1st Report from the
Big Data Task Force

Charles P. Holmes

Chair, BDTF

March 10, 2016

Topics

- Charter
- Membership
- Agenda of 1st meeting – Feb 16, 2016
- Legacy
- Work plan
- Highlights from the 1st meeting
- Soap Box

Charter 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.

Membership of the BDTF

Name	Dept./Center	Organization
Charles Holmes - Chair	Retired	Formerly NASA HQ
Reta Beebe	Dept of Astronomy	NMSU
Neal Hurlburt	Solar and Astrophysics Lab.	Lockheed Martin
James Kinter	Center for Ocean-Land-Atmosphere Studies	GMU
Clayton Tino	Software Architect	Virtustream / EMC
Raymond Walker	Institute for Geophysics and Planetary Physics	UCLA
Erin Smith – Exec. Sec.	SMD	HQ NASA (and Ames)

Plus two more members “well into the clearance process”, and two other nominations beginning the process.

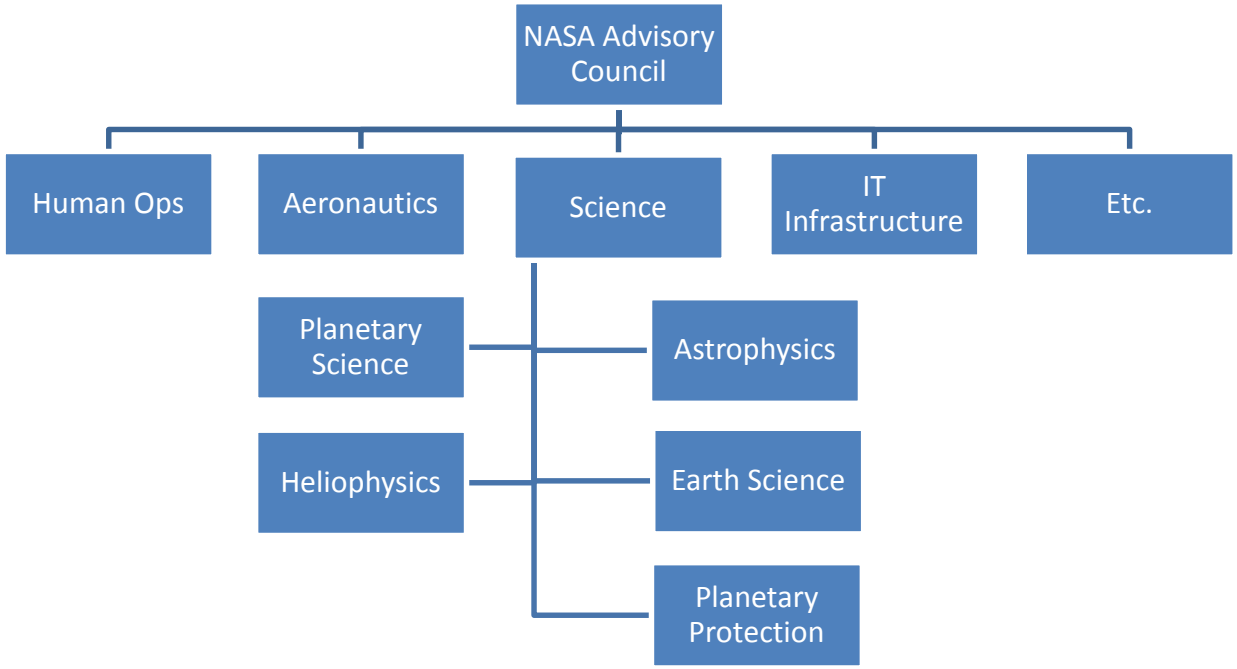
Agenda for the 1st Meeting of the BDTF

HQ NASA, Feb 16, 2016

• Opening Remarks / Introduction of Members	Dr. Erin Smith
• Big Data Charter / Subcommittee Feedback	"
• Legacy from NAC IT Infrastructure Cmte	Dr. Charles Holmes
• Greetings from the Science Committee	Dr. Bradley Peterson
• Heliophysics Big Data	Dr. Jeffrey Hayes
• Earth Science Big Data	Dr. Kevin Murphy
• Astrophysics Big Data	Dr. Paul Hertz
• Planetary Science Big Data	Dr. Michael New
• Supercomputing Big Data	Dr. Tsengdar Lee
• NSF's "Big Data Hubs and Spokes"	Dr. Fen Zhao, NSF
• Discussion on the Draft Work Plan	Dr. Charles Holmes
• Public Comment	
• Discussion / Findings / Recommendations/Work Plan/Future Meetings	

Big Data Task Force:
Legacy from
NAC IT Infrastructure Committee

NAC Structure 2010 - 2013



Last slide of the last presentation
By Chair of the ITIC Presentation to the NAC – Dec 12, 2013

NAC Committee on IT Infrastructure Recommendation #1 July 31, 2013

- Recommendation: The NASA NAC ITIC & Science Committees should collaboratively explore the existing and planned evolution of NASA's science data cyberinfrastructure that supports broad access to data repositories for NASA SMD missions. This exploration should be undertaken in the context of effective practices within NASA, other Federal agencies, as well as industry and research institutions.

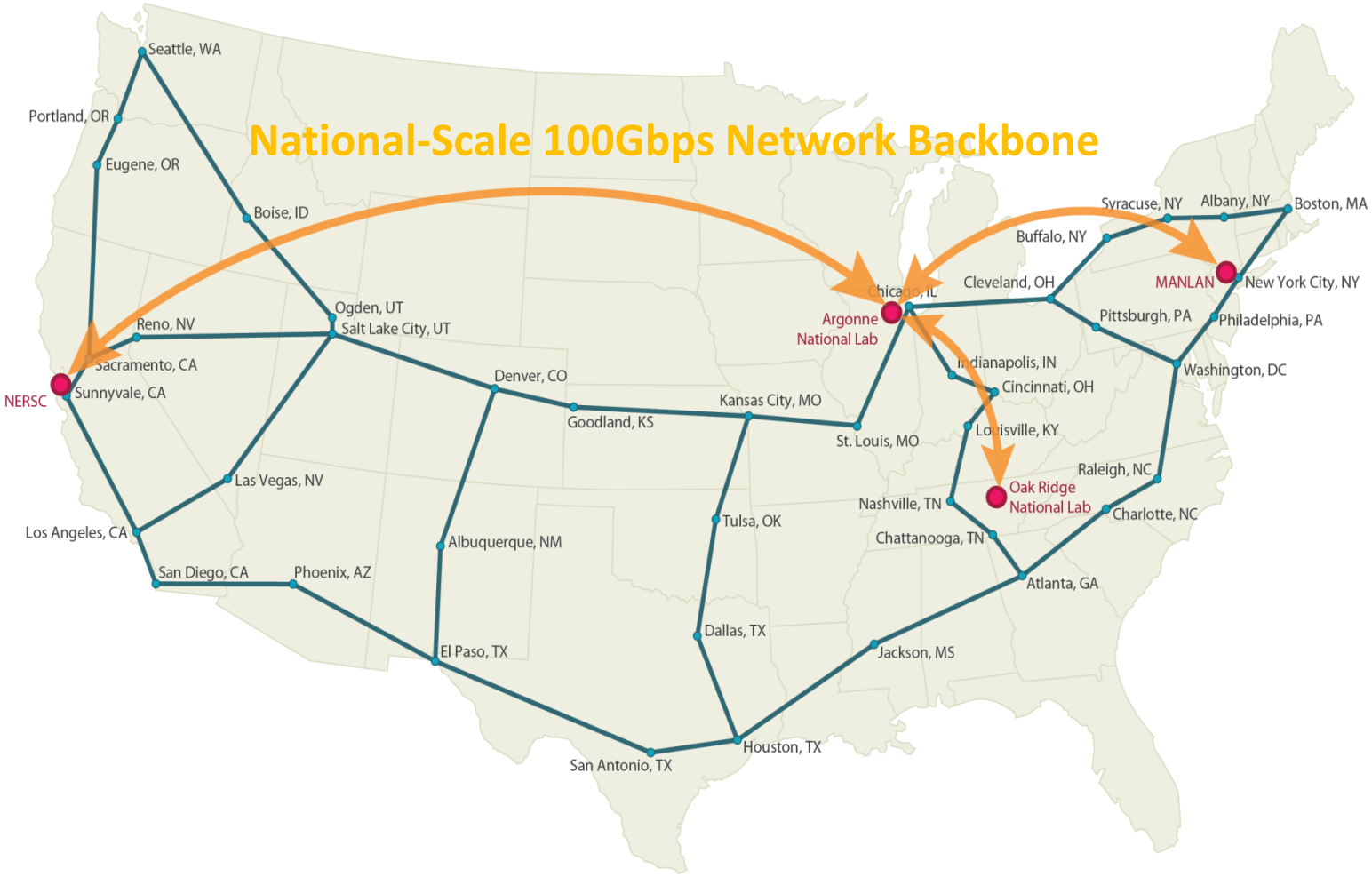
Wording Agreed to by Both ITIC and Science Committees
July 31, 2013

Work Will Continue as Big Data Taskforce Under Science Committee

NAC Committee on IT Infrastructure Recommendation #1

- Recommendation: To enable NASA to gain experience on emerging leading-edge IT technologies such as:
  - Data-Intensive Cyberinfrastructure,
  - 100 Gbps Networking,
  - GPU Clusters, and
  - Hybrid HPC Architectures,
  - we recommend that NASA aggressively pursue partnerships with other Federal agencies, specifically NSF and DOE, as well as public/private opportunities.
  - We believe joint agency program calls for end users to develop innovative applications will help keep NASA at the leading edge of capabilities and enable training of NASA staff to support NASA researchers as these technologies become mainstream.

