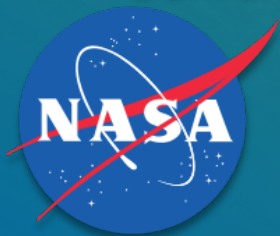# Data Ecosystem for the Joint ESA-NASA Multimission Algorithm and Analysis Platform

Kaylin Bugbee[1], **Aaron Kaulfus[1]**, Jeanné le Roux[1], Aimee Barciauskas[2], Rahul Ramachandran[3], Manil Maskey[3], Amanda Whitehurst[4], Alyssa Harris[2], Anthony Lukach[2], Christopher Lynne[5], Laura Duncanson[6], Marco Lavalle[7]

(1) University of Alabama in Huntsville (2) Development Seed (3) NASA Marshall Space Flight Center (4) ASRC Federal Technical Services (5) NASA Goddard Space Flight Center (6) University of Maryland (7) NASA Jet Propulsion Laboratory

Open Science – "the idea that scientific knowledge of all kinds should be openly shared as early as it is practical in the discovery process" (University of Cambridge)

Science 2.0 – "all scientific culture, including scientific communication, which employs features enabled by the Internet"

What does open science look like in the 2.0 era?
Open Science = Open Data + Open Software + Open Information Sharing

Scientists collaborate through sharing these components & research can begin with any one of these aspects

*Scientists want and need to work more collaboratively and openly in the 2.0 era*

Created by Olena Panasovska
from Noun Project

Global Aboveground Terrestrial Carbon Dynamics Research Community

- Understanding biomass distribution and changes in biomass is critical to understanding the global carbon cycle and global climate change
- Space-based data can improve understanding by providing (Herold at al)
  - Improved geographic coverage
  - Ongoing period of temporal observations
- These data, in combination with field and airborne observations for calibration and validation, can greatly improve the community's understanding
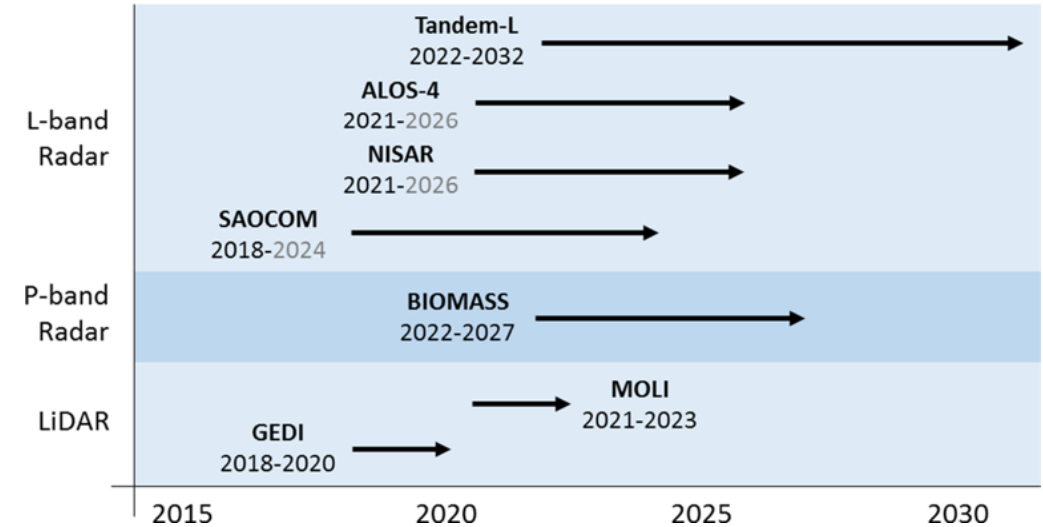
Upcoming space-based missions which may contribute to biomass monitoring.
Image Credit: Herold, M., Carter, S., Avitabile, V. et al. Surv Geophys (2019) 40: 757.
https://doi.org/10.1007/s10712-019-09510-6

Obstacles to Research

- Relevant data are dispersed, making global analysis of data from various sources difficult
- Not all data and algorithms are openly shared
- Data for biomass research are heterogeneous in nature
  - Multiple spatial scales
  - Various observation geometries (footprints vs pixels vs plots)

