CICompass

Funded by the National Science Foundation, Grant #2127548

AI Strategies for Improving Data Management Components and Preparing AI-ready Data that meets FAIR and Data Provenance

by Giri Prakash

Making the Major Facilities Data Lifecycle FAIR to Provide AI-Ready Data

March 1, 2022
CI Compass Cyberinfrastructure for NSF Major Facilities Workshop

USC Viterbi School of Engineering Information Sciences Institute
INDIANA UNIVERSITY
UNIVERSITY OF NOTRE DAME
THE UNIVERSITY OF UTAH
TEXAS TECH UNIVERSITY
renci

AI Strategies for Improving Data Management Components and Preparing AI-ready Data that meets FAIR and Data Provenance

GIRI PRAKASH

Oak Ridge National Laboratory

https://www.arm.gov

2022 NSF Cyberinfrastructure for Major Facilities Workshop: Getting Together, Working Together, March 01, 2022
palanisamyg@ornl.gov

Observations Support Atmospheric Research

MISSION:

Provide the climate research community with strategically located atmospheric observatories to improve the understanding and representation in earth system models of clouds and aerosols and their interactions with the Earth's surface.

The World's Foremost Ground-Based Atmospheric Observing Facility

Atmospheric Radiation Measurement Facility

Since 1992, the world-leading facility for measurements of cloud & aerosol properties, & their impacts on Earth's energy balance

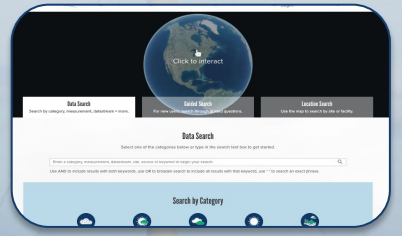
Network of 3 fixed-location & 3 mobile observatories

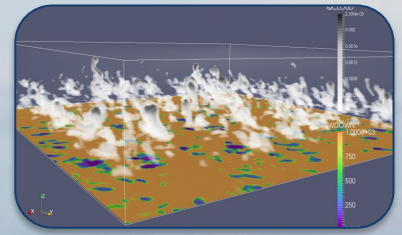
Piloted & unmanned aerial measurement platforms

Extensive data management infrastructure

Freely available data products to support atmospheric research & model development

Large-eddy simulation (LES) model simulations and analysis tools





Comprehensive measurements across diverse climate regimes

Serves the international climate research community and has close collaboration with Atmospheric System Research (ASR)

