GHRC Overview and Highlights

Rahul Ramachandran

GHRC DAAC Manager/MSFC ST11

2017 GHRC User Working Group Meeting

Sept 26-27, 2017

NASA

itsc
INFORMATION TECHNOLOGY AND SYSTEMS CENTER

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Thank You!

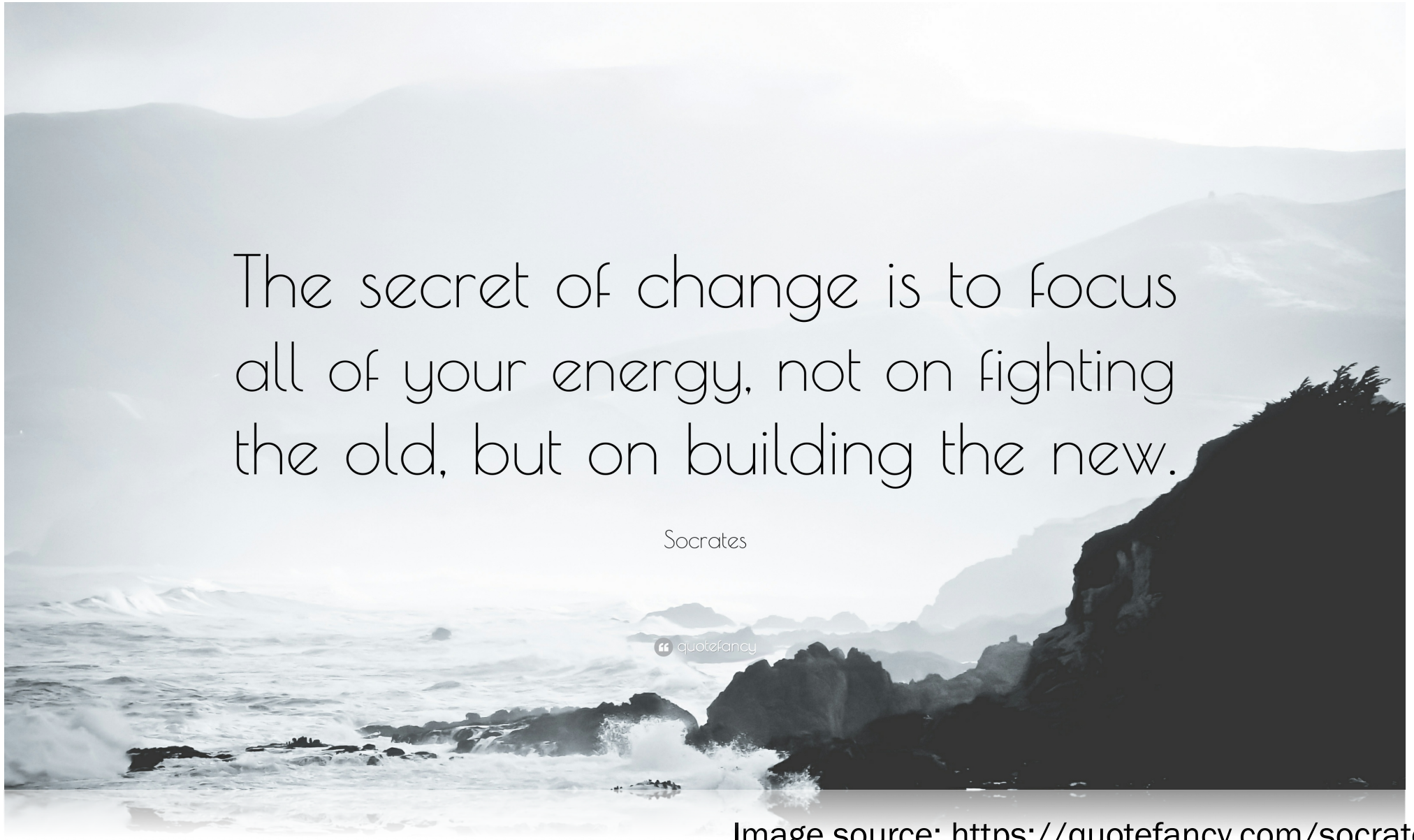


Image source: https://quotefancy.com/socrates-quotes

Overall Organization

EARTHDATA Powered by EOSDIS

ABOUT DATA COMMUNITY RESOURCES

Search datasets, news, articles, and information

HQ Earth Science Data Systems (ESDS) Program › About the ESDS Program

HQ ESDS Program

- HQ Earth Science Data Systems (ESDS) Program
- About the ESDS Program
- ESDS Policies
- Data and Information Policy
- Data Processing Levels
- Data Rights & Related Issues
- Open Source Policy
- New Missions
- Data Management Plan Guidance
- New Missions Requirements
- Program Components
- Continuous Evolution
- Program Review Findings and Recommendations

More Resources

About the ESDS Program

The Earth Science Data Systems (ESDS) Program is responsible for:

- Actively managing NASA's Earth science data as a national asset
- Developing data system capabilities optimized to support rigorous science investigations and unique needs of multiple science disciplines
- Processing instrument data to create Earth System Data Records (ESDRs)
- Upholding NASA's policy of free, full, and open sharing of all data, tools, and ancillary information for all users
- Engaging members of the Earth science community in the evolution of data systems

Alignment with NASA Strategic Plan

Mission Statement

Program Charter

Continuous Evolution

Collaborations

HQ: https://earthdata.nasa.gov/earth-science-data-systems-program/about-the-esds-program

EARTHDATA Powered by EOSDIS

ABOUT DATA COMMUNITY RESOURCES

Search datasets, news, articles, and information

About EOSDIS › ESDIS Project

ESDIS Project

The Earth Science Data and Information System (ESDIS) Project is a part of the Earth Science Projects Division under the Flight Projects Directorate at the Goddard Space Flight Center (GSFC).

The ESDIS Project manages the science systems of the Earth Observing System Data and Information System (EOSDIS). EOSDIS provides science data to a wide community of users for NASA's Science Mission Directorate.

The ESDIS Project is responsible for:

- Processing, archiving, and distributing Earth science satellite data (land, ocean, atmosphere, cryosphere, human dimensions, and calibrated radiance and solar radiance data products)
- Providing tools to facilitate the processing, archiving, and distribution of Earth science data
- Collecting metrics and user satisfaction data to learn how to continue improving services provided to users
- Ensuring scientists and the public have access to data to enable the study of Earth from space to advance Earth system science to meet the challenges of climate and environmental change.
- Promoting the interdisciplinary use of EOSDIS data, including data products, data services, and data handling tools to a broad range of existing and potential user communities.

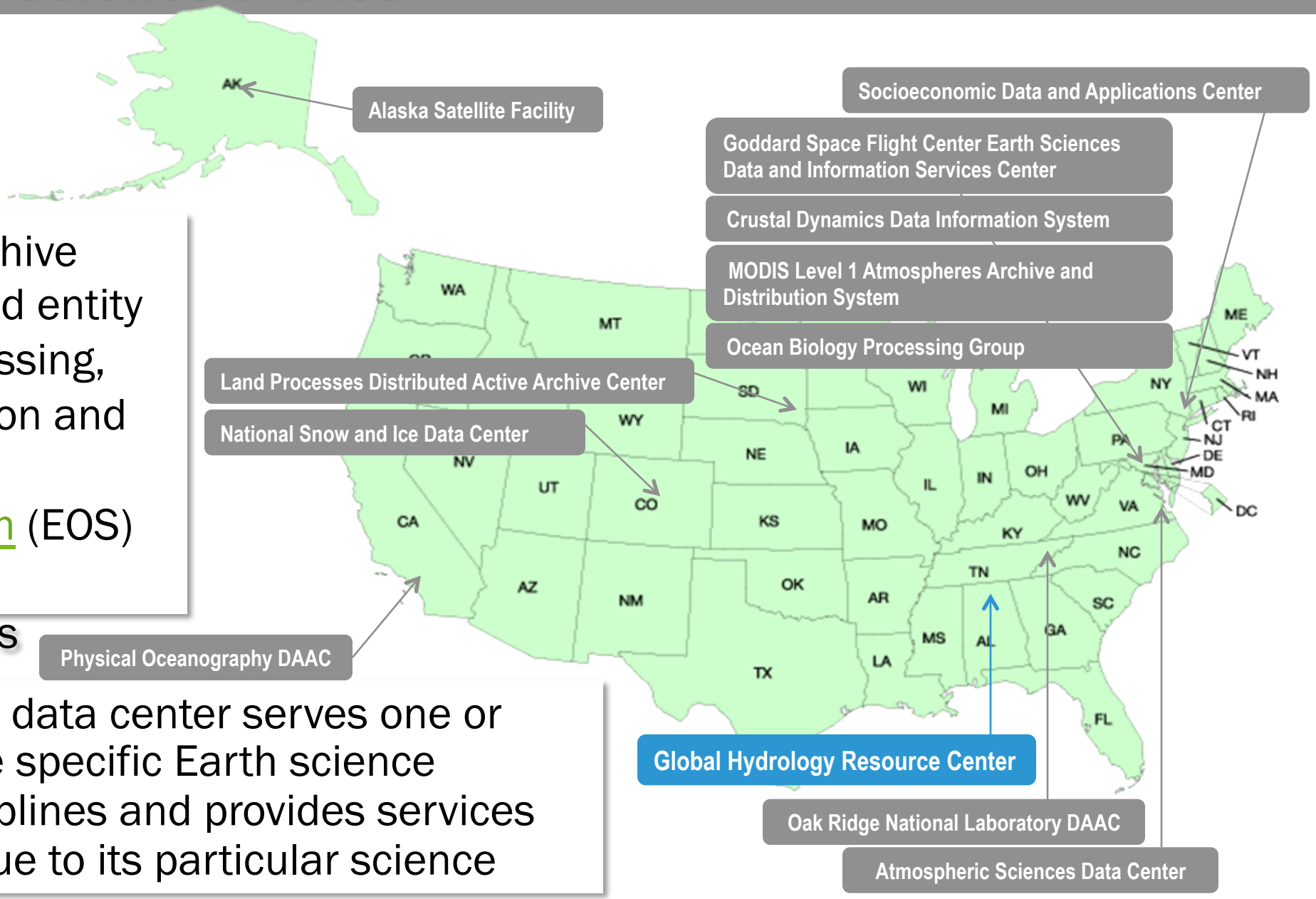
For information on various components within EOSDIS, visit the Science System Description page.

