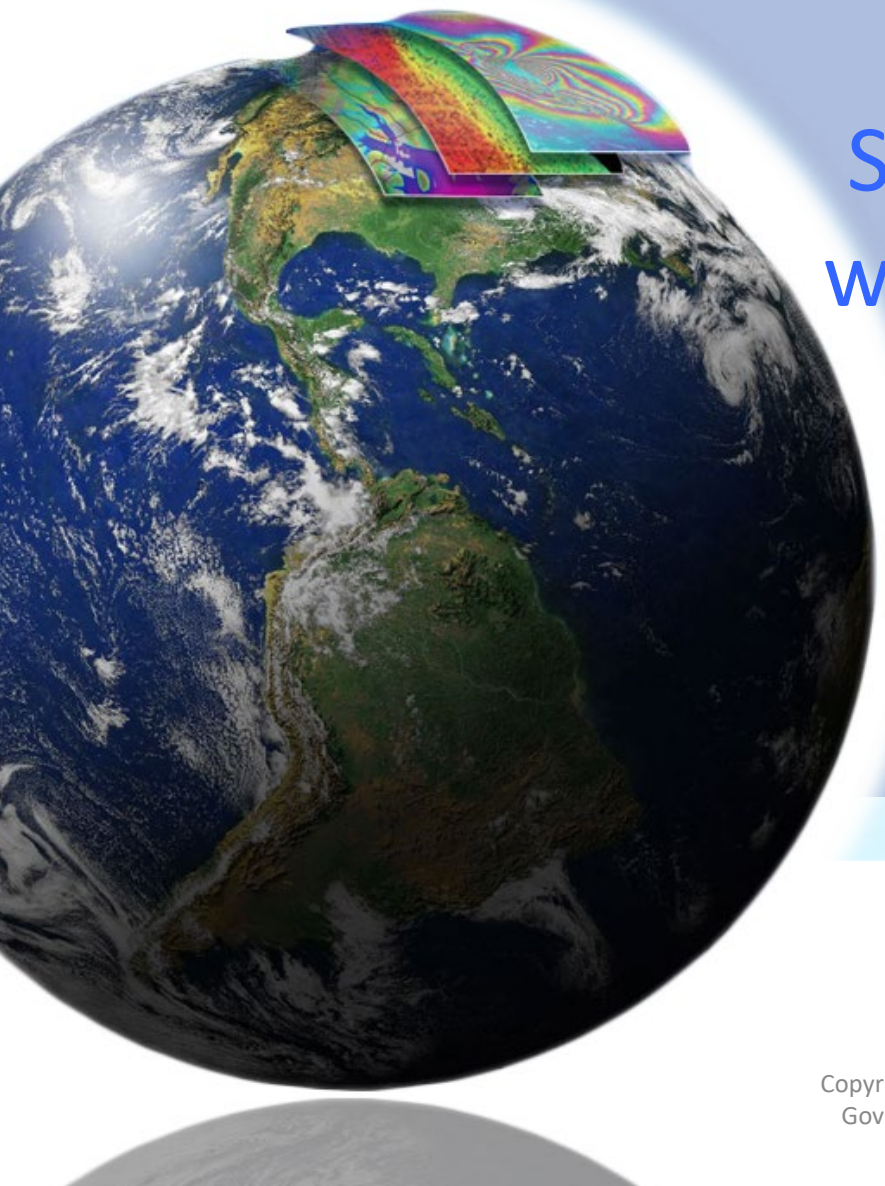


Large-scale Science Data Systems for SAR Missions, with On-Demand Machine Learning and Analysis-Ready Services

24th annual Ground System Architectures Workshop (GSAW)

Wednesday, March 4, 2020

Hook Hua (SDS Architect, NISAR, SWOT, MAAP, ARIA, UnitySDS)
Jet Propulsion Laboratory, California Institute of Technology.

Overview

- Large-scale science data systems (SDS) for next big NASA SAR missions
- Need for Analysis Ready Data (ARD) and Analysis Optimize Data ~~Storage~~ Services (AODS)
- Applications of machine learning for analysis on large SAR data streams
- On-Demand analysis via collocated algorithm development and analysis with large-scale SDSes

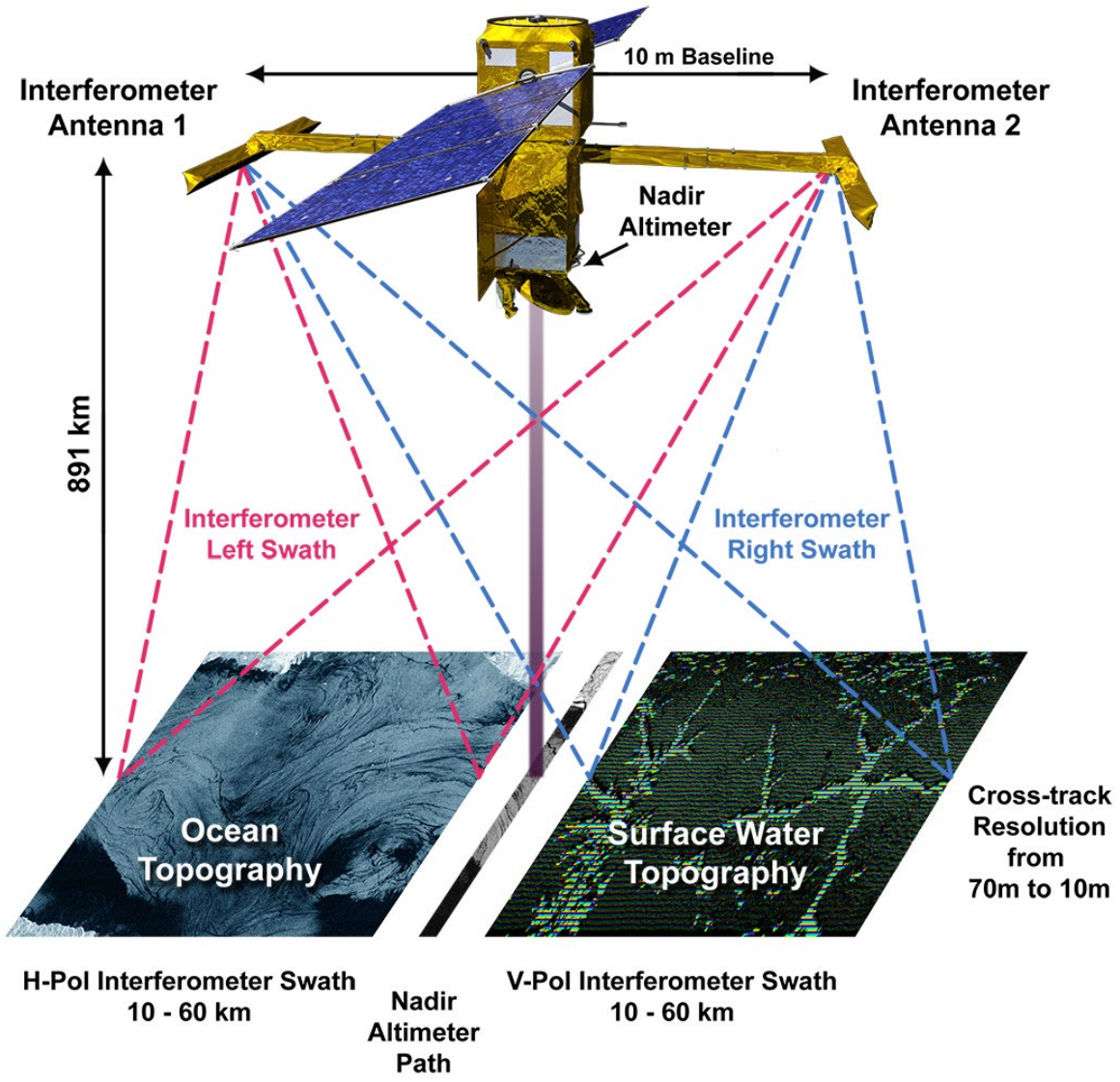
SWOT Mission Concept

Surface Water and Ocean Topography (SWOT)

Oceanography: Characterize the ocean mesoscale and sub-mesoscale circulation at spatial resolutions of 15 km and greater.

