6th & Final Report of the
Big Data Task Force

Charles P. Holmes
Chair, BDTF
November 28, 2017

BDTF Membership

Name	Dept./Center	Organization
Charles Holmes - Chair	Retired	Formerly NASA HQ
Reta Beebe	Dept of Astronomy	NMSU
Eric Feigelson	Center for Astrostatistics	Penn State U.
Neal Hurlburt	Solar and Astrophysics Lab.	Lockheed Martin
James Kinter	Center for Ocean-Land-Atmosphere Studies	GMU
Chris Mentzel	Program Director Data-Driven Discovery	Gordon & Betty Moore Foundation
Clayton Tino	Software Architect	Virtustream / EMC
Raymond Walker	Institute for Geophysics and Planetary Physics	UCLA
Jerry Smith – Exec. Sec.	SMD	HQ NASA

BDTF at JPL, Nov 3, 2017

What is our take away message?

1. There are a few areas that SMD is behind:
   - Participating in the science data super highway
   - Leadership in data science applications and education
2. SMD needs to view its science data archives as the sources of half its new science.
3. There are other areas were SMD needs to start applying leadership in a measured, systematic way:
   - Modeling workflows
   - Server-side analytics

Today's Report

- Four white papers with accompanying findings (4) and recommendations (4).
- Are the Science Data Archives Ready to Meet Future Challenges?
- Finding on courses offered by the Frontier Development Labs.
- Recommendation on organizing SMD's Data Science and high-performance computing programs.
- Recommendation on DS&C expertise on SMD Advisory Committees

And don’t forget

Four recommendations submitted in July:

1. NASA Participation in DOE’s Exascale Computer Program
2. Joining the National Data Superhighway
3. Joint Program with NSF’s Big Data Innovation Regional Hubs and Spokes
4. SMD Data Science Applications Program

Today's Report

- Four white papers with accompanying findings (4) and recommendations (4).
- Are the Science Data Archives Ready to Meet Future Challenges?
- Finding on courses offered by the Frontier Development Labs.
- Recommendation on organizing SMD's Data Science and high-performance computing programs.
- Recommendation on DS&C expertise on SMD Advisory Committees.

Scope of the Big Data Task Force

The scope of the Task Force includes all NASA Big Data programs, projects, missions, and activities. The Task Force will focus on such topics as exploring the existing and planned evolution of NASA's science data cyber-infrastructure that supports broad access to data repositories for NASA Science Mission Directorate missions; best practices within NASA, other Federal agencies, private industry and research institutions; and Federal initiatives related to big data and data access.

Abstracted from the Terms of Reference, Ad Hoc Task Force on Big Data, signed by the Administrator on Jan. 8, 2015.

BDTF Work Plan

(Fall 2015)

1. Present and approve the work plan
2. Survey and nominate topics for study.
3. Choose 3 to 4 topics.
4. Define problem and approach to each study.
5. Produce products:
   - Research,
   - Organize and develop positions,
   - Draft white papers and
   - Finish present results white paper with accompanying slide presentations.

BDTF Work Plan (Fall 2015)

1. Present and approve the work plan. ← *HQ, Feb 2016*
2. Survey and nominate topics for study. ← *GSFC, June 2016*
3. Choose 3 to **4** topics. ← *GSFC, June 2016*
4. Define problem and approach to each study. ← *ARC, Oct 2016*
5. Produce products:
   - Research, ← *HQ, March 2017*
   - Organize and develop positions, ← *HQ, March 2017*
   - Draft white papers and ← *Telemeeting, June 2017*
   - Finish present results white paper with accompanying slide presentations. ← ***JPL, Nov 2017***

BDTF Focus Topics

The TF reviewed the list of (~30) candidate topics for study. We selected these topics:

1. *Data accessibility*
2. *Data science methodologies*
3. *Modeling workflows*
4. *Server-side analytics*

BDTF White Papers - General Outline

- Executive Summary
- Background
- Statement of the Problem
- Example(s)
- Approach(es) for Solutions
- Findings and Recommendations

DATA ACCESSIBILITY – MAKING NASA SCIENCE DATA MORE USEABLE

Data Accessibility

NASA SMD is mandated to assure the highest quality of data possible while dealing with very large and increasing data volumes containing ever more complex data.

In addition, they are faced with trying to meet ever-increasing demands from the science data user community.

The four science divisions each have specific challenges that result from the types of science they support, the types and complexity of the data, the difficulty of acquiring observations and the amount of data and the size and diversity of their user communities.

In this report, the BDTF reviewed the current state of the data activities in each of the disciplines and the challenges each faces and made recommendations for improvement.

