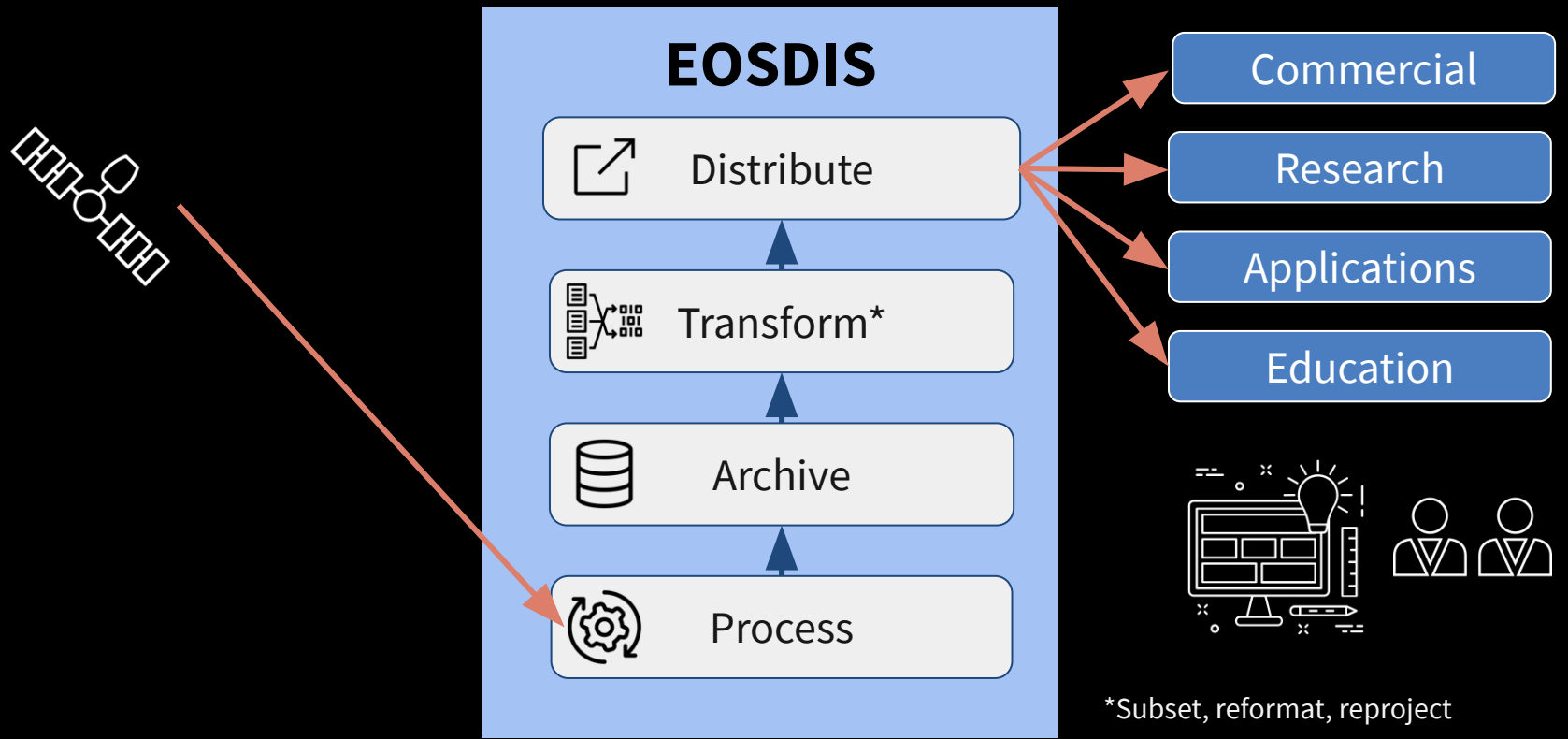
Opportunities for Accelerating Research in the Cloud

