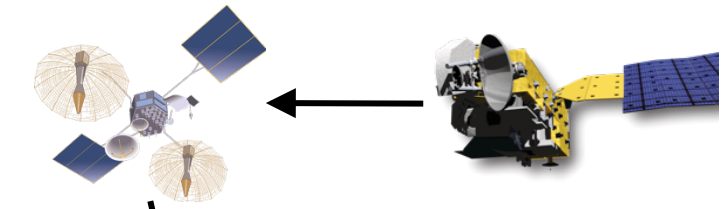
Big Data in the Earth Observing System Data and Information System

Chris Lynnes (System Architect)
Katie Baynes (System Architect)
Mark McInerney (Deputy Project Manager)
NASA/GSFC, ESDIS Project

Earth Observing System Data and Information System (EOSDIS)



data downlink

EOSDIS

capture and clean

process

archive

subset

distribute

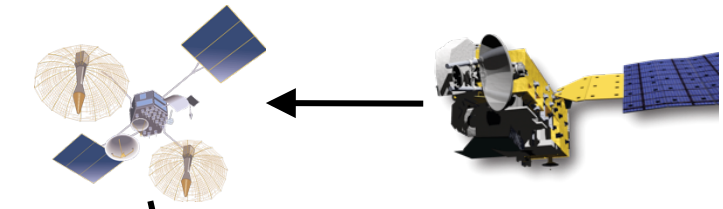
Research

Applications

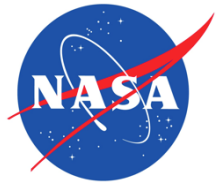
Education

Users

Earth Observing System Data and Information System (EOSDIS)



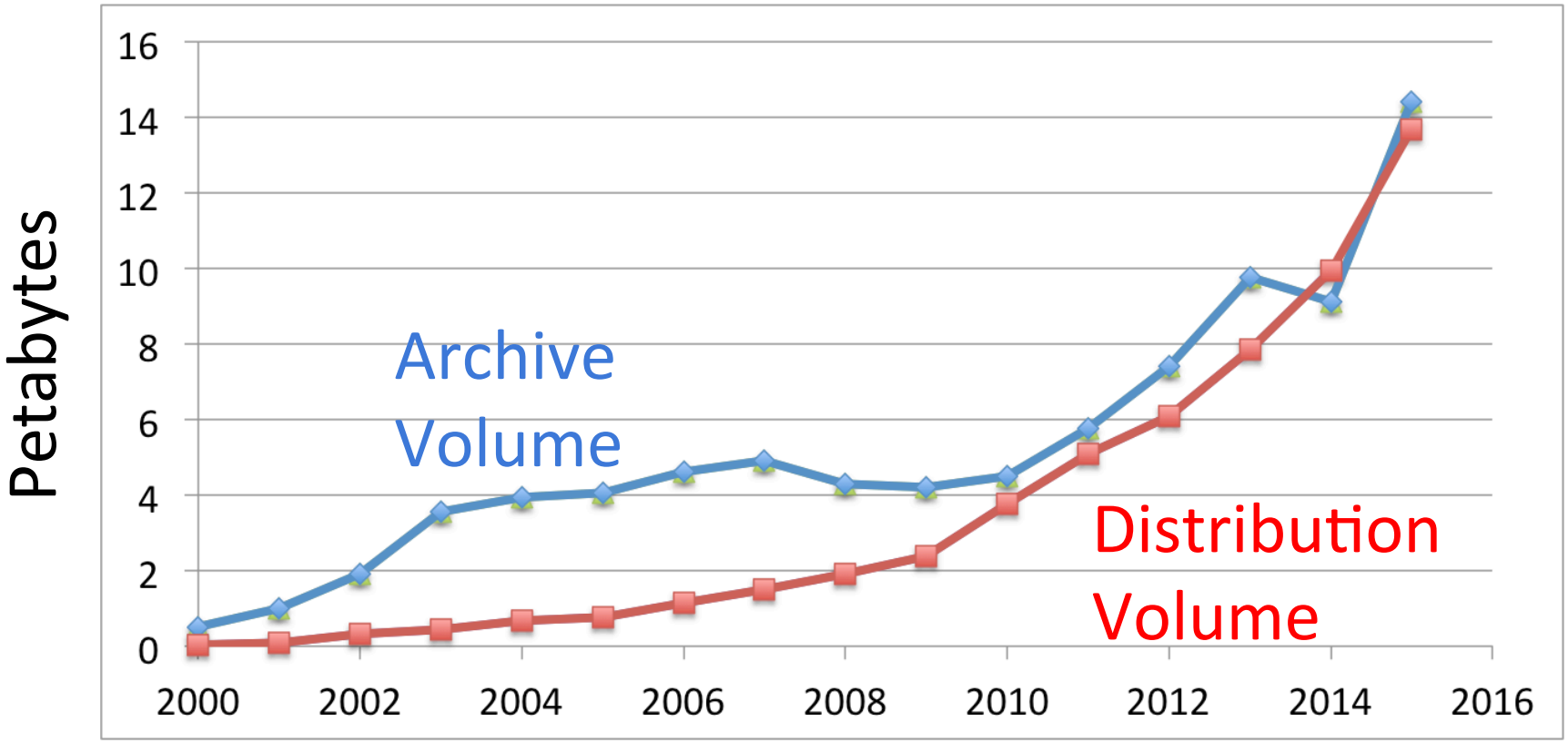
Take Home Message Preview



1. Cloud prototypes are underway to tackle the *Volume* challenge of Big Data…
2. ….But advances in computer hardware or cloud won't help (much) with *Variety*
3. Interoperability standards, conventions, and community engagement are the key to addressing *Variety*

V is for...