Data Accessibility

The Big Data Task Force review of the NASA science archives noticed that while the archives were in general proactive in encouraging missions to submit data, the quality of the metadata describing the data and the calibration in some cases were inadequate for successful analysis of the data.

Other observations include:

- Science data volumes from ever more sophisticated instruments are growing.
- While users can find the data products using the online systems at the archives many are very difficult to use.
- Even domain experts often find the data difficult to use. A major part of the problem with using the data results from the complexity of the instruments which leads to very complex data products.
- The key to having archival data products that the science community can readily use is well calibrated data and metadata that clearly describe the data products in the archive.
- Recently the Cassini mission to Saturn has augmented the metadata by including detailed user's guides.
- These text documents have been very successful in aiding users as they work with the very complex data from Saturn.

Data Accessibility

Recommendation:

- 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.

Data Accessibility

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.

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.

DATA SCIENCE: STATISTICAL AND COMPUTATIONAL METHODOLOGIES FOR NASA'S BIG DATA IN SCIENCE

Data Science: Methodologies

Background:

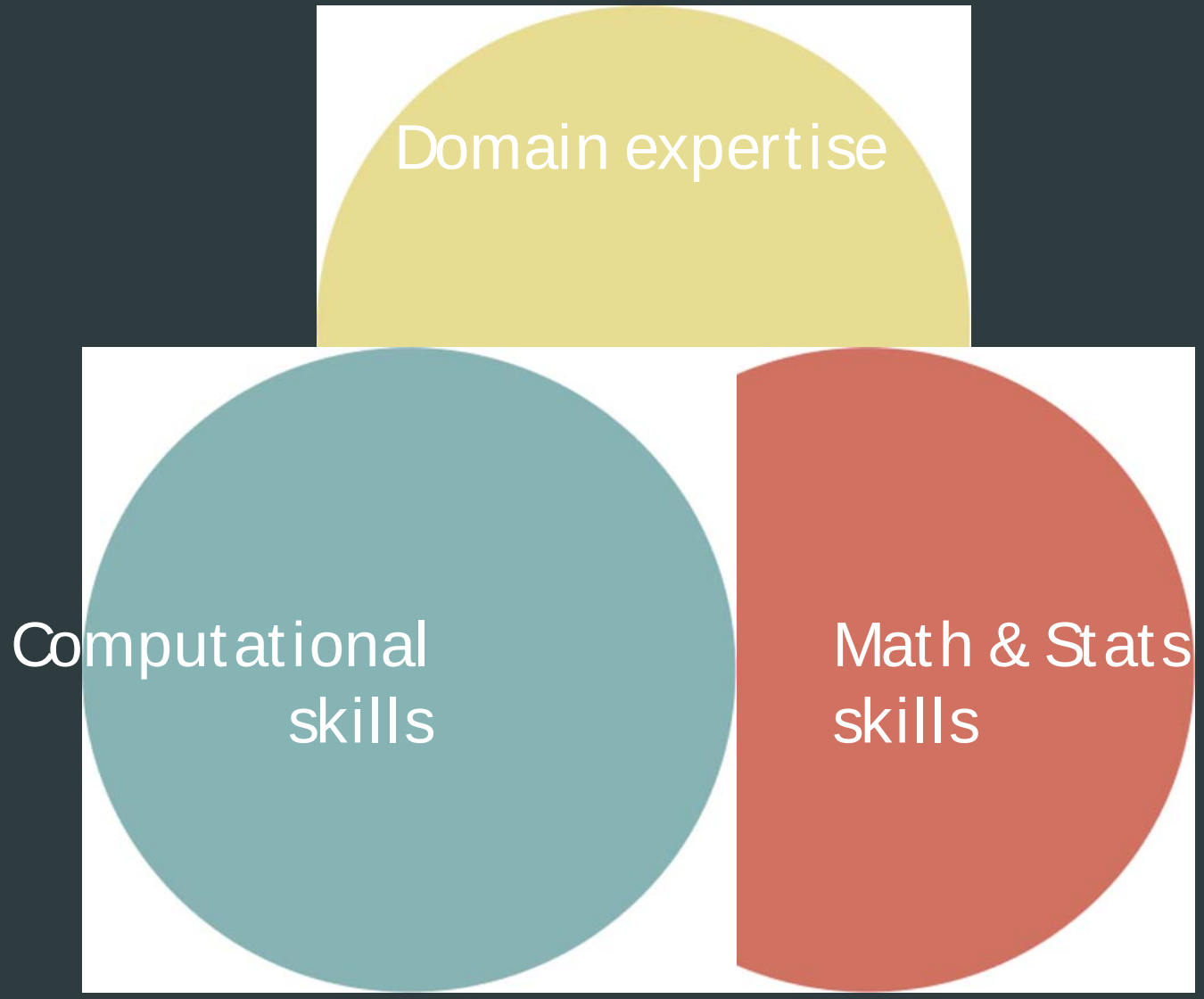
- Big Data and High Performance Computing deeply permeates NASA's Earth and space science endeavor.
- However, NASA science is not institutionally well-integrated with the large community of scholars and experts who develop the methodologies underlying modern Data Science: computer scientists, information technologists, applied mathematicians and statisticians.
- Technology transfer from and collaboration with these scholarly communities is insufficient to meet NASA's needs, both within NASA centers and in the wider U.S. space science research community.