Source: Jim Mather

Comprehensive Sets of Measurements Deployed in Diverse Climate Regimes

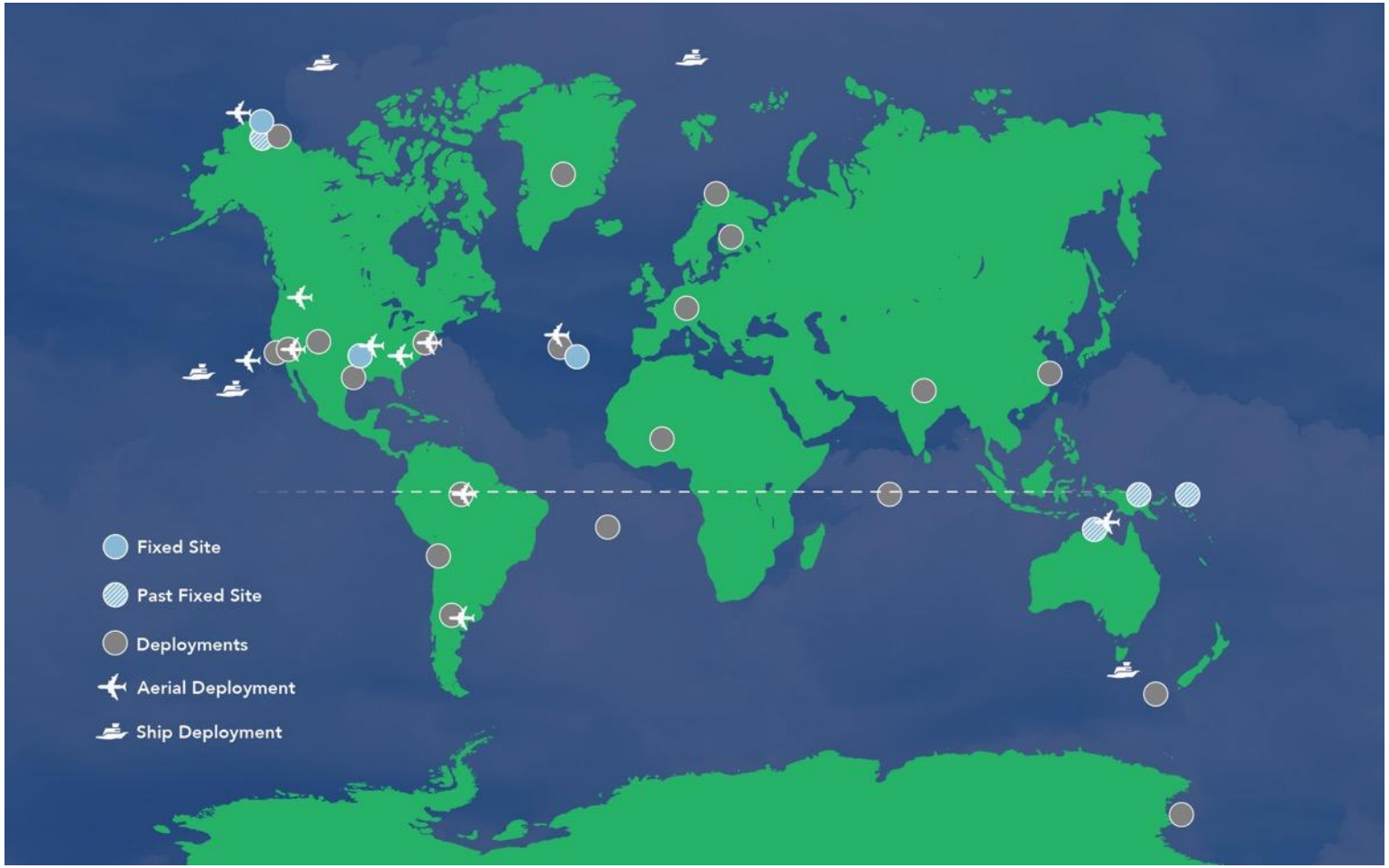
Background atmospheric state

Surface energy balance

Aerosol and hydrometeor profiles

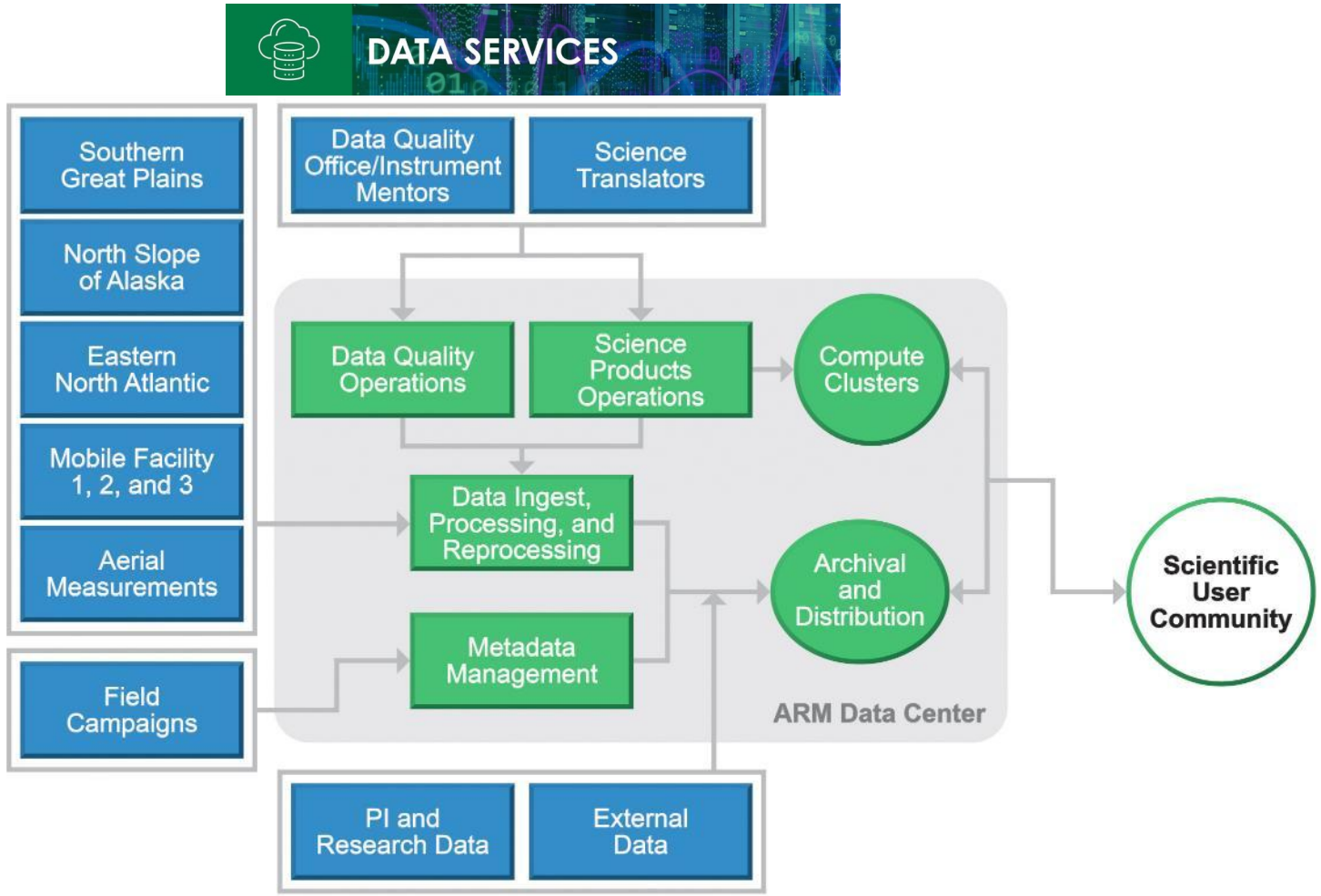
Near-surface aerosol properties

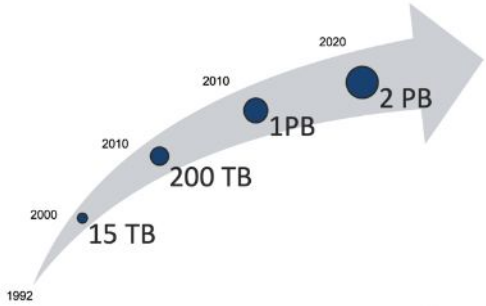
Aerial measurements



End-to-End Data Services

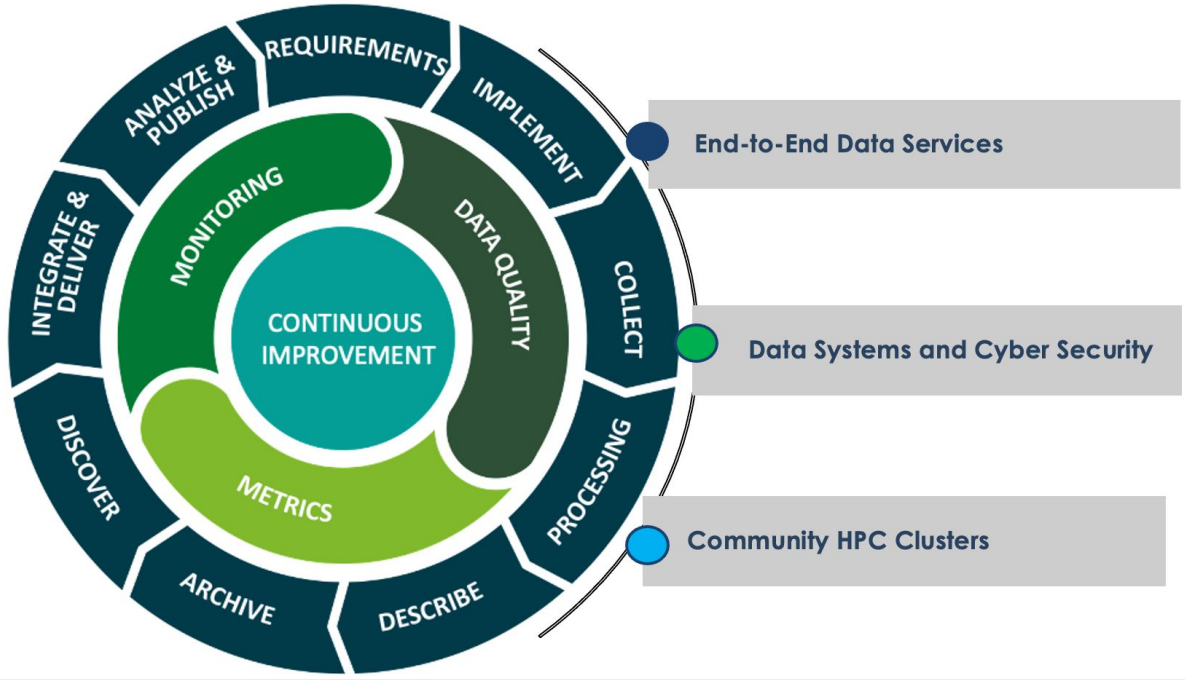
