Big Data in the Earth Observing System Data and Information System

Chris Lynnes (System Architect, Code 586)
Katie Baynes (System Architect, Code 586)
*Mark McInerney (Deputy Project Manager, Code 423)*
NASA/GSFC, ESDIS Project

Earth Observing System Data and Information System (EOSDIS)

EOSDIS

data downlink

capture and clean → process → archive → subset → distribute

distribute → Research
distribute → Applications
distribute → Education

Users

Earth Observing System Data and Information System (EOSDIS)

EOSDIS

Research

Applications

Education

Users

distribute

subset

archive

data downlink

capture and clean

process
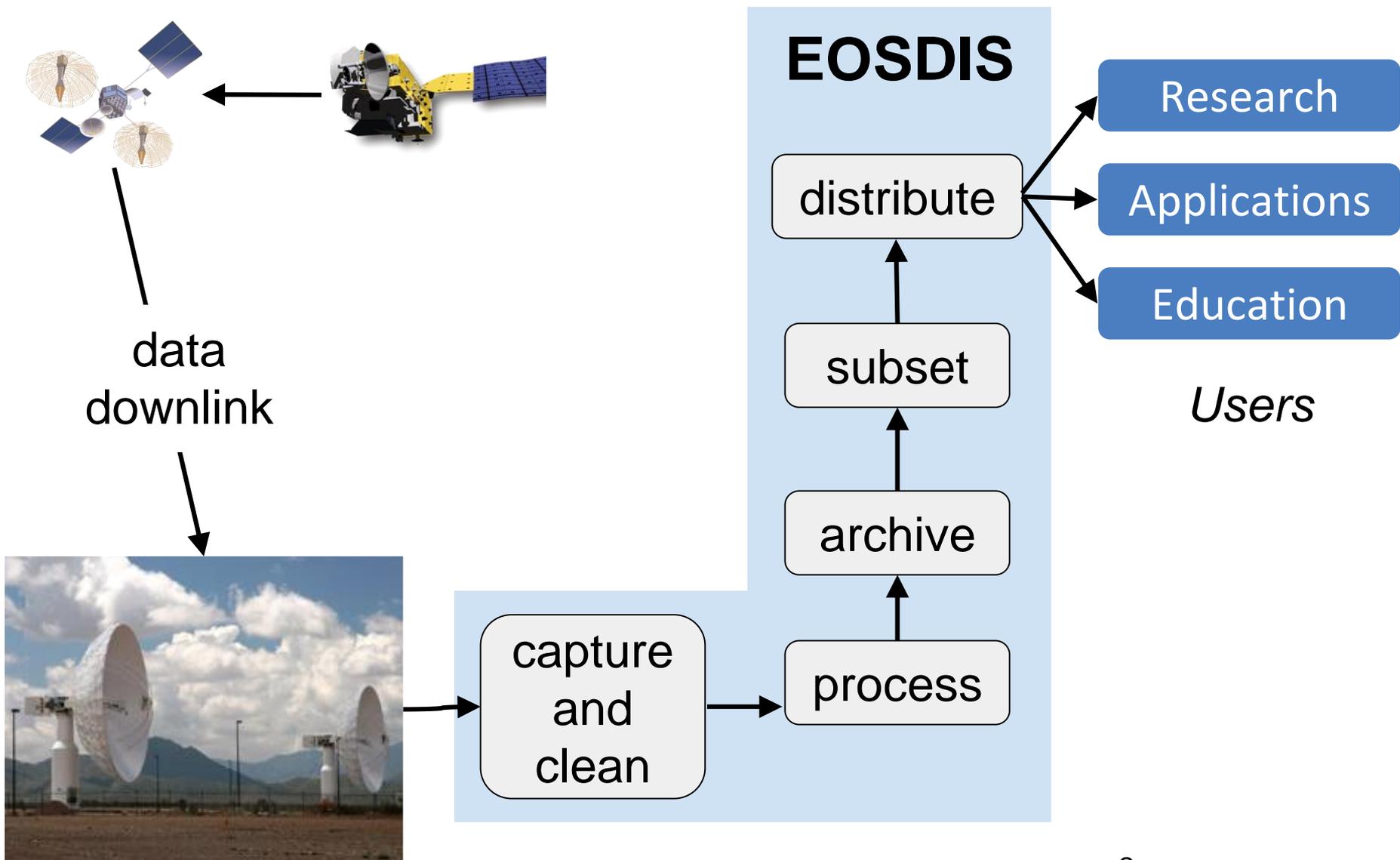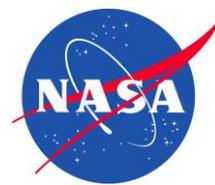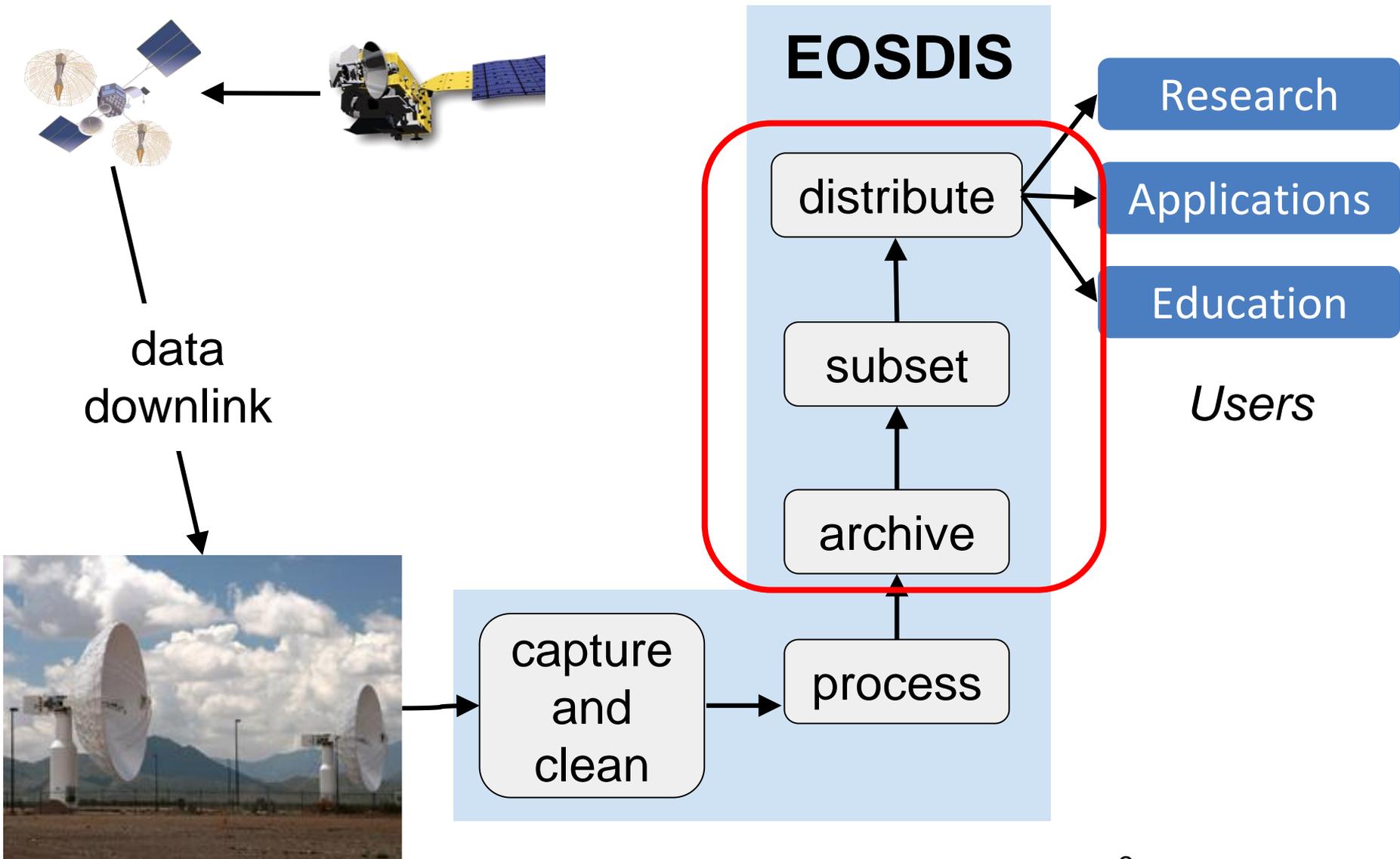
Take Home Message Preview
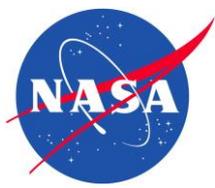
1. Cloud prototypes are underway to tackle the _Volume_ challenge of Big Data...