Data Science: Methodologies

The emerging term for such cross-disciplinary activities -- *Data Science* -- can be represented as the intersection of three skill areas: domain expertise (like space science), mathematics and statistics, and computational skills. Data science is a rapidly emerging area of study and practice that has witnessed amazing acceleration in recent years with foundations in statistics and computer science over several decades. However, NASA science is not a leader in this movement. Substantial benefits can accrue with new analysis procedures at a fraction of the cost of the total satellite missions, and yet these techniques are not universally employed across every mission. Promising areas of methodological improvements for space science include:

- Statistical procedures including: False Discovery Rate for multiple hypothesis testing; Uncertainty Quantification for computationally expensive modeling; model selection and residual analysis following regression fits
- Machine learning, data mining and computational intelligence including: Deep Learning neural nets, pattern recognition, decision trees and Random Forests
- Data cleaning and curation including: data aggregation, normalization, synchronization, assimilation, and improved meta-data descriptions of datasets
- Data visualization to improve understanding including: visual exploration tools, sonification, haptic data sensing techniques, and virtual reality.

How to succeed in Data-Driven Discovery – the right skill mix



Credit: Chris Mentzel, Gordan and Betty Moore Foundation

Data Science: Methodologies

Finding: 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.

Finding: 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.

Data Science: Methodologies

Recommendations: 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.

Data Science: Methodologies

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.

MODELING WORKFLOWS

Modeling Workflows

Modeling is a major aspect of the Earth and space science research conducted by NASA and other Federal agencies. The development of numerical models of the Earth system, planetary systems or astrophysical systems is essential to the linkage of theory with observations. Furthermore, the optimal use of observations that are often quite expensive to obtain and maintain typically requires assimilation or other forms of objective analysis that involves numerical models.

"Modeling workflows" refers to the set of activities undertaken by a research team to perform a set of calculations from a large computer code incorporating a theoretical physics model(s). The activities include data preparation, integrating the model(s), tuning parameters and algorithms, iterating intermediate results, and reduction of results for examination, analysis and publication.

Modeling Workflows

Despite its importance, the manner in which NASA and other federal agencies model Earth, Earth-Sun, planetary and interstellar systems, generally described herein as “modeling workflow”, is not keeping pace with the rapid advancement of information technology and high-performance computing, methodological innovations, or the rapid growth of data volumes and complexity.

Modeling workflows are largely ad hoc and little changed from when they were first conceived decades ago.

The modeling workflows in place today have reached the point where time between generations of Earth and space system models is measured in years.

Scientists are encountering: increasing difficulty confronting models with new data sources; insufficient scalability to petascale computing and data volumes; lack of reproducibility; and performance limitations due to lack of data compression or brittleness of aging data formats.

Modeling Workflows

Finding: 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

Finding: 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

Modeling Workflows

Recommendation: 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.

Modeling Workflows

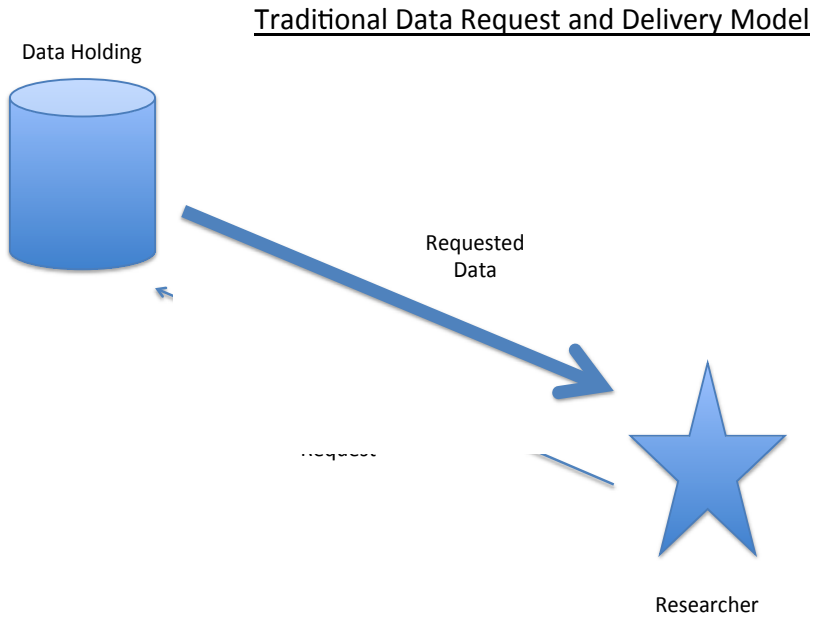
Rationale: Workflows designed and implemented several decades ago are no longer capable of keeping pace with the growing complexity of the model development cycle, the exponentially growing volume of input, output and validation data, and the rapidly advancing computing environment for both high-performance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

