

Jet Propulsion Laboratory  
California Institute of Technology

# The Advanced Rapid Imaging and Analysis for Monitoring Hazards (ARIA-MH) Science Data System

Earth Science Technology Forum (ESTF2014)

Thursday, October 30, 2014

Hook Hua<sup>1</sup>, Susan Owen<sup>1</sup>, Sang-Ho Yun<sup>1</sup>, Paul Lundgren<sup>1</sup>, Angelyn Moore<sup>1</sup>, Piyush Agram<sup>1</sup>, Gian Franco Sacco<sup>1</sup>, Eric Fielding<sup>1</sup>, Paul Rosen<sup>1</sup>, Frank Webb<sup>1</sup>, Mark Simons<sup>2</sup>, Alexander Smith<sup>1</sup>, Brian Wilson<sup>1</sup>, Timothy Stough<sup>1</sup>, Peter F. Cervelli<sup>4</sup>, Michael Poland<sup>3</sup>

<sup>1</sup> Jet Propulsion Laboratory

<sup>2</sup> California Institute of Technology

<sup>3</sup> USGS Hawaiian Volcano Observatory

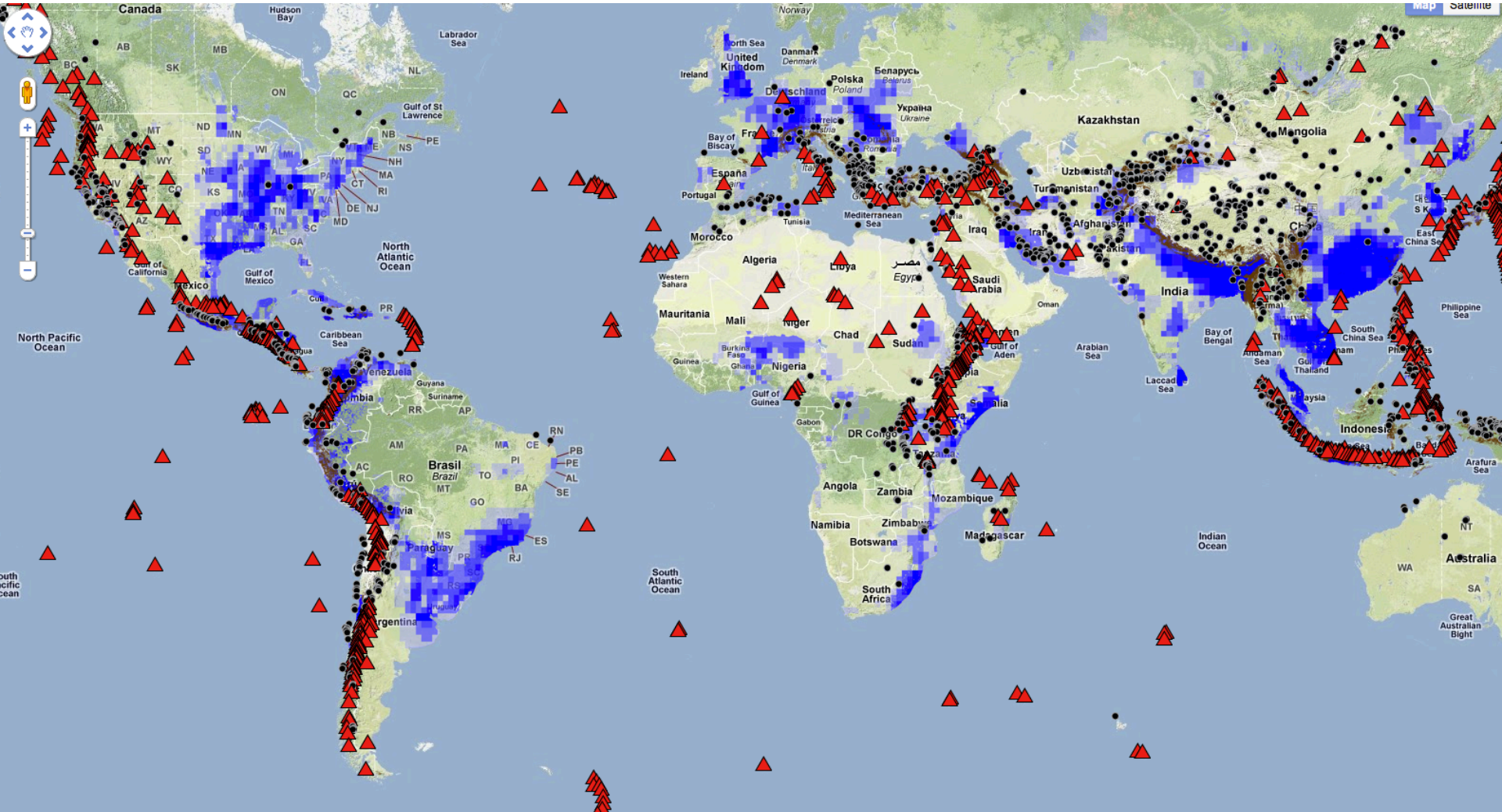
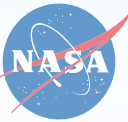
<sup>4</sup> USGS Menlo Park Science Center

Copyright 2014 California Institute of Technology.

Government sponsorship acknowledged.