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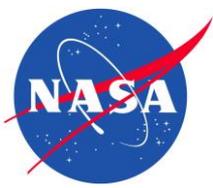
…Volume
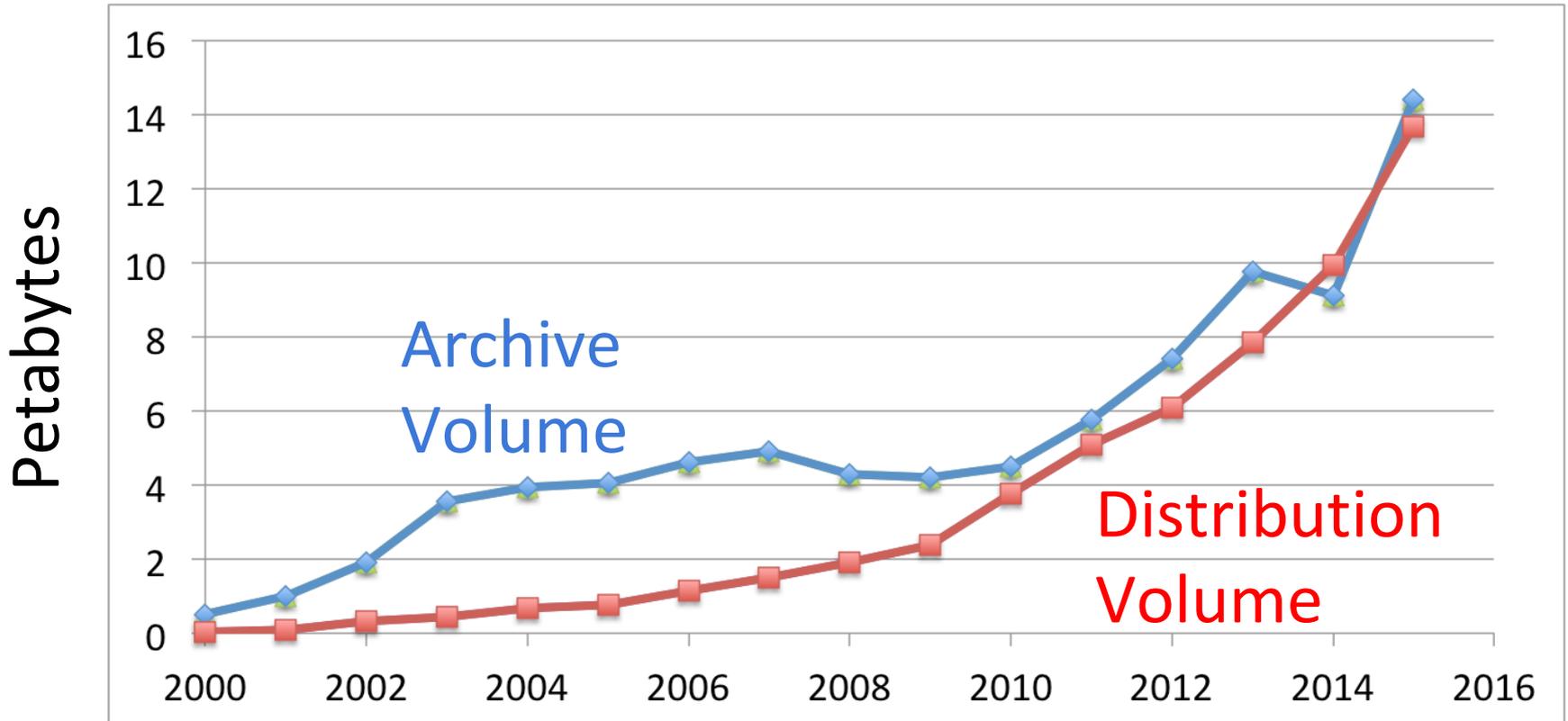
Big Data Indicators
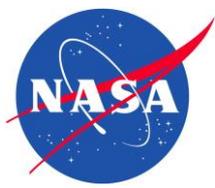
| EOSDIS FY2015 Metrics	
Unique Data Products	9,462
Distinct Users of EOSDIS Data and Services	2.6 M
Average Daily Archive Growth	16 TB/day
Total Archive Volume (as of Sept. 30, 2015)	14.6 PB
End User Distribution Products	1.42 B
End User Average Daily Distribution Volume	32.1 TB/day

EOSDIS Cloud Prototypes
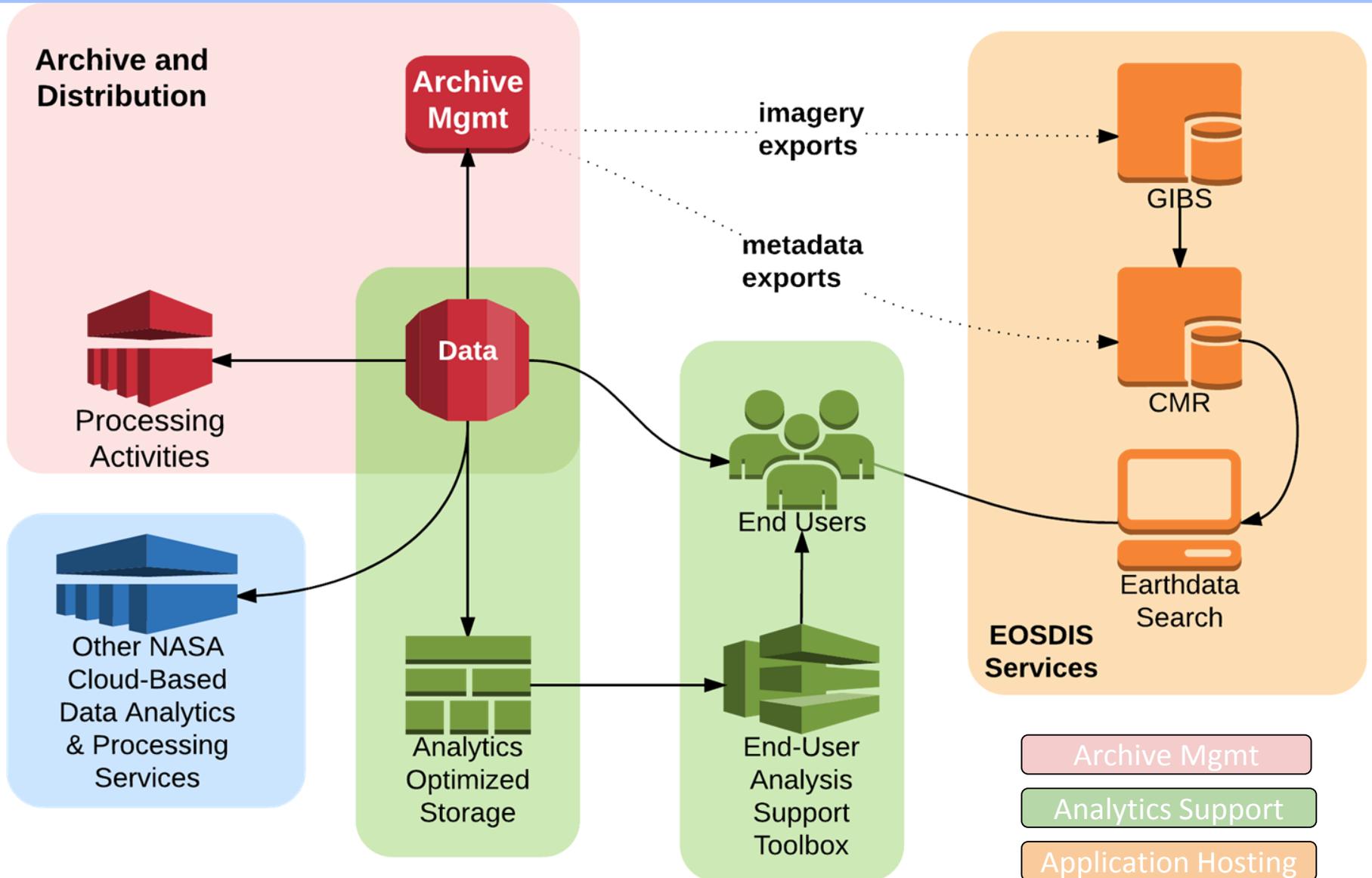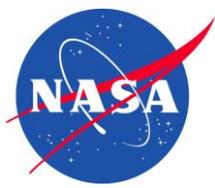
Archive Cloud Prototypes

Archive and Distribution
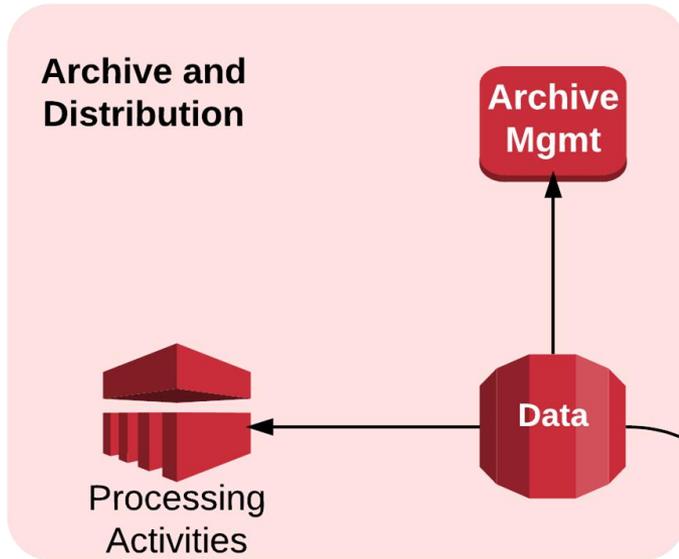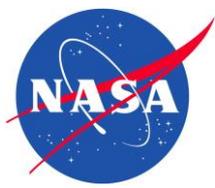
Benefits from Archive in the Cloud

