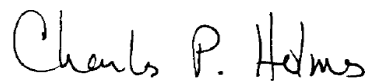


**Ad-Hoc Task Force on Big Data
of the
NASA Advisory Council Science Committee**

Meeting Minutes

**June 28-30, 2016
NASA Goddard Space Flight Center**



Charles P. Holmes, Chair



Erin C. Smith, Executive Secretary

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

Table of Contents

Introduction	3
GSFC Science and Exploration Directorate	3
CCMC	5
HPC and Climate Model Data	6
Climate Analytics	8
Big Data Projects Outside NASA	9
Data Policy Evolution in Heliophysics	10
Discussion	12
Findings Discussion	14
NASA Astronomy Archive	14
MAST	16
HEASARC	18
EOSDIS	19
PDS	21
SPDF	23
SDAC	24
OCIO ITD	25
Discussion	26
Work Plan Topics	27
Cloud Computing Initiative	27
Findings and Recommendation Concurrence	29

Appendix A- Attendees

Appendix B- Membership roster

Appendix C- Presentations

Appendix D- Agenda

June 28, 2016

Introduction

Dr. Erin Smith, Executive Secretary of the NASA Advisory Council (NAC) Ad-Hoc Task Force on Big Data (BDTF), called the meeting to order and provided administrative details that govern Federal Advisory Committee Act (FACA) meetings. She introduced Dr. Charles Holmes, Chair of the BDTF, who then introduced Dr. Colleen Hartman, Director of Science and Exploration Directorate at Goddard Space Flight Center (GSFC).

GSFC Science and Exploration Directorate

Dr. Hartman presented an overview of the directorate at GSFC, which employs over 500 civil servants, 600 post-doctoral students, 1500 contractors, and many young student interns. Among the notable scientists who have performed research at GSFC are Piers Sellers, Nobelist John Mather, and Compton Tucker. GSFC is centered on fundamental science questions in all four science areas, based on their respective Decadal Survey goals. The center has a very diverse mission portfolio, and has conducted hundreds of successful flights, producing much interrelated data that needs to be analyzed. Currently, the space sciences are very keen on extrasolar planets, which will require knowledge from all four science divisions at NASA; the existence of exoplanets is no longer a single dimensional question. The Sciences and Exploration Directorate at Goddard is ensuring that teaming across the center between scientists and engineers, to ensure value for the taxpayer.

Dr. Hartman highlighted some recent Goddard missions: the International Space Station (ISS)'s Neutron star Interior Composition Explorer (NICER) payload, which is looking at spinning stars. NICER is a low-cost, class D mission. The mirrors for the James Webb Space Telescope (JWST) are in the process of being installed at the center, after which the spacecraft will move on to Northrup Grumman at the end of the year for further testing. JWST will launch in 2018, enabling new abilities such as the observation of gaseous planet atmospheres. Dr. James Kinter asked if the data centers were maximizing the scientific return on investment. Dr. Hartman affirmed that the science visualization staff regularly performs scientific data mining for presentation to the public, providing imagery that is both informative and spectacular. The Hubble Space Telescope imagery enthralled both the laymen and the scientist. In response to a question, Dr. Hartman confirmed that NASA civil servants were being denied access to supercomputing assets at the National Science Foundation (NSF), and that there was a Science Mission Directorate (SMD) FACA group that is dealing with the issue. Dr. Holmes felt that perhaps the BDTF could look into this matter.

Task Force Member Reports

Dr. Holmes presented the goals of this, the second meeting of BDTF: hear from the projects, continue the discussion of the NSF Big Data Hubs project, converge topics for Task Force studies over the next 18 months, and generate findings and recommendations for the BDTF report to the Science Committee on 26 July. For future meetings, Dr. Holmes wished to expand the individual Task Force reports to about an hour, after conducting the appropriate research.

BDTF members introduced themselves in advance of their individual reports.

Dr. Holmes began by describing reporting to the Science Committee in March, for which he received positive feedback. He also visited the Goddard Institute of Space Science (GISS: New York, NY) and discussed their big data needs, as well as NSF's Northeast Big Data Hub at Columbia University. He also worked on considering topics for BDTF projects, and tuned in to two "Big Data" webinars, and reported that the webinars were too high-level and covered topics that were not germane to the BDTF.

Dr. Reta Beebe reported that she had talked with the leader of the Midwest Big Data Hub and got the impression is that they're very active. She referred to a group of planetary scientists in Europe who had proposed to the European Union an initiative to integrate planetary science in Europe, and subsequently got funded for a five-year period. She felt that the top-down association had put a big burden on this group, which did not provide enough funding and tools, and that perhaps this NSF effort may result in a similar negative interaction.

Dr. Kinter reported visiting NSF Headquarters and meeting with Program Managers (PMs) on the Hubs project. NSF is now in the midst of reviewing solicitation for the Spokes component of the NSF Big Data project. Currently, Dr. Kinter is thinking about the workflow of the Task Force as it pertains to his focus on Earth system modeling, and generating data through observations and models. Scientists are in one place and data is someplace else; this situation needs to be better coordinated.

Dr. Raymond Walker assessed the California Institute of Technology's Big Data effort, which he also described as being concerned with high-level issues. Caltech is organizing a December workshop to look in particular at Big Data and computational issues. Dr. Walker followed up his effort with a librarian, Christine Borgman, at the University of California at Los Angeles (UCLA), which is considering hosting a giant archive for space-derived big data.

Dr. Neal Hurlburt reported on the West Coast Hub, and having reached out to the Moore foundation. He shared his concerns about the limitations of a top-down approach. He received limited information on an SDO experiment, and on the Daniel K. Inouye Solar Telescope (DKIST) that will eventually generate 10TB of data per day.

Dr. Clayton Tino reported a positive session with the Southeast Data Hub at the Georgia Institute of Technology, which was trying to bring industry and academia together to seed small projects. As to metrics, there is still a struggle to define how to quantify data projects.

Dr. Eric Feigelson described his efforts to organize the first international conference on astroinformatics under International Astronomical Union (IAU), in an attempt to harness a decade of dispersed activity that must be consolidated. The conference aims to bring in engineers to talk to scientists. There are a handful of engineers who are interested in working with astronomers, which are too few, but the conference

represents a start. There was also a meeting of statisticians at Carnegie-Mellon representing a group relevant to the cause.

Community Coordinated Modeling Center

Dr. Maria Kuznetsova, Director of the Community Coordinated Modeling Center (CCMC), gave an overview of the center. CCMC was established in 2000 as an essential element of the National Space Weather Program, to help bridge the “Valley of Death” between research and space weather applications. CCMC was a game-changing solution that pioneered the path from research to operations. The center serves as a hub for the collaborative advancement of space weather prediction capabilities, and ingests, assesses, and disseminates data for operations at the National Oceanic and Atmospheric Administration (NOAA) and the Department of Defense (DOD). It also provides prototyping services for NASA missions. Assets and services include hosting of models, providing simulation services, performing assessments of models, developing tools for dissemination (actionable displays, databases, flexible infrastructure tools), and providing space weather services for NASA missions as well as hands-on education. There is an expanding collection of more than 80 models whose domains are from sun to Earth. CCMC deals with the complexity of physics, platforms, computer languages, and compilers, and is characterized by flexibility, diversity, openness and inclusiveness: a “Super Store” of services. Two signature services provided by CCMC are a Runs-on Request System, an interactive system to serve advanced models to the international research community, and the Integrated Space Weather Analysis System, which provides real-time data for space weather assessments.

Runs-on Request, begun in 2000, has had more than 16,000 simulation runs, and has provided data for more than 200 publications. Using the Kameleon software suite, it addresses metadata, standardization, access and interpolation, and facilitates access to space weather models hosted at CCMC, to enable scientific discovery. One application of Kameleon has been used in collaboration with New York Planetarium to visualize models on the dome. Results for a set of solar science simulations are currently available for online visualization of magnetic reconnection events. An integrated space weather analysis system (iSWA.ccmc.gsfc.nasa.gov) is available for use, in which the user can search and drag to build his or her own display; it can be used on Android and iPhone. The app had 22,000 users in 2015. CCMC has recorded 2.6 million mobile space weather apps downloads (more than 100 k files per day) with 90 million files registered and archived to date. CCMC is also testing predictive capability before the onset of the event, as seen in the coronal mass ejection (CME) scoreboard, and event-based evaluations to trace model improvement; other validation measures include a sanity check tool kit for real-time runs.

CCMC has representation across the space weather community, including domain experts, space weather forecasters and analysts, and software developers and systems engineers, many of which represent overlapping domains. Between NSF and NASA, CCMC has about 13 FTEs, operating as one team. There is a CCMC Advisory Group chaired by BDTF member Dr. Walker; this group has suggested improving visualization, and better managing the archive simulations. CCMC has numerous partners including

the European Space Agency (ESA), the University of New Hampshire, and Dartmouth University, and connects to the community through biennial workshops, the annual NASA Robotic Mission Operations Workshop, and participation in NASA's Living With a Star (LWS) program. CCMC continues pursuing objectives, including addressing a need for a hub that allows the community to interact more effectively. The center is in the process of redesigning its database based on SPASE and IMPeX data formats.

CCMC's computational infrastructure is comprised of supercomputing clusters (2200 CPUs); dedicated servers, workstations and virtual machines (1800 CPUs), 1.2 PB of data storage; and a 1-10 GB network/SEN support, all of which is imperiled by budget cuts. NASA wants to keep the system fast and flexible, and is committed to keeping the same footprint. CCMC may eventually need access to other systems to produce future products, such as for LWS, and will need increased storage capacity for large files such as particle distribution data (typically on the order of a TB).