Researchers want to share algorithms and data easily and quickly since these algorithms are highly customizable

What can we do to support the open data, information and compute needs of this research community?
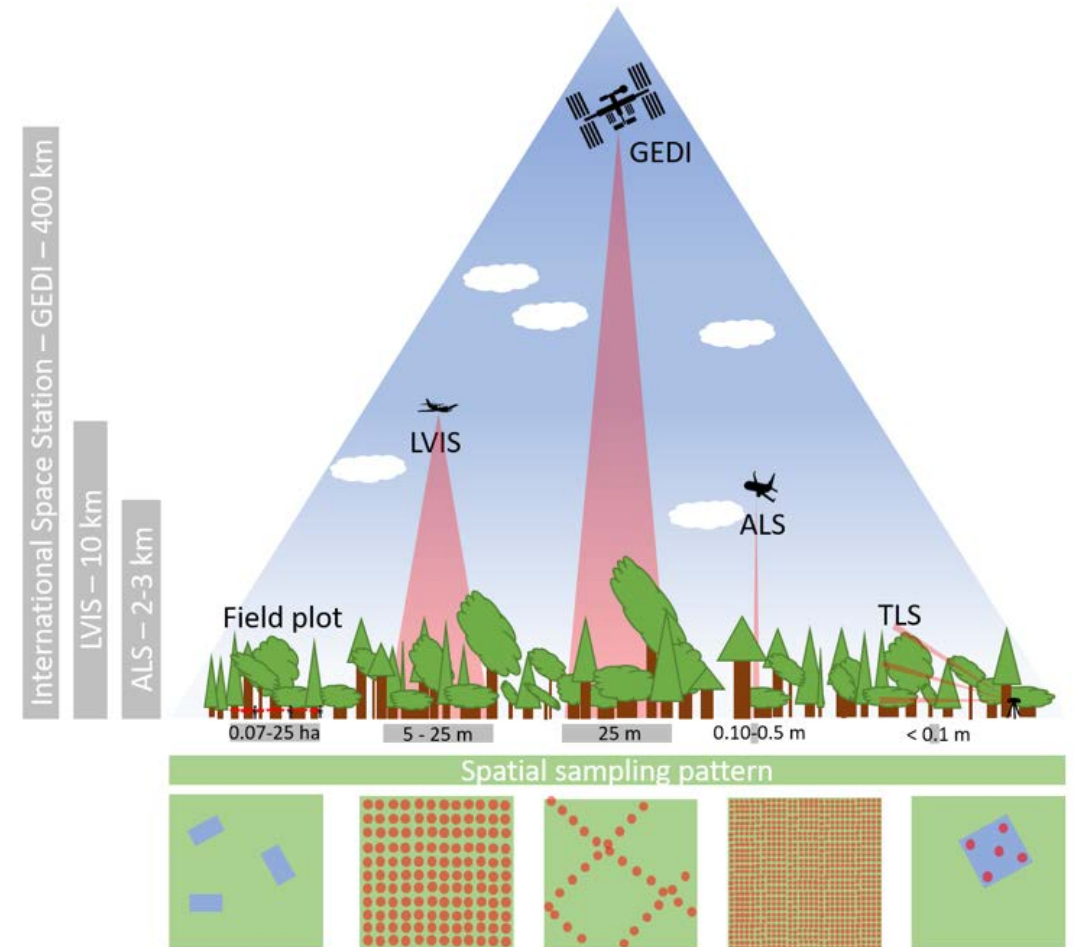
An example of various spatial sampling patterns used by the biomass community.
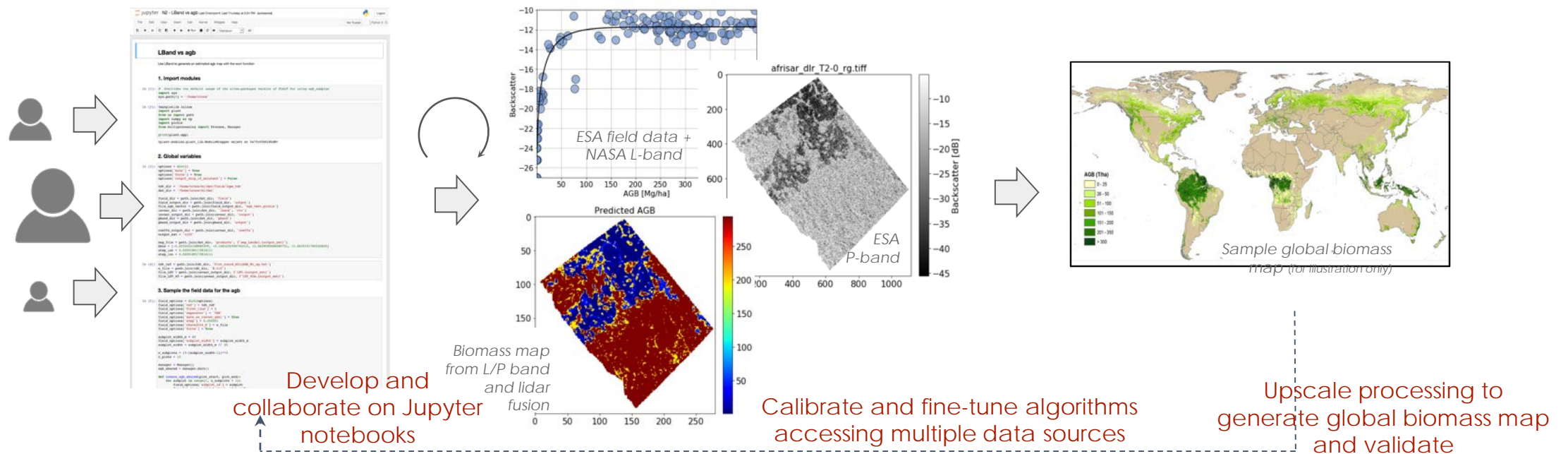Image Credit: https://gedi.umd.edu/science/calibration-validation/