The ESDIS Project Supports		
Science System Elements	Distributed Active Archive Centers (DAACs)	12
	Science Investigator-led Processing Systems (SIPS)	15

GSFC: https://earthdata.nasa.gov/about/esdis-project

NASA's Earth Science DAACs

A Distributed Active Archive Center is a NASA-funded entity charged with the processing, archiving, documentation and distribution of Earth Observing System (EOS) satellites and field measurement programs

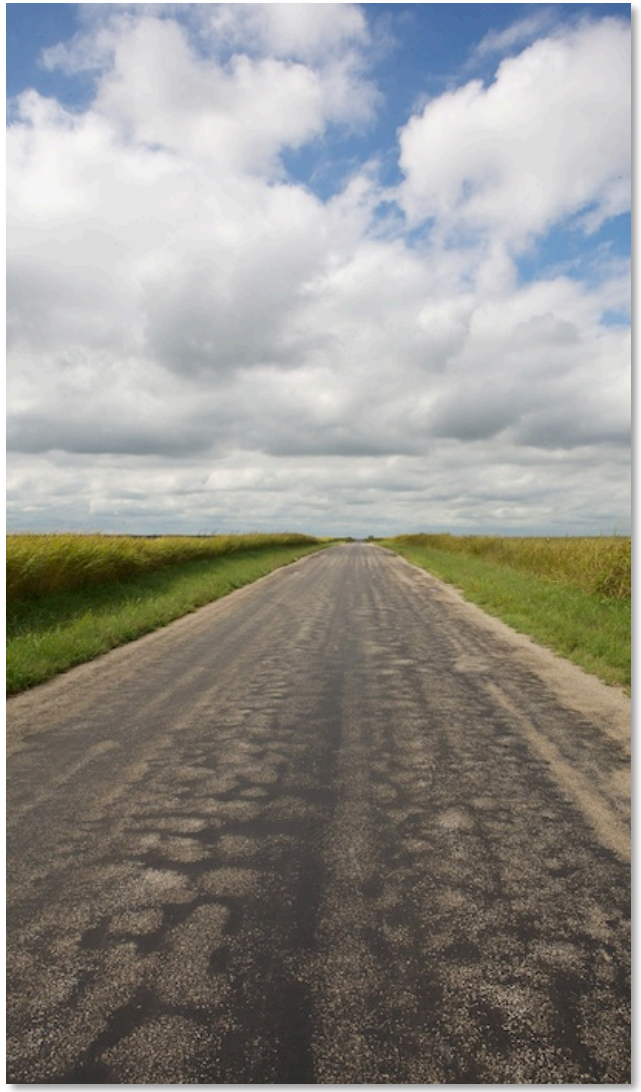


Each data center serves one or more specific Earth science disciplines and provides services unique to its particular science

Vision Statement

GHRC subscribes to the NASA ESDS Vision:

Make NASA's *free* and *open* Earth science data *interactive, interoperable* and *accessible* for research and societal benefit today and tomorrow.



Mission Statement

To provide a comprehensive active archive of both data and knowledge augmentation services with a focus on *hazardous weather, its governing dynamical and physical processes, and associated applications*.

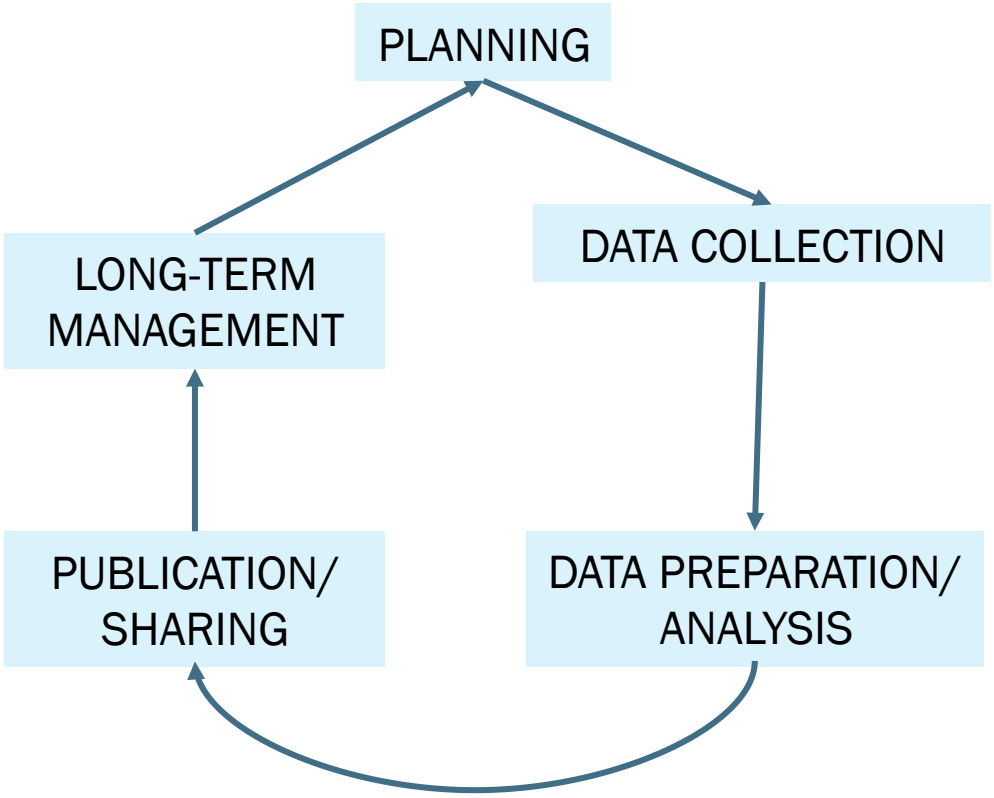
Focus on *lightning, tropical cyclones, and storm-induced hazards* through integrated collections of satellite, airborne, and in-situ data sets.