Partnering Opportunities with DOE: ARRA Stimulus Investment for DOE ESnet



Source: Presentation to ESnet Policy Board

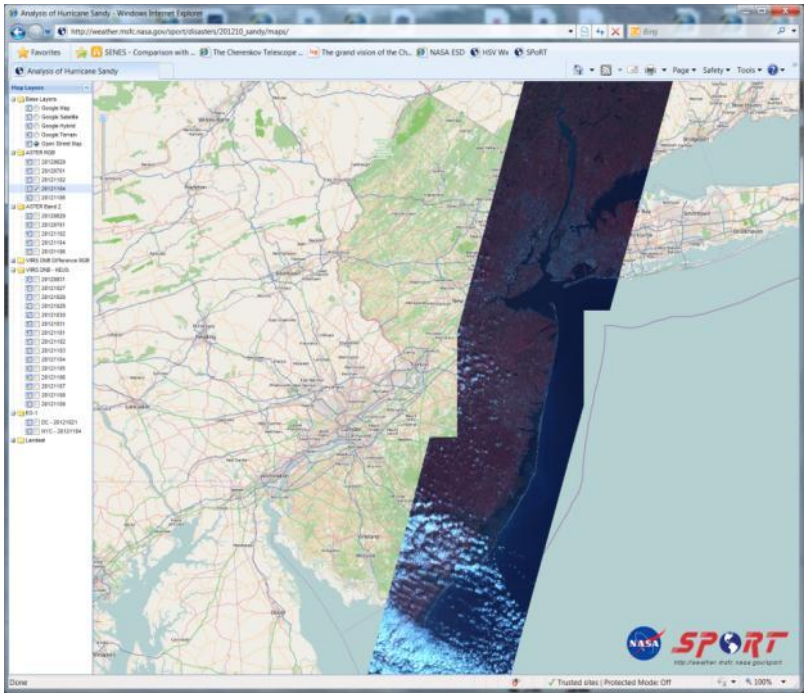
Web Services to Support Disaster Applications

Short-term Prediction Research and Transition Center

Need for access to data and products supporting disaster applications "anytime and from any place"

SPoRT Web Services

- tiled imagery for a "Google Earth" roam and zoom
- web-based applications - tiled web service link
- Android and IPhone "apps"



Tiled web service for Hurricane Sandy

Crowdsourcing Science: Galaxy Zoo and Moon Zoo Bring the Public into Scientific Discovery

More than 250,000 people have taken part in Galaxy Zoo so far.
In the 14 months the site was up Galaxy Zoo 2 users helped us make over 60,000,000 classifications. Over the past year, volunteers from the original Galaxy Zoo project created the world's largest database of galaxy shapes.

www.galaxyzoo.org

Re-organization of the NASA Advisory Council (Memo signed April 28, 2014)

The NASA Administrator shall establish the following Council committees, subcommittees, and task forces:

- Aeronautics Committee.
- Human Exploration and Operations Committee.
- Science Committee.
  - Astrophysics Subcommittee.
  - Earth Science Subcommittee.
  - Heliophysics Subcommittee.
  - Planetary Protection Subcommittee.
  - Planetary Science Subcommittee.
  - ***Ad Hoc Task Force on Big Data.***
- Technology, Innovation, and Engineering Committee.
- Institutional Committee.
- Ad Hoc Task Force on Science, Technology, Engineering, and Mathematics (STEM) Education.

HIGHLIGHTS OF PRESENTATIONS TO THE BDTF ON FEB 16, 2016

National Aeronautics and Space Administration

Astrophysics




Big Data Task Force of the NAC Science Committee
Washington DC
February 16, 2016

Paul Hertz
Director, Astrophysics Division
Science Mission Directorate
@PHertzNASA

WFIRST

Wide-Field Infrared Survey Telescope

Wide-Field Infrared Survey Telescope

Top priority of 2010 Decadal Survey

Science themes: Dark Energy, Exoplanets, Large Area Near Infrared Surveys

Mission: 2.4m widefield telescope at L2; using existing hardware, images 0.28deg^2 at 0.8-2μm

Instruments (design reference mission): Wide Field Instrument (camera plus IFU), Coronagraph Instrument (imaging/IFS)

Phase: Currently in pre-formulation

http://wfirst.gsfc.nasa.gov/

CURRENT STATUS:

- Completed Mission Concept Review (MCR) held in December 2015
- Formulation Science Investigation Teams selected in December 2015; first Formulation Science Working Group meeting in February 2016
- Planning for Key Decision Point A (KDP-A) in Feb 2016
  - Official start of formulation phase
  - Supported by FY16 appropriations
  - SMD Program Management Council January 26, 2016
  - Agency Program Management Council on February 17, 2016
- Industry RFI released July 2015; RFP for industry studies released in January 2016
- Other activities include:
  - Technology development for detectors and coronagraph (with STMD); prototyping key parts
  - Assessment of telescopes + risk mitigation
  - Mission design trades; performance simulations
- Maturing key technologies by FY19
  - H4RG infrared detectors for widefield imager
  - Internal coronagraph for exoplanet characterization
  - Milestones on road to achieve TRL-5 by end of CY16, TRL-6 by end of CY18; reports made public

Astrophysics: Science and Data Archives

Challenges identified by 2015 Senior Review:

- The infrastructure and the technological approaches that are being used will certainly be obsolete at the end of the next 4-5 year review cycle.
- Network bandwidths available to the data centers will soon be two generations behind the current standard for research internet.
- Data centers need to raise concerns about sustainability where they exist, regardless of budgetary constraint

  http://science.nasa.gov/media/medialibrary/2015/07/08/NASA-AAPR2015-FINAL.pdf

- Cloud computing and associated commercial services.
  - It is clear that some of our services cannot be migrated yet, while we do utilize clouds for some services.
  - It may well be that cloud computing is a good fit for creating the huge simulations we need (for Euclid and WFIRST, and incidentally also LSST), and also for running joint processing.
  - It is a matter of how well matched the science requirements are to the commercial services, first their technical services but also their charging model (it may be cheaper to work with DOE supercomputers).