Hydrology: To provide a global inventory of all terrestrial water bodies whose surface area exceeds $(250m)^2$ (lakes, reservoirs, wetlands) and rivers whose width exceeds 100 m (rivers).

- To measure the global storage change in fresh water bodies at sub-monthly, seasonal, and annual time scales.
- To estimate the global change in river discharge at sub-monthly, seasonal, and annual time scales.



NISAR Mission Concept

NASA-ISRO SAR Mission (NISAR)

A dedicated U.S. and Indian InSAR mission, in partnership with ISRO, optimized for studying hazards and global environmental change.

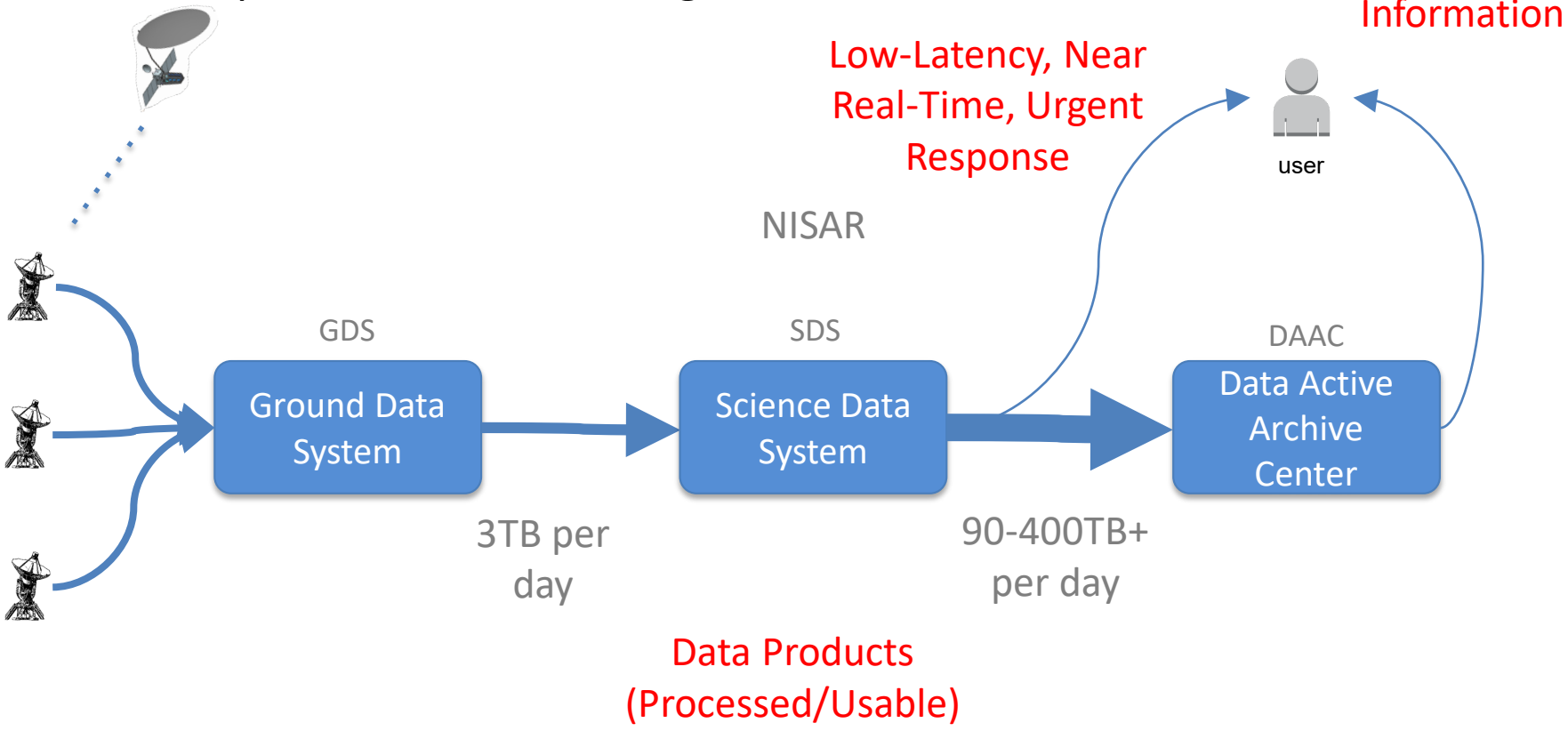
NISAR Characteristic:	Would Enable:
L-band (24 cm wavelength)	Low temporal decorrelation and foliage penetration
S-band (12 cm wavelength)	Sensitivity to light vegetation
SweepSAR technique with Imaging Swath >240 km	Global data collection
Polarimetry (Single/Dual/Quad)	Surface characterization and biomass estimation
12-day exact repeat	Rapid Sampling
3-10 meters mode-dependent SAR resolution	Small-scale observations
3 years since operations (5 years consumables)	Time-series analysis
Pointing control < 273 arcseconds	Deformation interferometry
Orbit control < 500 meters	Deformation interferometry
>30% observation duty cycle	Complete land/ice coverage
Left/Right pointing capability	Polar coverage, North and South
Noise Equivalent Sigma Zero ≤ -23 db	Surface characterization of smooth surfaces

Figure 1 – NISAR radar characteristics, as of Oct. 2015.

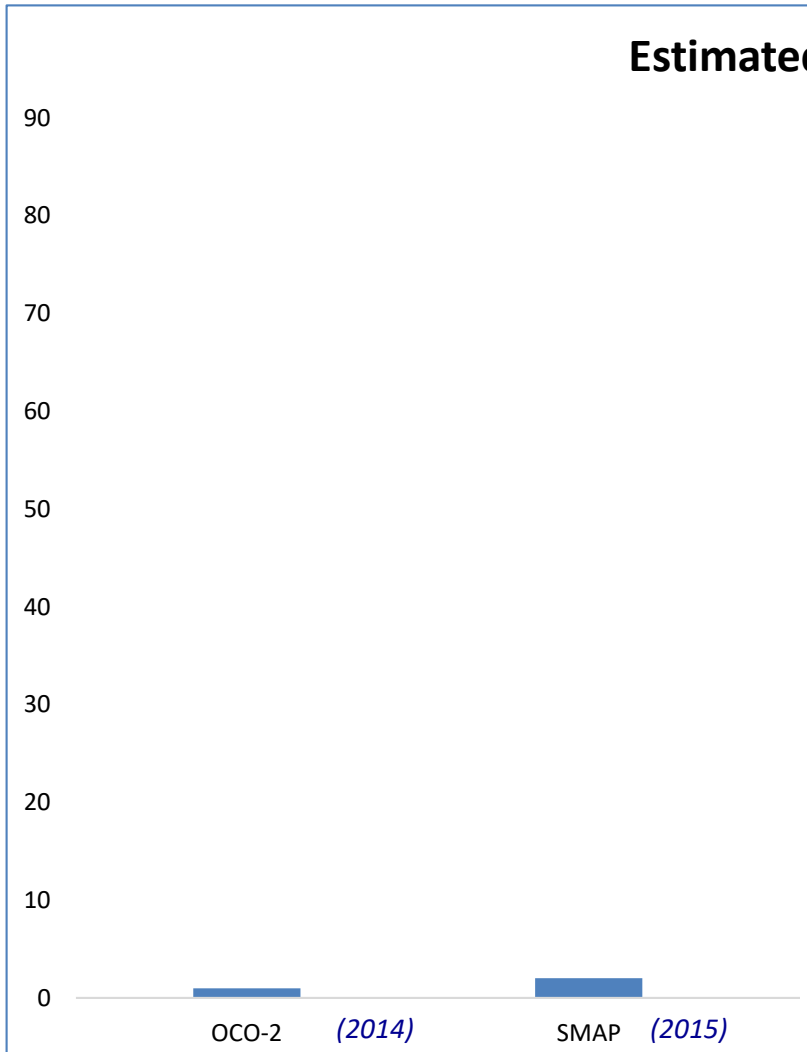
The NASA-ISRO Synthetic Aperture Radar (SAR), or NISAR, Mission will make global integrated measurements of the causes and consequences of land surface changes. NISAR will provide a means of resolving highly spatial and temporally complex processes ranging from ecosystem disturbances, to ice sheet collapse and natural hazards including earthquakes, tsunamis, volcanoes, and landslides.