‣ 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

Archive and Distribution

Archive Mgmt

Data

Processing Activities

Other NASA Cloud-Based Data Analytics & Processing Services

Analytics Optimized Storage

End Users

End-User Analysis Support Toolbox

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

Distinct Science Products Distributed

Instrument Variety

Spectrometer

Radiometer

Sounder

Interferometer

Polarimeter

GPS

Radar

Lidar

Accelerometer

Limb Scanner

Imager

Microwave

Infrared

Synthetic Aperture

Precipitation

Scatterometer

Profiler

Doppler

Altimeter

Microwave

Infrared

Visible

Ultraviolet
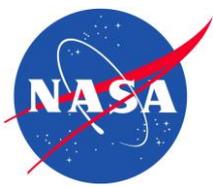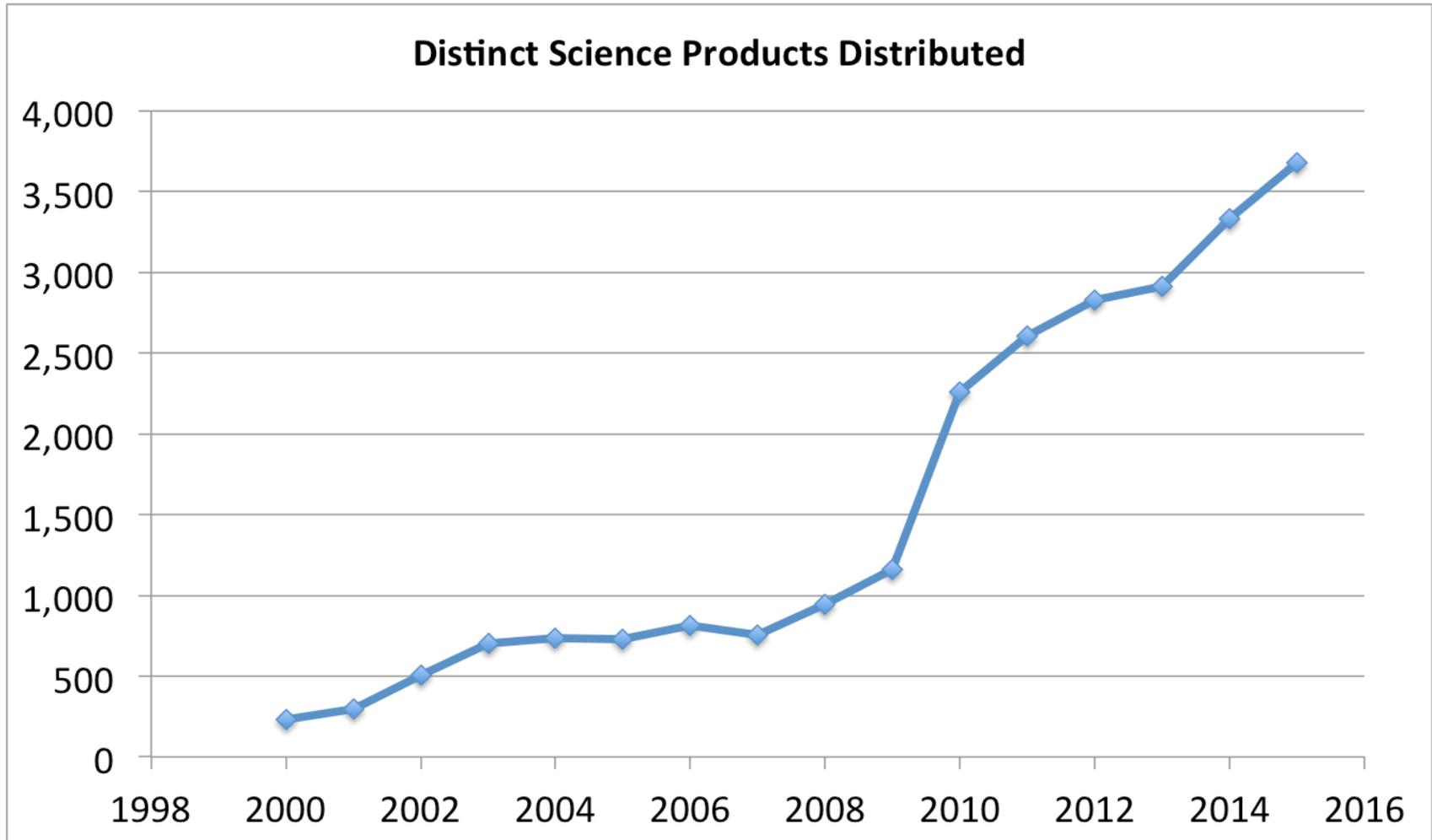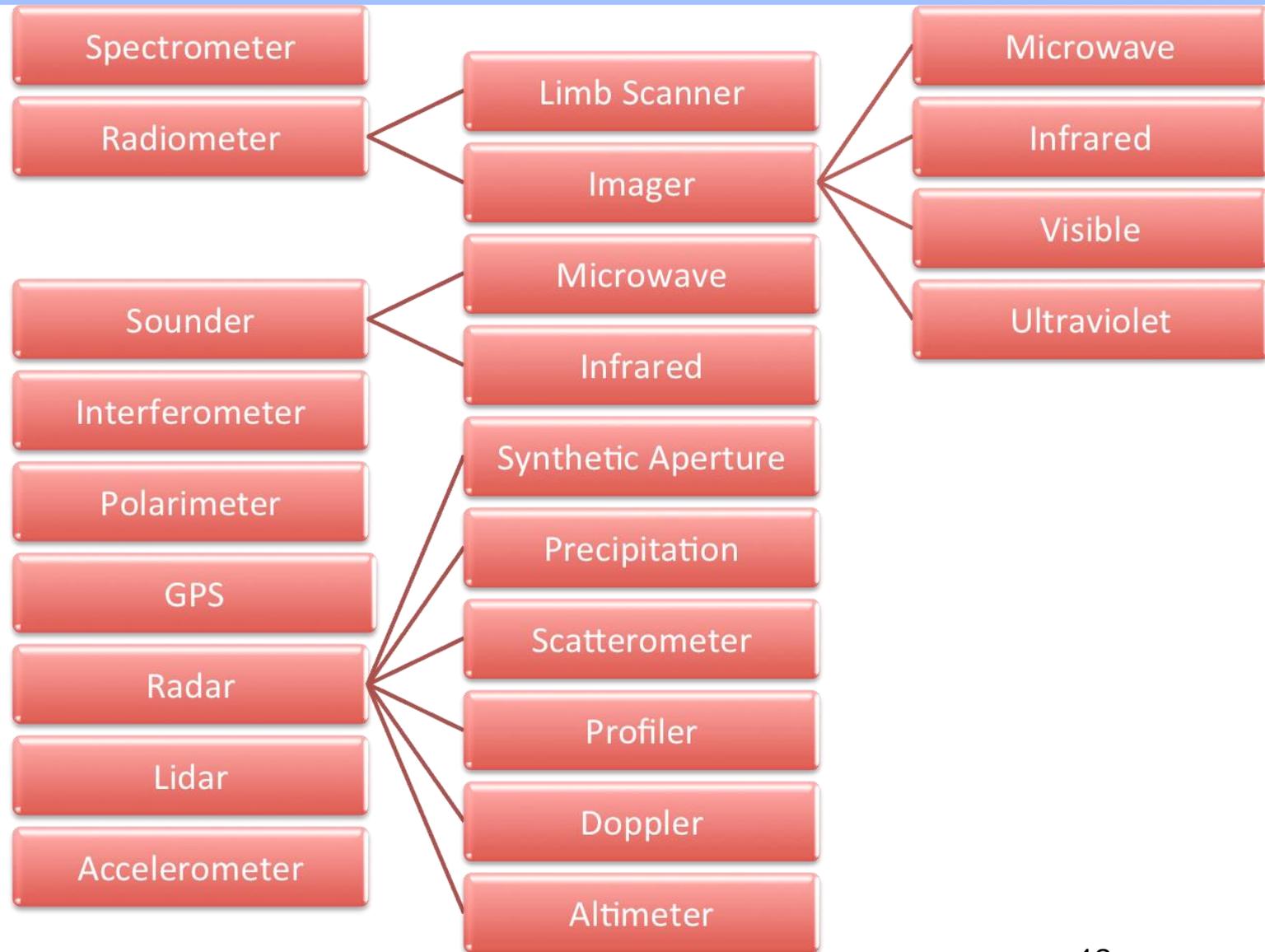
Satellite Instrument "Footprints"
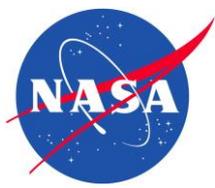
Example Imaging Footprint
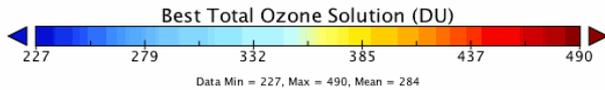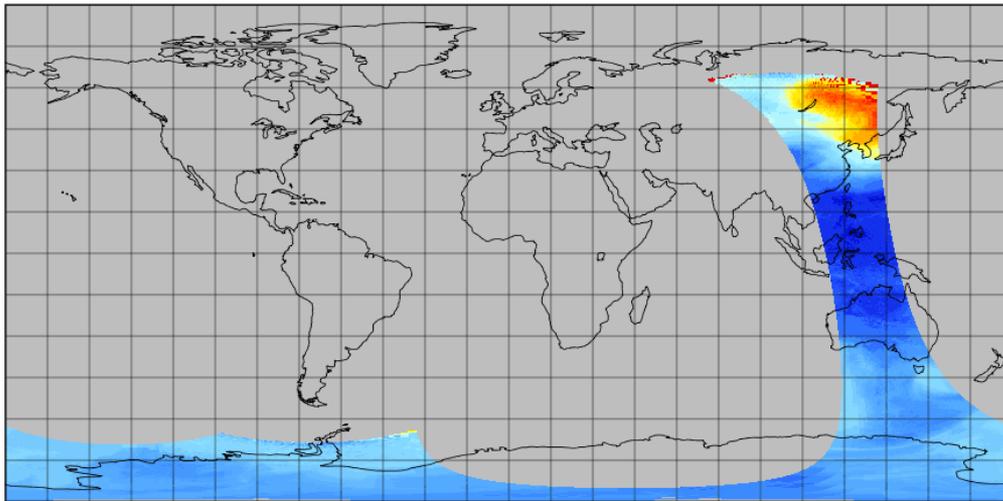
Example LIDAR footprint
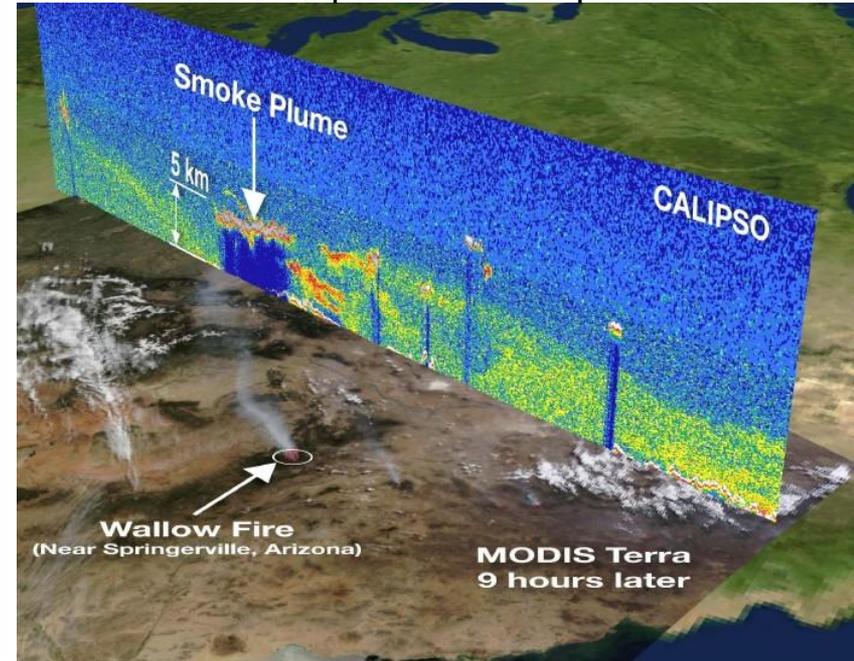
Example Limb Scanning Footprint
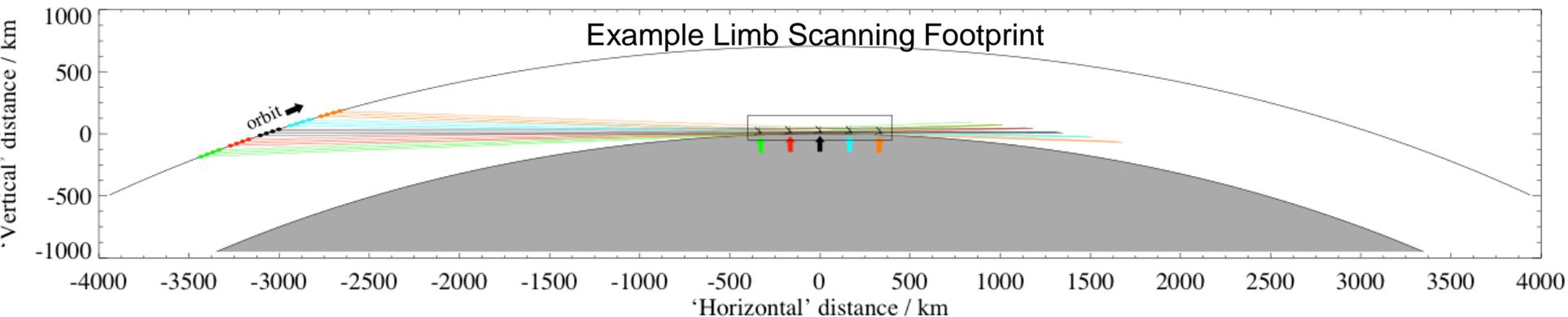
Microwave Limb Scanner (from Algorithm Theoretical Basis Document, Livesey and Wu, 1999)