**ESTO**  
Earth Science Technology Office

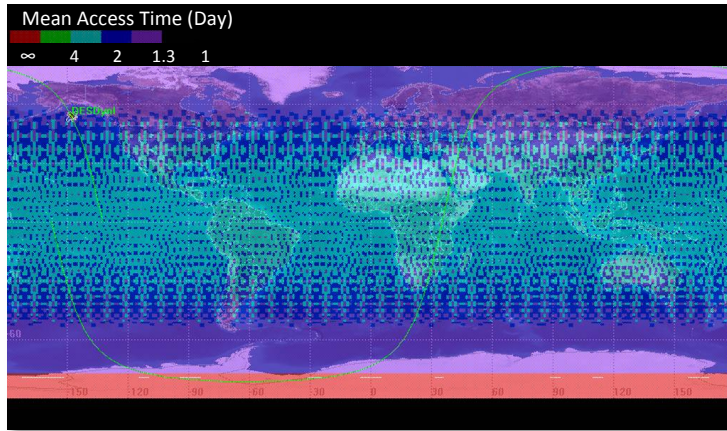
# Natural Hazards



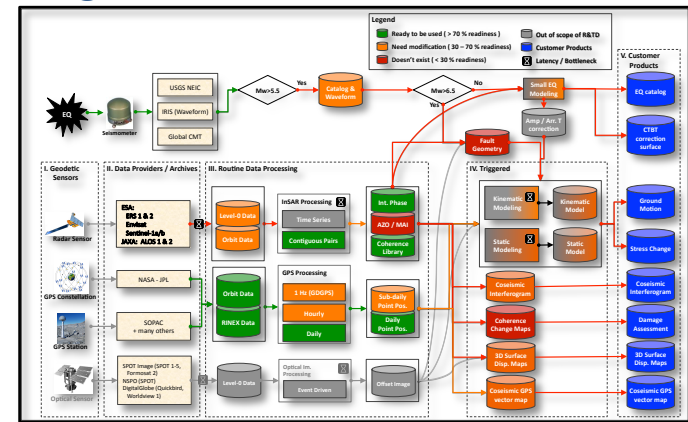
- ▲ known or inferred Holocene volcanoes (Smithsonian Global Volcanism Program)
- shallow earthquake epicenters (>50 km depth) on land with a magnitude of 6.5 or higher since 1976
- extreme flood events (Dartmouth Flood Observatory , from 1985-2003)
- subject to landslides (Norwegian Geotechnical Institute and UNEP-Grid Geneva).



# The Challenge of Leveraging Remote Sensing for Disaster Response

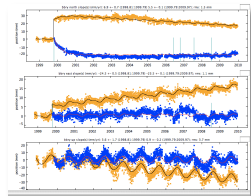


Automated data system are required to analyze large quantities of data from NASA NISAR (formerly DESDynI), other satellite missions, and rapidly expanding GPS networks.

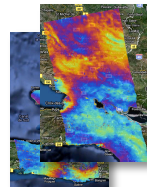


Going from Artisan to Automation: Use system engineering approach to translate specialized data analysis into operational capability.

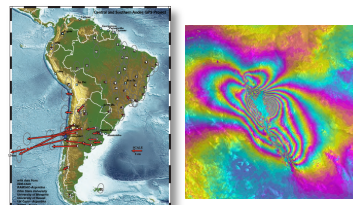
Demonstrate response to hazards with standardized set of data products for decision & policy makers.