SERVER-SIDE ANALYTICS

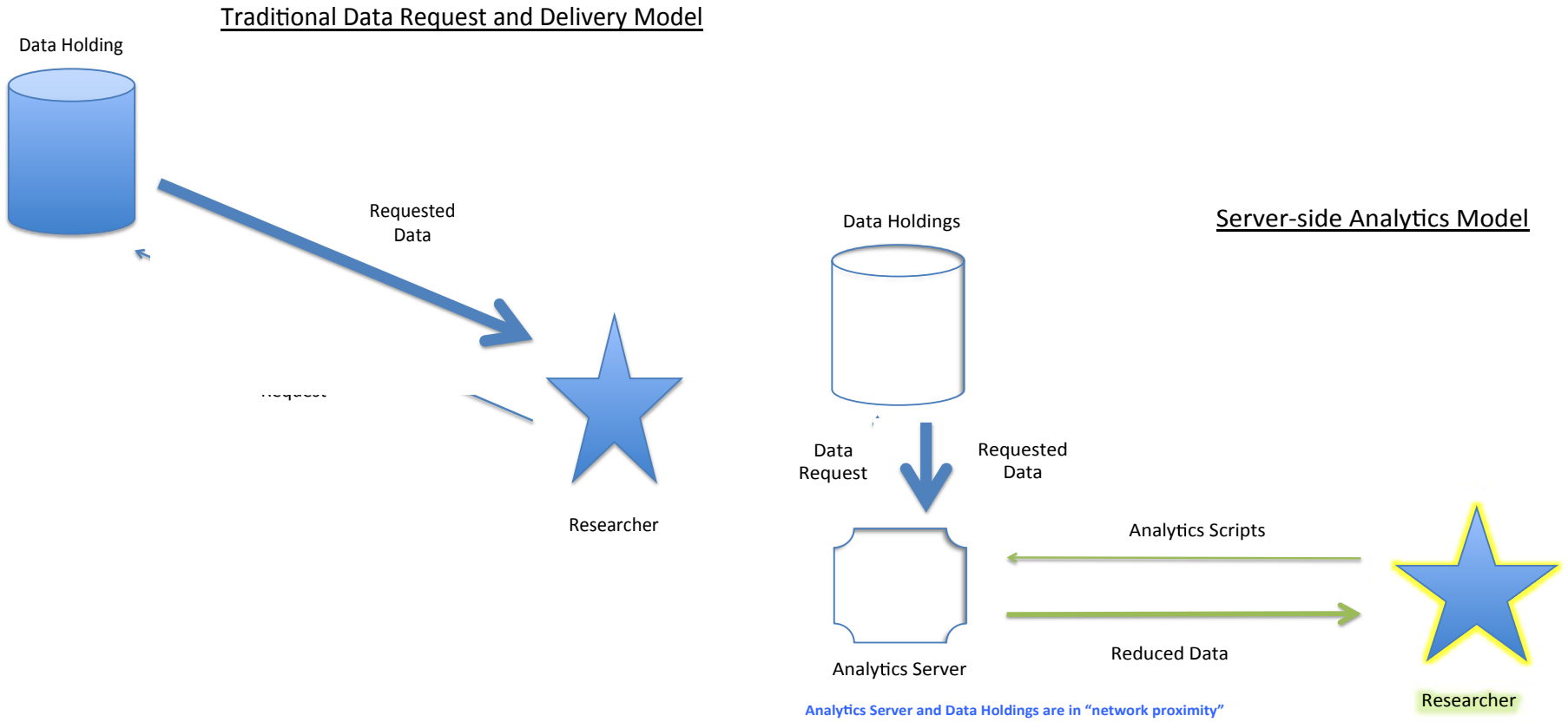
Sever-side Analytics

“Bringing Code to the Data”



Sever-side Analytics

“Bringing Code to the Data”



Server-side Analytics

The case for Server-side Analytics:

- 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 analysis result is delivered to the client.
- Server-side analysis saves a lot of time if:

  (size of result) << (size of data operated on)

Server-side Analytics

As an example, one NASA-funded university research group recently acquired a petabyte data storage system.

