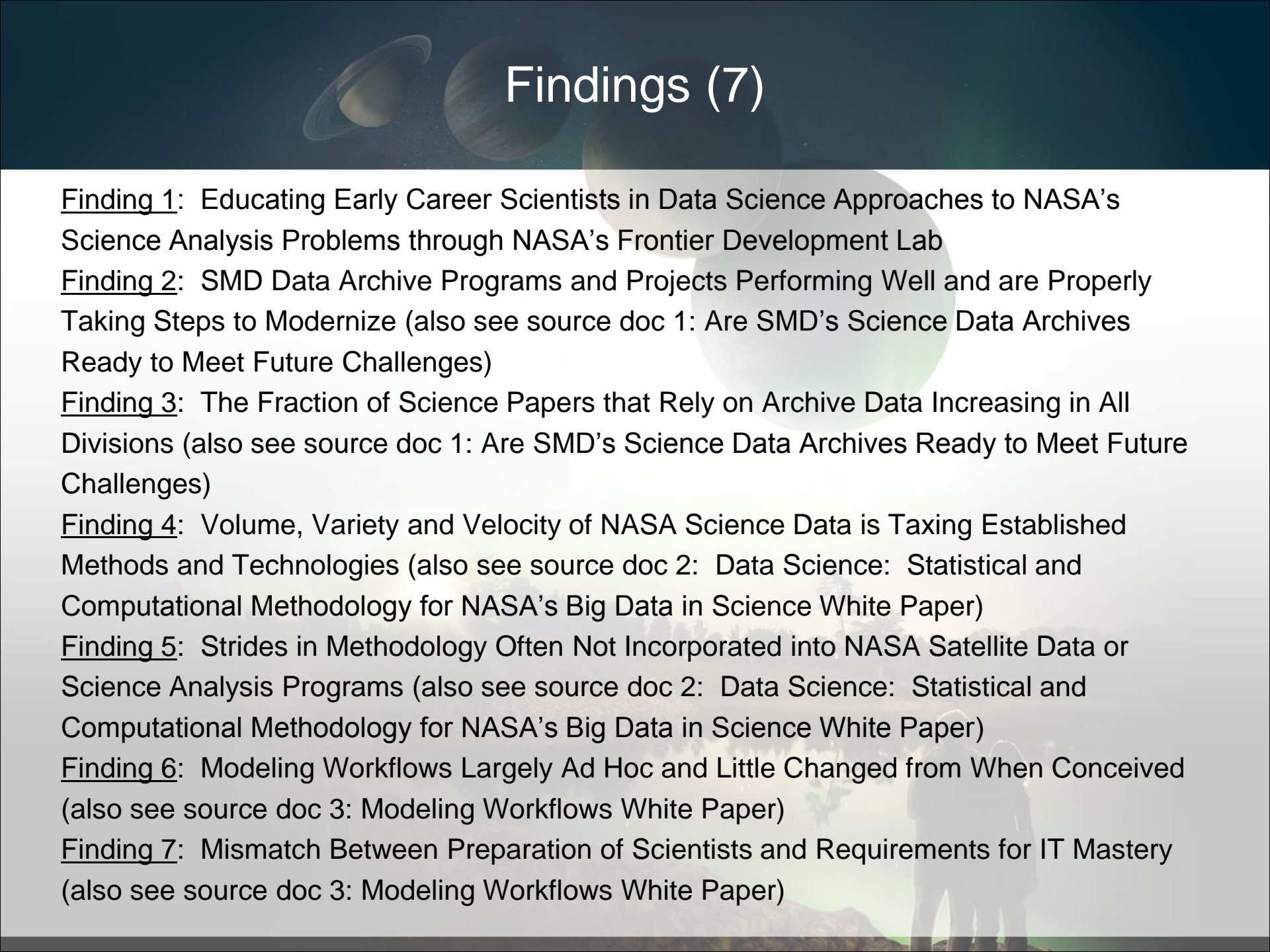




Big Data Task Force (BDTF) Final Findings and Recommendations



January 2017

The background of the slide is a dark blue gradient. In the upper left, there are faint images of planets, including Saturn with its rings and Jupiter. In the lower right, there is a silhouette of a person standing and looking at a large screen or monitor. The title 'Findings (7)' is centered at the top in a large, white, sans-serif font.

Findings (7)

Finding 1: Educating Early Career Scientists in Data Science Approaches to NASA's Science Analysis Problems through NASA's Frontier Development Lab

Finding 2: SMD Data Archive Programs and Projects Performing Well and are Properly Taking Steps to Modernize (also see source doc 1: Are SMD's Science Data Archives Ready to Meet Future Challenges)

Finding 3: The Fraction of Science Papers that Rely on Archive Data Increasing in All Divisions (also see source doc 1: Are SMD's Science Data Archives Ready to Meet Future Challenges)

Finding 4: Volume, Variety and Velocity of NASA Science Data is Taxing Established Methods and Technologies (also see source doc 2: Data Science: Statistical and Computational Methodology for NASA's Big Data in Science White Paper)

Finding 5: Strides in Methodology Often Not Incorporated into NASA Satellite Data or Science Analysis Programs (also see source doc 2: Data Science: Statistical and Computational Methodology for NASA's Big Data in Science White Paper)

Finding 6: Modeling Workflows Largely Ad Hoc and Little Changed from When Conceived (also see source doc 3: Modeling Workflows White Paper)

Finding 7: Mismatch Between Preparation of Scientists and Requirements for IT Mastery (also see source doc 3: Modeling Workflows White Paper)

Finding 1: Educating Early Career Scientists in Data Science Approaches to NASA's Science Analysis Problems through NASA's Frontier Development Lab

In its 2nd year NASA's Frontier Development Lab (<http://www.frontierdevelopmentlab.org>) is proving its value at training early career professionals/students to apply modern data science techniques to sticky analysis problems confronting NASA science and exploration programs. This organization lead by the SETI Institute is sponsored by the NASA's Space Technology Mission Directorate and in part by the Science Mission Directorate. The FDL organizes 8-week summer schools for cadres of early career professionals and graduate school-level students entering into space science fields and computer science. At its Nov 2017 meeting, the BDTF heard of the successes of FDL's 2017 program for solar storm prediction and space weather interactions. The BDTF finds that this type of program aligns with its recommendations to NASA that there needs to be more education and workshops dedicated to introducing modern data science methodologies as approaches for improving the discoveries in its vast science data archives. The BDTF further notes that these types of early career workshops often create collaborations that can last lifetimes as the students enter into their professional careers.