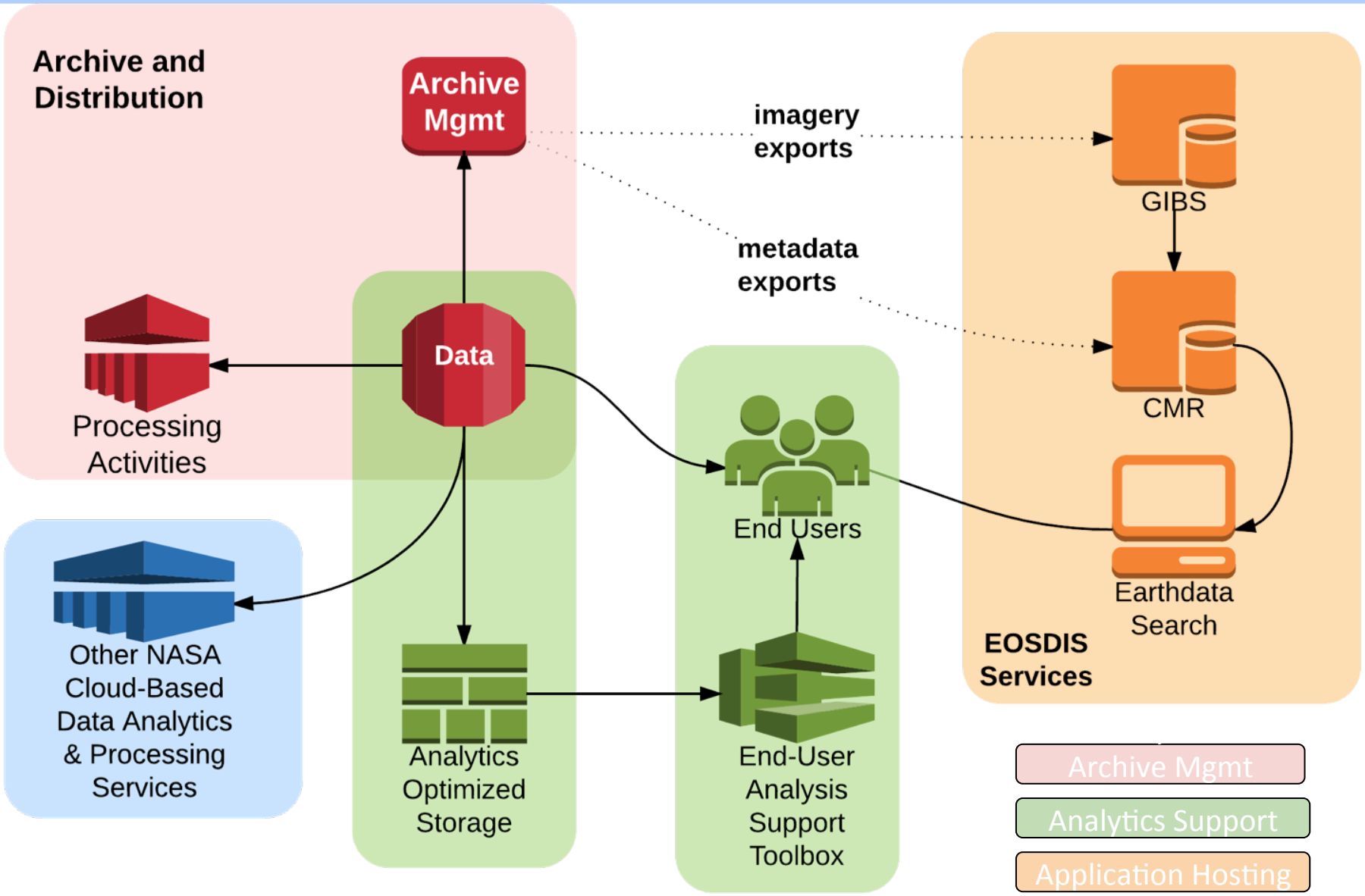
...Volume



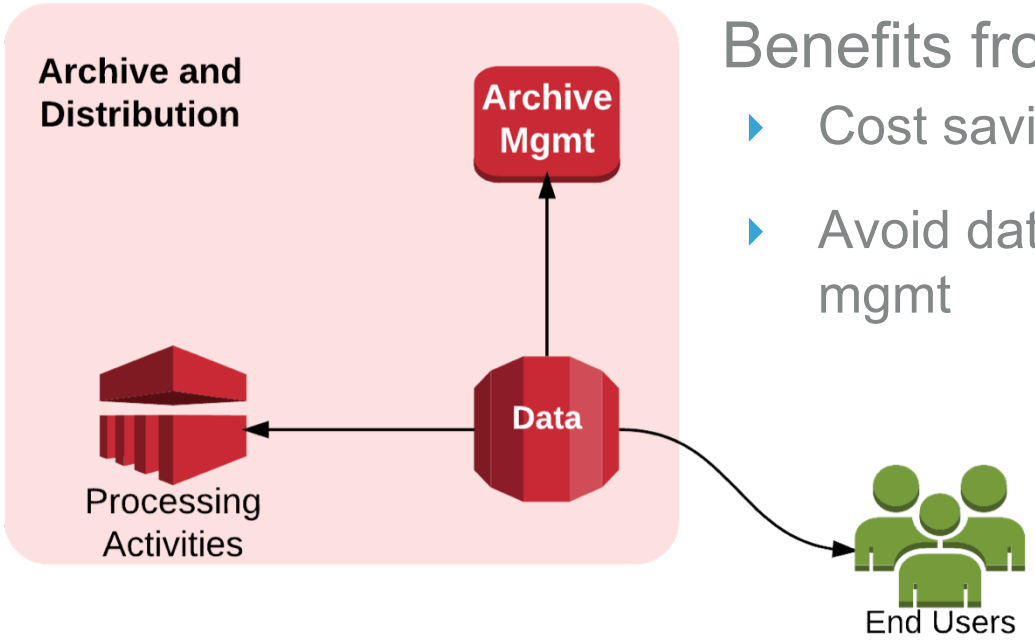
Big Data Indicators

EOSDIS FY2015 Metrics	
Unique Data Products	9,462
Distinct Users of EOSDIS Data and Services	2.6 M
Average Daily Archive Growth	16 TB/day
Total Archive Volume (as of Sept. 30, 2015)	14.6 PB
End User Distribution Products	1.42 B
End User Average Daily Distribution Volume	32.1 TB/day

EOSDIS Cloud Prototypes



Archive Cloud Prototypes

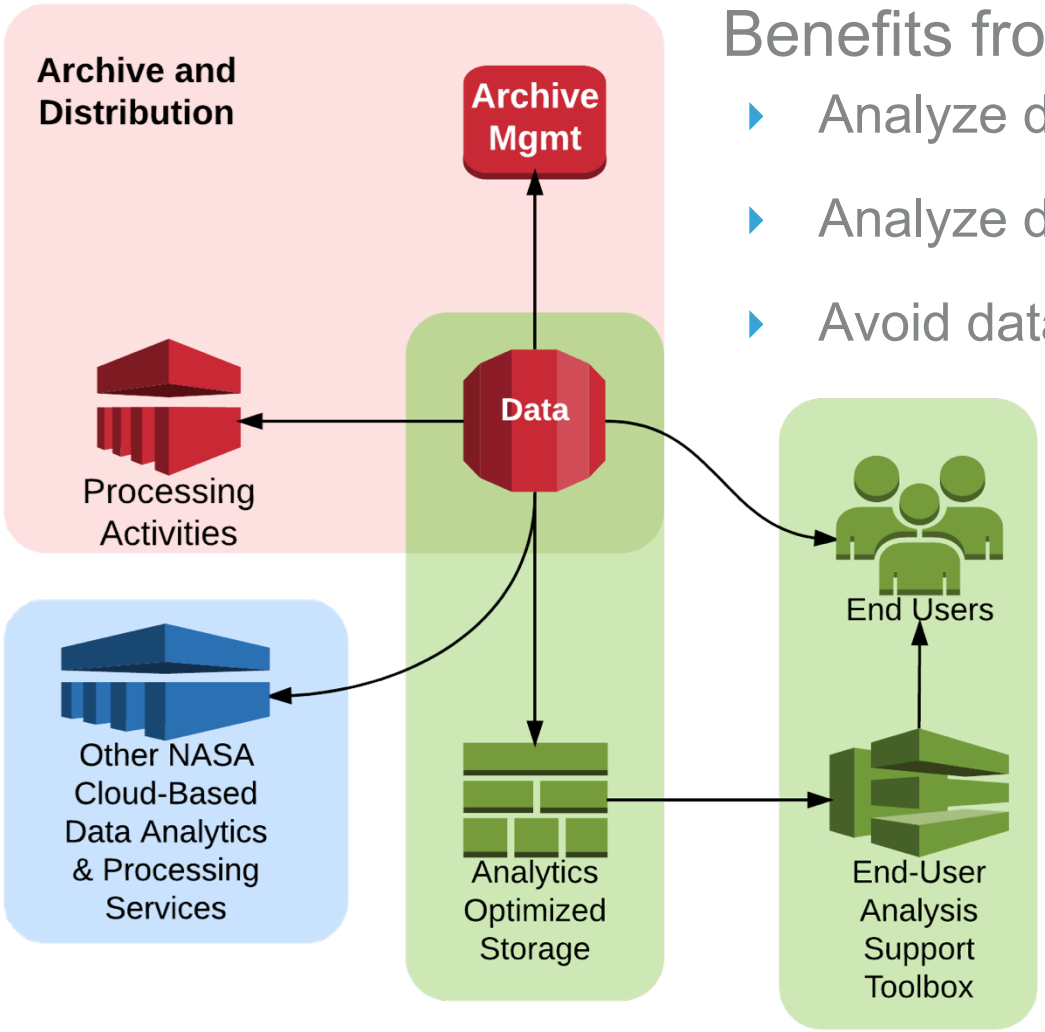


Benefits from Archive in the Cloud

- Cost savings for storage of Big Data?
- Avoid data downloading and local data mgmt

- Alaska Satellite Facility Web Object Storage prototype
  - Distribute Sentinel radar data from Amazon storage
- Global Imagery Browse Service in the Cloud
- Ingest and Archive management prototype

Cloud Analytics Prototypes



Benefits from Cloud Analytics

- Analyze data at scale
- Analyze datasets together easily
- Avoid data downloading and local mgmt

Analysis support toolbox to attract users to cloud analytics

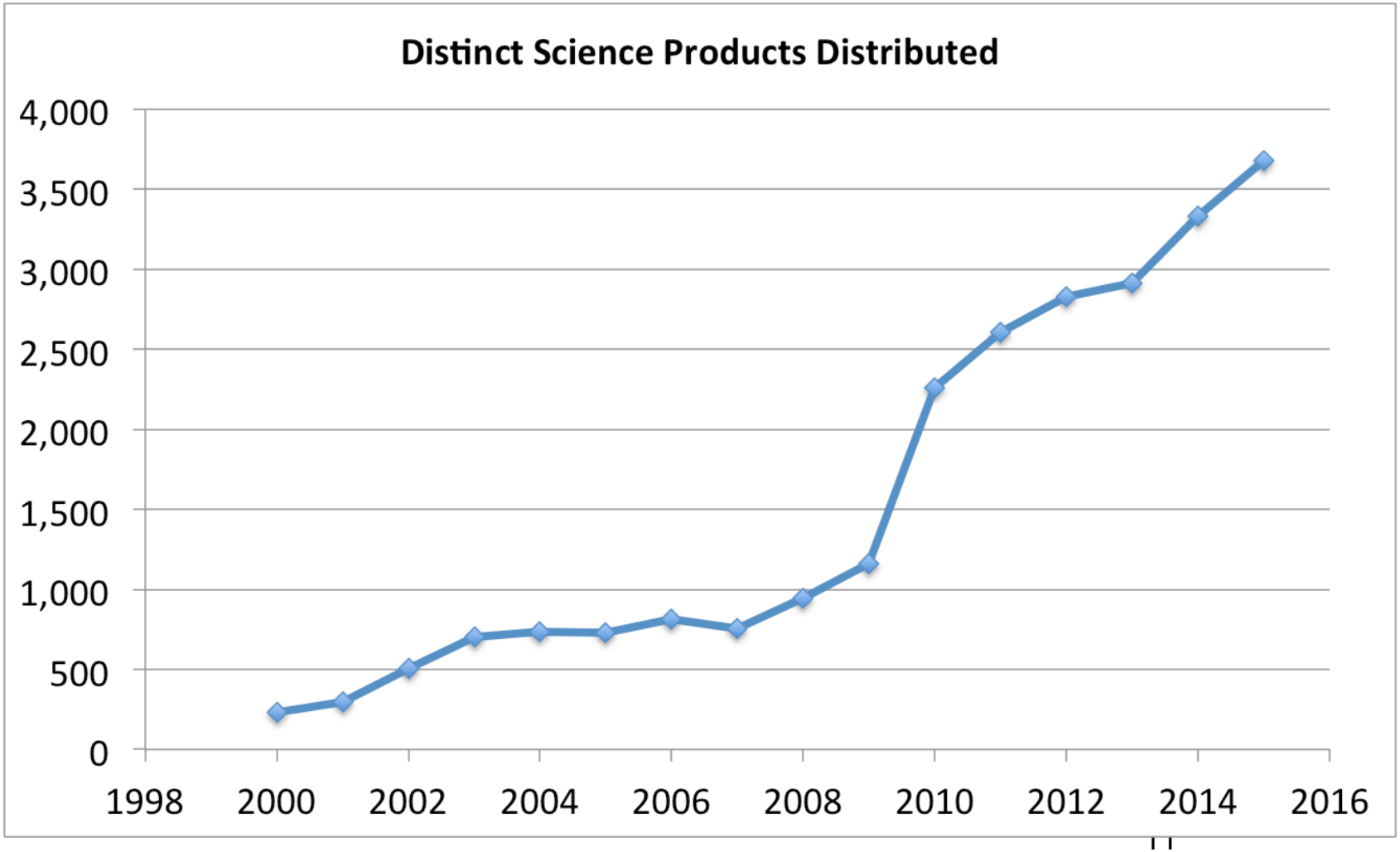
- Community open source tools
- DAAC-developed tools
- Cloud analytics examples and recipes
- Initial cross-DAAC proof of concept in progress based on Python + Jupyter Hub

