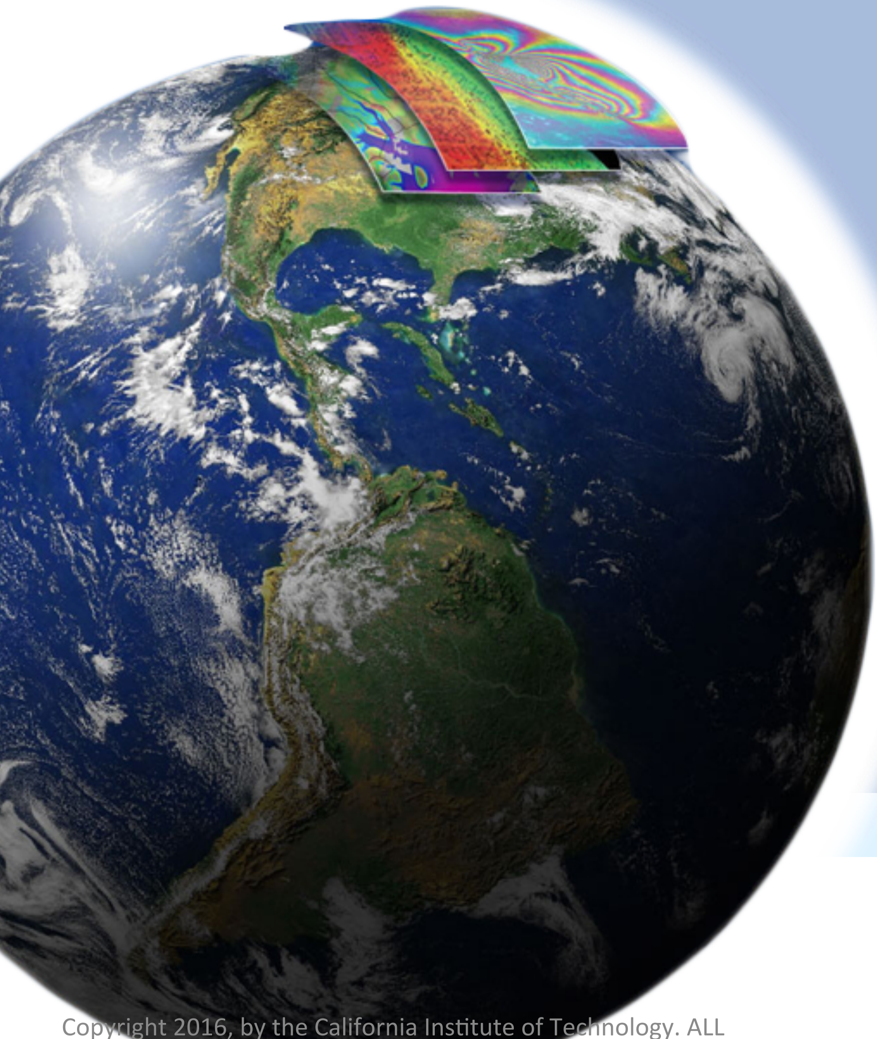
Jet Propulsion Laboratory
California Institute of Technology



Agile Big Data Analytics of High-Volume Geodetic Data Products for Improving Science and Hazard Response

2016 Earth Science Technology Forum (ESTF2016)
Annapolis, MD
Wednesday, June 15, 2016

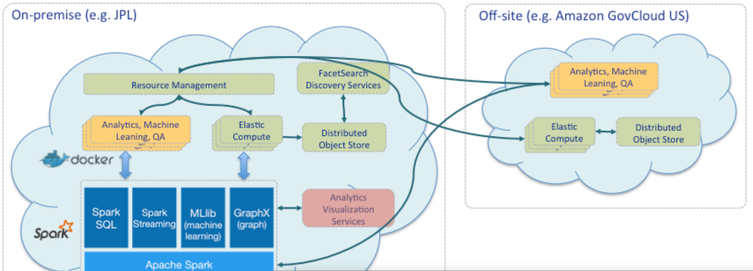
Hook Hua[1], Susan Owen[1], Gerald Manipon[1], Gian Franco Sacco[1], Piyush Agram[1], Brian Bue[1], Michael Starch[1], Lan Dang[1], Justin Linick[1], Eric Fielding[1], Sang-Ho Yun[1], Paul Lundgren[1], Angelyn Moore[1], Paul Rosen[1], Zhen Liu[1], Eric Gurrola[1], Tom Farr[1], Vincent Realmuto[1], Frank Webb[1], Mark Simons[2], Pietro Milillo[1]

[1] Jet Propulsion Laboratory
[2] California Institute of Technology

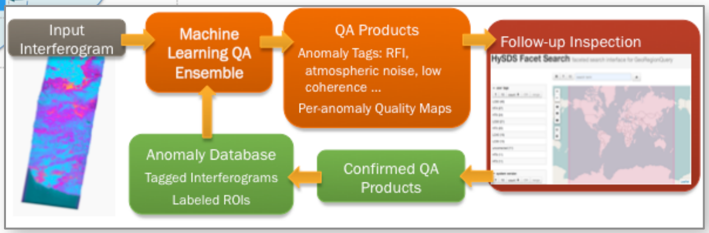
Agile Big Data Analytics of High-Volume Geodetic Data Products for Improving Science and Hazard Response

Objective

- Develop an advanced hybrid-cloud computing science data system for easily performing massive-scale analytics of geodetic data products. This includes:
  - Improving the quality of automated data product generation of high-volume and low-latency NASA Solid Earth science data products to support hazards monitoring.
  - Enabling end-user analysis to be performed on increasing collections of InSAR and GPS data in order to improve the understanding, quality, and features of the data



Scalable Big Data Analytics with hybrid-cloud computing



Machine learning-based QA to detect anomalies.

Approach:

- Leverage prior AIST investments on our hybrid-Cloud Computing data system and augment with capabilities for Big Data Analytics for scaling up onto PB-scale datasets.
- Leverage metrics for testing and comparing different processing strategies and time series analysis on large and comprehensive data sets.
- Employ machine-learning approaches to automate QA on massive scales across our generated high-volume data products.

Key Milestones

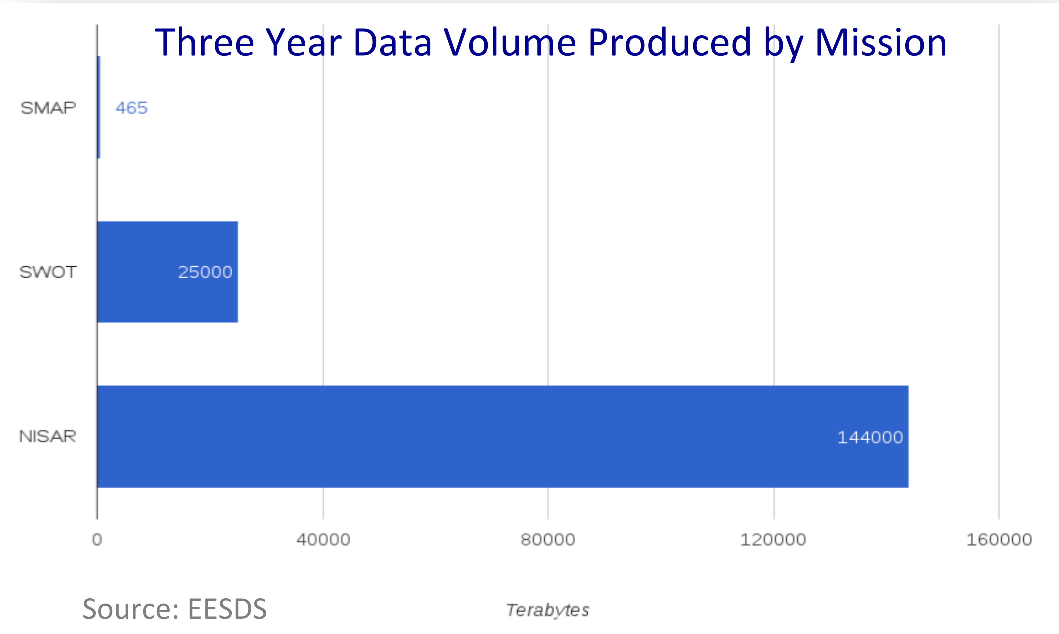
• Augment hybrid-Cloud Science Data System for running analytics.	12/15
• Machine Learning-based Analytics for QA	06/16
• Science Analytics for InSAR Coherence Time Series and Troposphere Corrections.	12/16
• Visualization of Analytics via faceted analytic metrics and fast browse	12/16
• Expanded Machine Learning-based Analytics for QA	06/17
• Science Analytics for InSAR and GPS Timeseries Analysis	06/17

TRL_{in} = 3 TRL_{out} = 5

Flood of Multi-sensor SAR Data

- Remote sensing SAR
  - ALOS PALSAR 1 and 2
  - RADARSAT 1 and 2
  - TerraSAR-X
  - COSMO-SkyMed (CSK)
    - CSK constellation of 4 satellites has acquisition capacity of 450 frames/day for each satellite
      - 1 frame = 40 x 40 km swath, 3m resolution, 1.2 Gb
  - Sentinel 1A/1B
    - C-band SAR
    - 1.8TB/day L0 data, each
  - NISAR (2021)
    - L-band SAR
    - Nominally **~95 TB per day**
    - Peak at **~150TB per day**
  - SWOT (2021)
    - Nominally **~16 TB per day**
    - Peak at **~71TB per day**
  - Etc.

- Airborne SAR
  - UAVSAR
  - AfraSAR
  - Etc.