- The group intends to do some very complex analyses of data that are residing at NCAR, NASA Ames and the Texas Advanced Computing Center.
- The new local data storage is providing a great opportunity to bring together massive data sets that are stored at disparate locations.
- Despite using of sophisticated data transfer optimizations tuned in coordination with the transmitting data centers, the current estimate is that just to download these requested data sets from the 3 remote locations ***will take upwards of 4 months!***

Server-side Analytics

More rationale:

- Reduces duplication by multiple groups and encourages community sharing.
- Increases interoperability of diverse data sets.
- The analyzing and assimilating large observational data sets will become more efficient leading to more rapid scientific discoveries.

Server-side Analytics

Many examples are being worked on.

- The NASA research community has explored with some prototyped analytics processing and, in some cases, gone on and implemented specific SSA applications.
- The APPENDIX has discussions on the following projects: GrADS, MERRA/AS, OceanWorks, Western States Water Mission, SWOT and NISAR science data systems, researcher's tools for accessing and analyzing SDO data, SciServer, the NOAO DataLab and GAVIP.

Server-side Analytics

BDTF 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.

Most importantly, SMD needs to take a leadership role to ensure that new SSA architectures are implemented where most needed and other proposals for lower priority implementations are reserved.

Server-side Analytics

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) << (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 Level 1 requirements.

Today's Report

- Four white papers with accompanying 4 findings and 4 recommendations
- Are the Science Data Archives Ready to Meet Future Challenges?
- Finding on Courses offered by the Frontier Development Labs.
- Recommendation on organizing SMD's Data Science and High-performance computing programs.
- Recommendation on DS&C expertise on SMD Advisory Committees

Are SMD's Science Data Archives Ready to Meet Future Challenges?

The BDTF had some concern on each project's approach to preparing for the future. Each data archive project was asked to describe their approach. In particular, they were asked to respond to three specific questions:

- What are the processes for planning for future (5-10 years) capabilities of your service? How and from whom do you gather input for this planning process and where does input typically come from? What new feature(s) do you really want to implement?
- What feature(s) of your service would you like to stop performing? How do you gather input for making such decisions and where does input typically come from? What is preventing you from stopping?
- What steps you are taking to make your data interoperable with allied data sets from other data sites in and out of NASA? How do you find allied data sets and what criteria make data sets candidates for enabling interoperability?

Are the Science Data Archives Ready to Meet Future Challenges?

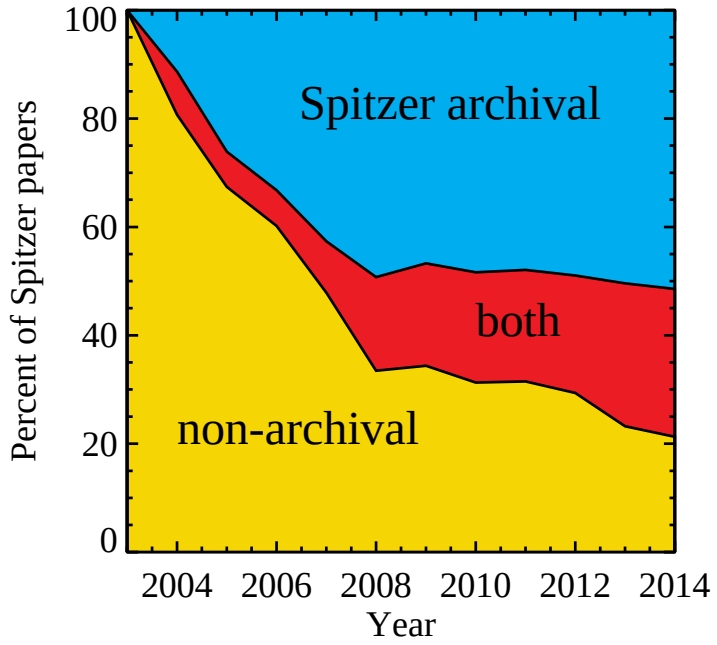
Finding: The BDTF finds that SMD's data archive programs and projects are performing quite well and are properly taking steps to modernize and meet future challenges. Like all infrastructure projects they could use increased budgets and will put them to use wisely. Selective implementations of recommendations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

Finding: 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%.

Universal Experience

The use of data in NASA archives can **double the science** of the original mission.

