (BDTF), called the membership to order and made some administrative announcements. Dr. Charles Holmes, Chair of the BDTF, opened the inaugural meeting of the BDTF. Introductions were made around the table.

Charter/Subcommittee Feedback

Dr. Smith presented an overview of the Task Force, which was created in response to a number of White House directives on the Big Data concept, which related to the purviews of NASA's Heliophysics and Earth Sciences divisions (HPD and PSD), which engage in the study of solar activity and solar storms, and weather forecasting. The administration also expressed a great deal of interest in the interoperability of data sets, and related uses of Big Data. Successful applications of science in these areas will require the breakdown of subdiscipline stovepipes, and the interoperability of NASA data sets with those of the National Oceanic and Atmospheric Administration (NOAA) and the US Geological Survey (USGS), making data available to numerous end users such as emergency response and disaster relief agencies. Big Data may also enable the identification of actionable science information, making data useful for unforeseen applications. Big Data also means different things to different users, and for specific data-handling tools, data formats, and the creation of data standards. Applications vary for the Astrophysics (supernova models), Planetary (identifying exoplanets, galaxy formation), and Heliophysics divisions (one target/many missions, coronal mass ejections, radiation environment for human exploration). NASA's Earth Science Division has been managing and exploiting Big Data for many years in creating climate models, and for societal applications such as drought forecasting and disaster response. Many NASA spaceborne measurements are currently being used to improve air quality decision support systems in Texas, and in producing accurate cloud formation models. HPD data and engineering data are being fed into an Integrated Radiation Protection System, to help determine how to get to acceptable risk figures for radiation exposure in human exploration.

The terms of reference (TOR) for the BDTF form a broad charter, which can be described as examining what the community as a whole is doing in Big Data, as well as what other agencies are doing, and identifying what can be done better. The intent is to catalogue best practices in NASA and other federal agencies, as well as in private industry, research institutions, and academia. One of the final products may be a white paper reporting out findings and recommendations. A major challenge for the Task Force will be to define what the term 'big data' means to the various communities; to an astronomer it is an archive issue. To HPD and ESD, it is interoperability issues and engineering. Other challenges will be to determine the most useful and efficient architectures, storage modes, data accessibility, data rates, data security, and intellectual property requirements. How do we communicate what data sets are saying, and how do we train people in use of data sets? It is a dynamic area. To date, the BDTF has

completed its ethics training and is in the process of signing on its last two members to round out the committee.

The NAC Science Committee has provided feedback to the BDTF, namely to acquire more representation from commercial entities and other non-NASA sciences, as well as to consider ground-based sciences that may have produced scientific data; Feedback was also to look at data visualization; data permanence; and data usage. The Science Committee has asked that the BDTF act as a go-between for community, and to find links and leverage points with existing efforts on big data. The Science Committee also recommended that BDTF invite people from the NASA archives, NASA Ames Research Center, simulation experts, modelers, and industry partners. Within disciplines, practitioners should be able to understand themselves within their subfields, and to allow for cross-pollination between subfields. The BDTF has also been asked to find the best way to gather feedback so that the Science Committee and its subcommittees can benefit from this effort (survey to industry members, town halls, e.g.).

The NAC Science subcommittees would like the BDTF to address data usability, management and access, utilization (including real-time), analysis and data mining of large data sets, algorithm and statistics development, data curation, archiving tools and technology, visualization (such as hyperwall), and using state of the art information technology (IT) systems and tools. Other questions to address: What opportunities are there in big data? Which subject matter experts (SMEs) should be consulted? What kind of products are desirable?

Dr. Holmes noted that given the extensive shopping list, he wished to devise a work plan to use the limited time available, in order to distill the Task Force output into something valuable. As to the term “interoperability,” he challenged Dr. Smith to fine-tune this definition, as it is a wide-open topic. He believed that innovation comes from the bottom up, and worried that “interoperable” raises some red flags for the creation of top-down management. Dr. Clayton Tino worried about “needs for future use,” which would require a fundamental understanding of data formats; it is nearly a non-solvable problem to make data understandable to all communities. Dr. James Kinter commented that interoperability tends to become a catchall phrase for simulation and modeling, best practices, and interoperability between discipline scientists (including metadata and documentation). Dr. Reta Beebe noted that “data mining” connotes something magical and is a major question. Externally, people think that data mining is magically done. Data sets are so different, particularly in Planetary Science, that data mining becomes a major problem. Dr. Holmes reiterated his belief in the bottoms-up approach, and to allow successes from this approach to replicate through other scientific areas.

Legacy from NAC IT Infrastructure Committee

Dr. Holmes gave an overview of the BDTF’s history, having served as vice chair of the NAC Information Technology Infrastructure Committee (ITIC), which stood from 2010-2013. Its main affiliation was with the NASA Chief Information Officer (CIO), but it had ties across NASA as well, in areas such as cybersecurity. The NAC recommended that both the ITIC and the Science Committee explore an approach to improve access to

NASA science data repositories, with that exploration to include best practices, etc., that have been translated to the present TOR for the BDTF. In Fall 2013, the NAC advisory committee structure was revamped, cybersecurity was put under the aegis of a new committee, and the work of the former ITIC now continues with the current Big Data Task Force, reporting to the Science Committee.