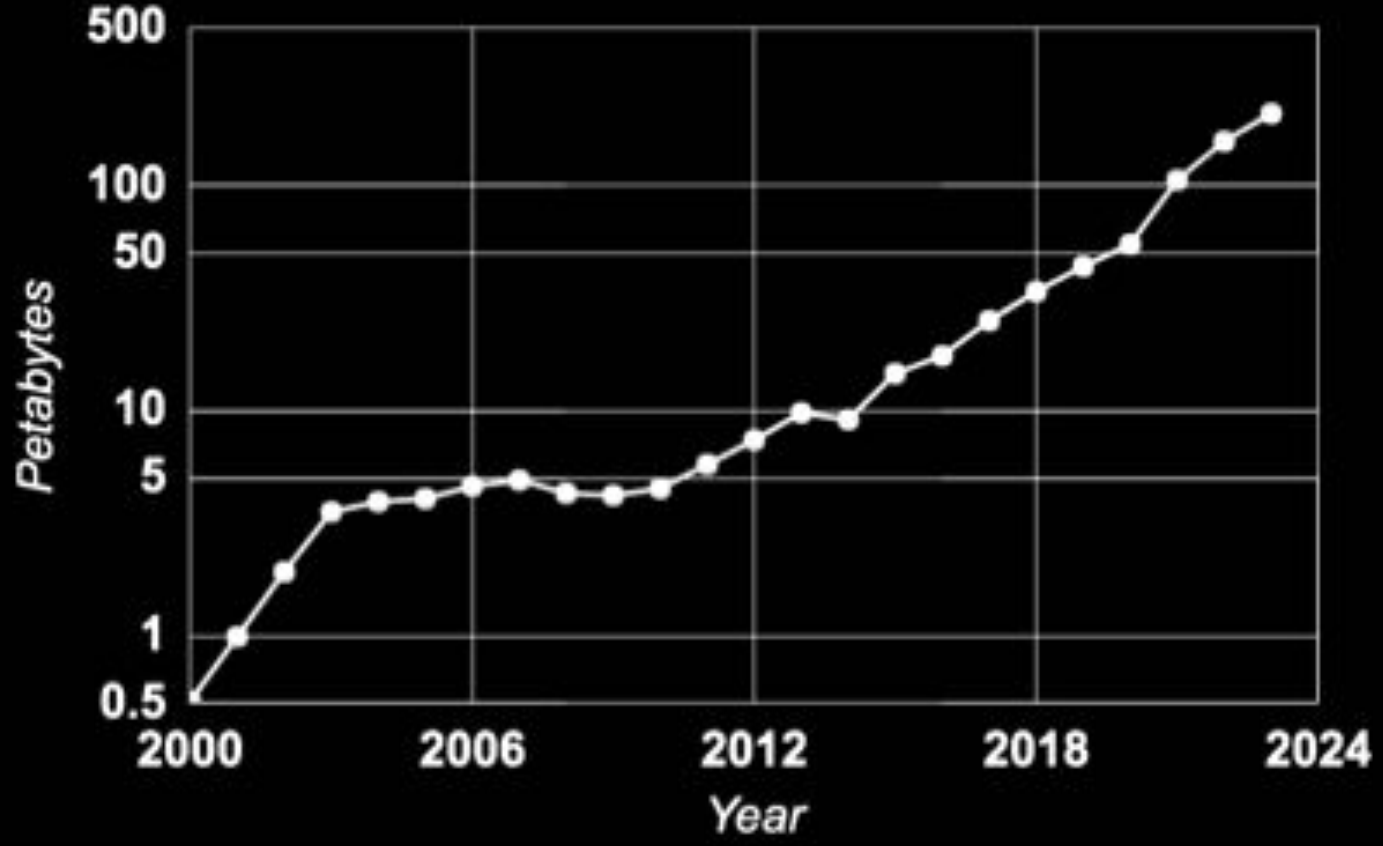
Chris Lynnes, NASA

Earth Observing System Data and Information System

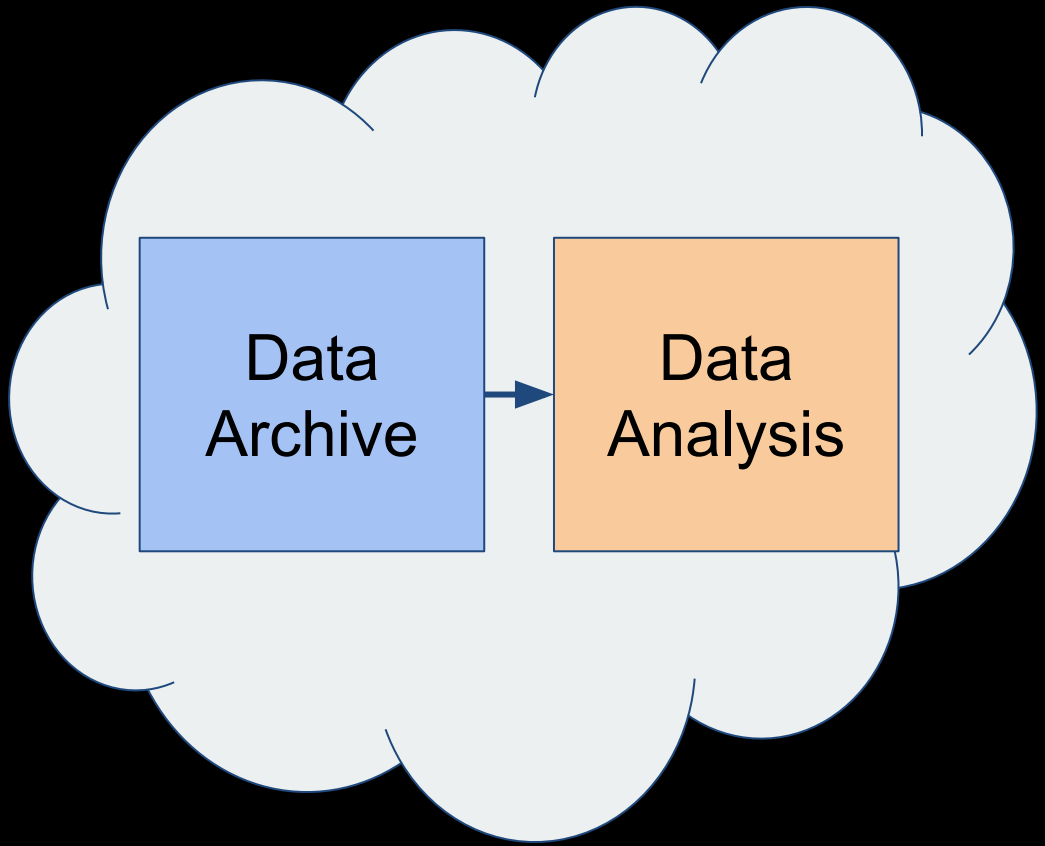


*Subset, reformat, reproject

Impending Archive Growth in EOSDIS



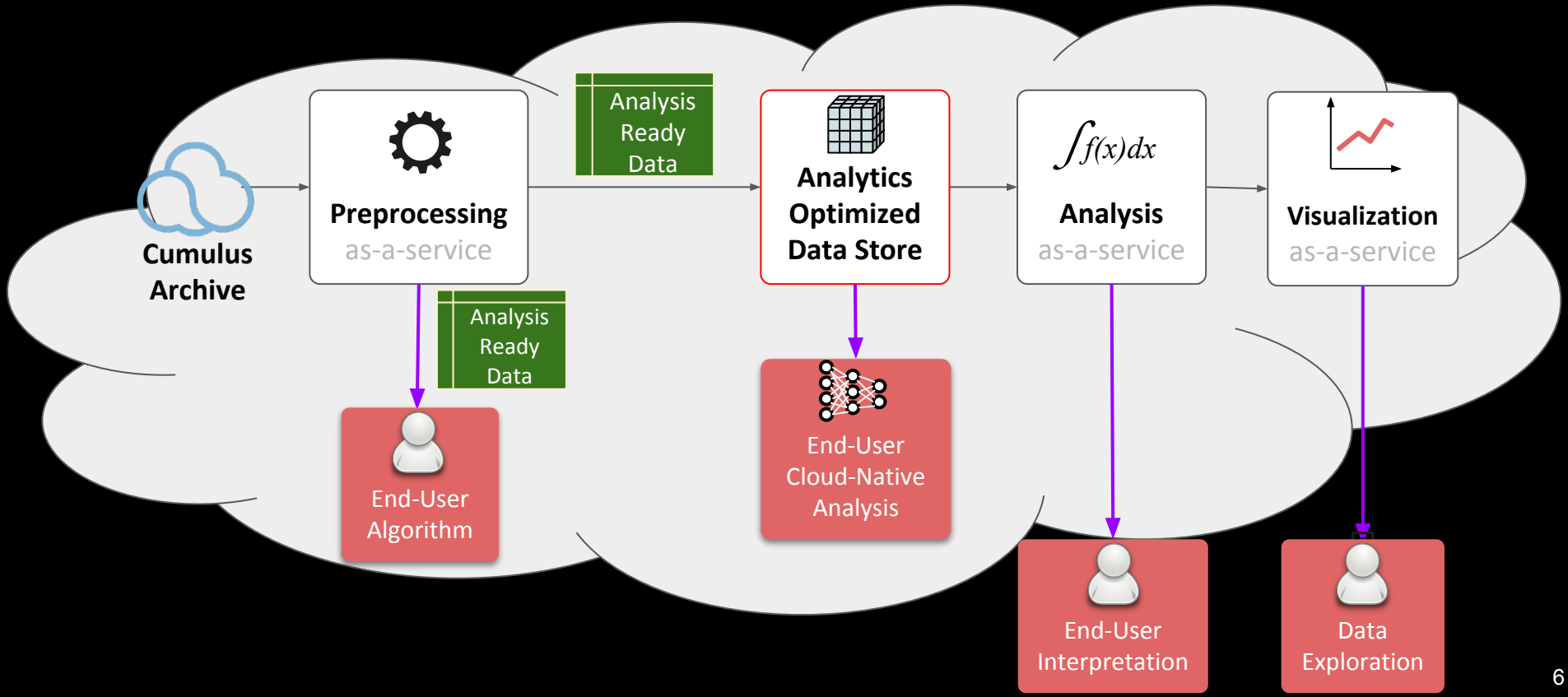
Data-Proximal Analysis



SWOT Analysis of Cloud

Strengths	Weaknesses
Scaling S3 costs Virtual co-location	Egress cost Storage access granularity
Opportunities	**Threats**
Faster services Synchronous machine interface Mixed-data services Mixed-service workflows Analysis-in-place	Unbounded demand-driven cost

Data Analysis Workflow



Analysis Ready Data (ARD)

Analysis Type	Visualizations	Challenges
Time Series	Animation, Hovmoller, Line Plot vs. Time	Orthogonal to typical data organization
Machine Learning	Classification Map, Segmentation Map	Need labeled data (usually)
Data Fusion	Super-resolution Map	Need co-registered data How much are we making up?

ARD vs. Analytics Optimized Data Store

ARD: Changes Data Values

Common Criteria

- Georeferenced
- Calibrated
- Orthorectified, reprojected, regridded
- Quality filtered