Finding 2: SMD Data Archive Programs and Projects Performing Well and are Properly Taking Steps to Modernize

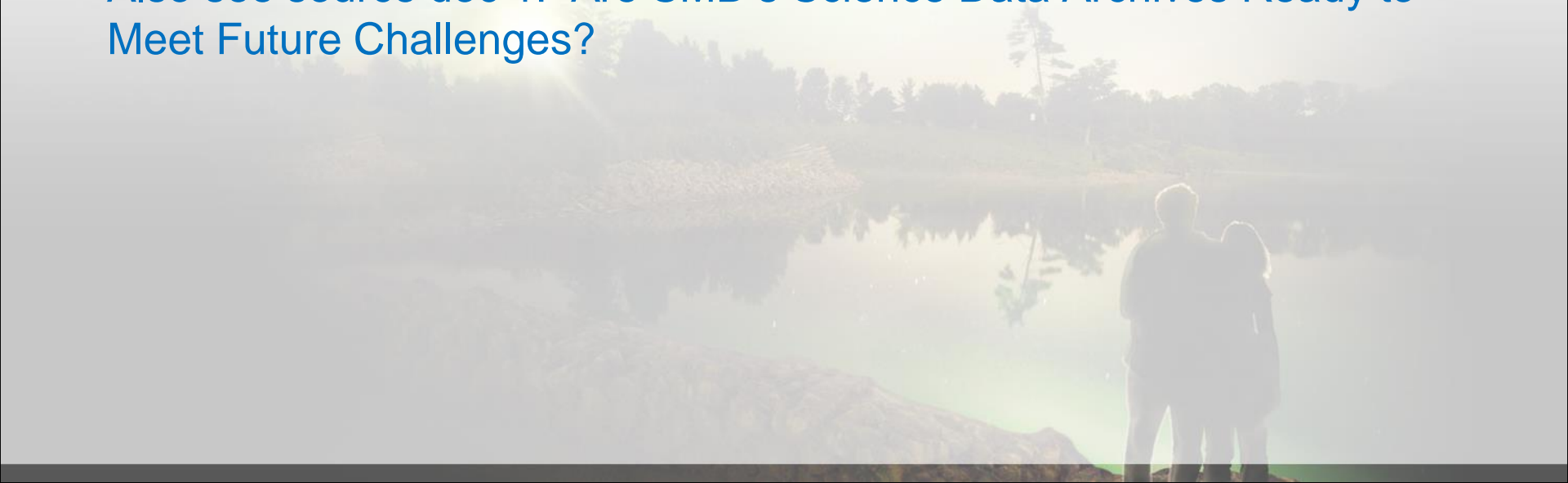
The BDTF finds that SMD's data archive programs and projects are performing quite well and are properly taking steps to modernize and meet future challenges. Like all infrastructure projects they could use increased budgets and will put them to use wisely. Selective implementations of recommendations coming from peer reviews and the communities served are important to consider when augmenting what are generally steady-state funding profiles.

Also see source doc 1: [Are SMD's Science Data Archives Ready to Meet Future Challenges?](#)

Finding 3: The Fraction of Science Papers that Rely on Archive Data Increasing in All Divisions

The BDTF finds that the fraction of science papers, often from outside the original science team, that rely on archive data is increasing in all divisions, and is rivaling the fraction of papers based solely on new mission data and in many cases, exceed 50%.

Also see source doc 1: [Are SMD's Science Data Archives Ready to Meet Future Challenges?](#)



Finding 4: Volume, Variety and Velocity of NASA Science Data is Taxing Established Methods and Technologies

The volume, variety and velocity of NASA science data is taxing established methods and technologies. The problem arises both from data generated by science missions and from computationally intensive simulations supporting these missions.

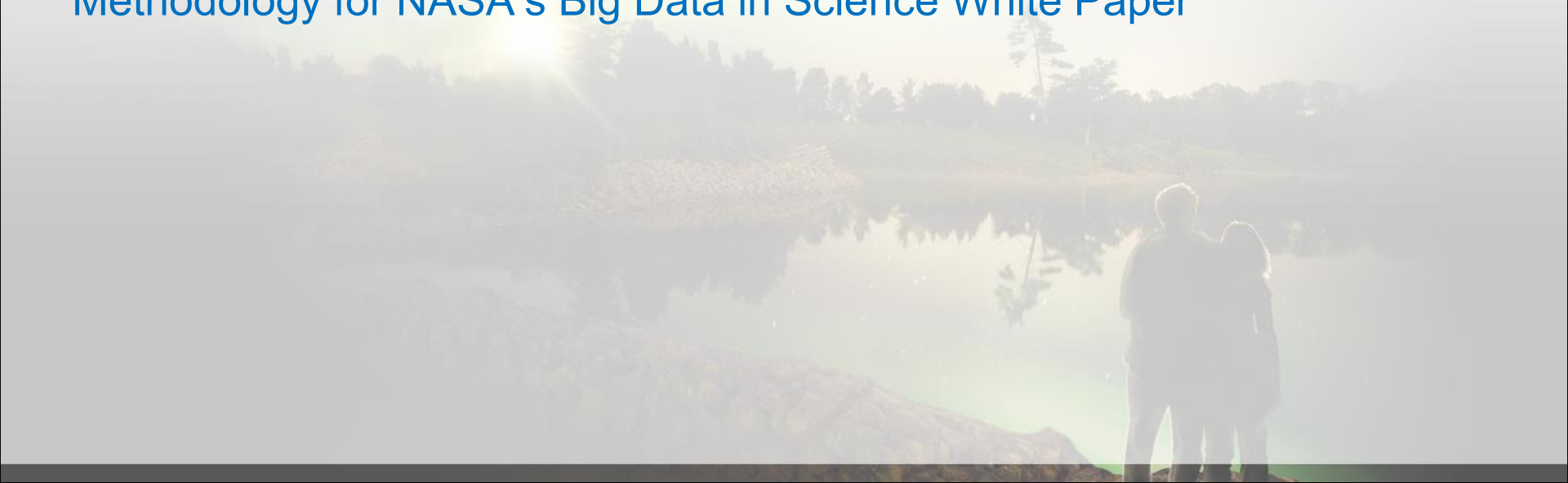
Also see source doc 2: [Data Science: Statistical and Computational Methodology for NASA's Big Data in Science White Paper](#)



