Building A Data Ecosystem:
A New Data Stewardship Paradigm For The Multi-Mission Algorithm and Analysis Platform (MAAP)

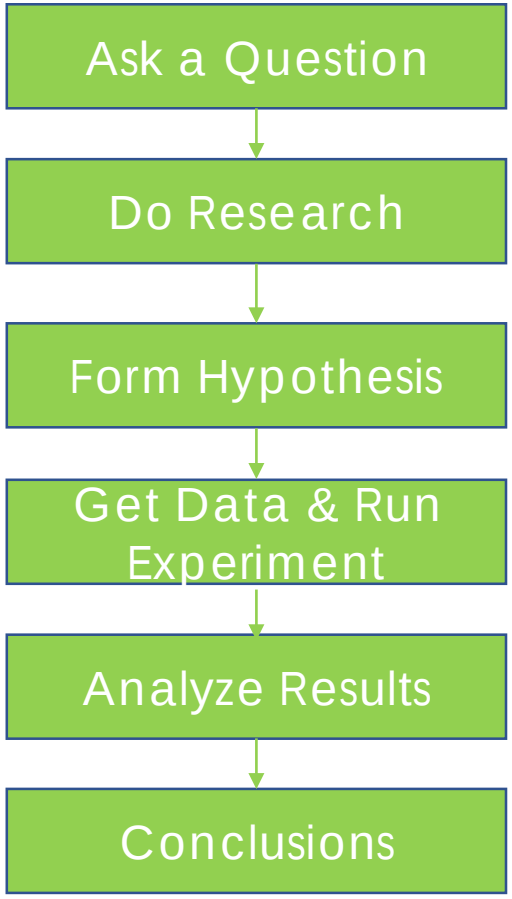
Kaylin Bugbee[1], Manil Maskey[2], Aimee Barciauskas[3], Rahul Ramachandran[2], Aaron Kaulfus[1], Jeanné le Roux[1], Jeffrey Miller[1], Iksha Gurung[1],Amanda Whitehurst[4], Chris Lynnes[5]

(1) University of Alabama in Huntsville (2) NASA Marshall Space Flight Center (3)DevelopmentSeed (4) ASRC Federal Technical Services (5) NASA Goddard Space Flight Center

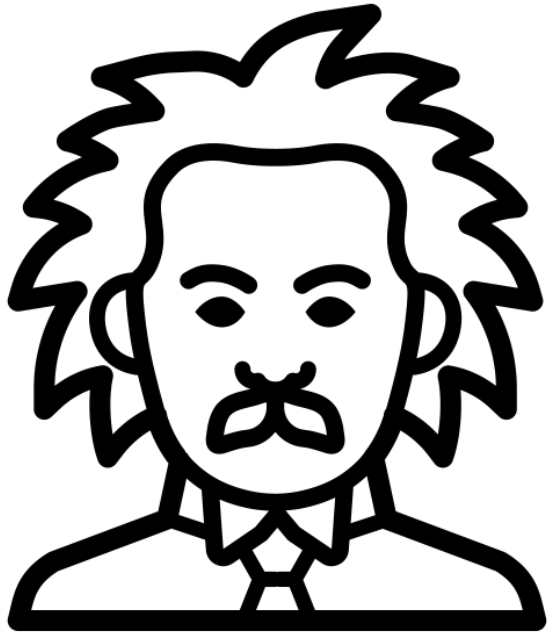
IMPACT
Interagency Implementation and Advanced Concepts Team

Introduction: How We Think About Science

Traditionally taught that scientific research is a linear process
Reinforced by the **scientific method**