One of the first recommendations of the former ITIC was that NASA should take advantage of assets in the Federal government, such as GPU clusters, cloud computing under the National Science Foundation (NSF), and other sponsorship. ITIC also recommended that NASA improve the cyber infrastructure that supports Agency science. One of the findings of the ITIC notes that NASA science data does not sit in one place but is distributed across NASA centers, at USGS, industry, and universities. NASA data centers are discipline-focused, and are managed in this way. The number of science publications coming out of these centers is growing dramatically. Education and Public Outreach continues to tap into these data stores, sometimes directly, and sometimes through a group that processes it for the general public. The Department of Energy (DOE) has set up a backbone throughout the country with many nodes not far from the NASA centers; it would be good to leverage this pipeline, as well as a 10-Gps network research that links research innovation laboratories.

Use of NASA supercomputers at both Goddard Space Flight Research Center (GSFC) and Ames Research Center (ARC) is growing. The Earth Observing System Data and Information System (EOS-DIS) is also growing in its data product distribution. Web services to support disaster applications, such as the Short-term Prediction Research and Transition (SPoRT) Center at Marshall, are transitioning research data to the operational weather community. The Solar Dynamics Observatory (SDO) is revolutionizing the way we understand the sun, and is collecting roughly a petabyte of data per year, with 5 petabytes per year worth of processing. There has been a two-order-of-magnitude jump in what solar physics had been ingesting previously from older missions such as Hinode. NASA's Multimission Archive at Space Telescope (MAST) is showing almost exponential growth, and which will grow even more when future telescope missions come on-line. There are 200-plus apps in the Apple iStore that will return from a search on NASA; many of these apps are in high demand from the public, and pull processed results out of NASA's data stores. More than 250,000 people have taken part in NASA's Galaxy Zoo program. In 2012, the Office of Science and Technology Policy (OSTP) sent out a memo to the public announcing a Big Data Initiative, earmarking \$200M to be spent on improving access to the government's big data stores. In 2013, there were more memos and Executive Orders coming out on this issue, but NASA was missing from the list of recipients (DOE, Department of Defense, and others); so it must be asked- where did NASA miss the boat? Dr. Holmes noted an ITIC finding in November 2012, that NASA acquire fiber-optic pathways to support current and future data, and a recommendation that they buy rather than own these pathways.

Discussion

The committee discussed a draft work plan to determine how the BDTF would move forward. Dr. Holmes felt that the BDTF shouldn't address the areas of data searchability

and availability, proprietary periods, long-term archiving, and other frequent requests that are made of NASA's data stores, feeling that processes are already in place for this at NASA. The BDTF should break new ground instead, and should survey the community, choose 3 to 4 topics, and produce products. The BDTF should form a concise problem statement, research, organize and develop positions, form a consensus, and draft and present results in a white paper (4-6 pp) accompanied by a slide presentation. Because the BDTF expires in December 2017, there are only 4-5 more face-to-face meetings in advance of each of the future Science Committee meetings in which to develop findings and recommendations to take to the Science Committee. To this end, the Task Force should also hold teleconferences as appropriate. Dr. Holmes reviewed his duties as Chair as primarily being the representative to the Science Committee, and closed with the thought: "Do good, work hard, NASA needs us."

Dr. Ray Walker agreed that data availability/searchability did not require a hard look, but noted that as data volumes get larger, it will be necessary to figure out the pieces we want to use; in this sense the issue is still important to consider. Dr. Holmes invited Dr. Walker to write up an actionable recommendation on the issue and send it to Dr. Smith. Dr. Tino commented that there are model-level, internal, and external use domains; what is it that are we actually trying to do? He agreed to write up an item on this question. Dr. Kinter said that it seems that by definition, Big Data means the biggest and baddest data sets; in that respect, we typically we see accessibility as a way to aggregate and analyze data from an entire data set (petabytes); very few users will have the resources to operate data sets of such magnitude. The Task Force should also think about facilitating the analysis of data sets that are too big to move and too big to analyze *in-situ*. Dr. Holmes agreed to revise the work plan with the additions of the written contributions, and to look at areas that can be extended beyond the state of work; the BDTF needs to look at benchmarks regarding this issue.

HPD Big Data

Dr. Jeffrey Hayes presented areas of concern for the Heliophysics Division (HPD) in terms of Big Data needs. HPD studies the sun's variance, the response of geospace, and the Sun-Earth system's impacts on humanity. To do this, HPD engages in the science of space weather, tries to understand the interconnections between the Sun and Earth, and develops knowledge to improve the prediction of extreme events such as major coronal mass ejections (CMEs). The mission portfolio includes a research and analysis (R&A) line, an Explorers mission line, along with Living with a Star, Solar Terrestrial Probes, and the sounding rockets program. Mission investment is guided by the Decadal Surveys and NASA's advisory bodies. The HP System Observatory includes numerous satellites such as IRIS, Wind, STEREO, the Van Allen probes, and the Interstellar Boundary Explorer (IBEX). Within the current missions and the operations budgets, there is a certain amount of funding for data archiving, and the creation of standards and accessibility. Dr. Hayes felt that most missions were able to respond quickly to decisions on data archiving and curation. Senior Reviews address the scientific merits of HPD missions every two years, and take into account the accessibility, usability and utility of data (including archiving after the mission is complete). As a result, the data pipeline is doing very well.