Terra Incognita

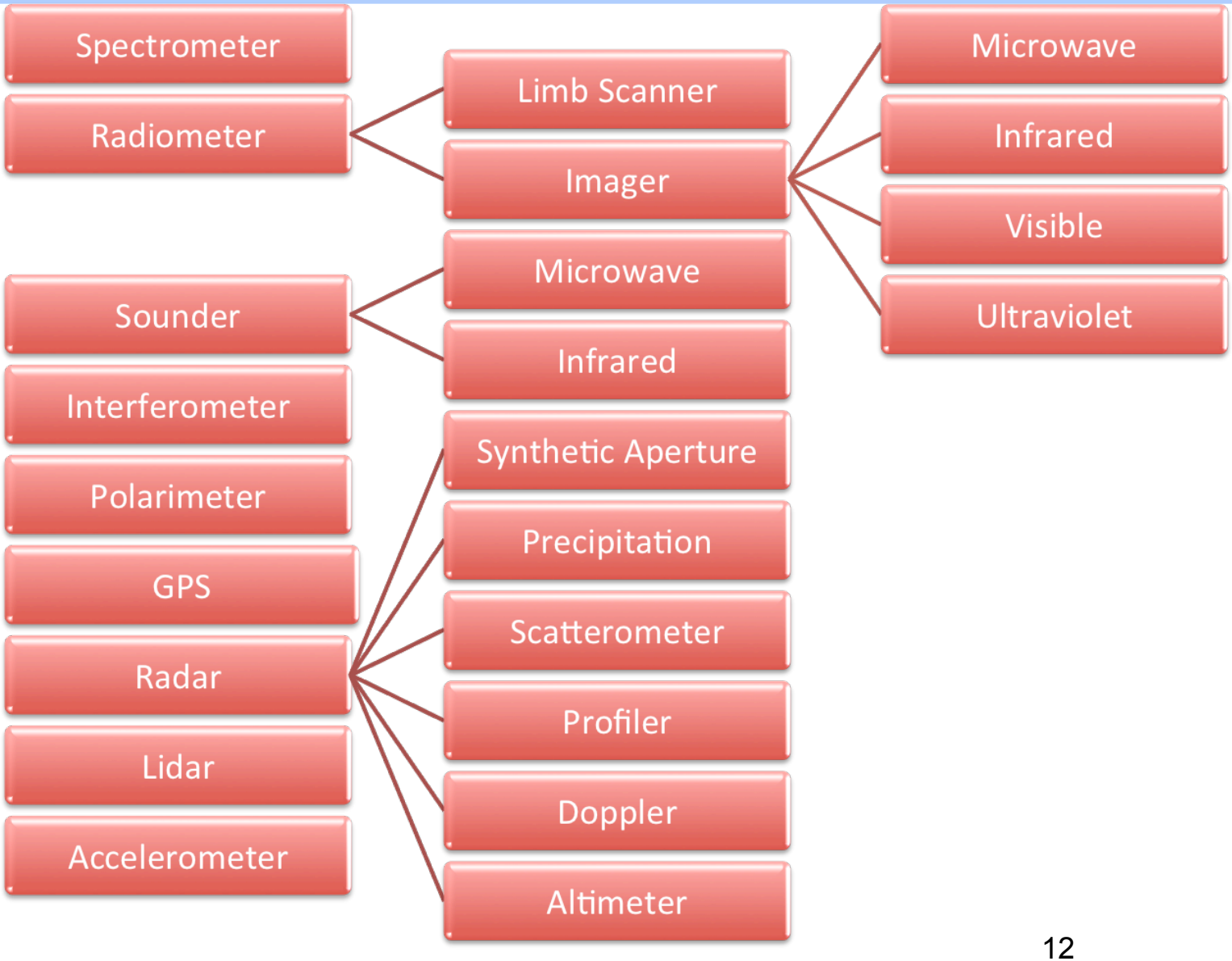
1. Vendor Lock-in
2. Future storage costs
3. Uncapped egress costs
4. Security Restrictions
5. Network trust

V is for...

...Variety

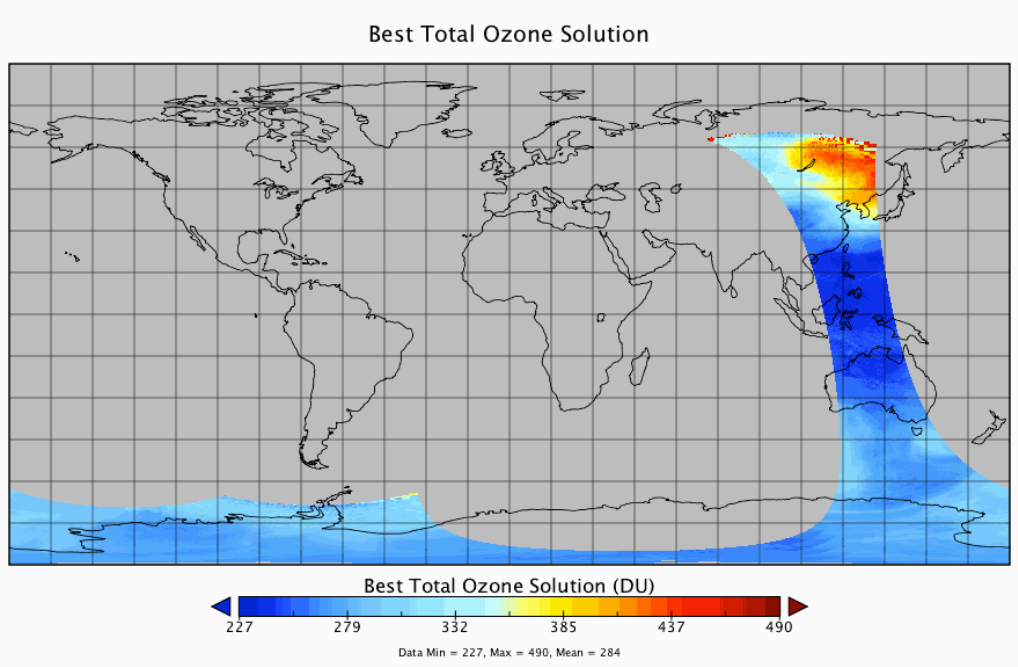


Instrument Variety

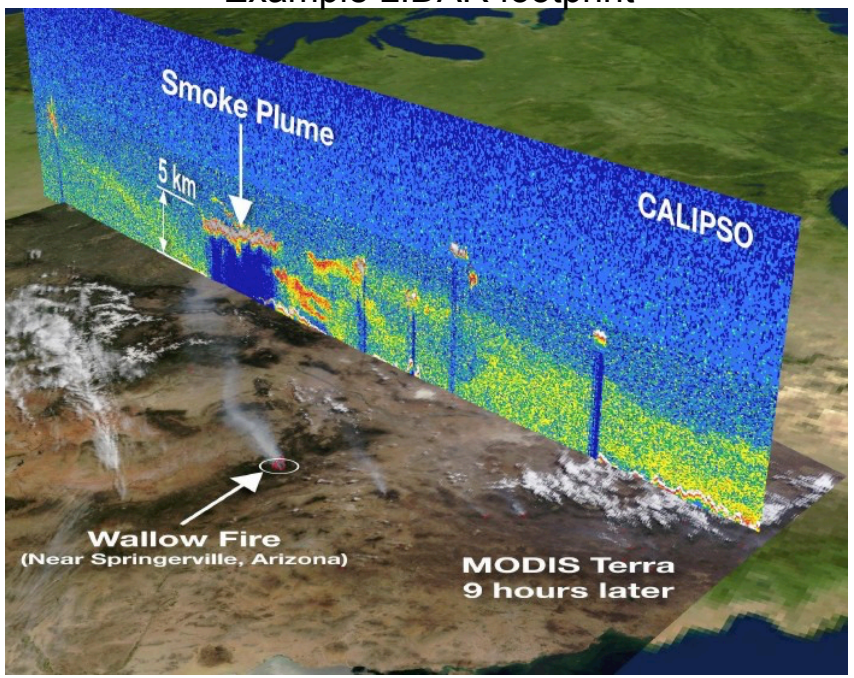


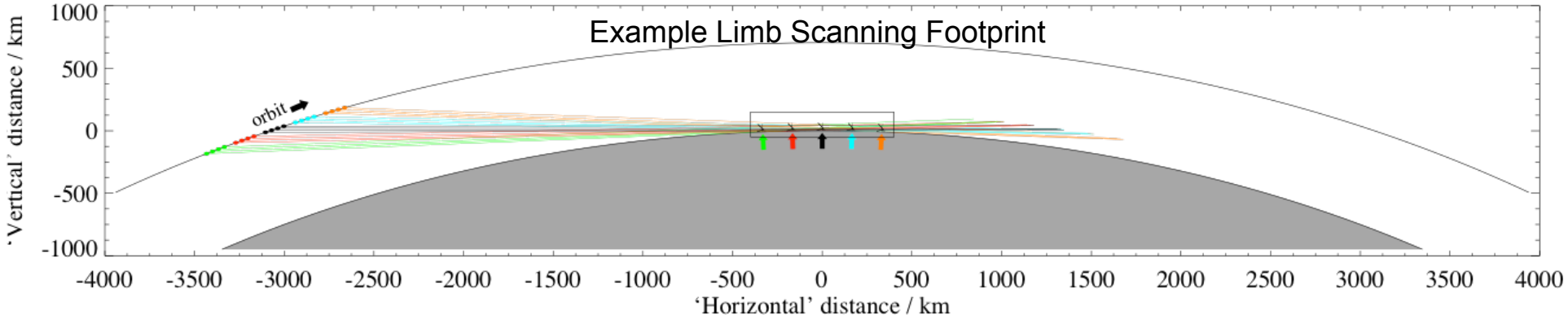
Satellite Instrument "Footprints"