In summary, CCMC is an asset for the international space weather community, a fast-response unit for emerging community needs, an access point for state-of-the-art capabilities, and a venue for dissemination of research results. It serves as an actual and virtual repository for models, simulation results, and space weather products; a playground for scientists; a hub for collaborative research and development; and a resource for hands-on education. Its current challenges include security for remote access. Dr. Kuznetsova indicated that despite having some NSF funds, which covers education and the needs of the atmospheric community, CCMC still lacks access to NSF supercomputing assets. NASA has recently established one contact person to try to resolve the problem, but there remain some conflict-of-interest issues. Dr. Tino asked how specific implementation is to the CCMC platform, given possible access to faster networks. Dr. Kuznetsova replied that users tell CCMC what they need, such as compilers or platforms, and CCMC builds a custom service for each requester using virtual modeling (VM), mirroring, or sometimes custom building a physical model. However, not everything can be solved by VM. Asked if the hub is open or closed, Dr. Kuznetsova said the model developers are the owners, and that CCMC doesn't distribute source code. Everything is based on handshake agreements, and everything CCMC develops is open for use. Dr. Hurlburt asked how the framework is maintained over the long run. Dr. Kuznetsova replied that a rapid prototyping center didn't work, which is why it was kept in-house. CCMC is validating models for NOAA, for instance, and is trying to streamline the procedure. NOAA looks at the Super Store and chooses what it wants. As for old models, when the models stop working, CCMC upgrades it, but it does keep the old models. What CCMC maintains depends on the customer needs. Dr. Holmes commented that the bottom-up approach of CCMC constituted a success story, to which the Air Force, NSF, and NASA had contributed, when it became apparent that CCMC needed a funding line and a review process. The CCMC participates in the Heliophysics Division's senior reviews and they pass with flying colors each time.

High Performance Computing (HPC) and Climate Model Data

Dr. Daniel Duffy, High Performance Computing (HPC) Lead, presented a briefing on the High-End Computing program at NASA. The NASA Center for Climate Simulation (NCCS)

provides an integrated, high-end computing environment to support the specialized requirements of climate and weather modeling, and to enable big science beyond NASA. It holds about 100 PB in total storage. Generally speaking, NCCS takes small input and creates a large output, such as Observing System Simulation Experiments (OSSE) runs for the Global Modeling and Assimilation Office (GMAO) for new remote-sensing platforms. GEOS model is a cubed sphere; its current operational forecast is running at 27km resolution using about 27 million grid points; its target is 13km resolution. NCCS performs dynamic downscaling assessments for such events as Northeast winter storms, and midcontinent mesoscale convection systems (MCS). NASA downscaling models include the NASA Unified-Weather Research and Forecasting (WRF) models and Modern-Era Retrospective analysis for Research and Applications (MERRA-2) replays. Petabytes of data are involved in each of these runs. Dr. Duffy displayed movies of output from the GMAO, showing carbon dioxide movement at 12km altitude, column concentrations of sulfur dioxide from both natural features and power plants, and even shipping tracks. A 1.5km-resolution global simulation with Goddard Earth Observing System-5 (GEOS-5), executed on roughly 1000 nodes of the NCCS Discover SCU10 cluster, is able to show dust storms, convective cells, dust transport across the Atlantic, fires in central Africa, and sea salt concentrations. Also seen are deep and shallow convections of cumulus clouds, fires, and point sources, demonstrating the complexity and maturity of global climate models. Resolution is fine enough to show rain bands in hurricanes.

NCCS has a total peak capacity of 3.5 Pflops with 90,000 computer cores, and in terms of storage has grown from 4PB to 40 PB over four years. Science requirements are growing in the GMAO portfolio (MERRA, GEOS) and data volume will continue to grow as well. To increase the GEOS-5 model resolution to 100m resolution, just for atmosphere, 14 billion grid points will be needed, and 0.1 PB RAM. Each doubling of resolution requires eight times the number of grid points, hence eight times the memory. Dr. Duffy displayed some downscaling data analysis examples, including water maps from LandSat, and cloud climatology record from the Moderate Resolution Imaging Spectroradiometer (MODIS).

Different infrastructure is also needed for analytics compared to HPC; it needs an infrastructure that enables traditional use of data, therefore NCCS is evolving its services and has developed two platforms; one is an Advanced Data Analytics Platform (ADAPT; a cloud). ADAPT is best used for inherently parallel processing of big data. ADAPT is not an archive. Current data being stored in ADAPT comes from LandSat, MODIS, MERRA, and NGA. This is for NASA research use only. NCCS is also developing a Data Analysis Storage System (DASS) to accommodate both HPC and Data Analytics. This is where research projects will reside, using both traditional HPC storage software with a server, and JBOD commodity-based hardware, using the same infrastructure. To help these work together, NASA is using a George Mason University (GMU)-developed spatiotemporal index and a Hadoop File System, which supports the structured scientific data. Thus far, a test run using average global temperatures over many years has shown that that the system is working. The challenge is to be able to perform data analysis at every layer, and to be able to efficiently move data to the appropriate layer

for computation. To the “5 Vs” of data, NCCS also needs to be adding lifecycle and data security.

Dr. Walker asked: what happens to impressive results? Where’s the archive? Dr. Duffy explained that the official projects get archived, while nature runs do not. However, the OSSE runs get re-done every three years. The observational data is retained and can always be re-run. Dr. Kinter asked how decisions were made about where data sets reside, and how to field requests. Dr. Duffy noted that HPC allocations are done at the Headquarters level, via two calls per year. Data analytics will be managed similarly, and the program works closely with users to answer those questions. Dr. Holmes asked if the program has an organized approach for feedback. Dr. Duffy said the program holds monthly meetings with major users, integrated product teams, and tiger teams in a continual collaboration. Allocations change every 6 months. In addition, the program is leveraging as much interoperable software as possible.

Climate Analytics

Dr. John Schnase gave an overview on NASA’s climate analytics as a service. Earth Science Data Analytics refers to the processes and workflows useful for processing data specifically for scientific needs. Climate analytics serves the climate research community primarily, given that it operates on climate model data rather than Earth observational data. Currently, it performs the front-end data assembly part, concentrating on reducing data for science applications that best serve the climate research community. Right now, the analytics effort is focused on finding answers to questions that we know to ask, rather than knowledge discovery that involves finding answers to questions we may not know to ask. The analytics tier of NASA’s climate analytics software stack is used to produce the most commonly used information products. The goal is to do simple things fast, do work near the data source, apply capabilities to an interesting and useful climate science data set, create a service that enables community construction of advanced capabilities, improve efficiencies in upstream data processes of climate science workflows, and create a context for developing more advanced approaches. To date, NASA has focused on its development efforts on reanalysis data.

Part of the climate analytics rationale is in response to the International Panel on Climate Change (IPCC), a good place to go to see how the research community uses climate data. IPCC’s fifth assessment report contained many findings, but relatively few classes of findings; most IPCC findings involve statements about the values of climate variables, climatologies, and trends. These represent a small collection of data attributes that are the basis of an enormous amount of intellectual work in the discipline, and are essentially centered on classic descriptive statistics, such as maximum, minimum, average, variance, difference. High-performance climate analytics technologies that can compute basic, descriptive statistics can then be used to compute more complex data products of value, such as climatologies, anomalies, and trends. MERRA Analytic Services operates on NASA’s Modern-Era Retrospective Analysis for Research and Applications (MERRA) and is an example of these technologies in use.. The system combines high-performance cluster computing and storage with

MapReduce analytic technologies to implement its climate analytic services. Its canonical operations take as input a variable name, spatial extent and temporal extent, and then performs arithmetic, extracts spatiotemporal subsets, and models the data by computing the descriptive statistics and extended products described above.

MERRA Analytic Services's canonical operations enable virtual collections of realizable objects. It makes possible a converged approach to analytics and archive management. Its capabilities can be accessed through a RESTful Web service interface, based on an Open Archival Information Service (OAIS), which can be easily interfaced with standard web services. It can also be accessed through Python scripts that call on the Climate Data Services Python library. A sample use of this system replicated the data processes of a recent paper, Single Reanalysis Estimation of the Contribution of Irrigation to Precipitation (Wei, et al). At 8.4TB in size, at one time this task would have taken days, but with upfront data assembly, it took minutes. More complex case studies are under way (e.g., global energy and water balance).

"Simple and fast" can make a difference. NASA is achieving up to a three orders-of-magnitude reduction in the time it takes to assemble data for many tasks. In a hypothetical example based on the work that underlies the IPCC effort, an estimated 135 person-years of data assembly and reduction work could be reduced to an aggregate effort of 2 months. Dr. Kinter asked if there were any plans for future analytic capabilities, such as linking with linear algebra and statistical libraries. Dr. Schnase said that yes, he has graduate students getting ready to jump on these techniques in future ROSES proposals.

Advances in Big Science Data Projects Outside NASA

Dr. Larry Smarr presented a briefing entitled: Creating a Science-Driven Big Data Superhighway, describing a long-running activity that originated with the NSF. Given the physics of optical fibers, this effort should be able to create a system that rapidly connects science and researchers using a conduit that is separate from the Internet. This effort has been building for 15 years, and began as NSF's OptIPuter Project: Demonstrating How SuperNetworks Can Meet the Needs of Data-Intensive Researchers. In August 2003, the network achieved a 18Gbps file transfer out of the available 20Gbps. At the same time, the Department of Energy's was building ESnet's Science "DMZ," to develop data transfer nodes to create a secure system for rapidly transferring data between the nodes. This was followed by the creation of a Big Data Freeway on the University of California-San Diego (UCSD) campus using NSF grants and dedicated 10Gbps optical fibers. The network essentially gave to each scientist the bandwidth equivalent to that used by the 30,000 users on the campus Net. Flash I/O Network Appliances (FIONAs) are being used to transfer data in support of sequencing the microbiome. NSF has funded over 130 campuses to build local Big Data freeways (\$500K awards). The current goal is to create a regional DMZ out of the campus DMZ, building on 15 years of member investment in CENIC, which is an ideal backplane to link all these California campuses together, in collaboration with private sector and multiple cloud vendors.

Dr. Smarr described being in the current process of pulling together the Pacific Research Platform (PRP), creating a regional end-to-end, science-driven Big Data superhighway system, acting as Principal Investigator (PI) on a \$5M grant initiated in October 2015. FIONAs were used to link up the system. As of March 2015, the PRP has logged good performance (ranging at various nodes from 5Gb/s out of 10, to 36 Gb/s out of 40). The PRP point-to-point bandwidth map has shown huge improvement in the last six months and now has more or less uniform coverage. All the participants have signed letters of commitment to use this facility and will report on how it improves their science. PRP is being driven by multi-site data intensive research, including particle physics, biomedical “omics,” earthquake engineering, telescope surveys, and visualization sciences.