Finding 5: Strides in Methodology Often Not Incorporated into NASA Satellite Data or Science Analysis Programs

The enormous strides in methodology from statistics, applied mathematics and computer science of recent decades are often not incorporated into NASA satellite data or science analysis programs. High standards for analysis methodology and algorithms are not set consistently for analysis pipelines within NASA mission centers, for science analysis software maintained by NASA archive centers, or for extramural science programs funded by NASA.

Also see source doc 2: [Data Science: Statistical and Computational Methodology for NASA's Big Data in Science White Paper](#)

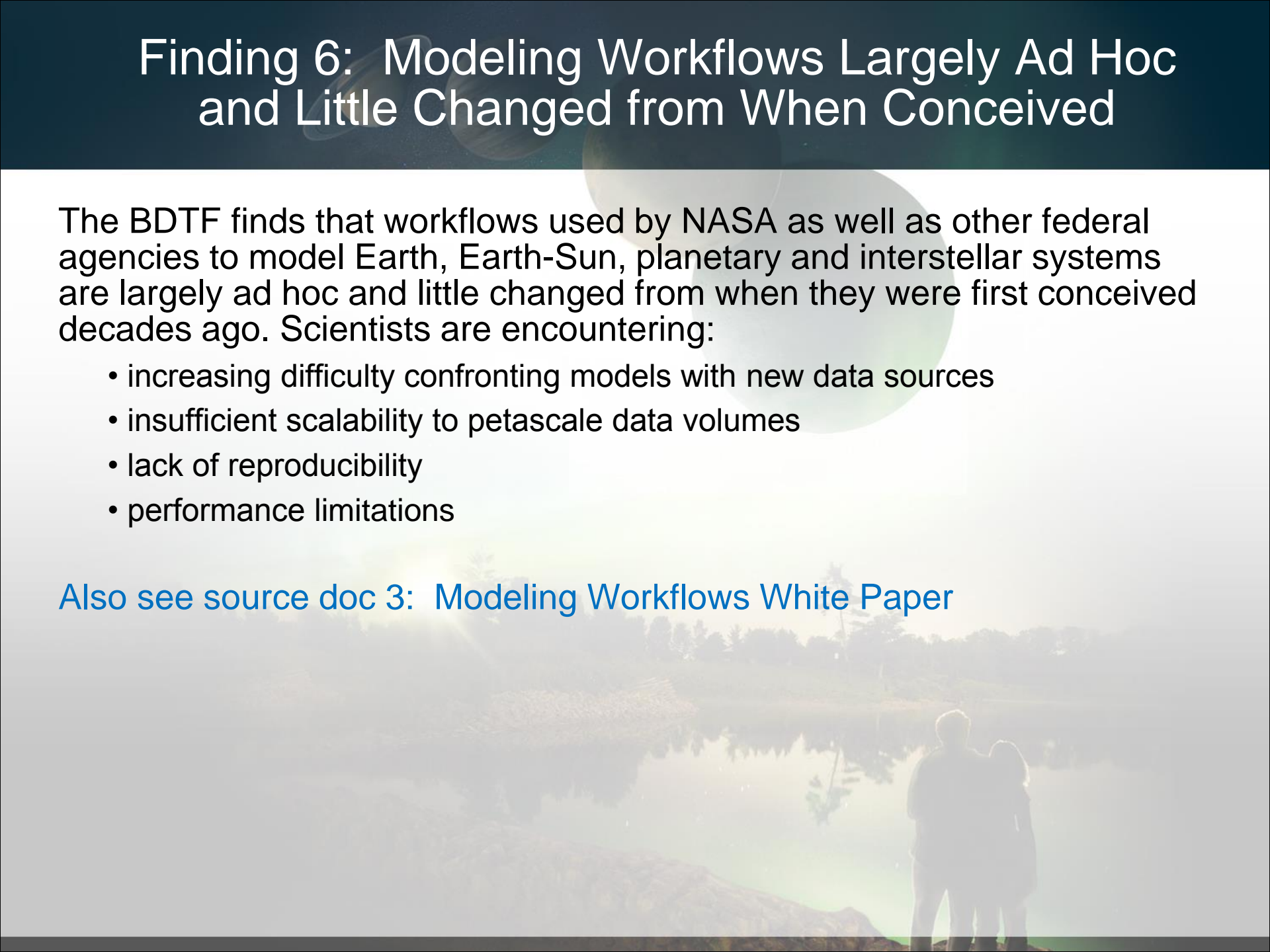


Finding 6: Modeling Workflows Largely Ad Hoc and Little Changed from When Conceived

The BDTF finds that workflows used by NASA as well as other federal agencies to model Earth, Earth-Sun, planetary and interstellar systems are largely ad hoc and little changed from when they were first conceived decades ago. Scientists are encountering:

- increasing difficulty confronting models with new data sources
- insufficient scalability to petascale data volumes
- lack of reproducibility
- performance limitations