About 70-80% of HPD data come from extended mission phases. The sun varies in a roughly 22-year cycle; all of these HPD missions operating simultaneously are beginning to enable the understanding of a very complex system. The average cost of a Heliophysics satellite operation is \$2.9M annually. The Solar Data Analysis Center (SDAC) and Space Physics Data Facility (SPDF) are the active archives for HPD and run at about \$3.3M per year. There is also a ROSES element amounting to about \$1M a year. Thus, the total to curate the data is about \$4.5M per year, plus some money in the mission lines themselves. Dr. Hayes noted that “Scientists want all the data all the time, forever.” In the early 2000s, the Decadal Survey came out with a priority for a Virtual Observatory, in which the idea was to collect all the data (both Astrophysics and Heliophysics) and make it universally accessible through common standards. At the time, Astrophysics had one standard, and Heliophysics had multiple standards. Over the last 20 years, NASA has been trying to get these standards in line, and Dr. Hayes felt that good progress was occurring in this area.

Heliophysics has an explicit policy that established standards, which are FITS, CDF, and NetCDF. NASA is in a much better place than it was 10 years ago in terms of standardization. HPD has also restored a large fraction of data from its older missions, and has been systematically examining old archives and restoring data archives and datasets of scientific interest. For any metadata, it is necessary to get everyone to agree on key words. HPD has gotten good buy-in, and users can now use the Space Physics Archive Search and Extract (SPASE) metadata wrappers to do an inventory, search by date or event, etc., to help do system science. The process has gotten a lot better, and appears to be going faster. HPD’s three most recent missions are successfully using the SPASE metadata wrappers. The first data from Magnetospheric Multiscale (MMS), for example, will be available on SPDF on March 1.

HPD is starting to get terabytes of data - this is a new experience. There are 800 TB from SDO to date, and the volume is growing. HPD is now looking at storing 1 PB in the SDAC; this data volume will probably triple or quadruple as future missions come online. Stanford University will not always support SDAC; at some point the data will have to be brought back to NASA. Dr. Hayes felt that putting data on the cloud was still an iffy prospect, and cited a recent accidental deletion of stored data as one of its potential drawbacks. Solar project data volume growth, in terms of both lifetime data volume and data rate, will continue to grow. The question is where and who will store it, and how will it be moved around? HPD can’t throw data away because Heliophysics science needs the context.

Data policy is working well. HPD has a registry and inventory of the data, and is constantly updating. Legacy datasets have pretty much completed their extractions. Now HPD is concentrating on standards. A future challenge is how to use the SPASE metadata, how to use the data, and how to make it accessible to the non-expert user. Remote sensing vs. *in-situ* measurements are very different and these differences must be taken into account. For modeling, how do we archive useful, powerful comparisons? At this point, models do not have a standard; we are working toward it. As we move away from

the Virtual Observatory concept to a more consolidated way of getting data out, we must focus on metadata and links to generic access methods, and avoid stovepiping. The interdisciplinary aspects of data will be addressed by a larger group. Dr. Hayes noted that the Virtual Observatory concept did not fail, but the technology has since moved on.

Dr. Holmes asked Dr. Hayes to identify HPD needs from the BDTF standpoint. Dr. Hayes replied that one useful finding acknowledging the value of standards. The other issue of concern for him was the unfunded mandate about keeping versions of data in perpetuity. There is a NASA policy in response to the OSTP about public accessibility and publications, however the worrisome issue is whether the reference data in a paper has certain pedigree that may or may not be preserved in the archive. Who owns the final data? Which version of the software? There is never enough disk space. Another useful finding would be a statement that having data active, on-line, is a good thing. Data, especially taxpayer-funded data, shouldn't be buried in someone's desk drawer. NASA tends to get pushback from principal investigators on this issue- they feel their data is proprietary. Dr. Hayes agreed to write up an item for Dr. Smith.

Dr. Kinter commented that there is no data standard for models, and that this is a challenge for the future; he wondered how much interaction there is between the Heliophysics community and the tropospheric and weather communities. Dr. Hayes felt there was not much interaction, certainly not at the tropospheric level. There are meetings ongoing, however, and HPD would be open to anything the other communities have that can be used. The variables may be different, but it is something that could be explored. Dr. Walker mentioned that the National Science Foundation (NSF) is looking into data assimilation. Dr. Holmes noted that the community had looked at compatibility between Earth Science and Heliophysics data ten years ago, and stopped because of data sparseness. Dr. Neal Hurlburt agreed that the effort was still at the case study-level. IRIS is a good example of where we were forced to use models. Dr. Kinter noted that there are also ocean data assimilations that have a similar problem with data sparseness. The tropospheric problem has moved well during the last decade, and can accommodate data sparseness a little better. GSFC has some expertise here. Dr. Holmes asked Dr. Kinter provide POCs at Goddard. Dr. Walker mentioned that the Planetary Data System (PDS) has begun a study of archiving models, as well as the Community Coordinated Modeling Center (CCMC), and European work in both Heliophysics and Planetary at the University of Paris; these can provide useful Lessons Learned.

Science Committee Greetings

Science Committee Chair, Dr. Bradley Peterson, addressed the committee, thanking members for their important contributions. He noted that time was a pressing issue, and urged the BDTF to focus on finding commonalities and best practices across the subdisciplines, and building on the existing infrastructure only if it is useful. He asked the membership to regard the NASA budget is a zero-sum game, as NASA will buy in to recommendations only if they are affordable, or whether they are worth giving up something for. Eating into the budget for missions and research would be an undesirable outcome. Dr. Peterson suggested that the BDTF consult with subcommittee

chairs when useful, in order to iterate ideas across the Science Committee, subcommittees, and BDTF.