Three Year Data Volume Produced by Mission

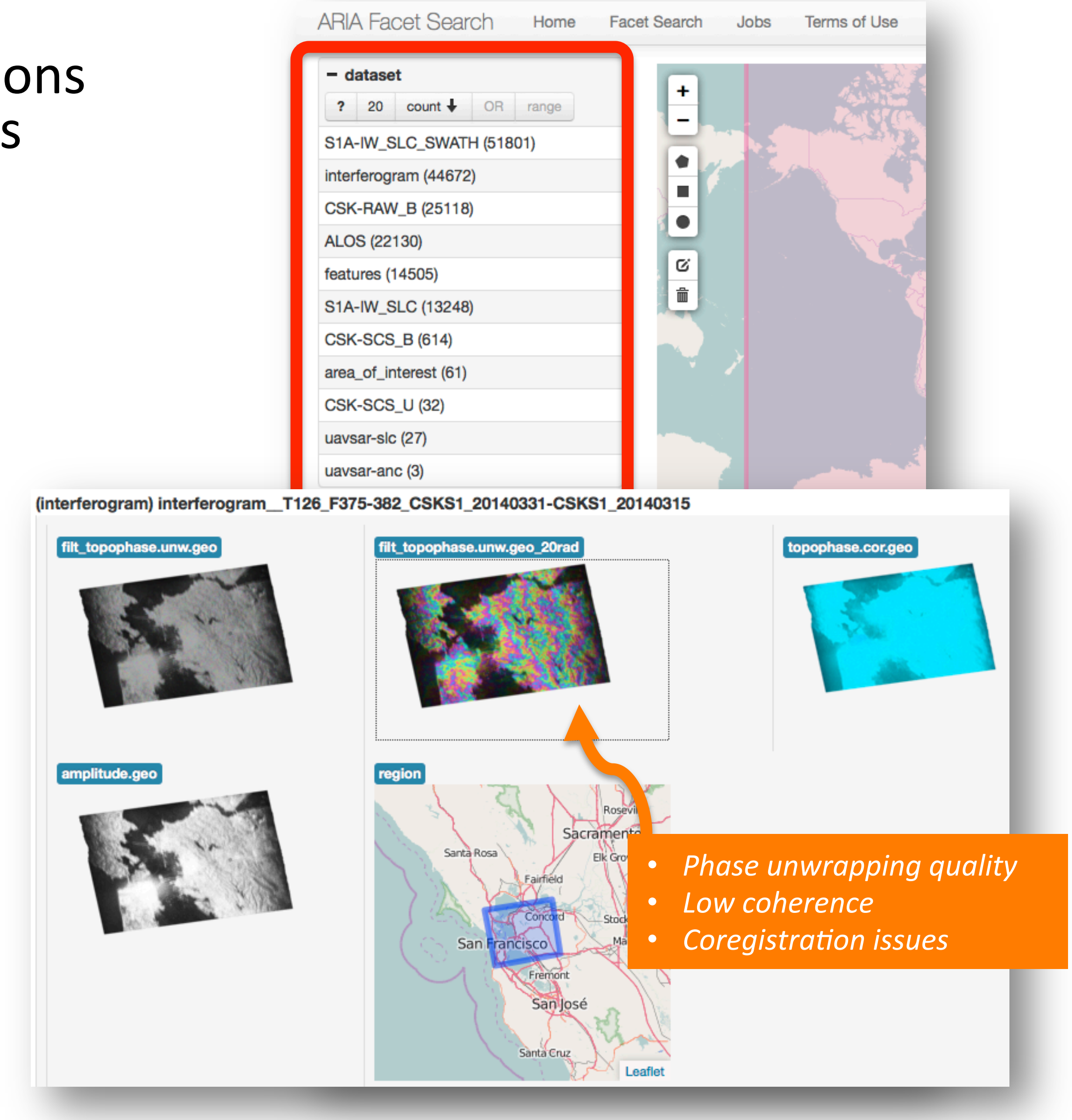
Source: EESDS

Voluminous SAR Data and QA

- **Voluminous** SAR Data
  - Increasingly large collections of L0 to L2 interferograms
  - Derived L3 data products
    - Solid earth, ecosystems, cryosphere

- Various **quality issues** associated with data, processing, instrument, terrain

Need for scalable and automated quality assessment in [SAR] science data systems


ARIA Facet Search | Home | Facet Search | Jobs | Terms of Use

− dataset

? | 20 | count ↓ | OR | range

S1A-IW_SLC_SWATH (51801)
interferogram (44672)
CSK-RAW_B (25118)
ALOS (22130)
features (14505)
S1A-IW_SLC (13248)
CSK-SCS_B (614)
area_of_interest (61)
CSK-SCS_U (32)
uavsar-slc (27)
uavsar-anc (3)

(interferogram) interferogram__T126_F375-382_CSKS1_20140331-CSKS1_20140315

- *Phase unwrapping quality*
- *Low coherence*
- *Coregistration issues*

Scalable Hybrid Cloud Science Data System with Machine Learning-based QA

- **Agile** and **large-scale** analysis
  - End-to-end ML is a key analysis capability that was scaled up in cloud computing
  - Runs *on private cloud* (OpenStack) and *public cloud* (AWS)
  - Migration to *Containers*

- **Crowd sourced faceted labeling**
  - Used as "truth" training data
- **Machine Learning (ML)** on SAR data
  - Feature extraction, model training, prediction
  - Initially applied to CSK sensor data stream
- **Quality assessment (QA)**
  - phase unwrapping

Collaborative QA and Faceting

COLLABORATIVE TAGGING FOR QUALITY ASSESSMENT

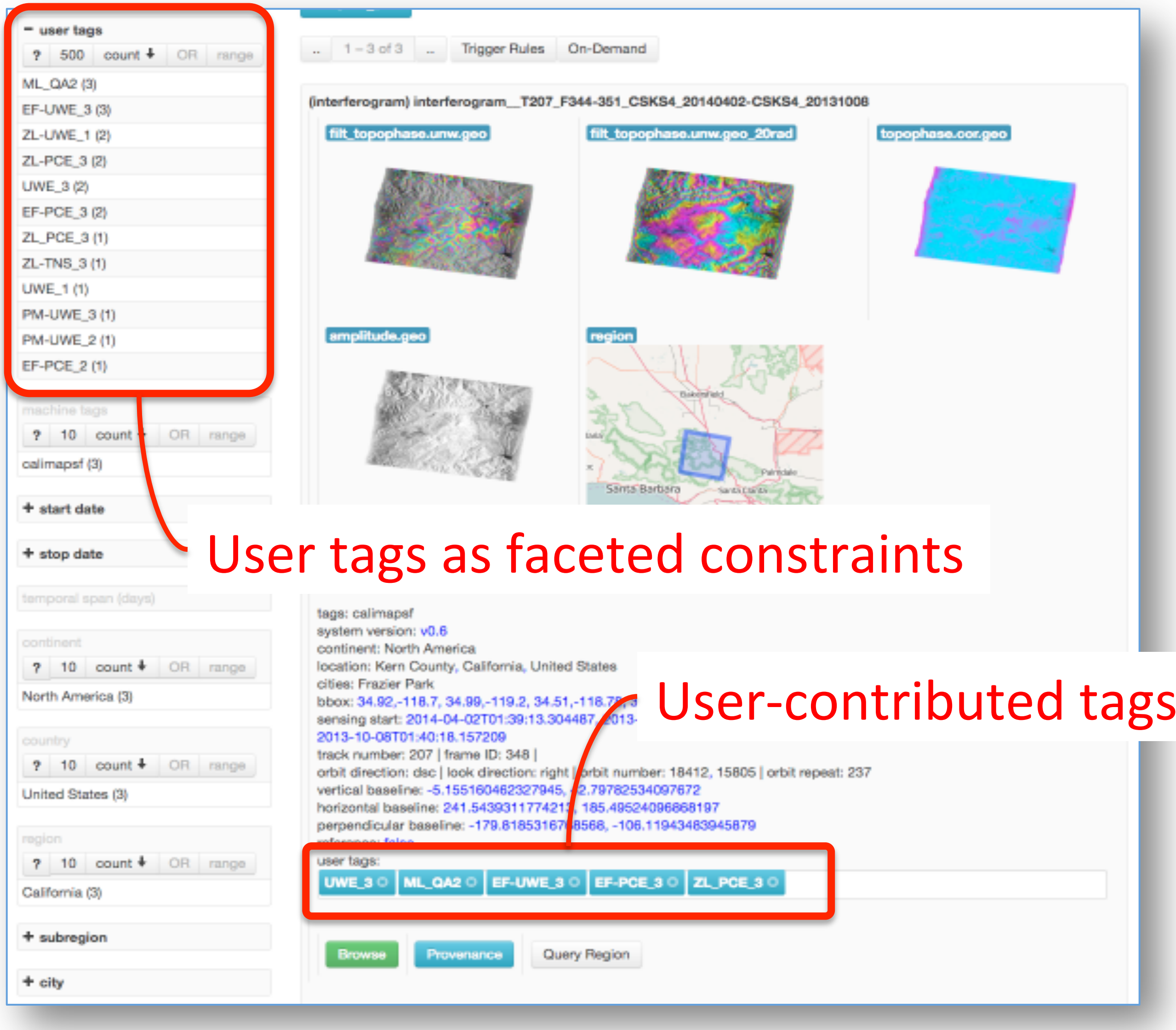
Quality Assessment (QA) Metrics

- Leverage user tagging of data products to help identify quality issues
- Identifies issue and intensity
- Ensure identification of both good and bad qualities

- Intensity Levels
  - 0: clean
  - 1: mild
  - 2: medium
  - 3: severe

Tag	Metric
UWE_{intensity}	unwrapping errors
TNS_{intensity}	troposphere noise - stratified
TNT_{intensity}	troposphere noise - turbulent
ION_{intensity}	ionospheric noise
ORB_{intensity}	orbital ramps
DME_{intensity}	DEM error
PML_{intensity}	process errors - missing lines
PMR_{intensity}	process errors - misregistration
PCE_{intensity}	process errors - chirp extension noise
RFI_{intensity}	radio frequency interference
COH_{intensity}	coherence

QA as Faceted User Tags



ML_QA2 (3)
EF-UWE_3 (3)
ZL-UWE_1 (2)
ZL-PCE_3 (2)
UWE_3 (2)
EF-PCE_3 (2)
ZL_PCE_3 (1)
ZL-TNS_3 (1)
UWE_1 (1)
PM-UWE_3 (1)
PM-UWE_2 (1)
EF-PCE_2 (1)

calimapsf (3)

North America (3)

United States (3)

California (3)

(interferogram) interferogram__T207_F344-351_CSKS4_20140402-CSKS4_20131008

tags: calimapsf
system version: v0.6
continent: North America
location: Kern County, California, United States
cities: Frazier Park
bbox: 34.92,-118.7, 34.99,-119.2, 34.51,-118.7
sensing start: 2014-04-02T01:39:13.304487, 2013-
2013-10-08T01:40:18.157209
track number: 207 | frame ID: 348 |
orbit direction: dsc | look direction: right | orbit number: 18412, 15805 | orbit repeat: 237
vertical baseline: -5.155160462327945, -2.79782534097672
horizontal baseline: 241.5439311774212, 185.49524096868197
perpendicular baseline: -179.8185316768568, -106.11943483945879
reference: false
user tags:
UWE_3 ML_QA2 EF-UWE_3 EF-PCE_3 ZL_PCE_3