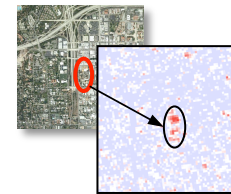
Temporal Records of Ground Deformation



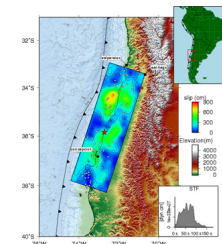
Spatial Maps of Ground Deformation



Coseismic Ground Deformation

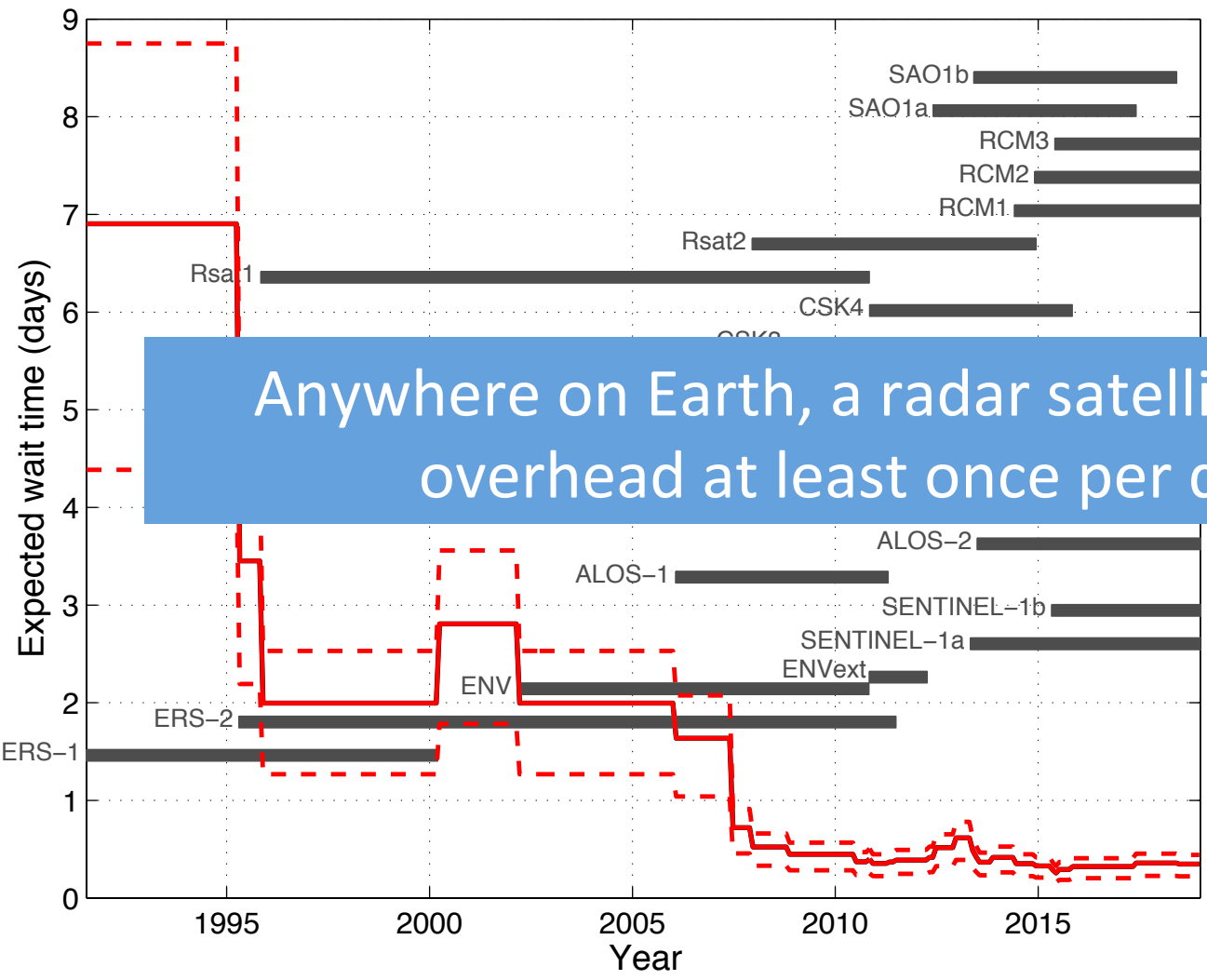


Coseismic Damage



Earthquake Models

# Data Acquisition Latency of InSAR Missions



**Expected wait time until the first SAR satellite to visit after an event**

Ascending + descending orbit

Anywhere on Earth, a radar satellite passes overhead at least once per day.