Large Data Streams

- GDS handles global downlinks and ground network
- SDS generates science data products
- DAACs provide access, storage, and services to end-users



Next Generation Earth Science Remote Sensing Mission SDSes

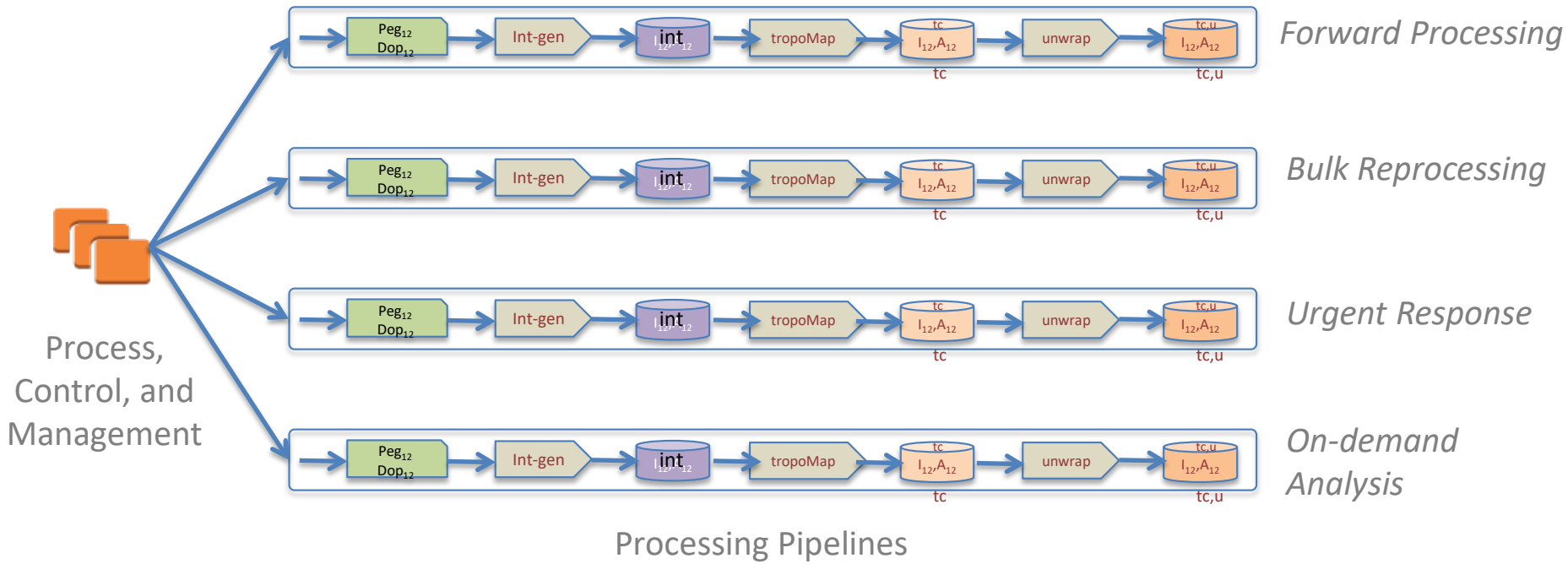


The volume of data produced is more complex and larger than previous missions

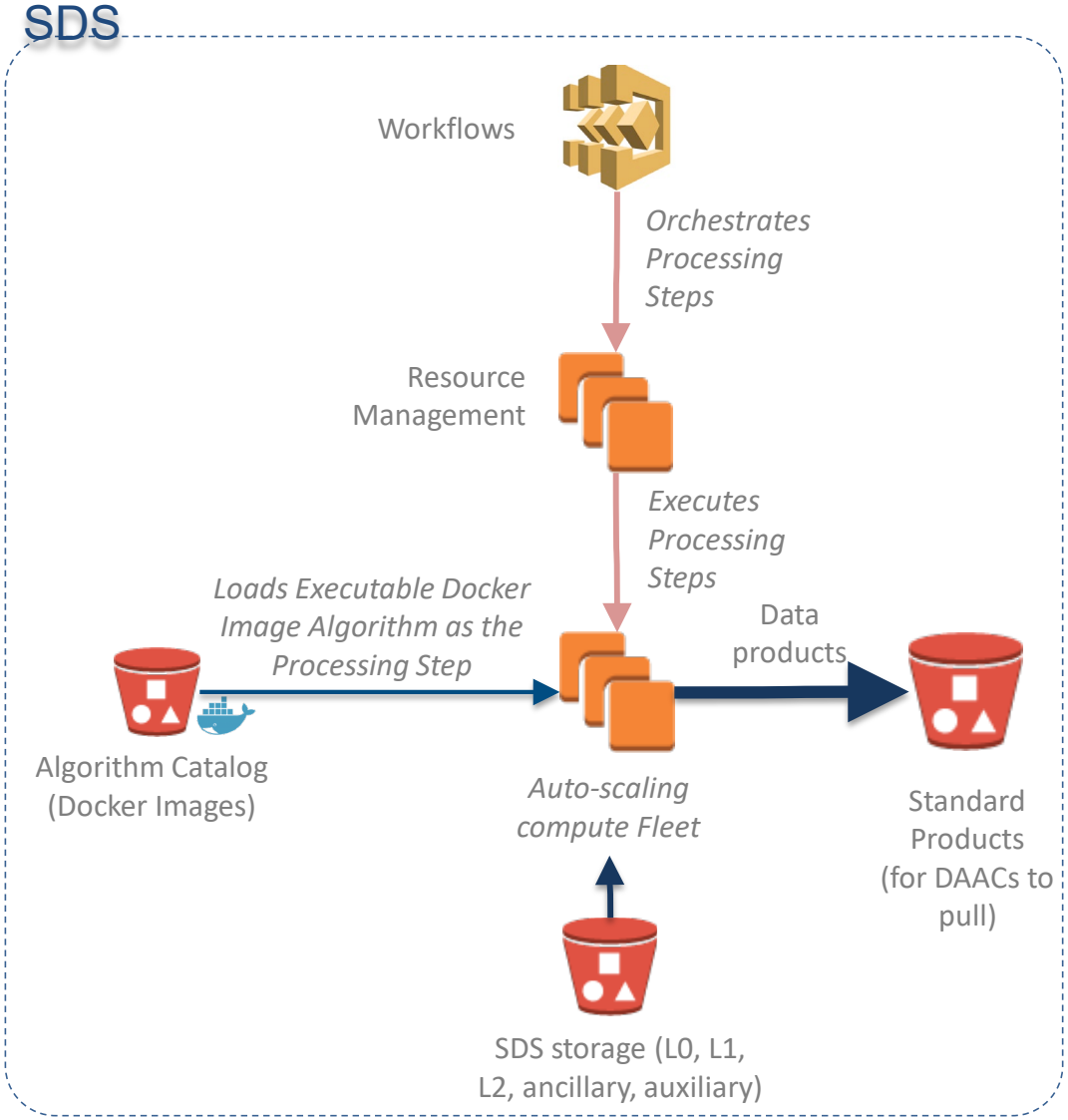
Data *storage, processing, movement, costs, and agility* are the biggest challenges