http://ghrc.nsstc.nasa.gov/

DAAC Role in Supporting Science

Research Data Life Cycle

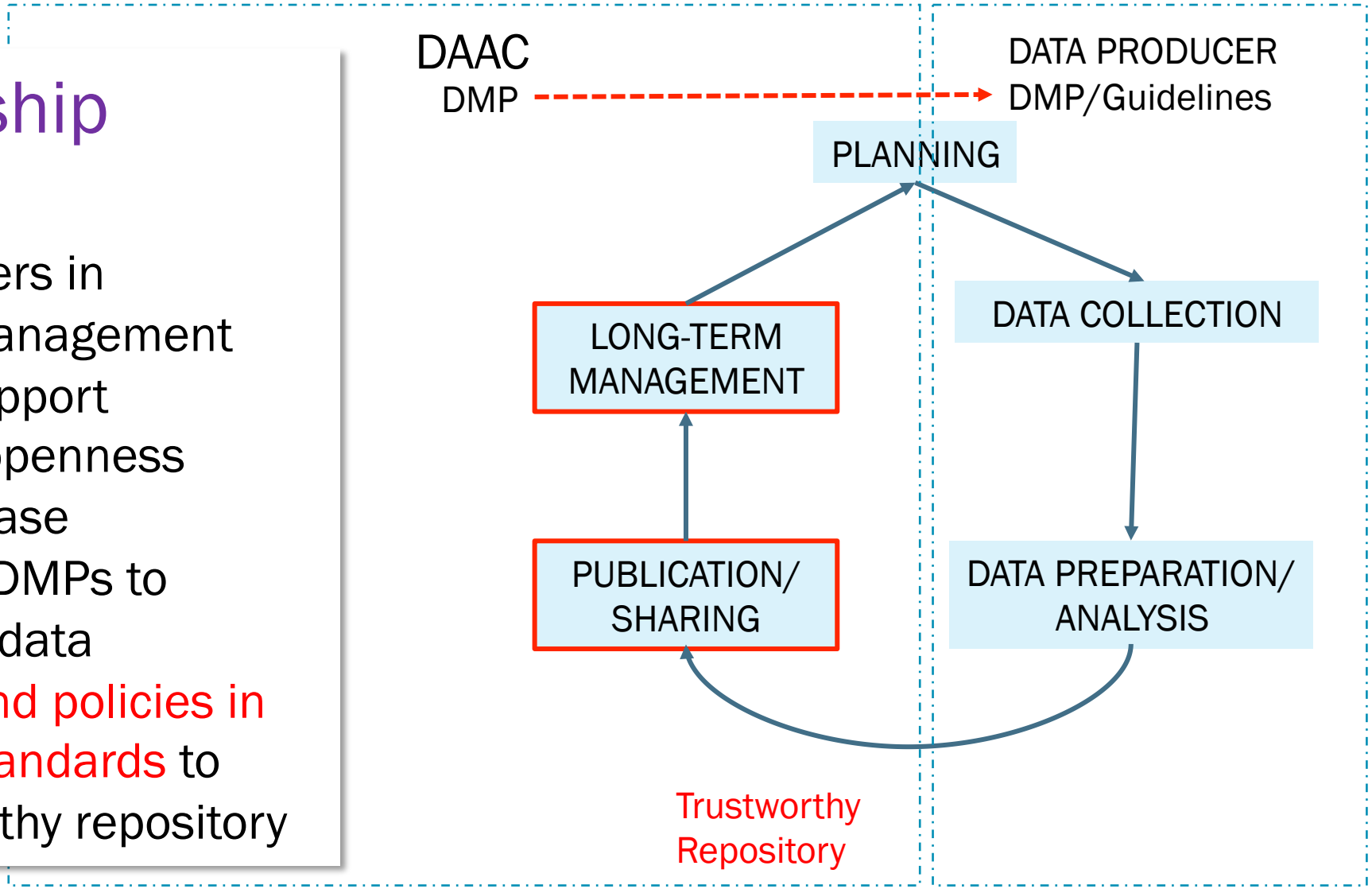
- Can be depicted in multiple ways
- Represents the various stages that data go through during a research project
- Data management occurs across the entire research data lifecycle



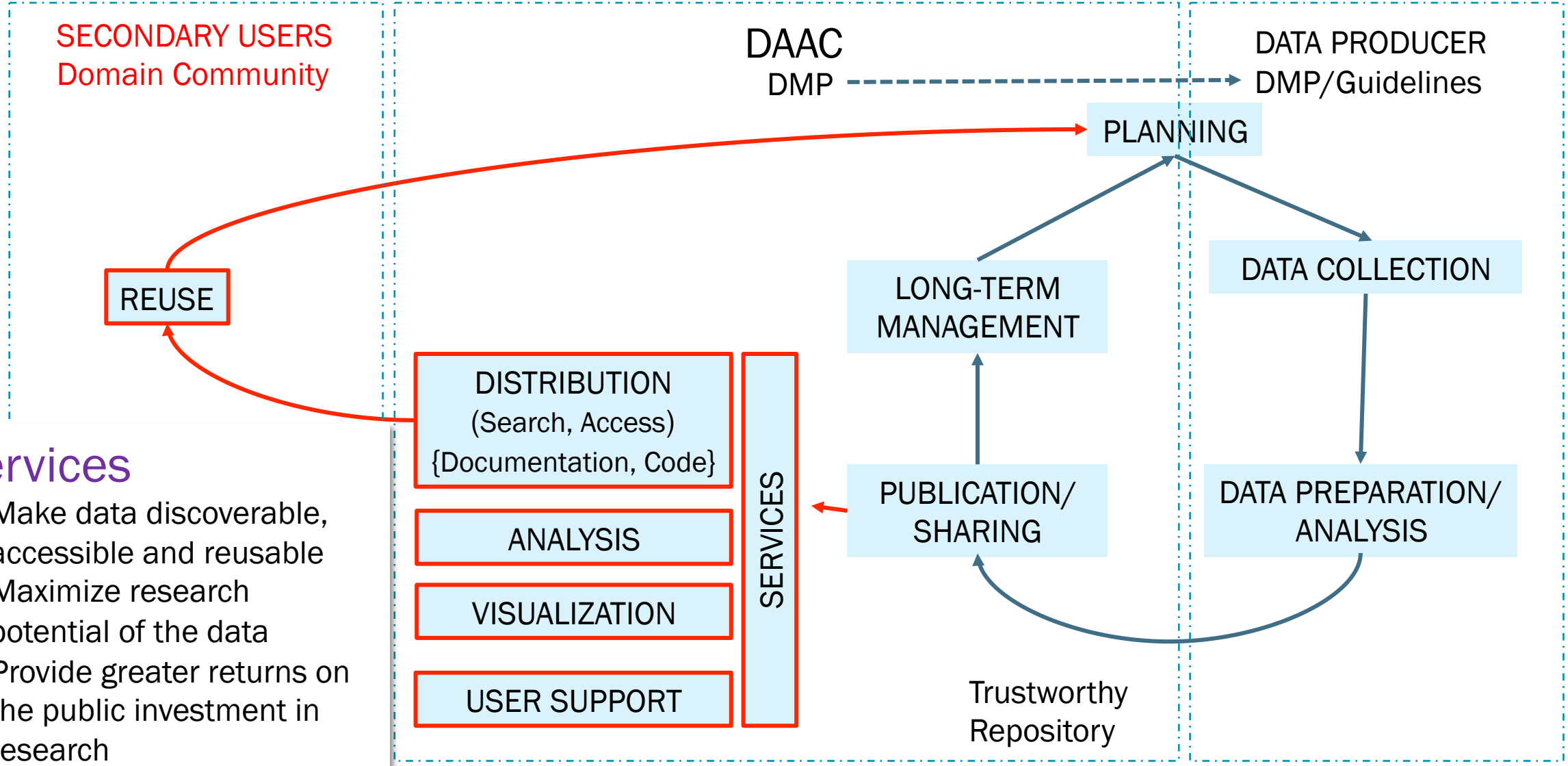
DAAC Role in Supporting Science

Data Stewardship Responsibility

- Assist data producers in developing Data Management Plans (DMPs) to support transparency and openness during research phase
- Use internal DAAC DMPs to efficiently manage data
- Utilize workflows and policies in accordance with standards to serve as a trustworthy repository