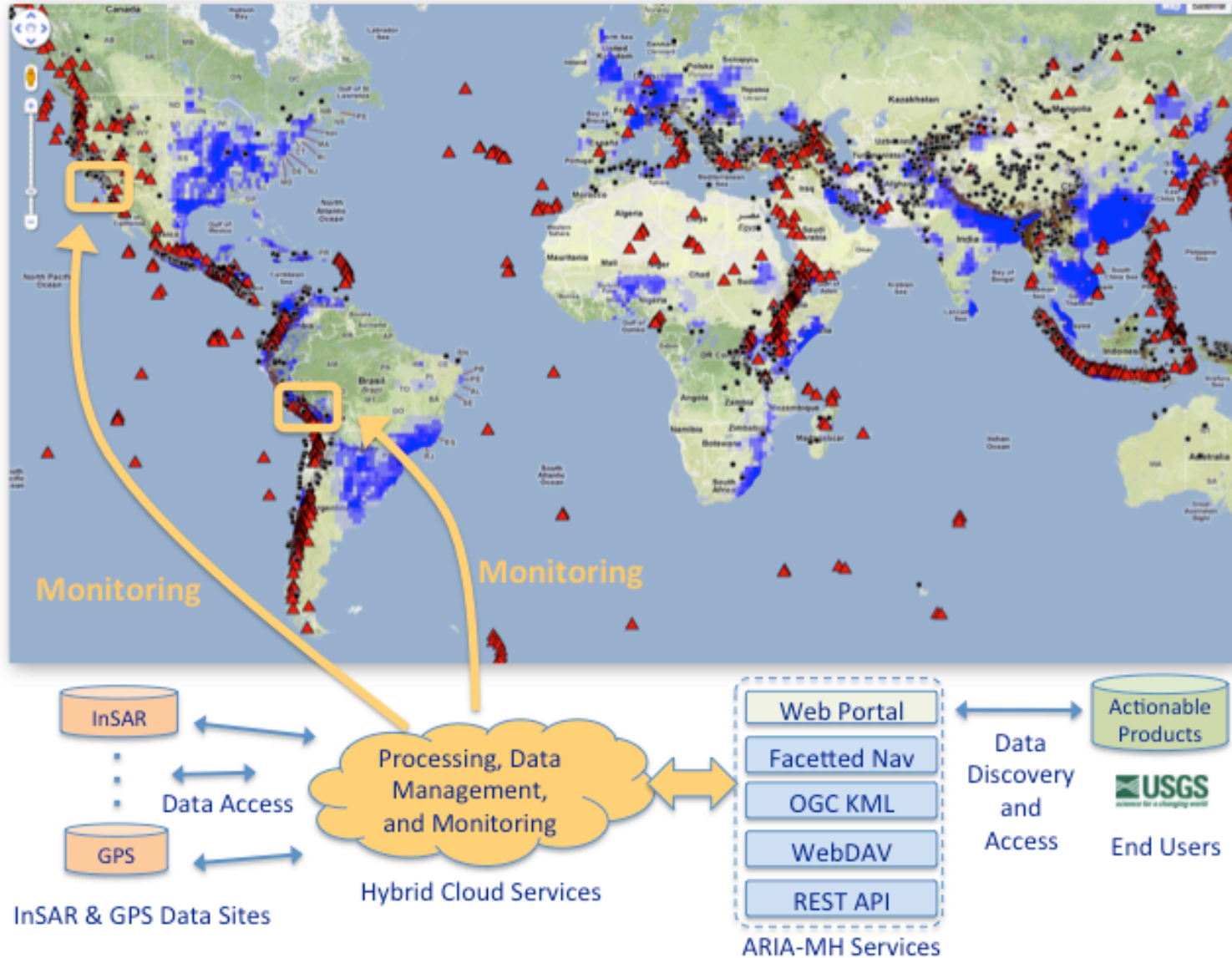
Present: 15 hours

2020: 8 hours

Source: Sang-Ho Yun (JPL)



# Elastic Processing, Data Management, and Monitoring



- **High-volume, low-latency**, and automatic generation of NASA Solid Earth science data products (InSAR and GPS) to support hazards monitoring.
- Enabling both **science and decision-support** communities to monitor ground motion in areas of interest with InSAR and GPS data.
- Leverage and geographically optimize **hybrid Cloud**-based processing and data management of geodetic data products
- **Monitoring and Subscriptions**
  - Event streams from USGS NEIC
  - Data product streams
- **Conditional actions** for triggering of geodetic data processing
- **Situational awareness** for
  - Near real-time information
  - Data system health



# Near Real-Time Big Data Streams



- JPL/Caltech/ASI collaboration effort opens flood of COSMO-SkyMed (CSK) data for select regions
  - Provides access to CSK X-band Level-0 SAR radar data from the Italian Space Agency (ASI)
  - *COSMO-SkyMed (CSK) data provided as part of a **technical collaboration** between JPL-Caltech and the Center for Earth Observations (CIDOT), Italian Space Agency (ASI)*
  - 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
  - Example: San Andreas Fault region of California
    - 580 GB of raw data every 16-day cycle, about 500 frames.
    - Downstream derived data products **increase data volume 50X+**
- NISAR
  - L-band SAR
  - Deliver ~85TB data products per day to DAACs (~**1GB/sec sustained**)
- Sentinel 1A/1B
  - C-band SAR
  - 1.2TB/day raw data
- Data movement and storage concerns for handling global-scale coverage