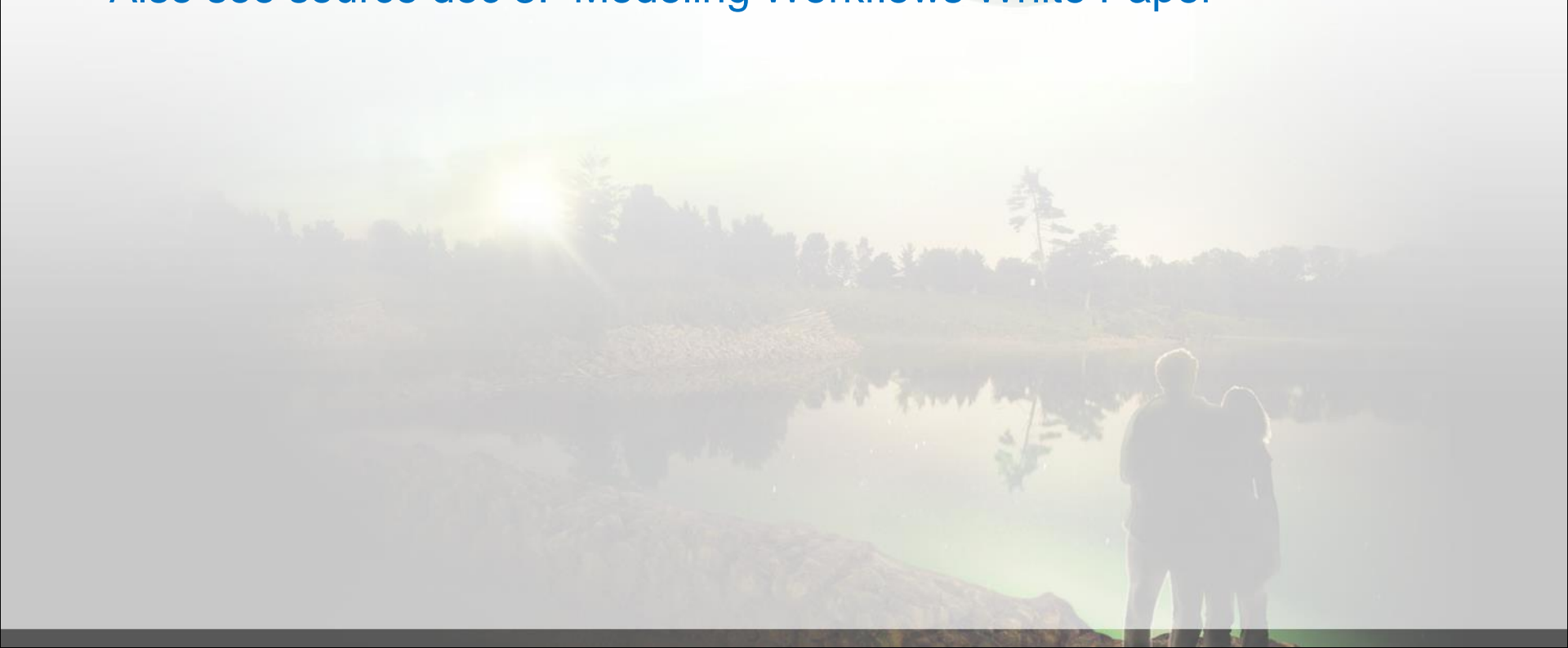
[Also see source doc 3: Modeling Workflows White Paper](#)



Finding 7: Mismatch Between Preparation of Scientists and Requirements for IT Mastery

The BDTF finds that there is a mismatch between the preparation of scientists involved in modeling Earth and space systems and the requirement for information technology mastery in a fast-paced open-source software development environment.

Also see source doc 3: [Modeling Workflows White Paper](#)



Recommendations (11)

Rec 1: SMD Should Manage its Data Archives at Same Rank as the Flight Missions in its Portfolio (also see source doc 1: Are SMD's Science Data Archives Ready to Meet Future Challenges)

Rec 2: Incorporate Data Science and Computing Advisory Positions in the SMD Advisory Committees

Rec 3: Establishing a Data Science and Computing Division in SMD

Rec 4: Necessary Changes in Training, Proposal and Mission Reviews, and Implementation of the Critical Capabilities that Data Science Algorithms Provide (also see source doc 2: Data Science: Statistical and Computational Methodology for NASA's Big Data in Science White Paper)

Rec 5: Making NASA's Archived Science Data More Usable and Accessible (also see source doc 4: Making NASA's Archived Science Data More Usable White Paper)

Rec 6: NASA Should Make Prioritized Investments in Computing and Analysis Hardware, Workflow Software and Education and Training to Accelerate Modeling Workflows (also see source doc 3: Modeling Workflows White Paper)

Rec 7: Implementing Server-side Analytics Architectures (also see source doc 5: Server-Side Analytics White Paper)

Rec 8: NASA Participation in DOE's Exascale Computer Program

Rec 9: Joining the Nation's Science Data Superhighway (also see source doc 6: Joining the Science Data Highway)

Rec 10: Joint Program with NSF's Big Data Innovation Regional Hubs and Spokes

Rec 11: SMD Data Science Applications Program Position and Directed Funding

Rec 1: SMD Should Manage its Data Archives at Same Rank as the Flight Missions in its Portfolio

Given the empirical evidence that community use of these archives leads increasingly to discovering new science and combined with the aggregate budgets, the BDTF recommends that as a policy SMD should view the set of its data archives at the same rank as the flight missions in its portfolio.

Rationale: NASA data archives and centers are playing an increasing role in generating new scientific results as tools come available to discover and extract new meaning from existing data sets. The combined budgets of the data archives are comparable to that of mid-level missions during implementation! Promoting their status relative to new missions will provide to senior management both visibility and the focusing of appropriate resources for optimizing NASA science programs.

Consequences of not adopting recommendation: Valuable knowledge buried within science.

Also see source doc 1: [Are SMD's Science Data Archives Ready to Meet Future Challenges?](#)

Rec 2: Incorporate Data Science and Computing Advisory Positions in the SMD Advisory Committees

In staffing the Science Committee and the four disciplinary Science Advisory Committees, SMD should ensure that at least one appointment on each of these committees is reserved for an expert who is a routine user of high-performance computers (NASA's or others), is active in employing modern data science methodologies, and/or is deeply involved in the science operations of large, complex scientific data archives.¹ It is important but perhaps not necessary that these appointees be or have been active in one of SMD's science endeavors.