Providing powerful and adaptable computing resources to meet data analysis challenges





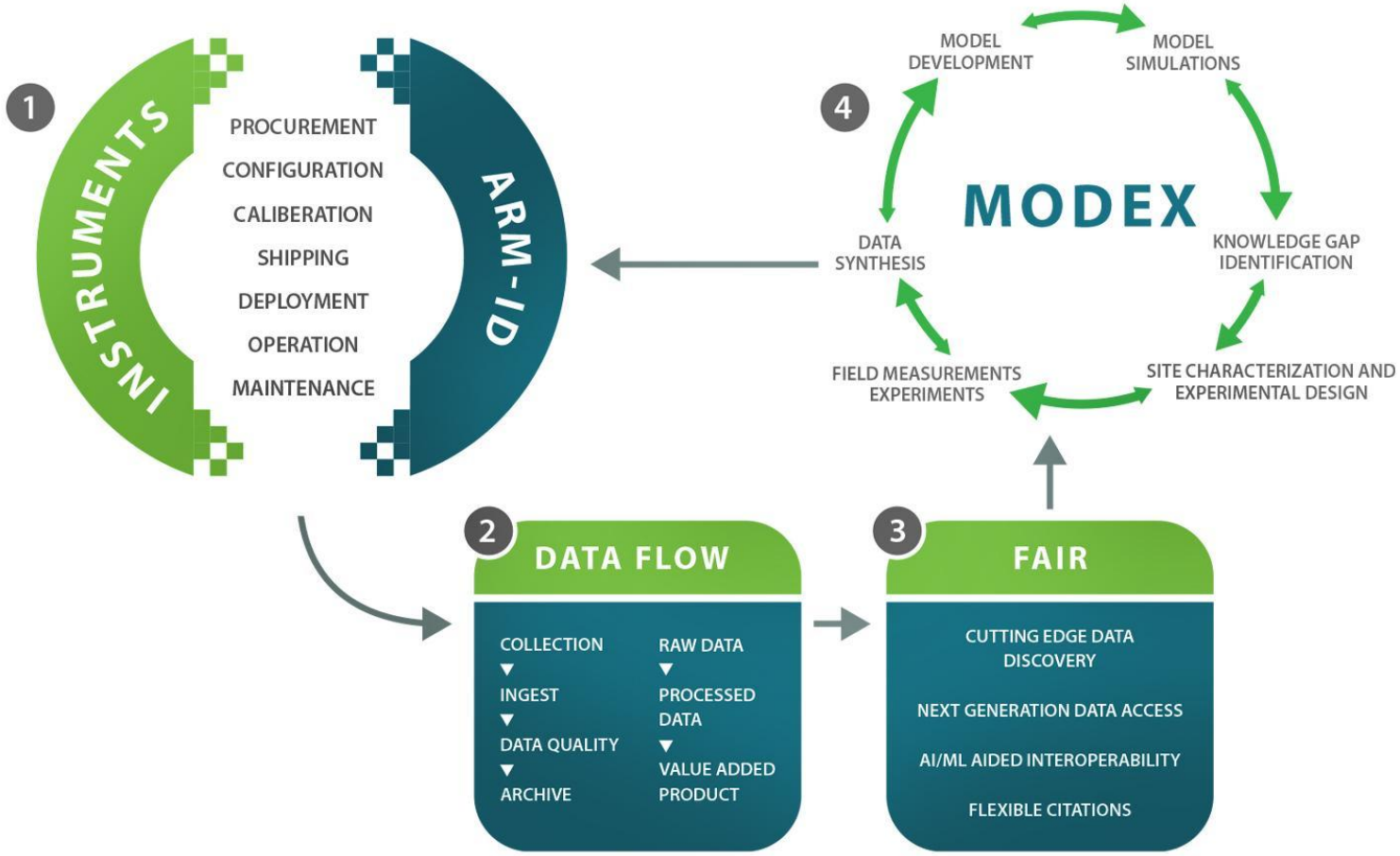
ARM Data Center (ADC)

- Provides a robust integrated data and computing ecosystem to advance understanding of atmospheric radiation
- Key components include:
  - Data management, operations, and monitoring
  - Data archive and distribution
  - Cyberinfrastructure
  - High-performance computing (HPC) environment
  - User metrics
  - Data analytics and visualization