Example Imaging Footprint



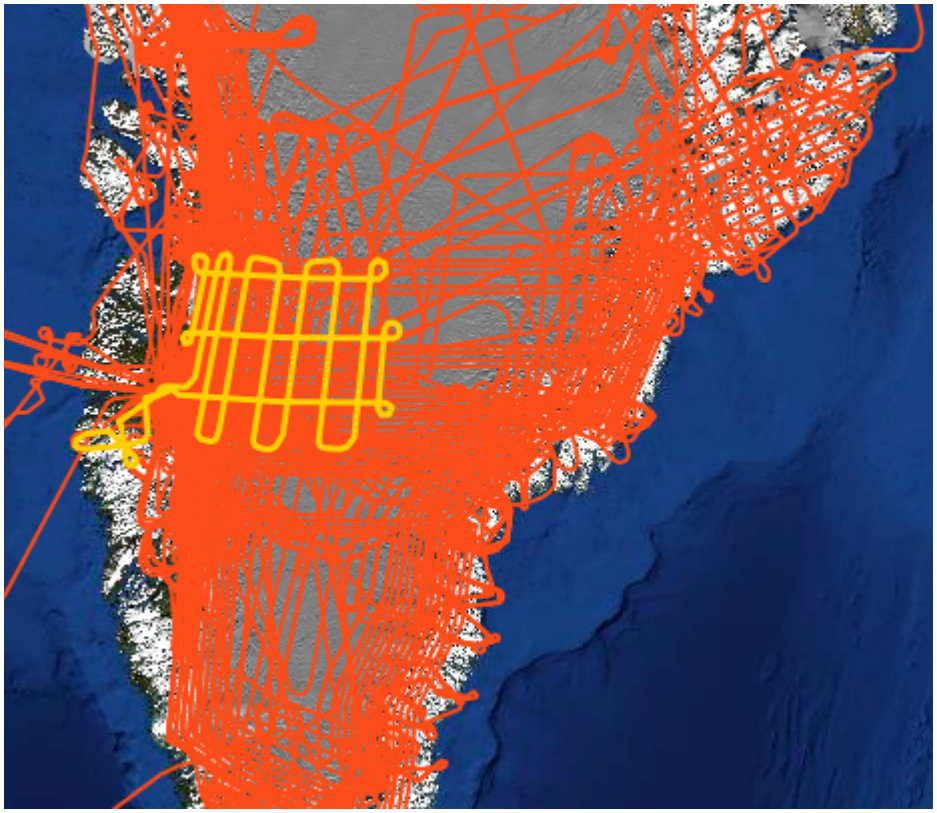
Example LIDAR footprint



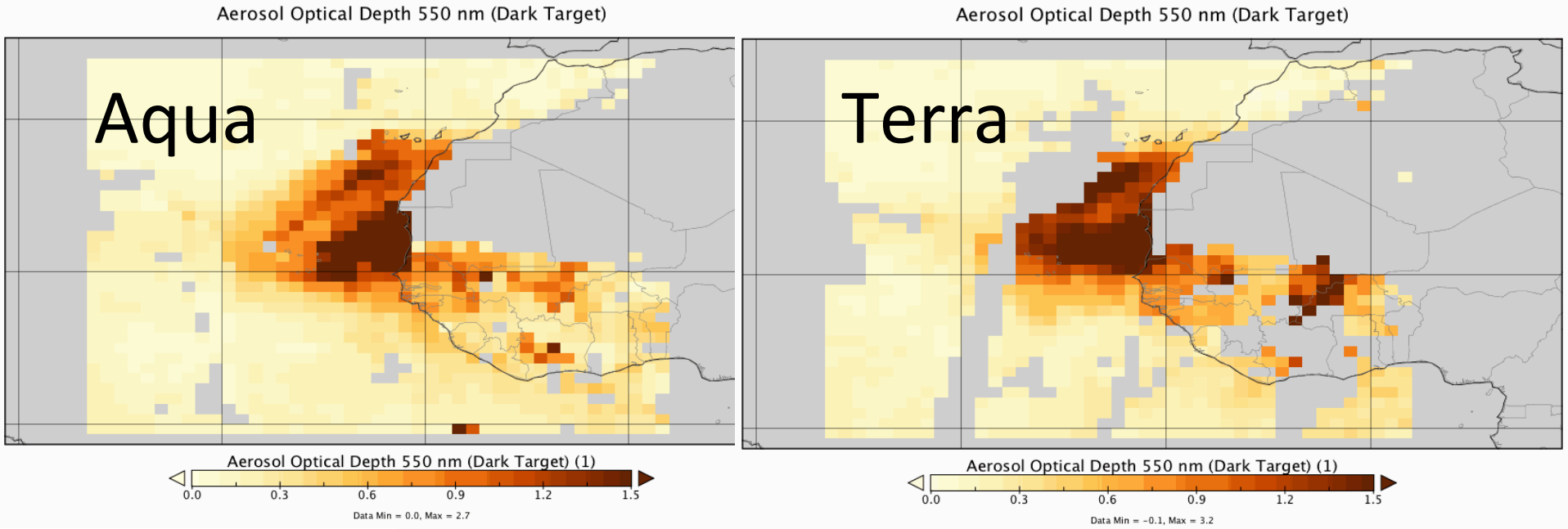


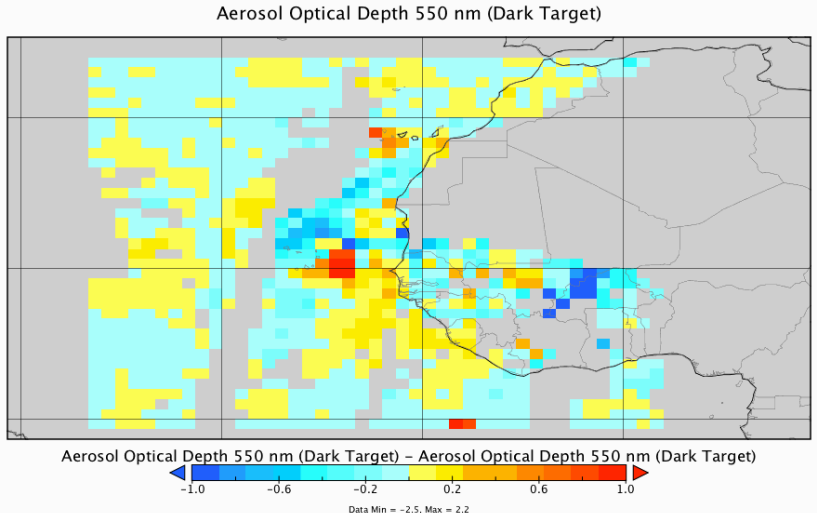
Microwave Limb Scanner (from Algorithm Theoretical Basis Document, Livesey and Wu, 1999)