Major Reasons for the Recommendation: The new approaches appropriate to extending and upgrading the data analysis and computing activities are emerging and being implemented widely both inside and outside of the NASA science domain. These new approaches often involve new terminologies, technologies, methodologies, workflows, etc. Experts engaged in these approaches should be regularly participating in SMD's FACA committees so as to represent, interpret, translate, and reach out as necessary on these matters.

Rec 2: Incorporate Data Science and Computing Advisory Positions in the SMD Advisory Committees (cont.)

Consequences of No Action on the Recommendation: New ideas for modifying SMD policies or programs relative to its data science and computing activities risk not benefitting from the expertise that could be on hand within the FACA committees to deliberate, review and provide feedback to the advisee organizations.

Background: The subject matter undertaken by the Big Data Task Force has taken on a new urgency throughout modern society in the corporate, educational and government worlds as the tradecraft of Big Data takes hold. It was the TF's goal and charge under its Terms of Reference to bring these ideas forward.

¹ Persons with command of all three topical areas are rare, but it is feasible to find experts in two out of the three areas.

Rec 3: Establishing a Data Science and Computing Division in SMD

SMD should establish a new division whose responsibilities include:

- (1) High-Performance Computing including the projects for the High-End Computing Capability (HECC) at Ames and the NASA Center for Climate Simulation (NCCS) at Goddard as well as activities leading to the future such as participating in DOE's Exascale Computing Program.
- (2) The Data Science Applications Program which will promote bringing modern data science methodologies into SMD's data analysis worlds including the science operations of SMD's missions.
- (3) High-capacity data communications – connecting the NASA science world to the National Data Superhighway.
- (4) Initiate projects to develop server-side analytics services that will lead to “bringing code to the data” while relieving congestion involving serving archive data and caching the data at the broadly-distributed research institutions.
- (5) Initiate projects aimed at modernizing and streamlining SMD's sponsored science models to improve both efficiency of the human resources needed perform calculations as well as on the hardware hosting the models.
- (6) Lead in SMD's actions to implement recommendations resulting from the current study by the National Academy of Sciences “to establish an open-code and open-models policy.”
- (7) Coordinate with the science data archive programs within the four science divisions on matters involving SMD data policies, standards, and interfaces to new services developed under the responsibilities listed above.
- (8) Serve as SMD's point-of-contact for coordination and participation on matters of high performance computing and big data operations with external organizations both within NASA HQ (e.g., the OCIO and the Space Technology Mission Directorate) as well as other Federal departments and agencies including the NSF, DOE, NOAA, DOD, etc.

SMD's science data archiving projects such as ESDIS, HEASARC, PDS, SPDF, etc., must remain under the direction of their current science divisions and intimately connected with the missions providing data

Rec 3: Establishing a Data Science and Computing Division in SMD (cont.)

for archiving as well as the science communities that they are serving. But it is vital that these projects establish good working interfaces to the new activities that will be established under the aegis of this new division. It is inevitable that some new projects initiated by this new division will have impacts at the data archives. These impacts will be both technical (e.g., new software, new connections, etc.) as well as programmatic. So establishing the working interfaces at the beginning will be vital to the success in developing and implementing the new capabilities.

It will be the responsibility of SMD senior management to ensure that the working interfaces are being set up and that programmatic stove pipes involving the Directorate's data and computing programs are not established and sustained. It is important that the new division is tightly integrated with the other four science divisions and, on the other hand, that it does not create a new set of "stovepipes" within the SMD science domain. This new division will use the tools that already are widely adopted within SMD to initiate and carry on its new projects: competitive solicitations via ROSES, community workshops, presence at scientific meetings and conventions, focused community feedback, etc.

Major Reasons for the Recommendation: There are many opportunities and demands waiting for concerted efforts to enhance the discovery of new science from the expansive science data holdings acquired by NASA from its flight missions and associated projects - past, present, and future. Through its recommendations the BDTF has brought forward several approaches, there are most likely more. A focused effort directed by and from a central organization reporting to the AA for Science is the best way to move forward and provide balanced opportunities in all of SMD's science areas. Also, this recommend approach will guard against duplications of effort that can occur in a distributed management regime.

Including the high-performance computing in this new division makes the most sense because of the roles these computers play in the system of analytics that will be enhanced by the activities undertaken from the new division. Having a centralized programmatic focus will enhance its visibility to both SMD and non-SMD communities to ensure a balanced approach to advance forward in this endeavor.

Rec 3: Establishing a Data Science and Computing Division in SMD (cont.)

Consequences of No Action on the Recommendation: There are many “big data” activities occurring under the sponsorship of NASA’s Science Mission Directorate (SMD). The BDTF has witnessed many approaches that are being explored at several of NASA’s centers and in the broader research community. Within the NASA centers, most of these approaches are being paid out of discretionary accounts, on a time-available basis, or by non-NASA funding sources. With a few notable exceptions, there is little organized efforts sponsored directly from SMD to move ahead. By not adopting this recommendation, these efforts will remain fragmented, uncoordinated and very likely unimplemented.

Background: During its 22-month tenure, the members of the Ad Hoc Big Data Task Force (BDTF) have had the privilege of observing and interviewing many “big data” activities occurring under the sponsorship of NASA’s Science Mission Directorate (SMD) as well as occurring in other Federal departments and agencies and in non-governmental organizations. This has permitted the TF to come forward with 14 findings, 12 recommendations, and 4 white papers involving some of the more important topics for moving forward SMD’s data and computing activities. With this in mind, the BDTF finds that SMD lacks the central focus to implement and carry forward the ideas expressed by this advisory activity. Also the TF finds that the program activities for SMD’s high-performance computing activities have been successful and have good forward motion. It is well known that the HPC activities serve many interests sponsored by HQ both within SMD and in other directorates. We conclude that a more central management focus for the HPC programs would be beneficial to keeping its momentum as well as maintaining a balanced approach to serving its diverse scientific and engineering customers.