National Aeronautics and Space Administration

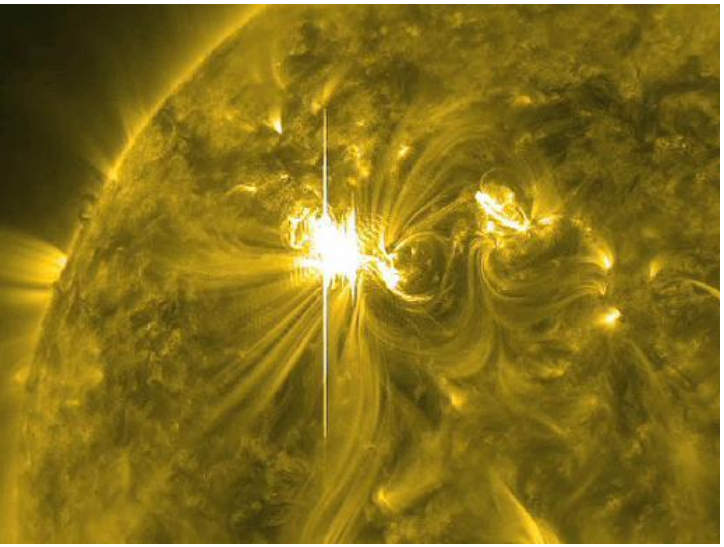
Ad hoc Big Data Task Force

February 16, 2016

Jeffrey J.E. Hayes
Heliophysics Division
Science Mission Directorate

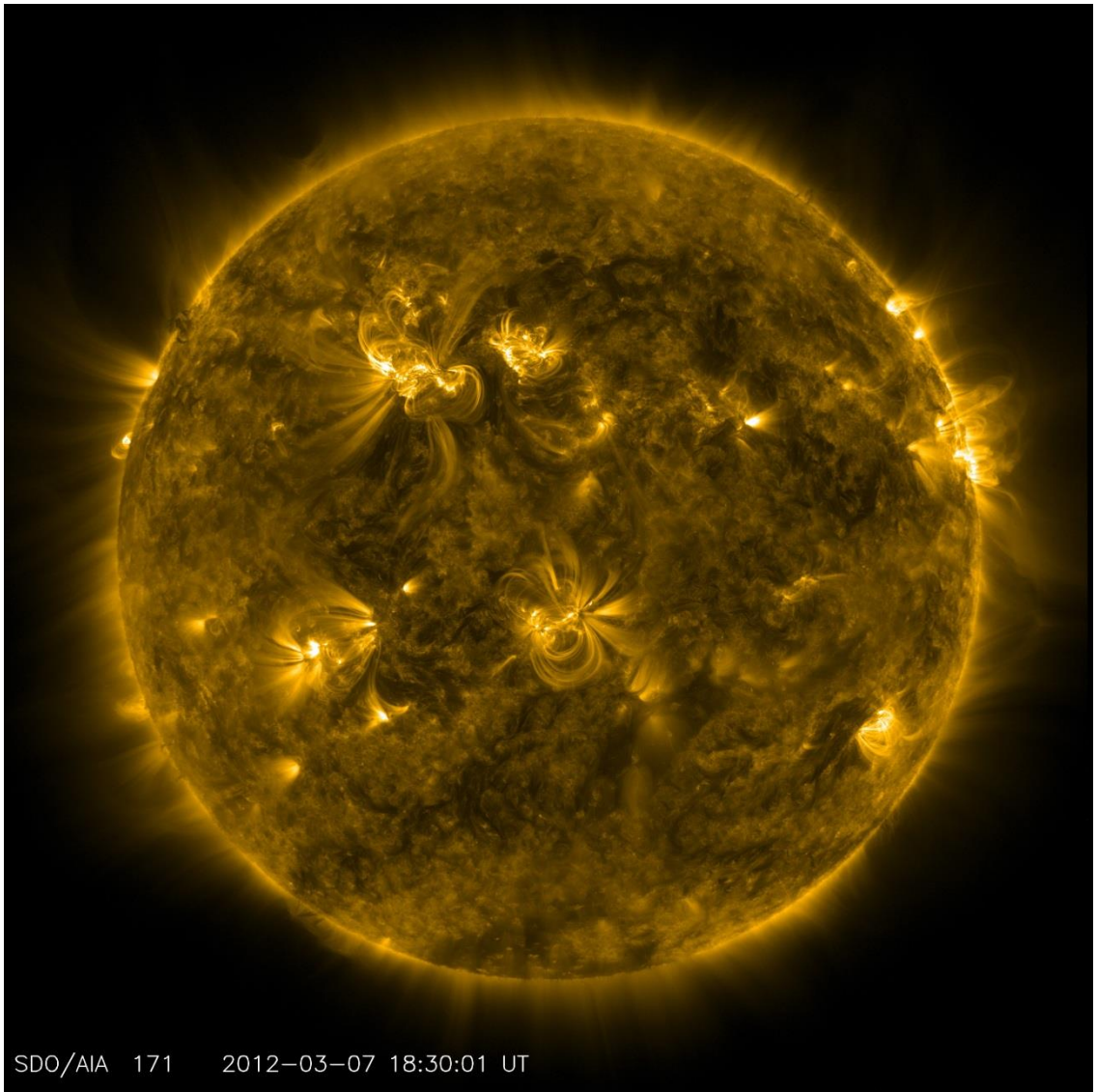
Solar Dynamics Observatory
4096x4096 AIA Camera – 57, 600 Images/Day

JSOC is Archiving ~5TB/day From 6 Cameras
Leads to over 1 Petabyte per year!



March 6, 2012 X5.4 Flare from Sunspot AR1429 Captured by the Solar Dynamics Observatory (SDO)
in the 171 Angstrom Wavelength

Credit: NASA/SDO/AIA



SPDS 1990

NASA

What scientists want to do

- Model physical processes
- Study physical interactions
- Use multi-source data

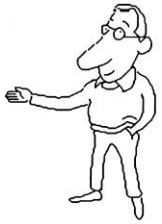
NASA

What scientists don't want to do

- Wonder if they have all the data available
- Wait a long time to get the data
- Spend time and money getting data into a useable format
- Spend time and money fighting computers

NASA

What scientists want a data system to do



- Locate relevant data
- Access data quickly
- Create versatile data sets
- Use the data easily

NASA

What data systems managers want scientists to do

- Hold to agreed upon standards
- Submit all relevant data
- Document all relevant data
- Report location of relevant data

Viewgraphs (!) from the first "Space Physics Data System" meeting in 1990. Our needs have been remarkably constant. We now do this!

The Current PDS

The hierarchy shown is misleading; actual operations are collaborative, as befits a federation.

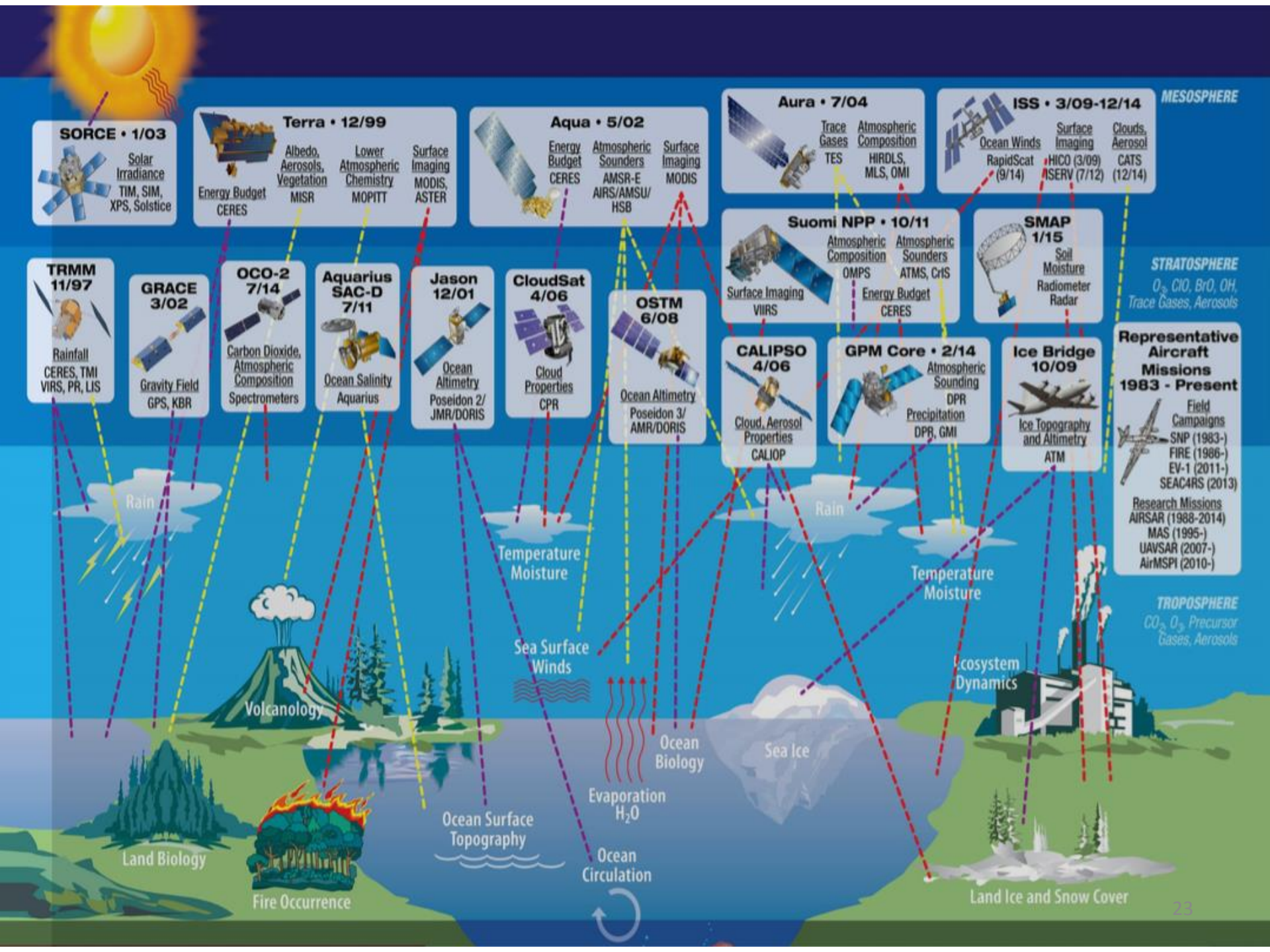
Performance reviewed this month by non-PDS peer reviewers.