Concurrent Processing Pipelines

- Concurrently keep up with:
  - Forward processing (“keep up”)
  - Bulk (re)processing
  - Urgent response
  - Near-real time (NRT)
  - On-demand analysis

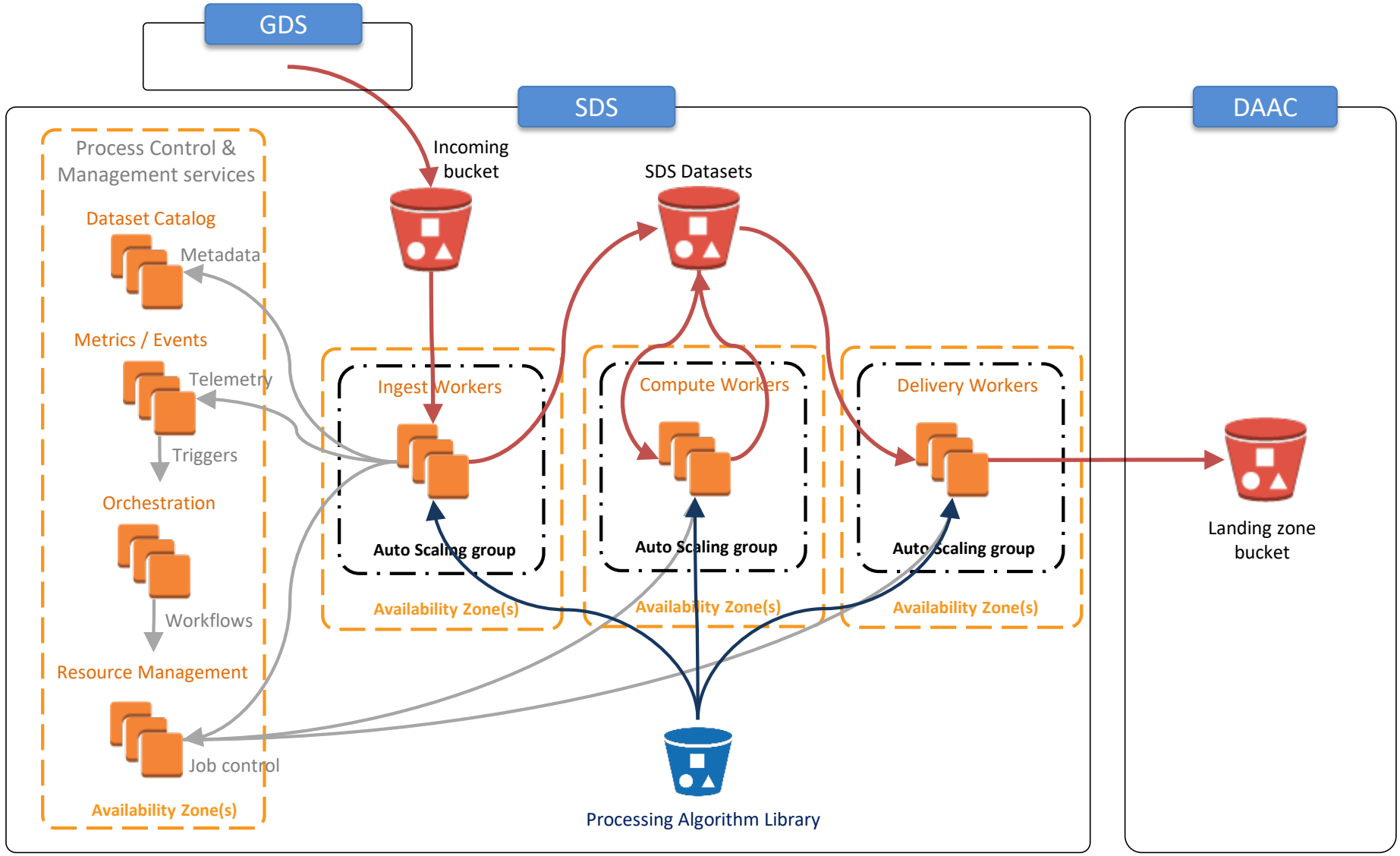


On-Demand Processing Step

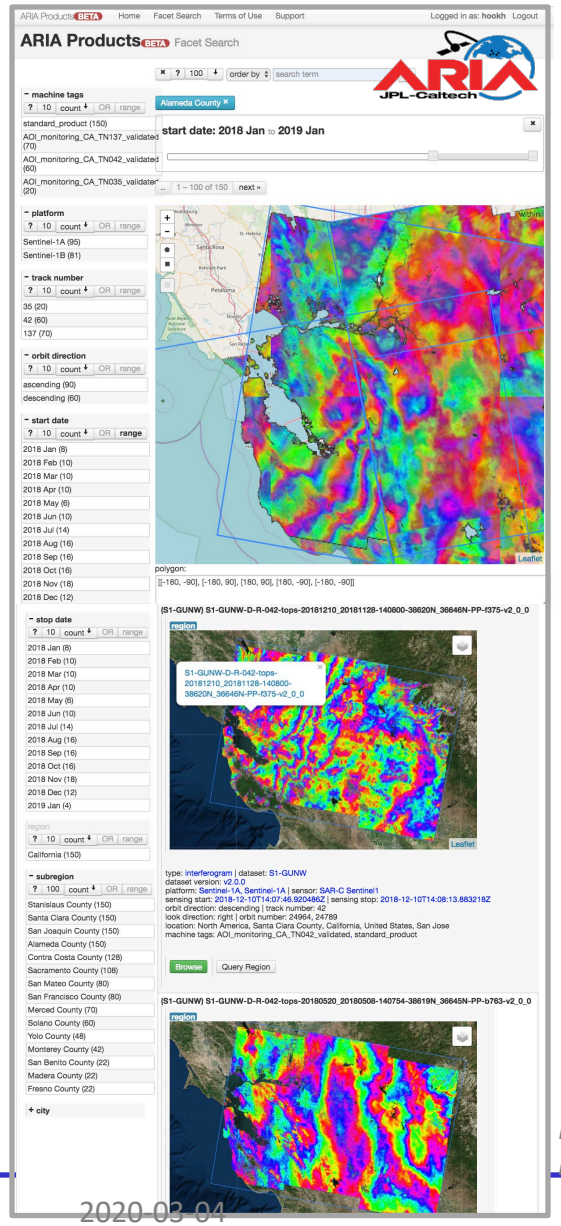


- Product Generation Executives (PGEs) are the algorithm implementations that generate the data products.
- Represented as **Docker** images
- Allows algorithms to be **versionized, archived, preserved, and resurrected** for data processing
- PGEs are stored in an "*Algorithm Catalog*"
- Dynamically loaded into the SDS' compute fleet of workers for science data processing
- Used to generate **on-demand GeoTIFFs, CoGeo**, etc.
- On-demand processing is **more than just PGEs**. It includes production rules, anc/aux, etc.