This is the trend in all SMD's science data archives!



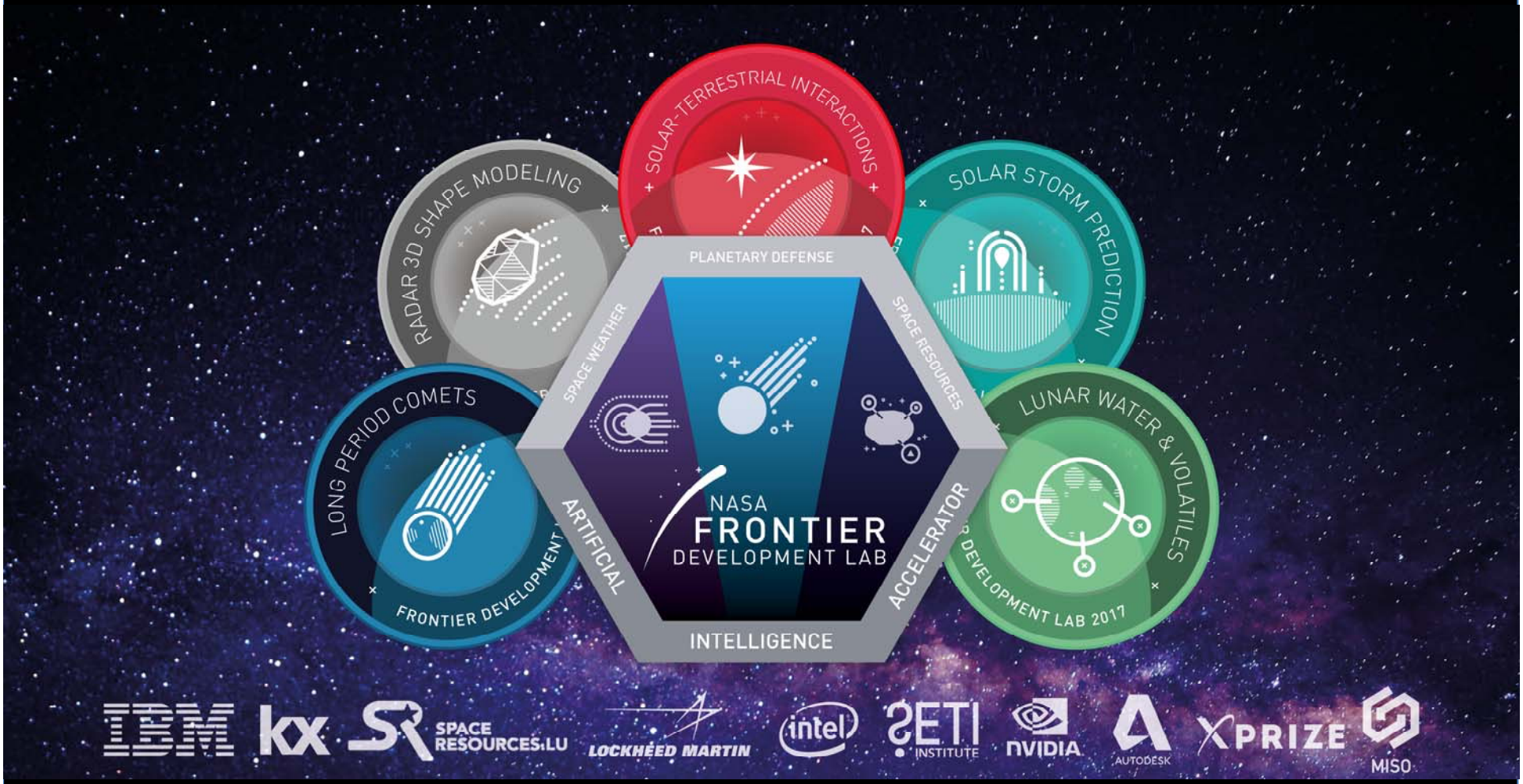
Taken from IPAC's presentation to the BDTF on 11/2/2017

Are the Science Data Archives Ready to Meet Future Challenges?

Recommendation: Given the empirical evidence that community use of these archives leads increasingly to discovering new science and combined with size of 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 data archives may be overlooked or only rediscovered by new missions at great expense.



(http://www.frontierdevelopmentlab.org)

Educating Early-Career Professionals in Data Science Approaches to NASA's Science Analysis Problems

In its 2nd year NASA's Frontier Development Lab 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 formal, long term education as well as more short-form 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.