Imagery-specific

- Atmospherically corrected

SAR-specific

- Radiometrically terrain corrected
- Despeckled

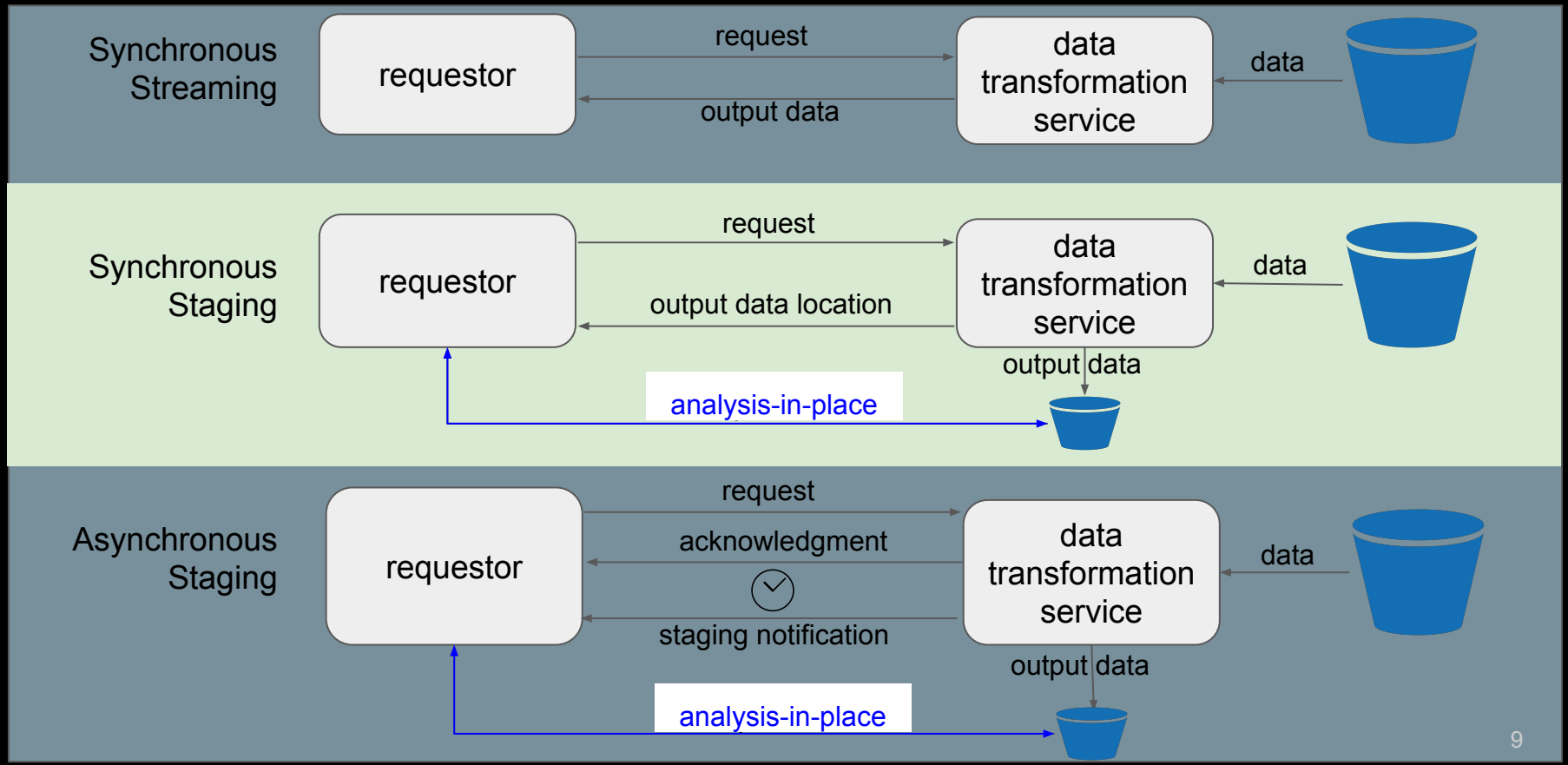
Parameter Retrievals

- Aerosols
- Sea Surface Temperature
- ...and thousands of others

AODS: Arranges/Stores Data Values

- Scalable Databases
  - MongoDB
  - SciDB
  - Parquet + Athena
- Scalable Filesystems
  - HadoopFS
  - ClimateSpark
- File Formats
  - zarr
  - Cloud-optimized Geotiff