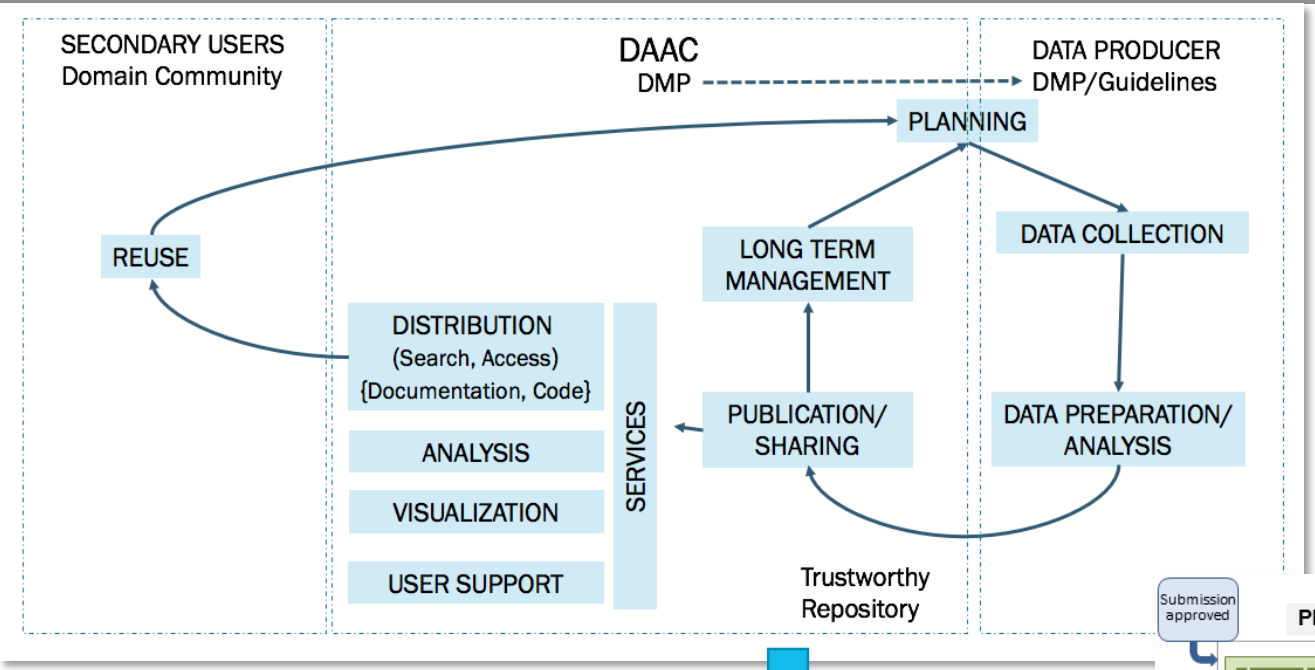
DAAC Role in Supporting Science



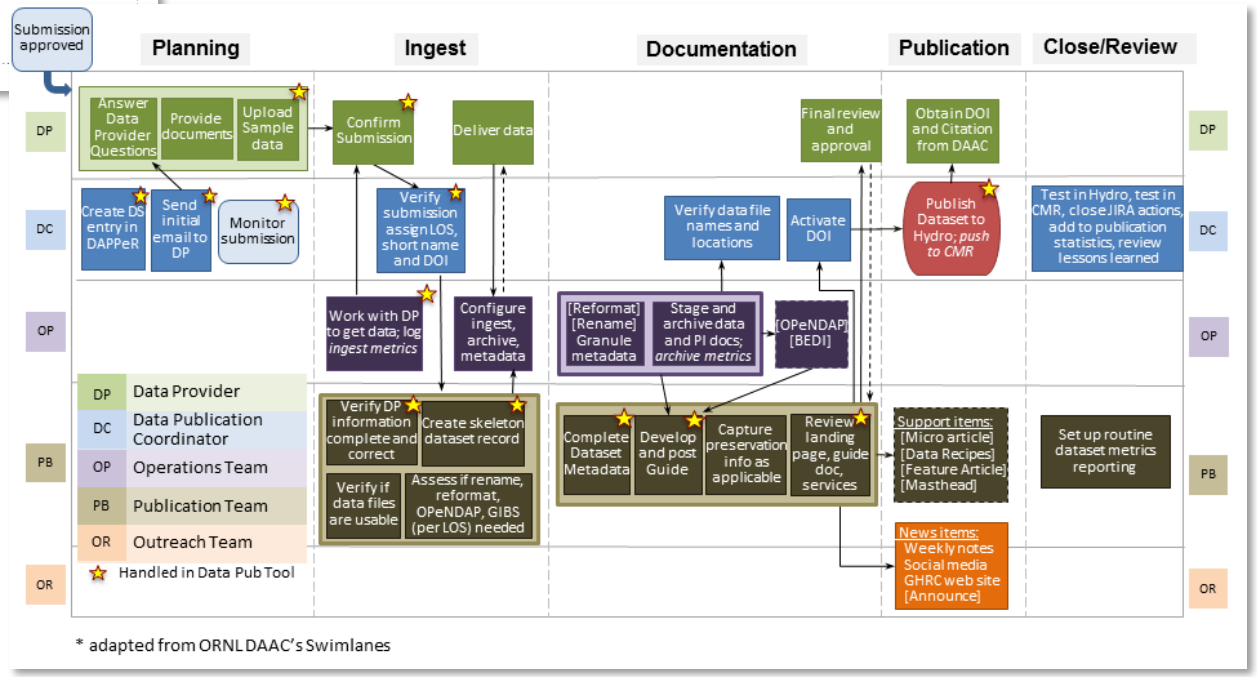
Services

- Make data discoverable, accessible and reusable
- Maximize research potential of the data
- Provide greater returns on the public investment in research

Creating a Common Process for Different Data Sources

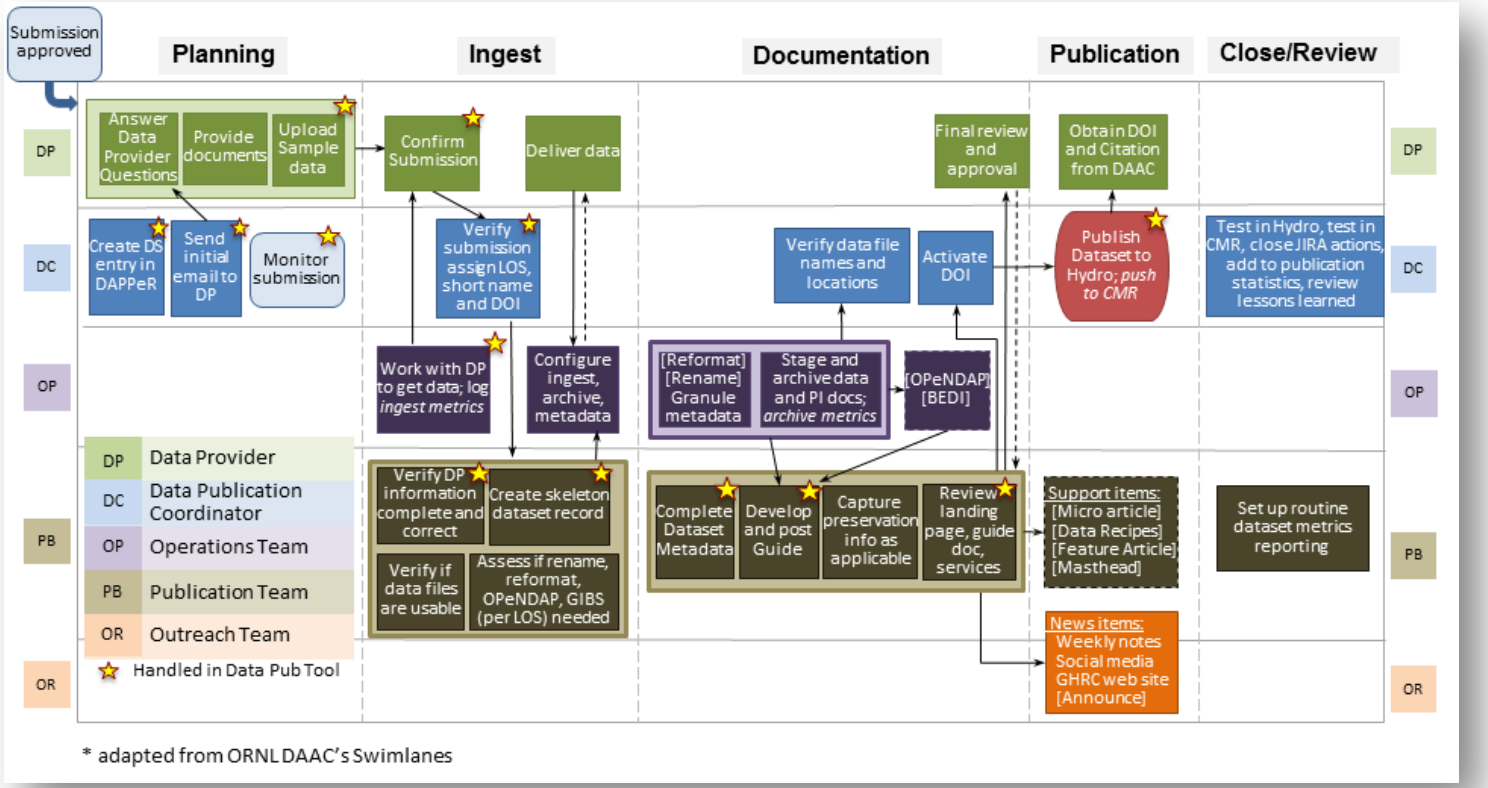


1. Assigned Satellite Mission (LIS)
2. Assigned Field Campaign (GPM-GV)
3. SIPS/MEASURES Program
4. Recommendation from the User Community
   - UWG/ESDIS/HQ Approval



* adapted from ORNL DAAC's Swimlanes

Standardized Internal Process



- Improve efficiency
  - Handle more data using limited resources
- Improve quality assurance steps
- Better serve all our stakeholders
  - User Community, Data Producers, ESDIS Project, and HQ

State of GHRC (FY 17)