Scientist conducting research **alone** in a lab



Created by Maxim Kulikov from Noun Project

Current Data Infrastructure

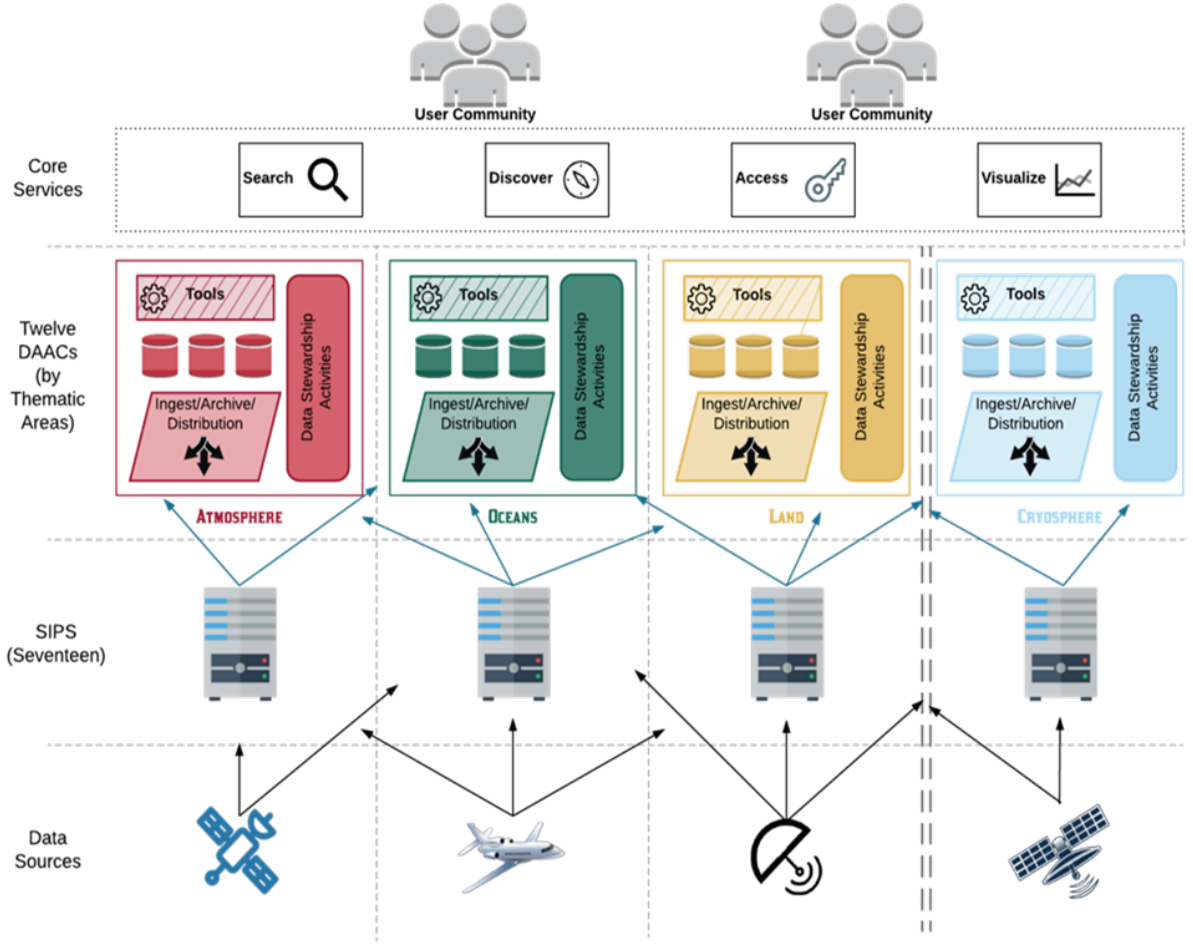
Data infrastructure is static, linear as well

Due to previous infrastructure constraints

Siloed components that require coordination

On demand generation of products is limited, not dynamic

Each component has different capabilities, services



Current NASA EOSDIS Architecture

Current Data Stewardship Model

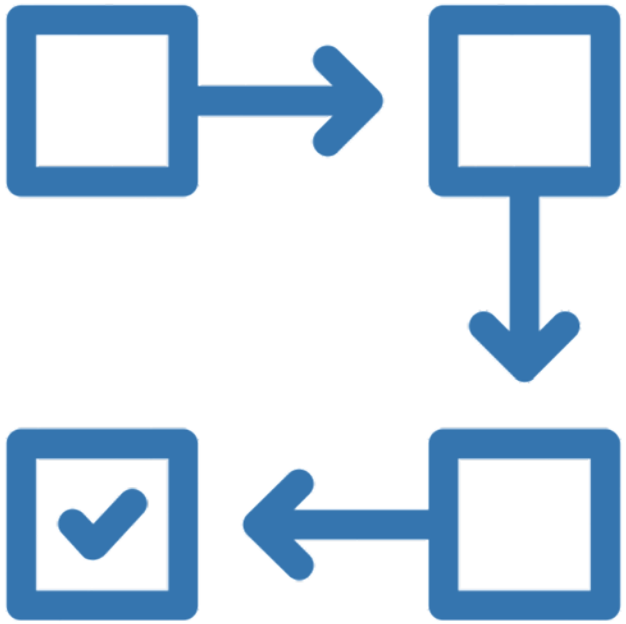
Data stewardship model is also narrow, linear

Data publication metaphor is limiting (Parsons and Fox, 2013)

- Discrete, well-described data sets
- Focus on preservation – less attention on issues of latency, rapid versioning, reprocessing, computational demands

Focus is on the data as the end goal with all other components or information supporting the data

- Software or code
- Experimental data
- Methods



Created by Adrien Coquet from Noun Project

Reality of Scientific Research

Scientific research is non-linear, interconnected

Open Science = Open Data + Open Software + Open Information Sharing

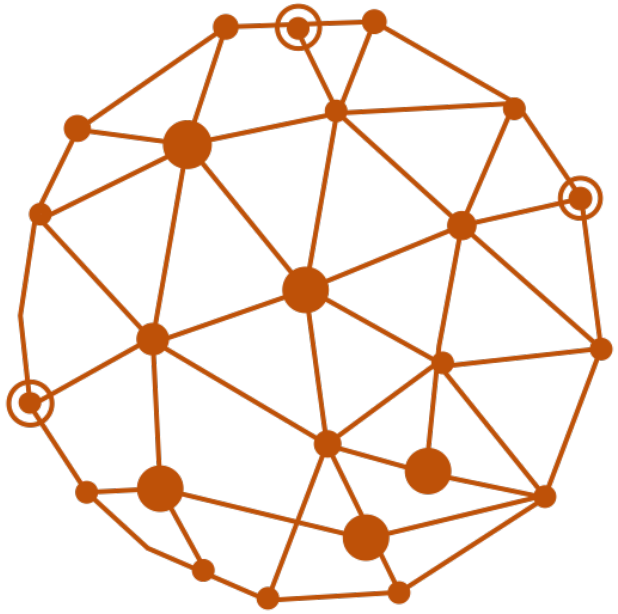
Scientists collaborate through sharing these components