Big Data and Earth Science

Dr. Kevin Murphy presented an overview of the Earth Science Data Systems program, and stated that regardless of varying definitions of big data, Earth Science has it, as well as a large user base. Objective 2.2 of the 2014 NASA Strategic Plan informs the usage of Earth Science data to form a view of Earth that can be used across disciplines: ocean, atmosphere, cryosphere, etc. and their interactions.

The Earth Observing System Data and Information System (EOSDIS) is the largest component of the Earth Science data system, and is associated with the competitively selected programs, Making Earth System data records for Use in Research Environments (MEaSUREs) and Advancing Collaborative Connections for Earth System Science (ACCESS). EOSDIS works internationally and among the federal agencies to get data to the public, and processes data from level 0 to higher products to make available to users. EOSDIS was initiated in 1990, incorporating heritage data sets in 1994 from satellites, aircraft and *in-situ* sensors (e.g. flux towers), and was designed to handle a terabyte of data per day. EOSDIS reprocesses data quite often as instruments deteriorate or as better signal processing methods become available. There are about 15 petabytes (PB) of data currently available, all of which interoperate with other agencies and archives through established standards. EOSDIS has a distributed framework, and has had an open data policy since 1997. The system generates biophysical products and geolocates them, and distributes to the end users. EOSDIS has an extensive volume of data represented in over 9200 data types, which range over human dimensions, land, atmosphere, ocean dynamics and the cryosphere. The system works closely with missions in formulation and development in order to prepare data plans.

EOSDIS is spread out over the US. Mission data are processed by Science Investigator-led Processing System (SIPS), which are then passed along to the Distributed Active Archive Centers (DAACs) to support the user base.

DAACs are located at host organizations that are widely recognized by the community, and each DAAC has a working group that help to direct how the DAACs work. There is also a Program Scientist within each DAAC that roughly aligns with each subdiscipline. The two components overseeing the DAACs are primarily Headquarters for management and the Goddard Space Flight Center (GSFC) for implementation. The Earth Science Data and Information System (ESDIS) manages the coordination of EOSDIS activities to avoid duplication of efforts. ESDIS holds annual meetings and continually takes input through weekly teleconferences and annual meetings with DAACs managers and DAAC systems engineers. Roughly 160-180 people go to the annual meetings.

The EOSDIS infrastructure also ties together users and DAACs through earthdata.nasa.gov, a common metadata repository (CMR), Global Imagery Browse Services (GIBS), EOSDIS Metrics System (EMS), and various user support tools. EOSDIS performs an annual customer satisfaction survey, and also has DAAC User Working

Groups, which receive regular feedback. EOSDIS metrics from 2015 show 9462 unique data products, and 2.6M distinct users of EOSDIS data and services. EOSDIS distributes about twice as much data as it ingests. In 2015, the system received an ACSI score of 77 (considered very good). The trend for product delivery is increasing.

EOSDIS converts high-value products into imagery, such as the NASA Worldview website, which uses data from the Aqua/Terra/Moderate Resolution Imaging Spectroradiometer (MODIS) satellites, and NOAA's Visible Infrared Imaging Radiometer Suite (VIIRS). Worldview works much like Google Earth; users can zoom in and go back in time. Users can also overlay data, such as the SO₂ cloud over an erupting volcano, and find specific data such as fire hot spots. EOSDIS holds Senior Reviews to evaluate the various subsystems to evaluate performance and scientific merit.

Dr. Walker noted the many highly derived data products, and asked how EOSDIS kept up with evolving algorithms. Dr. Murphy explained that standard products are produced in collections, and EOSDIS is currently going from MODIS collection 5 to collection 6, reprocessing data. Collection 5 will be maintained until collection 6 is complete. Science teams will determine when the new collection is done. Dr. Holmes asked what the BDTF could for Earth Science. Dr. Murphy felt that NASA received little recognition for this important work, as it is generally not well understood. The data product ramp is currently limited by adapting to input from new instruments. EOSDIS has to put algorithms closer to the data in a way that allows unimpeded access to products; how to do this is still an open question. NASA also needs to learn how to work with commercial high-performance computing groups, maybe. Dr. Hurlburt asked how many of the 2.9 M distinct users were part of the active (science) community. Dr. Murphy replied that people who use a lot of the data will frequently use all of it (operational users who use Level 1 data). The numbers of graduate students, etc., are hard to estimate. Dr. Kinter asked how EOSDIS dealt with the budget realities. Dr. Murphy noted that EOSDIS recognizes the need to develop or adopt standardized-enough components to allow people to develop their own tools, a strategy that saves both time and effort. NASA doesn't want to be the first adopter or the last. The strategy depends on the community. EOSDIS keeps the principle of open application programming interfaces (APIs), and open access. The community is well aware of the data policy. Dr. Walker asked about the extent of which NASA provides interoperability in its joint work with NOAA. Dr. Murphy explained that NASA operates with NOAA on a catalogue level, uses open software sourcing, shares observations, and works closely with NOAA on the Climate Initiative and in the airborne program.

Supercomputing Big Data

Dr. Tsengdar Lee, Program Manager of the Earth Science Division Supercomputing Program, presented an overview of the program, and the NASA vision for future computing services. NASA has two supercomputing centers, one at Ames Research Center (ARC), which serves the entire agency) and one at GSFC, which serves primarily Earth Science. ARC supports agency-wide activities, from launch vehicles to general relativity.