Aircraft and In Situ



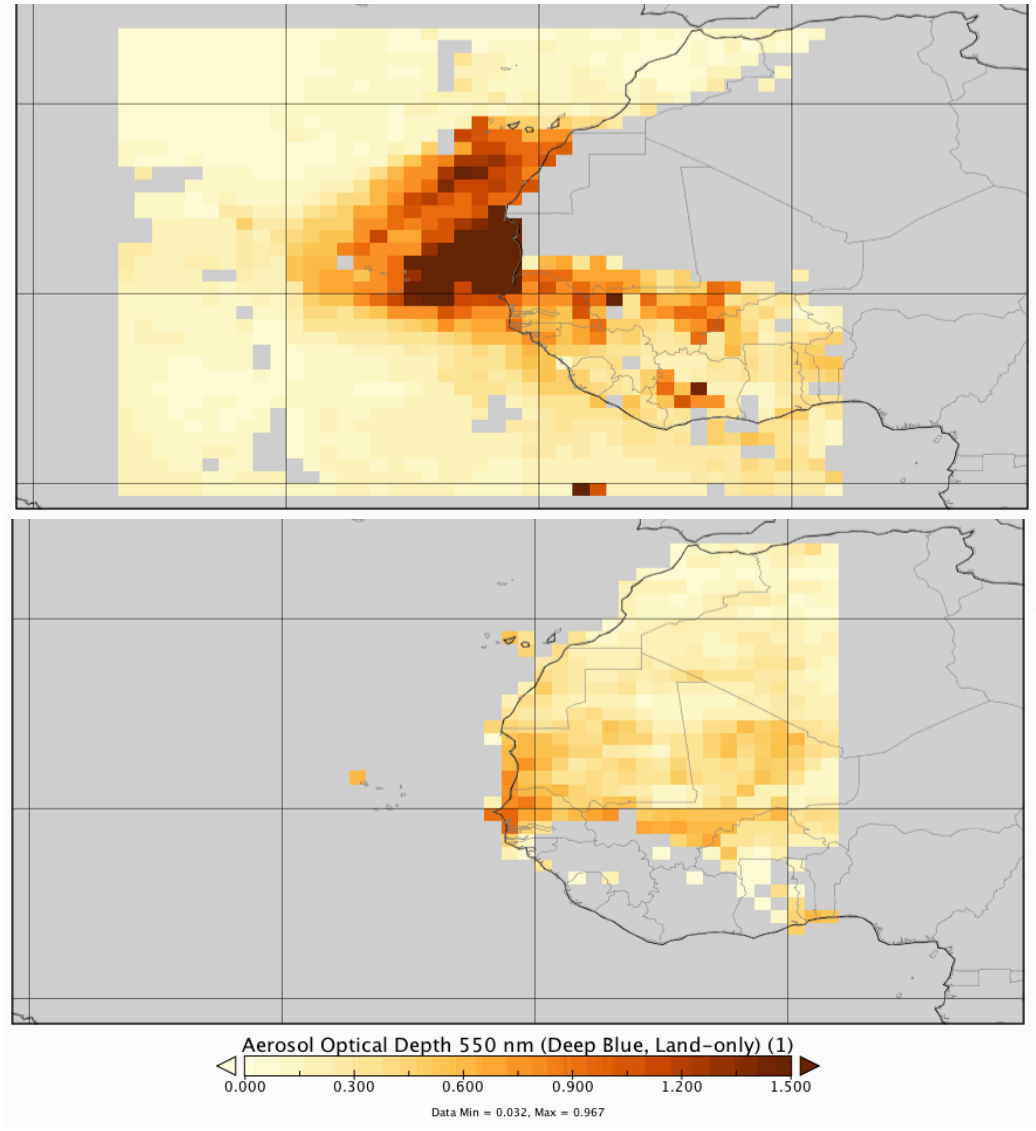
Same Instrument, Different Satellite





Aqua - Terra

Same Instrument+Satellite, Different Algorithm



MODIS on Aqua
Aerosol Optical Depth

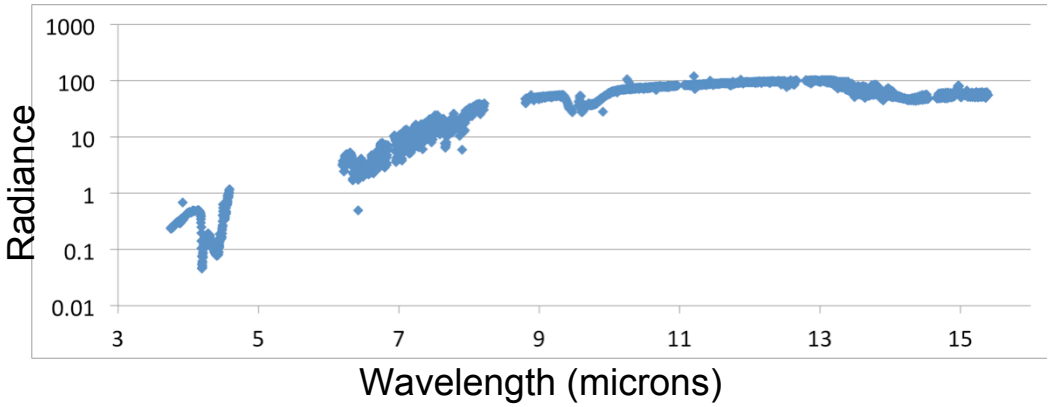
Dark Target Algorithm

Deep Blue Algorithm

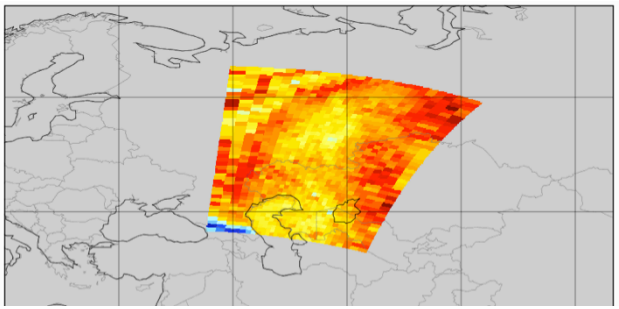
Processing Levels

AIRS data for 2011-08-11

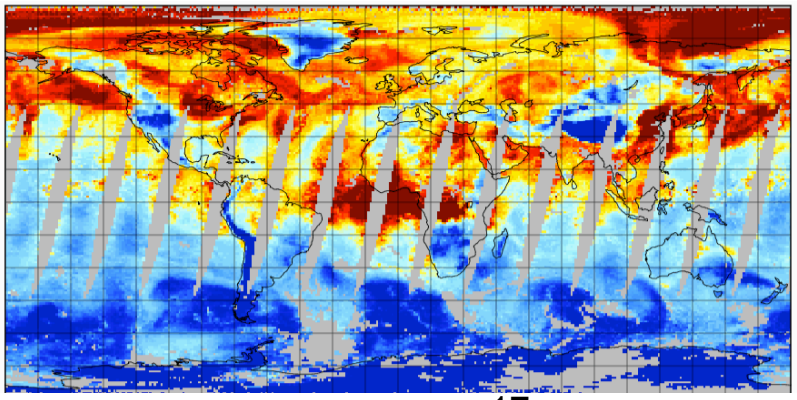
Level 1B
Calibrated radiance at a pixel



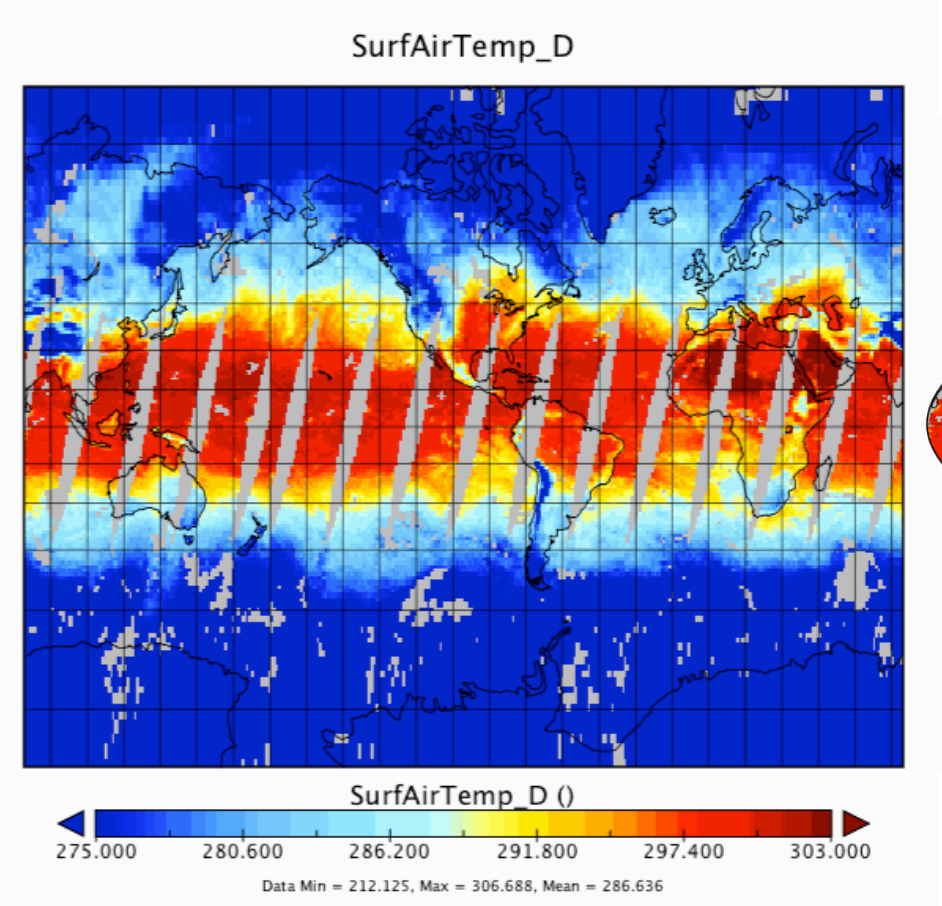
Level 2
Carbon monoxide for one scene

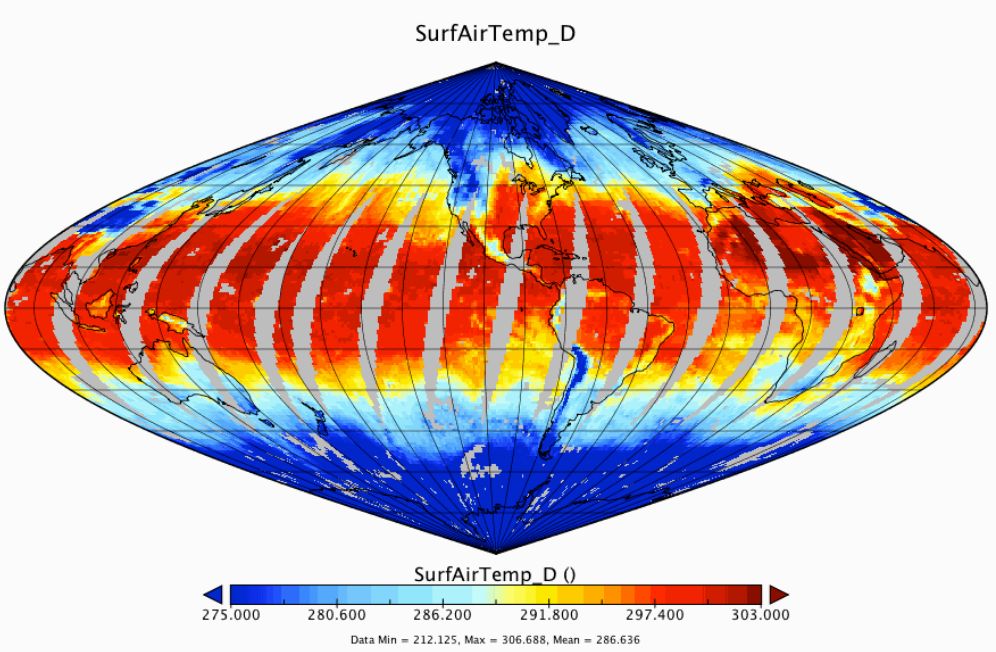


Level 3
Global carbon monoxide for one night

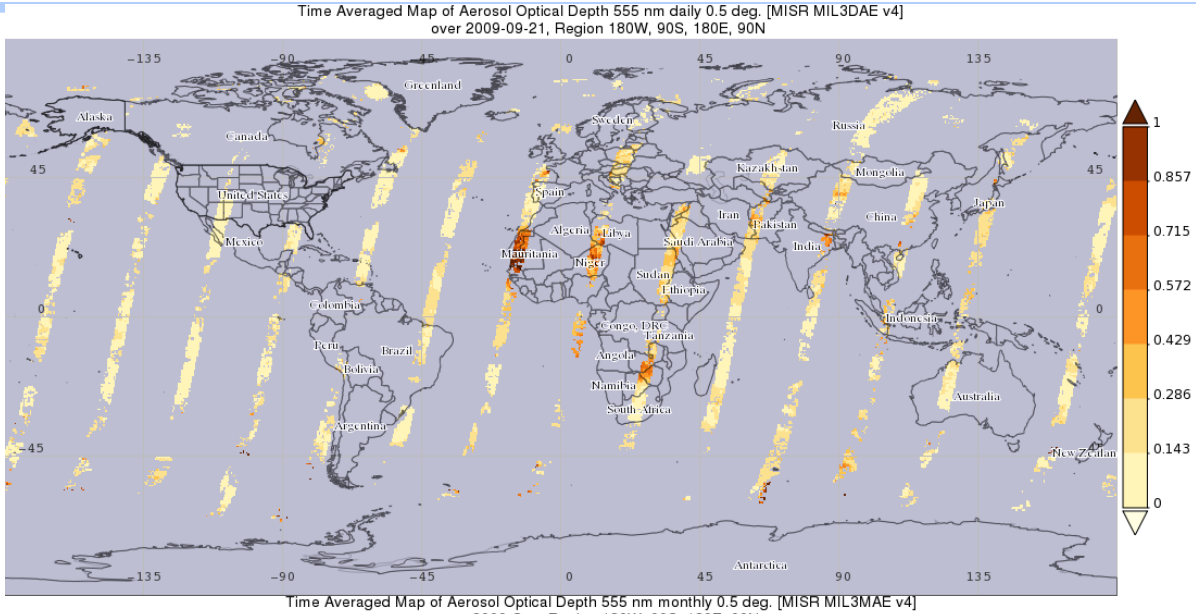


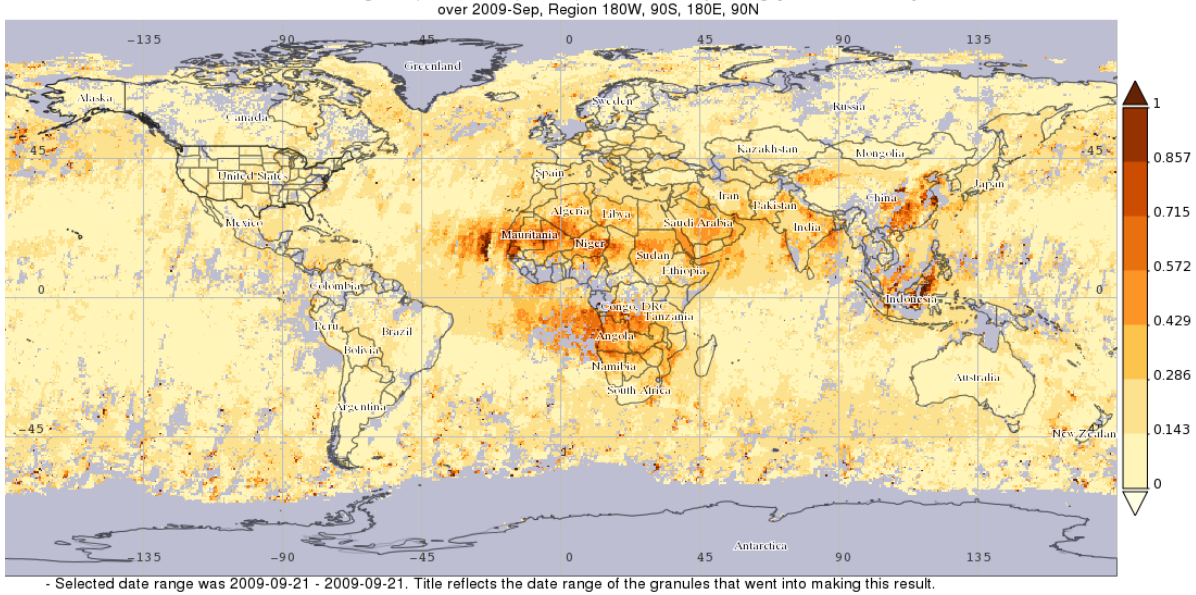
Projections