Research can begin with any one of these

Technological innovations make collaboration easier

- Easier to share data and code
- Cloud computing and other technologies for data intensive computing

We need a new data system and management paradigm that reflects the scientific paradigm



Created by Olena Panasovska
from Noun Project

New Data Model: Data Ecosystem

Ecosystems:

- Are adaptive in complex ways
- Have niches that evolve
- Are healthy when diverse
- Evolve in response to components

Data ecosystem:

- ‘the people and technologies collecting, handling, and using the data and the interactions between them” (Parsons et al 2011)

- Reflects ever evolving, collaborative nature of scientific research

NASA recognizes the value of this approach

- **Prototyping a new platform to explore further**

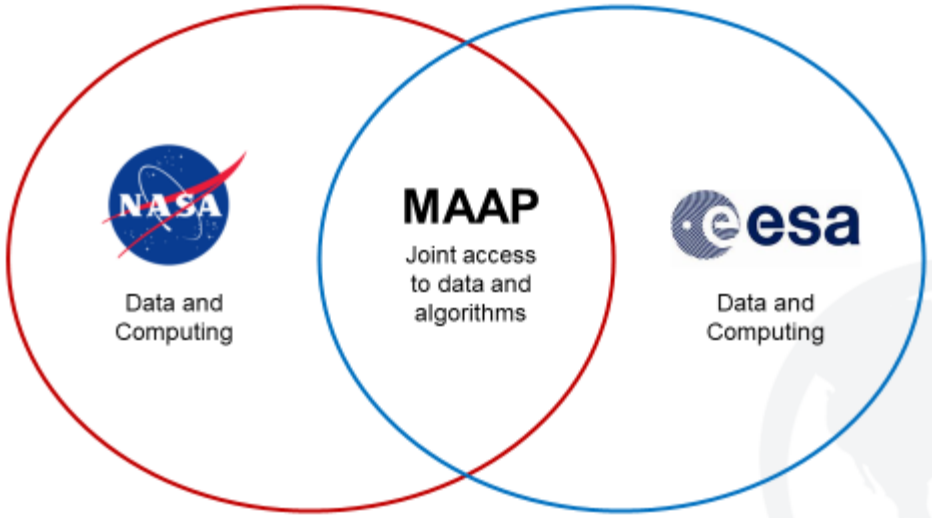
Created by Nithinan Tatah from Noun Project

What is MAAP?

- The MAAP 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

- The MAAP's long term vision:
  - Clearly connect data, algorithms, software and results to support the global aboveground terrestrial carbon dynamics research community
  - Encourage community collaboration by
    - Providing collaborative work environments
    - Making it easy to share data, algorithms and software to collaborators and the MAAP
  - Encourage interoperability between ESA and NASA by providing joint access
  - ***Build a cloud based data system to support open science***
  - ***Data ecosystem approach makes this possible***