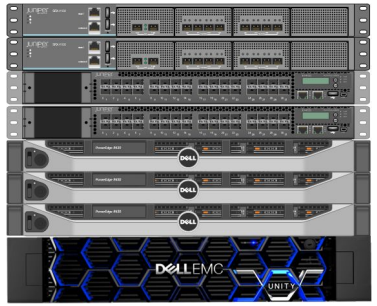
Data Services – Looking Ahead

- Enabling Digital-twin for the instruments
- MODEX integration
- FAIR for next Generation data access and AI/ML interoperability



Data Collection Systems and Edge Computing

- Scalable data systems, hardware and software solutions proven effective over multiple generations/deployments
- Real-time data access to enable data reduction and edge computing
- Hardware (SSDs and Data deduplication) to support onsite data analysis

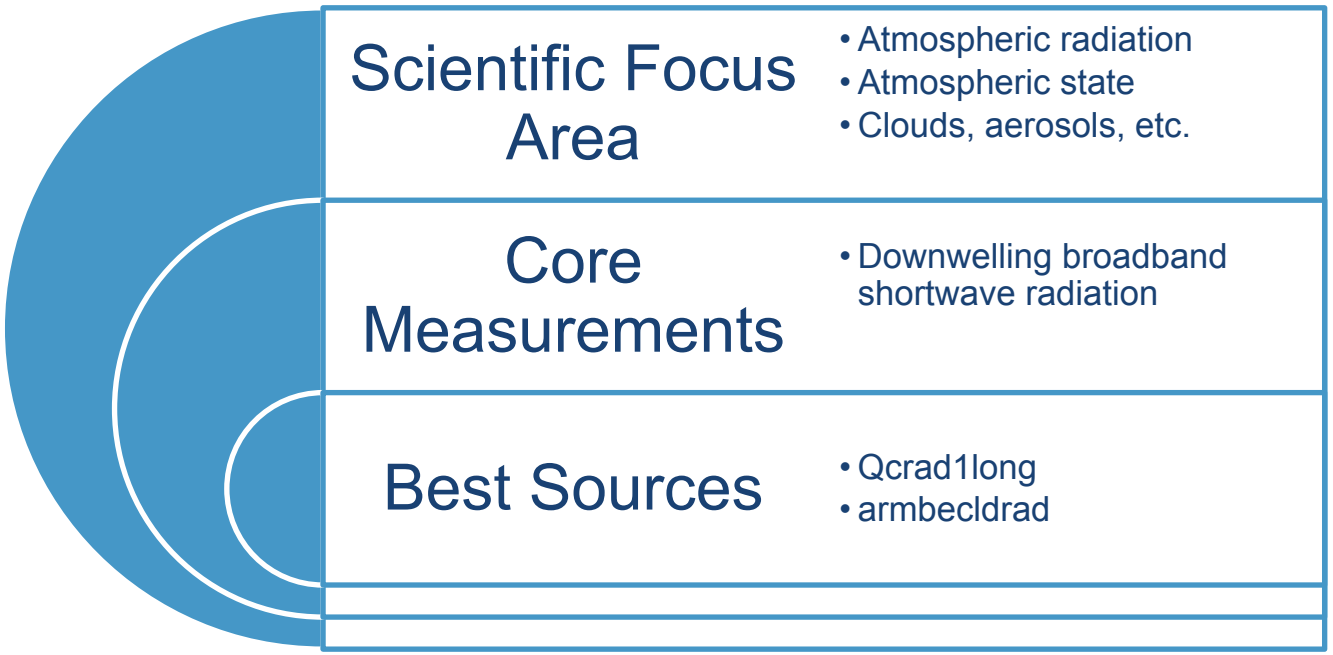


DATA COLLECTIONS	EDGE COMPUTING	SUPPORTING SERVICES
Directly from Instruments	Local Data Storage	System/Process Monitoring
From Complex Instruments	Real-Time Processing	Instrument System Configuration
Scheduled or Streaming	Data Distribution/ Transfer	Remote Access

All communications are encrypted, inspected, and logged.
Data can also be encrypted at rest.

Data Recommender System – A solution for Data Discovery Challenge

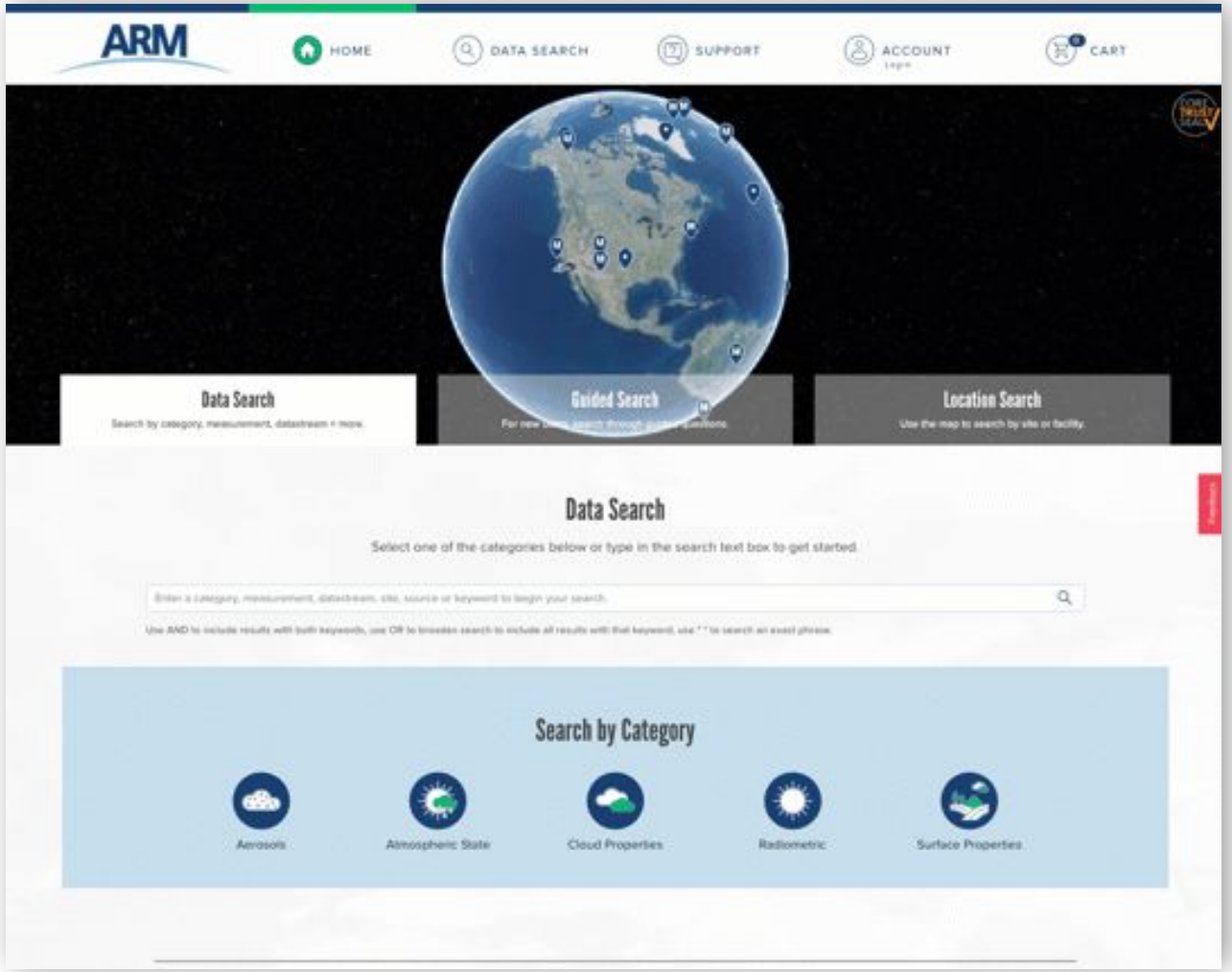
- Recommends best data sources for the core measurements
- Criteria for recommendations include:
  - Quality
  - Temporal and spatial coverage and resolution
  - Applicability for the research needs
- Process include input from subject matter experts