All of the BDTF recommendations have the goal of enhancing the production of new science results. We believe that enabling this goal is a systems issue involving the data archives, high-performance computing and the glue binding this system: data analysis methodologies, high-capacity communications, more efficient operations of science data search and delivery, and streamlining the work flows in performing theoretical model calculations.

Rec 4: Necessary Changes in Training, Proposal and Mission Reviews, and Implementation of the Critical Capabilities that Data Science Algorithms Provide

The BDTF recommends that NASA SMD make the necessary changes in training, proposal and mission reviews, and implementation of the critical capabilities that data science algorithms provide.

- NASA SMD should organize and fund professional development in statistics and informatics, both for its internal scientists and for the wider Earth and space science communities. This includes organizing training workshops, producing on-line training materials, attending methodology conferences, and hiring expert consultants.
- Specifications and performance reviews for mission science operations software development should include high standards for computational algorithms and statistical methods with evaluation by cross-disciplinary experts.
- NASA SMD should ensure modern software engineering (for example: agile, fast iteration processes for creation and modification of software systems) is applied to its sponsored development and maintenance projects. This may include active involvement of data science professionals as staff or consultants.

Rec 4: Necessary Changes in Training, Proposal and Mission Reviews, and Implementation of the Critical Capabilities that Data Science Algorithms Provide (cont.)

Rationale: Data science and modern software engineering methods provide powerful insights and allow for fully leveraging the large, real-time, and complex datasets coming from NASA SMD missions and its research programs. These methods often come from adjacent fields from the normal SMD focus areas, and therefore require cross-disciplinary training, collaborations, and other technology transfer.

Consequences for not adopting the recommendation: NASA science missions will not adequately leverage data to extract the full value from the datasets made available. Inadequate methodologies will result in weak recovery of the science potential, lower quality results, and higher costs for analysis. Furthermore, there is potential for inefficient software maintenance practices.

Also see source doc 2: [Data Science: Statistical and Computational Methodology for NASA's Big Data in Science White Paper](#)

Rec 5: Making NASA's Archived Science Data More Usable and Accessible

The BDTF recommends that near the end of the prime mission NASA conduct a review of data entering the archives including the quality of the calibration and the metadata describing the mission. Spacecraft and instrument status may have changed during the mission and should be updated. In addition, we recommend that at this time the missions prepare or update user's guides for the data products from each instrument detailing their use. These are important steps in making the data truly useable and essentially extending the effective life of the mission.

Rationale for the recommendation: As missions age instrument states change and poorly calibrated data can get into the archives. Including calibration reviews in a major review such as that at the end of the prime mission will improve this by catching errors and updating documentation and calibration tables. Space instruments have become so complex that using the data can be very challenging especially for those with limited resources from small grants. User's guides have proven to be a straightforward and effective way to make the data more useful and essentially extend the missions beyond their active lifetimes.

Rec 5: Making NASA's Archived Science Data More Usable and Accessible (cont.)

Consequences of no action: Without the calibration review poorly calibrated data will continue to be mixed into the archives. Without the user's guides more scientist time and effort is needed to learn the complex instruments and use the data. Some studies for which a given data product is appropriate will not be feasible given limited resources.

Also see source doc 4: [Making NASA's Archived Science Data More Usable White Paper](#)



Rec 6: NASA Should Make Prioritized Investments in Computing and Analysis Hardware, Workflow Software and Education and Training to Accelerate Modeling Workflows

The BDTF recommends that NASA should make prioritized investments in computing and analysis hardware, workflow software and education and training to substantially accelerate modeling workflows. NASA should take the lead to make substantial increases in:

- accessible and affordable computing and data storage capacities
- software modernization
- resources to develop new data analysis paradigms
- education and training workshops, scientific conferences and journal special collections to effect a culture acceptance of the importance of workflow development and management.

Immediate efforts that can contribute to accelerating modeling workflows include:

- adopting a systems approach to designing workflows
- modularization
- identifying and implementing concurrency in workflows and algorithms
- automation of repetitive steps in modeling workflows.

Longer-range investments whose potential NASA should investigate include:

- virtualized environments
- research on memory and processing scalability
- lossy data compression and more advanced methods for signal detection
- data-centric science gateways
- platforms for sharing workflows
- automation of the creation of workflows.

Rec 6: NASA Should Make Prioritized Investments in Computing and Analysis Hardware, Workflow Software and Education and Training to Accelerate Modeling Workflows

Rationale: Workflows designed and implemented several decades ago are no longer capable of keeping pace with the growing complexity of the model development cycle, the exponentially growing volume of input, output and validation data, and the rapidly advancing computing environment for both high-performance computing and large-memory data analysis.

Consequences of not adopting the recommendation: The outdated modeling workflows employed by NASA will waste valuable high-performance computing, data storage and human resources and will increasingly hamper progress.

Also see source doc 3: [Modeling Workflows White Paper](#)



Rec 7: Implementing Server-side Analytics Architectures

NASA's archives of science data are growing larger and incorporating more features. In many cases efforts needed by users of these data are being hampered by the inefficiencies associated with moving large data sets from the archives across the internet into the storage local to the user. A new paradigm is beginning to appear which will permit operations on the data prior to transmission. These operations reduce the amount of data that will be transmitted. There are several terms being used to describe the new paradigm: data analytics, climate analytics, serverside analytics, etc., which all have common features – namely bring data analysis code to the data. The BDTF has studied the case for adopting SSA implementations where needed and provides the following recommendations:

- Publish an RFI via NSPIRES to ascertain where there are centers of SSA development that could be useful for NASA-sponsored research.
- Convene workshops for the NASA science community to show off the SSA services that are under development and get feedback that could point to where new developments are needed,
- Solicit via ROSES for demonstrations or prototypes of SSAs that have the most promise of solving “log jam” situations in priority areas.
- For those projects that demonstrate potential at significantly improving the delivery of tailored processed data to the researcher, SMD should take steps to make those projects operational.

Rec 7: Implementing Server-side Analytics Architectures (cont.)