Time Aggregation





- Selected date range was 2009-09-21 - 2009-09-21. Title reflects the date range of the granules that went into making this result.

Aerosol Optical Depth at 555 nm from Multi-angle Imaging Spectro-Radiometer

Daily

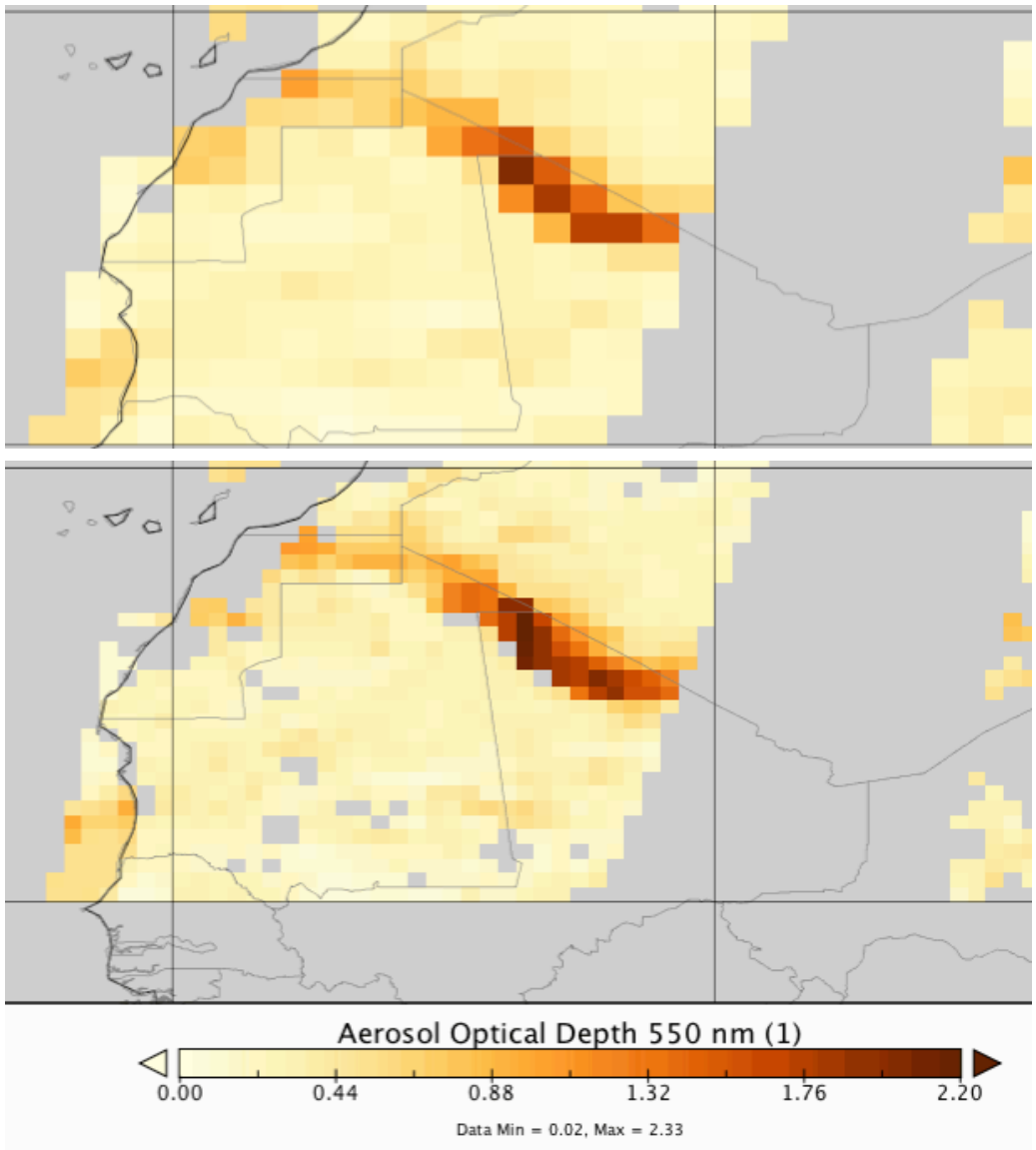
Monthly

Spatial Aggregation

SeaWiFS Deep-Blue Aerosol Optical Depth 2006-10-06

1.0 Degree Resolution

0.5 Degree Resolution



Data Formats

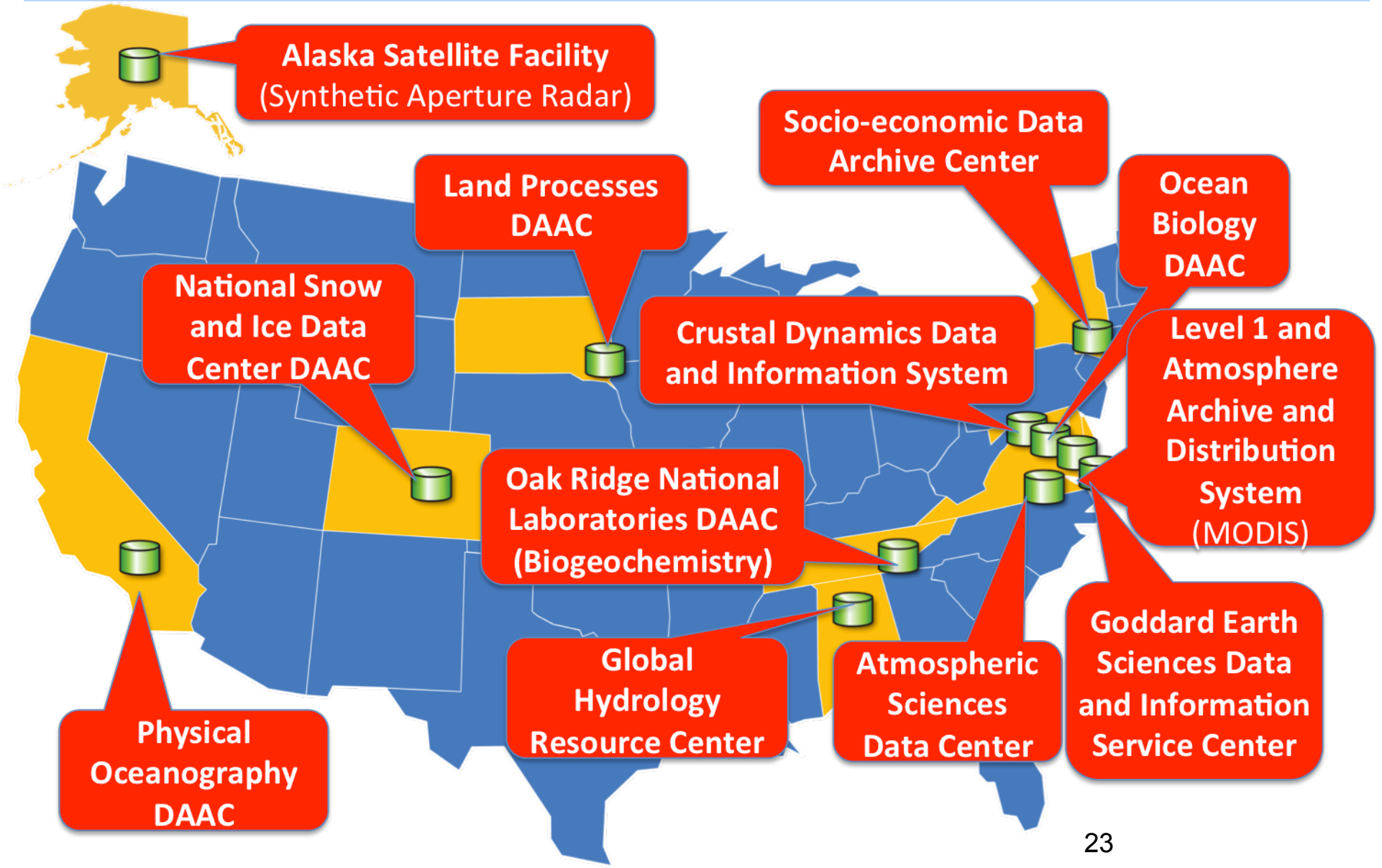
- Self-Describing API-Based
  - Hierarchical Data Format (HDF)
  - network Common Data Form (netCDF)
- Additional conventions
  - HDF-EOS
  - Climate-Forecast coordinates
- Other Standards
  - Gridded Binary (GRIB)
  - ICARTT (Airborne)
- Binary
- ASCII