NASA MAAP Data Ecosystem

MAAP Data Ecosystem Conceptual Vision

MAAP ARDs and AODS

Standard MAAP Data

User Shared Data

User Shared Software and Algorithms

MAAP Data Ecosystem Conceptual Vision

MAAP ARDs and AODS

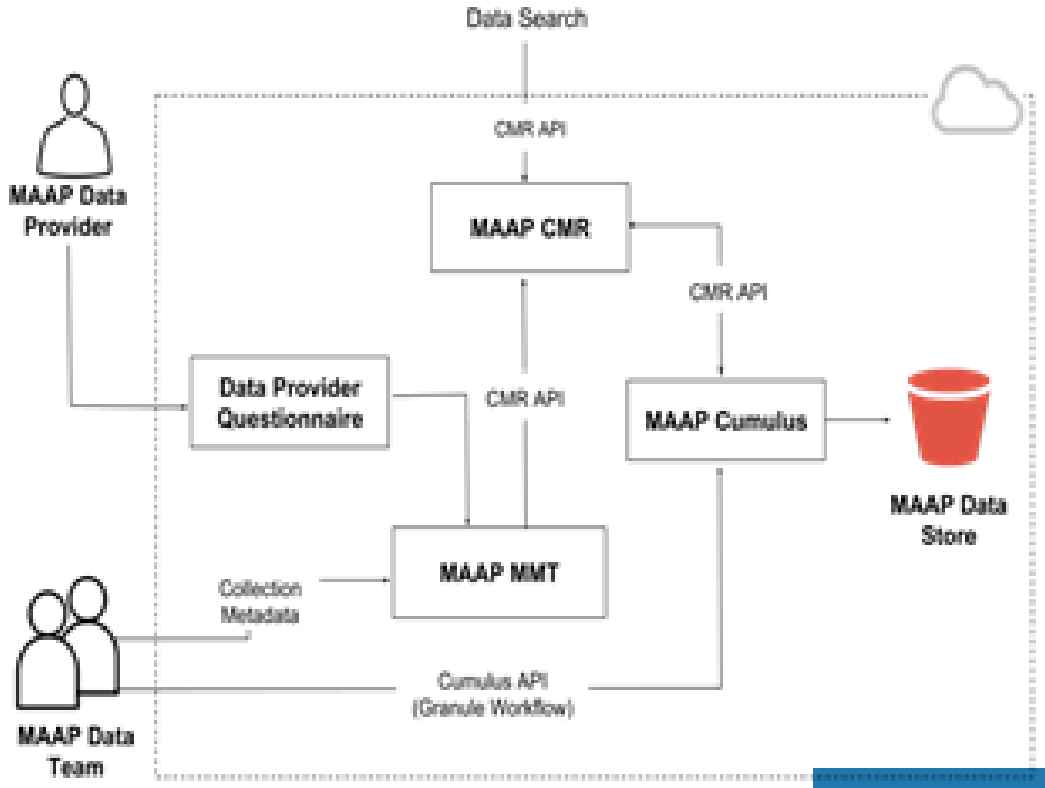
Standard MAAP Data

User Shared Data

User Shared Software and Algorithms

Standard MAAP Data Products

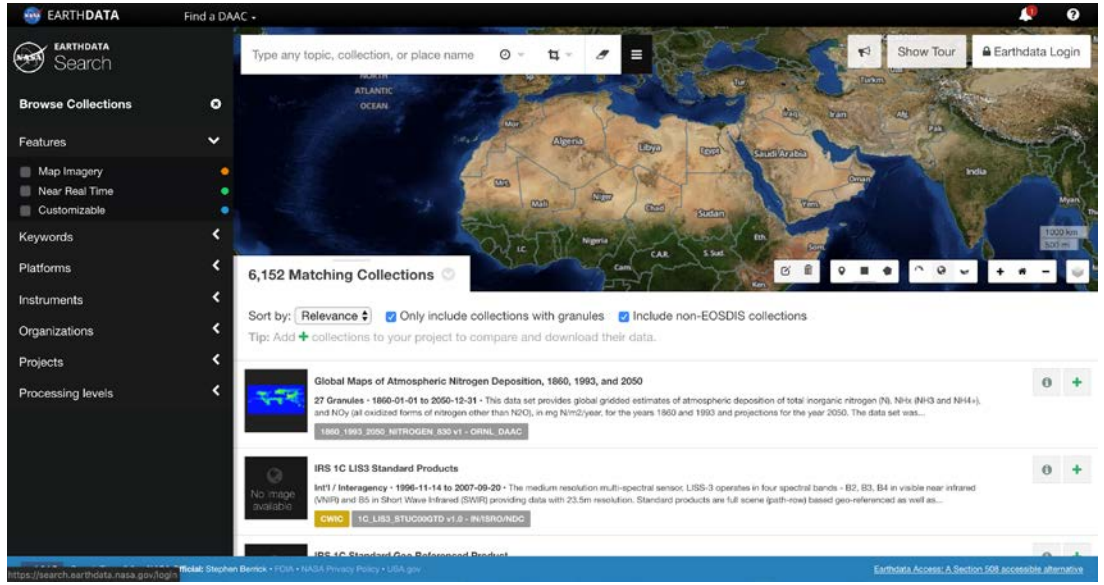
- Standard data products identified by SMEs for inclusion
- Data stewardship system reuses open source components developed by NASA's ESDIS project
  - Common Metadata Repository
  - Metadata Management Tools
  - Cumulus
- Replicates data publication process similar to those followed by NASA's DAACs
- 2 workflows
  - MAAP Data Team ingests publicly available data into
  - MAAP Data Team collects relevant information from data providers for data that is not publicly available.



COMMON METADATA REPOSITORY

Standard MAAP Data Products: Interoperability

- Facilitating discovery of biomass relevant data via a centralized location
  - Allows data from various organizations to be quickly discovered
  - Data includes primary mission data and supporting ancillary data
  - Highly curated data holdings encouraging data reuse
- ESA and NASA are contributing metadata to a single repository
  - MAAP Common Metadata Repository (CMR)
  - Meta(data) may be discovered via an API or the Earthdata Search client
  - Additional metadata information provided to support biomass search needs
  - ***Significant contribution to data ecosystem***