User tags as faceted constraints

User-contributed tags

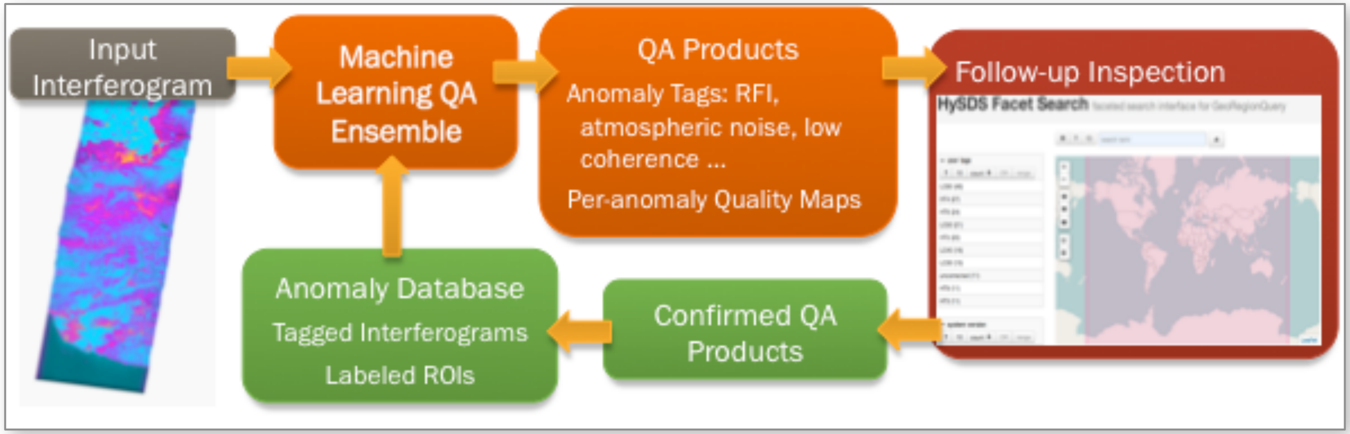
Tag	Count
UWE_0 (pristine)	613
UWE_1 (minor)	491
UWE_2 (major)	340
UWE_3 (severe)	70
Total	**1514**

Quality Assessment

MACHINE LEARNING

ARIA Machine Learning for Automated Quality Assessment

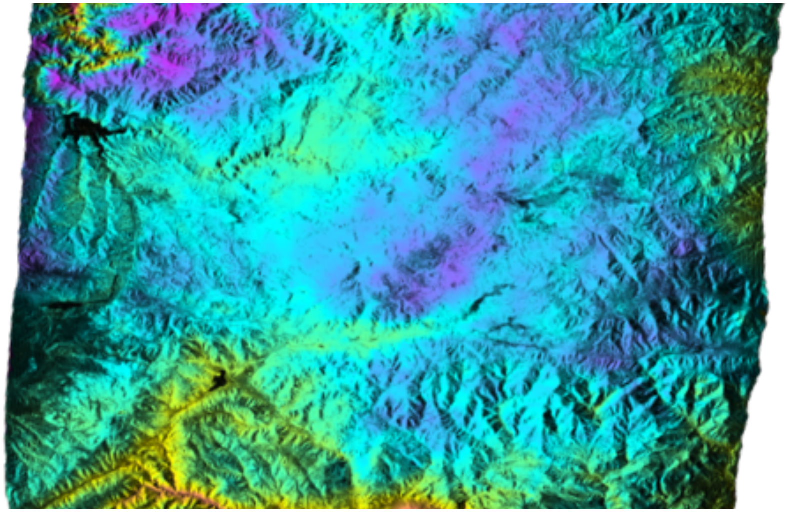
- An advanced hybrid-cloud computing science data system for easily performing massive-scale analytics of geodetic data products.
  - Enables end-user analysis to be performed on increasing collections of geodetic data in order to improve the understanding, quality, and features of the data
  - Employ machine-learning approaches to automate QA on massive scales across our generated high-volume data products.
- Novel capabilities
  - Large-scale and machine-learning-based analytics on QA



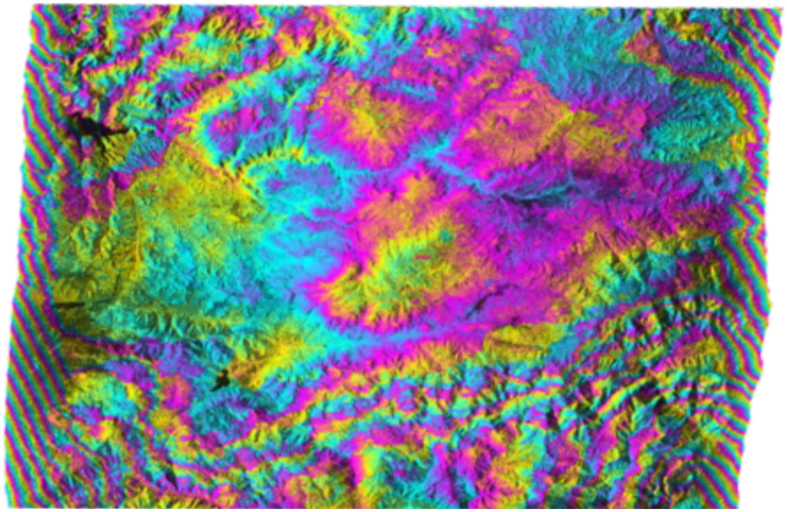
Machine learning-based QA to detect anomalies.

Phase Unwrapping Error Detection

- **Goal: provide scientists a tool to identify/filter out interferograms with phase unwrapping errors**
- Unwrapping errors identified as most common / troublesome interferogram artifact
- Traditionally identified by manual (visual) inspection
- **Approach**: train machine learning classifier to predict p(error|I): probability of unwrapping errors in interferogram I

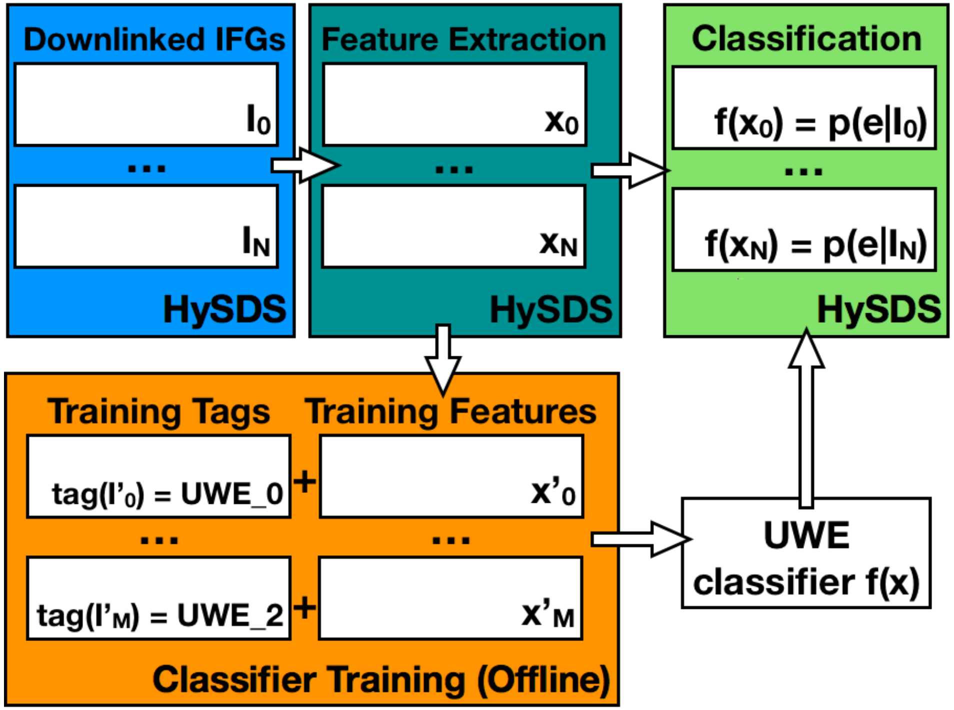


Accurate Unwrapping
p(error|I)→0%



Severe Unwrapping Error
p(error|I)→100%

Scalable Machine Learning Approach



- Classifier **training**
  - Offline with GPUs
  - Leveraging collaborative user tags
- Extract **features** from SAR interferogram data stream from science data system
- **Classification** on data stream to get **predictions**
- Online components are **scalable** in cloud computing environment

Feature Construction / Extraction

Quality Metric	Features
Unwrapped phase smoothness	Locations of rapid local changes in phase gradient, gradient values adjacent to discontinuities
Unwrapping "Barriers" (connected components)	# of connected components percentage of (high-coherence) pixels covered by the largest components
Cyclic Residues	% of (high-coherence) pixels with 2π cycles among their neighbors
Topography Removal Quality	Correlation between unwrapped phase & (SRTM DEM) topography

Mapped unwrapping quality metrics defined by science team to numerical features for training ML classifier

Feature extraction / prediction of P(error|I) automatically triggered upon downlink in HySDS system

Baseline Accuracy / P(error|I)

- ML predictions of p(error|I) capture unwrapping error severity
- Potentially viable tool to speed up science return

Tag	Count	Mean P(error|I)	Accuracy
UWE_0 (pristine)	613	36.95	58%
UWE_1 (minor)	491	60.31	83%
UWE_2 (major)	340	69.46	92%
UWE_3 (severe)	70	70.65	90%

- >90% accuracy identifying major/severe errors (red text)
- Low UWE_0 accuracy (blue text) due to confusion between UWE_0 and UWE_1 tags classes (lower priority filtering)

UWE Detection with Deep Neural Networks

Deep Neural Networks: state of the art for image classification tasks

Input Data	Classifier	Precision	Recall	F1
Derived IFG Features (quality metrics)	Random Forest (**baseline**)	*75.0*	*86.9*	*80.5*
Browse Images (8bit, 3band RGB)	Neural Network (MLP) (2 layers)	71.2	96.0	81.6
	ConvNet (3 layers)	72.3	64.8	67.8
	ConvNet (6 layers)	73.8	80.0	76.3
Masked Interferogram (32bit, 1band)	Deep ConvNet (6 layers)	**75.3**	**91.2**	**82.5**
Masked Interferogram Tiles (250x250) (32bit, 1band)	Deep ConvNet (5+1 layers)	**76.1**	**89.4**	**82.75**

Overall: marginal improvement in prediction accuracy over baseline (blue)

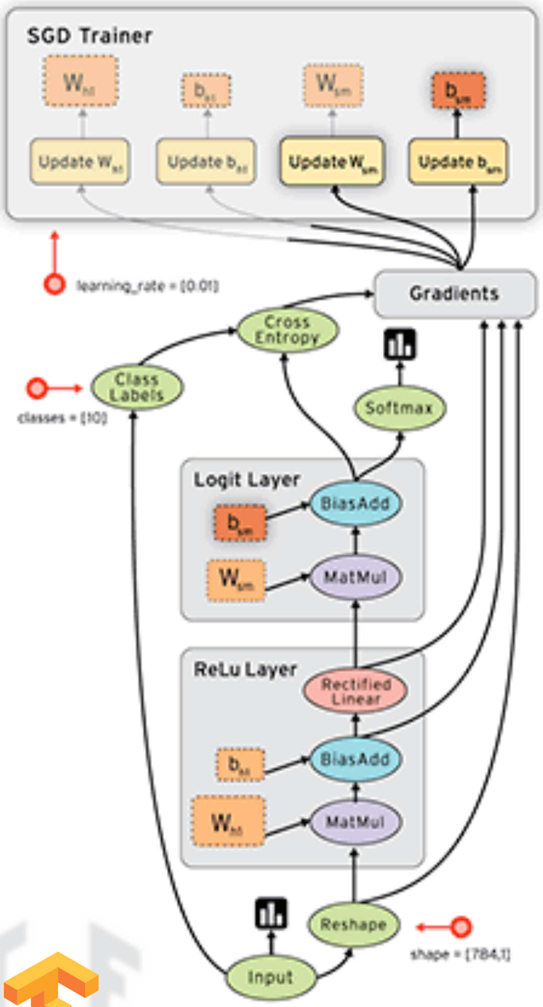
Mixed Approach with Tensor Flow

In November 2015 Google open sourced its software for Machine Learning called Tensor Flow. The software is mostly meant for image recognition and classification using Convolutional Neural Networks (ConvNets), although the same approach has been used for other tasks such as speech recognition, **ontology**, text prediction etc.