In August 2015, the NASA Flagship computer, Pleiades, reached a half billion SBUs (computing cycles) delivered accumulatively from 2008, translating to nearly \$300M of services, at a cost of roughly 26 cents per SBU in 2015. NASA continues to grow the system, relying on Moore's law to go forward (Dr. Lee noting that some argue that the law has come to its end). Scientific and engineering efforts will grow, thus NASA will have to come up with a user policy because the system has become oversubscribed. The ROSES selection process is now being tightly coupled to the availability of computing time. For Earth Science imaging and modeling, the system can push the resolution down to 1.5 km currently; the holy grail of atmospheric science is 0.5 km. The workload is changing, shifting into data processing. As an example, the Kepler mission is using Pleiades to support validation for new exoplanets. This has become the primary avenue for producing discoveries in that area. Data assimilation systems are being used to create physically consistent long-term data sets, from 1979 to the present, and are also downscaling to higher resolution data for climate studies. The Orbiting Carbon Observatory (OCO-2) is presenting data processing challenges. NASA is doing a data re-processing campaign with new algorithms, with about 60% of this work being done on the supercomputer and 40% on the Amazon cloud. High End Capability Computing (HECC) is being used to clear 5 years of an unmanned aerial vehicle synthetic aperture radar (UAVSAR) data processing backlog, to reduce latency. Processing is moving into the big data area, pitching high-performance computing against Large Scale Internet. Can high-performance computing (HPC) be used as a private cloud? How do we put together an architecture to process, analyze and mine data?

Currently, data storage and data management is the core of the business, with data in the middle, and all the service and processing surrounding the data set. A Science Cloud architecture ideally provides an agile, high level of support, with the system owning the data, using a data management system, data analytics service, openstack, etc. NASA is constantly looking at new technologies: cloud and virtualization, high-performance object store, and SciDB (the latter heavily supported by DARPA). The science benefit of a science cloud has helped to validate many types of measurements, such as global fires. Coupling HPC and cloud computing can create a best-of-breed computing service environment. HECC's path to growth is constrained at present; NASA has maxed out the infrastructure in terms of facilities, building, water, and electricity, and is engaged in a study on how to build next-generation data centers. Drs. Holmes, Walker, and Hurlburt expressed concerns about user constraints, given that 70-80% of the program's workload requires a tightly coupled process. Dr. Lee agreed to write a statement on this state of being for use by the BDTF. He added that certain types of workloads could be cloud-computed, and NASA is exploring those options as well. Dr. Clayton Tino asked if Dr. Lee had any sense of the capacity the program was losing due to mixed mode services. Dr. Lee replied that NASA was doing the mixed workload because of the demand. Some of the projects didn't plan for their HPC use, and need to do a better job of such planning in the future.

Astrophysics and Big Data

Dr. Paul Hertz, Director of the Astrophysics Division (APD) presented Big Data needs as viewed by the Astrophysics community. Astrophysics addresses the evolution of the

universe, the origin of galaxies and stars and the question of whether we are alone in the universe. The APD is driven by the Decadal Surveys, science roadmaps, and implementation plans to support its ability to handle large data questions. Sixty percent of the budget supports developing space missions, 20% operations, another 5-10% is dedicated to research and development. Data archives are funded as an infrastructure investment. APD's current suite of missions run from many small missions such as Neutron star Interior Composition Explorer (NICER), to the large space telescopes, Hubble and the future James Webb Space Telescope (JWST). The next large flagship after JWST is Wide-Field Infrared Survey Telescope (WFIRST), whose prime science is to understand dark energy and dark matter, which can only be done by measuring the small impact these forces have had in the history of the universe, by looking at large swaths of universe; i.e. looking at large amounts of data to see small perturbations. Thus WFIRST will be computationally intensive. WFIRST will be looking at millions of galaxies, searching for evidence of microlensing, which is also computationally intensive. Euclid, a European mission with similarities to WFIRST, will also create large data sets. Another future ground-based observatory is the Large Synoptic Survey Telescope (LSST). All three of these projects will be combining their data in pixel-by-pixel analysis. The various agencies are studying the best way of carrying out this data processing, a decade in advance of the need. A white paper on this topic can be found at [\[\[arxiv.org/abs/1501.07897\]\]](https://arxiv.org/abs/1501.07897); Jain et al; The Whole is Greater Than the Sum of the Parts.

All NASA Astrophysics science data are open to the community, and all data centers go through the Senior Review process every two years. All astrophysics archives share a set of common protocols and standards, allowing the user community to combine data from multiple ground and space observatories. The NASA Astrophysics Virtual Observatory (NAVO) manages the protocols, while NSF funds the tools. The three Astrophysics archives manage the NAVO backbone. APD recently held a Senior Review of the archives, and recommended that they become more proactive and aggressive about evolving into the future (increasing bandwidth, keeping up with technological advances, preparing for large volumes of data). Some types of computing might be more expensive in the cloud, and it must be determined which are which.

NASA and NSF are currently funding theoretical and computational Astrophysics networks (TCAN). Dr. Hertz was not aware of any issues thus far on getting time on NSF supercomputers. (Dr. Lee noted that NASA civil servants can't typically get on NSF supercomputers, but university Principal Investigators can.) Another computationally intensive area is laboratory Astrophysics: interpreting x-rays from Chandra, far infrared data from Herschel, and visible-to-ultraviolet Hubble spectral lines. These atomic line calculations are needed for creating line catalogues. Dr. Tino asked if underestimation of computing time were a theme in APD. Dr. Hertz explained that processing Kepler data has been more computationally intensive than was appreciated at the beginning of the mission, but that a new mission, Transiting Exoplanet Survey Satellite (TESS), which has a similar data product to Kepler, had planned accordingly to Lessons Learned on the need for anticipating computing time. Dr. Lee noted that NASA is also making tighter connections between HPC and the budget-planning process. In terms of

recommendations, Dr. Hertz noted that Astrophysics was a minority user of HPC, and was interested in areas where it could leverage existing assets, or in commercial or other research that can improve Astrophysics science. APD has partnered with DOE in the past, when they are interested in the science problem. DOE is not interested in exoplanets, but it is interested in dark energy and dark matter, therefore APD will be working with them on joint WFIRST-Euclid-LSS analysis.