Re-competed in August.

Scale of the PDS

Node	Data Volume (TB)
Atmospheres	3.0
Cartography and Imaging	825.0
Geosciences	165.0
NAIF	0.5
Planetary Plasma Interactions	7.5
Ring-Moon Systems	1.6
Small Bodies	3.1
TOTAL	1,000.7

- Total volume is currently ~1 PB.
- Note, though, that no user ever needs to access, search, download, or process the entirety of the PDS.
- Almost all computations on data are performed on individual workstations.

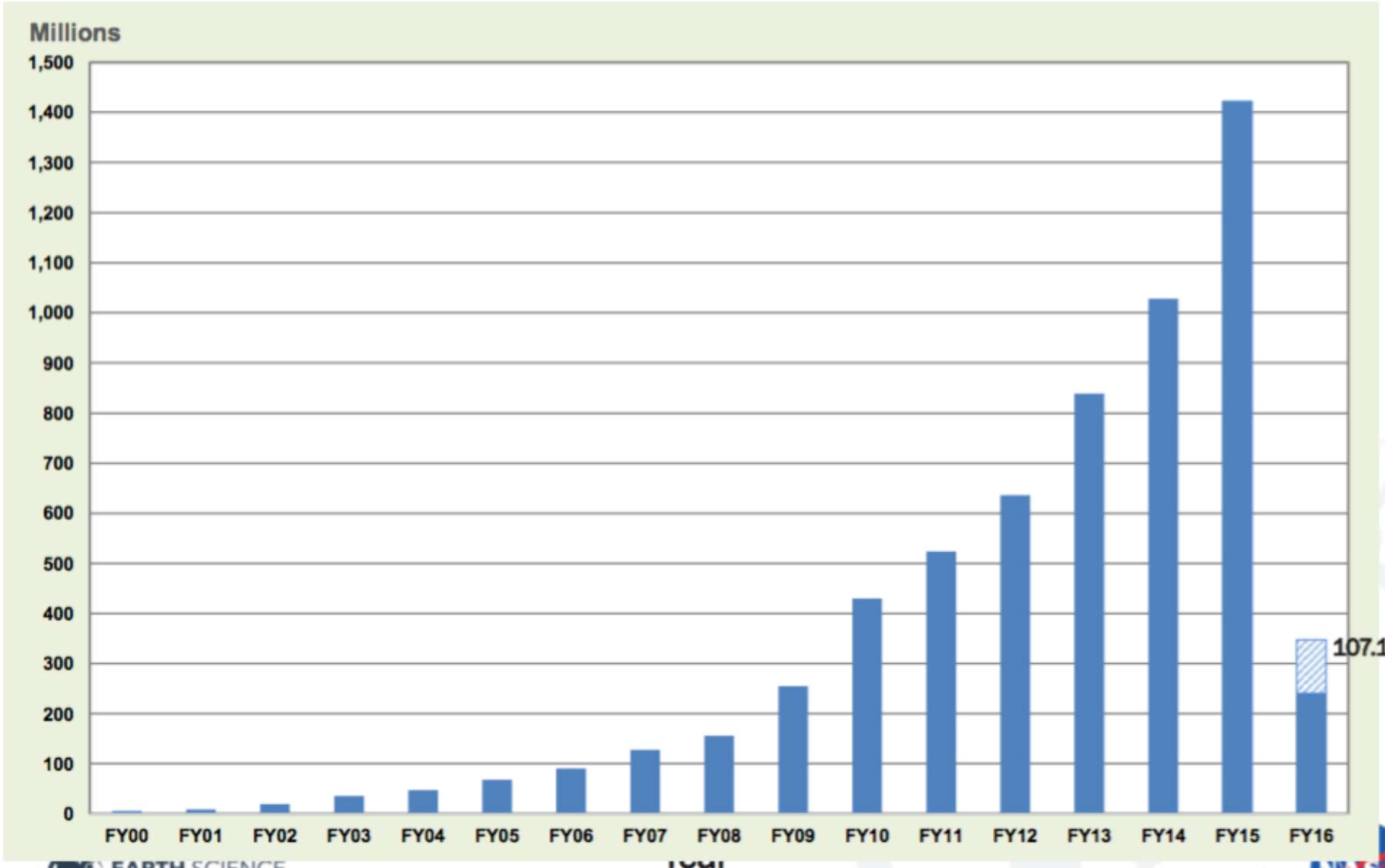

SORCE • 1/03
Solar Irradiance
TIM, SIM, XPS, Solstice
Terra • 12/99
Energy Budget
CERES
Albedo, Aerosols, Vegetation
MISR
Lower Atmospheric Chemistry
MOPITT
Surface Imaging
MODIS, ASTER
Aqua • 5/02
Energy Budget
CERES
Atmospheric Sounders
AMSR-E AIRS/AMSU/ HSB
Surface Imaging
MODIS
Aura • 7/04
Trace Gases
TES
Atmospheric Composition
HIRDLS, MLS, OMI
ISS • 3/09-12/14
Ocean Winds
RapidScat (9/14)
Surface Imaging
HICO (3/09) RSERV (7/12)
Clouds, Aerosol
CATS (12/14)
MESOSPHERE
TRMM 11/97
Rainfall
CERES, TMI VIRS, PR, LIS
GRACE 3/02
Gravity Field
GPS, KBR
OCO-2 7/14
Carbon Dioxide, Atmospheric Composition
Spectrometers
Aquarius SAC-D 7/11
Ocean Salinity
Aquarius
Jason 12/01
Ocean Altimetry
Poseidon 2/ JMR/DORIS
CloudSat 4/06
Cloud Properties
CPR
OSTM 6/08
Ocean Altimetry
Poseidon 3/ AMR/DORIS
Suomi NPP • 10/11
Surface Imaging
VIIRS
Atmospheric Composition
OMPS
Atmospheric Sounders
ATMS, CrIS
Energy Budget
CERES
SMAP 1/15
Soil Moisture
Radiometer Radar
STRATOSPHERE
O_3, ClO, BrO, OH, Trace Gases, Aerosols
CALIPSO 4/06
Cloud, Aerosol Properties
CALIOP
GPM Core • 2/14
Atmospheric Sounding
DPR
Precipitation
DPR, GMI
Ice Bridge 10/09
Ice Topography and Altimetry
ATM
Representative Aircraft Missions 1983 - Present
Field Campaigns
SNP (1983-)
FIRE (1986-)
EV-1 (2011-)
SEAC4RS (2013)
Research Missions
AIRSAR (1988-2014)
MAS (1995-)
UAVSAR (2007-)
AirMSPI (2010-)
TROPOSPHERE
CO_2, O_3 Precursor Gases, Aerosols
Rain
Temperature Moisture
Rain
Temperature Moisture
Sea Surface Winds
Volcanology
Ecosystem Dynamics
Ocean Biology
Sea Ice
Evaporation H_2O
Land Biology
Ocean Surface Topography
Ocean Circulation
Fire Occurrence
Land Ice and Snow Cover