Aircraft and In Situ
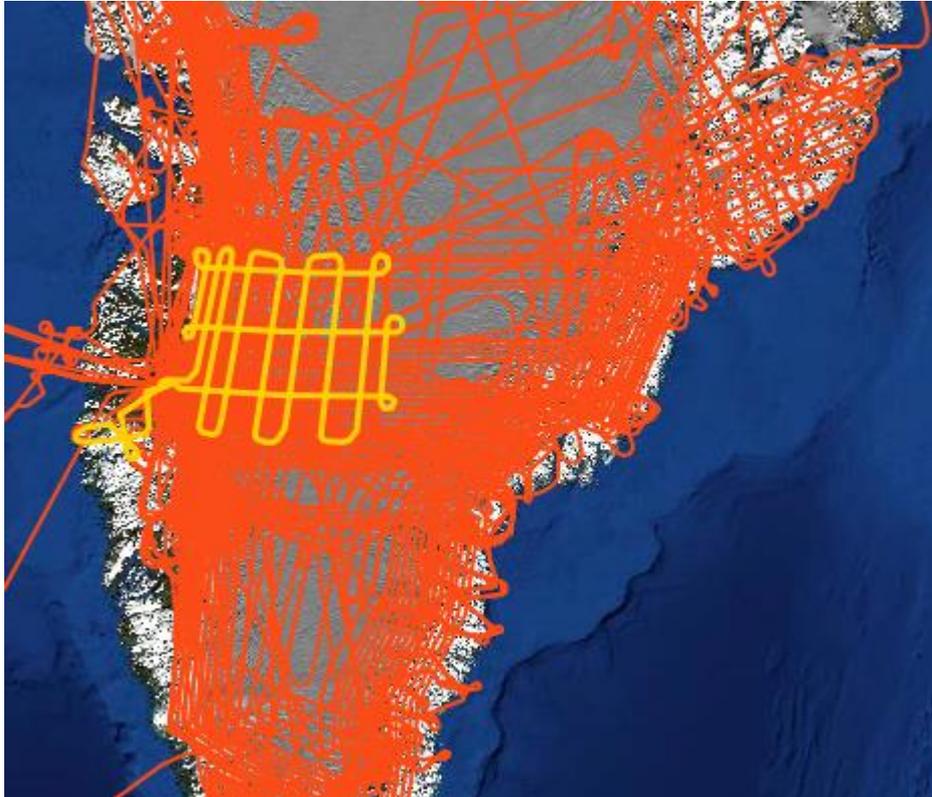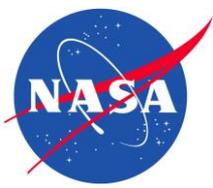
Same Instrument, Different Satellite
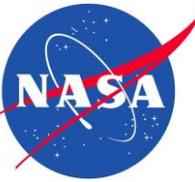
Aerosol Optical Depth 550 nm (Dark Target)

Aqua

Aerosol Optical Depth 550 nm (Dark Target) (1)
0.0   0.3   0.6   0.9   1.2   1.5
Data Min = 0.0, Max = 2.7
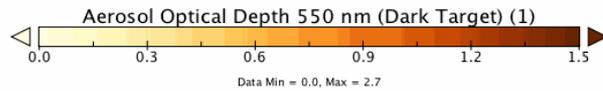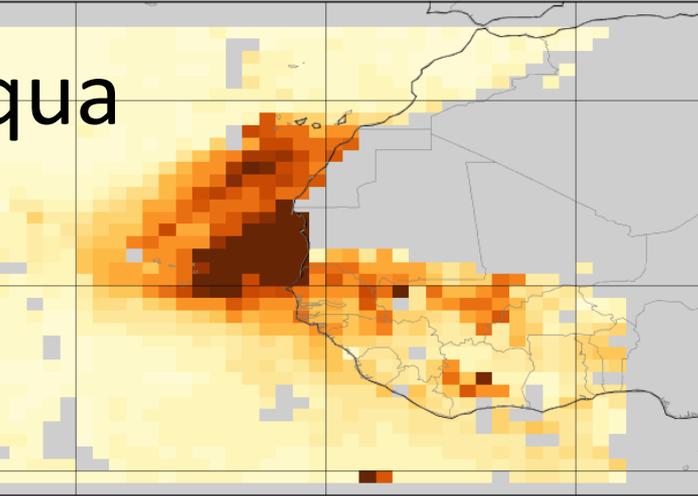
Aerosol Optical Depth 550 nm (Dark Target)

Terra

Aerosol Optical Depth 550 nm (Dark Target) (1)
0.0   0.3   0.6   0.9   1.2   1.5
Data Min = −0.1, Max = 3.2

Aerosol Optical Depth 550 nm (Dark Target)

Aqua - Terra

Aerosol Optical Depth 550 nm (Dark Target) − Aerosol Optical Depth 550 nm (Dark Target)
−1.0   −0.6   −0.2   0.2   0.6   1.0
Data Min = −2.5, Max = 2.2

15
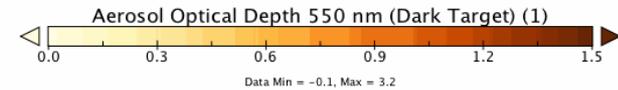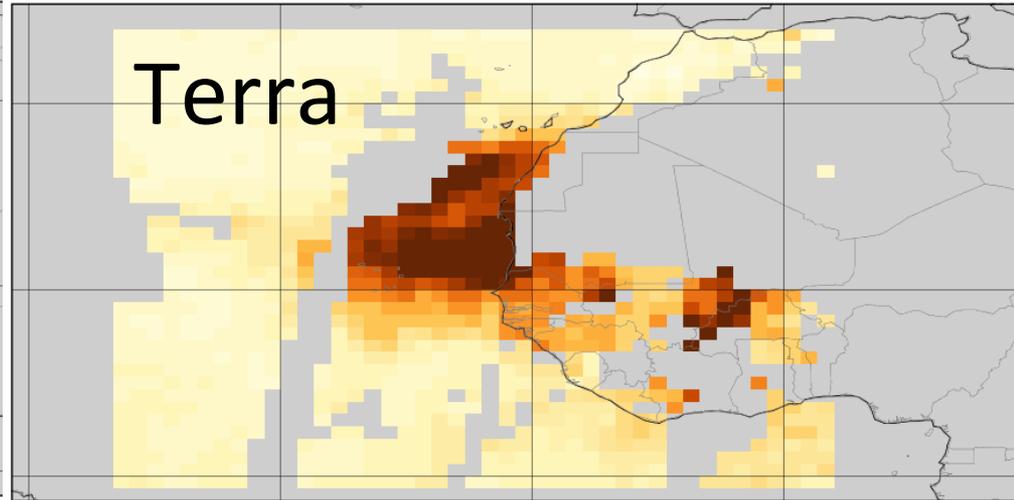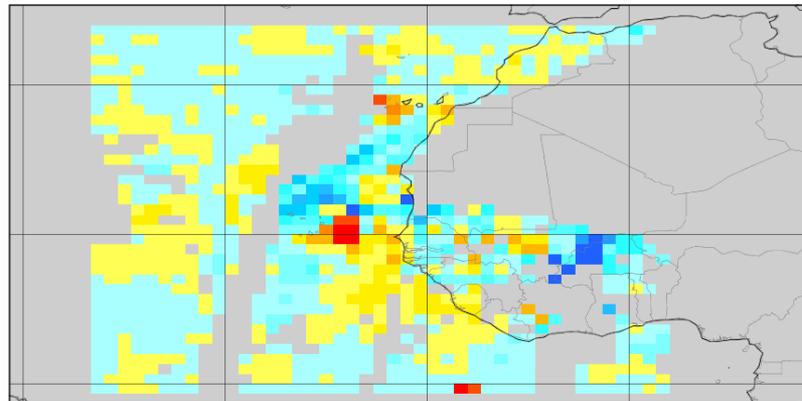
Same Instrument+Satellite, Different Algorithm
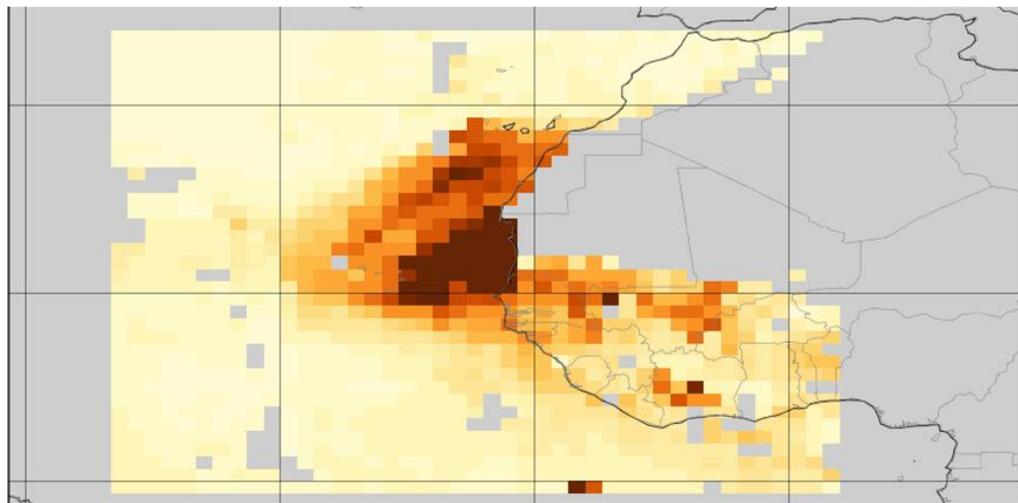
MODIS on Aqua
Aerosol Optical Depth