Solutions to the Variety Problem

1. Interoperable discipline-focused DAACs
2. Common Metadata Repository
3. OPeNDAP* data services
4. Community engagement

*Open-source Project for a Network Data Access Protocol

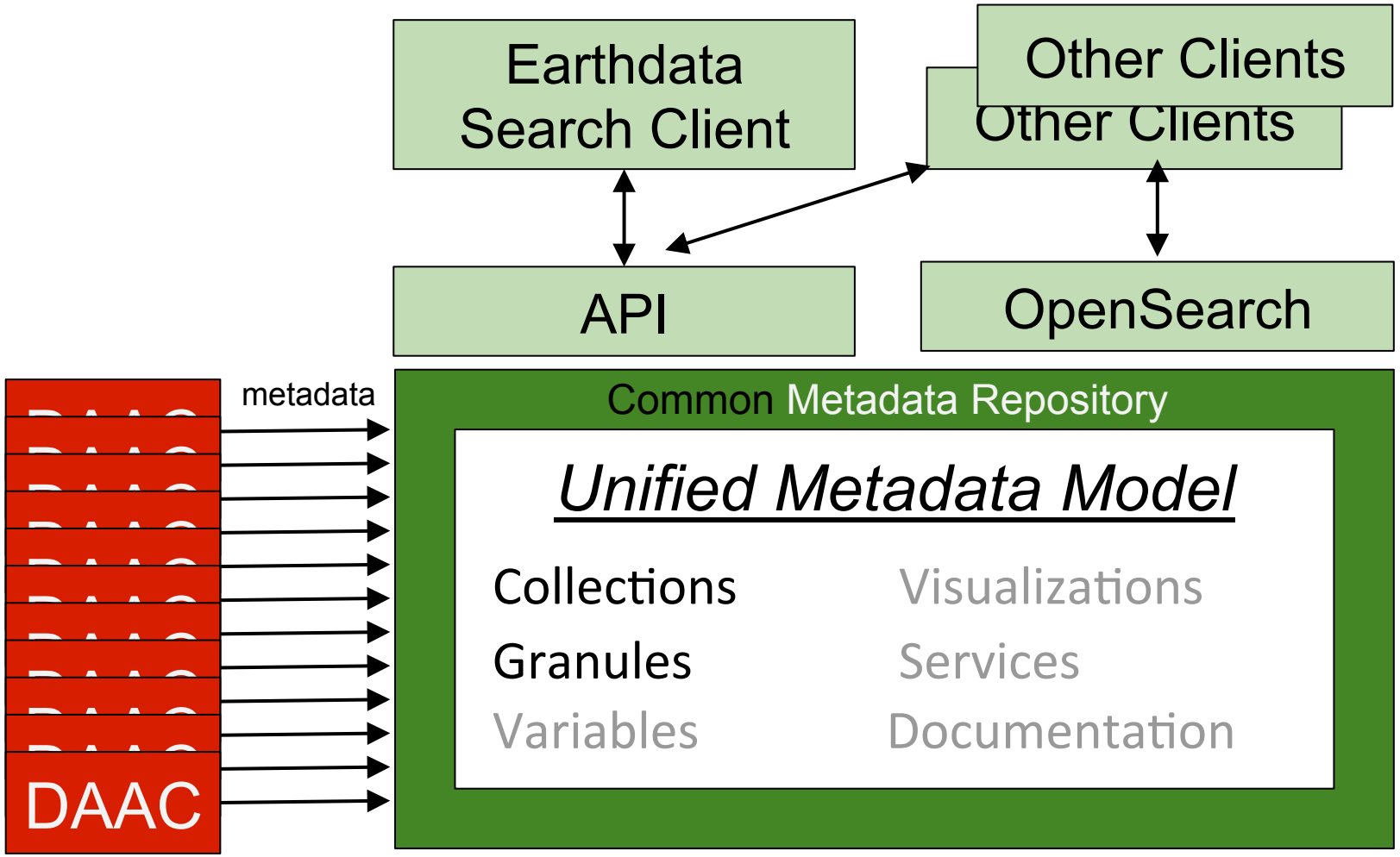
Discipline-Focused Distributed Active Archive Centers (DAACs)



Different DAACs have different "-Spheres of Influence"

DAAC	Atmo	Hydro	Cryo	Litho	Bio	Anthropo	Sub-Specialty
Alaska Satellite Facility			✓	✓			SAR
Atm. Sciences Data Center	✓						
Crustal Dynamics Data Info Sys				✓			Space geodesy
Global Hydrology Resource Ctr		✓					Weather events
Goddard Earth Sciences DISC	✓	✓					
Land Processes DAAC					✓	✓	
L1 and Atm Archive & Dist Sys	✓						MODIS, VIIRS
Nat. Snow Ice Data Ctr DAAC			✓				
Oak Ridge Nat Lab DAAC					✓		Field experiments
Ocean Biology DAAC		✓			✓		
Physical Oceanography DAAC		✓					
Socioeconomic Data Arch Ctr						✓	

The Common Metadata Repository presents a consistent catalog for discovery of data from multiple DAACs

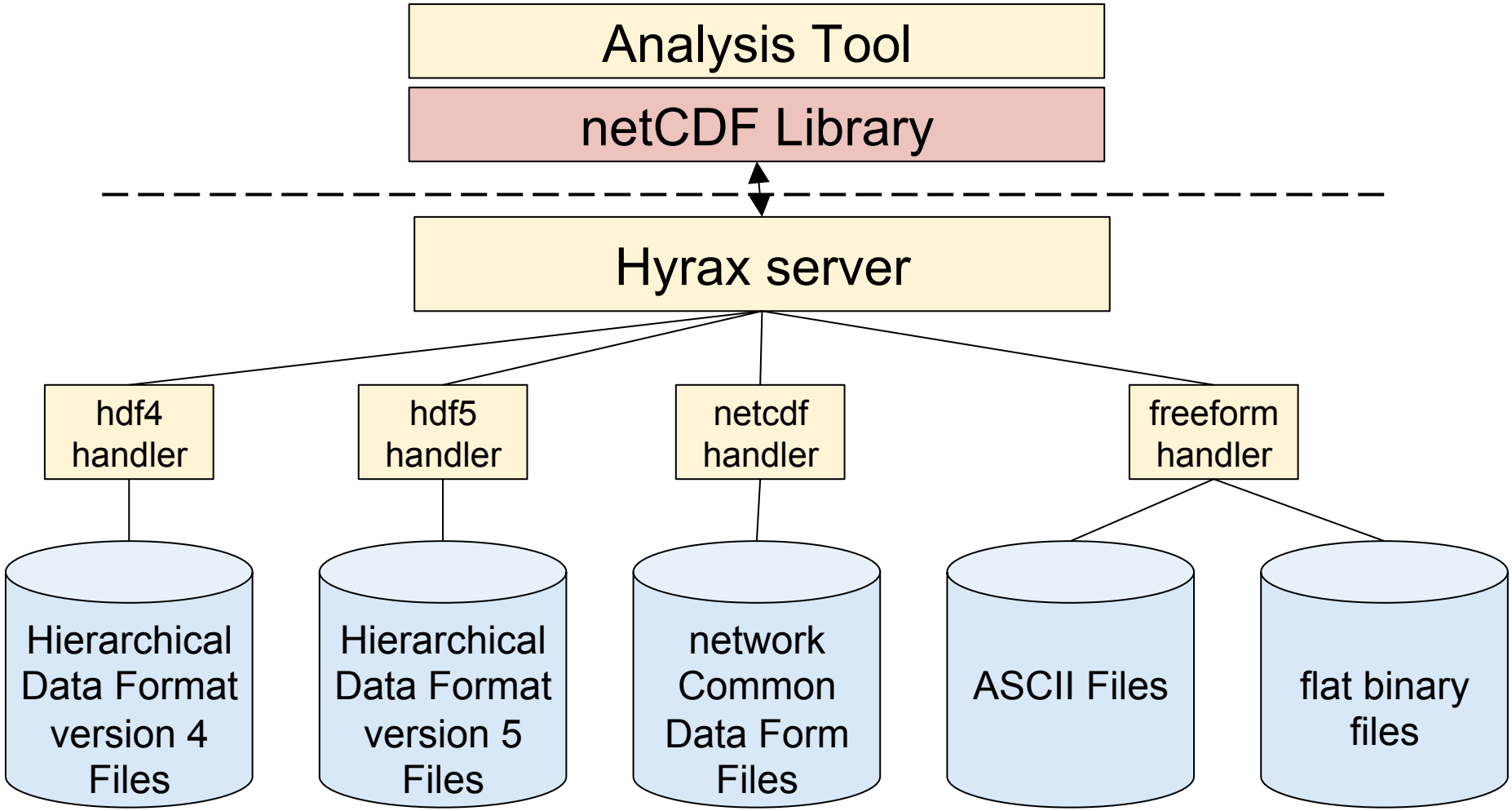


One Metadata System to rule them all,
One Metadata System to find them,
One Metadata System to bring them all
And in cyberspace bind them