# Near Real-Time Data Source



## CSK Scene Footprints: Rolling 1-day view

Satellite observation of granules from Italian X-band Level-0 SAR radar data



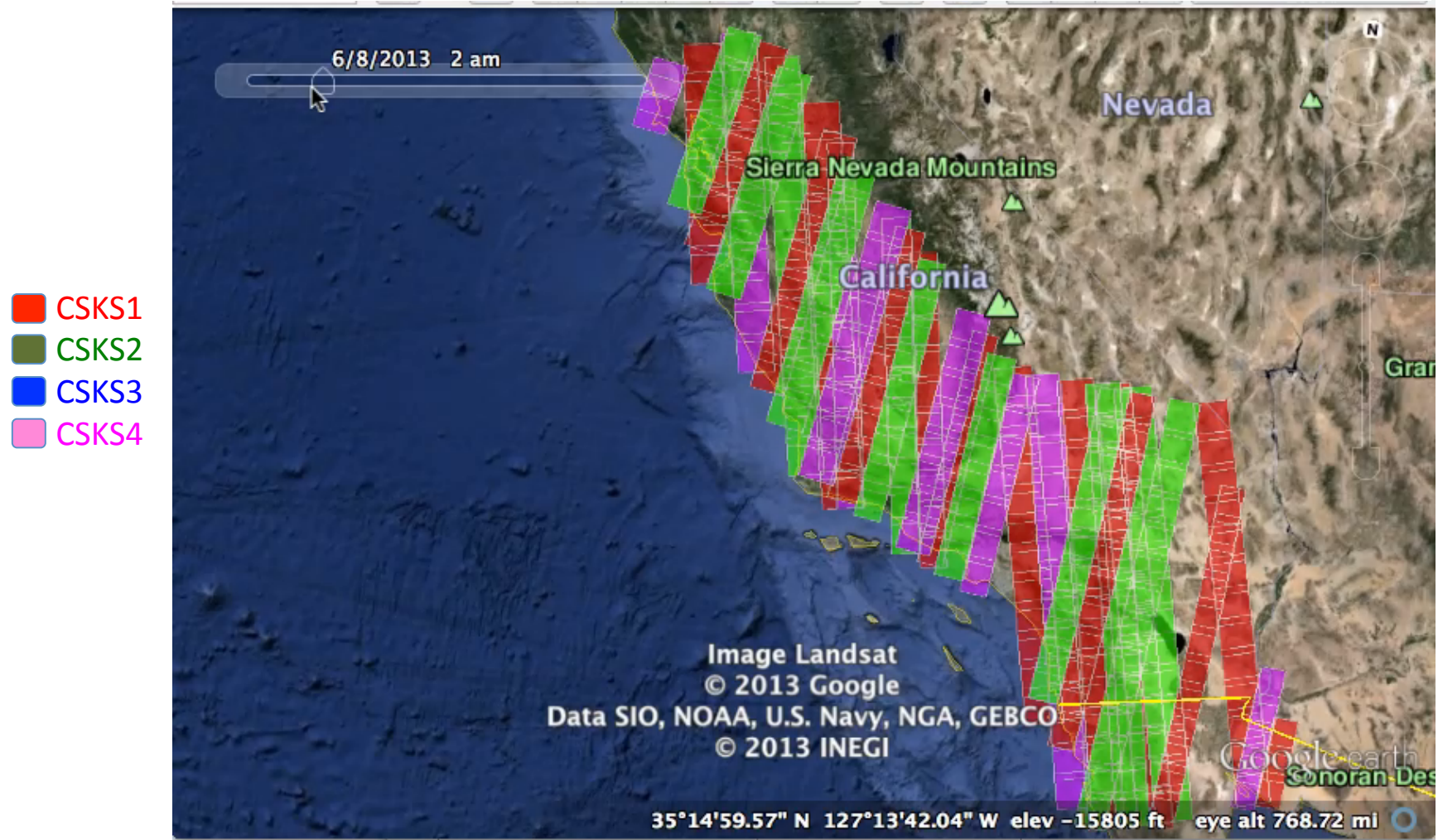


# Near Real-Time Data Source

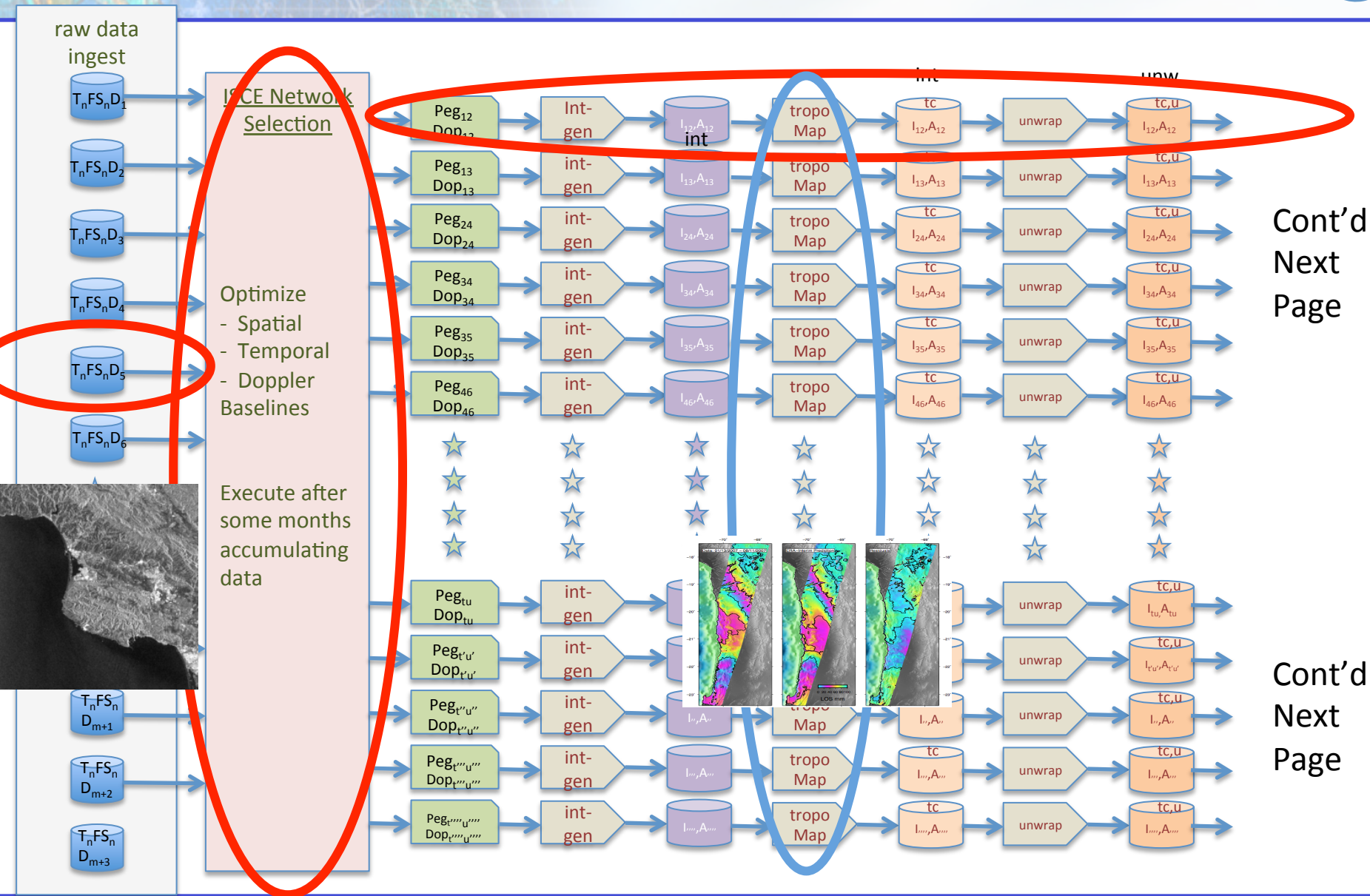