Dark Target Algorithm

Deep Blue Algorithm

Aerosol Optical Depth 550 nm (Deep Blue, Land−only) (1)

0.000   0.300   0.600   0.900   1.200   1.500

Data Min = 0.032, Max = 0.967
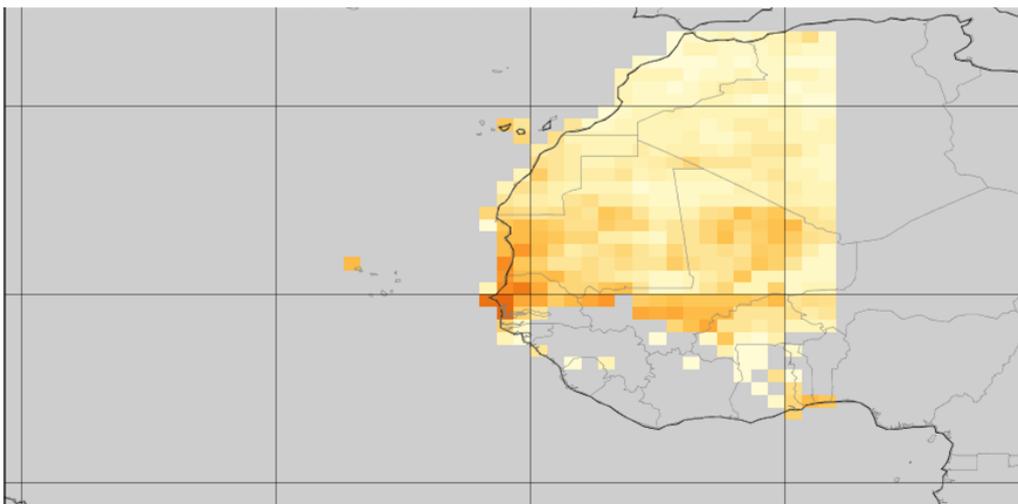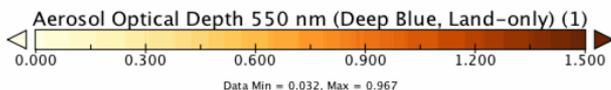
Processing Levels
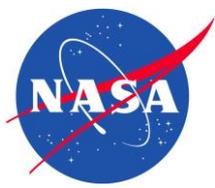
AIRS data for 2011-08-11

Level 1B
Calibrated radiance at a pixel
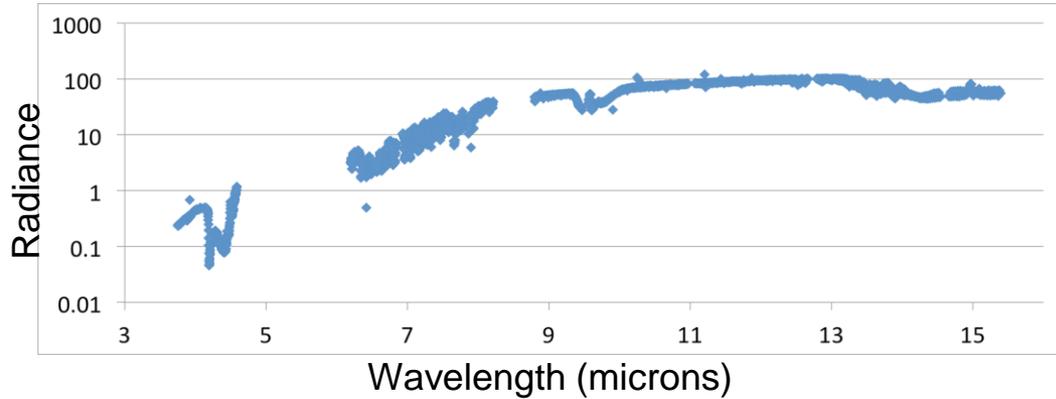
Level 2
Carbon monoxide for one scene
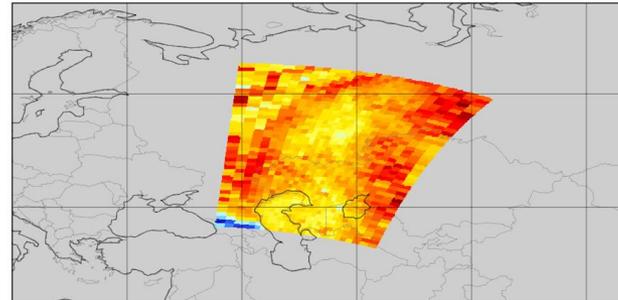
Level 3
Global carbon monoxide for one night
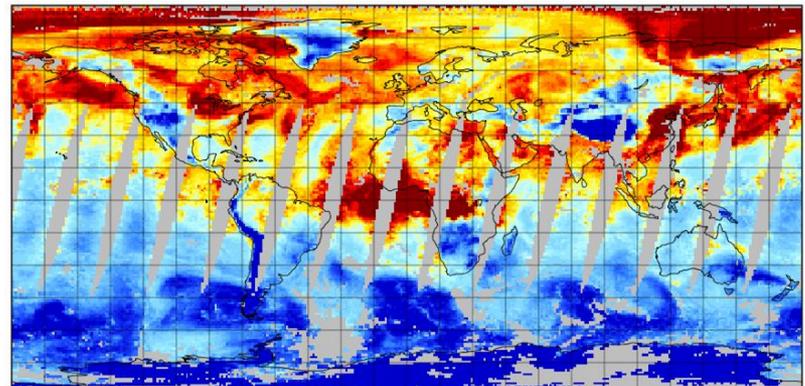
Projections
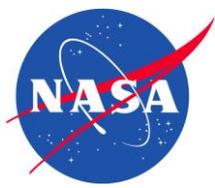
SurfAirTemp_D

SurfAirTemp_D ()

275.000   280.600   286.200   291.800   297.400   303.000

Data Min = 212.125, Max = 306.688, Mean = 286.636

18
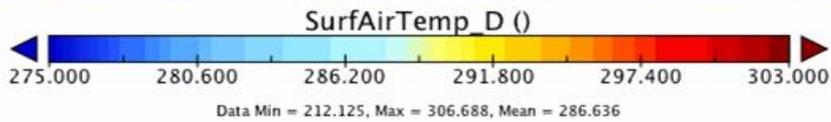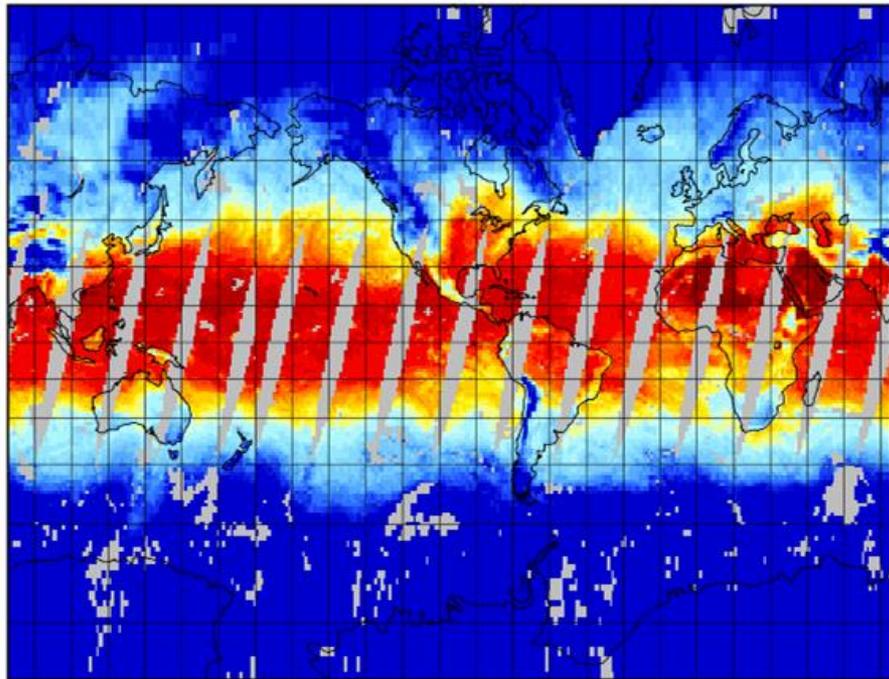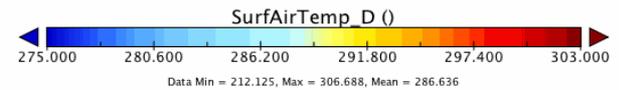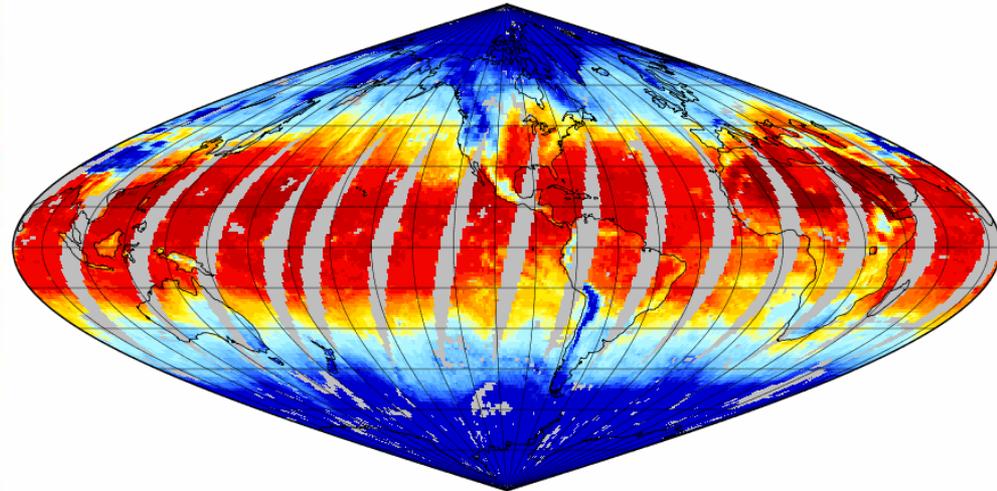
Time Aggregation