MAAP Data Ecosystem Conceptual Vision

MAAP ARDs and AODS

Standard MAAP Data

User Shared Data

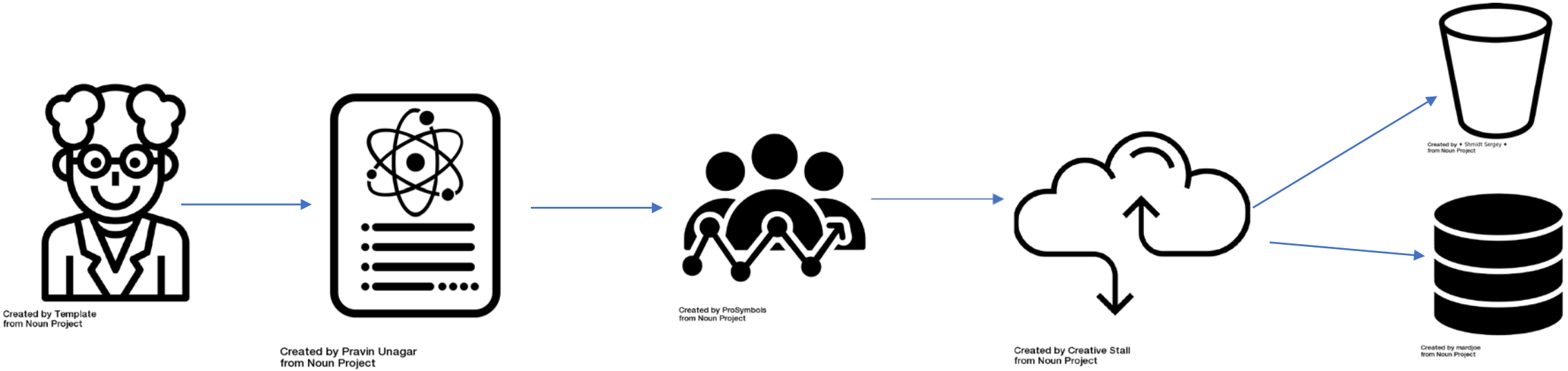
User Shared Software and Algorithms

MAAP User Shared Data

- Enabling quick and easy data sharing with MAAP users while still enabling data discovery across the MAAP data ecosystem
  - Users can share data with select collaborators
  - Can share data more broadly to MAAP CMR so users can discover it
  - Supports MAAP open data policy
- To make data sharing easier, MAAP will leverage creative ways to capture metadata info
  - Capturing information from the data itself
  - Streamlining metadata needs to lower burden on user
  - Developing best practices guidance for users sharing data including
    - Recommended file naming conventions
    - Recommended file formats

Created by Thomas' designs from Noun Project

MAAP User Shared Data



User is Ready to Share Data

User Provides Data Information

User Submits Information

User Data Ingested

User's Data Available in MAAP

MAAP Data Ecosystem Conceptual Vision

MAAP ARDs and AODS

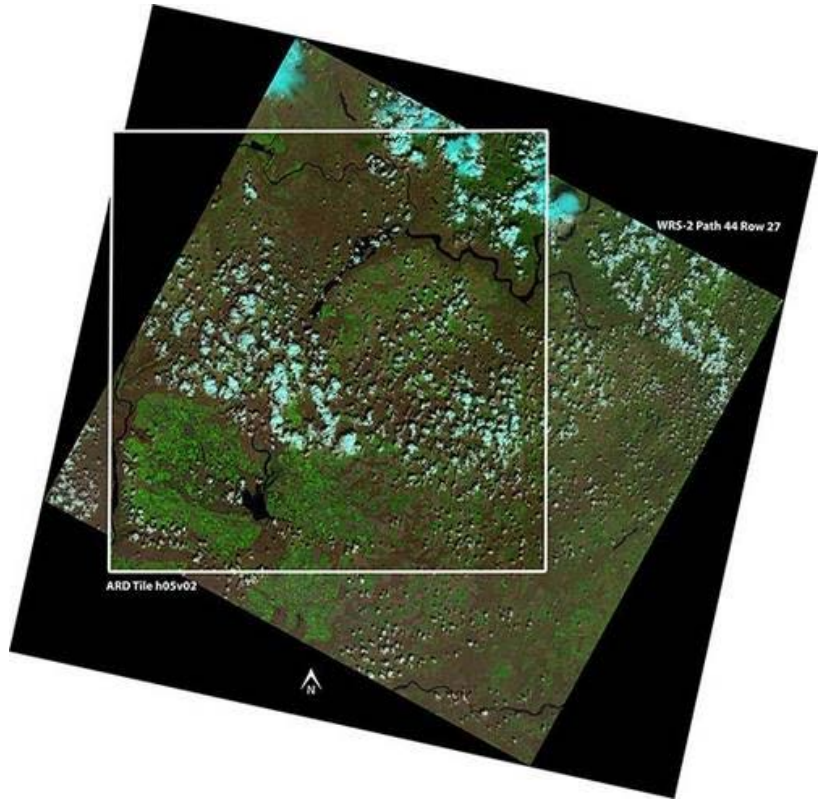
Standard MAAP Data

User Shared Data

User Shared Software and Algorithms

MAAP ARDs and AODS

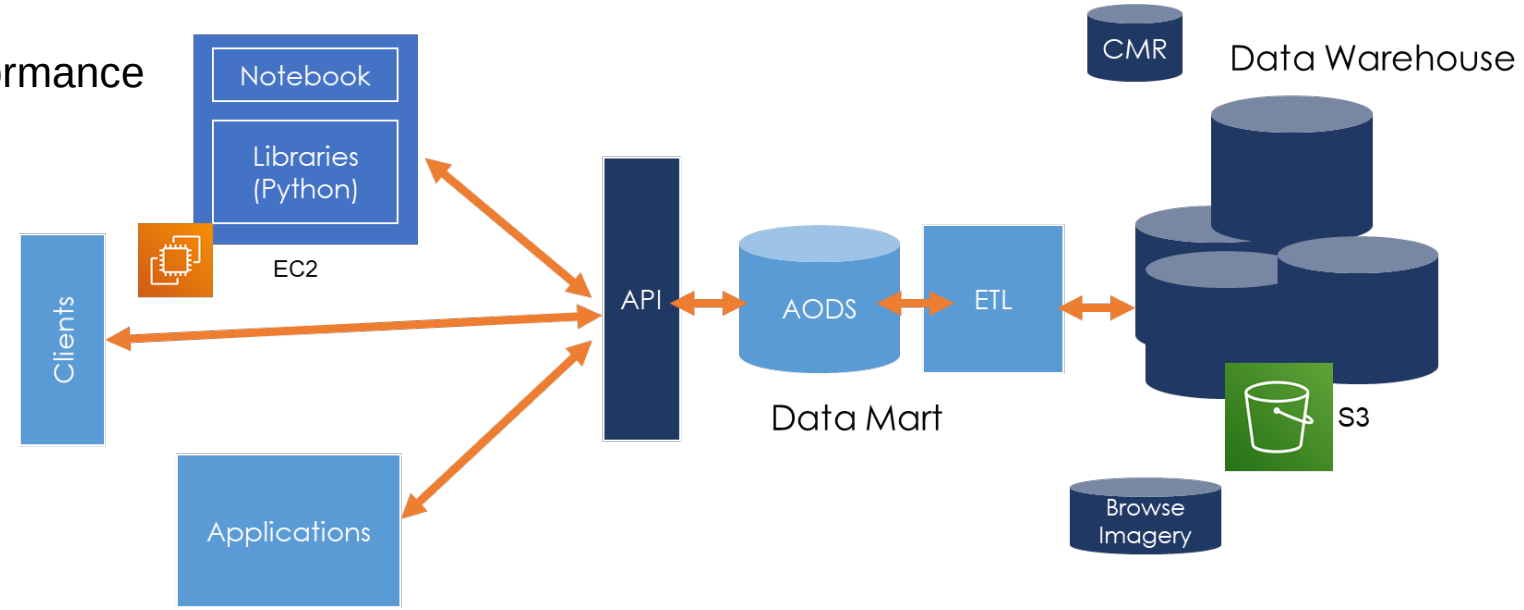
- Analysis Ready Data
  - “data products processed in such a way as to ease preprocessing and analysis burden”
- Includes processing to a common
  - Projection
  - Pixel resolution
  - Tile size
  - Variable unit
- Can also include some common preprocessing steps such as
  - Atmospheric corrections
  - Cloud masking
- Advantages of ARDs
  - Enables interoperability
    - Speeds up development of algorithms, processes (Zhu et al)
    - Easier adoption of data into decision making workflows
  - More time for science and analysis, less time spent on preprocessing