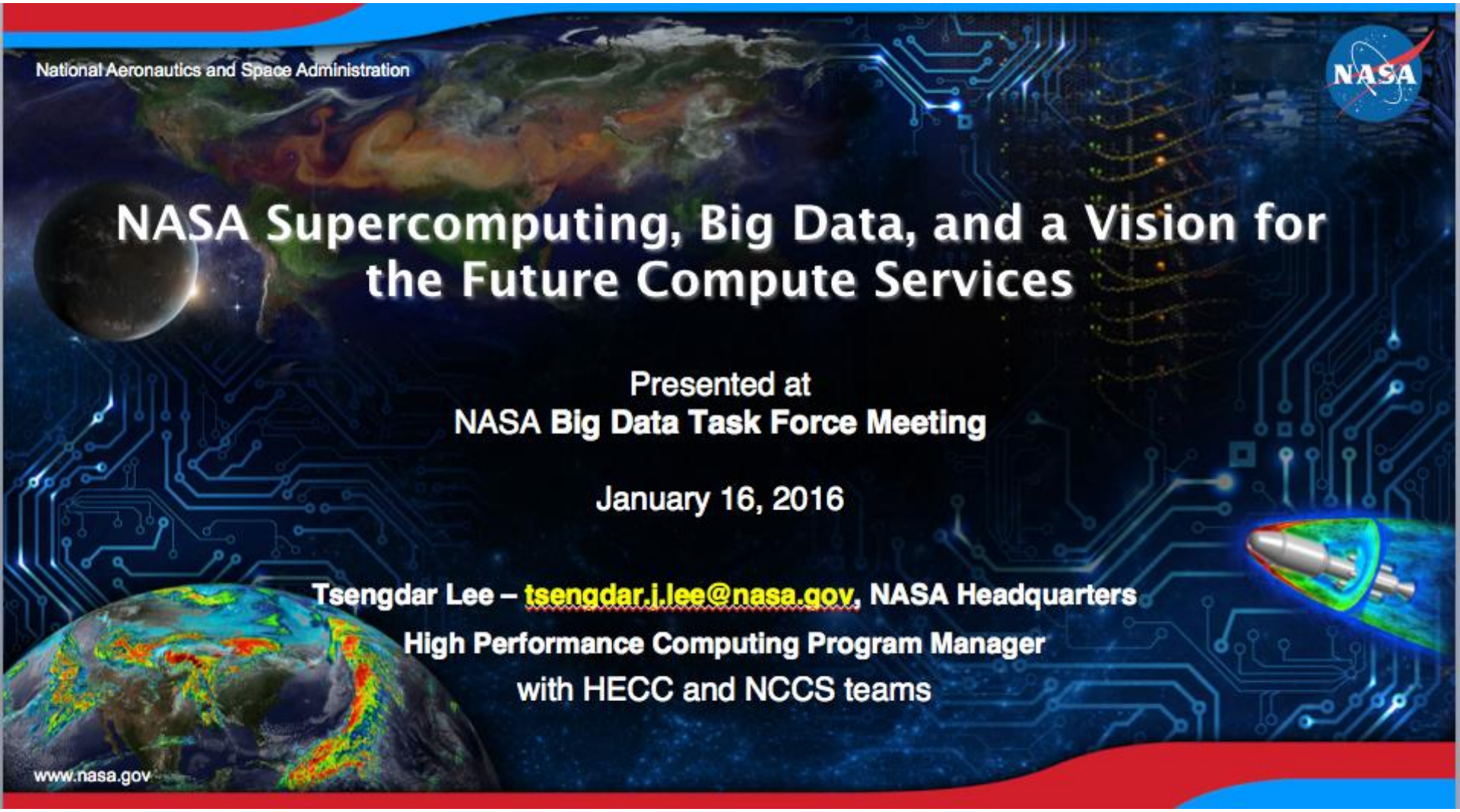
Extensive Data Collection

- EOSDIS data collection includes over ~9200 data types
  - Land
    - » Cover & Usage
    - » Surface temperature
    - » Soil moisture
    - » Surface topography
  - Atmosphere
    - » Winds & Precipitation
    - » Aerosols & Clouds
    - » Temperature & Humidity
    - » Solar radiation
  - Ocean Dynamics
    - » Surface temperature
    - » Surface wind fields & Heat flux
    - » Surface topography
    - » Ocean color
  - Cryosphere
    - » Sea/Land Ice & Snow Cover

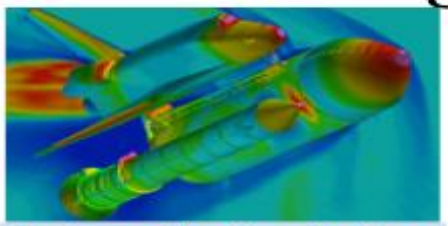
- Human Dimensions
  - » Population & Land Use
  - » Human & Environmental Health
  - » Ecosystems

EOSDIS Products Delivered: FY00 – Dec.'15

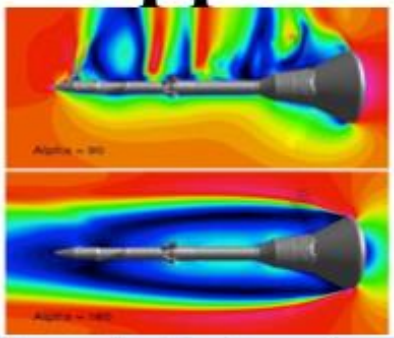



National Aeronautics and Space Administration
NASA
NASA Supercomputing, Big Data, and a Vision for the Future Compute Services
Presented at
NASA Big Data Task Force Meeting
January 16, 2016
Tsengdar Lee – tsengdar.j.lee@nasa.gov, NASA Headquarters
High Performance Computing Program Manager
with HECC and NCCS teams
www.nasa.gov

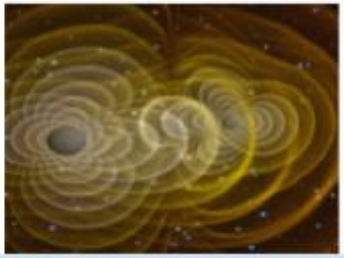
Strategic Support for NASA Programs



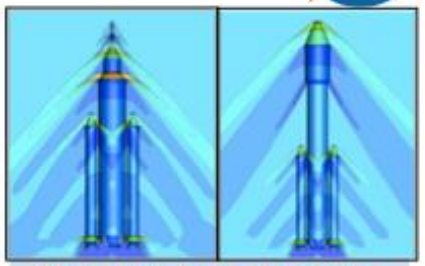
External tank redesign



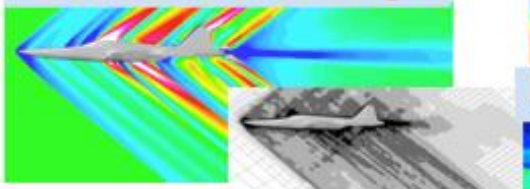
Launch abort system



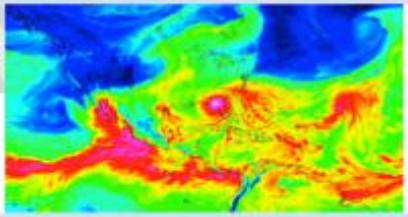
Merging black holes



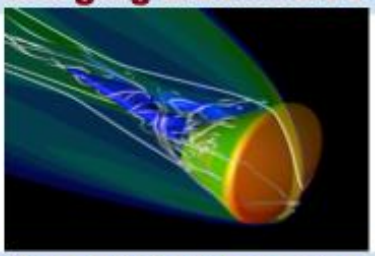
SLS vehicle designs



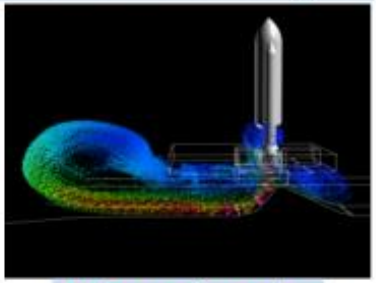
Sonic boom optimization



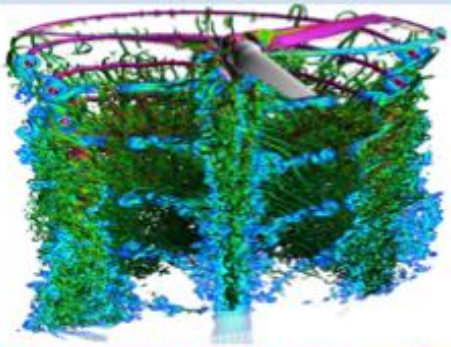
Hurricane prediction



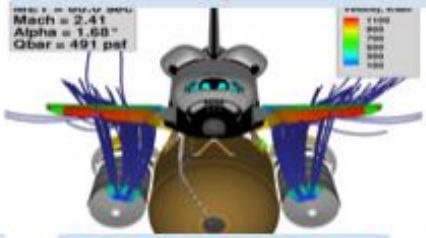
Orion/MPCV reentry



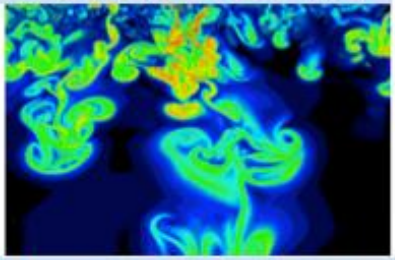
Flame trench



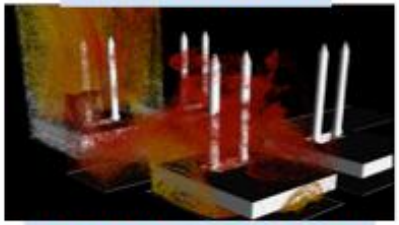
Rotary wing aerodynamics



Debris transport



Solar magnetodynamics



SRB burn in VAB

BIG DATA REGIONAL INNOVATION HUBS & SPOKES

Accelerating the Innovation Ecosystem

Fen Zhao
Staff Associate, Strategic Innovation
CISE Directorate, Office of the Assistant Director

THE HISTORY BEHIND BD HUBS

The National Big Data R&D Initiative & Data to Knowledge to Action (Data2Action)

MAR 2012 — Launch
OSTP and NITRD Agencies kick off National Big Data R&D Initiative with new federal programs totaling $200M

MAY 2013 — Big Data Partnerships Workshop
Industry, academia, and government representatives gathered to learn about current Big Data partnership and brainstorm new ideas

NOV 2013 — Data2Action
90 organizations announce 29 new Big Data partnerships supported by $100M in non-federal funds

JUN 2014 — Partnerships Bear Fruit
Partnerships update NITRD on midterm outcomes from announced projects

MAR 2015 — BDHubs
NSF initiates BDHubs effort to sustain and scale up collaborative Big Data innovation activities

Goals of the Big Data Hubs initiative

- Enables teams of data science researchers to come together with domain experts, with cities and municipalities, and with anchor institutions to establish and grow collaborations that will accelerate progress in a wide range of science and education domains with the potential for great societal benefit.
- The BD Hubs will be sites for transitioning research into practice. They will also educate and train the next-generation workforce in data science.
- The projects from this first phase of the program will help establish the governance structure of the BD Hub consortia, support the recruitment of executive directors and administrative staff for each BD Hub and begin developing approaches for inter BD Hub collaborations.