Algorithm Library to Cloud Processing

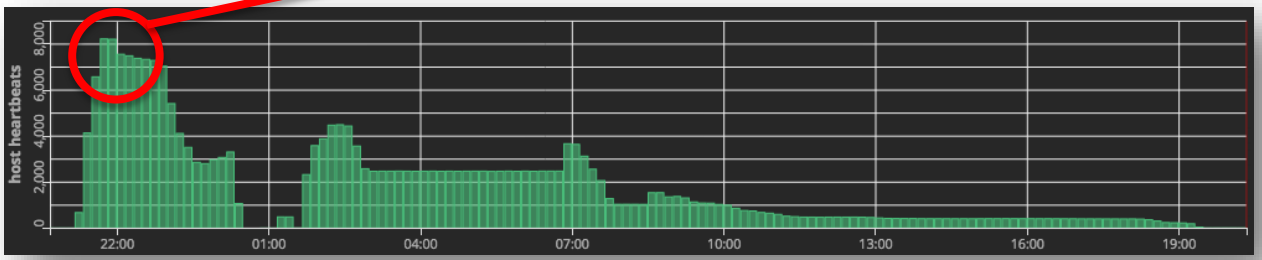


Mitigate Engineering Risk for NISAR SDS



- *SDS demonstrated cloud can support high rates*
- Prototyped NISAR SDS under "Getting Ready For NISAR" (GRFN) project
- Processing and Delivery of L2 Data Products
  - Successful demonstration of data processing and deliveries from JPL SDS to ASF DAAC
  - Demonstrated up to NISAR 5X rate in AWS
- Towards L3 Time Series
  - L3 in NISAR Cal/Val
  - Processing entire Sentinel-1A/B coverage period
- Multiple SDS Processing Modes
  - Bulk (re)processing
  - Forward "keep-up" processing (running 24/7)
  - On-demand processing

NISAR 5X (430TB+ per day)
~8,300 compute nodes
260K+ cores
2PB+ of scratch disk



Live SDS production at https://aria-products.jpl.nasa.gov/

Sentinel-1A/B Processing with Diversified Portfolio of AWS "Spot Market" Compute Instances

- Leverage excess capacity for cost reduction
- Spot Instances → Robustness
- Use Auto Scaling Fleet to spread compute across a diversity of ec2 instance types
- Different compute instance types yields different price-performance ratio

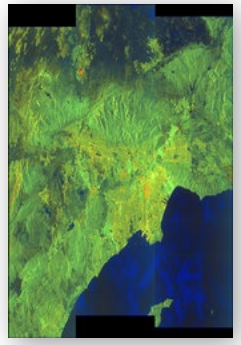
Name	API Name	Memory	vCPUs	Instance Storage	Network	On-demand	Spot (us-west-2)
C5 High-CPU 9xlarge	c5.9xlarge	72.0 GiB	36 vCPUs	EBS only	10 Gigabit	$1.530 hourly	$0.55 hourly
C5 High-CPU Quadruple Extra Large	c5.4xlarge	32.0 GiB	16 vCPUs	EBS only	Up to 10 Gigabit	$0.68 hourly	$0.33 hourly
C3 High-CPU Quadruple Extra Large	c3.4xlarge	30.0 GiB	16 vCPUs	320 GiB (2 * 160 GiB SSD)	High	$0.840 hourly	$0.25 hourly
C3 High-CPU Eight Extra Large	c3.8xlarge	60.0 GiB	32 vCPUs	640 GiB (2 * 320 GiB SSD)	10 Gigabit	$1.680 hourly	$0.49 hourly
I3 High I/O Quadruple Extra Large	i3.4xlarge	122.0 GiB	16 vCPUs	3800 GiB (2 * 1900 GiB NVMe SSD)	Up to 10 Gigabit	$1.248 hourly	$0.50 hourly
I3 High I/O Extra Large	i3.xlarge	30.5 GiB	4 vCPUs	950 GiB NVMe SSD	Up to 10 Gigabit	$0.312 hourly	$0.10 hourly

NASA Data Product Levels

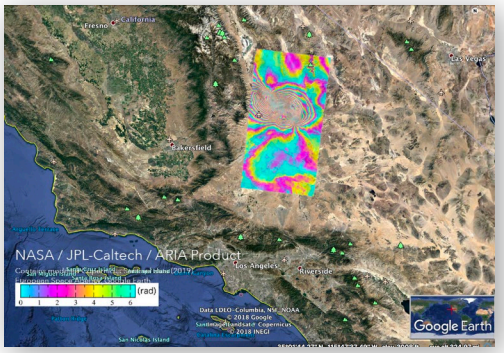
Information Increases

Information Decreases

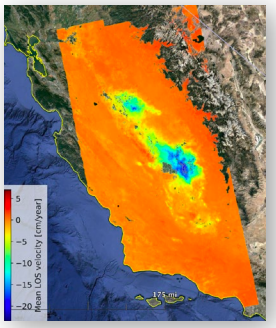
- Level 1: calibrated pixel radiances



- Level 2: geophysical parameters



- Level 3: time series, [globally] gridded averages

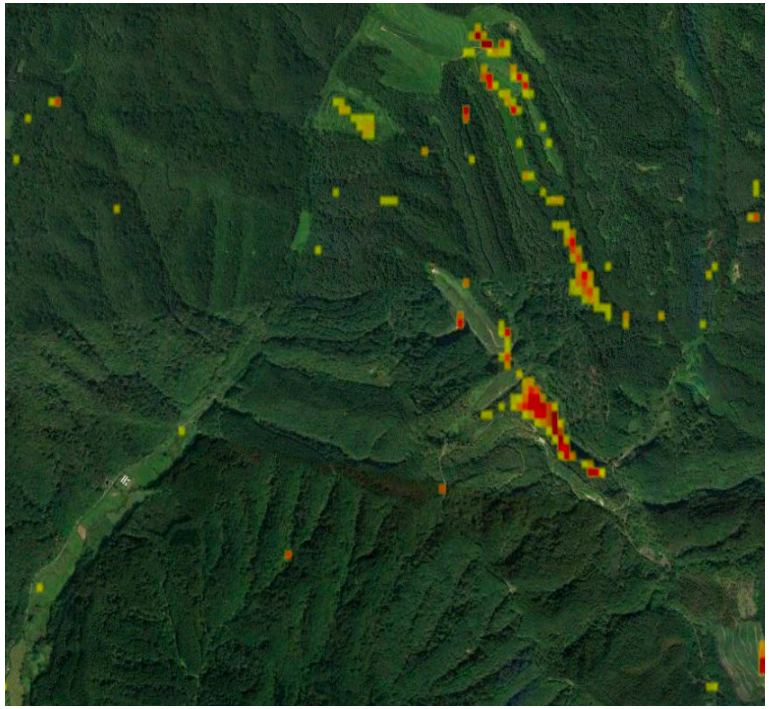


Multitemporal High Resolution L2 SAR Stacks

- Trends in retaining high resolution L2 data stacks for multi-temporal analysis

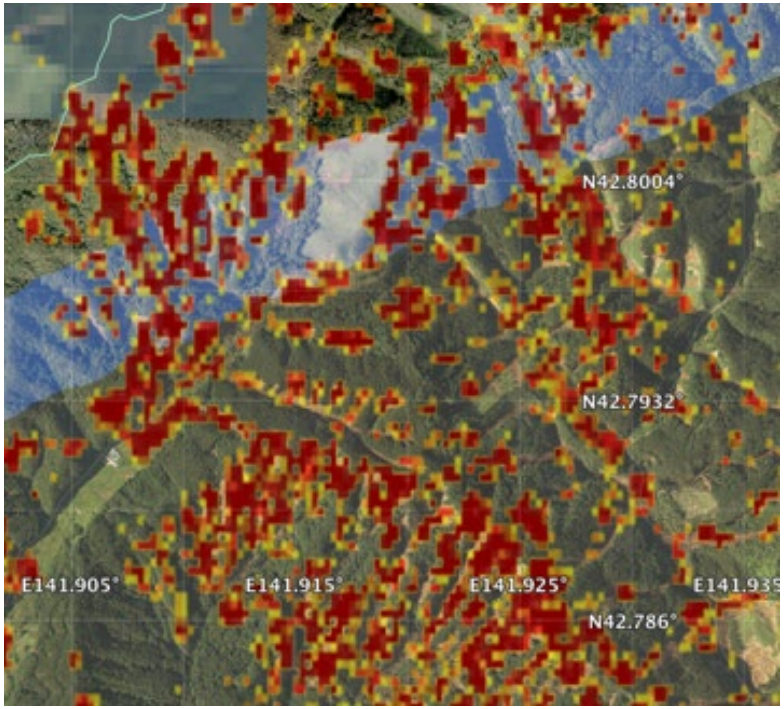
Source: Sang-Ho Yun (JPL)