## CSK Scene Footprints: Rolling 16-day cycle view

Satellite observation of granules from Italian X-band Level-0 SAR radar data



# Interferogram Processing Workflow (1 of 2)



Cont'd  
Next  
Page

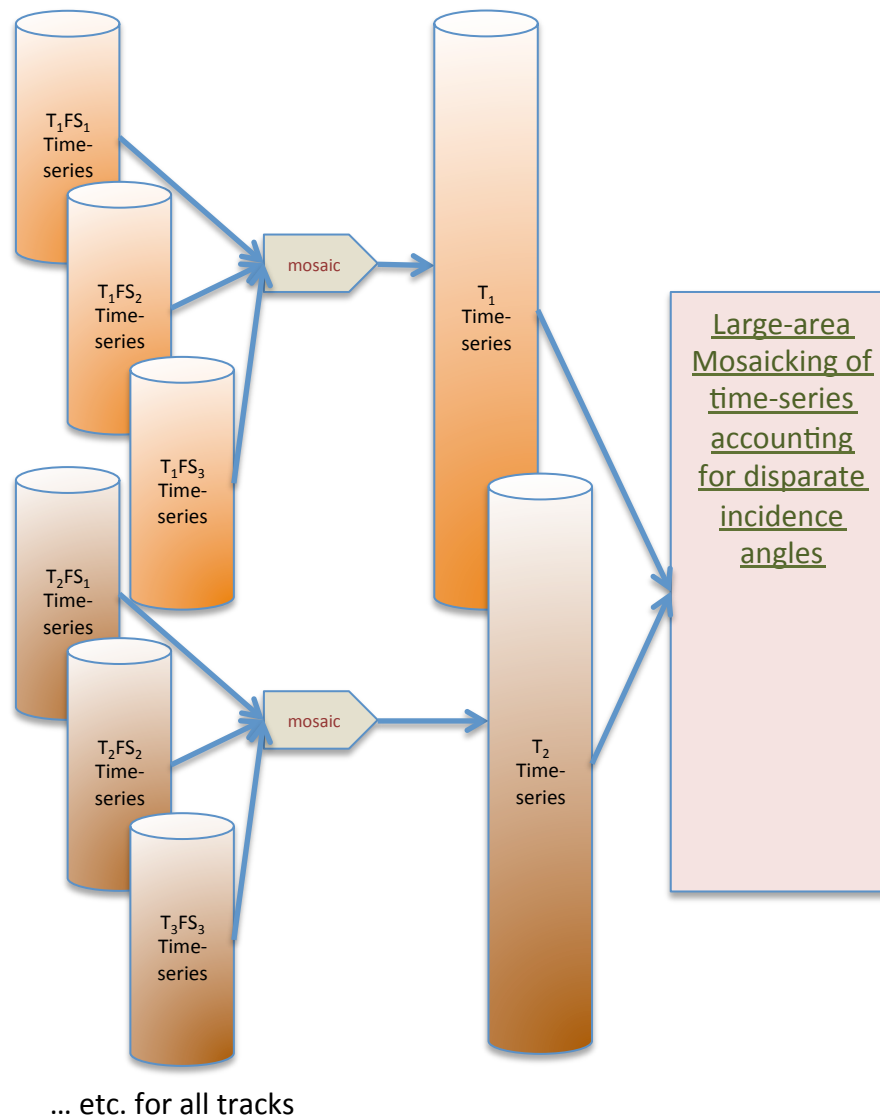
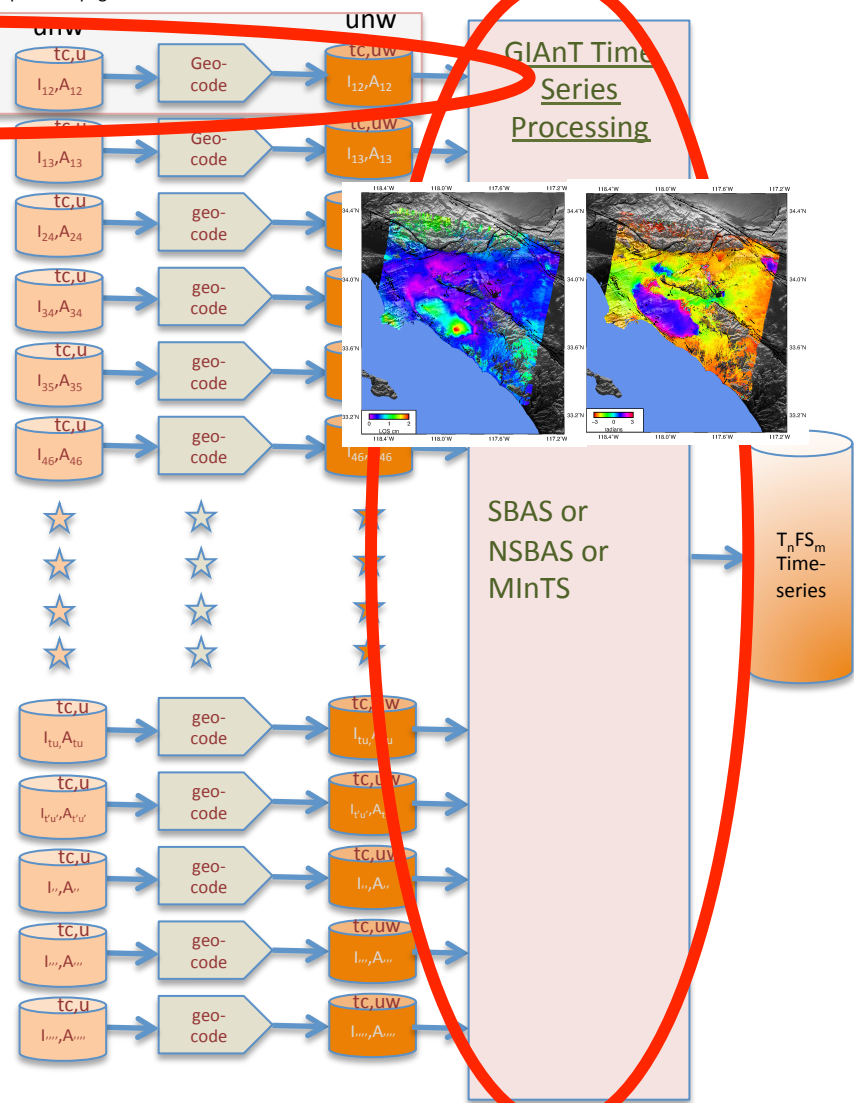
Cont'd  
Next  
Page