In all events members of organizations who provide SMD data archive services need to participate. It is most important that representatives, ideas, and approaches from the broader data user communities need to participate and in certain cases lead in the implementations. Sound science leadership from SMD will be needed to ensure that new SSA architectures are implemented where most needed and other proposals for lower priority implementations are reserved. The use of community peer groups and advisory working groups will be valuable in providing priority rankings for new concepts and proposals.

Rationale for the recommendation: The rationale for moving to server-side analytics architectures include:

- Reading data over the internet is slower than reading data from a local file.
- Some analysis tasks require a lot of data: “Compute the mean annual cycle from a 100-year climate model run.”
- Analysis expression is evaluated at the server, where the data reside on disk.
- Only the analytics result is delivered to the client.
- SSAs can save a lot of time if: (size of result) \ll (size of data operated on).

Consequences for not adopting the recommendation: We are already seeing examples where data analysis communities are restricted by the bandwidth of the transmission from the data archives. This problem will get worse due to ever increasing data volumes. If we do not take action now, we will not realize the full potential for producing new science from these data sets and possibly not meeting L1 requirements.

Also see source doc 5: [Server-side Analytics White Paper](#)

Rec 8: NASA Participation in DOE's Exascale Computer Program

The Exascale Computing Project (ECP) was established as part of the National Strategic Computing initiative. The goal of the project is to accelerate delivery over the next decade of an exascale computing system that integrates hardware and software to achieve 50 times more performance than the 20-petaflops machines available today. The Department of Energy is the lead Federal agency along with the Department of Defense and the National Science Foundation. The ECP is using a co-design approach that includes hardware technology, scalable software and science and mission application development. NASA science can greatly benefit from this approach but to do so NASA needs to be closely involved in the ECP development. The NASA HQ Program Officer for High Performance Computing is participating in top-level coordination and planning meetings. We recommend that NASA continues participating in this important project and becomes more fully involved by providing modeling and expertise to support the ECP's Modeling Development activities.

Major Reasons for the Recommendation:

Exascale computing has the capability to greatly enhance the science return from NASA missions. For instance, over the past two decades the steady improvement of high performance computing has greatly increased our ability to more realistically model and simulate the natural environment. Models and simulations are now regularly used as tools in analyzing observations from NASA missions. Exascale computing offers the exciting possibility of increasing our capability by more than an order of magnitude. The next generation of simulation and modeling codes will need this enhanced capability. For instance, in space plasma physics three-dimensional particle-in-cell codes are now being used but for ideal systems and frequently with unrealistic ion to electron mass ratios.

Rec 8: NASA Participation in DOE's Exascale Computer Program (cont.)

With the increase in capability that exascale computing will bring, we will be able to model more realistic systems based on observations with realistic mass ratios. In the Earth sciences, models of the global atmosphere, initially, and more recently models of the entire Earth system, have been employed in data assimilation mode to develop four-dimensional regular gridded state estimates over time and to project the future state of the Earth system. Expectations for the application of exascale computing to the problem of Earth system simulation and projection are that critically important features that are parameterized in current generation models – such as clouds, ocean eddies, landscape variations, and sea ice polynyas – will be explicitly represented. Crossing that threshold has the potential to dramatically increase Earth system model accuracy and reliability. The holistic approach of the ECP with applications, software and hardware being considered together as part of the development process is exciting. NASA science can greatly benefit from this approach but to do so NASA needs to be closely involved in the ECP development. NASA science is now listed as a potential user of exascale computing but by becoming a participating partner NASA can make sure that the resulting system will address NASA's needs.

Consequences of No Action on the Recommendation:

Large scale computing has greatly increased the scientific yield from NASA missions. The ECP project is the only Federal development project that promises to bring this new capability online. Exascale computing is necessary for the next generation of NASA science models and simulations. Failure to participate in this initiative will slow the progress of NASA science.

Rec 9: Joining the Nation's Science Data Super Highway

SMD should establish a temporary 2- to 3-year staff position to focus on providing access to the Nation's Science Data Super Highway provided by DOE's ESNet, the PRP, and allied services such as the Internet2. This program officer would investigate in detail the requirements and opportunities for extending the Nation's Science Data Super Highway to many of NASA's sponsored research groups. This officer would sponsor a program of ~\$3-5M/year which would be divided among two elements:

- (1) a ROSES solicitation so that research groups can propose to acquire necessary hardware and software in order to set up their own Science DMZ that would connect into the larger campus and national science data networks and
- (2) provide requirements and funds to the Agency's CSO (Communication Services Organization) to acquire necessary links - where vitally needed - to bridge to junction points in the high-speed national networks.

Initially the program officer might be recruited as an IPA from one of the several organizations participating in the ESNet or Pacific Research Platform. After two years, SMD should evaluate the need to continue this position and funding needs.

Major Reasons for the Recommendation:

Over the past six years the Federal Government has been investing heavily in upgrading transmission capabilities for science data into many science research groups throughout the

Rec 9: Joining the Nation's Science Data Super Highway (cont.)

nation – connecting science research groups to science archives, collaborating groups and data to/from high-performance computing centers. Many NASA-supported research groups are being left out of this new way of doing business. In order to spread Science DMZs into the NASA-sponsored research groups, there needs to be a cognizant program officer at NASA HQ who is on top of the details, coordinating at the agency/department levels and overseeing a grants program that permit the research groups to get funding to implement their Science DMZ.

Thirty years ago at the dawn of the Internet age, NASA Science led the way in extending Internet connectivity to the Nation's science research centers. New infrastructure upgrades for science data transmission are taking place with NASA Science largely standing by. It's not too late for NASA Science to participate and provide opportunities for many of its sponsored research groups to join the Data Super Highway.

Consequences of No Action on the Recommendation:

NASA-sponsored research groups will be left behind in this new era of 100 Gbps transmission. Yet it is NASA science data that are at core of the attacks on some of the nation's most challenging science problems.

[Also see source doc 6: Joining the Science Data Super Highway](#)

Rec 10: Joint Program with NSF's Big Data Innovation Regional Hubs and Spokes

NASA/SMD should:

- (1) alert its research community via the listserv messaging capability in NSPIRES that this NSF program is up and operating and that NASA research PIs may benefit from making contact with one or more of the participants in the Hubs and Spokes infrastructure and discussing data analysis problems that they are facing, and,
- (2) consider establishing a joint research solicitation with the NSF that will bring together NASA research PIs with the appropriate elements of the Regional Hubs and Spokes to attack some of the daunting problems in analyzing large and complex data sets arising from its vast fleet of operating science missions.