What is MAAP?

- The MAAP, or Multimission Algorithm and Analysis Platform, is a virtual environment dedicated to the unique needs of sharing and processing data from relevant field, airborne and satellite measurements related to ESA and NASA missions
  - Jointly managed by ESA and NASA and accessible to designated ESA and NASA scientists.
  - Initially populated with pre-launch and complimentary data from other projects.
- Science focus is to improve the understanding of global terrestrial carbon dynamics & to support algorithm development
- Addresses a need expressed by the science community to more easily share and process data collected by NASA and ESA activities

MAAP Goals

- **Seamless open access** to airborne, spaceborne and field ESA and NASA data for biomass mapping
- **Ability to upscale user's algorithms** from small regions of interest to global scale
- **Collaboration** on end-to-end calibration and validation of higher-level science products

ESA field data + NASA L-band

afrisar_dlr_T2-0_rg.tiff

ESA P-band

Predicted AGB

Sample global biomass map *(for illustration only)*

Biomass map from L/P band and lidar fusion

Develop and collaborate on Jupyter notebooks

Calibrate and fine-tune algorithms accessing multiple data sources

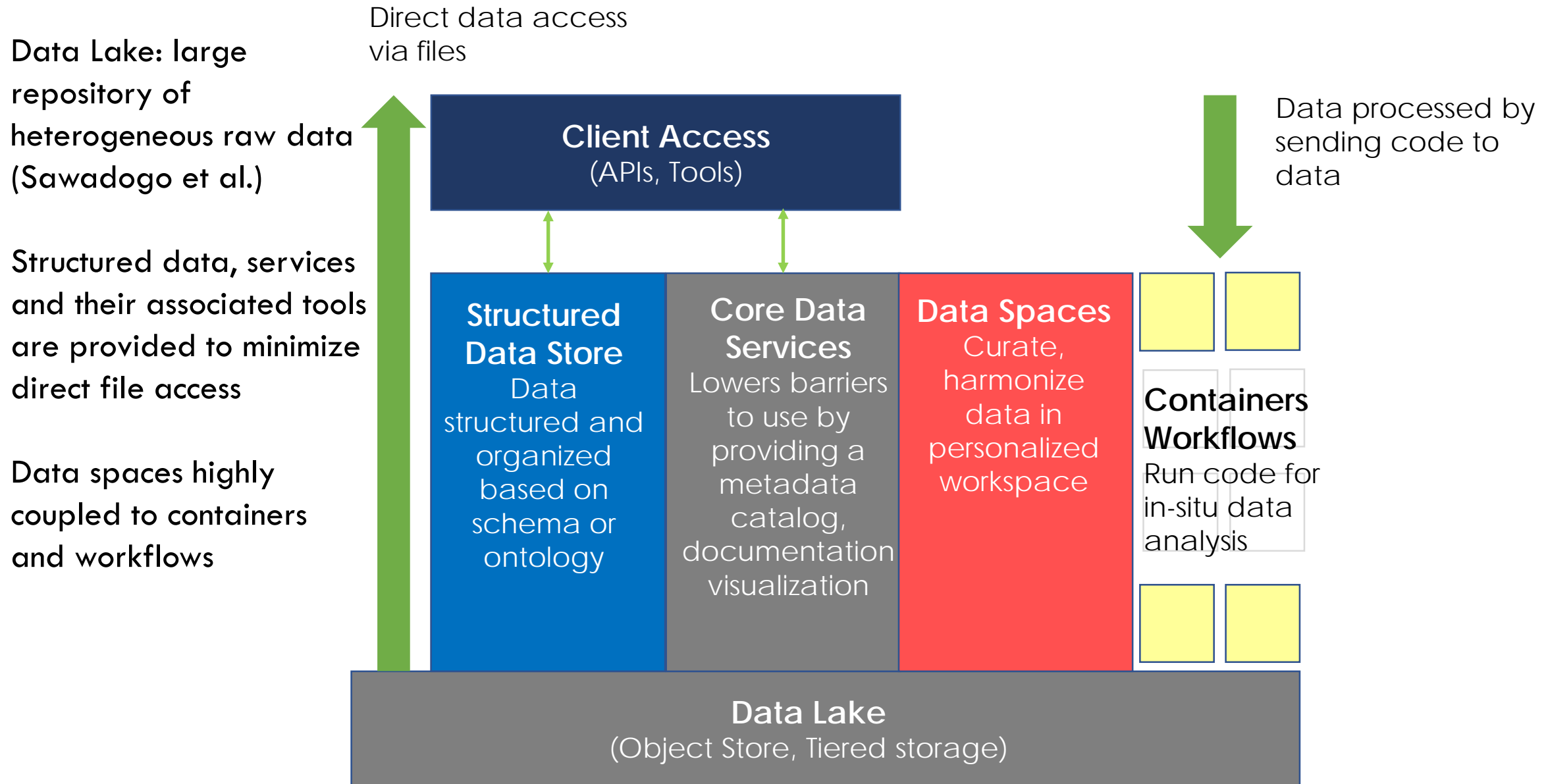Upscale processing to generate global biomass map and validate

NASA MAAP Data Ecosystem

'the people and technologies collecting, handling, and using the data and the interactions between them" (Parsons et al 2011)

MAAP Data Ecosystem Architecture

Direct data access via files

Data Lake: large repository of heterogeneous raw data (Sawadogo et al.)

Structured data, services and their associated tools are provided to minimize direct file access

Data spaces highly coupled to containers and workflows

Data processed by sending code to data

Client Access
(APIs, Tools)

Structured Data Store
Data structured and organized based on schema or ontology

Core Data Services
Lowers barriers to use by providing a metadata catalog, documentation visualization