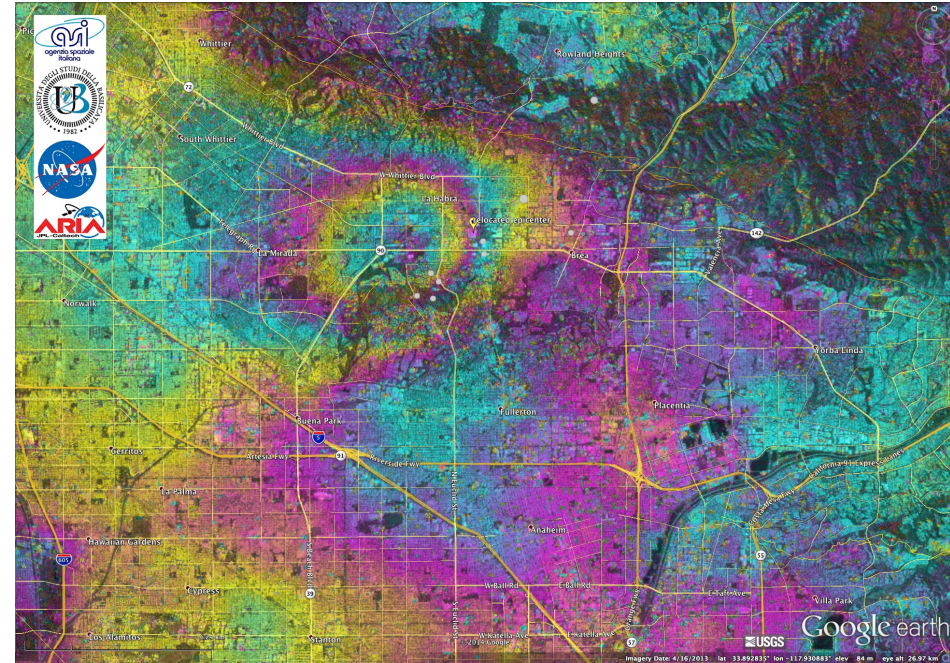
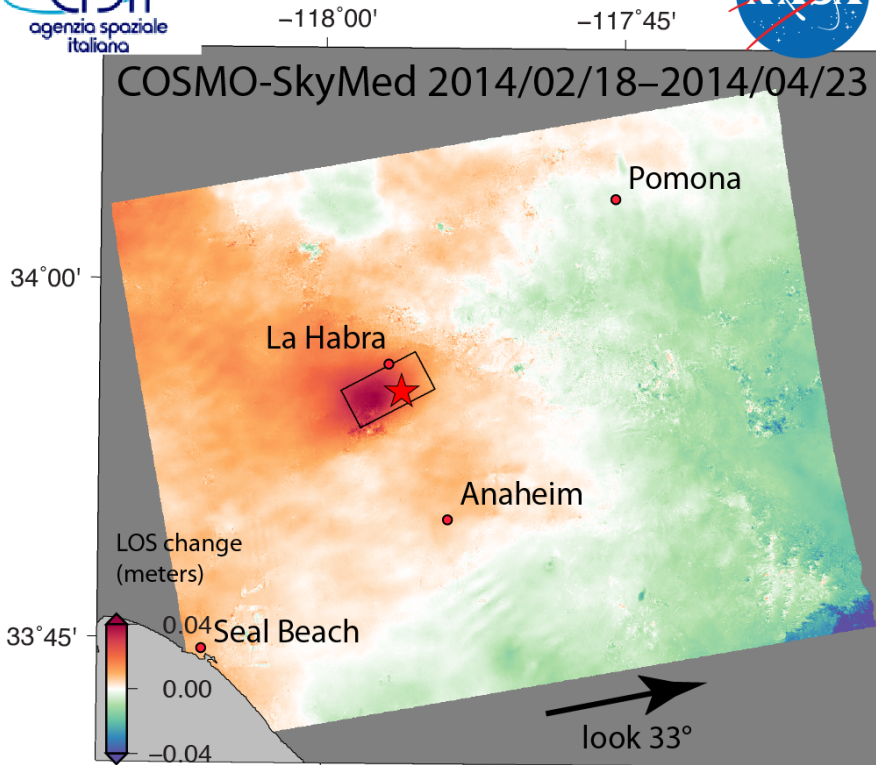
# Interferogram Processing Workflow (2 of 2)