Data Science and Computing Division

Recommendation: SMD should establish a new division that would focus on cross-cutting data science and computing projects and whose responsibilities would include:

(1) Managing High-Performance Computing Program including the High-End Computing Capability at Ames and the NASA Center for Climate Simulation at Goddard as well as activities leading to the future such as participating in DOE's Exascale Computing Program. •

(2) Establishing 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) Connecting the NASA science world to the National Data Superhighway. *

(4) Initiating projects to develop sever-side analytics services where most needed. •

(5) Initiating projects aimed at modernizing and streamlining SMD's sponsored modeling workflows. •

(6) Leading 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) Coordinating 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) Serving 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 as well as other Federal departments and agencies. * •

* BDTF Recommendation – July 22, 2017 • BDTF Recommendation – Nov 28, 2017

DS&C Divison (cont'd)

Rationale: 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 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.

SMD Projects Supporting Archives and Computational Capabilities

	FY16	FY17	FY18	FY19	FY20	FY21	FY22
Total	272.8	280.9	279.9	276.6	284.0	292.8	301.3
Planetay Science	17.3	16.9	17.9	18.3	18.9	19.1	19.2
Planetary Data System	15.0	14.5	15.4	15.6	16.2	16.3	16.4
Science Data and Computing	2.3	2.4	2.5	2.7	2.7	2.8	2.8
Astrophysics Data Curation and Archival	18.7	16.9	18.8	18.9	18.9	18.9	18.9
Heliophysics	3.3	3.4	3.5	3.6	3.4	3.5	3.5
Space Physics Data Archive	2.3	2.3	2.3	2.3	2.3	2.3	2.3
Solar Data Center	1.0	1.1	1.2	1.3	1.1	1.2	1.2
Earth Science	233.4	243.7	224.7	235.9	242.7	251.3	259.8
High-End Computing Capability	43.6	44.7	32.5	47.5	48.4	49.1	50.4
Scientific Computing	21.8	21.3	21.8	21.5	22.2	22.5	23.2
Multi-Mission Operations	168.0	177.7	170.5	166.8	172.1	179.6	186.2
Modular Supercomputing Facility			15.0				

DS&C Division- Resources

- Incorporate existing staff and budgets for “High-End Computing Capability”

- Create two staff positions:
  - Data Science Program Scientist – permanent
  - Science Data Network Officer – temporary
  - Both initially staffed by IPAs
- Work with two Federal agencies:
  - DOE (Exascale computing & Science Data Superhighway)
  - NSF (Big Data Hubs and Spokes)
- Create 3 ROSES Programs
  - Data Science Applications - $10M/yr
  - Big Data Hubs & Spokes – none specified-negotiate w/ NSF
  - Science Data Super Highway - $3-5M/yr

DS&C Division (cont'd)

Consequences: 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

No stovepipes!

It will be the responsibility of SMD senior management to ensure that the working interfaces are 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 stove pipes within the SMD science domain.

Data Science and Computing Advisory Positions

Recommendation: In staffing the Science Committee and the four thematic 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. It is important but perhaps not necessary that these appointees be or have been active in one of SMD's science endeavors.

DS&C Advisory Positions (cont'd)

Rationale: 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.

Consequences: 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.

Chuck's

The Soapbox

Scope of the Big Data Task Force

The scope of the Task Force includes all NASA Big Data programs, projects, missions, and activities. The Task Force will focus on such topics as exploring the existing and planned evolution of NASA's science data cyber-infrastructure that supports broad access to data repositories for NASA Science Mission Directorate missions; best practices within NASA, other Federal agencies, private industry and research institutions; and Federal initiatives related to big data and data access.

Abstracted from the Terms of Reference, Ad Hoc Task Force on Big Data, signed by the Administrator on Jan. 8, 2015.