Operational Improvements

- Infrastructure upgrades
  - Oracle to Postgres (cost savings)
  - Re-architecture and Database Clean-up (preparing for the future)
- Formalized internal data processes via a tool (DAPPER)
  - Improved data publication efficiency
  - Improved data, metadata, and documentation consistency and quality
  - "Living" process documentation (Kaizen – continuous improvement)
- Metadata Improvements from the ARC Review Team
  - Analysis and Review CMR (ARC) Team tasked with finding metadata issues across all DAACs
  - GHRC should be 100% complete by end of FY17

Community Engagement

- Systematic engagement with Science Teams
- Collaborations within MSFC Earth Science Office
  - MSFC/SPoRT
  - DEVELOP (SERVIR) Program

Key Datasets Published

- Initial ISS LIS datasets
- LIS/OTD Gridded Lightning Climatology Data Collection revisions (10 datasets)
- All core HS3 datasets
- 50% of core OLYMPEX datasets

Looking to the Future (FY18 and Beyond)

- Improved operational processes and data publication rates to augment data portfolio without impacting budget
- Strategic acquisition of data based on HQ/ESDIS directives as well as portfolio gaps
- Cross-DAAC Collaborations
  - Data Analysis and Visualization Portal for Atmospheric Science datasets
    - Define requirements (GESDISC, GHRC)
  - Improved Airborne Data Management (ASDC, GHRC)
  - Common Data Publication Workflow Framework (ORNL, GHRC, NSIDC, GESDISC)
    - Integration of DAPPER, SaUS
- *GHRC data migration to the cloud*
  - Transition to using Cumulus as the ingest, archive tool
  - Provide all GHRC data on AWS cloud by end of FY18

Migration to the Cloud: Background

https://cdn.pixabay.com/photo/2016/06/06/12/33/clouds-1439324_960_720.jpg

http://www.opencodez.com/wp-content/uploads/2017/02/cloud-computing.png

Cloud Context: EOSDIS Architecture

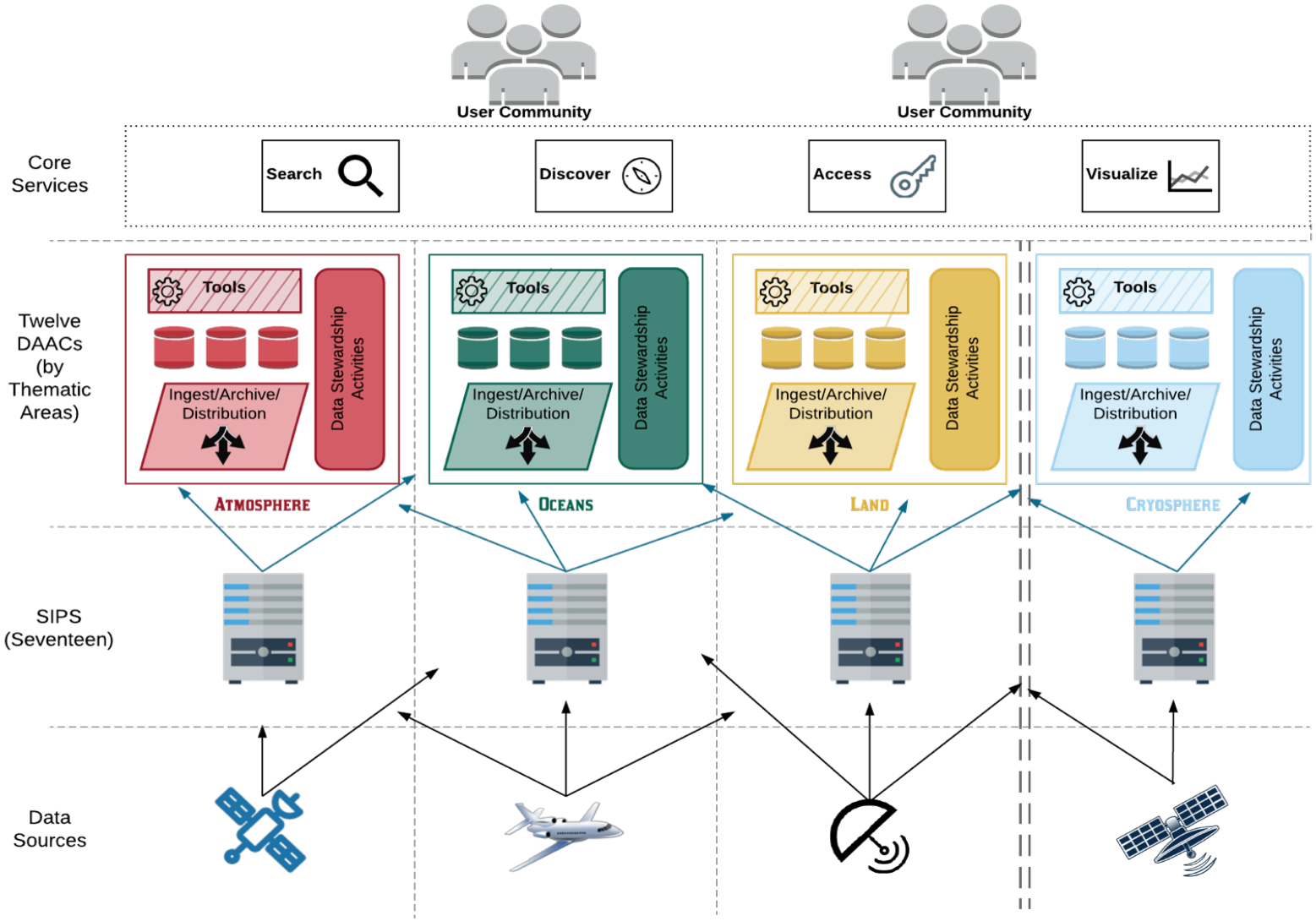
Characteristics (+)

Optimized for archive, search and distribution

Expert user support

Easily add new data products and producers

Predictable



Characteristics (-)

Uneven levels of service and performance

Significant time to coordinate interfaces

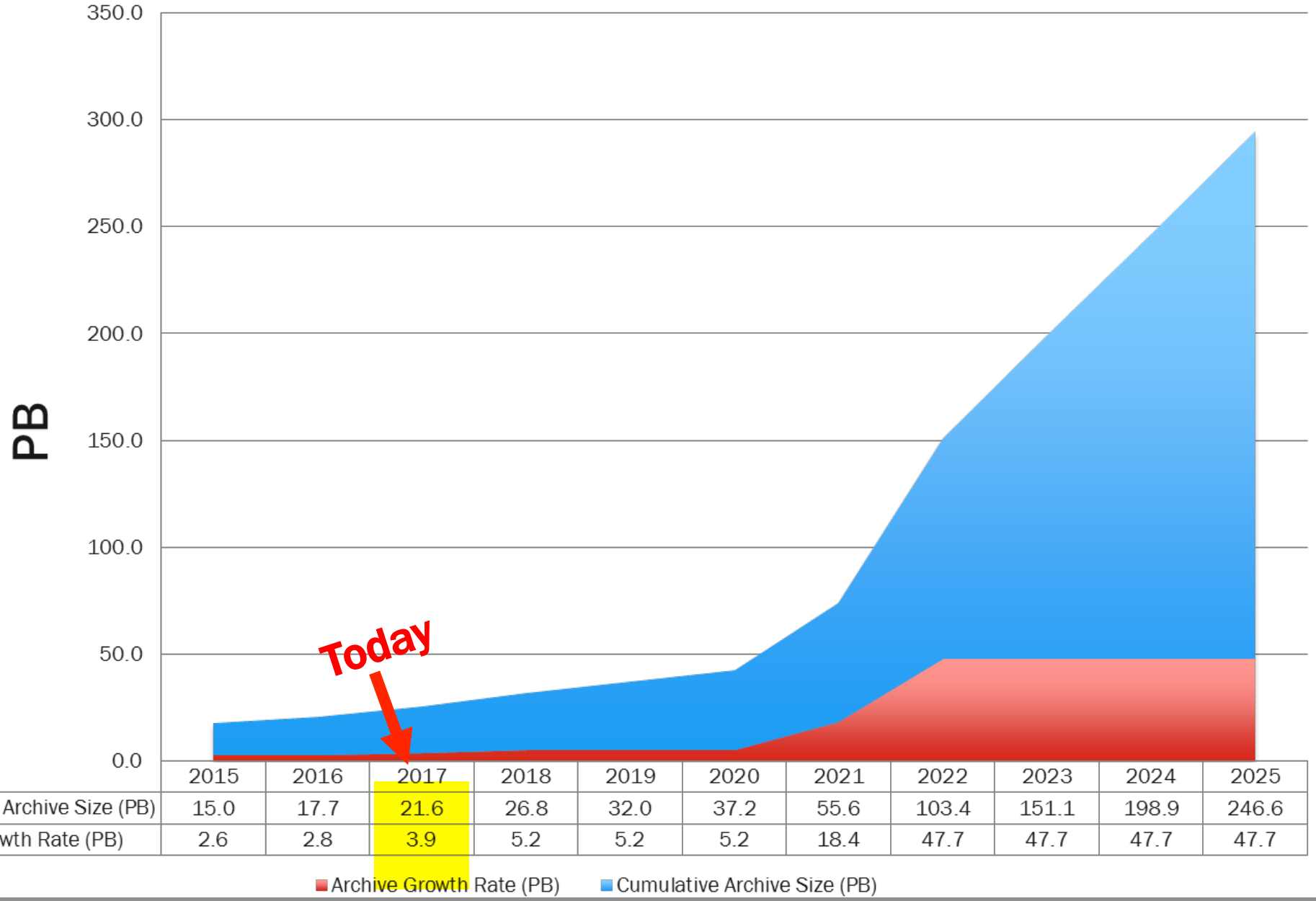
Limited on-demand product generation and end-user processing capabilities