On accepting this recommendation, SMD should appoint a program scientist to contact the NSF Project Officer to begin developing the NSPIRES listserv message as well as entering into preliminary discussions for a joint-solicitation.

Major Reasons for the Recommendation:

This NSF Big Data Regional Innovation Hubs and Spokes program has established a network of expertise available to take on problems in data analyses that may be facing many PIs of NASA-sponsored research activities.

This should be a win-win for both communities – NASA PIs finding and applying new and emerging methodologies to their data analyses; NSF-sponsored Big Data PIs applying their craft to Earth and space science problems.

Rec 10: Joint Program with NSF's Big Data Innovation Regional Hubs and Spokes (cont.)

Consequences of No Action on the Recommendation:

By not making the connections, either informal or formal, between NASA-sponsored research PIs and the participants in the NSF's Big Data Hubs and Spokes project, there will be missed opportunities of having the nation's experts in modern data science techniques applying their craft to finding new science results in NASA's large and complex science data holdings.

Background: In 2015, the National Science Foundation initiated a series of solicitations designed to promote and scale up collaborative big data innovation activities¹. These activities are now in place applying modern methods for analyzing large and complex data sets to problems across many diverse disciplines including science, medicine and public health, engineering, city operations, etc. The awards from this program are promoting alliances of experts in computer science, applied mathematics and statistics with experts in the domains of the problems under consideration. The program formally known as Big Data Regional Innovation Hubs and Spokes established a nationwide network focused universities and involving researchers from across academia, government, and industry. The BDTF has observed the establishment of this network, interviewed participants at many levels and attended data science workshops organized by the Regional Hubs. The BDTF finds that this network is a new and important resource that could be employed to enhance NASA's sponsored science research – both at the mission science levels as well as for the Research and Analysis (R&A) investigators.

Rec 11: SMD Data Science Applications Program

SMD should establish a permanent position: “Data Program Scientist.” This person would have several responsibilities needed to bring data science approaches to fruition within NASA’s science domain:

1. Direct an annual research program of about \$10M entitled “Data Science Applications” and solicited via ROSES. This would include monitoring and reporting progress and adaptations resulting from the awarded work.
2. Convene workshops for NASA-sponsored research PIs and associates including students on applying data science methodologies to data analysis problems associated with the data obtained from NASA’s science missions.
3. Participate in NASA’s mission development processes:
 - i. Review all draft AOs to ensure that the data science perspective is properly represented in the solicitation,
 - ii. Assist SMD’s mission program scientists in nominating prominent data scientists to participate in science review panels for proposed missions and/or instruments, and
 - iii. During a mission’s implementation phases work with HQ program scientists to ensure that appropriate data science methodologies are being properly incorporated and reviewed for implementation in the mission’s science operations segments.
4. Interact with other NASA HQ directorates and offices concerning their implementations of data science methodologies that may be of common interest to SMD.

The “Data Science Applications” solicitation would be open to all SMD science themes. SMD might consider that initially the ROSES “Data Science Research” solicitation would include a joint solicitation with NSF’s Big Data Regional Innovative Hubs and Spokes activities. (See BDTF recommendation: “Participation in NSF’s Big Data Regional Hubs”.)

Rec 11: SMD Data Science Applications Program (cont.)

The Data Program Scientist is a new position working with the Directorate's program scientists and the broader NASA-sponsored research community. SMD may want to consider initially filling this position with a data scientist who is actively engaged in applying the methodologies. For example, a candidate from one of the NSF-sponsored Big Data Regional Hubs might be able to come aboard want as an IPA.

This position is independent of SMD's standing data archive programs. Since search, access and retrieval of data are important considerations when applying data science methodologies, the Data Program Scientist will need to work closely with the SMD's data archive program officers to represent requirements for preparing and delivering data to research groups applying the data science methodologies.

Major Reasons for the Recommendation: Modern data science techniques for analyzing large and complex data sets by and large are not being applied to the analyses of NASA's extensive data sets. The BDTF is recommending a top-down approach for getting NASA research PIs and their teams to start applying modern data science methods in their science analyses. This push would be applicable to broad research community funded via the ROSES solicitation process as well to NASA's science missions.

Consequences of No Action on the Recommendation: In many cases the analyses of NASA science data will not reach their full potential without incorporating the modern data science approaches. NASA-sponsored research groups will continue to fall behind the state-of-the-art of

Rec 11: SMD Data Science Applications Program (cont.)

applying modern data-analytical methods. This risks that valuable new science results will not be gleaned from NASA's extensive data holdings.

Background: Large, complex data sets are nothing new to NASA's science research program. The aggregate of NASA's science data holdings is growing at rates following Moore's Law. This is nothing new. What is new is that more and more "new" science is being discovered from accessing these data sets by researchers not connected with the original science teams. But the growths in size, complexity, and utilization of data archives are not singular to NASA. Many other endeavors such as medicine, engineering, city and industrial operations, forensic science, to name a few, are facing similar growths. There is a big difference: these other endeavors are vigorously applying with great success new data analysis methodologies to dig out important results from the data.

An emerging area, Data Science, is being organized to attack the analyses of large and complex data sets. The area is an amalgam of computer science, applied mathematics and statistics working together with domain specialists to attack the problems at hand. Methodologies including techniques such as deep learning, convolutional neural nets, Bayesian statistics, etc., are being applied. Though still in its formative years, Data Science may soon become its own scholarly discipline.

The BDTF finds that the SMD research programs by and large have not embraced adopting these new methodologies. One reason for this is lack of emphasis and sponsored opportunity from NASA HQ.

The BDTF also finds that SMD needs a program officer, with experience in the data science methodologies, to assist mission program scientists in bringing new science missions online to ensure that data science methodologies are being considered and incorporated into the back end of the science discovery process.