Public comment period

No comments were noted from the online audience. At NASA Headquarters, Tripp Corbett made some comments from the vendor perspective, saying that he was noting a bit of disconnect, as tools are available at NSSC that should be more widely circulated. At a recent NASA meeting, he had heard a briefing on working with the cloud-computing community in a budget-conscious way, and agreed to send more specific information to the BDTF.

Other Federal Big Data Initiatives (NSF)

The NSF Big Data Hubs Program director, Dr. Fen Zhao, briefed the BDTF by phone on her program, which is funded at about \$20M year. There are related programs at NSF that look at Big Data infrastructure, pilot and implementation efforts, and Education-related activities such as the Big Data Work Force (\$30M a year looking at traineeships). The Big Data Hubs program looks at the complex relationships between data projects, end users, and commercial entities, and involves cross-disciplinary efforts and data sharing across the research ecosystem.

The inspiration for BD Hubs came from OSTP's 2012 Big Data Initiative, in which a Big Data Partnerships Workshop initiative resulted in 29 new partnerships, with 90 organizations participating, representing areas such as energy, health care, and finance. The initiative chose various issues such as climate change and personalized healthcare, and NSF initiated the BD Hubs effort to allow these partnerships to gel. BD Hubs was launched in March 2015, with four hubs in four regions of the US, and made awards in September 2015 (Columbia University in the Northeast, Georgia Tech and x in the South, UIUC in the Midwest, and University of SD, UC Berkeley, and the University of Washington in the West). Hubs are differently constructed consortia; the current phase is allowing hubs to start up their activities. The projects are called BD Spokes, which represent specific activity within each topical area, such as a platform for sharing neuroscience data. The spokes are funded at \$1M over three years, and are meant to leverage existing efforts. The Hubs are currently organizing drafts for each spoke, and full proposals are due this month. A large number of ideas came in on smart cities, and Internet of Things; the food/energy/water nexus; and human healthcare. NSF intends to fund these proposals this fiscal year, and there are latent projects waiting in the wings that can help transition some of these ideas to practice. NSF hopes to do this again next year. Dr. Holmes offered kudos to NSF for setting up this open-ended effort. Dr. Zhao noted that there is an end goal of sorts, as each Hub is responsible for generating 29 projects at the end of three years. This idea is not completely novel at NSF. The Foundation hope to fund each spoke for a second three years, to have them become self-sustaining. A similar effort was undertaken under US-Ignite, to support networking. The

idea is to look for the unknowns, as interesting things can happen in these large, multiple collaborations. Everyone brings their own physical infrastructure, and also tries to identify service providers. Dr. Holmes noted that most of the Hubs were geographically close to NASA PIs. Drs. Holmes and Zhao agreed that a closer collaboration would be ideal.

Planetary Science Big Data

Dr. Michael New, Program Scientist for the Planetary Data System (PDS), presented the needs of Big Data from the planetary perspective. Most planetary data work is based at GSFC. Planetary Science Division (PSD) data policies state that all science data returned from planetary missions belongs to the public domain. Any exclusive data access cannot exceed six months. In funded science research, any data necessary to replicate published research results, that are also the product of a NASA award, must be made immediately available to the public. The planetary data environment includes PDS, the Planetary Cartography Program (PCP; USGS), Minor Planets Center (MPC; Harvard) and the Astromaterials Curation Facility (ACF; Johnson Space Center). Data ranges from ground-based assets, individual investigators, mapping, data analysis (e.g., trajectories), sample returns, ANSMET (Antarctic meteorites), to atmospheric dust. The output of the PDS is primarily to taxpayers, educators and talented amateurs. At the ACF, NASA stores space-exposed hardware, lunar samples, cosmic dust samples, and Hayabusa (comet) samples. NASA is currently re-engineering its sample catalogue to make these samples available on line. The MPC is responsible for small bodies, and the orbits of minor planets and comets. The PCP maintains the cartographic capability for mapping the planets and the Moon, and develops and maintains the Integrated System for Imagers and Spectrometers (ISIS), which enables things like spectrographic maps of Io. ISIS is preparing to incorporate an open-source visualization tool, the SPICE-based Cosmographia. ("SPICE" is a NASA information system and its use extends from mission concept through post-mission data analysis, and it helps to correlate individual instrument data sets with those from other instruments on the same or on other spacecraft.)