Data Spaces
Curate, harmonize data in personalized workspace

Containers Workflows
Run code for in-situ data analysis

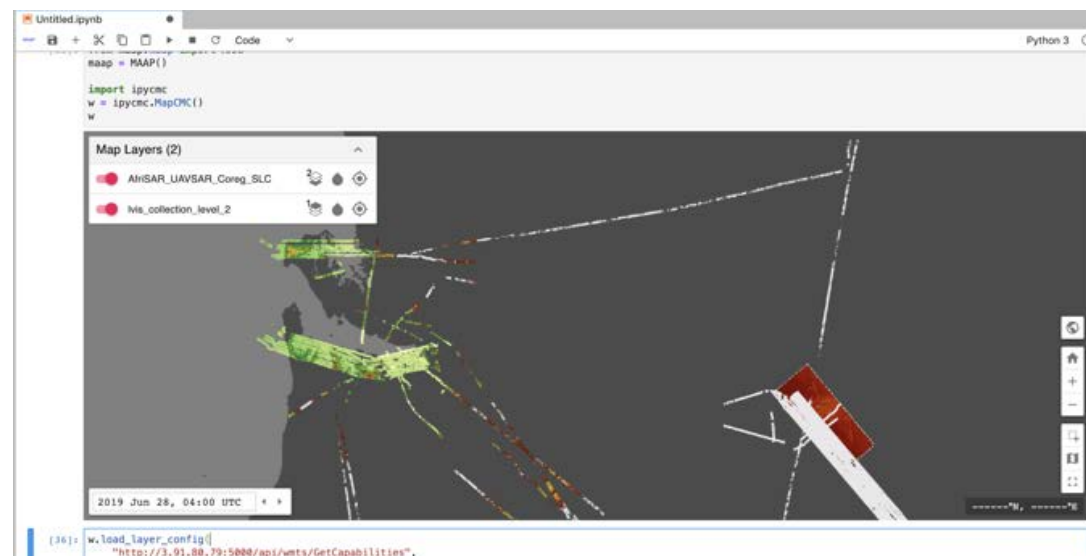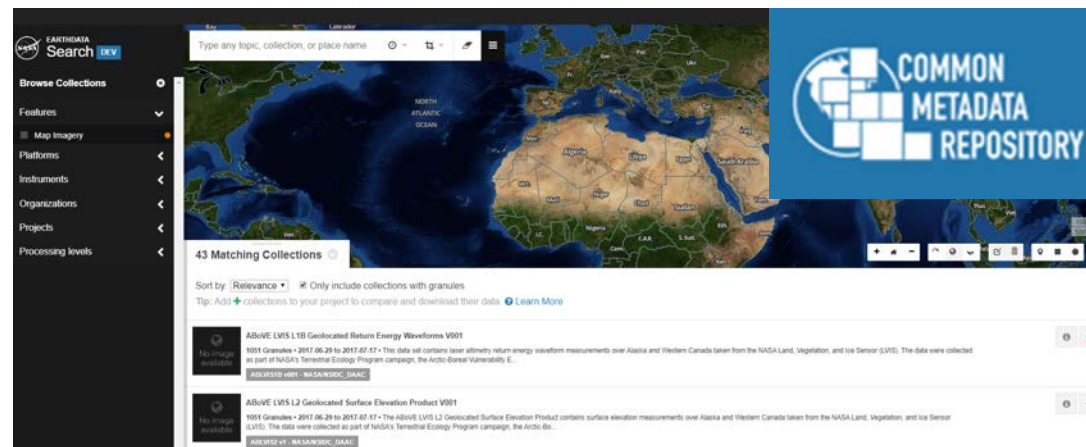## Data Lake
(Object Store, Tiered storage)