A Superhighway workshop in October 2015 hosted 130 attendees from 40 organizations. The first application will be the distributed IPython/Jupyter Notebooks, from University of California-Berkeley (UCB) to UCSD, across the PRP. The Cancer Genomics Hub at Santa Clara now has a data flow of 15 Gb/s. Another area will be telescope surveys: one project that is processing 300 images per night, and the other is a dark energy spectroscopic instrument. These efforts are applicable to NASA, obviously. The Large Synoptic Survey Telescope will track 40B objects and will produce 10M alerts per night, within one minute of observing. IBM is planning to build a dedicated exascale computer for the world’s largest, million-antennae radiotelescope. There are also experiments with PRP for the Large Hadron Collider Data Analysis using a portion of the West Coast Open Science Grid, aggregating a PB of disk space and Pflops of compute, connected at 10-100 Gbps, allowing transparent computation on data at home institutions and systems. In preparing for further climate change in California, UCSD climate researchers are planning to use the pipeline to download results from the National Center on Atmospheric Research (NCAR) in order to make regional climate change forecasts. The PRP will also be useful for downscaling supercomputer climate simulations to provide high-resolution predictions for California over the next 50 years, helping to plan for future electrical grid needs in response to hotter summers. The next step is to build a global research platform building on this effort.

Dr. Holmes asked Dr. Smarr to suggest any lessons learned for NASA. Dr. Smarr felt it would be good to figure out where there are similar efforts in NASA in employing things like DMZs, to allow university end-users of NASA data to link in to NASA rapidly and securely. If there were other internal experiments on the NASA Research and Education Network (NREN), it would be great to have NASA engaged with the PRP, as NSF intends to build this out as a national system. Dr. Holmes commented that while NASA doesn’t have the national charter (for this effort), it is true that there are people at NASA who are very cognizant of this effort, but are not active because there is no requirement base coming out of the science organizations. That said, a DMZ pipeline is a good effort to push for at NASA, and may be a potential finding. Dr. Smarr complimented Dr. Duffy’s briefing and expressed interest in his slides. He added that were likely to be researchers at SMD that would be natural colleagues with PRP, and he welcomed them with open arms. He added that the NASA Chief Information Officer was interested in this effort several years ago. Dr. Smarr offered to send out the PRP proposal to BDTF members,

and noted that there is a lot of unused capacity in existing national laboratory facilities (e.g., Oak Ridge, Argonne) that can be taken advantage of; this is another way to improve connectivity.

Data Policies

Dr. Holmes gave a briefing on the subject of Heliophysics data practices and how they had transitioned during a particular solar cycle (23). Reasons behind this transition of practice included developments such as the rise of the Internet and the influence of Moore's law, as well as a shift in attitude among PIs, and specific directions levied from Headquarters onto the community. The Heliophysics environment in the 1990s was limited to only six satellites in operation. From 1996 through 2008, many more satellites were launched, resulting in much more data covering the area from the Sun to the ionosphere. Data were stored in NASA-funded data stores, some of which were located at universities and national research laboratories. The virtual observatories grew, and resident archives were created to store data after missions ended. In 2007, the Heliophysics Division (HPD) approved a Science Data Management Policy: be modern, take advantage of current technology, evolve the existing technology, provide open access, and produce independently useable data (supported by easy access, analysis tools, documentation).

After the policy had been in place for some time, Dr. Holmes was asked to survey the research community with six questions. First, he asked if there were significant changes in data production and data access. The answer was a resounding "Yes." Before the policy change, PIs owned their data in personal fiefdoms. Each instrument represented a point observation in a time series, and data distribution was done by mail. After the policy change, the PIs acted as stewards of data sets, and provided open data sets with rapid availability. This open, Web-based distribution also provided valuable feedback to PI (hundreds of eyes see more than two, hence improved quality control, and verification and validation), as well as multi-platform analyses of extended complex phenomena.

Asked if there had been changes in collaborations and multiplatform studies, the surveyed population strongly agreed that the multiplatform analysis was much easier to manage. They felt that technology, combined with program guidance from Headquarters, all helped to bring about the cultural shift from PI ownership to stewardship; and the evolving role of data standards, as well as metadata standards.

In exploring the question of the respondent's role in the evolution of these changes, analysts, modelers, advisory committees, instrument teams provided unanimous affirmation of the data policy. In addition, respondents agreed that the policy helped to promote advances in the Heliophysics discipline itself, as closer collaboration and openness boosted the ability to create deep knowledge about the Sun-Earth system; most of the low-hanging fruit had been picked, and the evolution allowed for a deeper dive. Other comments received in Dr. Holmes's survey indicated that most felt that the evolution must continue, and that the Virtual Observatories had been valuable. One downside however is that one must beware of ignorant or careless data users.

In summary, the change in policy contributed to a significant transition in Heliophysics science data, with many benefits and some unintended consequences. Dr. Holmes was surprised that Headquarters had played a large role in guiding data behavior, and that community peer pressure had not seemed to be that important in driving the observed change. The fact is that the environment is enabling multiplatform, multidisciplinary investigations, and Dr. Holmes felt that the message behind the policy change was that Headquarters has a job to do, and should do it. Dr. Walker commented that at the time of the transition, PIs didn't want to be first or last (in adopting changes), and Headquarters policy effectively eliminated this hesitancy. Dr. Hurlburt added that the fact that the document was not onerous was a great help. Dr. Holmes noted also that he had received much community feedback before the policy was published. Dr. Beebe agreed that prior to the policy change, the Heliophysics community had been in an entrenched area, and didn't even discuss magnetic fields.

Discussion

Dr. Feigelson began the general discussion by asking how legacy efforts could best be integrated with new ones. Dr. Holmes noted that the Senior Review process could be used to determine whether such items as legacy software have outlived their usefulness. Archives are also subject to peer review and could include software capability and services. Dr. Kinter observed that there's a difference between an observational data set and a software system used to access and analyze it. Dr. Tino said that rapid prototyping, in the commercial world, typically doesn't go further than the initial effort. Dr. Walker noted that CCMC deals with a lot of runs-on-request (ROR) queries, and it appears that's what the community finds most valuable. In space weather applications, ROR is potentially used as a predictive tool. Dr. Tino said he would argue that both tasks are very similar and that ROR is really an extension of rapid prototyping- it's expensive to maintain infrastructure for some services. Dr. Holmes felt that CCMC operated on quite modest funding, considering the breadth of its operations. Dr. Beebe commented that you can archive an algorithm, but not a code; this is a big gap in planetary science; Headquarters had to write an overguide to require Cassini to write user guides, which turned out to be a very wise use of funds. Dr. Feigelson noted that post-facto algorithmic documentation is also important, which maybe enough for a finding. The idea of distinguishing algorithms from software is important, and the community should be informed that algorithms are substantive. Drs. Feigelson and Beebe agreed to write up a recommendation on the issue.

Dr. Kinter felt that the Duffy presentation appropriately recognized the need for higher resolution and its attendant memory issue, as well as recognizing that the data must be combined with tools (data analytics) and made broadly accessible to be useful for big data applications. Dr. Tino agreed that the roadmapping effort was also impressive and agreed to write a finding on the subject. He also mentioned that NOAA had partnered with Amazon or some other commercial provider to solve problems with hosting data sets and to provide compute at the same locality. Dr. Hurlburt added that data sets are getting too big to move; that which takes months to process at NASA currently could take days, given the right tools. Dr. Kinter noted that a particular problem for state

universities is that agencies are reluctant to provide compute clusters to universities; it is not a scalable solution. Dr. Schnase felt that the gains being made now are in computation, which will eventually make the end product easy to move. Alternatively, one could use a round of processing to create a smaller universe that can be shipped out as an image. There can be a tiered solution.

Discussion of NSF Big Data Hubs

The committee summarized their individual meetings with Hub representatives.

Dr. Kinter reported having had a good meeting at NSF with Drs. Fen Zhao and Alejandro Suarez, in regard to the Hubs and Spokes program; they are focused on providing socialization services, and putting together needs and capabilities. Spokes solicitations are still being reviewed, and awards will be announced shortly. Dr. Kinter tried to drill down on possible interactions with NASA. He learned that NASA PIs can submit proposals to the Spokes program, such as taking a Big Data problem and solving it in a vertical stack, and using the Spoke to outsource. Joint solicitations are also possible, although any such solicitation would have to be flowed under a specific NSF directorate, such as Geophysics. NSF is also interested in solving problems that have some societal benefit, such as hazards. Metrics of success for NSF will be the number of partnerships.

Dr. Tino described his experience with the Southeast Big Data Hub at Georgia Tech, meeting with Renata Rawlings-Goss. The Southeast Hub is trying to increase the impact of its investment in areas such as health care, energy and material science. There is also an effort to increase coordination. The Hub is holding workshop meetings with industry representatives, such as Georgia Power, big healthcare groups, etc., and is pitching research topics. They are also training academics to speak about subjects interestingly, how to talk to industry and demonstrate the utility of their research areas. Georgia Tech is planning to build a space in Atlanta that will be evenly divided between academics and industry, and is also launching an institute-level initiative. The Southeast Hub technically covers an area that includes Virginia through Texas. Georgia Tech is partnered with the University of North Carolina.

Dr. Hurlburt reported meeting with the West Hub, which is focused on medicine and a technology element, and which has not done much yet. Their current efforts include reaching out to industry and exploring the science of data-driven story-telling. Dr. Beebe added that the West Hub, spread out over a number of universities, hadn't started Spoke development, and said she heard rumors of a couple of states making moves to carry out desert research.