Duplication of infrastructure

Duplication of services and software

Slide source: Kevin Murphy

EOSDIS Data Growth



	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Cumulative Archive Size (PB)	15.0	17.7	21.6	26.8	32.0	37.2	55.6	103.4	151.1	198.9	246.6
Archive Growth Rate (PB)	2.6	2.8	3.9	5.2	5.2	5.2	18.4	47.7	47.7	47.7	47.7

Slide source:
Kevin Murphy

Simplified Cloud Architecture

Characteristics (+)

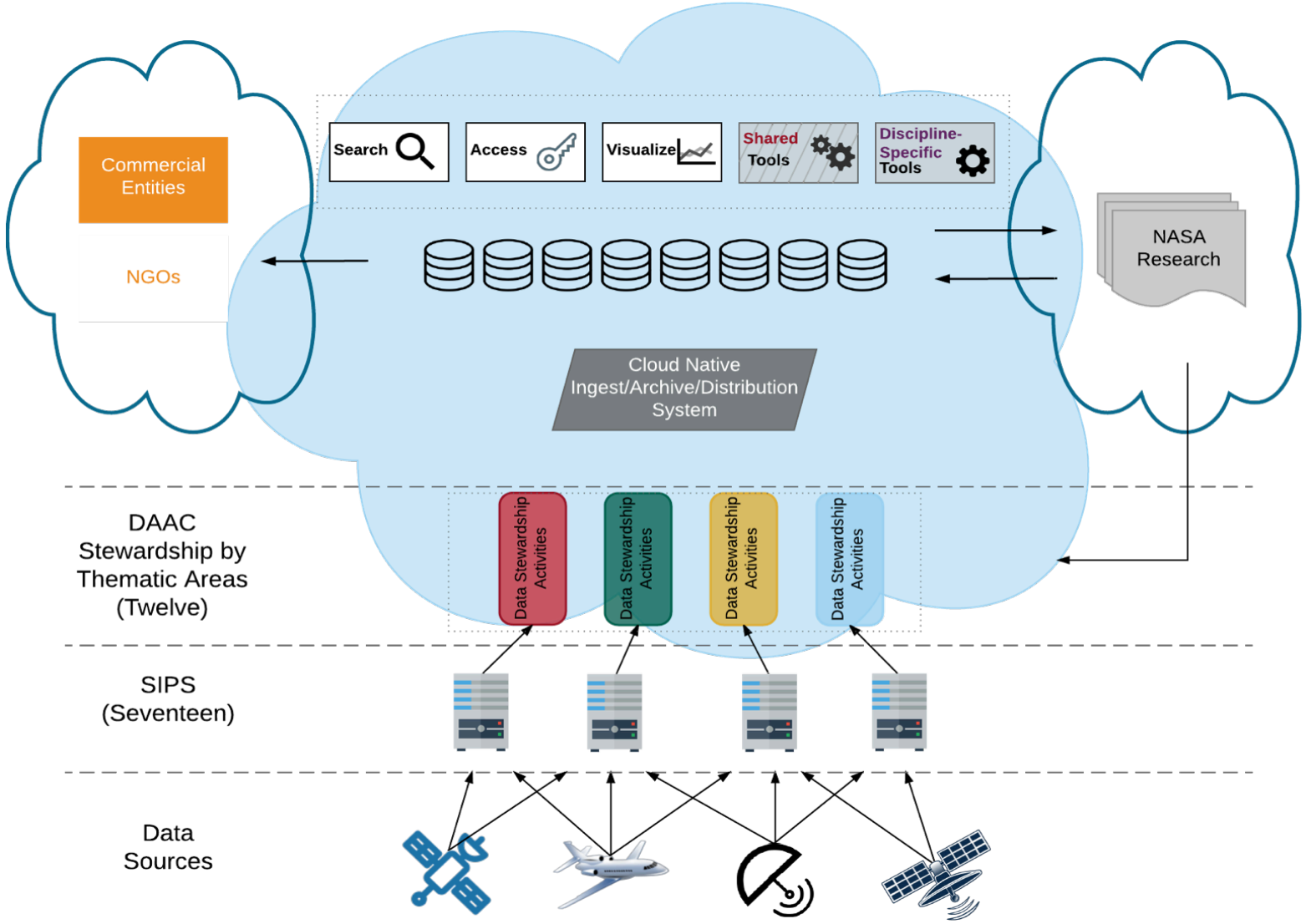
Processing next to data – buy by the yard for anyone

No duplication of data - Data available to all DAACs and users

Expert user support

Easily add new data producers

Apply new technology more easily (e.g. machine learning)



Characteristics (-)