PDS is a federated archive, with data distributed across the country; its discipline nodes were recently re-competed. Management of the system as a whole is also based on a federated model. Planetary data are managed by planetary SMEs. Data is physically stored at the nodes, and the deep archive is maintained at the NASA Space Science Data Coordinated Archive (NSSDCA). The Navigation and Ancillary Information Facility (NAIF) implements standards and tools that are needed to understand the motion of celestial objects. In planetary data sets, everything is moving relative to everything else: spacecraft, instrument, Earth, and Sun, all of which need time conversion standards. The collection of these variables is called Observation Geometry (OG). The current PDS is distributed across six nodes, which after a recent competition are now in their first year of a 5-year Cooperative Agreement. The PIs at each node collectively form a management council, and provide input about standards and decision-making. PDS-4 has just recently been rolled out. It is an XML-based, model-driven, service-oriented model, and a modern technical foundation for planetary science data. Existing PDS-3

products will be converted to PDS 4 when practical and sensible. The European Space Agency and JAXA' planetary data systems are both adopting PDS-4 standards.

The total volume of PDS is about 1 PB. Almost all computations are performed on individual workstations. PDS has just started its next 10-year roadmap, and will be announcing an opportunity to self-nominate in early March. Areas of improvement to be addressed in the roadmap are to include: simplifying and improving the pipeline; improving search capability; developing more useful metrics; improving tools for archiving small data sets; and improving archive preparation and documentation, especially for non-mission data providers. Relevant websites are: naif.jpl.nasa.gov and pds.nasa.gov

Dr. Hurlburt asked about PDS metrics. Dr. New admitted to having poor metrics of usage and users, and noted that the roadmap effort would help to identify the metrics PDS wants, and to adapt the system to provide them. Dr. Beebe commented that the international planetary data alliance accepted SPICE as their data tool at their last meeting, a favorable indicator. Dr. New, when asked about Big Data needs, allowed that there were not many specific areas in planetary, with the exception of magnetospheric and plasma data, or when generating very high-fidelity gravity models. The lunar gravitational mapping mission, GRAIL, is currently working on a gravity field model on the HPC. He hadn't heard about any issues with pipeline associated with the GRAIL work. Dr. New felt the BDTF could direct a question to the Agency as to how it would like to handle the storage of grant data. PSD needs a clear direct statement on this issue, which needs to be informed at the Agency level because it will be a response to an OSTP directive. There are 1500 grantees in PSD; it would take a labor-intensive effort to store all their data. Another question is what kind of data PDS is expected to archive. Dr. Holmes noted that the directive applies to the other disciplines as well, and instructed Dr. Smith to note this as an issue. A meeting participant noted that the grant disposition question was being addressed in the roadmapping task, entailing a community-based reappraisal of the subject over the next 6-9 months.

Discussion

Dr. Holmes followed up briefly with Dr. Lee on HPC, and asked what visibility existed for the program, and what the chances for collaboration with DOE Exascale might be. Dr. Lee identified himself as Chair of the High-End Computing Interagency Working Group (HECIWG), but noted that the Exascale computing facility is under National Strategic Computing Initiative, a different governance. The HECIWG is meeting monthly at the moment, and Dr. Lee felt he could start vectoring the discussion in their direction. He noted that DOE sets up a process for eligibility; a task needs to have a certain profile, and x number of cores. The gate for eligibility to get on the DOE's leadership computing systems, however, is higher than NASA's entire system. NASA is far behind NSF and DOE in the supercomputing arena. NASA's leading system is less than 5 Tflops. Dr. Holmes considered that BDTF make a finding on the matter, as NASA is working on projects of national significance. Dr. Tino asked if Exascale was specifically designed to solve DOE problems, with specifically implemented architecture. Dr. Lee reported that DOE has a co-design concept, and they bring in an application that works on the exascale system.

They are considering climate-change as a co-designed system. DOE doesn't have the interoperability requirement. Dr. Walker commented that DOE has specific problems, while NASA is more broad. Dr. Holmes noted that DOE is addressing both astronomy and climate, and that while some of the scales are different, the physics are similar. Dr. Tino felt that NASA should either focus on products and services, or accept generality. Dr. Holmes suggested NASA managers address utilization models at future meetings. Dr. Kinter asked about what HPC would use Big Iron for after its nominal 3 years of operation.. Lee said that NASA plans to repurpose Big Iron after 3 years, back into a generalized cluster. NASA is still limited by facilities re: power and cooling. Dr. Holmes asked Drs. Tino and Kinter to write a talking point on the facilities issue.

BDTF members raised some general topics for further exploration. Dr. Tino noted that each of the presenters had adopted some form of standard, illustrating that people recognize that standards do matter. From a management standpoint, however, the subdisciplines had inconsistent metrics on users, and questioned why archives had to be maintained, in the absence of usage. Dr. Walker explained that some data have extremely long lives; every time we get a new mission to Jupiter, for instance, Voyager and Pioneer data sets are in demand again. It's critical that some of these data sets be safeguarded. Dr. Holmes noted that the Senior Review might be a vehicle for determining which data should be kept. Dr. Hurlburt suggested user metrics inform these sorts of judgments. Dr. Tino felt user surveys were not always effective, and that metrics on actual use would be more useful in getting smart on what data to store. Dr. Holmes asked Dr. Tino et al. to flesh this out thought and do more research in advance of the next meeting. Dr. Beebe added that one also needs to consider the intrinsic sizes of communities and their stability; they also tend to move around when major missions arise.

Dr. Holmes was surprised at the lack of a clear vision for the future and asked Dr. Hurlburt to write a finding on this topic. Dr. Holmes asked Dr. Smith to sound out the Science Mission Directorate to determine the level of concern over grant data storage. Dr. Beebe reported that it was a major concern that has already reached the top level of the administration, which had established workshops for people preparing for federal grants. Dr. Holmes gave an action to Dr. Smith to clarify Dr. Murphy's statement on the use of open source software, and asked BDTF members to examine the NSF nodes of the BD Hub effort, to determine how close they are to co-located NASA PIs.