Dr. Holmes described his visit to the Northeast Regional Data Hub, where he met with the PI and Executive Director. The Northeast covers Maryland to Boston, and includes Penn State and the Ivy League consortium. The Hub is still in the process of organizing, but the PI expressed interest in Dr. Holmes's insight into the NASA research community. Spokes have not been selected yet. Dr. Holmes also spoke to Max Bernstein of ROSES, whose main job is to publish solicitations. He was very open to the idea of a cooperative solicitation between NSF and NASA. ROSES also serves as a messaging

service to the NASA research community, and Dr. Bernstein said he would be happy to broadcast Hub and Spoke information to the community. Dr. Holmes suggested tabling the issue until the Spokes are chosen, although the Task Force may want to entertain a draft finding to bring to the Science Committee. Dr. Smith agreed to vet this idea with the Executive Secretary of the Science Committee, and Dr. Holmes tasked himself to write a draft.

Findings and Recommendations discussion

Drs. Walker and Tino agreed to draft commentary on how SMD could be more proactive in pushing or incentivizing the NASA research community to interconnect and coordinate.

Dr. Hurlburt commented on the superiority of the Jupyter notebook concepts, Julia, and iNotebooks over the older workflow systems. The new system enables more interactive usage on networks, better captures the workflows, and facilitates communication to the research communities. Dr. Feigelson regarded Jupyter as a powerful vehicle for education and for software methodologies; it is interactive, includes multilingual codes, and can be hyperlinked. The Jupyter notebook can also be embedded into a blog that allows commentary and discussion. It is also multiplatform, and scalable to massively parallel. He felt it was an excellent fit and well suited to implementation by young people.

Dr. Holmes pointed out that his briefing on the evolution of data policies was meant as a pep talk. Dr. Hurlburt was disappointed with the lack of ensemble studies in the Heliophysics field, and felt some new frameworks could help bring about breakthroughs. Dr. Walker, referring to on magnetospheric studies, noted that looking frequently at plasma distribution functions used to mean a lot of re-processing, and that the big difference now is that missions such as Magnetospheric Multiscale (MMS) process these PDFs regularly, making them easy to access. This is a major step forward in magnetospheric physics from 2008, and is also fairly well documented.

Dr. Holmes reviewed his list of topics from the February 2016 meeting and asked members to champion one topic, and provide input for developing a work plan. Dr. Smith captured modifications on the list to send out to the group. Dr. Feigelson suggested that one can separate data reduction from the science analysis that comes later, and that this is applicable to all the divisions of NASA science. Is data reduction part of Big Data? Should it be part of Task Force purview? Dr. Smith agreed that dividing data reduction from science analysis is a useful concept, as it “gets to the wavy line before getting to the analysis of the wavy line.”

June 29, 2016

Opening remarks

Dr. Erin Smith opened the meeting. Dr. Holmes introduced the day's topics, commenting on a statement having recently emerged from the Astrophysics data archive review, and expressing grave concern about the data centers' infrastructure and the risk of its obsolescence.

NASA Astronomy Archive Report

Dr. Hashima Hasan reported on the results of a Senior Review on the Astrophysics archives, which include the High Energy Astrophysics Science Archive Research Center (HEASARC; extreme UV, x-ray), Mikulski Archive for Space Telescopes (MAST; visible to UV), Infrared Science Archive (IRSA), and the Infrared Processing and Analysis Center (IPAC). There are also specialist archives integrated into other archives, such as the NASA Extragalactic Database (NED; multiwavelength data and bibliography of objects beyond the Milky Way), NASA Exoplanet Archive (NEA), NASA Astrophysics Virtual Observatory (NAVO), and the Astrophysics Data System (ADS), a high-value research facility.

External reviews of the archives are held every 3-4 years. The latest were held in 2008, 2011, and 2015, and the next review is planned for 2019. Findings from 2011 resulted in a major restructuring of exoplanet data holdings. In 2015, the report identified several challenges, finding that the infrastructure and technological approaches being used will be obsolete within 4-5 years, and that network bandwidths available to data centers will soon be two generations behind the current standard for the research Internet. The latest archives to be reviewed were ADS, MAST, HEASARC, IRSA, NED, and NEA. The motivation for the review was to ensure that the Archives keep abreast of new technologies and techniques, develop innovative strategies, etc.

The review committee had broad expertise, including archivists in Earth and planetary science. Reviewers also requested that archives provide prioritized objectives, against which their progress could be assessed annually. Overall evaluations ranged from Good to Excellent. Responses to concerns identified by review include an augmentation of the ADS budget to improve services and functionalities. MAST is currently giving a high priority to increasing the speed of its network. NASA also modestly augmented the NED budget to enable a proposed Physics within Galaxies program and a Machine Learning experiment. HEASARC and IPAC are cognizant of the issues identified by reviewers. Key IPAC servers have been using a Gb/s pipeline for several years.

Dr. Walker asked about interoperability issues. Dr. Hasan deferred his question to Dr. Alan Smale's briefing later in the meeting. Dr. Holmes noted that one concern is the look forward 5 to 10 years from now, and asked how new ideas and new infrastructure were to be implemented. Dr. Hasan commented that in terms of speed and bandwidth, infrastructure refreshment is being addressed. IPAC, for example, is being augmented to include Cloud computing. Dr. Holmes added that complexity is also an issue; the BDTF concern is that Headquarters can be overwhelmed by day-to-day issues as well as incoming data from many missions. Dr. Feigelson noted that the VAO was an unusual NSF/NASA project, which was eventually viewed as troubled, and was disbanded. Operationally it did many good things, but the science coming out of VAO was weak.

ESO de-funded its virtual observatory for the same reason. The effort has now been assumed by the NASA archive system, and this makes sense, as NASA has a permanent staff and an established archive system. He asked Dr. Hasan to address the new NAVO from a management point of view, and asked whether the weaknesses that caused the demise of the former observatory been addressed. Dr. Hasan replied that this very question was raised in a 2014 review, and since that time NAVO has put much more emphasis on management, and gave GSFC the lead. GSFC has appointed a PM and a Project Scientist to manage the NAVO like a NASA archive. NAVO will be reviewed next year.

Mikulski Archive for Space Telescopes

Dr. Rick White, PI for MAST, presented a talk in place of Dr. Marc Postman. MAST was established in 1990 with the launch of the Hubble Space Telescope (HST). MAST is the archive for four active missions including HST and Kepler, and will also be the science operations center for JWST. It holds data from many legacy missions including Galaxy Evolution Explorer (GALEX), International Ultraviolet Explorer (IUE), and Far Ultraviolet Spectroscopic Explorer (FUSE). In the future MAST will house data from the Transiting Exoplanet Survey Satellite (TESS), Wide-Field Infrared Space Telescope (WFIRST) and JWST. MAST also holds some ground-based archives such as the Guide Star Catalog, and the results of the PanSTARRS survey (2PB). MAST data are diverse in data types (images, spectra, etc.), scale (small to large missions) and processing levels, and they encompass many different missions and instruments. HST alone has 12 different instruments and 17 kinds of detectors. The current “Era of Surveys” has been marked by the increasing sophistication of archive users, whose queries are crossing over archive and wavelength boundaries. MAST must evolve, especially with the launch of WFIRST, scheduled for the mid-2020s.

For MAST, Big Data means data whose raw form is so large that it must be qualitatively changed in the way it is reduced, stored and accessed. MAST contains 800 TB at present; storage will increase greatly with JWST, which in turn will be dwarfed by WFIRST. The price of storage is decreasing, while supercomputer and Internet speeds are increasing. Technology growth is outpacing astronomical imaging. The real limit is the capacity of the human brain, according to some. Big Data challenges for MAST will include automated source classification; often incomplete, multiwavelength data; time-domain analysis; model-fitting to large or complex data; and disentangling data.

Science uses cases to help prioritize MAST advances, as in computer needs, network needs, and common hardware requirements. In the current infrastructure of the Space Telescope Science Institute (STScI), there are five science servers. Servers can't easily be reconfigured to match different applications. Current large mission systems tend to be highly physical and have fixed capacity. JWST mission systems are being built with the idea that there will be a much more flexible, easily scalable, virtual system. In terms of bandwidth, MAST supports a mix of 1Gb/s to 10Gb/s, and is working on increasing internal bandwidth to 100 Gb/s. Plans are being made to upgrade standard Internet from 500Mb/s to 1Gb/s in the short term and to 10Gb/s in the longer term. Storage is currently 8PB, of which about 800TB is available online. Most internal storage is

enterprise-level and relatively inexpensive to support. PanSTARRS is supported by a custom system built by Aspen, which is basically a RAID system. Right now, MAST is trying to move everything to EMC systems that can be readily expanded.

Archives of the near future will be more than a collection of data. They will require a collection of services that will range from simple to complex, such as extracting information from spectra, stacking, etc. There will be a need for application programming interfaces. Cloud computing is clearly the wave of the future, and MAST is keeping an eye on its costs. While using the Cloud is still more expensive, it has the advantage of not needing to pay the hardware refresh costs. Given the amount of astronomical data in MAST, it will need to have server-side analyses, and MAST is now looking at ways to support advanced scripting capabilities.

The archive is currently looking at proposed capabilities by science case. Cases such as lensed galaxy detection or TESS supernovae searches will require a PB of memory and several thousand cores. MAST infrastructure is going to need significant improvements in internal and external bandwidth; and it is currently hampered by the lack of good network staff. It is hard to keep qualified staff on NASA payroll when they can readily earn so much more elsewhere. Multiple virtualized dynamically configurable nodes with upwards of 1000 cores each are needed within two years, and multiple 10k core systems will be needed in five years. The workspace must support a variety of tools, scripting language and software (MAST is working with Johns Hopkins University in this area), and must have good internal bandwidth between workspace, data sources and CPU cores. In five years, STScI will need to invest in a hybrid Cloud, possibly with JHU as well, and will need to explore, select and deploy machine learning architectures to support archive researchers. The institute will also need to support automated spectral feature classification tools and connect them to data processing. Server-side scripting will allow users to refine queries in real-time without having to download data for query and repetition. NASA's Citizen Science program is another interesting Big Data product, taking advantage of human pattern recognition abilities for such tasks as classifying bubbles in interstellar gas, parallelizing the human side in order to improve machine algorithms. NASA has initiated the Barry M. Lasker Data Science Postdoctoral Fellowship at MAST, and is also looking to increase the skill sets of staff on server-side projects, and develop expertise with networks and HPC, machine learning and automated classification techniques. In response to recommendations on organizational structure, MAST is establishing a Data Science Mission Office at STScI, which will have an archive Program Scientist and Project Technologist. There is a new position associated with this development, to head the Mission Office. These changes are also being driven by JWST mission priorities. A recent MAST Big Data report is available online. A Discovery in Astronomically Big Data workshop will be held February 27-March 2, 2017 at STScI. Workshop goals are to learn about new scientific discoveries, exchange skills for discovery methods in datasets, assess and reduce cultural barriers, and develop partnerships.