MAAP Data Lake

- Data publication process into data lake similar to those followed by NASA's archives
  - Over **30 datasets** ingested for pilot MAAP
  - Data products identified by SMEs
- Reuses open source components developed by NASA's ESDIS project
  - Data ingested into MAAP data lake enabled by Cumulus – *available on the cloud*
  - Metadata Management Tool to author and curate metadata
  - Metadata published to Common Metadata Repository (CMR) to facilitate discovery of objects in the data lake
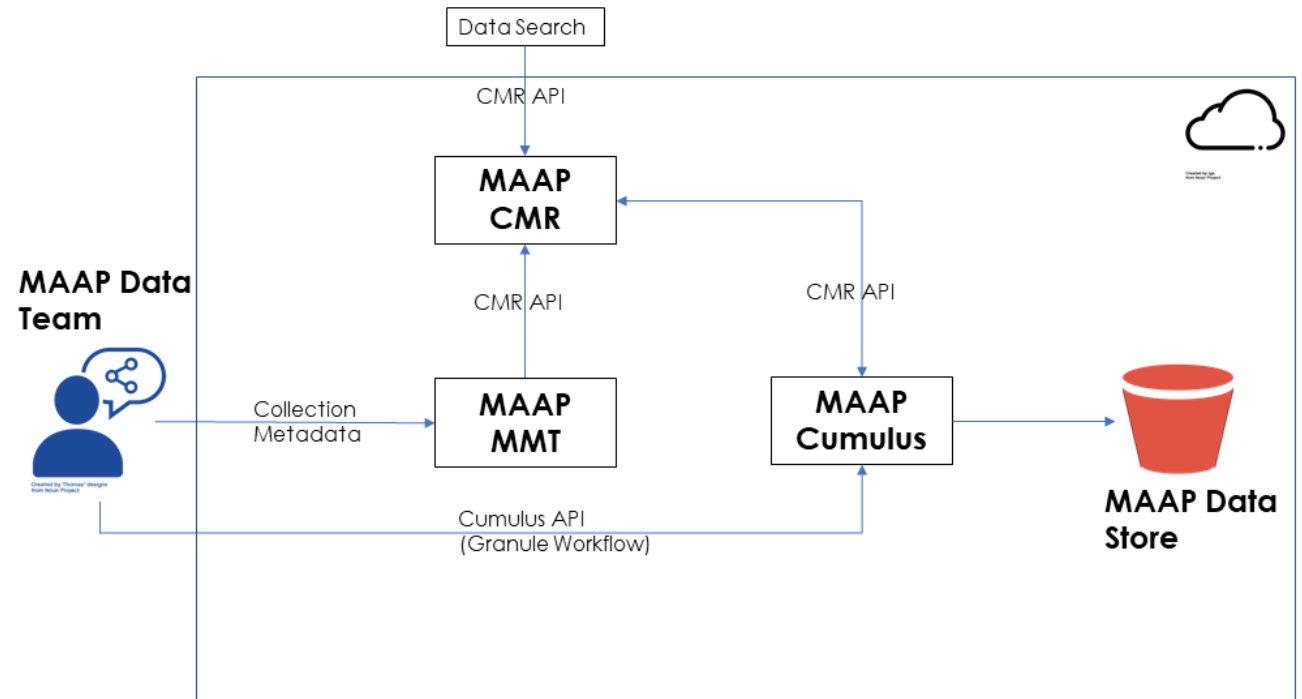
MAAP Core Data Services

- Discovery of biomass relevant data via a centralized catalog
  - ESA and NASA are contributing metadata to a single repository (CMR)
  - Meta(data) may be discovered via an API in the data spaces environment or via the Earthdata Search client
  - Additional metadata information provided to support biomass search needs
- Other core services include:
  - Visualization
    - Fast browse capabilities
  - Documentation
    - Links to data documentation, relevant publications
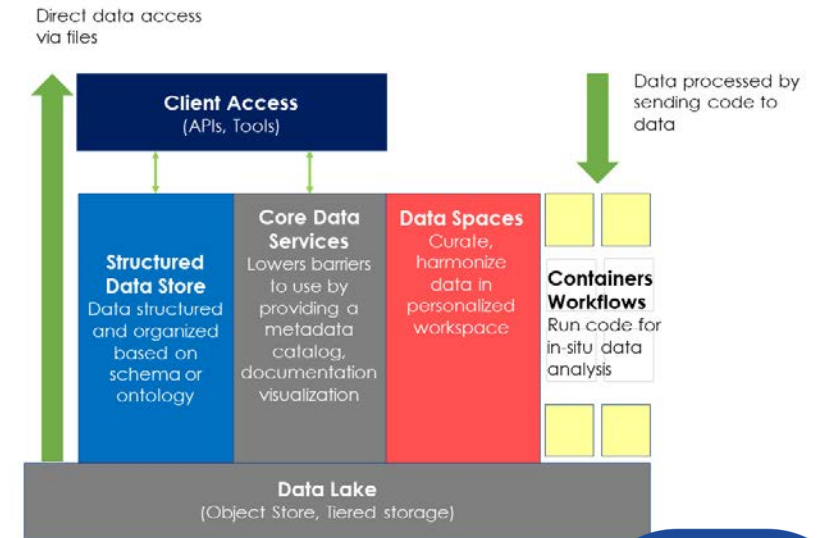
Connecting Data Spaces to Core Services