We tried to combine the power of image classification of a ConvNet in Tensor Flow with the speed up resulting from pre-computed features.

The features adopted in this approach are:

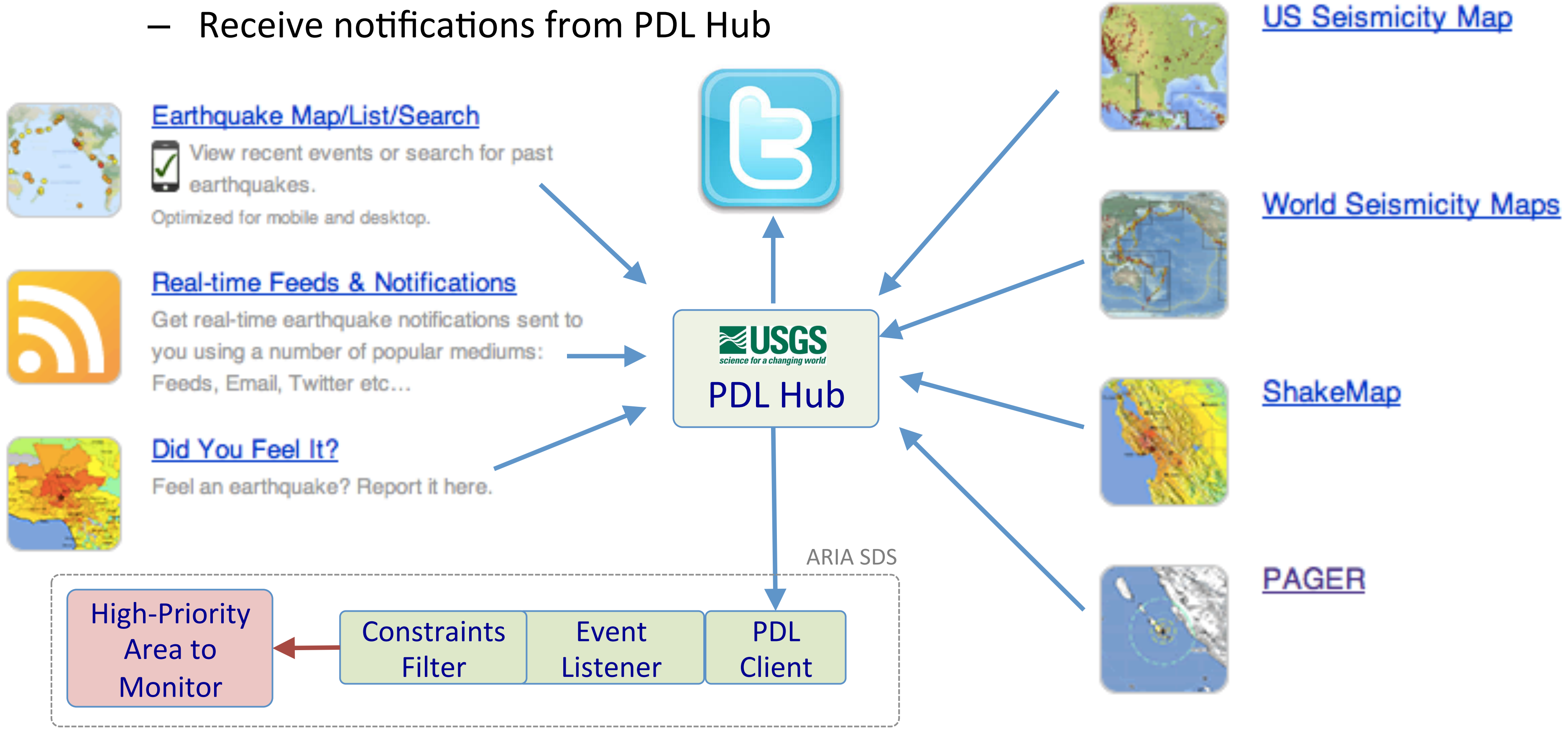
- Average and Standard Deviation of coherence (two values).
- Average and Standard Deviation of the phase gradient (two values).
- Number of connected components (regions that are consistently unwrapped) necessary to cover 50 % and 90 % of the image (two values).



EVENT RESPONSE

Event Notifications

- Leverage USGS National Earthquake Information Center (NEIC)'s Product Distribution Layer (PDL)
  - Pub-sub notification of events and data products
  - Publish events/data to PDL Hub
  - Receive notifications from PDL Hub



Earthquake Map/List/Search

View recent events or search for past earthquakes.

Optimized for mobile and desktop.

Real-time Feeds & Notifications

Get real-time earthquake notifications sent to you using a number of popular mediums: Feeds, Email, Twitter etc…

Did You Feel It?

Feel an earthquake? Report it here.

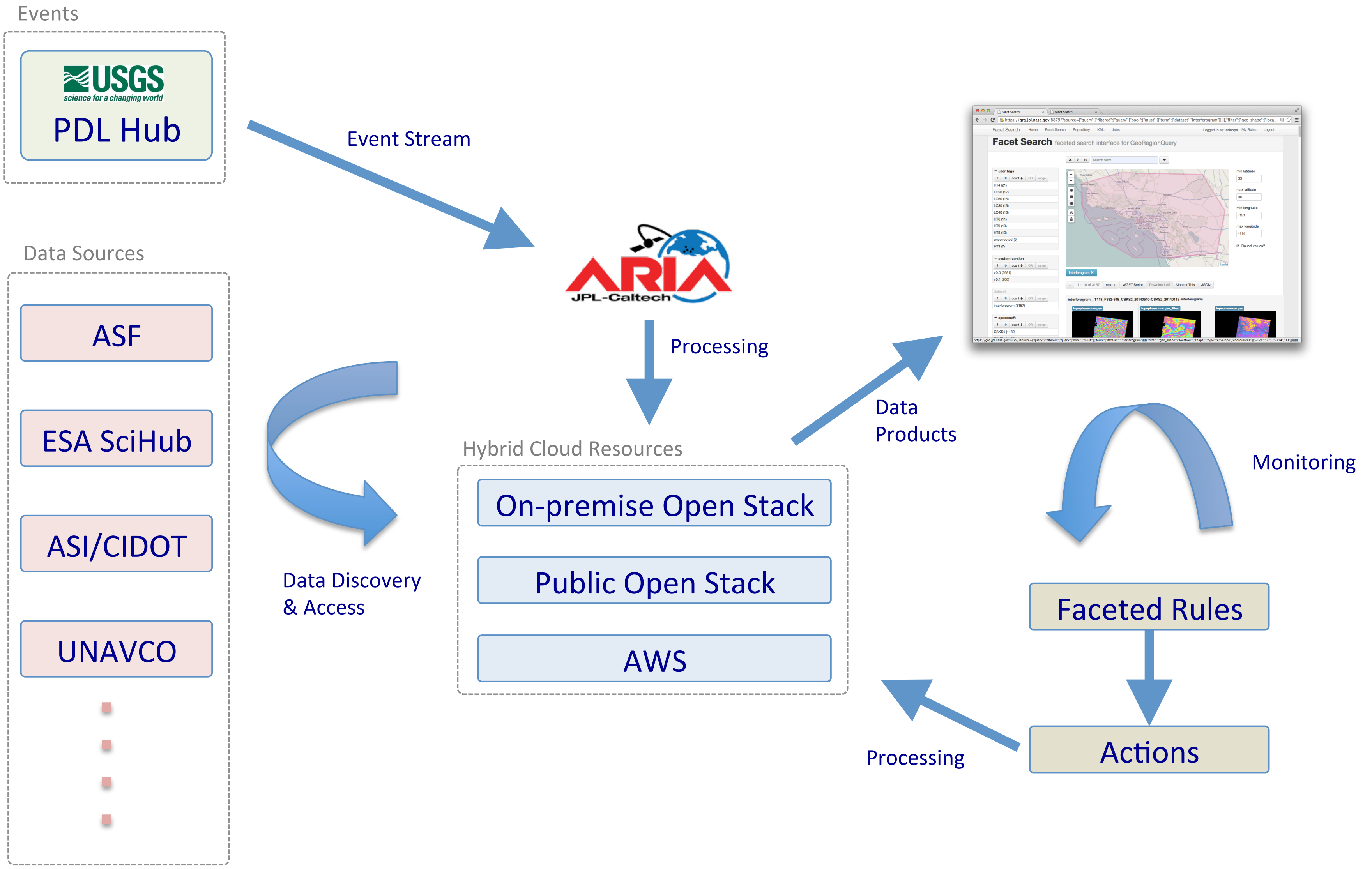
US Seismicity Map

World Seismicity Maps

ShakeMap

PAGER

Monitoring & Event Response



Facet Search | Facet Search

https://grq.jpl.nasa.gov:8879/?source={"query":{"filtered":{"query":{"bool":{"must":[{"term":{"dataset":"interferogram"}}]}},"filter":{"geo_shape":{"loca...