The Landsat ARDs are the most recognized example of ARDs. This image shows the ARD tile versus the original image.

MAAP ARDs and AODS

- ARDs are helpful for individuals conducting research or for users integrating data into decision making tools
- ARDs are not necessarily scalable for big data analysis tools
- To continue to evolve, MAAP will also be exploring data engineering through creating Analytics Optimized Data Stores (AODS)
  - Optimized for big data analysis tools
  - Preprocessed to meet goals
  - Focus on improved/efficient performance



MAAP Data Ecosystem Conceptual Vision

MAAP ARDs and AODS

Standard MAAP Data

User Shared Data

User Shared Software and Algorithms

MAAP User Shared Software and Algorithms

- Encourage sharing of algorithms and software to the broader MAAP community
- Create metadata to support sharing of and creating 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

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.)

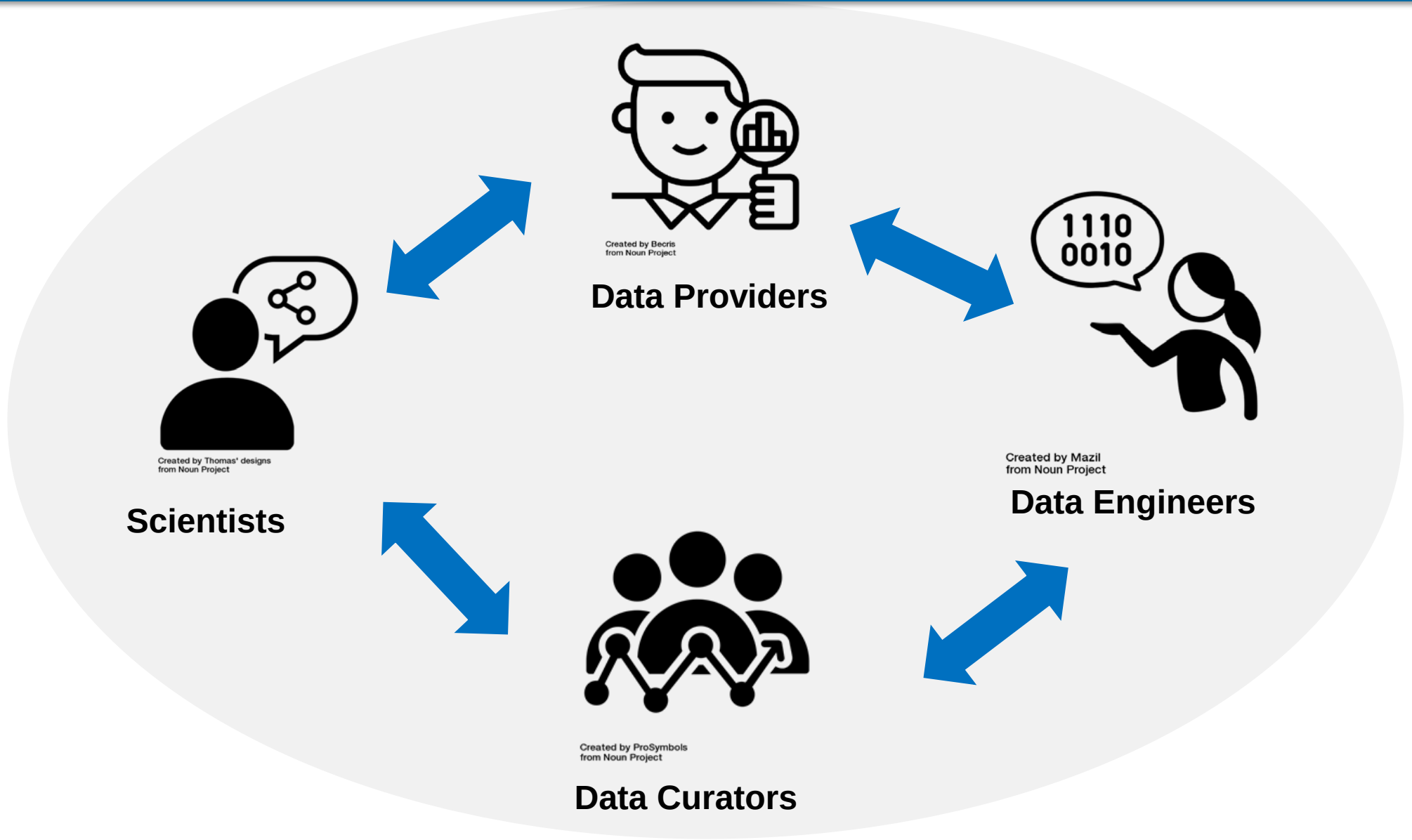
3.3 Algorithm Input Variables

Name	Unit
total Rayleigh radiance at the TOA	rad

3.4 Algorithm Output Variables

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