PIs: Maggie Davis (ORNL) & Scott Collis (ANL)

Modern Data Discovery – Seamless Utilization of AI/ML

- Modern big-data software architecture with Continuous Integration (CI)
- Intelligent search capabilities based on the actual data, guided search for user comfort
- Recommendations, data tagging based on epochs or golden periods
- Near real-time access via secured webservices (API access)

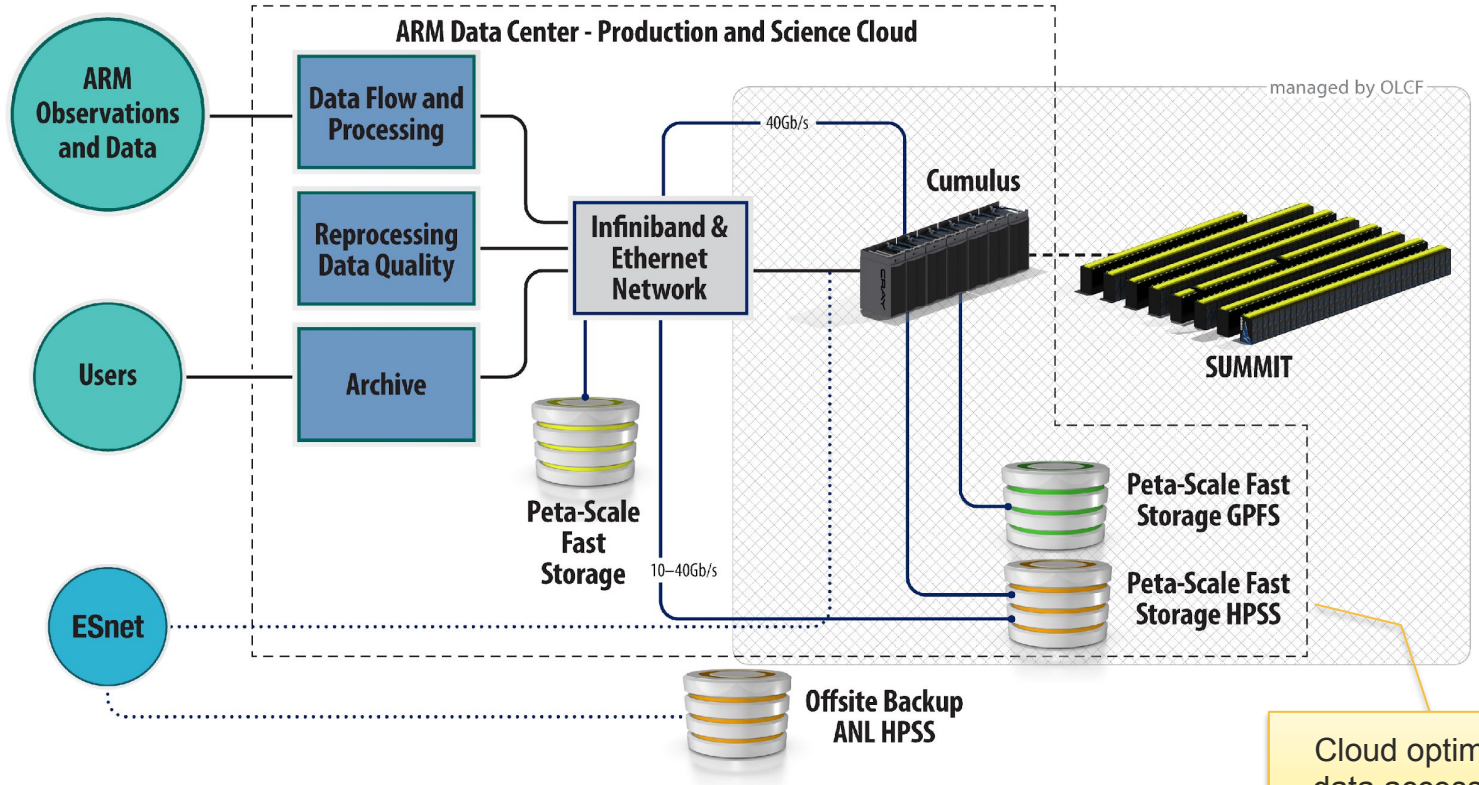


Next-Gen Data and Computing Infrastructure