Dr. Walker asked if MAST had prioritized its tasks. Dr. White's personal perspective was that the new senior lead for the Data Science Mission Office will help create these

priorities. In the near term, MAST is focusing on the internal and external network, and seems to have a grip on storage. A flexible data center is working to assemble a large computing component. MAST is hoping for more opportunities in machine learning, and is looking for more connections to the community. MAST gets a lot of external input from committees that focuses more on short-term needs and requirements. It doesn't get a lot of visionary advice from user committees, which is an area where BDTF might prove helpful. MAST continues to piggyback a lot on JHU advances. MAST is pretty well integrated with the NASA computing system, although much communication happens at a low level. Dr. White was not aware of how much communication there was with centers such as Ames Research Center. Dr. Feigelson suggested that MAST consult more with engineering expertise in academia, to tap on experience. He recommended that MAST and NASA consult with an IEEE astronomy task force, and maybe put a small amount of budget funds toward consulting fees for these high-level experts. Dr. White welcomed this idea. Dr. Tino, speaking from his engineering background, felt that engineering could solve a science problem, but not necessarily build a platform. MAST's issue lies not in just solving one science problem; it has to scale up efforts in order to solve many different science problems. There needs to be a combination of systems engineering, subject matter expertise and development expertise to accomplish this goal.

High Energy Astrophysics Science Archive Research Center

Dr. Alan Smale, Director of the HEASARC, presented a briefing on the center, which was established in 1990 to support the broad science goals of the Physics of the Cosmos theme. HEASARC is now holding 90-plus TB of data. HEASARC's function is to maintain mission data sets and keep them secure and accessible, and is a research center as well as an archive. Active missions covered by HEASARC include Chandra, Fermi, the X-ray Multi-Mirror Mission (XMM), Hitomi, and Swift. Future missions include the Neutron star Interior Composition Explorer (NICER) on the ISS, and the Explorer missions in Phase A, Imaging X-ray Polarimetry Explorer (IXPE) and Polarimeter for Relativistic Astrophysical X-ray Sources (PRAXyS). NICER will produce data at the rate of about 30 to 40GB per year. HEASARC is not experiencing the huge rise in data volume seen in some other areas of astronomy, but has supported a broad range of missions over the years, dating from the beginnings of x-ray astronomy. The center curates complete mission sets in FITS and OGIP standard formats. Data and metadata are logically arranged, easy to retrieve, and ready to analyze. HEASARC also has readily available in-house experience, and enhances the return on investment for NASA missions and technologies. The most popular data downloads are from Fermi and Swift, but older data sets continue to be popular. Legacy astronomy data is still considered valuable, and papers are still being published on missions that have been over for ten years. HEASARC programmers coordinate one to two major public releases per year of the HEASoft data analysis software package. HEASoft is open source, and in 2015 had approximately 5000 registered institutional users. HEASARC is also seeking involvement with future Probe-class missions.

In response to the 2015 Senior Review of the archives, Dr. Smale agreed that the challenges that had been identified are real. However, HEASARC stays current with new

technologies, and the budget contains provisions for updates to servers, data storage and other necessary infrastructure. Currently, network management is transitioning to the Marshall Space Flight Center MSFC. Dr. Smale agreed that data centers need to raise concerns about sustainability, and averred that he would not be shy in making these concerns known. There is no current danger of saturating the network, as the existing system is adequate for the data it provides. Dr. Smith commented that OCIO is probably a better forum than SMD for the exertion of any necessary pressures. Dr. Holmes noted that the OCIO is the implementer of requirements.

In response to BDTF questions, Dr. Smale said HEASARC planned generally to a three-to-five-year horizon, and less so for a 10-year horizon, as it is not known what missions will be directed or competed. The center undergoes regular programmatic review processes such as the Senior Review, communicates often with Headquarters, receives regular user group input and carries out intra-Goddard discussions with code 700, and uses special working groups. Currently, its highest priorities are in improving data visualization and increasing interoperability; and there are no services it wishes to terminate.

NASA is conducting an implementation study of how HEASARC might operate in the Cloud. There is no cost reason to avoid going to the Cloud, but there are issues of security and data integrity. There are many details to work out, including how the center would charge for Cloud services. HEASARC envisions a gradual transition and transfer of deep archive copy from Iron Mountain to Cloud, and support for key services using the Cloud at relatively low cost. In terms of inter-archive cooperation, collaboration is the usual mode rather than competition. The NASA archives (including HEASARC, MAST, and IRSA) have their own Executive Committee, the Astronomy Data Centers Executive Committee (ADEC), which holds monthly telecons and employs rotating chairmanship.

The NASA Astronomical Virtual Observatories is committed to sustaining the infrastructure of the former VAO. The NAVO mission is to facilitate maximum science return for NASA astronomy data, by developing consistent and comprehensive access to NASA data through Virtual Observatory (VO) protocols; representing NASA and US interest in developing astronomy standards; and maintaining infrastructure needed to exploit VO services. The NAVO is invisible to most users, yet is an important means for discovering and retrieving NASA astrophysics data. HEASARC also provides monitoring and validation tools that ensure that data and services remain available through NAVO. In sum, HEASARC effectively serves the broader science community through providing an archive.

Responding to a question on security, Dr. Smale said that the problem was typically that hackers tried to post on the website. The data on the public side is not proprietary; all other data is encrypted. Dr. Tino noted that if the concern is securing the access portal, NASA just needs to implement the site properly. Dr. Smale commented in closing that HEASARC also makes extensive use of the NASA-funded ADS to acquire publication data for metrics; ADS is essentially an online astronomy and physics abstract service. Dr.

Feigelson felt this latter fact was worthy of an encomium from BDTF, and was tasked to draft a finding.

Earth Science Data and Information System

Dr. Chris Lynnes, System Architect at the Earth Observing System Data and Information System (EOSDIS), gave a briefing on the system. EOSDIS captures, processes, archives, subsets and distributes data. At present, EOSDIS is considering a number of Cloud prototypes to tackle volume challenges, but advances won't help much with the variety of Earth Science data. The archive and distribution volume has gone beyond 15 PB of storage, and on average EOSDIS has been distributing the volume of the total archive. FY2015 metrics demonstrate almost 10,000 unique products, and the archive is growing at a rate of 16TB/day, comprising 1.4B distribution products (files).

Cloud prototypes may be able to provide value-added services, such as data analytics. NASA has a public Cloud offering that is run out of Amazon, under OCIO. NASA has also built a system called the Next Generation Application Platform (NGAP), with a hardened platform. NASA has just gotten approval to operate with the general public accessing EOSDIS data and migrating some services, such as browse.

A killer app for EOSDIS may well be Cloud analytics. EOSDIS wants to make it available to the average researcher, to allow analysis of data at scale, analysis of combined datasets, while avoiding data downloading and data management. NASA is experimenting with supporting a toolbox to attract users to Cloud analytics, such as community open-source tools, Distributed Active Archive Center (DAAC)-developed tools, Cloud analytics examples and recipes. An initial cross-DAAC proof of concept is in progress, based on Python and Jupyter. In the community, there is some eagerness to move to the Cloud, particularly with early career scientists who haven't invested in servers and other computing infrastructure.

Cloud risks are vendor lock-in and future storage costs (the latter is somewhat unknown- is Kryder's law flattening out?) uncapped egress costs (there are by-the-byte charges in the Cloud vendor structure), security restrictions (remains a persistent issue), and network trust (what is running in Cloud and connecting back to private resources?). NASA is tracking these issues.

Data variety continues to increase as time goes on. Data is coming from a variety of instruments. Synthetic aperture radar data is completely different from that which comes from a profiler. Images come in numerous wavelength regimes, which leads to very different footprints from satellite instruments, some of which are very difficult to visualize, and to integrate. Other issues arise within same instruments and satellites with different algorithms. Processing levels also present issues: there can be calibrated radiance at a pixel-level (Level 1B), carbon monoxide for one scene (Level 2), and global carbon monoxide for one night (Level 3). Products also come with different surface projections. Re-projection of Earth Science data also holds some risks; the science rationale behind the need for and reasons behind different projection modes is not well understood in the wider community. Time and spatial aggregations are another variety

problem. Data formats also vary widely: ASCII, binary, self-described API-based formats, and conventions used in the community such as HDF-EOS and climate forecast coordinates.

Some solutions to the variety problem lie in having interoperable discipline-focused DAACs, a common metadata repository, and other integrated solutions. There are 12 DAACs with different spheres of influence. Each DAAC serves one or two disciplines, and is able to understand the needs of the particular facilities they serve. The common metadata repository presents a consistent catalog for discovery of data from multiple DAACs. To address the file format issue, EOSDIS uses OPeNDAP, which enables access to data and smoothes out format heterogeneity, and supports subsetting. Data transformation options of several kinds can help with both variety and volume. The Big Earth Data Initiative (BEDI) is being used to improve dataset consistency across EOSDIS, and improve machine access to EOSDIS through a developer's portal. NASA has been doing a lot of community engagement to disseminate this information. Earth Science Information Partners (ESIP) works on variety and volume through clusters on discovery, information quality, Earth Science data analytics and Cloud computing. The Earth Science Data Systems Working Groups, composed of DAACs, is also working on the issue. There will be several workshops as well.

For now, EOSDIS pays a lot of attention to help tickets and uses advisory groups, webinars, and feedback from end-users to shape its response. The take-home message is that EOSDIS is looking to Cloud prototype efforts to find solutions for volume. For variety, it is using its experience and is working hard on interoperability issues.