Time Averaged Map of Aerosol Optical Depth 555 nm daily 0.5 deg. [MISR MIL3DAE v4]
over 2009-09-21, Region 180W, 90S, 180E, 90N
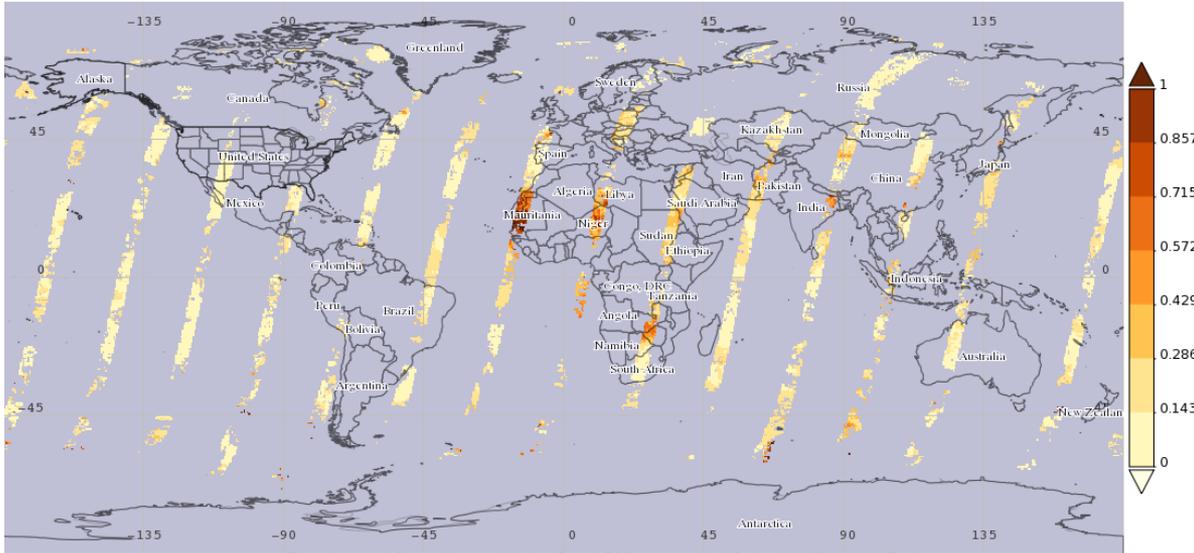
Aerosol Optical Depth at 555 nm from Multi-angle Imaging Spectro-Radiometer

Daily

Time Averaged Map of Aerosol Optical Depth 555 nm monthly 0.5 deg. [MISR MIL3MAE v4]
over 2009-Sep, Region 180W, 90S, 180E, 90N
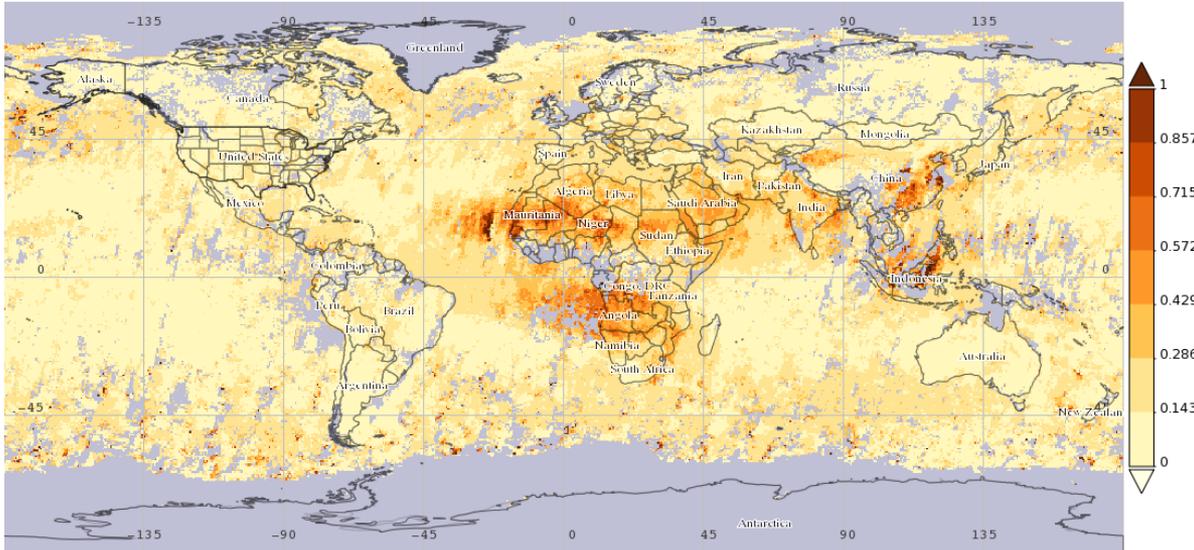
Monthly

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

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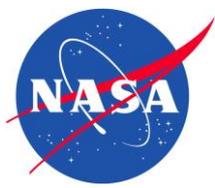
SeaWiFS Deep-Blue Aerosol Optical Depth 2006-10-06

1.0 Degree Resolution

0.5 Degree Resolution

Aerosol Optical Depth 550 nm (1)

0.00   0.44   0.88   1.32   1.76   2.20
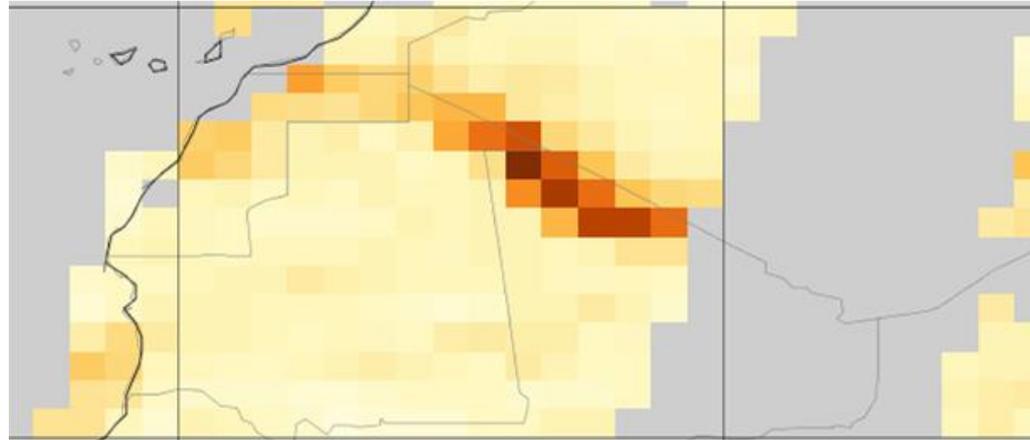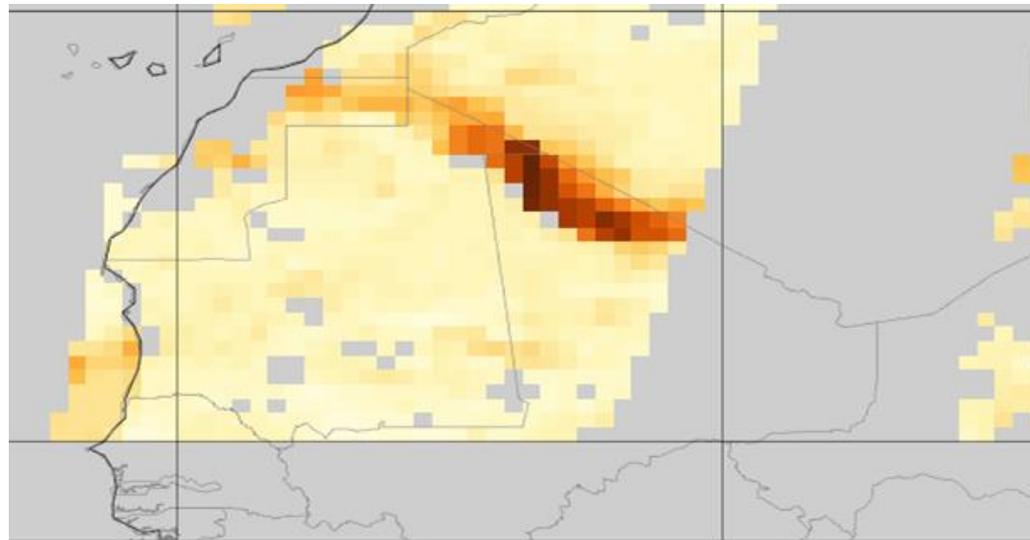
Data Min = 0.02, Max = 2.33
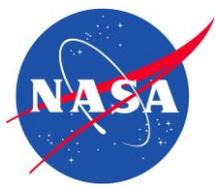
Data Formats

- **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
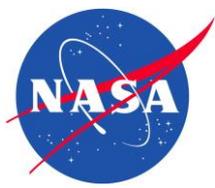
1. Interoperable discipline-focused DAACs

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)

Alaska Satellite Facility (Synthetic Aperture Radar)

Socio-economic Data Archive Center

Ocean Biology DAAC

Land Processes DAAC

National Snow and Ice Data Center DAAC

Crustal Dynamics Data and Information System

Level 1 and Atmosphere Archive and Distribution System (MODIS)