NSF's Big Data Regional Innovation Hubs

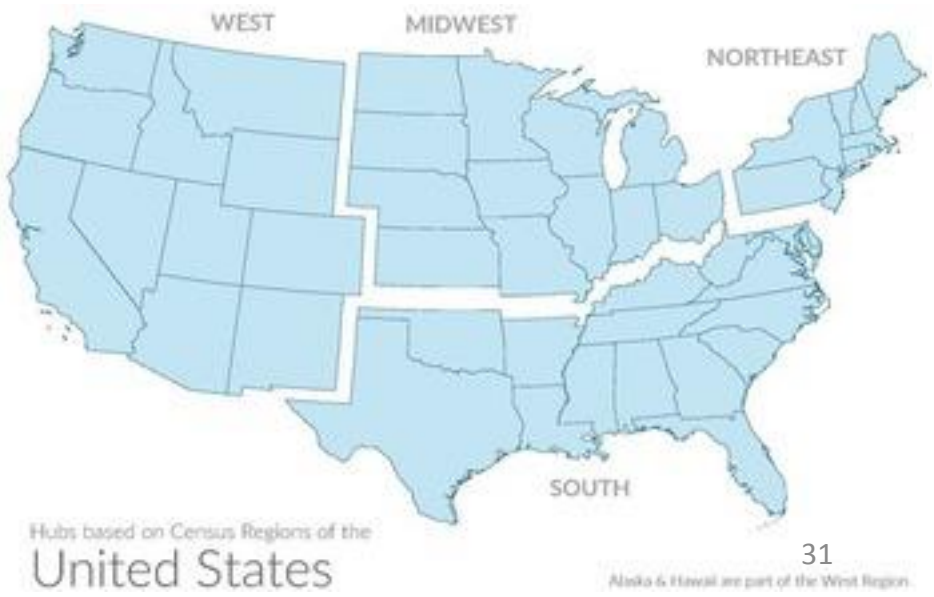
- The cover all 50 states and include commitments from more than 250 organizations--from universities and cities to foundations and Fortune 500 corporations--with the ability to expand further over time.

NSF budget ~$20M/year

(for five years?) to build

an infrastructure for

big data.



NSF's Big Data Hubs

The consortia are coordinated by top data scientists at

- Columbia University (Northeast Hub),
- Georgia Institute of Technology and the University of North Carolina (South Hub),
- the University of Illinois at Urbana-Champaign (Midwest Hub) and
- the University of California, San Diego, the University of California, Berkeley, and the University of Washington (West Hub).

Opportunity for NASA participation?

- Yes!
- NASA PIs may want to become involved with the regional hubs and spokes to overlay their some science tasks on this new infrastructure.
- BDTF members are taking a closer look at this activity. A subject for our next meeting.

Chuck's

The Soapbox

Draft Work Plan for the BDTF

Charles P Holmes

February 16, 2016

My View

This TF will recommend to NASA/SMD several courses of action intended to improve the science return from NASA's extensive data stores and which will enable new discoveries. One of the goals is to seek areas to leverage NASA's science resources with on-going projects in the government and industry.

My View – II

- Unless specifically requested by NASA or the Science Committee, I **don't** believe this TF should address these data topics
  - Availability – "Search-ability"
  - Proprietary periods
  - Long-term archiving
  - And other frequent questions of NASA's data stores
- Lets break new ground and propose new and exciting ways to find new science in NASA's data.

Draft BDTF Work Plan

- Survey
- Nominate topics
- Choose 3 to 4 topics
- Produce products
  - Concise statement of problem
  - Research
  - Organize and develop positions
  - Form consensus
  - Draft and present results in the form of a white paper with accompanying slide presentation.

Overall Schedule

- Terms of the TF members expire on Dec 2017.
- Plan for ~4-5 more face-to-face meetings in advance of the Science Committee Meetings.
  - Develop statements of advice to be presented to the Science Committee
    - recommendations and findings, etc.
  - Finalize “product” papers for submission to Science Committee.
- Hold telecons as appropriate to discuss progress on research, report on meetings, etc.

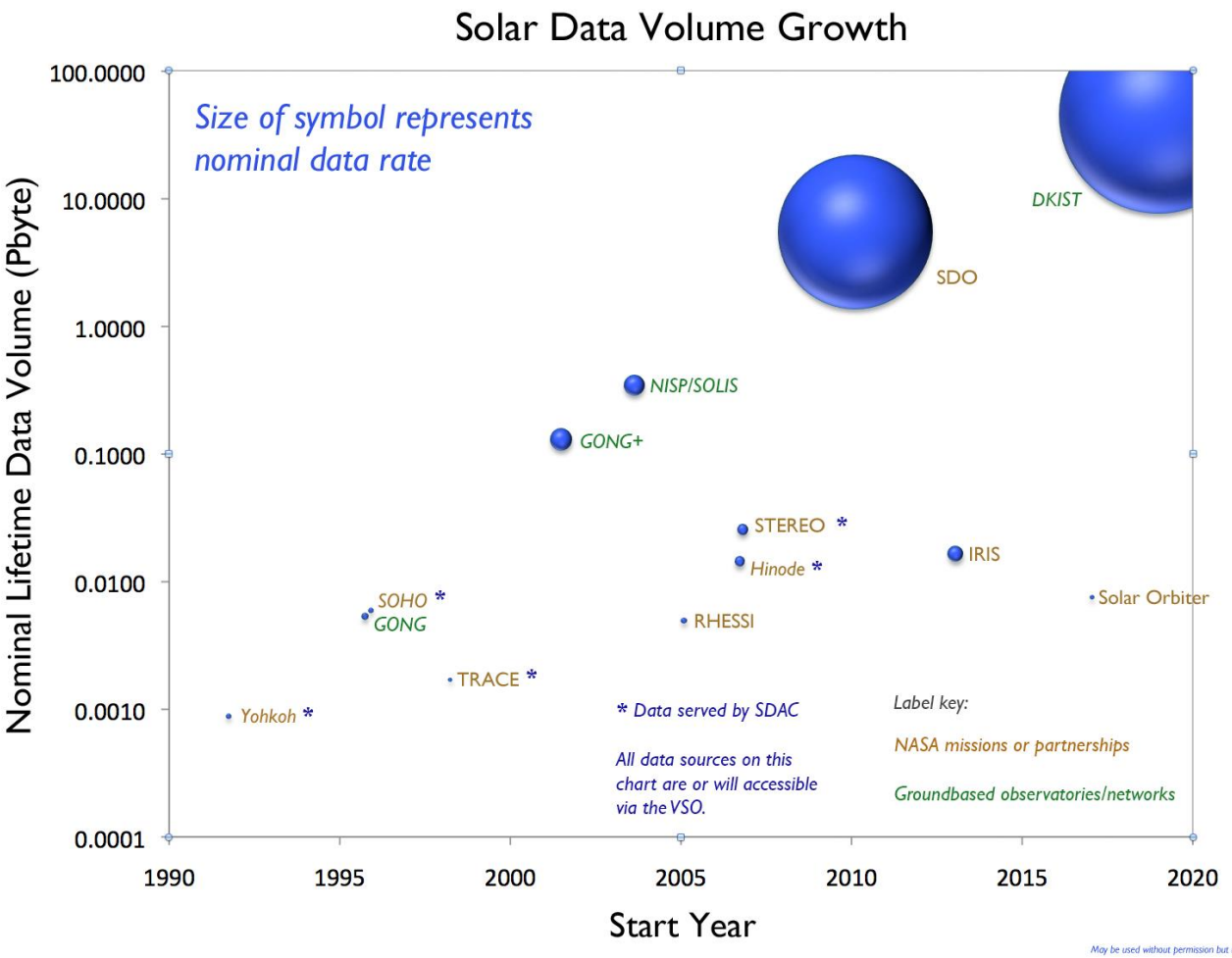
Possible Topic:
Long-term Planning of NASA's Big Data Capabilities