Development coordination

Cost management

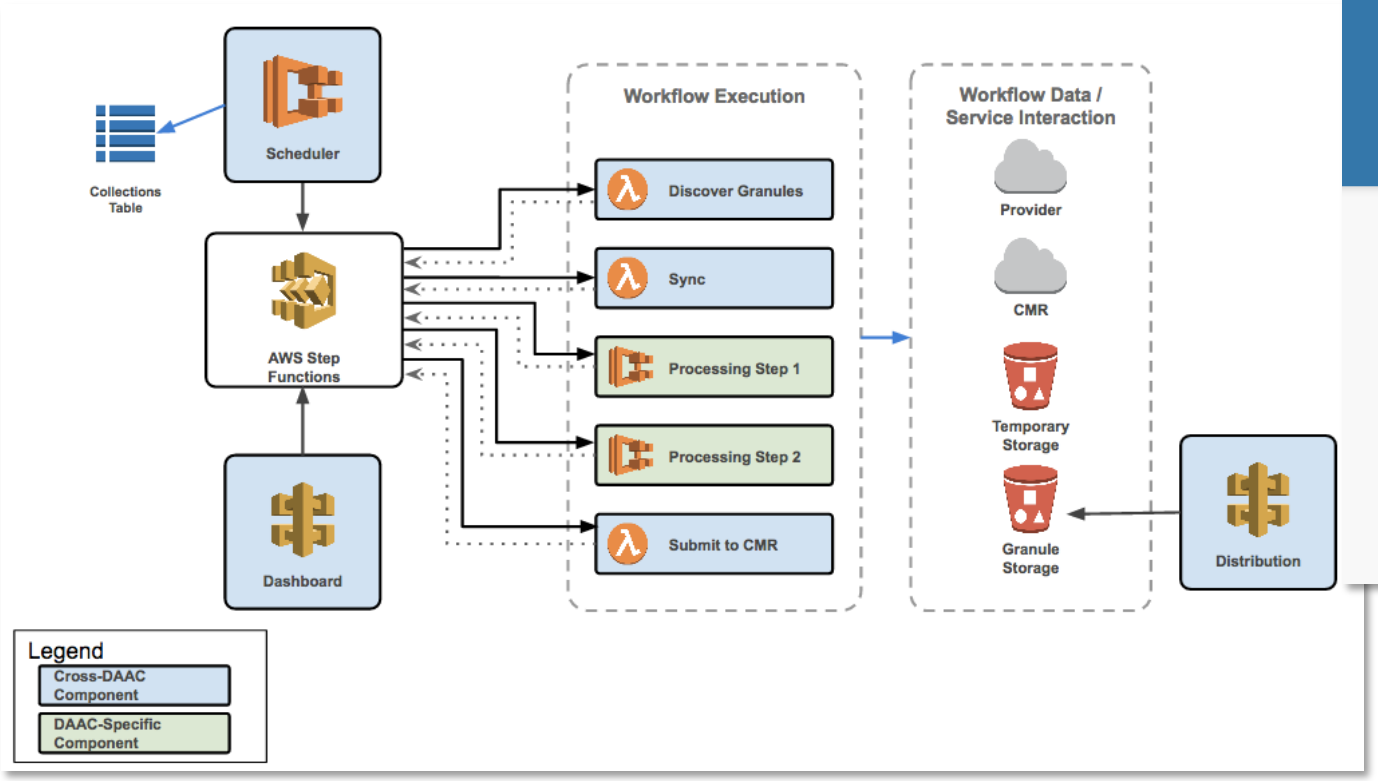
Security, ITAR, SBU, EAR99

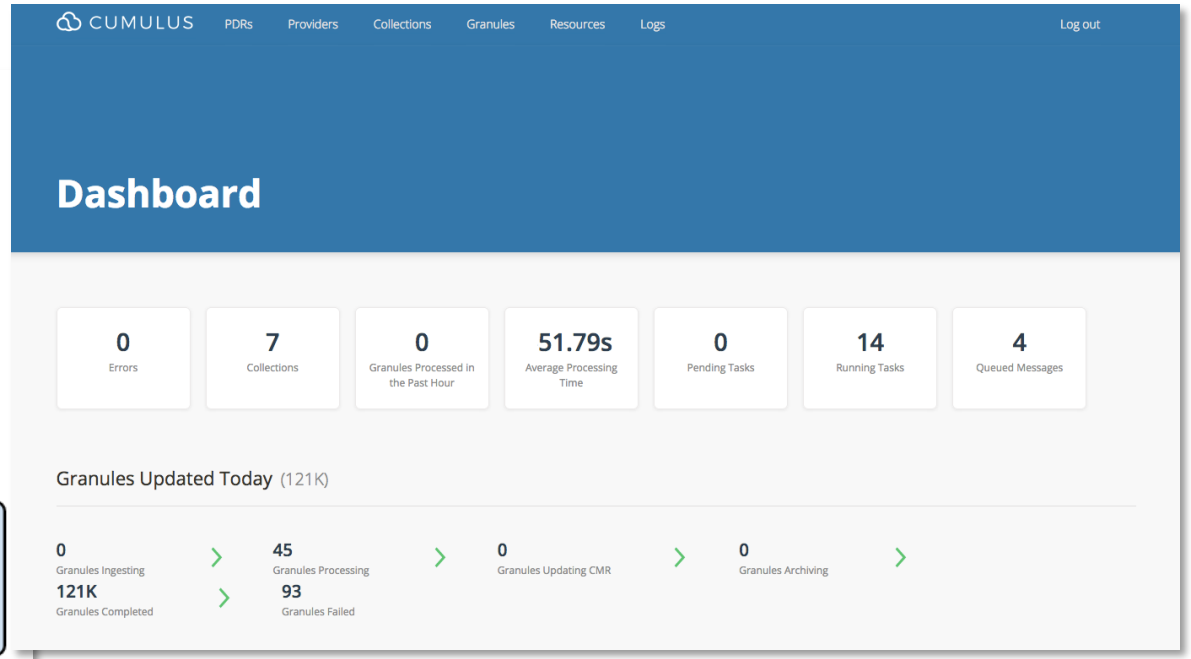
Business processes and skillsets

Vendor lock-in

Slide source: Kevin Murphy

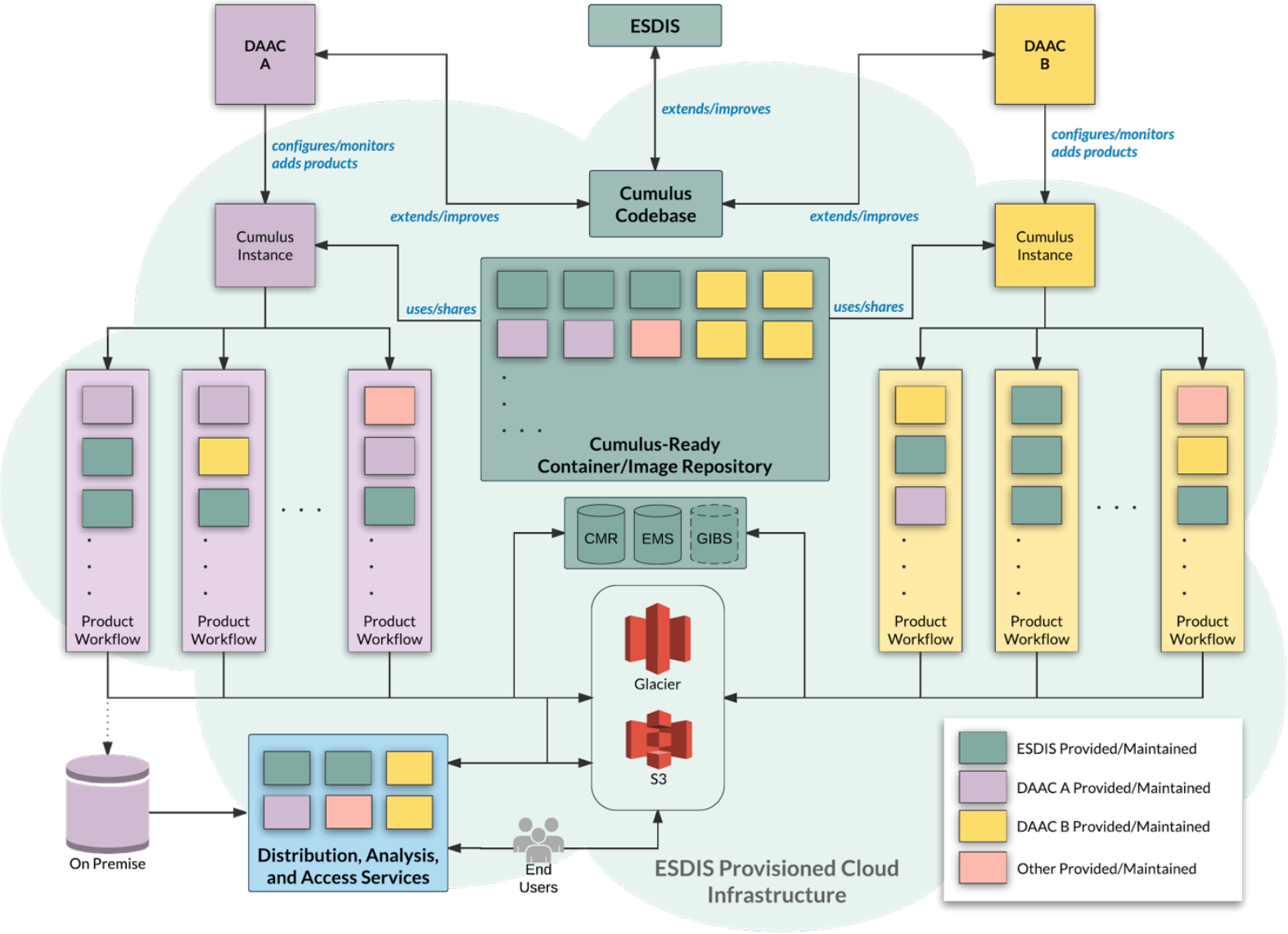
Cumulus Prototype





Cumulus Goal: Design and develop a functional "light weight" data ingest, archive, and distribution “cloud native” framework

Cumulus Deployment



DAACs will have access to the Cumulus code base and will use this code base to deploy and run Cumulus instances on a secure cloud (NGAP). Both the engineering and operations responsibilities fall on the DAAC.

Operational Model

- DAACs are responsible for most of the engineering and all of the operations
- ESDIS controls (allocates), monitors, and pays for the cloud resources
- ESDIS also contributes to the Cumulus core code base

Cumulus (Cloud Design) Trade-Offs

Advantages

- *Maintains thematic data stewardship expertise to support specific community needs*
- *Collocates computations and data to enable new science and application of large-scale analytics*
  - Enables opportunities for other communities to innovate (Commercial, NGO, etc.)
  - Collocation of data supports cross-thematic interdisciplinary science
- *Common framework reduces redundant tools/services*; enables sharing; enforces use of standards, uniform policies and processes
- *Brings to bear economies of scale*
- Enables consistent performance for services

Disadvantages

- *Perceived lack of control over data*
- *Will require a systematic transition*
- *Requires modifications in existing operations*
  - *New governance policies and procedures are needed*
- *Retraining existing staff with a different skillset*
- Cost of Dockerizing existing code
- Possible vendor lock-in

GHRC Cloud goals for FY18

- All forward processing using Cumulus
- All GHRC datasets available on S3
- Serve as pathfinders for other DAACs

GHRC User Working Group Mandate

Primary objectives include but are not limited to:

- Suggesting **improvements to enhance overall user experience** including discovery, access, and usability of data
- Suggesting new **research and development ideas** relevant to GHRC to support product/tool prototyping and generation
- Facilitating **communications with the general user community** and interested members of other communities
- Assisting GHRC in **prioritization and pursuit of new data holdings** within the bounds of budget and ESDIS mission constraints

Questions?