Modes of Service Fulfillment

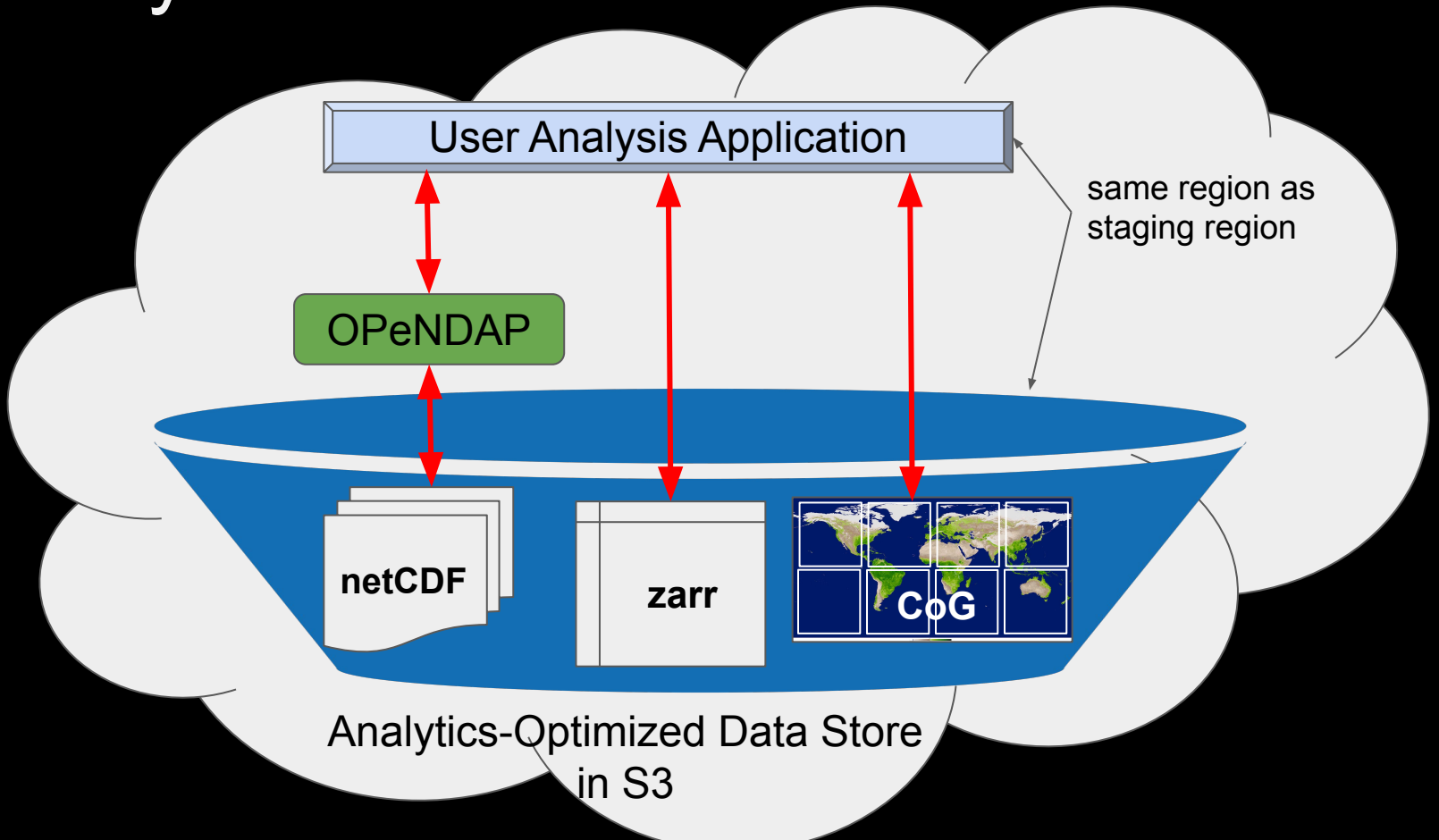


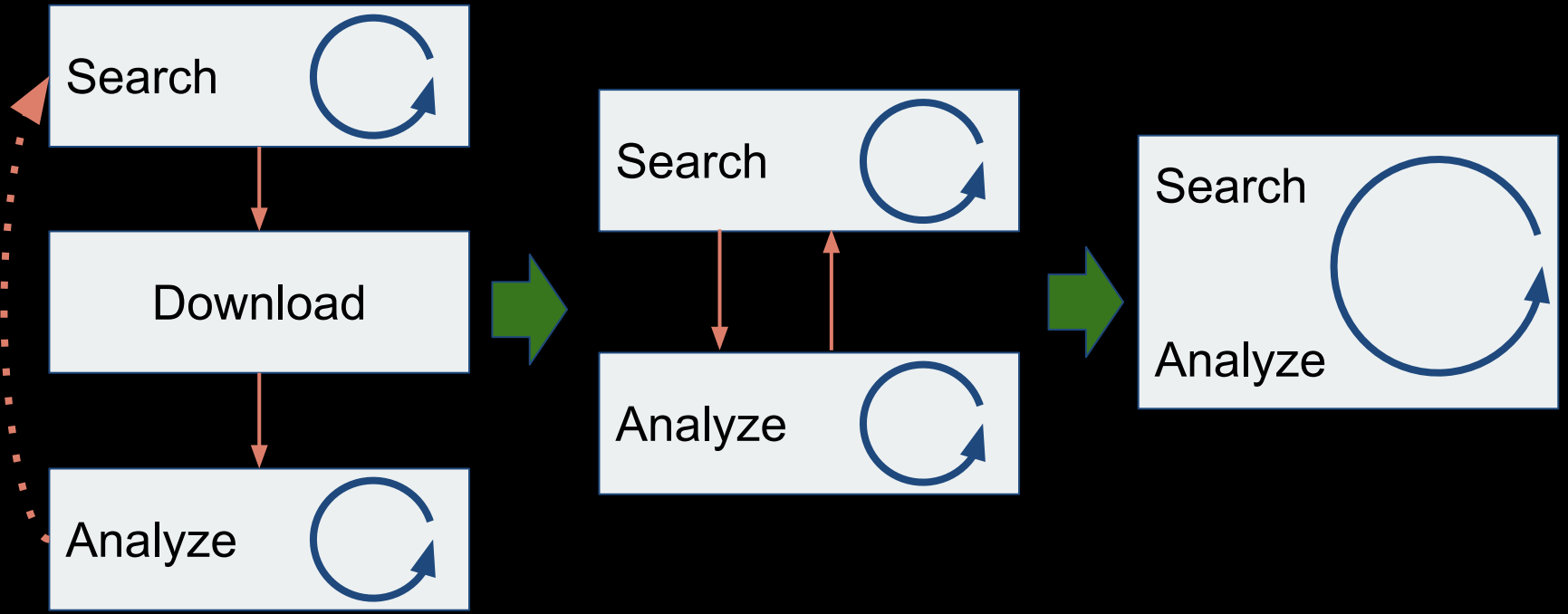
Pros and Cons of Different Modes

Synchronicity	Data Flow	Examples	Pros	Cons
Synchronous	Stream to client	OPeNDAP[1]	• Easiest machine interface	• Fast service reqt • Data egress • Single-file-out mode only
Synchronous	Stage to S3[2]	WCS[3] 1.1 “store=true”	• Easy machine interface • Handles multiple files • Analysis-in-place	• *Really* fast service reqt
Asynchronous	Stage to S3	HITIDE[4] AppEEARS[5]	• Unlimited number of files • Analysis-in-place	• Hard machine interface

[1] Open Source Project for a Network Data Access Protocol
[2] Simple Storage Service
[3] Web Coverage Service
[4] High-level Tool for Interactive Data Extraction
[5] Application for Extracting and Exploring Analysis Ready Samples

“Analysis-in-Place”




Search
Download
Analyze
Search
Analyze
Search
Analyze

Challenges

- For users:
  - Learning curve
  - Unintentional expenditures
- For data and service providers
  - Data egress costs
  - Cost management

Ongoing Initiatives

- How to Cloud Primer for Scientists
- Intern Pathfinder Projects
- Control mechanism for unintentional Costing