Users leverage the data lake and private data spaces to run analysis at scale via containers or workflows

MAAP enables quick & easy sharing of data from a user's data space to the broader data ecosystem

- Users can share data with select collaborators
- Share data more broadly from their private workspace to the MAAP data lake and CMR catalog so users can discover it

New Data Source

Requires data stewardship and corresponding core services for user shared data

Direct data access via files

Data processed by sending code to data

Client Access (APIs, Tools)

Structured Data Store Data structured and organized based on schema or ontology

Core Data Services Lowers barriers to use by providing a metadata catalog, documentation visualization

Data Spaces Curate, harmonize data in personalized workspace

Containers Workflows Run code for in-situ data analysis

Data Lake (Object Store, Tiered storage)

Created by Thomas' designs from Noun Project

MAAP Structured Data Store

- To continue to evolve, MAAP will also be exploring data engineering through offering structured data stores
  - Optimized for big data analysis tools
  - Preprocessed to meet goals
  - Focus on improved/efficient performance
- Current: Postgres database for calibration/validation field data
- Future: Analytics Optimized Data Stores (AODS)
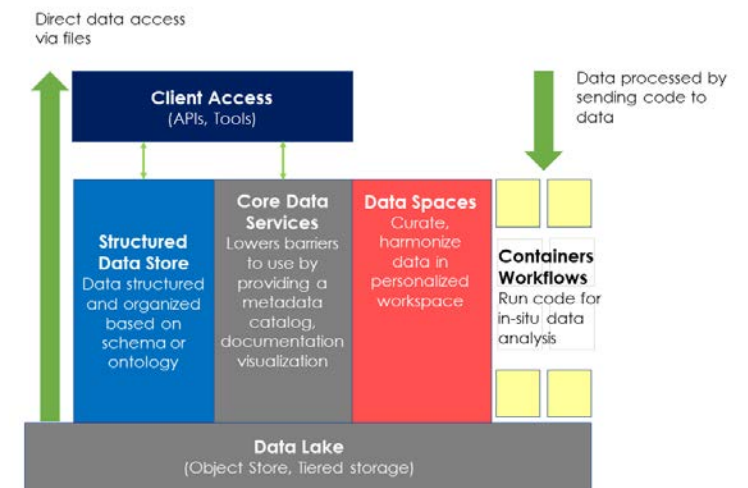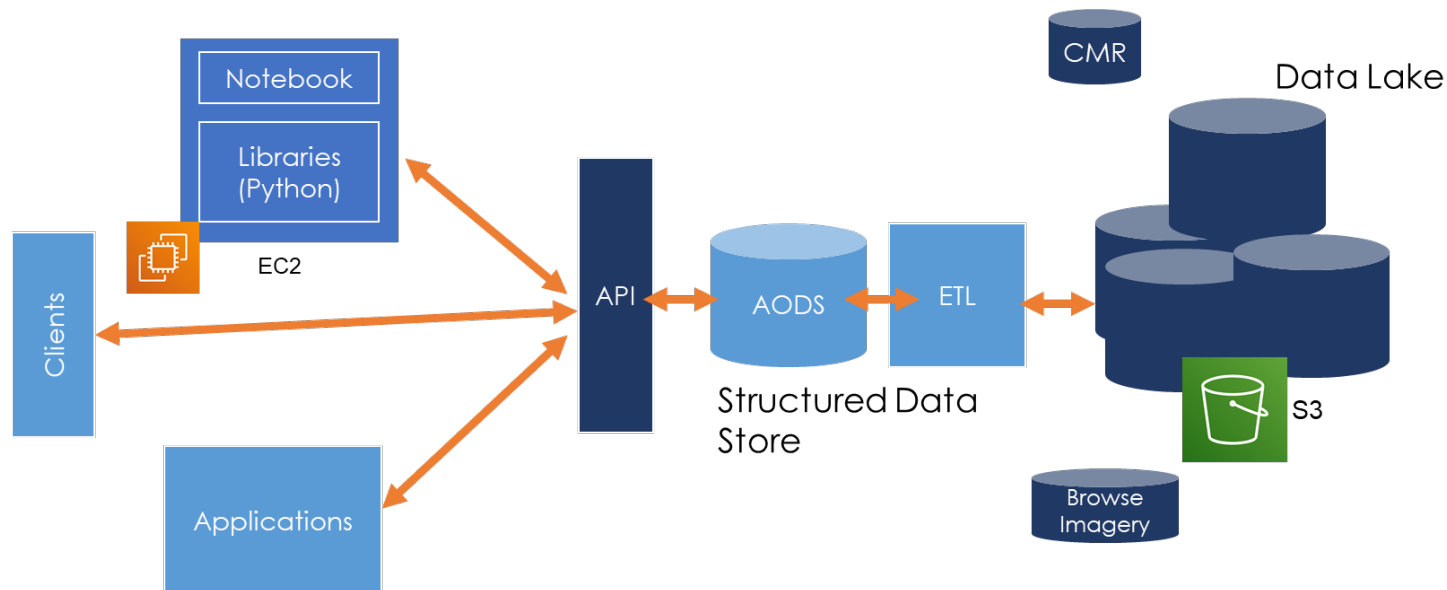
- New core service - encourage sharing of algorithms and software to the broader MAAP community
- Create metadata to support sharing of algorithms and create documentation of algorithm methodology
  - Scientific and mathematical rationale
  - Algorithm Theoretical Basis Documents (ATBDs) or similar
- Create metadata to support sharing of algorithm implementation software
  - Document inputs, outputs, processing steps, computation needed
- Metadata to be stored in MAAP CMR to facilitate discovery