NASA

itsc
INFORMATION TECHNOLOGY AND SYSTEMS CENTER

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

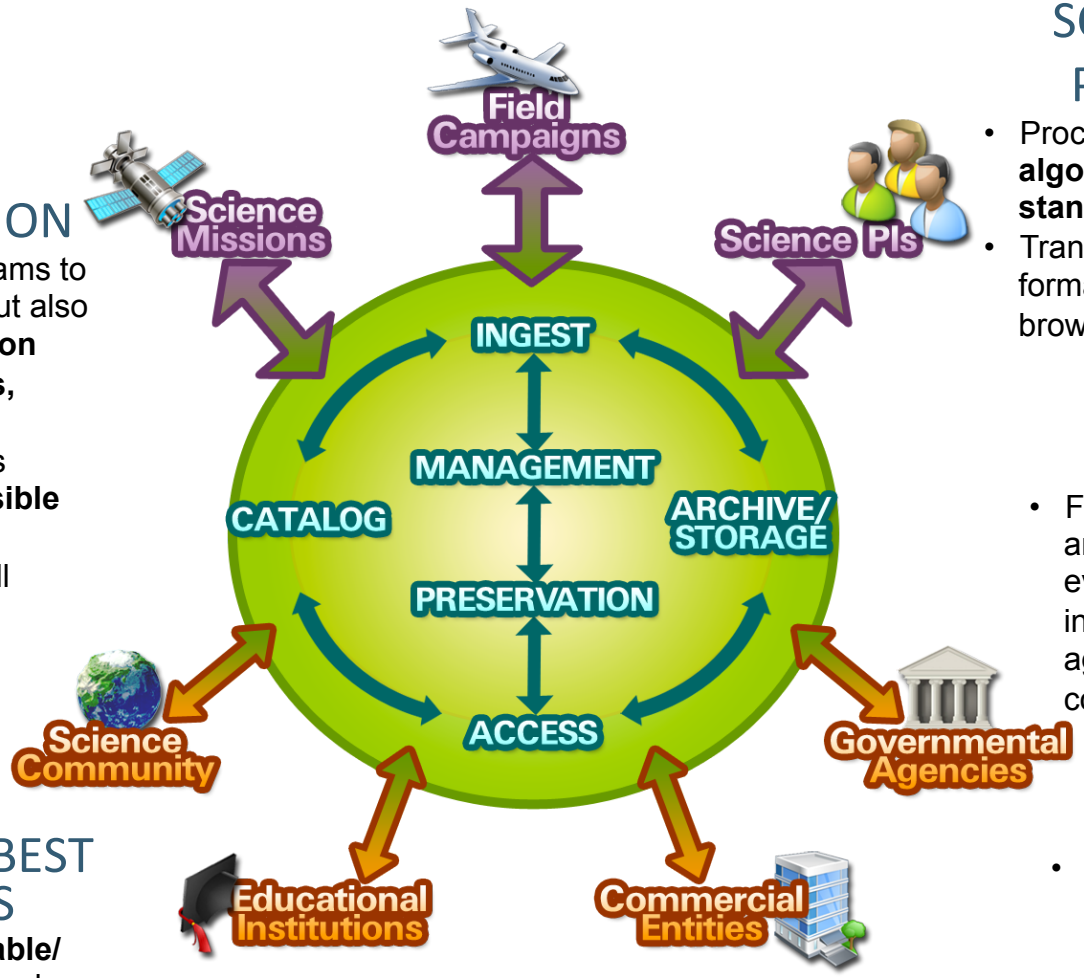
Data Stewardship Services

DOCUMENTATION

- Work with Science Teams to gather not only data but also **all relevant information including documents, papers**
- Ensure that the data is **discoverable**, **accessible** and "**independently understandable**" to all stakeholders without requiring experts

STANDARDS/BEST PRACTICES

- Ensure data is **usable/ interoperable** by tools



SCIENCE DATA PROCESSING

- Process data using **science algorithms to generate standard products**
- Translate data into standard formats, and generate browse imagery

ARCHIVE AND PRESERVATION

- Follow **documented policies** and **engineered** procedures at every step to insure data and information preservation against all reasonable contingencies

PROVENANCE

- Make the preserved data/ information available to all our stakeholder communities with **traceability** to support authenticity

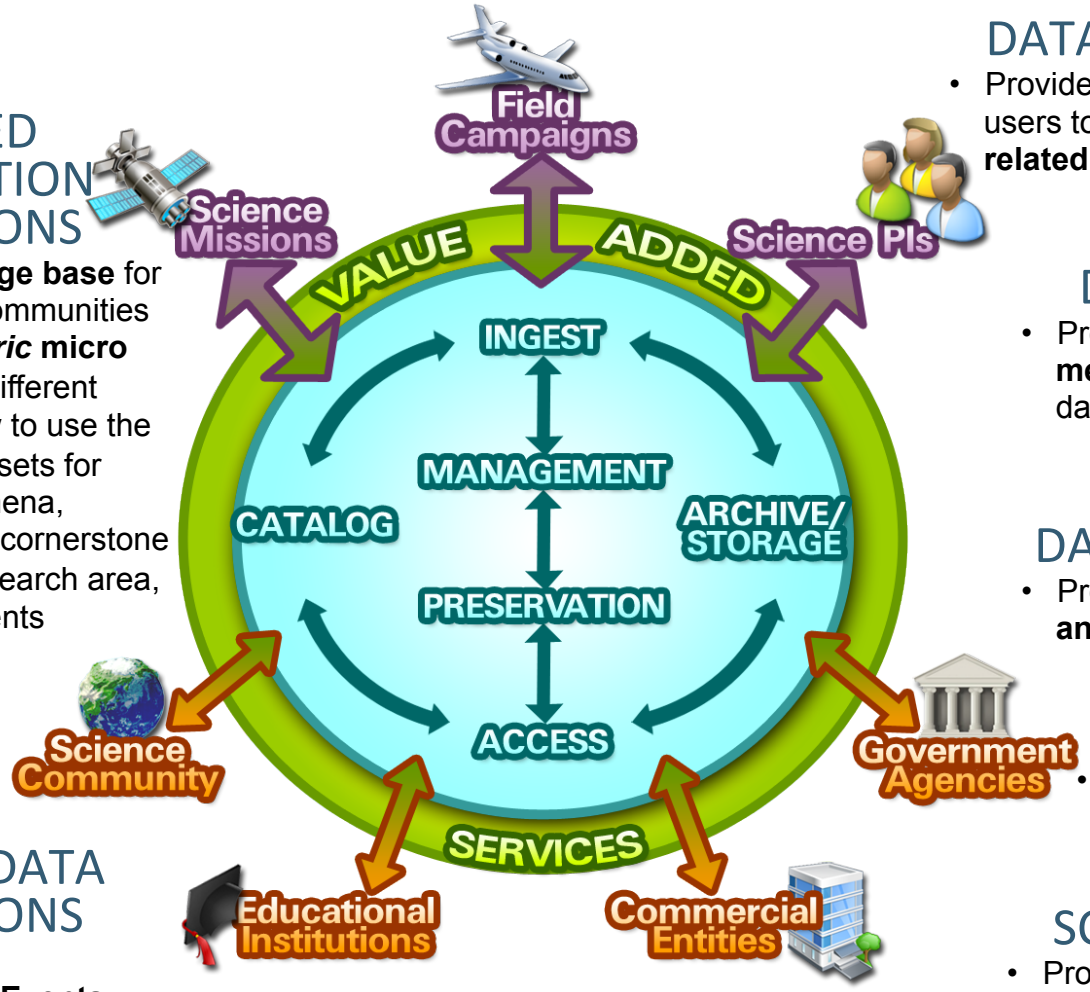
Serves as NASA's Earth science data stewards for scientific, educational, commercial, and governmental communities, with a focus on event/episodic data

Knowledge Augmentation Services

DATA & SCIENCE EXPERTISE

CURATED INFORMATION COLLECTIONS

- **Create a knowledge base** for our stake holder communities
- Provide ***data-centric*** **micro articles** covering different topics such as how to use the data, relevant datasets for studying a phenomena, instrument details, cornerstone publication in a research area, and interesting events



CURATED DATA COLLECTIONS

- Provide **Virtual Collections** for **Events of Interest**

TOOLS EXPERTISE

DATA DISCOVERY

- Provide tools that allow users to **discover data and related information**

DATA ACCESS

- Provide **multiple methods to access** the data

DATA EXPLORATION

- Provide tools to **visualize and analyze** the data

DATA USE

- Provide **data recipes/ code snippets** to allow user to use the data

SCIENCE PORTALS

- Provide **customized portals** for managing **field campaigns** and collecting data

Provides knowledge augmentation services for its datasets to serve stakeholder's needs