Oak Ridge National Laboratories DAAC (Biogeochemistry)

Global Hydrology Resource Center

Atmospheric Sciences Data Center

Goddard Earth Sciences Data and Information Service Center

Physical Oceanography DAAC

Different DAACs have different "-Spheres of Influence"
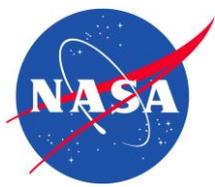
DAAC	Atmo	Hydro	Cryo	Litho	Bio	Anthropo	Sub-Specialty
Alaska Satellite Facility			✓	✓			SAR
Atm. Sciences Data Center	✓						
Crustal Dynamics Data Info Sys				✓			Space geodesy
Global Hydrology Resource Ctr		✓					Weather events
Goddard Earth Sciences DISC	✓	✓					
Land Processes DAAC					✓	✓	
L1 and 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

Earthdata Search Client

Other Clients
Other Clients

API

OpenSearch

DAAC
DAAC
DAAC
DAAC
DAAC
DAAC
DAAC
DAAC
DAAC
DAAC

metadata

Common Metadata Repository

Unified Metadata Model

Collections   Visualizations

Granules   Services

Variables   Documentation
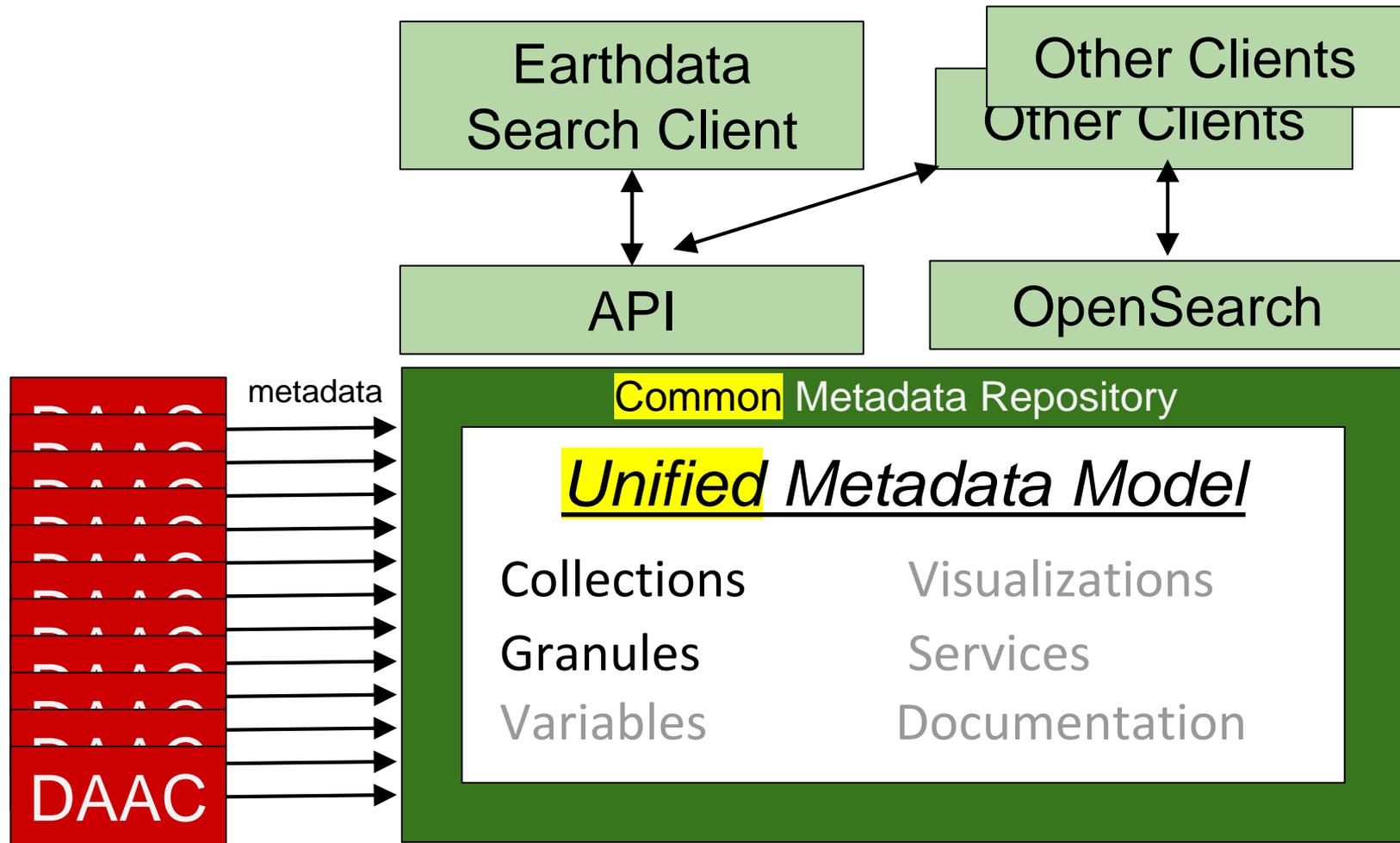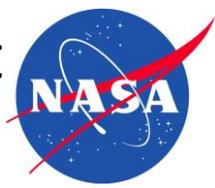
One Metadata System to rule them all,
One Metadata System to find them,
*One Metadata System to bring them all*
And in cyberspace bind them
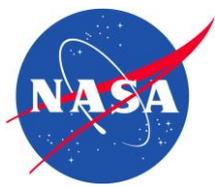
OPeNDAP

- 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

Analysis Tool

netCDF Library

- -

Hyrax server

hdf4 handler	hdf5 handler	netcdf handler	freeform handler

| Hierarchical Data Format version 4 Files	Hierarchical Data Format version 5 Files	network Common Data Form Files	ASCII Files	flat binary files
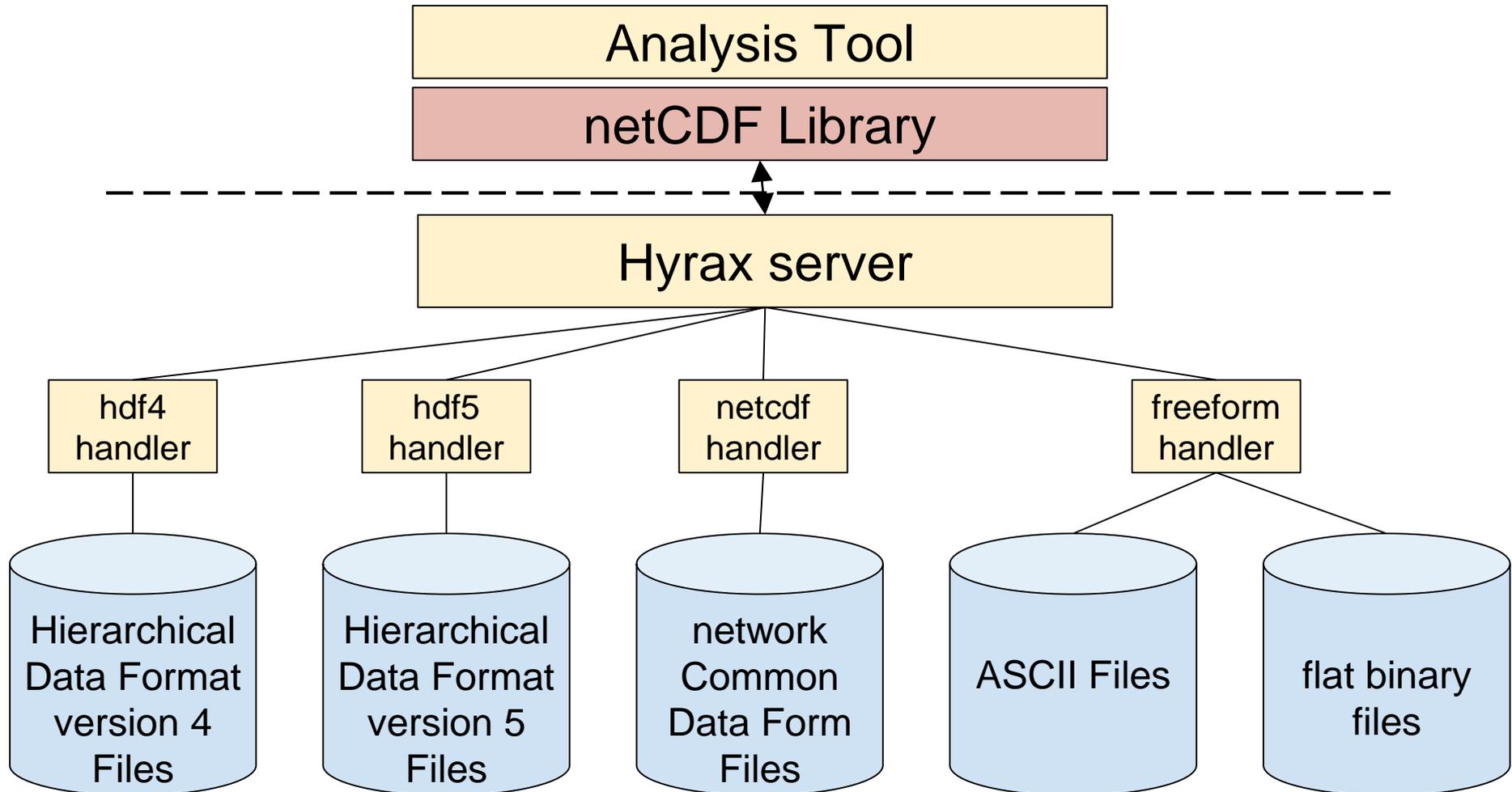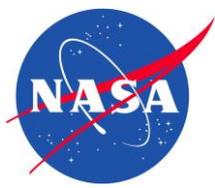
*Although the Hyrax implementation is shown, other OPeNDAP servers such as GrADS Data Server and THREDDS Data Server have similar capabilities but different architectures.