DPM1



Before/After Scenes
Processing: 1 hour
"Downloading": 1.5 hours

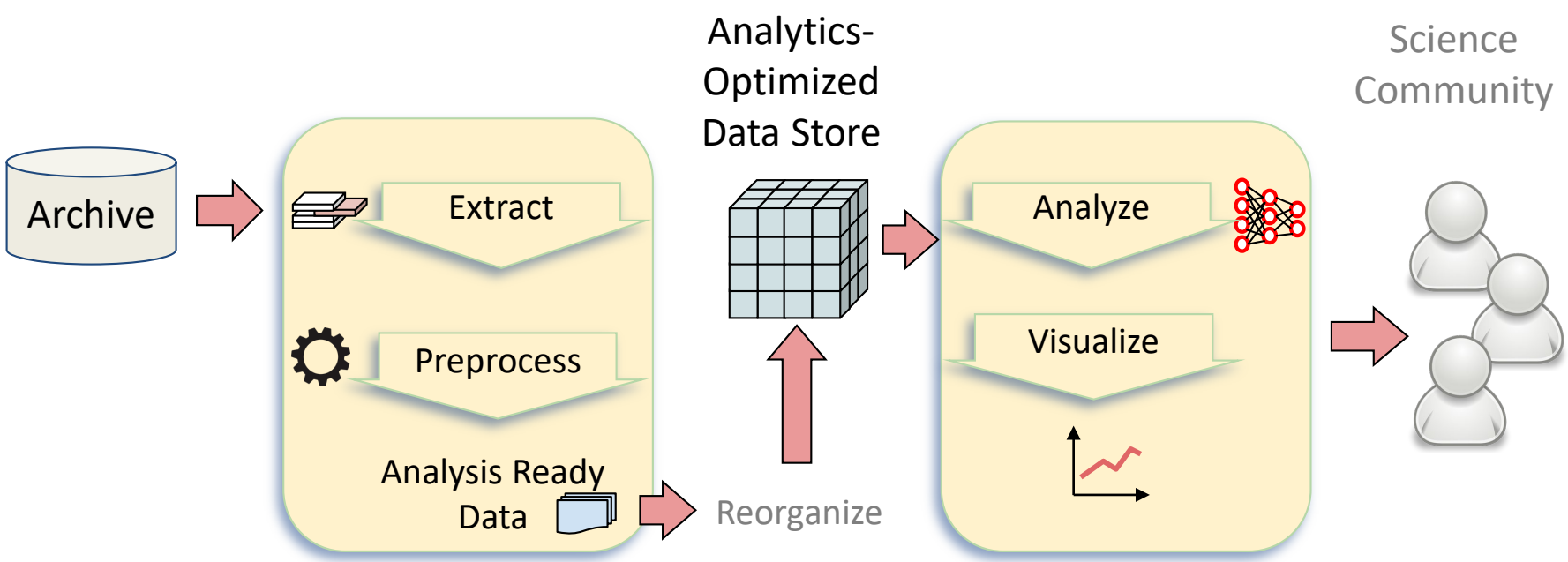
DPM2/3



Time Series of Scenes
Processing: 26 days → a few hours ?
"Downloading": 40 hours → 0.5 hours ?

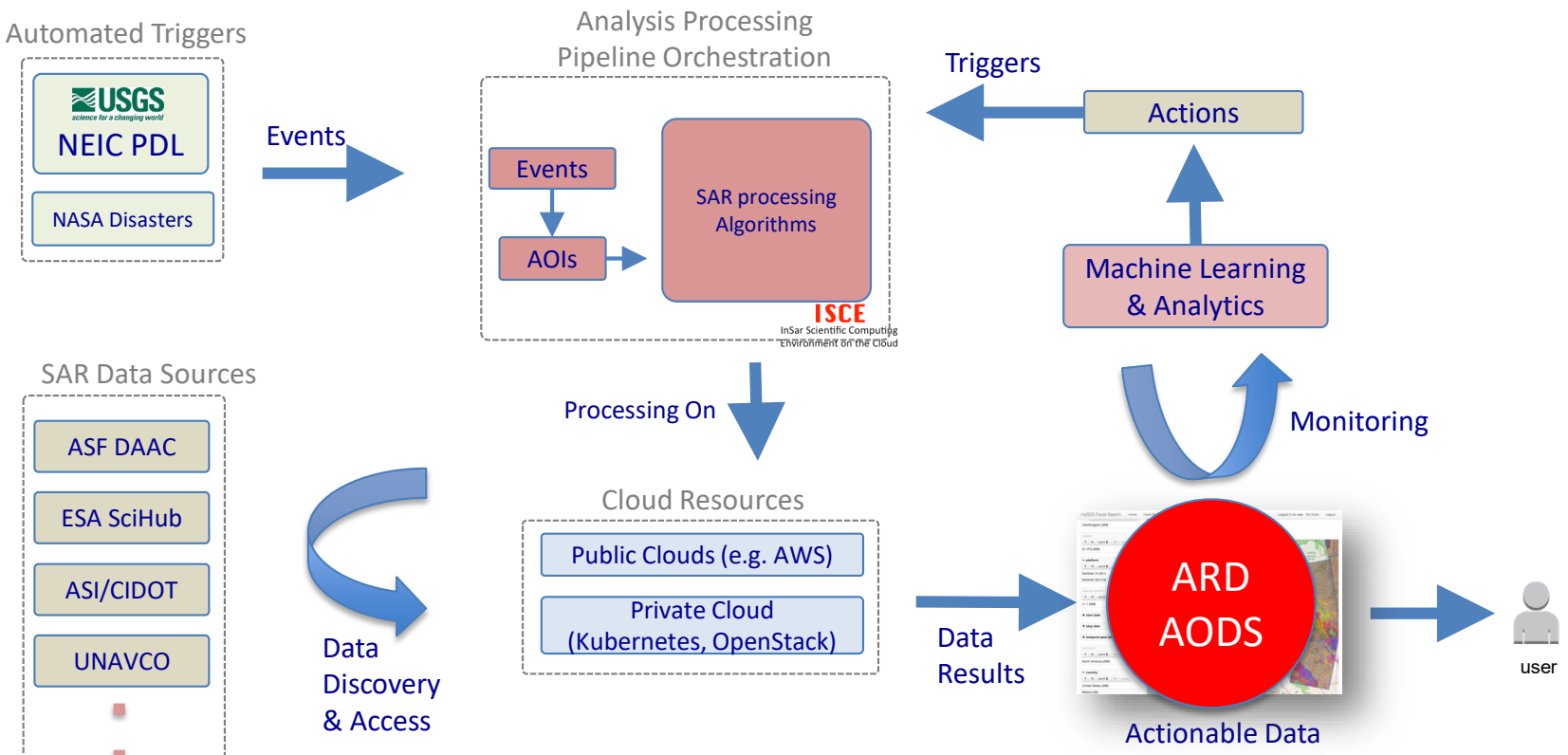
Landslides Triggered by the M6.6 Hokkaido Earthquake (Sept 2018)