MAAP People


Created by Becris
from Noun Project
Data Providers
1110
0010
Created by Mazil
from Noun Project
Data Engineers
Created by Thomas' designs
from Noun Project
Scientists
Created by ProSymbols
from Noun Project
Data Curators

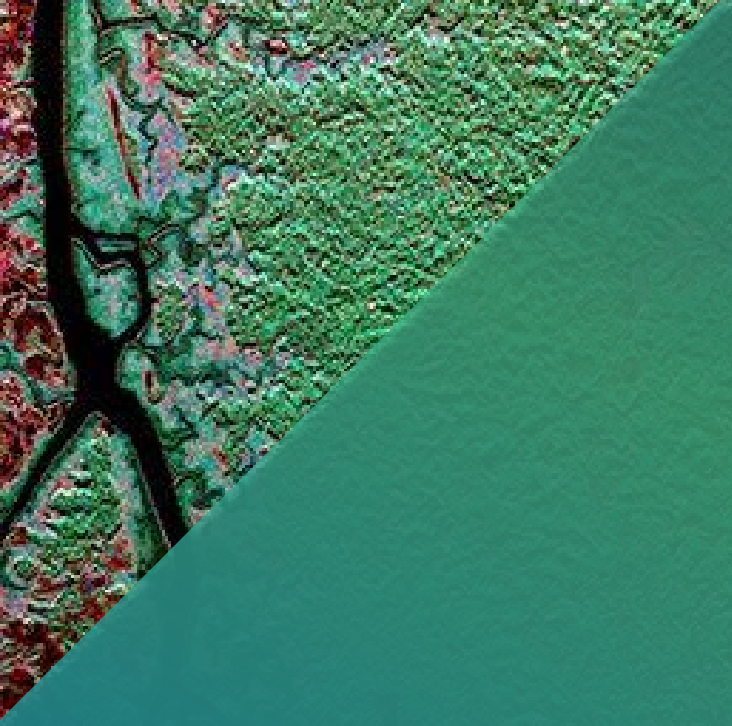
Discussion

- Reproducibility principles for ARDs and AODS
  - How do we document the scientific process with AODS?
  - Standardizing tools leveraging AODS may help ensure some consistency both for use and for communicating
  - Documentation may also help
    - Scripts, Jupyter notebooks etc that leverage AODS
    - Data recipes
- Data stewardship responsibilities in a cloud-based ecosystem
  - If data can be reproduced easily with software on the cloud, what are our stewardship responsibilities? Is it still necessary to follow the old paradigm of saving all data?
  - Trustworthiness
- Recognition or credit in this larger data ecosystem
  - Beginning to adopt 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

Conclusions

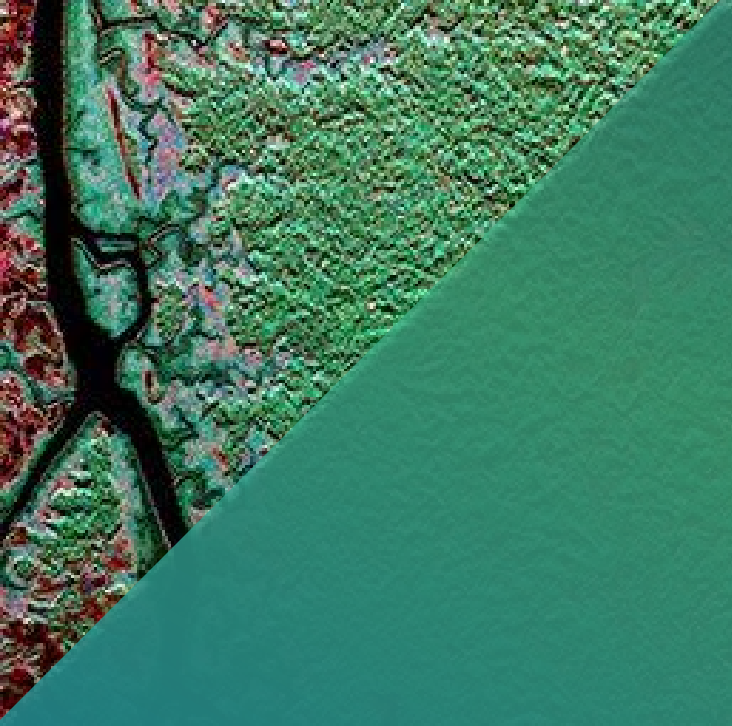
To be a true data ecosystem, MAAP will need to continue to adapt and evolve