Facet Search Home Facet Search Repository KML Jobs Logged in as: ariaops My Rules Logout

Facet Search faceted search interface for GeoRegionQuery

search term

user tags: HT4 (21), LC50 (17), LC60 (16), LC30 (15), LC40 (13), HT6 (11), HT9 (10), HT5 (10), uncorrected (9), HT3 (7)

system version: v0.0 (2951), v0.1 (206)

dataset: interferogram (3157)

spacecraft: CSKS4 (1180)

min latitude: 33
max latitude: 36
min longitude: -121
max longitude: -114
☑ Round values?

interferogram ✕

.. | 1 – 10 of 3157 | next » | WGET Script | Download All | Monitor This | JSON

interferogram__T118_F332-348_CSKS2_20140510-CSKS2_20140118 (interferogram)

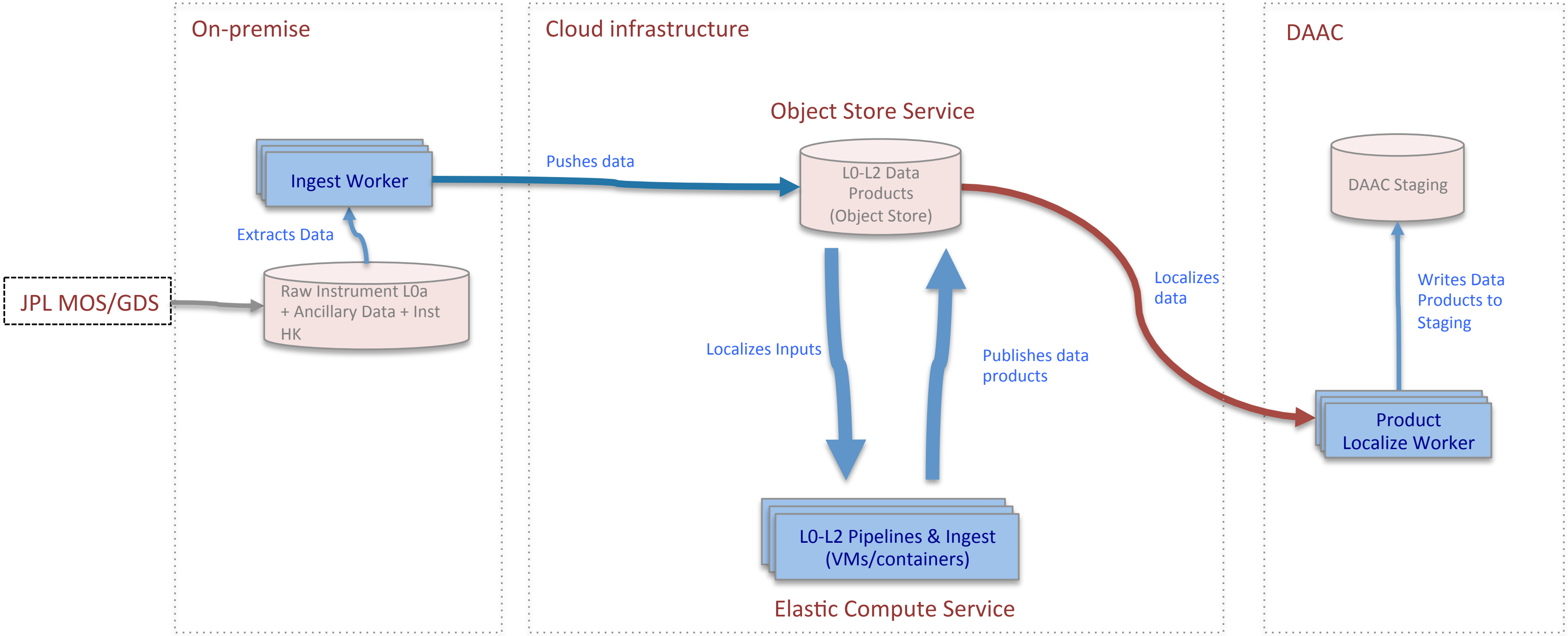
topophase.unw.geo | topophase.unw.geo_20rad | topophase.cor.geo

https://grq.jpl.nasa.gov:8879/?source={"query":{"filtered":{"query":{"bool":{"must":[{"term":{"dataset":"interferogram"}}]}},"filter":{"geo_shape":{"location":{"shape":{"type":"envelope","coordinates":[["-121","36"],["-114","33"]]}}}}...

AGILE BIG DATA HANDLING

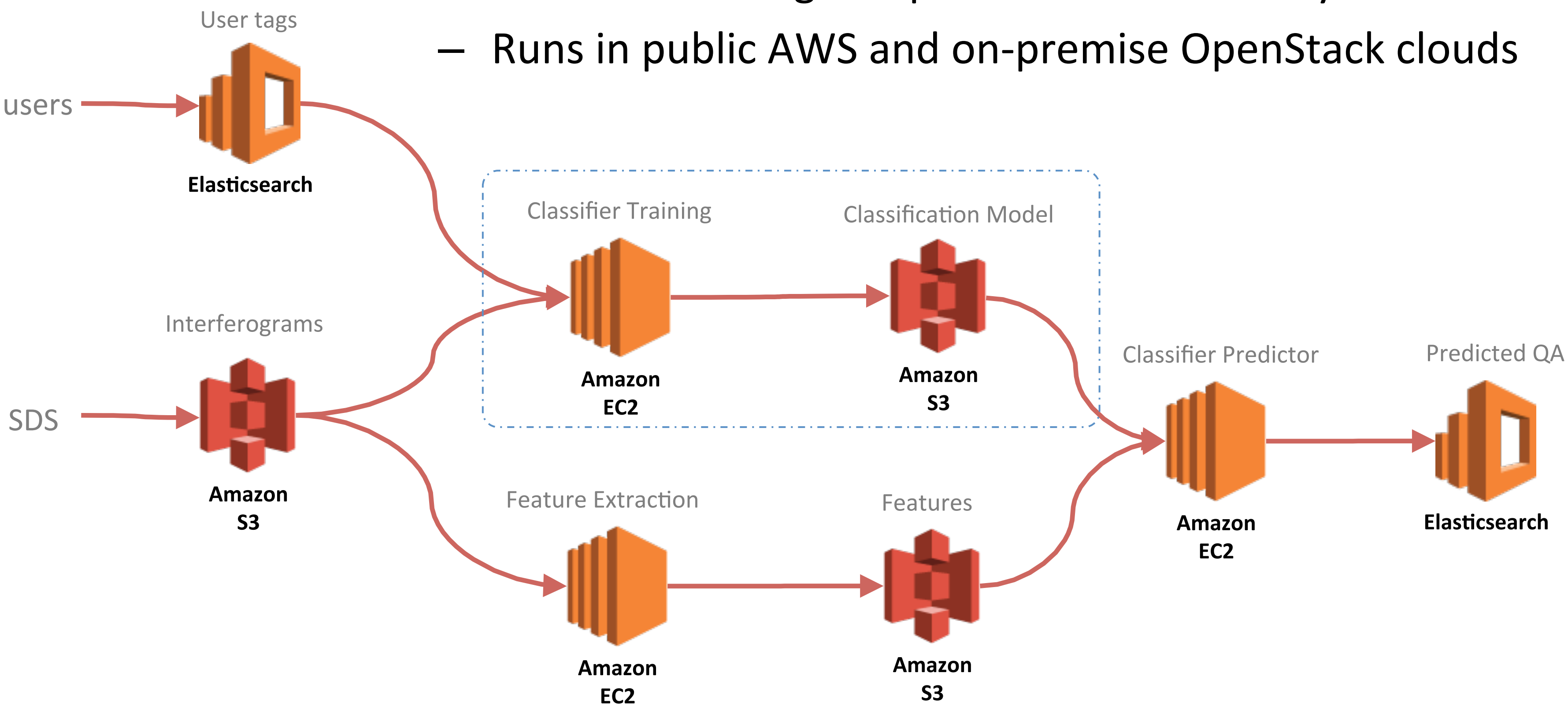
High-Level Cloud Approach for Science Data Processing

- **Stream input data into AWS S3 object storage**
- **Scale up compute nodes to run in AWS EC2**
- **Internal** SDS data throughput needs are scalable via cloud architecture
  - **Compute** instances can scale up to demand
  - Object storage can scale up **data volume** and **aggregate data throughput** to demand
- Asynchronously move results from AWS S3 back to on-premise facility
- Architectural components can be **collocated**



Machine Learning in Cloud-based Science Data Systems

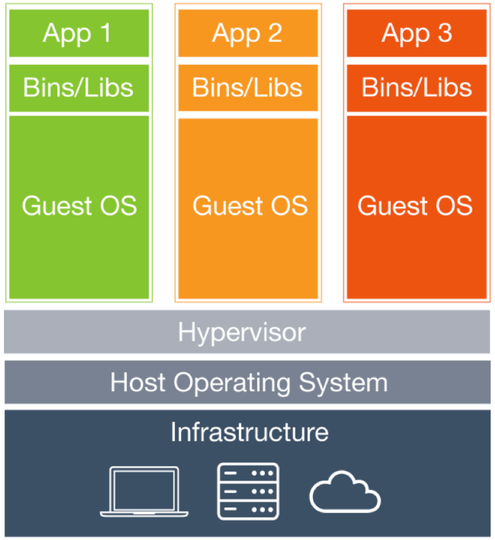
- Bringing machine learning on science data products to cloud-scales
  - Machine Leaning components scalable in HySDS
  - Runs in public AWS and on-premise OpenStack clouds



SDS Services and PGEs in VM and Containers

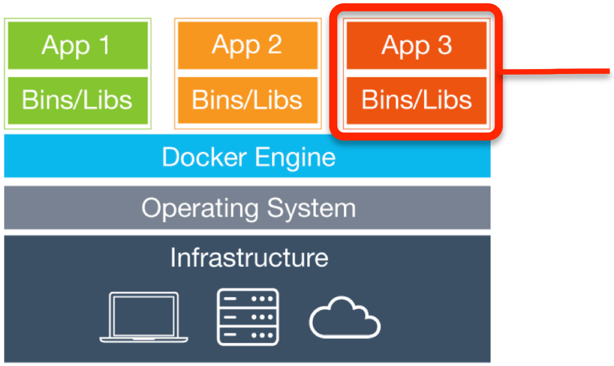
- Run SDS and PGEs in private and public cloud compute instances
- Containers run in compute instances
- Snapshots state of SDS services and PGEs

- SDS and PGEs can be packaged up into orchestrated Containers
- PGEs as Containers
  - L0-L3 PGEs
- Analytics as Containers too



Virtual Machines

Each virtual machines includes the application, the necessary binaries and libraries and an entire guest operating system - all of which may be tens of GBs in size.



Algorithms deployed in Containers

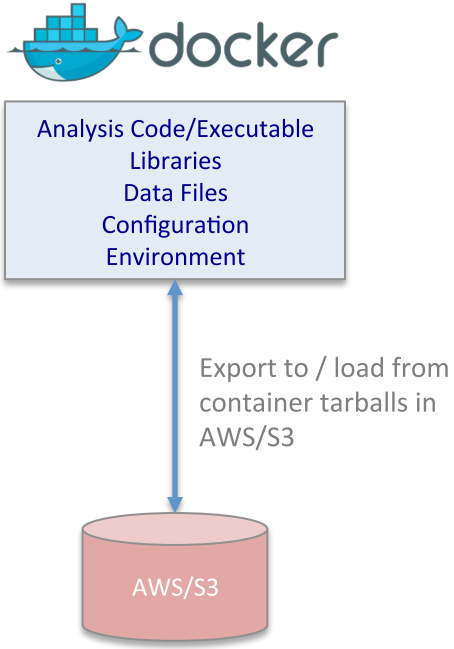
Containers

Containers include the application and all of its dependencies, but share the kernel with other containers. They run as an isolated process in userspace on the host operating system. They're also not tied to any specific infrastructure – Docker containers run on any computer, on any infrastructure and in any cloud.