Analysis-Ready Data vs. Analytics Optimized Data Stores



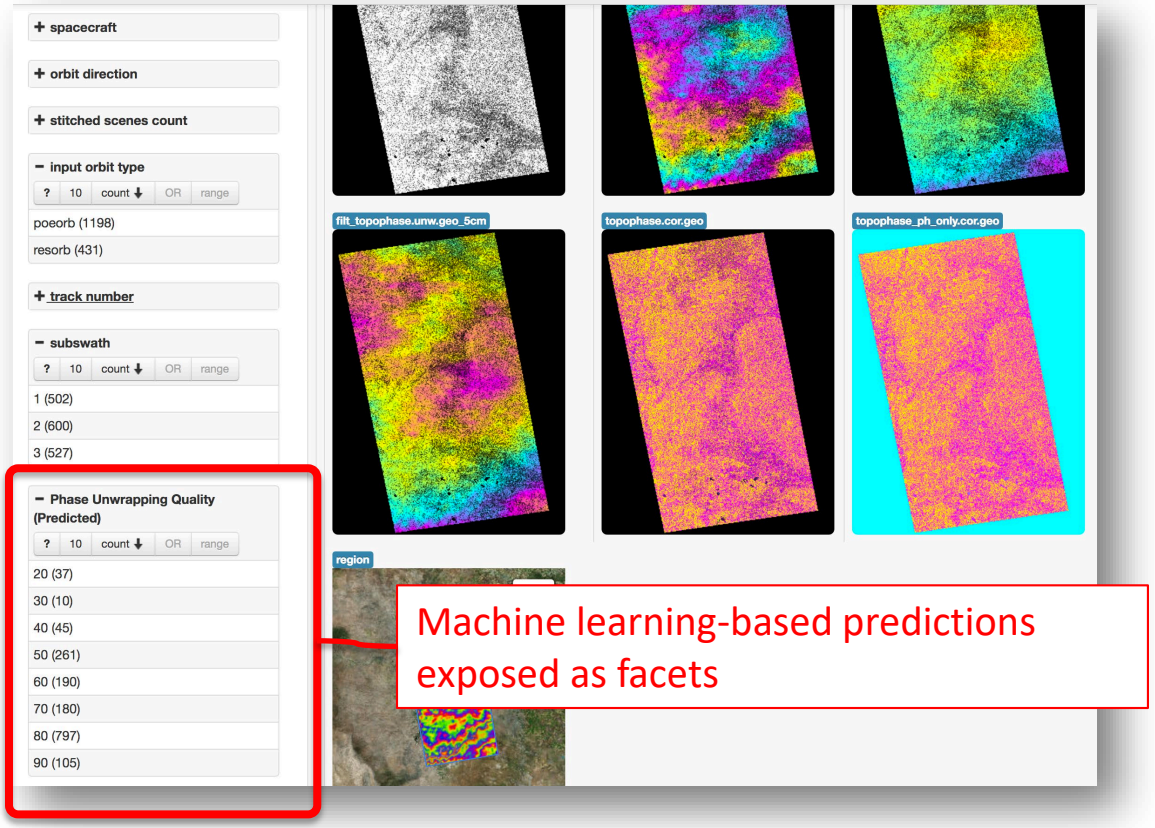
Source: Chris Lynnes (NASA ESDIS)

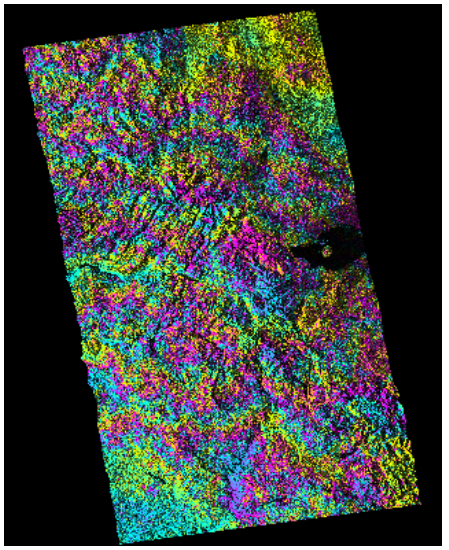
Automating Monitoring & On-Demand Analysis



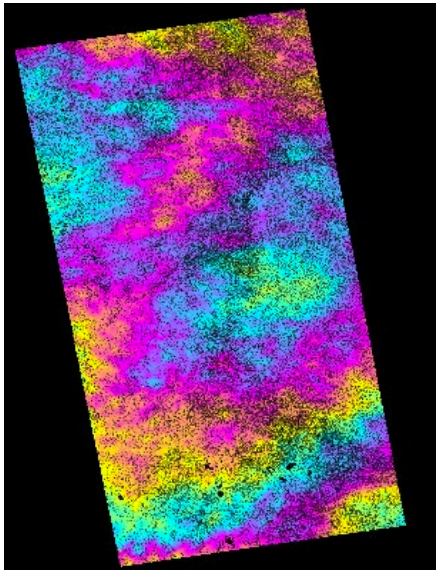
- Events triggers workflows of SAR analysis pipeline
- Continuous monitoring and analysis
- Urgent response: coregisted stack processing towards damage mapping
- Processing/Analysis scales up in cloud
- ML applied on science data product results
  - Triggers successive processing