- MAAP is a first step into a new data management paradigm
- Replicating ‘data publication’ model in cloud environment is a good first step
- Next steps towards ARDs, AODs, user shared data and software expand enrich the MAAP data ecosystem
- As people interact with new data and emerging technologies, MAAP data ecosystem will continue to evolve
  - Supporting collaboration and open science
  - Also need to consider reproducibility
- MAAP team will continue to be pathfinders for novel solutions in collaboration with other people in the MAAP ecosystem



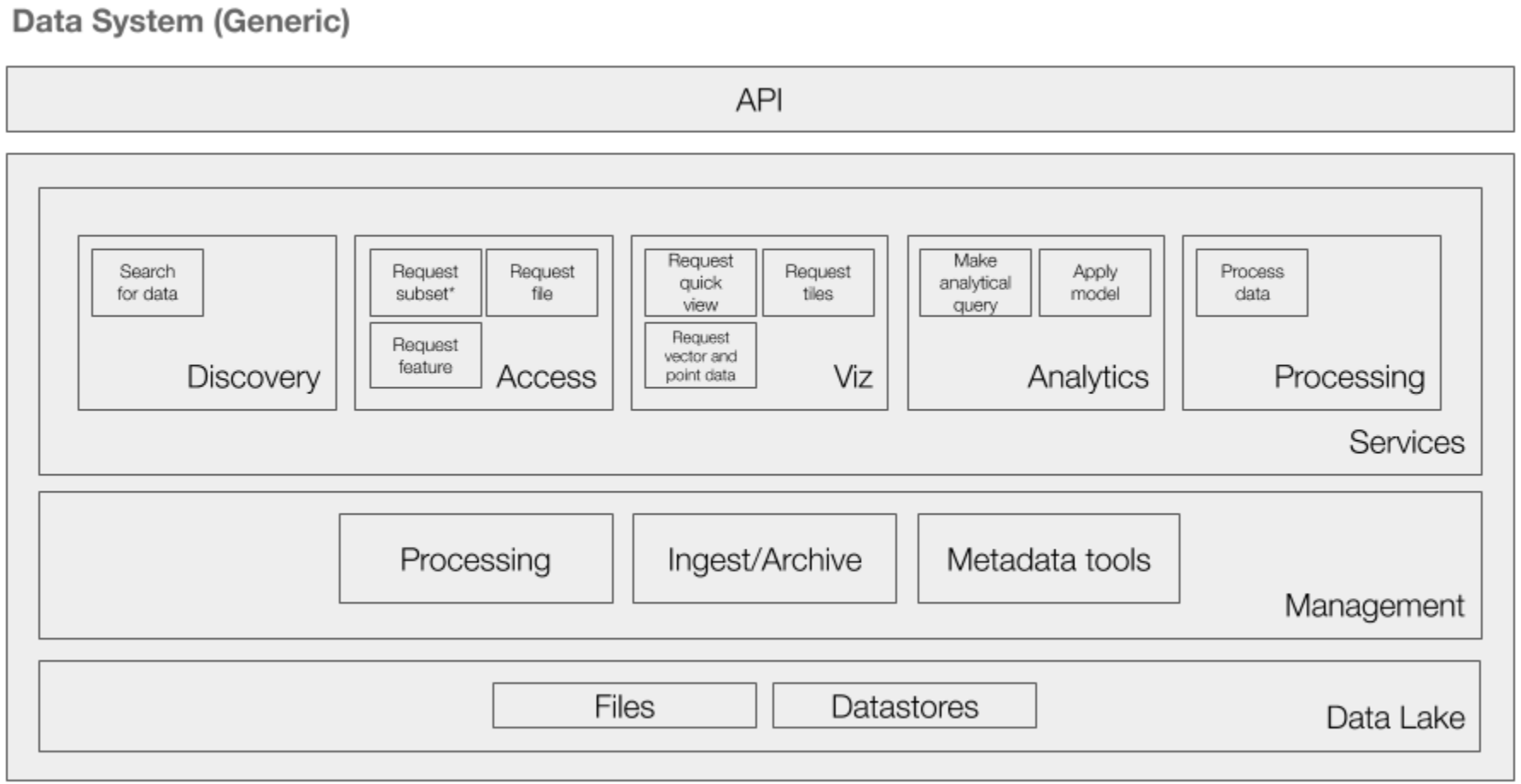
Questions?

Contact me at:
Kaylin.m.Bugbee@nasa.gov



Back Ups

MAAP Data Infrastructure



Future Work

- Implement software and algorithm metadata model into the MAAP data ecosystem to support sharing and documentation
- Investigate ways to support greater reproducibility
  - Current vision supports MAAP users sharing data, software, etc individually
  - Will want to support connections between these objects within the ecosystem
  - Want to also consider the option of sharing entire containers
    - How to document these with enough information for discoverability, usability
- Automation of user shared data workflow including scripts to help
  - Generate compliant data formats
  - Granule metadata
- Data stewardship and reproducibility principles for ARDs and AODS
  - How do we document the scientific process with AODS?
  - If data can be reproduced easily with software on the cloud, what are our stewardship responsibilities? Is it still necessary to follow the old paradigm of saving all data?