"Containerizing" Analysis Steps

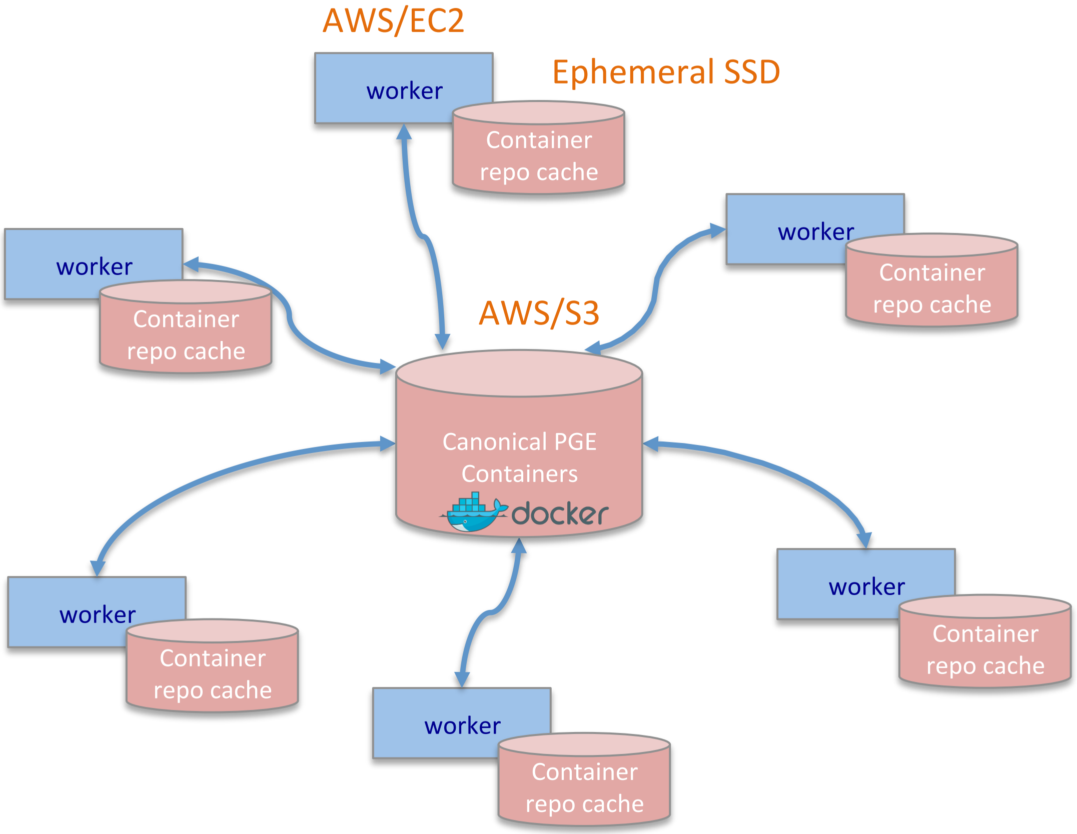
- **Containerizing**
  - Encapsulating analysis steps into more portable and self-contained Docker Containers
- **Agility**
  - Foster agility through rapid development and deployment of analysis steps
- **Portability**
  - Deploy analysis steps in private and public clouds
- **Scalability**
  - Large-scale deployment of Containers to compute fleet
- **Provenance**
  - Archive PGE Containers in AWS/S3
  - Reproduce all existing and prior versions of data analysis and production
  - *"use what you store, and store what you use"*
  - Re-run analysis by data system and DAAC

*"Docker containers wrap up a piece of software in a complete filesystem that contains everything it needs to run: **code, runtime, system tools, system libraries** – anything you can install on a server. This guarantees that it will always run the same, regardless of the environment it is running in."*



Container Orchestration for Big Data

- Concept
  - PGEs developed in Containers
  - PGE Containers managed by SDS
  - PGE Containers deployed at scale to compute workers
- SDS-DAAC scenarios
  - Addresses SDS to DAAC deliverable of PGEs
  - Enables any released PGEs to run *on-demand*
    - Retrieve, instantiate, and run all current and prior PGEs
- Big Data Scales
  - S3 object store archives *all versioned* PGE Containers
  - Use of Docker repo to *cache* PGEs—on each worker
    - Faster load times
  - ***Scales up to thousands of compute workers @ 100K+ cores***



Why is this important?

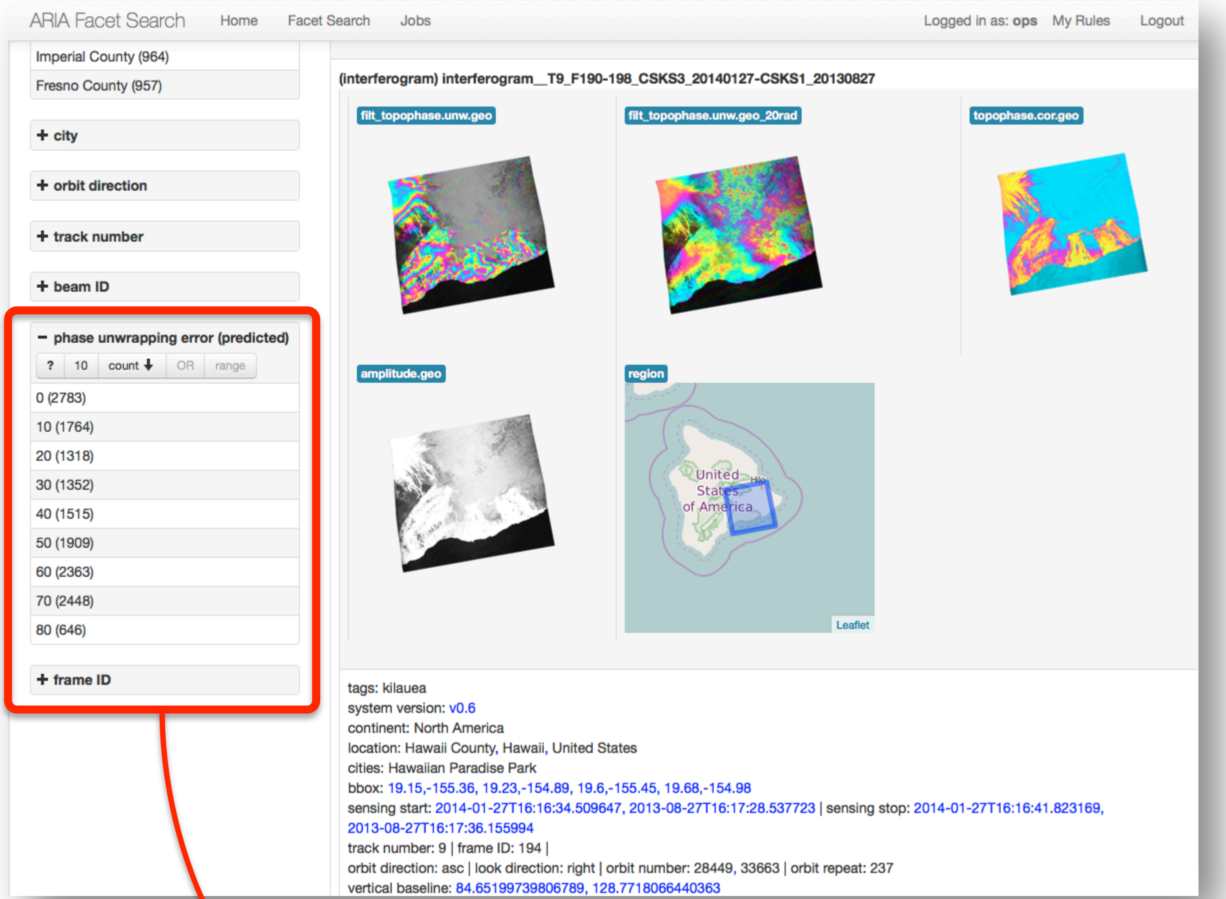
IMPACTS

Integration of ML to Cloud-based SDS

- **Automated QA can now be done for L2 interferogram production—at large-scales**
  - Model training on labeled interferograms
  - Feature extraction of interferograms generated
  - Generate prediction of quality based on pretrained model and features

- Scenarios
  - On-demand
  - Forward stream (keep up) data processing
  - Bulk reprocessing
  - Urgent response
    - Generating param sweep of coseismic ifg

Machine Learning Impact



Imperial County (964)
Fresno County (957)

0 (2783)
10 (1764)
20 (1318)
30 (1352)
40 (1515)
50 (1909)
60 (2363)
70 (2448)
80 (646)

(interferogram) interferogram__T9_F190-198_CSKS3_20140127-CSKS1_20130827

tags: kilauea
system version: v0.6
continent: North America
location: Hawaii County, Hawaii, United States
cities: Hawaiian Paradise Park
bbox: 19.15,-155.36, 19.23,-154.89, 19.6,-155.45, 19.68,-154.98
sensing start: 2014-01-27T16:16:34.509647, 2013-08-27T16:17:28.537723 | sensing stop: 2014-01-27T16:16:41.823169, 2013-08-27T16:17:36.155994
track number: 9 | frame ID: 194 |
orbit direction: asc | look direction: right | orbit number: 28449, 33663 | orbit repeat: 237
vertical baseline: 84.65199739806789, 128.7718066440363

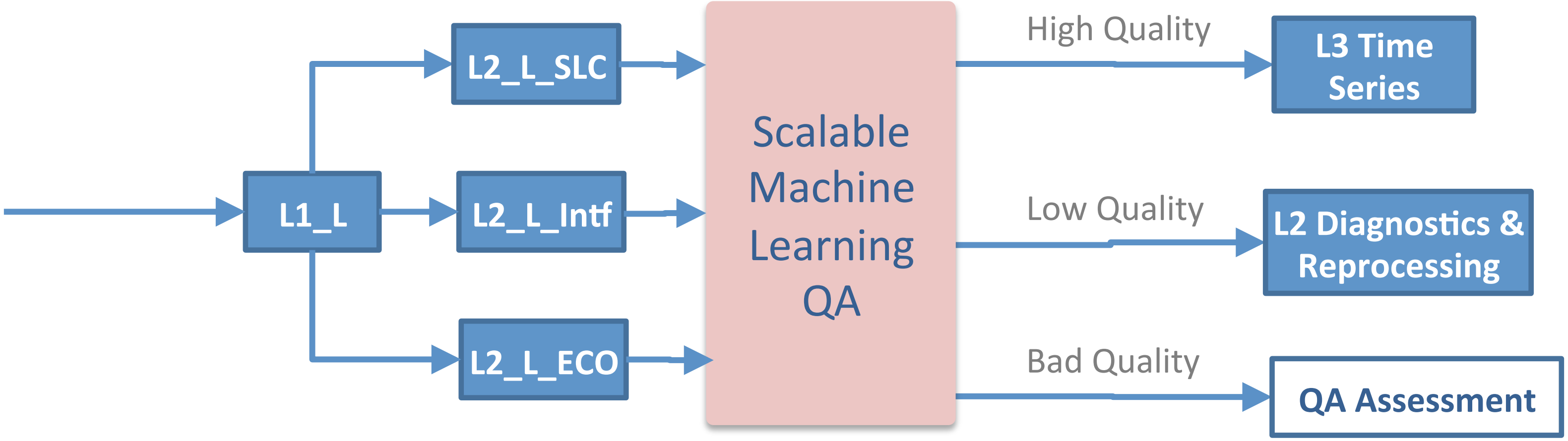
Machine learning-based predictions exposed as facets