Asked about OpenDAP scalability issues, Dr. Lynnes noted that for data transformations, one facet is missing, and that is server-side analysis. OpenDAP is caching data and using block storage. EOSDIS believes Cloud providers will really optimize http. Some other solutions are to break files up into Web object storage, or Hadoop files. For server-side analysis, the Jupyter Hub provides Python tools and a hub as a showcase, and runs analytics routines in the Cloud. The current system is Giovanni, which is used at 5 or 6 of the DAACs, where it provides arithmetic functions, climatology, and time series functions. Some scientists are uneasy about data products being produced in this way; it's regarded as "pushbutton" science. EOSDIS will probably have all of these capabilities in the long run.

EOSDIS has formed a User Needs Analysis Group, and is developing a database of all the data sources. The hope is to sort out the database results in August, and feed that input into a systems engineering group, which will then implement recommendations at the DAACs. Dr. Kinter commented that optimizing I/O through some types of compression of EOSDIS data has made a great difference in data analysis, and offered to write a finding to this effect. Dr. Holmes asked if interoperability of allied data sets between NOAA and NASA had been achieved. Dr. Lynnes said that the two agencies have worked together for a long time; the BEDI money helped fund some implementation of this interaction. NASA was the sole agency that got appropriations to improve the interoperability of data, but Dr. Lynnes expected that NOAA will be getting similar

funding down the line. Dr. Holmes asked Dr. Lynnes to provide suggestions to Dr. Smith, to incorporate into the BDTF report to the Science Committee.

Planetary Data Service

Dr. Ed Grayzeck presented a briefing on the growth of the Planetary Data Service (PDS), which stands at about 1PB today, and is expected to reach at least 2PB by 2022, particularly with numerous extended missions in operation. Planetary data also involves a lot of variety, a diverse community and diverse data sets. The PDS collects, archives and makes accessible the digital data and documentation produced from Solar System exploration from the 1960s to the present. PDS is a federated system with science discipline nodes that include atmospheres, engineering, geosciences, cartography and imaging sciences, plasma-planetary interactions, small bodies and ring-moon systems. The imaging sciences node is driving data growth, for the most part (lunar and Mars imagery). Currently, PDS has 40M data products from 625 unique instruments. The Lunar Atmosphere and Dust Environment Explorer (LADEE) produced 2 million data products. LADEE carried a laser communications demonstration that sent data from the Moon to Earth at >622 Gb/s. PDS also ties in with older data digitized from analog archives (Apollo, et al). Big Data challenges at PDS derive from variety of disciplines, moving targets and data; volume of data including provenance; and the federation of disciplines and international interoperability. These factors can affect data consistency and storage, computation, movement of data, and data discovery and distribution. There have also been issues with PDS versions 3 and 4 in data migration and overlap. PDS can eventually drop PDS3 tools, but it is critical that PDS adopt an architectural strategy that can scale up to support the size, distribution and heterogeneity of planetary data.

PDS4 is the upgrade from PDS3 that was developed to address architecture challenges. PDS4 is an explicit information architecture based on XML, using a hierarchy of dictionaries built to the ISO11179 standard. The PDS4 Information Model defines the data and its relationships, and is enabled by an Information-Model driven approach, where the Information Model is the cornerstone of the system. The Information Model can handle multiple disciplines, with multi-level governance that allows independent extensions, and employs multiple models integrated into an overarching ontology. PDS4 is enabled by a set of core software services for registration, search and distribution. PDS also involves international cooperation through the International Planetary Data Alliance, which was founded in 2006 and includes all major space agencies involved in planetary science data archiving. Its mission is to build compatible, interoperable planetary data archives. ESA's Planetary Science Archive has made a major investment and buy-in to use PDS4. While ESA has opened its access to planetary data, JAXA has been a bit less responsive to the open access movement. International collaborations that PDS4 will be supporting include LADEE, InSight, and BepiColumbo. The Planetary Cloud Experiment is under way with the Mars HiRise camera, although there is uncertainty in the cost model. Data movement can be a challenge (data to and from cloud) in this respect.

PDS4 allows user analytics: data classification (missions, instruments, targets) and

trend analysis. NASA is hoping to establish an international platform for planetary data archiving, data management and research. In addition, the PDS community is developing a 10-year PDS Roadmap, commencing with a meeting in Summer 2016, with the final Roadmap to be delivered in Summer 2017. The Roadmap will identify areas of improvement such as mission pipelines, search capabilities, and tool improvement. A Big Data workshop is planned for the 2016 IEEE conference. There are numerous other meetings, such as IPDA July 27-29; COSPAR July 30-Aug 7; an LPSC workshop; and a Planetary Interoperability workshop at the Division for Planetary Sciences of the American Astronomical Society, in the planning stages.

The community Roadmap looks out 10 years, and each discipline node has an assessment group, with priorities set by the PDS Management Council. Asked what features could be stopped, Dr. Grayzeck felt the PDS3 tool could be phased out, given input from the missions and community. PDS continues to take steps to make data interoperable, and is engaged in a project with MAST to provide pointers to HST data, as well as working with the Mars MAVEN mission.

Dr. Beebe commented that while researchers were used to PDS3, the problem was it was based on a level developed inside JPL. Opening up PDS4 to XML allowed international partners to share in the benefits, and this helped to resolve problems. She added that interoperability is the only way that planetologists can get their jobs done within the budget. It's one self-integrating system that works well and is poverty-driven. Asked if any nodes were having issues pushing data out, Dr. Lynnes reported that the Flagstaff microwave link data node (cartography and imaging) needs an upgrade. It is the slowest link, but a holder of much of the image data. Dr. Holmes suggested BDTF issue a positive finding on PDS.

Space Physics Data Facility

Dr. Bob McGuire briefed BDTF on the Space Physics Data Facility (SPDF), one of two active final archives in Heliophysics, which functions to serve and preserve data with metadata and software, and to understand past, present and future mission data. The NASA Space Science Data Coordinated Archive Center (NSSDC) is continuing recovery of older but useful legacy data from media. SPDF focuses on non-solar missions and data. The Heliophysics Data Environment provides the critical infrastructure. SPDF is also a Center of Excellence for enabling Heliophysics science, with an emphasis on multi-instrument, multimission science in context of other missions and data, and as enriching context for other data. Specific services include the Heliophysics Data Portal, SSCWeb and 4D Orbit Viewer, and OMNIWeb Plus. SPDF uses a Common Data Format (CDF), and also provides implementation guidelines. In the archive, there are master CDFs to update or add metadata and capabilities. The CDAWeb interface lists missions and instrument types and input parameters.

In FY16 mission and data highlights include ingesting and serving data from operating Heliophysics missions; acquiring data from past missions such as Polar VIS, Tether, NOAA missions that are Heliophysics-relevant (such as the Deep Space Climate Observatory, DSCOVR), and orbital information for the SSCWeb database. SPDF also

ingests and serves older data from NSSDC, and holds regular interactions and planning sessions with upcoming missions to support new ingests, new data and new capabilities. Metrics indicate that SPDF's usage is climbing slowly and steadily. About 33% of 2015 papers in AGU's JGT Space Physics acknowledged SPDF services or data; this is up from 25% in previous years. By volume, SPDF is about 120 TB, thus relatively small. Dominant data ingests are from MMS and the Van Allen probes. The Global-scale Observations of the Limb and Disk (GOLD) data rate will be comparable to MMS, which has been getting significant downloads.

SPDF supports many missions, some in a fairly complete fashion and others, less so. The facility deals a lot with re-processing data. In terms of complexity and heterogeneity, Heliophysics covers a wide scope of science problems and measurement techniques; currently the facility holds 1400 individual data sets and 30M files, with 25,000 individual parameters. Mission and community acceptance of data standards is key, and SPDF actively works with the missions in using these standards. As a result, missions have actually adopted stricter standards than proposed. The technical architecture has one public facing server with multiple hosts, and a pipeline of 10Gbps, with a private switching capability. Regular tape backups are made of data are made and stored with Iron Mountain.

SDPF is a team effort of data providers, instrument builders, and software scientists. SDPF expects to be able to manage currently expected data volumes, and support the wider Heliophysics data environment. It is currently supporting infrastructure for analysis software such as Autoplot and SPEDAS. SPDF will consider Cloud usage for the next major storage increment, the cost of which is still an unknown. The facility is considering how Cloud technology might support a long-term archive. Archiving will certainly require a long-term approach for older data. SPDF's current scope is observational and not model outputs, and is currently erring on the side of "doing something."

In answer to the Task Force's specific questions, Dr. McGuire said it was hard for SPDF to plan beyond 5 years; its priorities remain missions and data. The facility does not have an advisory group, and can identify nothing it wishes to cut back at this point in time. Interoperability depends on resources, priorities, and communication with the PMs.

Solar Data Analysis Center

Dr. Joseph Gurman gave a briefing on the Solar Data Analysis Center (SDAC), which is experiencing data growth. DKIST, a ground-based solar telescope on Haleakala, HI, is expected to far more produce archivable data than the Solar Dynamics Observatory (SDO). The SDAC doesn't have an entire Big Data archive; it holds tens of TB for Hinode and the Solar Terrestrial Relations Observatory (STEREO). Its largest single mission holding is SDO at 1PB. The archive is characterized by the "three Vs" and 7-8 wavelengths. It's hard to tell where it will be in five years, but at some point, the SDAC will be named the long-term archive for the entire SDO data set, which is not supportable by current storage architecture.

In responses to BDTF questions, Dr. Gurman characterized SDAC as mission-oriented, while VSO is user-oriented, and reaches out to the community to identify needed data services. New features are typically driven by available technology. SDAC would like to get out of specifying new storage on multiple tiers every 5-7 years, and would like to stop conflicting agency directives. Interoperability with solar data is already in place. Heliophysics data interoperability with other disciplines, however, will require the help of other specialists. SDAC data are currently interoperable with any other modern solar physics data (e.g., FITS, IDL, Python). In terms of allied data sets, Dr. Gurman felt SDAC was capable of interoperability. Eventually, SDAC will look to Cloud for storage, distribution of SDO data, dedicated hardware for large databases, and mission-oriented data. Currently, Amazon charges a lot for access to data. Upwards of 95% of the data with which we deal are in FITS format, a standard in astrophysics and solar physics. On SDO, the data from the HMI instrument (which produces about 49.5% of the SDO data flow) are stored in a database at the Stanford JSOC, but will have to be exported as FITS files (a capability of that database) when long-term archiving is implemented, because neither the SDAC nor any other facility has the staff to keep the unique database software going.