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

Map Projection

Format Conversion

Spectral Subsetting

Data Transformation

Spatial Subsetting

Resampling Parameters

Courtesy of B. Ramachandran, MODAPS/LAADS

Big Earth Data Initiative (BEDI)
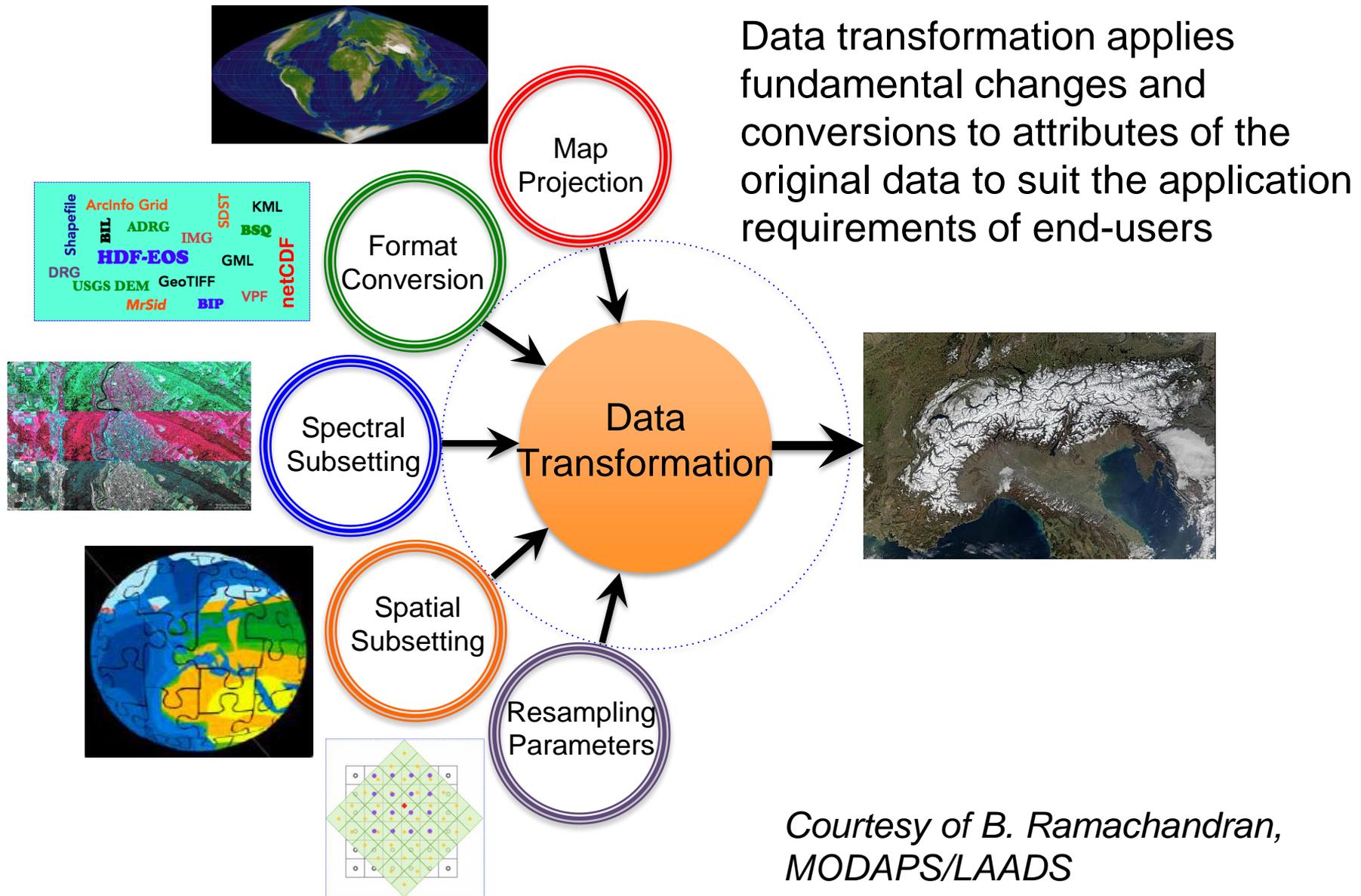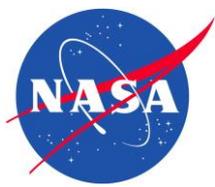
- OSTP-driven multi-agency effort
- Focus on datasets in Societal Benefit Areas
- Several interoperability aspects…

BEDI in EOSDIS
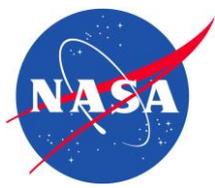
- **Improve dataset consistency across EOSDIS**
  - Metadata in Common Metadata Repository
  - Data in OPeNDAP

- **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
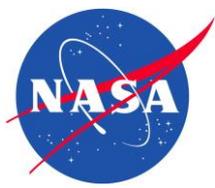
- ## Earth Science Information Partners (ESIP)
  - Variety: Clusters on Discovery, Information Quality
  - Volume: Clusters on Earth Science Data Analytics and Cloud Computing
- ## Earth Science Data Systems Working Groups
  - Formed of DAACs, ACCESS and MEaSUREs award winners
  - Variety: Working Groups 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
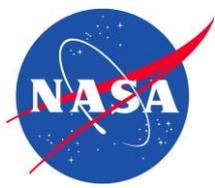
- Big Data Theme for both ESIP 2016 Meetings
- Co-Convening AGU 2016 session on Big Data Analytics
- Program committee for IEEE Workshop on Big Data in Earth and Planetary Sciences
- ESA's Big Data from Space (BiDS) workshops
  - "Improving Earth Science Data Discoverability And Use Through Metadata Relationship Graphs, Virtual Collections, And Search Relevancy"

User Needs from Community Sources

Large — Small (Sample Size) vs. Shallow — Deep (Depth & Detail of Insight)

- ACSI Survey Scores
- ACSI Survey Comments
- Applications Workshops
- Webinar Feedback
- Advisory Groups (ASAC, UWG)
- Help Tickets
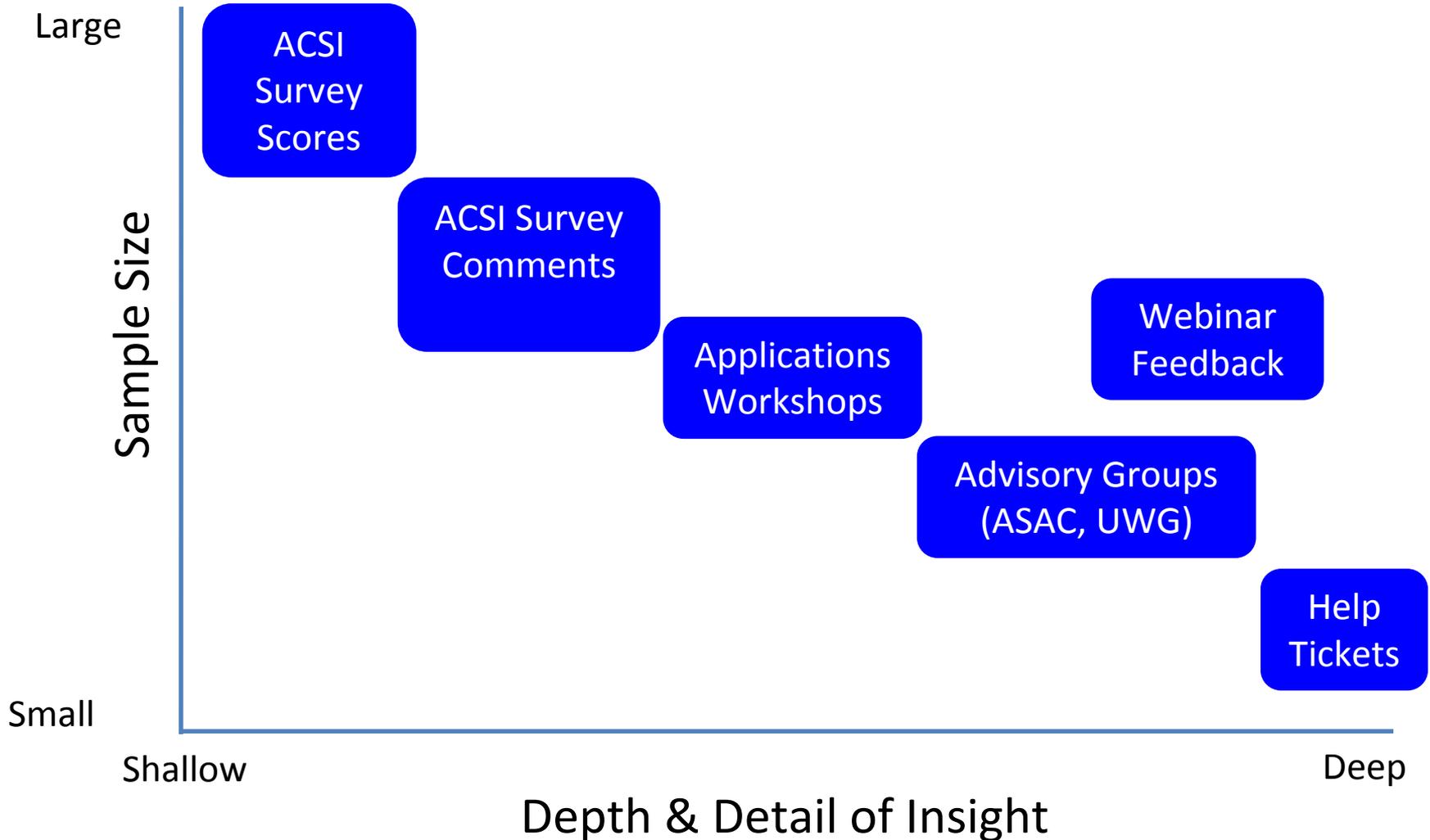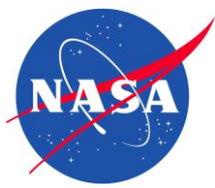
Take Home Message

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
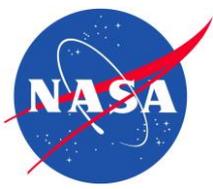
OPeNDAP Enhancements from the Big Earth Data Initiative
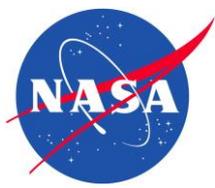
- More OPeNDAP for EOSDIS data
- More aggregation along time for data in OPeNDAP
  - Improved performance for aggregation in Hyrax