Machine Learning for SAR Data Quality Screening





Predicted **20%** quality for phase unwrapping



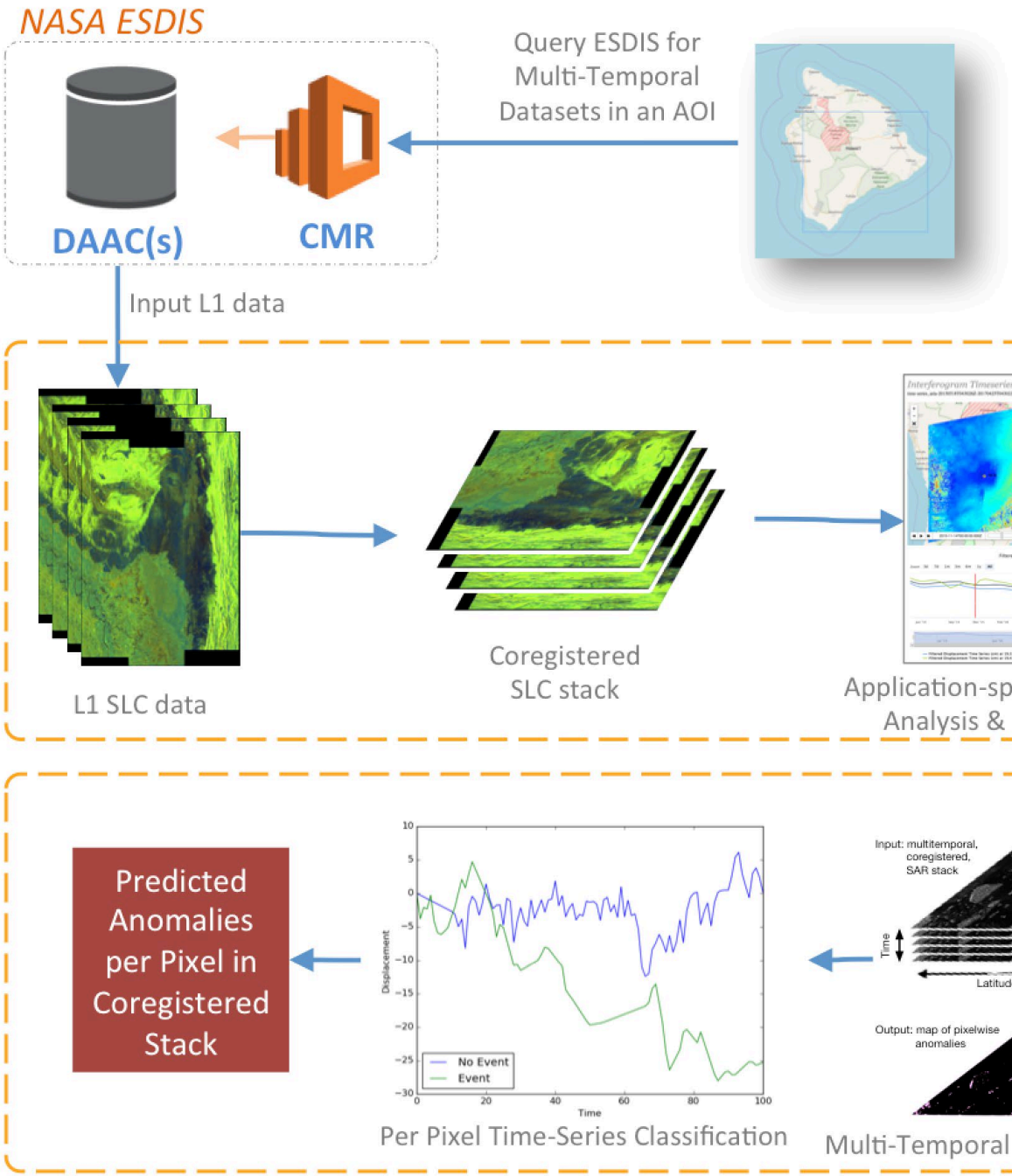
Predicted **80%** quality for phase unwrapping

- *Scalable assessment of data quality with InSAR phase unwrapping*
- Triage of “problematic” interferograms
- Selection of only high-quality interferograms for time series generation

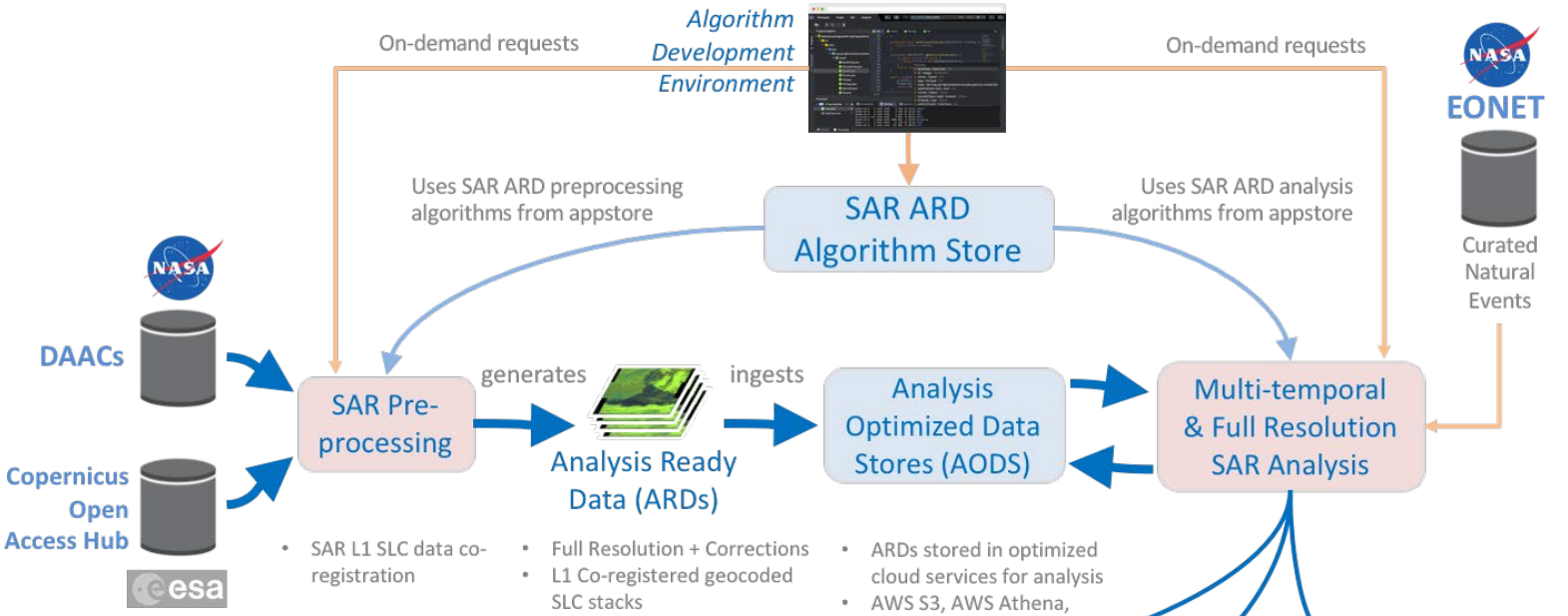
On-Demand Usage of ESDIS DAAC data holdings

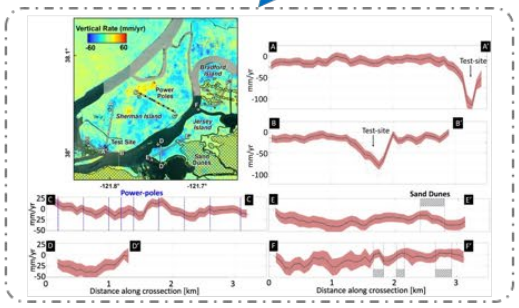
Cloud-Native Multi-Temporal Data Pre-Processing

Cloud-Native Machine Learning for Multi-Temporal Anomaly Detection

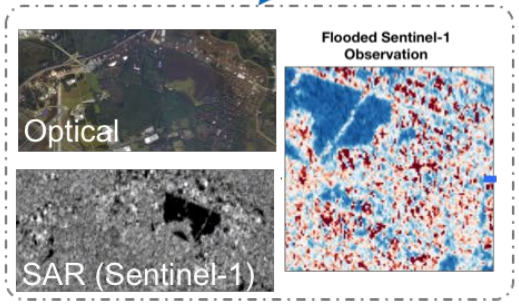


On-Demand Analysis with ARD/AODS

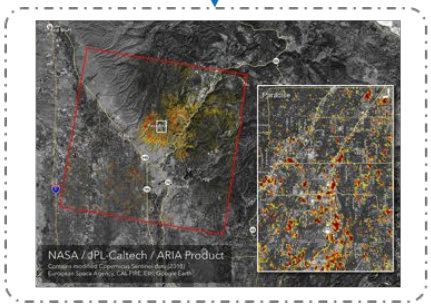




Full Resolution Time Series Analysis
(critical infrastructure monitoring, landslides, etc)



Multi-temporal-based Flood Assessment Maps



Multi-temporal-based Damage Proxy Maps

Key Points

- Increasing need for large-scale processing
- SDS already in the cloud
- ”Data Lake”
- For SAR applications--especially in change detection, observing shifts away from
  - Single scene analysis
  - Pair-wise analysis
- ...and towards high resolution multi-temporal L2 data stacks
- Need for collocated algorithm development environment in the cloud
  - Jupyter notebooks → on-demand processing
- To ARD and AODS

THANK YOU!