- Leveraging DOE Leadership Computing and commercial Cloud Capabilities

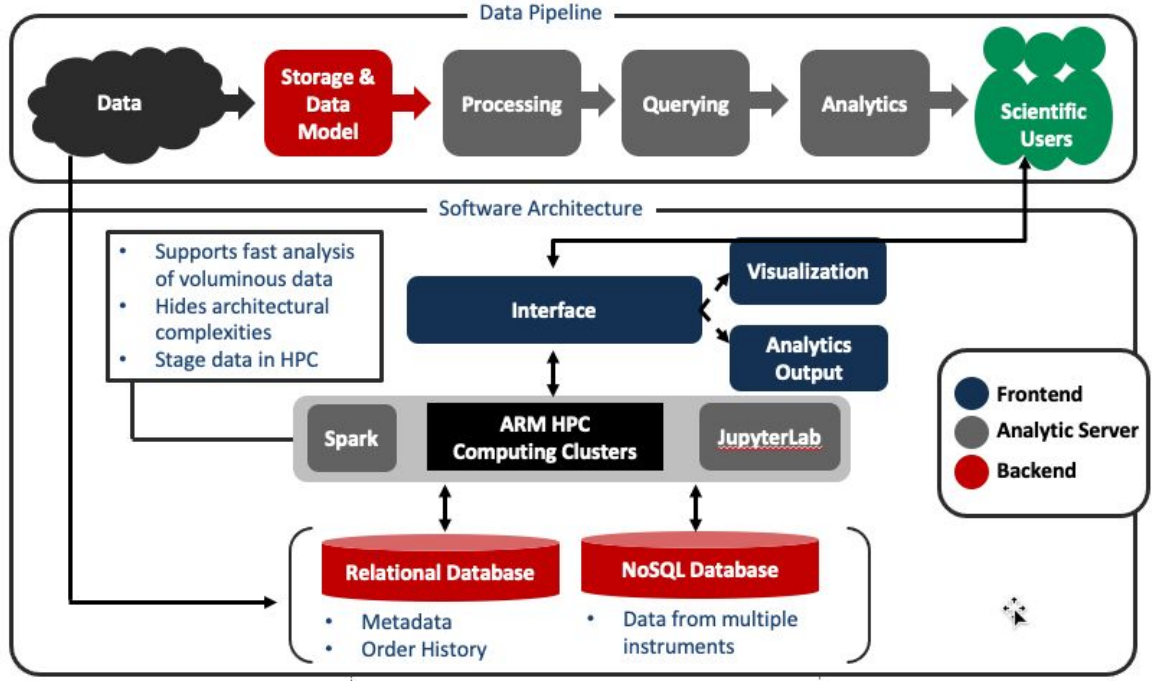
  https://www.arm.gov/capabilities/computing-resources

Offers computing infrastructure to support next-generation ARM model simulations, petascale data storage, and big-data analytics for atmospheric and climate science research.



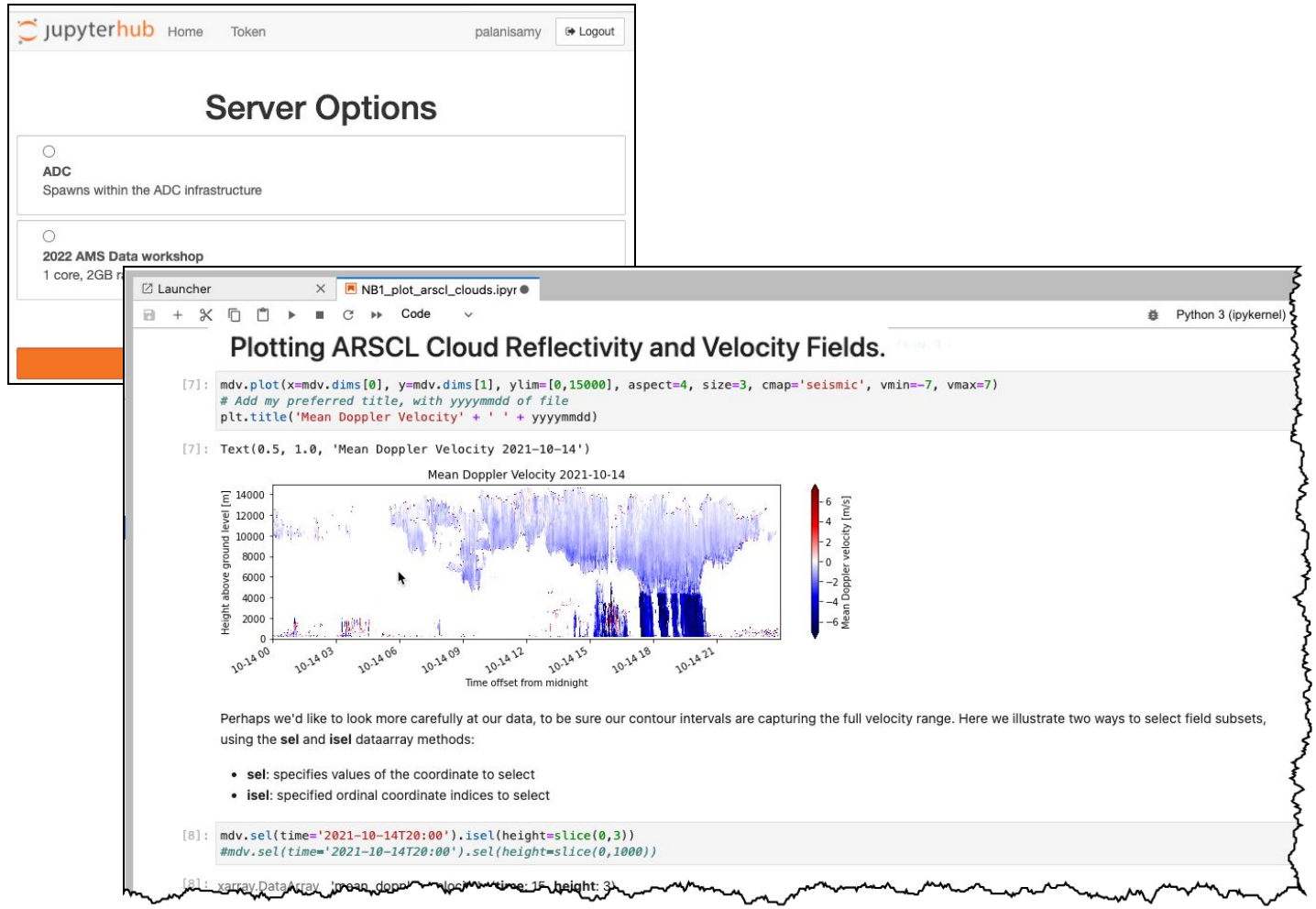
Workflow Improvements- AI-based Future Cyberinfrastructure for Data Analytics