- Sensors
  - CSK
  - Sentinel-1A (next)
- Labeling / tagging / social crowdsourcing
  - Thousands of user QA tags as training
- Classifiers
  - Random Forest, Convnet, Tensor Flow
- Feature extraction
- Modeling training
- Prediction: Integrated into faceted search

- *Use Case:* ***Selection of only high-quality interferograms for time series generation***

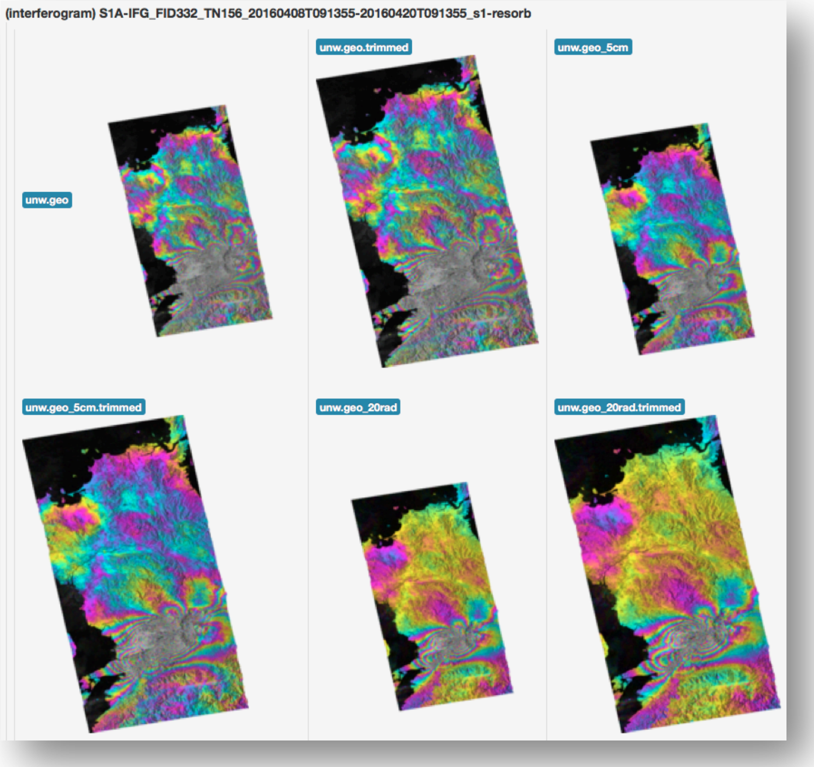
Benefits of Using ML for QA

- **Prioritize** accurate products, scientists can focus on science rather than filtering bad data or finding good/ interesting data
- **Automatically schedule reprocessing** for inaccurate products (e.g., bad unwrappings)
- Automated **diagnostics** of issues in processing pipeline issues

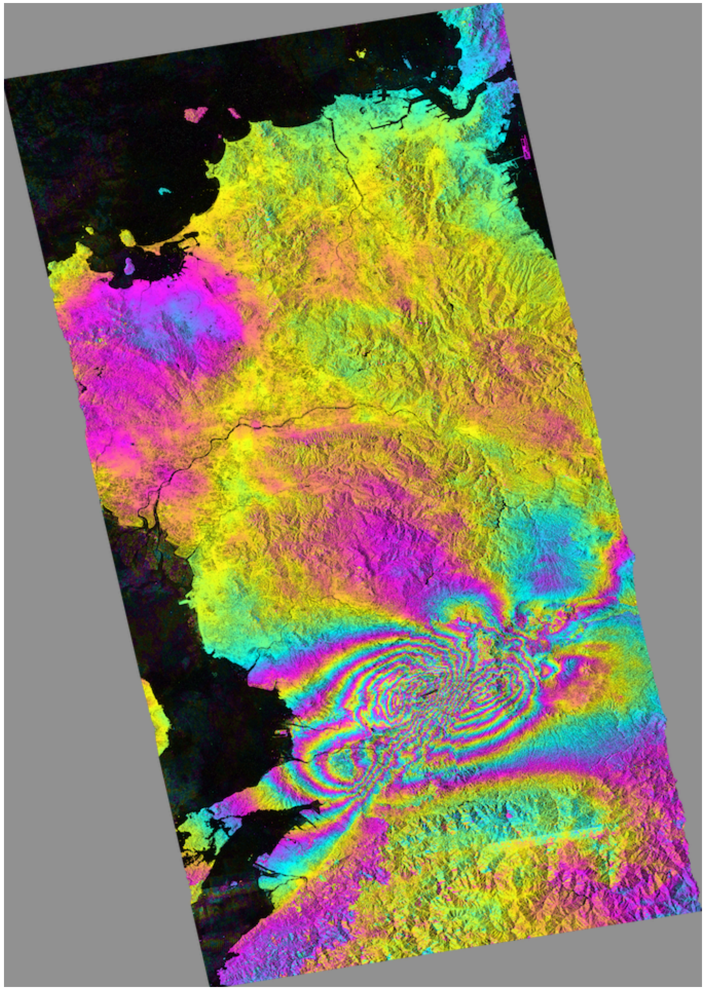


Event Response: Kumamoto Earthquake 2016

- Urgent response of Sentinel-1A L2 interferograms were **automatically processed—all in AWS**
  - 2-hour latency on cheaper m3.xlarge instances
  - ~30-40 minutes on larger c3.8xlarge instances
- Integration with **USGS NEIC PDL** earthquake event streams



(interferogram) S1A-IFG_FID332_TN156_20160408T091355-20160420T091355_s1-resorb



“Risk” Mitigation for NISAR and SWOT

- All cloud computing approach for SDS
- Multiple SDS venues (dev, I&T, ops fwd, ops bulk, on-demand)
- Forward processing in AWS
- Bulk processing in AWS
- On-demand processing in AWS
  - Virtual Data Products: not storing L1 data, regenerate on-demand
- Automated and scalable QA of science data products
- High volume data product generation
  - ~95TB/day for NISAR
- PGEs as Containers
  - Re-run any (current and prior) production pipelines
  - Post-mortem delivery to DAAC
- Cloud-based development
- IT Security

Impacts to NISAR and SWOT

- HySDS (funded under AIST 2011 and 2014) is the **leading approach selected** for NISAR and SWOT
  - – HySDS is the generic “core” SDS software
- HySDS adaptations
  - – ARIA and SAR SDP Foundry are adaptations of HySDS
  - – NISAR and SWOT SDSes will also be adaptions

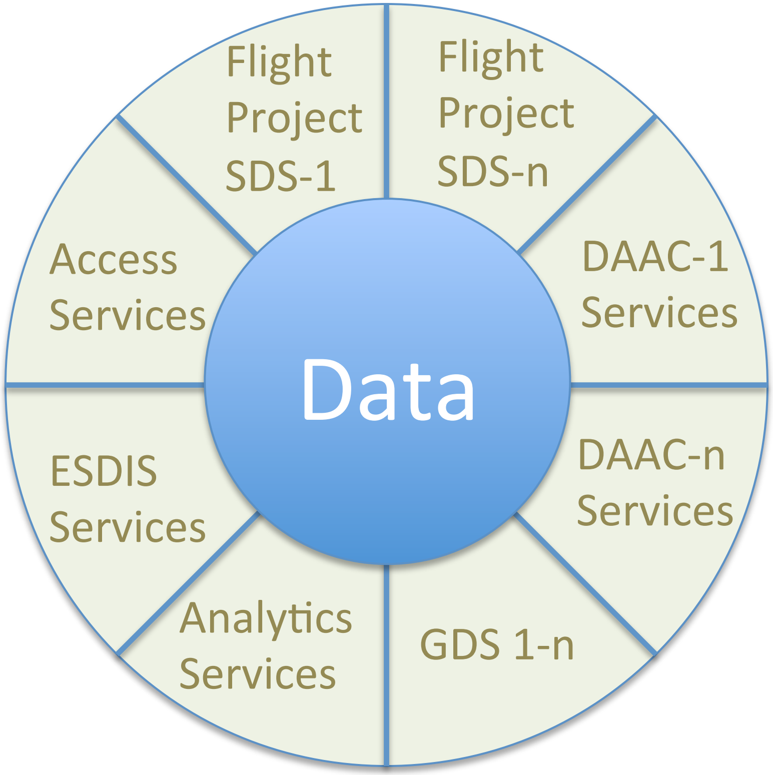
Next Steps

- **Machine Learning**
  - Expand to Sentinel-1A and Sentinel-1B
  - Additional QA metrics
- Additional **Analytics**
  - Coherence time series
  - Troposhere corrections
  - Fast browse of results
- SDS for Big Data mission needs
  - Dynamic **hot data caching** for improved scaling and cloud economics
  - High-availability (HA) support
  - Expand usage on AWS *spot market*

“Data Lake”-extended

- It’s about collocation!
- Minimize data movement
- Maximize user services
- Run on *public cloud provider* or at an *on-premise data center*

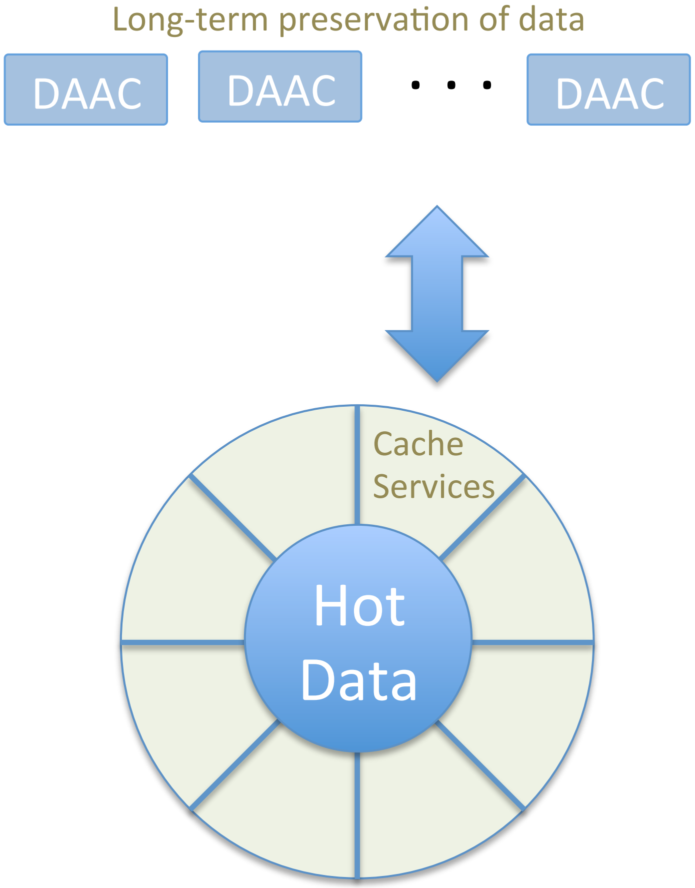
Reduce redundancy and foster ESDIS-wide services