Public Comment Period

No comments were noted.

Office of Chief Information Office (OCIO)- Innovation and Technology Division

Dr. Brian Thomas presented a briefing on data science from the OCIO perspective, whereby he described NASA as essentially a data science agency. Data science is defined as an interdisciplinary field for extracting knowledge from data, moving data to information to knowledge, and finally to wisdom. The goalposts keep moving in Big Data, in terms of all three Vs—volume, variety and velocity. The Decadal Survey in astronomy stated that the great discoveries are expected to come over the time domain, and the variety of data over time will be a key problem. Data scientists use sophisticated algorithms, math and statistical knowledge, and substantive expertise; it is often necessary to have more than one person to provide these three skills. NASA OCIO uses a Data Team and a Data Strategy (a white paper on data management), unified data lifecycles, data governance, data analytics lab, a data fellows program to bring in expertise, and data stewards. Dr. Thomas had spent 15 years away from Goddard, and was shocked to see how little IT had progressed in his absence.

The OCIO has numerous projects, one of which is aiding in organizing and integrating extravehicular activity (EVA) data integration. Other current projects include text analytics, i.e. trying to classify documents within their domains so that they are findable and searchable. Another is the Environmental Management System (EMS), which is building an intelligent system to assess astronaut health in remote space. NASA is also building microservices to get away from the data silo problems. NASA has all types of “Vs,” and variety is the hardest from a technical standpoint. NASA implementation of data analytics is uneven, and there is a need for much greater data sharing. Major Big Data concerns for data science include network throughput, code shipping ability, cost-

effective Cloud computing, infrastructure support for governance and quality, and discovery of data at the Agency.

Data silos and data provenance are the pressing issues, currently. Big Data will often require shipping analysis to the data. How do we describe our data better and work with it meaningfully? Earth Science has been exemplary in this area, and is hosting the NASA Earth Exchange (NEX) at Ames, an immersive environment where scientists can collaborate and easily access multi-mission data. NASA can do a better job at having the machine understand the data, with a little bit of machine learning or deep learning techniques, in getting scientific data to the scientist. Machine-learning efforts in Earth Science include publication of a common data model for 80 different missions. Earth Science taxes each mission to label and tag their data, and uses auto-tagging or contextual tagging for data in other disciplines.

Data fellows in the OCIO are currently housed at Headquarters; they visit centers to understand the various competencies and challenges, and carry out some of the described projects. The goal is to have them in-situ at the various centers. There is a lack of people with certain expertise within the centers; data science fellows can act as ambassadors into the NASA enterprises. The Agency needs to know how to fish for the required expertise in data science. To address interoperability, the OCIO Innovation and Technology Division holds hackathons, and has a “datanauts” program, in which links are made with all the publicly available data.

Discussion

Dr. Feigelson mentioned an organization called Statistical and Applied Mathematical Sciences Institute (SAMSI), which is an NSF-funded concern that periodically brings together discipline scientists, mathematicians, and statisticians. Its next meeting will be in astrostatistics, dealing with subjects such as time series analysis, gravity wave detection, and ground-based Big Data (surveys). He suggested SAMSI serve as a potential resource NASA could use to tap expertise and with which to collaborate. Dr. Holmes asked Dr. Feigelson to research how NASA could engage with this program, results of which would be discussed at the next meeting.

Members offered their opinions of the day’s proceedings. Dr. Feigelson felt the Astrophysics archives were viewed positively by the community, and rated reasonably by the Senior Reviews. Dr. Beebe seconded these thoughts, adding that the community liked the “invisibility” of the NAVO.

As to MAST, looking at data analytics and server-side analytics, Dr. Holmes felt there was a possible finding there. He also wanted to further investigate the inconsistencies between some facilities’ abilities to get desired bandwidth (1 Gbps vs. 10 Gbps). Dr. Tino felt some of the issue there was the backplane and infrastructure. He expressed his support for ESDIS moving to Cloud computing, and thought the BDTF could recommend that they look at non-public Cloud offerings for protected data sets. Dr. Kinter felt ESDIS had gone the farthest in their analysis of Cloud risks, as well as integrating compression into data formats. The HIRACS server—building functions under data types—is also a

very interesting way to subdivide the beast. OpenDAP is an issue, a layered protocol that introduces performance drag and may have a scalability issue; it's good to see ESDIS recognizes this and has an open mind for solutions.

Dr. Tino got the impression that NASA uses some architectures and data formats because they've been doing it for a long time, and felt there had to be a paradigm shift in how it examines data sets to be able to do better science. There should be a way to present data sets in Cloud-based architecture to support data analytics workloads.

Dr. Hurlburt agreed that an alternative approach was needed. Dr. Kinter noted that one solution in the modeling community is to store the data twice: e.g., one synoptic, and one in time-series form. The associated cost, of course, is at least 2x.

Dr. Holmes felt that the common message was that the data centers were all doing an outstanding job, and that BDTF produce a laudatory finding. Dr. Tino commented on ingrained behavior that drives some data issues, which is a generational issue: "some people like to download data." This ingrained behavior will hinder progress, and will also make it difficult for NASA to attract new talent. Referring to the OCIO briefing, Dr. Tino commented on the persistently stovepiped nature of science data. Big Data is really about business outcomes and science outcomes. The way NASA is constructed, it's a challenge across stovepipes. How do you know what's in each of those stovepipes to determine their needs? Dr. Tino felt that NASA needed to go beyond conferences for sharing data- proper application of Big Data methods has the potential to get insights across fields, to help tease out patterns and potential tools for discovery.

Work Plan Topics Discussion

Dr. Kinter shared his notes on work flow; he averred that NASA's Earth System modeling is based on a work flow developed in the 1970s; this observation was *apropos* of a conversation with Gavin Schmidt at GISS, based on a case study on data analytics re: tropical cyclones. These workflows are not scalable for optimizing time to solution. Technology is beginning to address this problem. NASA needs dedicated hardware with serious compute power, and must be able to assemble all the data sets for doing real server-side analysis. Instead of reinventing the workflow for every scientific problem, NASA needs to generalize some of this work. If you start with a hypothesis, and make a model to run and test, the data could all be put into an external loop; looping can be done much faster with some injection of observational data. Dr. Tino felt that Dr. Kinter's observations dovetailed with changing how people interact with technology—moving away from products to outcomes. It would be valuable to identify who's moving in the right direction, to figure out how to derive results from data without shipping data around. When Dr. Tino thought of Big Data, he felt it was about people getting results, rather than distributing data. He thought the CCMC and Climate Analytics work were steps in the right direction; they're providing abstractions. There also seems to be a consensus that data locality is a problem.

Dr. Feigelson's impression was that entire branches of NASA SMD don't have a Big Data problem. He had not heard complaints from Astrophysics concerning data processing.

The overall picture seems to be favorable. Dr. Holmes felt that some communities were starting to reach a bottleneck, which would worsen over time. Dr. Smith commented that right now, Astrophysics is able to hand off Level 2 data, but this won't be feasible as data volumes increase. Dr. Tino felt there was a war between abstraction and data manipulation. Dr. Holmes reported that Dr. Gavin Schmidt remarked that NASA needs to move away from data downloads as an effectiveness metric. Now it becomes the ratio of data sent out compared to raw data, processing as much raw data as possible, locally.

June 30, 2016

Cloud Computing Initiative

Ms. Karen Petraska, Program Executive for Computing Services in the OCIO, presented a briefing on the OCIO-provided Cloud service that has been set up for the benefit of the Agency. In 2011, the Office of Management and Budget (OMB) issued a Cloud-First initiative, requiring that each Agency develop apps in the Cloud within a certain time frame. Issues with security and searchability quickly became apparent during implementation of this initiative. To manage the sprawl and security of data, NASA decided on an enterprise-managed approach to Cloud-computing. This approach helped to unify the interpretation and fulfillment of requirements. NASA put a framework in place, analogous to setting up a data center, and let everyone use the capabilities. This constituted a top-down approach to structuring the service, and helped NASA focus on consumption of commercial Cloud services instead of building Agency clouds, with standardized Agency governance. OCIO performed considerable work on technical integration, with an eye to working interfaces only once. The primary provider is Amazon, with East and West Coast direct connects. When a NASA civil servant needs access, there are no new requirements. This ensures that NASA security operations have insight into the data going back and forth. NASA framed out patching, scanning, and continuous monitoring practices that have been blessed by security. Piv badges are used for access, which supports various authentication architectures into the Cloud. FedRamp compliance is used, enabling templates that people can build on. Procurement is done through the NASA Soup contract, which uses a pre-competed set of vendors, that also includes aggregated attention to contract details, such as data rights, so that individual users have the proper protections. NASA has adopted a pay-as-you-go approach.

Essentially, the enterprise approach created a new boundary, to enable the Cloud to function as a data center that does not have a physical presence. NASA uses the term Managed Cloud Environment (MCE) to describe a sort of a community or platform where the tools reside. All data is encrypted through VPN to the Amazon cloud that is integrated with NASA IT. There is also a tiered Cloud services architecture so that engineering or science can come in and build whatever they need. The wholesale and retail platforms provide core capabilities that extend across Agency centers, programs, projects and communities, to unify the delivery of services across the Cloud Services Network. Marshall has been migrating some of its business tools into the cloud, and there is also a General Purpose platform available.