- Provide transparent access to data anywhere users have resources
- Provide transparent access to resources wherever capacity exists
  - ***Machine learning and AI platform***: GPU-based computing platforms, software stack and workflows
  - Deep Learning software stack expertise and infrastructure to analyze large amount of observational and model data
  - Unified gateway to launch notebook on ARM HPC vs ADC computing nodes
  - Advanced HPC Viz/analytics such as ParaView



Enabling Data Analytics

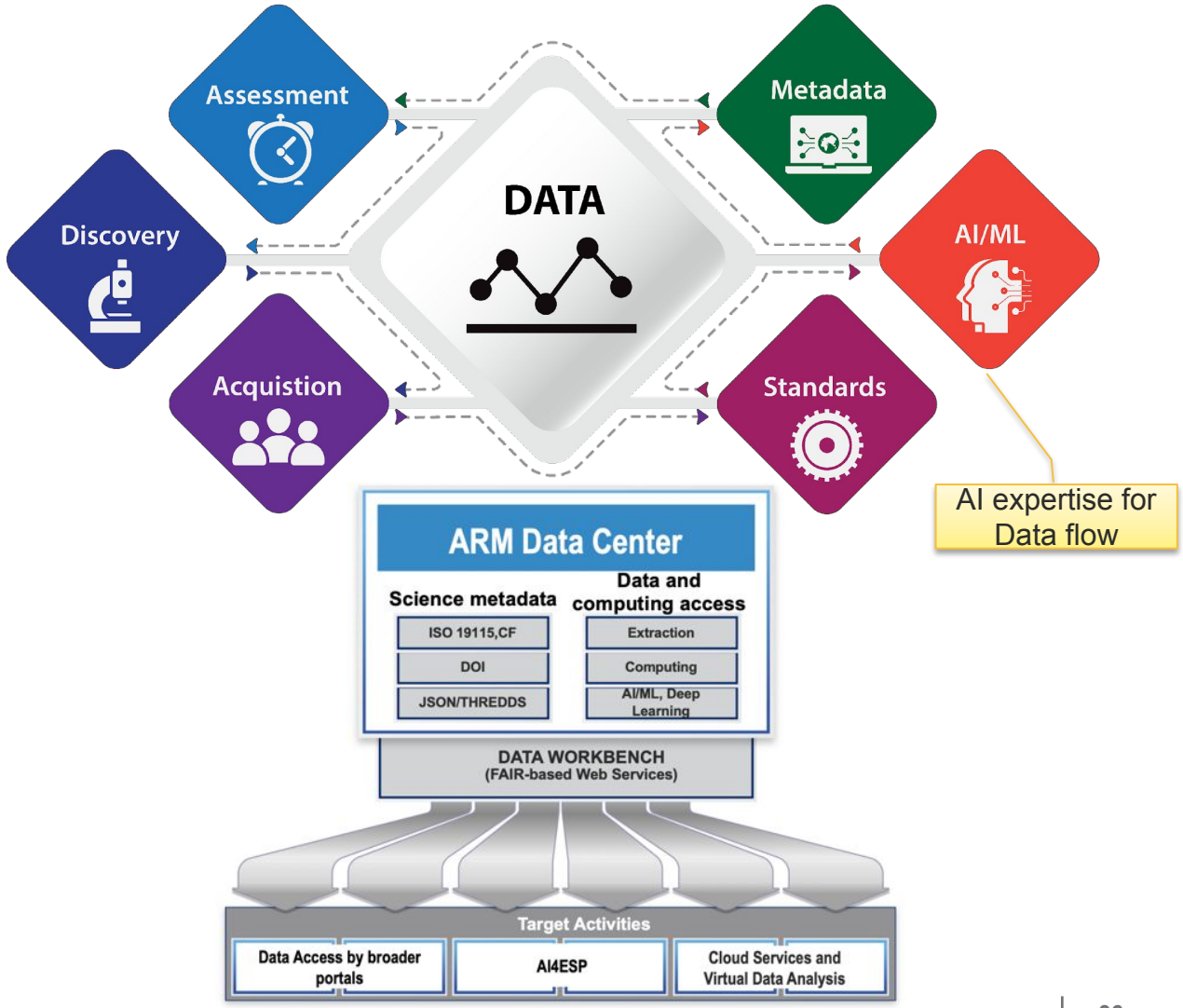
- Access to All ARM Data
- Spawn data analytics and processing to ARM HPC
- Trainings and tutorials
  - AMS 2022
  - https://www.arm.gov/data/work-with-arm-data/webinars/



Zach Price (ADC)