Enabling multi-disciplinary data approach for analysis

"Hot Data Lake"

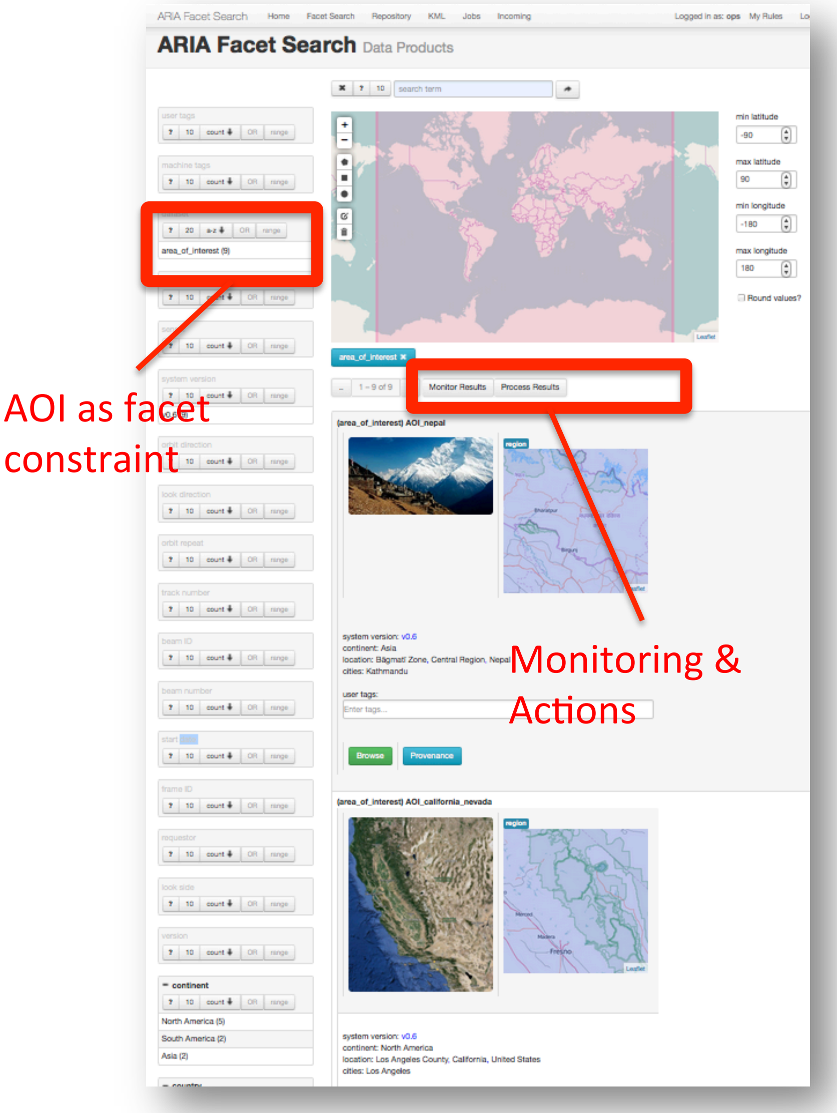
- Long-term **public cloud storage is expensive** at PB-scales
  - …Unless we can negotiate deals with cloud vendors
- Use object store data lake for ***"hot data"***
  - SDSes generate data into object stores
  - Object stores contain "fresh" / "hot" data as rolling storage
  - Offline moving of older data to DAAC for permanent storage
  - Automate caching of "hot data" back from DAACs



BACKUP

Automated Monitoring and Processing of AOIs

- AOI-based automated data discovery, access , and ingestion
  - ASI/CIDOT (CSK)
  - ESA SciHub (Sentinel)
  - ASF (ALOS, Sentinel, etc.)
  - UNACVO, GEO Supersites (CSK, ALOS, Sentinel, etc)
- AOIs in Faceted Search
  - Enables users to interact with AOIs
  - Supports reverse lookups of AOIs
    - Finds all data within an active AOI
    - Finds all AOIs within a certain region
- **Monitoring** data streams of AOIs
- **Actions** to automatically trigger processing to L2 interferograms



(area_of_interest) AOI_nepal

system version: v0.6
continent: Asia
location: Bāgmatī Zone, Central Region, Nepal
cities: Kathmandu

user tags:

(area_of_interest) AOI_california_nevada

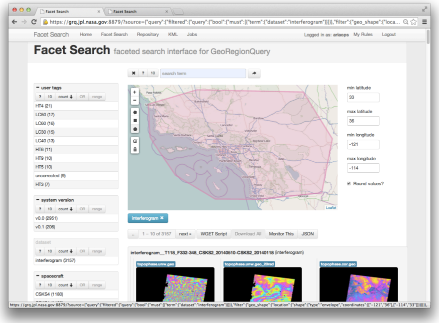
system version: v0.6
continent: North America
location: Los Angeles County, California, United States
cities: Los Angeles

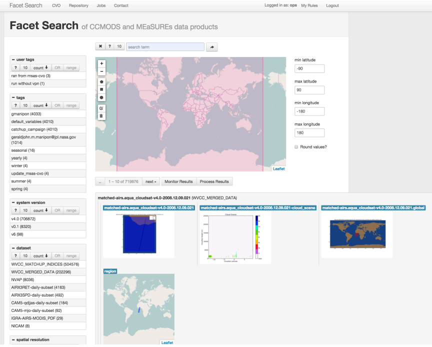
AOI as facet constraint

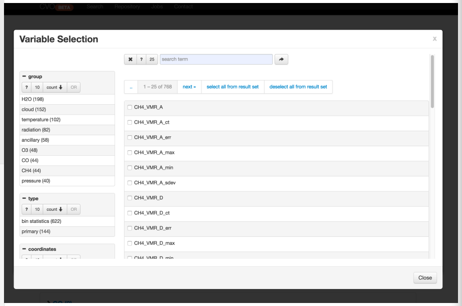
Monitoring & Actions

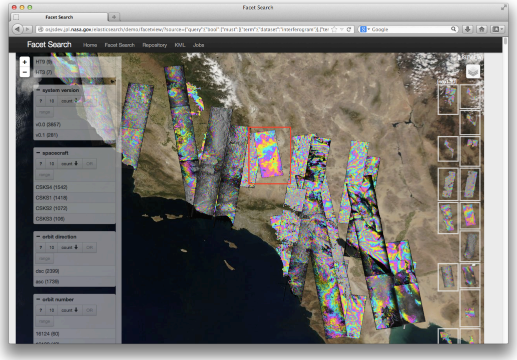
Analytics for Situational Awareness

- Science Data Products Faceted search and browse

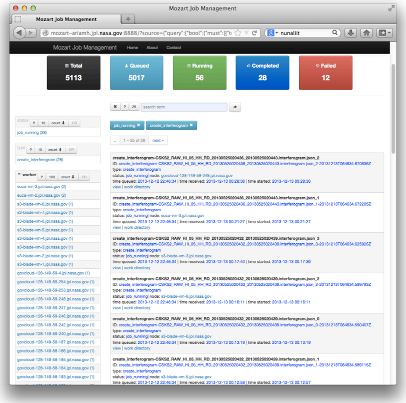

Facet Search faceted search interface for GeoRegionQuery


Facet Search of CCMODS and MEaSUREs data products

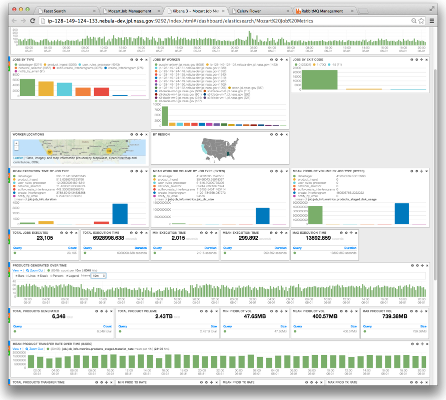

Variable Selection

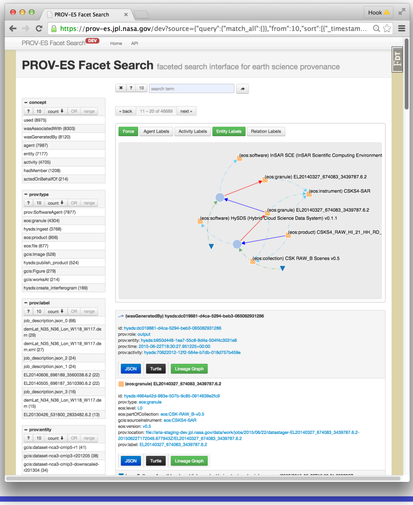

Facet Search

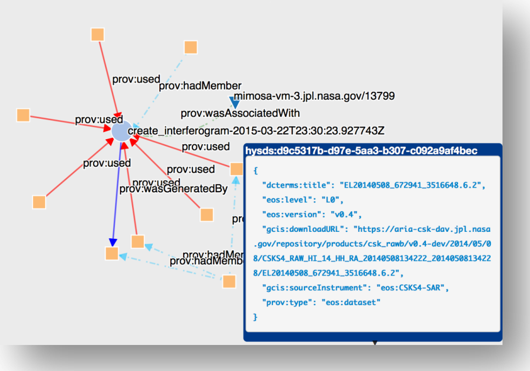
- Science Data System Faceted Metrics and PROV-ES provenance


Mozart Job Management

Total	Queued	Running	Completed	Failed
5113	5017	56	28	12




PROV-ES Facet Search faceted search interface for earth science provenance



hysds:d9c5317b-d97e-5aa3-b307-c092a9af4bec

```
{
  "dcterms:title": "EL20140508_672941_3516648.6.2",
  "eos:level": "L0",
  "eos:version": "v0.4",
  "gcis:downloadURL": "https://aria-csk-dav.jpl.nasa
.gov/repository/products/csk_rawb/v0.4-dev/2014/05/0
8/CSKS4_RAW_HI_14_HH_RA_20140508134222_2014050813422
8/EL20140508_672941_3516648.6.2",
  "gcis:sourceInstrument": "eos:CSKS4-SAR",
  "prov:type": "eos:dataset"
}
```