Solar Project Data Volume Growth

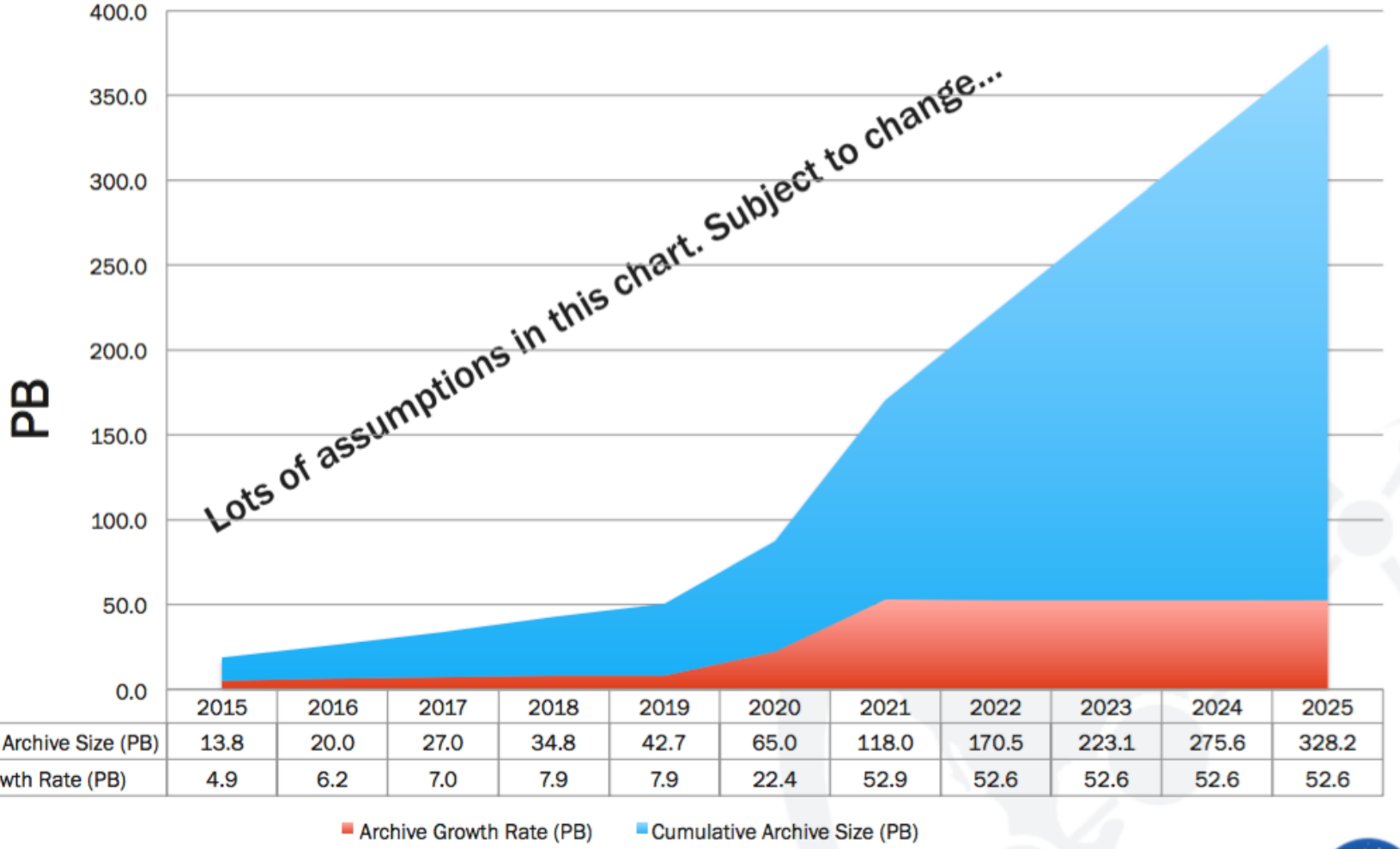
Both lifetime data volume and rate continue to grow

There is and will continue to be experience in the solar groundbased community as well as the NASA-supported community

Thus it makes sense to share experience and best practices between the two communities

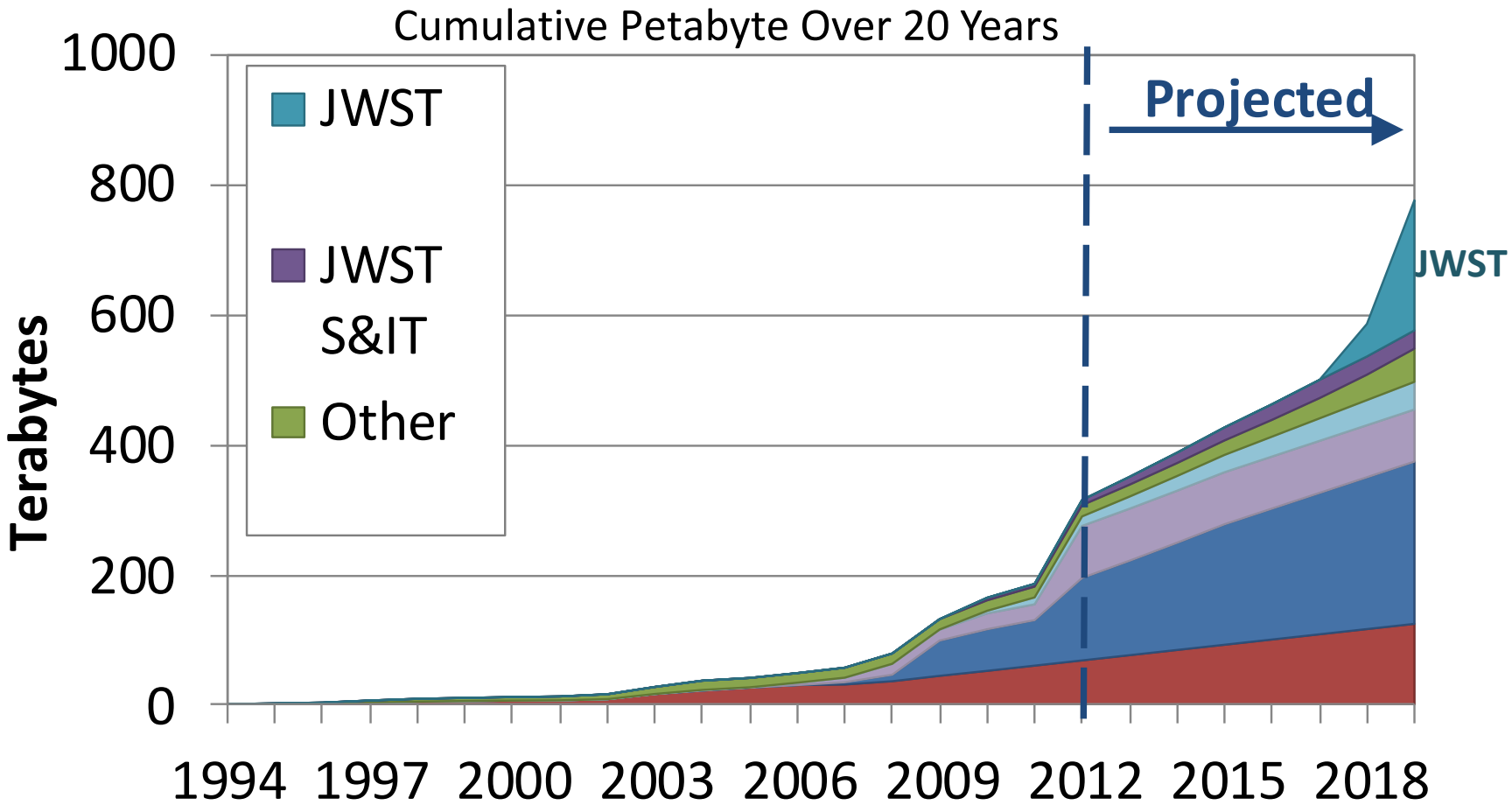


EOSDIS Archive Growth *Estimate* (Prime + Extended)