3.2 Mathematical Theory

The orbital characteristics of the proposed NASA HyspIRI (Hyperspectral InfraredImager) satellite were used to obtain the Sun zenith and az-imuthal angles at the time the sensor would fly over a point at (latitude, longitude) = (28.75N, 158.00 W) on *June 21.*

3.2.1 Assumptions

A coupled HydroLight-MODTRAN ocean-atmosphere radiative transfer code was used to compute the in-water and atmospheric radiances both justabove the sea surface and at the top of the atmosphere. (HydroLight is an under-water radia-tive transfer code.)

3.3 Algorithm Input Variables

Name	Unit
total Rayleigh radiance at the TOA	rad

3.4 Algorithm Output Variables

Name	Unit
TOA radiance due to scattering by aerosols only	Rad2
Output Variable	KHz

Discussion

- To be a true data ecosystem that supports open science, MAAP will need to continue to adapt and evolve
  - Extensible to new technologies, changes in data distribution methods
  - New community research needs
- Incentives, recognition or credit in a collaborative platform
  - DOIs for data as a form of recognition
  - Since science can happen using one or multiple parts of the data ecosystem, need to consider giving credit not just for data
- In the 2.0 open science era
  - Technology will continue to evolve and be more widely adopted
  - Data discovery, interoperability, accessibility may look quite different
  - ***Best practices at data repositories will need to continually evolve to support these changing needs and to support the broader open science movement***
    - ***MAAP explores both the technology and data repository best practices***

References

- Bartling S., Friesike S. (2014) Towards Another Scientific Revolution. In: Bartling S., Friesike S. (eds) Opening Science. Springer, Cham. https://doi.org/10.1007/978-3-319-00026-8_1
- Herold, M., Carter, S., Avitabile, V. et al. Surv Geophys (2019) 40: 757. https://doi.org/10.1007/s10712-019-09510-6
- Parsons, M., Oystein, G., LeDrew, E., De Bruin, T., Danis, B., Tomlinson, S. and D. Carlson. (2011). A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science. Journal of Information Science, Vol 37:6.
- Sawadogo, Pegdwendé Nicolas & Scholly, Etienne & Favre, Cécile & Ferey, Eric & Loudcher, Sabine & Darmont, Jérôme. (2019). Metadata Systems for Data Lakes: Models and Features.
- Scholarly Communication – Open Research. University of Cambridge. https://osc.cam.ac.uk/open-research. Accessed 11/20/19

Questions?

Contact:
Kaylin.m.Bugbee@nasa.gov
ak0033@uah.edu

Special thanks to all our collaborators on this project: Laura Jewell, George Chang, Hook Hua, Björn Fromnknecht, Clement Albinet, Valerie Dixon, Kevin Murphy