1\textsuperscript{st} 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

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

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 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

National-Scale 100Gbps Network Backbone

Source: Presentation to ESnet Policy Board

3/10/16 BDTF presentation to the Science Committee
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
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

---

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

---

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


• 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).
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
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

<table>
<thead>
<tr>
<th>Node</th>
<th>Data Volume (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atmospheres</td>
<td>3.0</td>
</tr>
<tr>
<td>Cartography and Imaging</td>
<td>825.0</td>
</tr>
<tr>
<td>Geosciences</td>
<td>165.0</td>
</tr>
<tr>
<td>NAIF</td>
<td>0.5</td>
</tr>
<tr>
<td>Planetary Plasma Interactions</td>
<td>7.5</td>
</tr>
<tr>
<td>Ring-Moon Systems</td>
<td>1.6</td>
</tr>
<tr>
<td>Small Bodies</td>
<td>3.1</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>1,000.7</strong></td>
</tr>
</tbody>
</table>

- 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.
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
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
Strategic Support for NASA Programs

- External tank redesign
- Sonic boom optimization
- Launch abort system
- Hurricane prediction
- Merging black holes
- Orion/MPCV reentry
- SLS vehicle designs
- 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

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

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

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 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)

Lots of assumptions in this chart. Subject to change...

<table>
<thead>
<tr>
<th>Year</th>
<th>Cumulative Archive Size (PB)</th>
<th>Archive Growth Rate (PB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>13.8</td>
<td>4.9</td>
</tr>
<tr>
<td>2016</td>
<td>20.0</td>
<td>6.2</td>
</tr>
<tr>
<td>2017</td>
<td>27.0</td>
<td>7.0</td>
</tr>
<tr>
<td>2018</td>
<td>34.8</td>
<td>7.9</td>
</tr>
<tr>
<td>2019</td>
<td>42.7</td>
<td>7.9</td>
</tr>
<tr>
<td>2020</td>
<td>65.0</td>
<td>22.4</td>
</tr>
<tr>
<td>2021</td>
<td>118.0</td>
<td>52.9</td>
</tr>
<tr>
<td>2022</td>
<td>170.5</td>
<td>52.6</td>
</tr>
<tr>
<td>2023</td>
<td>223.1</td>
<td>52.6</td>
</tr>
<tr>
<td>2024</td>
<td>275.6</td>
<td>52.6</td>
</tr>
<tr>
<td>2025</td>
<td>328.2</td>
<td>52.6</td>
</tr>
</tbody>
</table>
Multi-Mission Data Archives at STSI
Will Continue to Grow - Doubling by 2018

Projected Cumulative Petabyte Over 20 Years

- JWST
- JWST
- S&IT
- Other

Terabytes


3/10/16 BDTF presentation to the Science Committee

MAST Overview for NAC ITIC March 6, 2012
All Missions HEC Capacity Shares in SBUs

January 2016

- HECC-SMD: 7.62 M
- NCCS-SMD: 14.33 M
- HECC-Other MDs: 7.11 M

Millions of Standard Billing Units (SBUs)

HEC Reporting January 2016
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.
### SMD’s “Big Data” budgets ($M-FY’10)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Earth Science</th>
<th>Heliophysics</th>
<th>Planetary Science</th>
<th>Astrophysics</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computational methods</td>
<td>20</td>
<td>7</td>
<td></td>
<td>10</td>
<td>37</td>
</tr>
<tr>
<td>Archival data science</td>
<td>5</td>
<td>15</td>
<td>70</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>Tools for science inference</td>
<td>13</td>
<td>7</td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Data archives</td>
<td>45</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>85</td>
</tr>
<tr>
<td>Computational capabilities</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>143</strong></td>
<td><strong>24</strong></td>
<td><strong>25</strong></td>
<td><strong>100</strong></td>
<td><strong>292</strong></td>
</tr>
</tbody>
</table>

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!*
CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.

IT COMES FROM EVERYWHERE. IT KNOWS ALL.

ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS “BIG DATA.”

BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.

IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.

BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.

LET US PAY.

IS IT TOO LATE TO SIDE WITH EVIL? SHHHH! IT HeARS YOU.