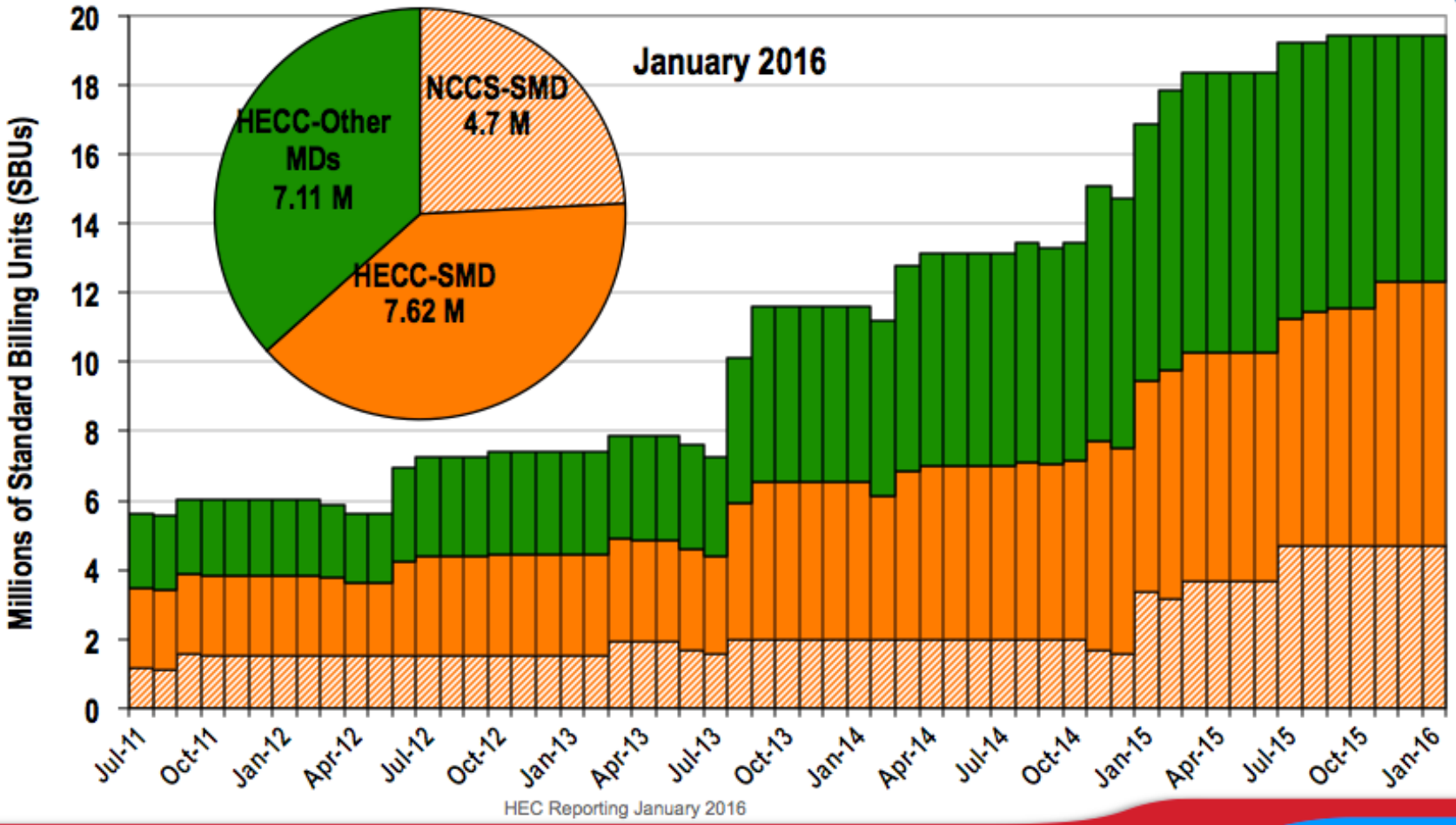


	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Cumulative Archive Size (PB)	13.8	20.0	27.0	34.8	42.7	65.0	118.0	170.5	223.1	275.6	328.2
Archive Growth Rate (PB)	4.9	6.2	7.0	7.9	7.9	22.4	52.9	52.6	52.6	52.6	52.6

Multi-Mission Data Archives at STSI Will Continue to Grow - Doubling by 2018



All Missions HEC Capacity Shares in SBUs



Initial Impressions

- SMD's data activities are projected to maintain or even accelerate their growth.
- Are there adequate planning activities focused on future needs and solutions for SMD's data infrastructure?
  - Infrastructure tends to have flat budgets that include upgrades/refreshes.
- Data project managers plan improvements based on Moore's law. But are we on a collision course?

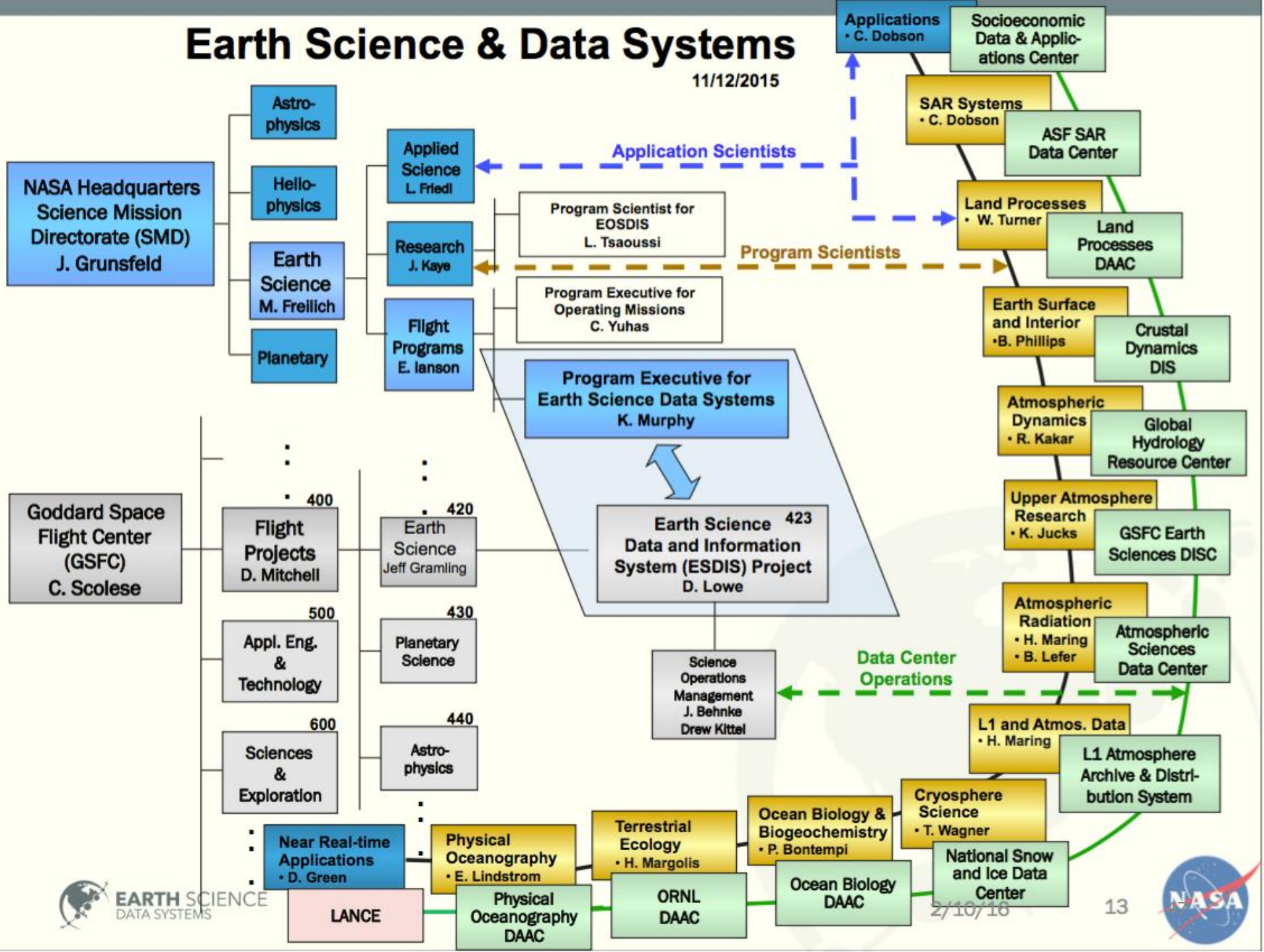
Be careful with “Interoperability”

- Raises many red flags.
- The domain of NASA’s science data stores can be characterized by many variables, many dimensions, ...
  - Inadequate specification, inadequate budget, …
- A general rule of thumb is our “community” cannot/will not accommodate a major top-down directive.

- Targeted bottoms-up projects for achieving/improving “localized” interoperability have a record of being successful.
  - Finite, well defined, affordable, and useful for specific, high-priority problems.

Earth Science & Data Systems

11/12/2015



SMD's "Big Data" budgets ($M-FY'10)

Topic	Earth Science	Heliophysics	Planetary Science	Astrophysics	TOTAL
Computational methods	20	7		10	37
Archival data science	5		15	70	90
Tools for science inference	13	7			20
Data archives	45	10	10	20	85
Computational capabilities	60				60
TOTAL	143	24	25	100	292

Table compiled by SMD in Dec 2011 using FY 2010 data.
Table does not include mission science operations activities which include level-0 and -1 data processing.

About 50% of this budget was competed!

Need I say more?