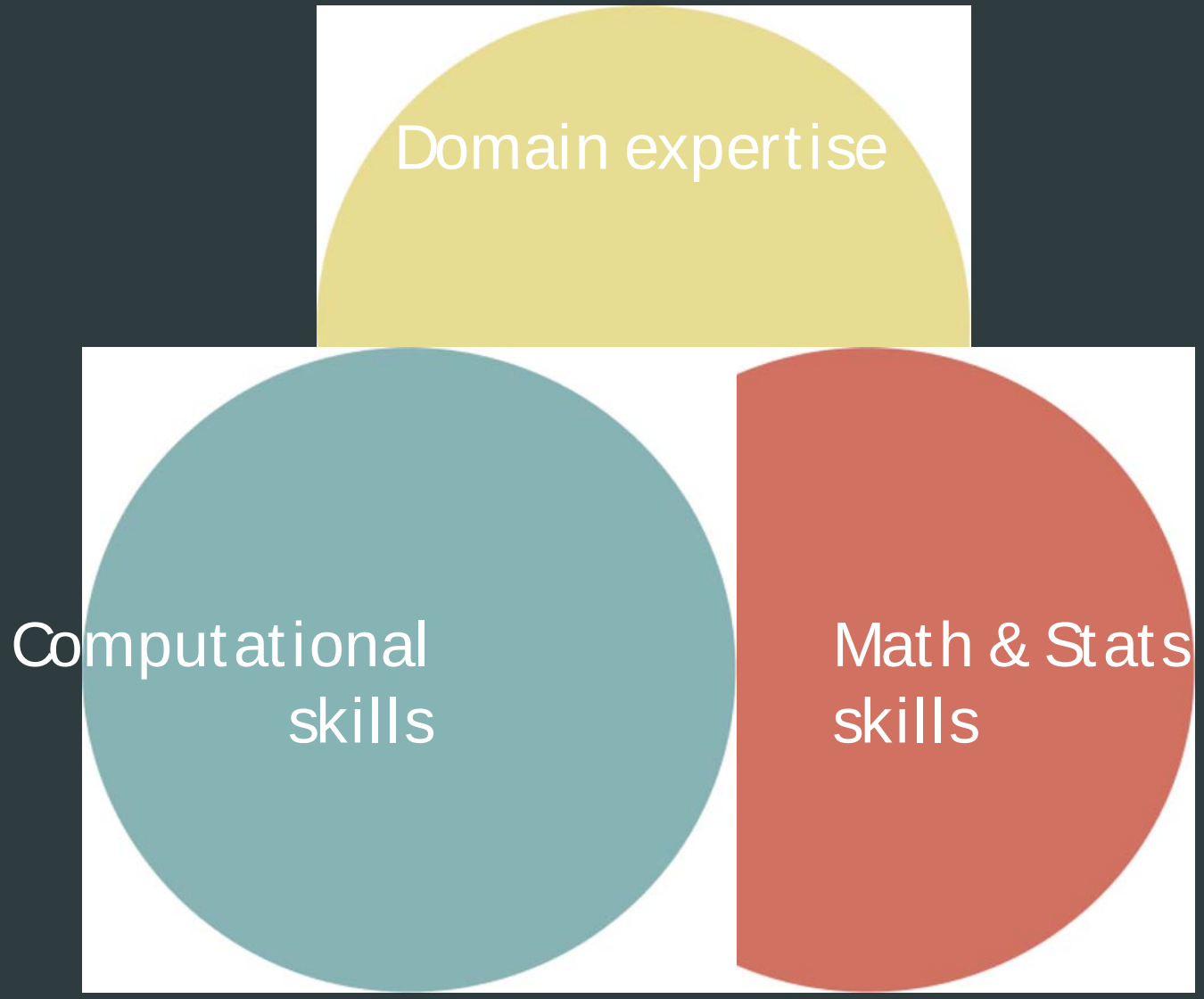
The BDTF by the #s

- Months in existence: 22
- Members of the BDTF: 9
- Number of Meetings: 6
- Number of presentations received: 68
- Number of pages in the minutes: 185
- Number of Findings: 14
- Number of Recommendations: 12
- Number of White Papers: 4

What is our take away message?

1. There are a few areas that SMD is behind:
   - Participating in the science data super highway
   - Leadership in data science applications and education
2. SMD needs to view its science data archives as the sources of half its new science.
3. There are other areas were SMD needs to start applying leadership in a measured, systematic way:
   - Modeling workflows
   - Server-side analytics

How to succeed in Data-Driven Discovery – the right skill mix



Credit: Chris Mentzel, Gordan and Betty Moore Foundation

Thank you's from the BDTF

- Erin Smith and Elaine Denning for setting us up.
- Gerry Smith for finishing us off!
- Brad Peterson and the members of the Science Committee for putting up with my presentations.
- Amy Reis for getting us there and back.
- Joan Zimmerman for recording 185 pages of minutes!
- And all the people who shared their time, exciting work and fantastic ideas with us.

BACK UP CHARTS

The BDTF Held Six Meetings

1.	Feb 16, 2016	HQ
2.	June 28-30, 2016	GSFC
3.	Sept 28-30, 2016	ARC
4.	Mar 6-7, 2017	HQ
5.	June 22-23, 2017	Tele-meeting
6.	Nov 1-3, 2017	JPL

Interviewed NASA programs & projects for large-scale data and computing projects – baseline and outlook for the future.

Interviewed project officers from these external organizations:

NSF, DOE, Moore Foundation, UCSD.

Members have interviewed many external organizations including Google, Simons Foundation, GISS, etc.

Made 6 reports to the Science Committee with:

14 Findings

12 Recommendations