Dr. Holmes asked that the next BDTF meeting take place at GSFC for 2.5 days in the April-May time period, and to perhaps consider a site visit to ARC in the future, to include some interaction with Silicon Valley. Dr. Smith reported that she would be working on an extension of the TOR, off-line. Dr. Holmes adjourned the meeting at 4:59 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 (via telecon/Webex)
Neal Hurlburt, Lockheed Martin
James L. Kinter, George Mason University (via telecon/Webex)
Clayton Tino, Virtustream, Inc.
Ray Walker, University of California at Los Angeles
Erin Smith, **Executive Secretary**, NASA HQ

NASA Attendees

Louis Barbieri, NASA
Dan Crichton, NASA JPL
Elaine Denning, NASA HQ
Deborah Diaz, OCIO NASA
John Evans, NASA
T. Jens Feeley, NASA HQ
Navid Golpayegani, NASA
Jeffrey Hayes, NASA HQ
Paul Hertz, NASA HQ
Tsengdar Lee, NASA HQ
Edward Masuoka, NASA
Duane McMahon, NASA
Tom Morgan, NASA HQ
Kevin Murphy, NASA HQ
Michael New, NASA HQ
Herbert Schilling, NASA
Grif Schilly, NASA
John Sprague, NASA OCIO
Elizabeth Yoseph, NASA

Non-NASA Attendees

Joseph Bredenkamp, NASA retired
Terry Blankenship, Booz Allen Hamilton
Jung Byun, Booz Allen Hamilton
Chiehsan Cheng, Global Science and Technology
Tripp Corbett, ESRI
Joseph Dohry, Booz Allen Hamilton
Alex Duner, Medill News, Inc.

Grace Hu, OMB
Eric Feigelson, Penn State University
Robert Kohon, Novetta
Bradley Peterson, OSU, Chair, NAC Science Committee
Amy Reis, Ingenicomm, Inc.
Alyssa Retski, Lobbyit.com
Marcia Smith, Space Policy Online
Connie Spittler, Global Science and Technology
Geordan Tilley, Medill News, Inc.
Joan Zimmermann, Ingenicomm, Inc.

Appendix B Membership

Dr. Charles P. Holmes, Chair NASA HQ (Retired)

Dr. Reta F. Beebe New Mexico State University

Dr. Neal E. Hurlburt Lockheed Martin Space Systems Company

Dr. James L. Kinter George Mason University

Dr. Clayton P. Tino Virtustream Incorporated

Dr. Raymond J. Walker University of California, Los Angeles

Dr. Erin Smith, Executive Secretary NASA Headquarters

Appendix C

Presentations

1. Big Data Task Force Charter/Subcommittee Feedback; *Erin Smith*
2. Legacy for the NAC Information Technology Infrastructure Committee; *Charles Holmes*
3. Heliophysics Division Big Data Needs; *Jeffrey Hayes*
4. Big Data and Earth Science; *Kevin Murphy*
5. Supercomputing and Big Data at NASA; *Tsengdar Lee*
6. Astrophysics Division Big Data Needs; *Paul Hertz*
7. Other Federal Big Data Initiatives (NSF); *Fen Zhao*
8. Planetary Science Big Data Needs; *Michael New*

Appendix D
Agenda

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

Inaugural Meeting
February 16, 2016

NASA Headquarters
Glennan Conference Room, 1Q39

Agenda
(Eastern Standard Time)

Tuesday, February 16

8:00 – 8:30	Opening Remarks / Introduction of Members	Dr. Erin Smith Dr. Charles Holmes
8:30 – 9:15	Big Data Task Force Charter / Subcommittee Feedback	Dr. Erin Smith
9:15 – 9:30	<i>BREAK</i>	
9:30 – 10:15	Legacy from NAC IT Infrastructure Committee	Dr. Charles Holmes
10:15 – 10:30	Discussion	
10:30 – 10:45	<i>BREAK</i>	
10:45 – 11:15	Planetary Science Big Data	Dr. Michael New
11:15 – 11:45	Heliophysics Big Data	Dr. Jeffrey Hayes
11:45 – 12:45	<i>LUNCH</i>	
12:45 – 1:00	Greetings from the Science Committee	Dr. Bradley Peterson
1:00 – 1:30	Earth Science Big Data	Dr. Kevin Murphy
1:30 – 2:00	Supercomputing Big Data	Dr. Tsengdar Lee

NASA Advisory Council Ad Hoc Big Data Task Force, February 16, 2016

2:00 – 2:30	Astrophysics Big Data	Dr. Paul Hertz
2:30 – 2:45	Public Comment	
2:45 – 3:00	Other Federal Big Data Initiatives (NSF)	Dr. Fen Zhao
3:00 – 3:10	<i>BREAK</i>	
3:10 – 3:30	Work Plan and Future Meetings	
3:30 – 5:00	Discussion / Findings / Recommendations	
5:00	<i>ADJOURN</i>	

Dial-In and WebEx Information

For entire meeting February 16, 2016

Dial-In (audio): Dial the USA toll-free conference call number 1-800-988-9663 or toll number 1-517-308-9427 and then enter the numeric participant passcode: 4718658. You must use a touch-tone phone to participate in this meeting.

WebEx (view presentations online): The web link is <https://nasa.webex.com>, the meeting number is 999 765 122, and the password is BigD@T@16.

** All times are Eastern Standard Time **