Data Interoperability for AI and Beyond

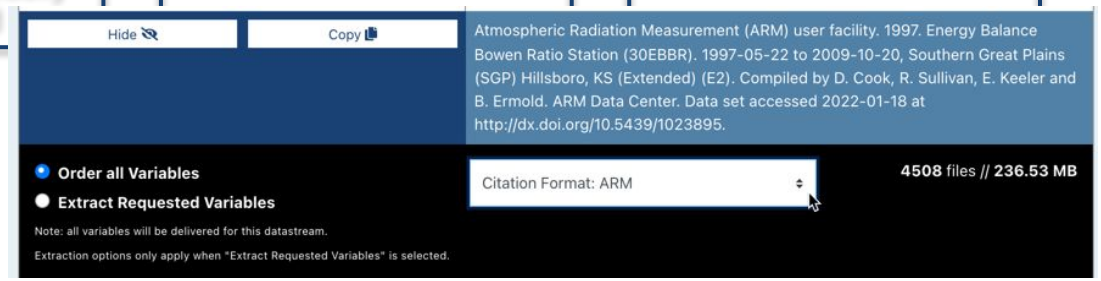
- Aims to achieve transformative knowledge discovery by providing modular capabilities
- Enable FAIR-based access to ARM data and computing for upcoming initiatives (e.g., AI4ESP)
- Integrates future cyberinfrastructure for data analytics
  - E.g., Enable machine-learning framework to support data interoperability from diverse sources (ARM, NEXRAD, MODIS, ECMWF, etc.)



Data Citation Strategy — In an Era of AI

- Enable metrics to quantify science impact and ensure data reproducibility
- Still evolving
  - Enabling time-based author credits
  - Additional citation formats
  - Nested citations
  - Data mashups

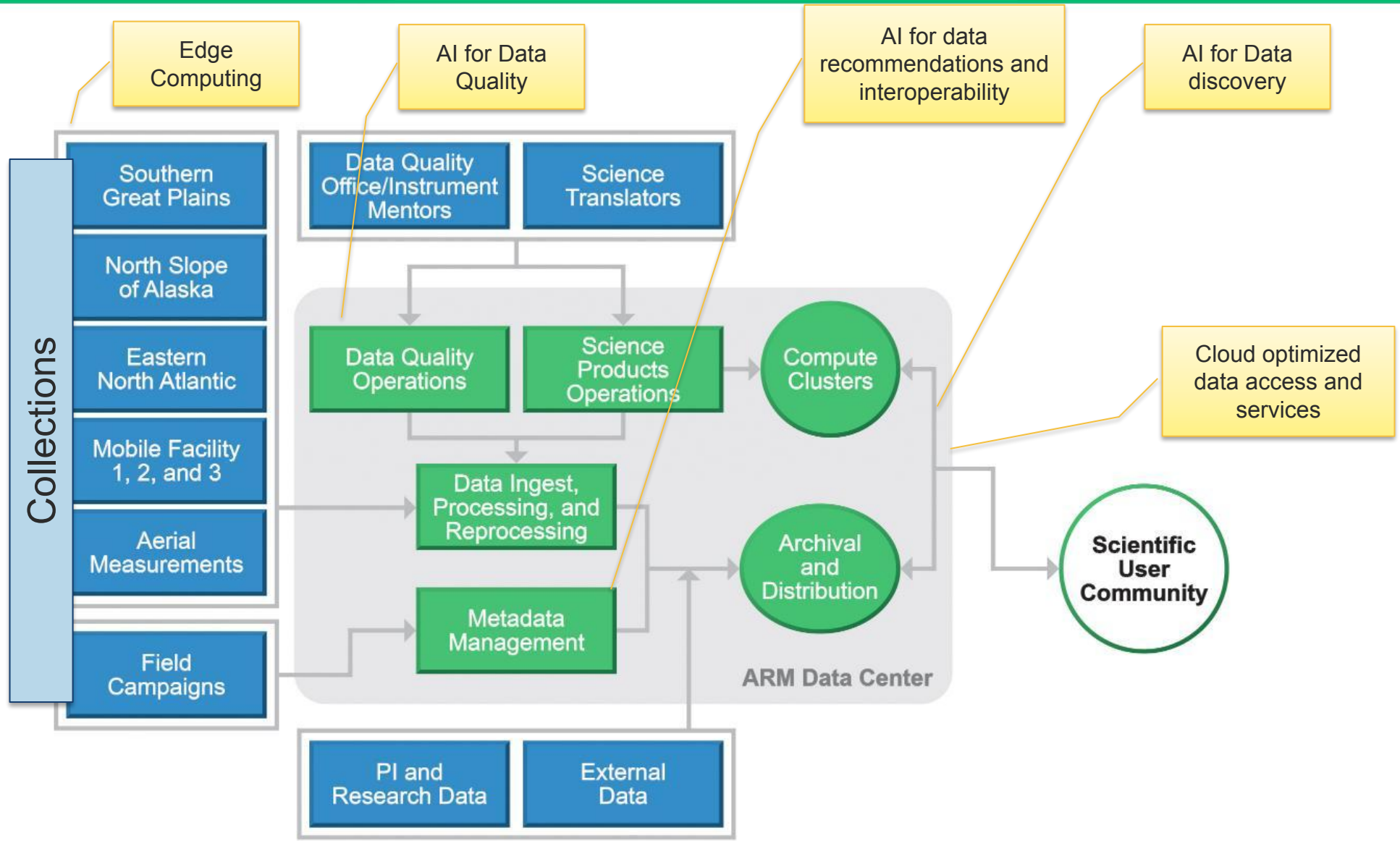
Benefits	Challenge	Strategy
▪ Allow users to cite exact ARM data used in their research/publication ▪ Allow ARM to provide proper data citation credits to the PIs and collaborators ▪ Allow future data users and the project to easily track the data used in various articles	▪ Millions of data files from over 11,000 data products ▪ Typically continuous datastreams, but some of them are from field campaigns	▪ DOIs are assigned at the data collection level ▪ Recommended citation structure



https://www.mdpi.com/154276

Prakash, G, B Shrestha, K Younkin, R Jundt, M Martin, and J Elliott. 2016. "Data Always Getting Bigger—A Scalable DOI Architecture for Big and Expanding Scientific Data." *Data* 1:11.

Looking Ahead..



Questions?

- https://www.arm.gov
- ”Ask Us”
- ARM Data Center: adc@arm.gov
- My contact: palanisamyg@ornl.gov