OPeNDAP

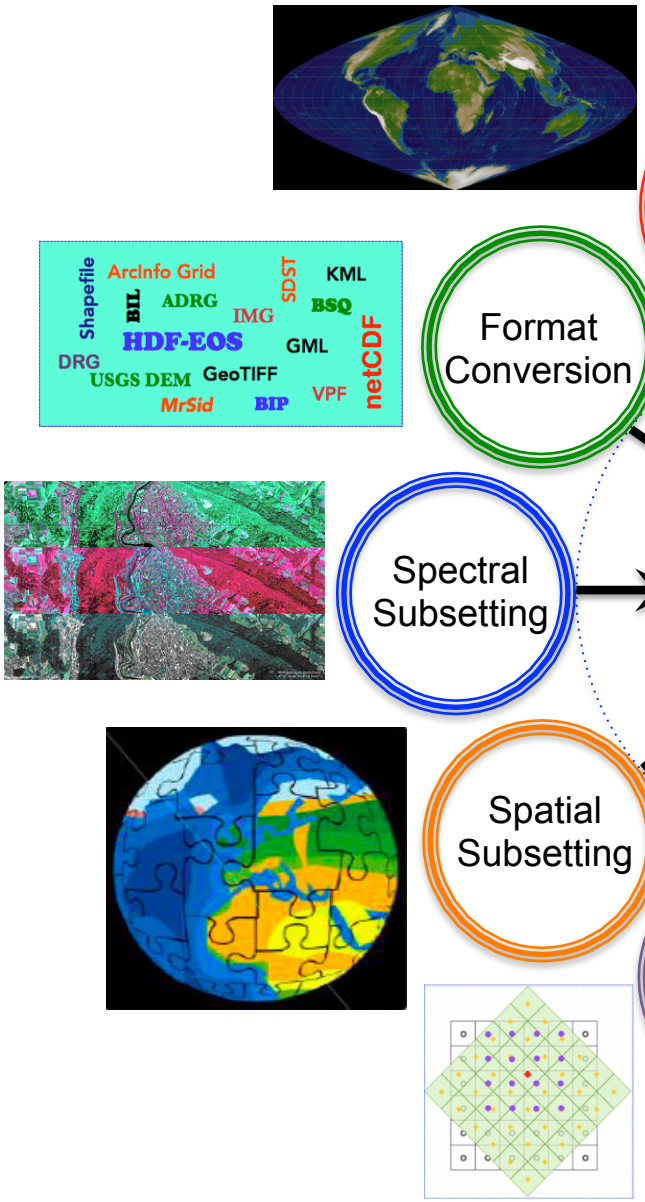
- Open-source Project for a Network Data Access Protocol
- High-performance network access protocol for complex science data
- Well-supported in Earth science community tools
  - Free: Panoply, IDV, McIDAS-V, nco,…
  - Commercial: ArcGIS, Matlab, IDL,…

OPeNDAP* access to data smoothes out format heterogeneity and supports subsetting



*Although the Hyrax implementation is shown, other OPeNDAP servers such as GrADS Data Server and THREDDS Data Server have similar capabilities but different architectures.

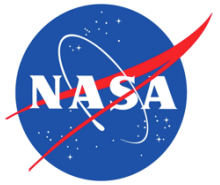
Data transformation options of several kinds can help with Variety and Volume



Data transformation applies fundamental changes and conversions to attributes of the original data to suit the application requirements of end-users

Courtesy of B. Ramachandran, MODAPS/LAADS

Big Earth Data Initiative (BEDI)



- OSTP-driven multi-agency effort
- Focus on datasets in Societal Benefit Areas
- Several interoperability aspects…

BEDI in EOSDIS

- Improve dataset consistency across EOSDIS
  - Metadata in Common Metadata Repository
  - Data in OPeNDAP
- Improve machine access to EOSDIS
  - Developers' portal
    - How To Access Common Metadata Repository
    - How to Access OpENDAP-served Data
  - OPeNDAP performance
  - OPeNDAP use with Cloud storage

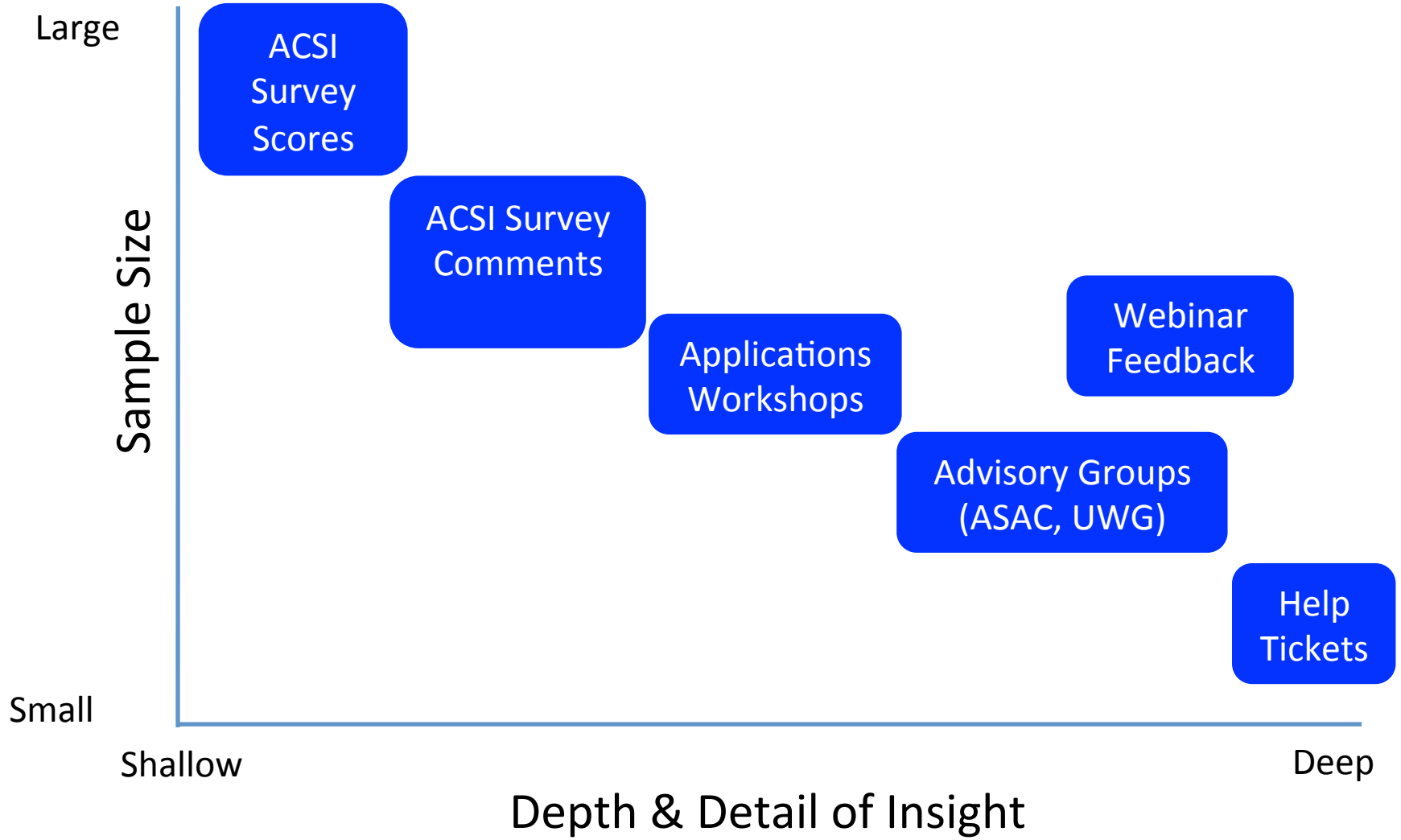
Community Engagement on Big Data

- Earth Science Information Partners (ESIP)
  - Variety: Clusters on Discovery, Information Quality
  - Volume: Clusters on Earth Science Data Analytics and Cloud Computing
- Earth Science Data Systems Working Groups
  - Formed of DAACs, ACCESS and MEaSUREs award winners
  - Variety: Working Groups on Dataset Interoperability, Search Relevancy
  - Volume: OPeNDAP Best Practices, Cloud Computing
- User Needs efforts
  - DAAC User Working Groups
  - American Customer Satisfaction Index survey
  - EOSDIS User Needs Analysis group

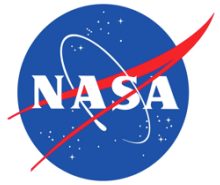
Big-Data-Community Engagement

- Big Data Theme for both ESIP 2016 Meetings
- Co-Convening AGU 2016 session on Big Data Analytics
- Program committee for IEEE Workshop on Big Data in Earth and Planetary Sciences
- ESA's Big Data from Space (BiDS) workshops
  - "Improving Earth Science Data Discoverability And Use Through Metadata Relationship Graphs, Virtual Collections, And Search Relevancy"

User Needs from Community Sources



Take Home Message



1. Cloud prototypes are underway to tackle the Volume challenge of Big Data…

2....But advances in computer hardware or cloud won't help (much) with Variety

3. Standards, conventions, and community engagement are the key to addressing Variety

Backup Slides

OPeNDAP Enhancements from the Big Earth Data Initiative

- More OPeNDAP for EOSDIS data
- More aggregation along time for data in OPeNDAP
  - Improved performance for aggregation in Hyrax