Ms. Petraska highlighted an innovative use of the Cloud, an Advanced Information System Technology MCE that provides AWS services to PI-led projects. The MCE is managed with a high degree of automation, and allows the PI to provide only a WBS-plus estimate. Dr. Holmes noted that this would be a good way to get access to HPC, and for allowing access for foreign nationals and non-NASA personnel with NASA grants. Dr. Mike Little interjected that this particular MCE uses a different cost-accounting model, and that it is a sociological problem—the current set of projects is learning how to estimate their costs in using Cloud computing. There are some tools to help them estimate the costs, but the objective is to get people smarter about estimating these costs in the next ROSES calls. The Cloud cost is about 1/100 of their costs, as opposed to overhead costs as typically computed in university grants. Dr. Holmes invited Dr. Little to give a briefing at the next meeting, centering on how the community can take advantage of this service.

Ms. Petraska indicated that the average user of the MCE saves about \$250K by leveraging the CSPO Cloud Services Framework, and by avoiding costs for security plan development through FedRamp compliance. NASA's future is in the Cloud; the general consensus is that within five years, 75% of all new projects would be born in the Cloud. NASA is already starting to see missions that are doing this. In addition, 100% of NASA's public data will be served up from the Cloud, and up to 40% of legacy systems will be migrated to the Cloud on a lifecycle modernization timetable. Success criteria for Cloud adoption are: on-board two communities per year for five years, and the demonstration of a rich representation of NASA's overall business being done in the Cloud. Dr. Kinter, remarking on the project's ambition, asked whether the new model supports the requirement for persistent exposure of data sets that are funded by NASA research projects. Ms. Petraska replied that the idea would be that the data would be publicly available. NASA is discussing cheaper types of storage; everyone is looking at these problems now. Dr. Feigelson said he hadn't heard much about OCIO engagement with the Cloud, and was happy to hear there is a centralized effort under way. Dr. Holmes observed that the buy-in process had not yet matured, presenting a challenge for the OCIO. He asked to see details of the migration plan in the future, which may incur unforeseen costs. Ms. Petraska noted that migration would be done on a case-by-case basis, and each project would have to address the risk individually. Dr. Tino asked how hardware refreshes were handled in the legacy system. Ms. Petraska indicated that this is where legacy cost comes in. She did feel many people at NASA were not exploring Cloud to the extent to they would find out where it resided at NASA, and that for now it was better to stand up the system, acquire some successes, and then socialize it. She reported she'd been seeing increasing acceptance of the system.

Work Plan Finalization

Drs. Kinter, Walker and Clayton were assigned to the topic of modeling work flows. Drs. Holmes, Hurlburt, Feigelson and Clayton were assigned to server-side analytics (for data reduction), addressing the climate studies, solar astronomy, and era of the WFIRST survey observatory.

Drs. Beebe and Walker were assigned data discovery.

Drs. Hurlburt and Feigelson were assigned improved data/science analysis methodology (and technology), or how to analyze data without touching the data.

Dr. Tino suggested as a further topic a taxonomy of Big Data, as an exercise in how to classify Big Data challenges and to determine some common thread— is it variety, velocity, volume? He took an action to work on this topic.

Dr. Kinter suggested a need for an over-arching vision or mission statement.

Dr. Holmes requested that each lead create a study plan by 19 July.

Findings and Recommendations

A potential finding on Education was transformed into an agenda item for the next meeting, based on Dr. Feigelson's suggestion for having NASA research scientists become minimally trained in computer science and statistics.

Members concurred on a finding on formatting mergers, and evolving the standard formats of NASA-relevant data. Dr. Holmes considered elevating this finding SMD-wide, because this evolution of data standards will be beneficial to the entire science community. Dr. Feigelson noted that the FITS format is regulated at the international level, and is therefore not a NASA responsibility.

Members concurred on a positive finding praising the Pacific Research Platform and Big Data Science, based on Dr. Smarr's presentation.

Members concurred on a finding on HPC bifurcation, looking at a major architectural shift in the way NASA does business.

Members concurred on a positive finding on ADS, reflecting universal appreciation for the service.

Dr. Holmes tabled a finding on algorithms, pending further review by Drs. Feigelson and Hurlburt.

Dr. Hurlburt noted that he had previously thought that the vision and plans for data management across divisions had been insufficient, and summarily withdrew this observation. Dr. Holmes agreed to express this very sentiment at the Science Committee. He noted also that he would include in his report a slide of the publications associated with MAST, while noting the share of archival papers is still increasing; this is an excellent statement on the value of MAST, and the other archives.

Conclusions

Dr. Holmes adjourned the meeting at approximately 11:30 am.

Appendix A Attendees

Ad Hoc Big Data Task Force Members

Charles P. Holmes, **Chair**, Big Data Task Force
Reta Beebe, New Mexico State University
Dr. Eric Feigelson, Pennsylvania State University
Neal Hurlburt, Lockheed Martin
James L. Kinter, George Mason University
Clayton Tino, Virtustream, Inc.
Ray Walker, University of California at Los Angeles
Erin Smith, **Executive Secretary**, NASA HQ

NASA Attendees

Myra Bambacus, NASA
Joe Bredekamp, NASA ret.
Marge Cole, NASA ESTO
Ed Grayzeck, NASA GSFC
Joe Gurman, NASA GSFC
Hashima Hasan, NASA HQ
David Liska, NASA STScI
Michael Little, NASA ESTO
Bill Knopf, NASA HQ
Masha Kuznetsova, NASA GSFC
Dawn Lowe, NASA
Christopher Lynnes, NASA
Bob McGuire, NASA GSFC
Karen Petraska, NASA OCIO
John Schnase, NASA GSFC
Aaron Roberts, NASA GSFC
Larry Roelofs, NASA GST
Alan Smale, NASA GSFC
Linda Sparke, NASA HQ
Brian Thomas, NASA HQ
Tina Tsui, NASA GSFC
Rick White, NASA STScI
Chiu Wiegand, NASA GSFC
Darrel Williams, NASA GST

Non-NASA Attendees

Amy Reis, Ingenicomm
Joan Zimmermann, Ingenicomm

Appendix C Presentations

1. Welcome to The Sciences and Exploration Directorate @ GSFC; *Colleen Hartman*
2. Community Coordinated Modeling Center; *Masha Kuznetsova*
3. HPC and Climate Model Data; *Daniel Duffy*
4. Climate Analytics as a Service; *John Schnase*
5. Advances in Big Science Data Projects Outside NASA; *Larry Smarr*
6. Lunch talk: Astrostatistics and Astroinformatics; *Eric Feigelson*
7. Evolution of Data Policies and Practices within NASA's Heliophysics Program (Solar Cycle 23); *Charles Holmes*
8. NASA Astronomy Archive Report; *Hashima Hasan*
9. Mikulski Archive for Space Telescopes; *Marc Postman (presented by Rick White)*
10. High Energy Astrophysics Science Archive Research Center; *Alan Smale*
11. Earth Science and Data Information Center; *Chris Lynnes*
12. Planetary Data Service; *Ed Grayzeck*
13. Space Data Physics Facility; *Robert McGuire*
14. Solar Data Analysis Center; *Joseph Gurman*
15. OCIO-Innovation and Technology Division; *Brian Thomas*
16. Cloud Computing Initiative; *Karen Petraska*

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

June 28-30, 2016

**NASA Goddard Space Flight Center
Building 28, Rm E210**

**Agenda
(Eastern Standard Time)**

Tuesday, June 28

9:00 – 9:30	Opening Remarks / Introduction	Dr. Erin Smith Dr. Colleen Hartman Dr. Charles Holmes
9:30 – 10:00	TF Member Reports	TF Members
10:00 – 10:30	Community Coordinated Modeling Center	Dr. Maria Kuznetsova
10:30 – 10:45	<i>BREAK</i>	
10:45 – 11:30	HPC and Climate Model Data	Dr. Daniel Duffy
11:30 – 12:00	Climate Analytics	Dr. John Schnase
12:00 – 12:45	Advances in Big Science Data Projects Outside NASA	Dr. Larry Smarr
12:45 – 1:45	<i>LUNCH</i>	
1:45 – 2:15	Data Policies	Dr. Charles Holmes
2:15 – 2:45	Discussion	
2:45 – 3:15	Discussion of NSF Big Data Hubs	
3:15 – 3:30	<i>BREAK</i>	
3:30 – 4:00	Scientific Visualization Studio Tour/Demonstration	Dr. Horace Mitchell

NASA Advisory Council Ad Hoc Big Data Task Force, June 28-30, 2016

4:00 – 5:00 Discussion

5:00 ***ADJOURN FOR DAY 1***

Wednesday, June 29

9:00 – 9:30 NASA Astronomy Archive Report Dr. Hashima Hasan

9:30 – 10:15 Mikulski Archive for Space Telescopes Dr. Marc Postman

10:15 – 10:30 ***BREAK***

10:30 – 11:00 High Energy Astrophysics Science Archive
Research Center Dr. Alan Smale

11:00 – 12:00 Earth Science Data and Information System Dr. Chris Lynnes

12:00 – 1:00 ***LUNCH TALK: Astrostatistics*** Dr. Eric Feigelson

1:00 – 1:45 Planetary Data System Dr. Ed Grayzeck

1:45 – 2:15 Space Physics Data Facility Dr. Robert McGuire

2:15 – 2:45 Solar Data Analysis Center Dr. Joseph Gurman

2:45 – 2:55 ***BREAK***

2:55 – 3:00 Public Comment

3:00 – 3:30 Office of Chief Information Office-
Innovation and Technology Division Dr. Brian Thomas

3:30 – 4:00 Discussion

4:00 – 5:00 Work Plan Topics Discussion Dr. Charles Holmes

5:00 ***ADJOURN FOR DAY 2***

Thursday, June 30

9:00 – 9:30 Cloud Computing Initiative NASA OCIO Ms. Karen Petraska

9:30 – 10:15 Finalize Work Plan Dr. Charles Holmes

NASA Advisory Council Ad Hoc Big Data Task Force, June 28-30, 2016

10:15 – 10:25	<i>BREAK</i>
10:25 – 11:25	Findings and Recommendations
11:25 – 11:30	Conclusions/Closeout
11:30	<i>ADJOURN</i>

Dial-In and WebEx Information

For entire meeting June 28-30, 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 997 975 025, and the password is BigD@T@16-2.

** All times are Eastern Standard Time **