Repeated from previous page



# M5.1 La Habra, CA Earthquake



2014 May 08 14:05:48 EJF geo\_fit\_topophase.los-N1

- ARIA-MH used for continuously monitoring for availability of new CSK scenes over La Habra with beam 5 and track 111.
- Automated interferogram processing
- Based on CSK 1.3cm X-band data

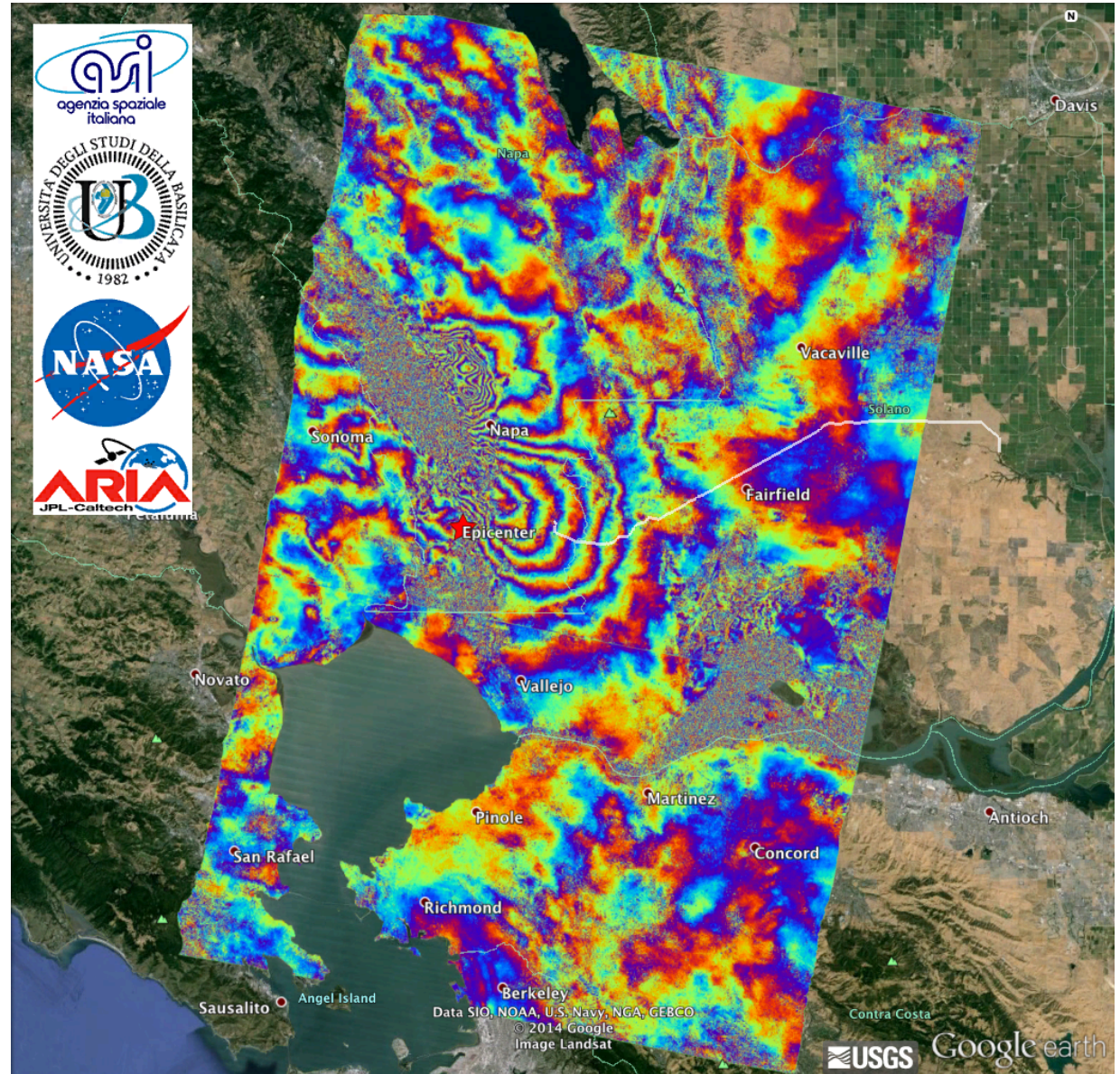




# M6.0 South Napa Earthquake Deformation Field



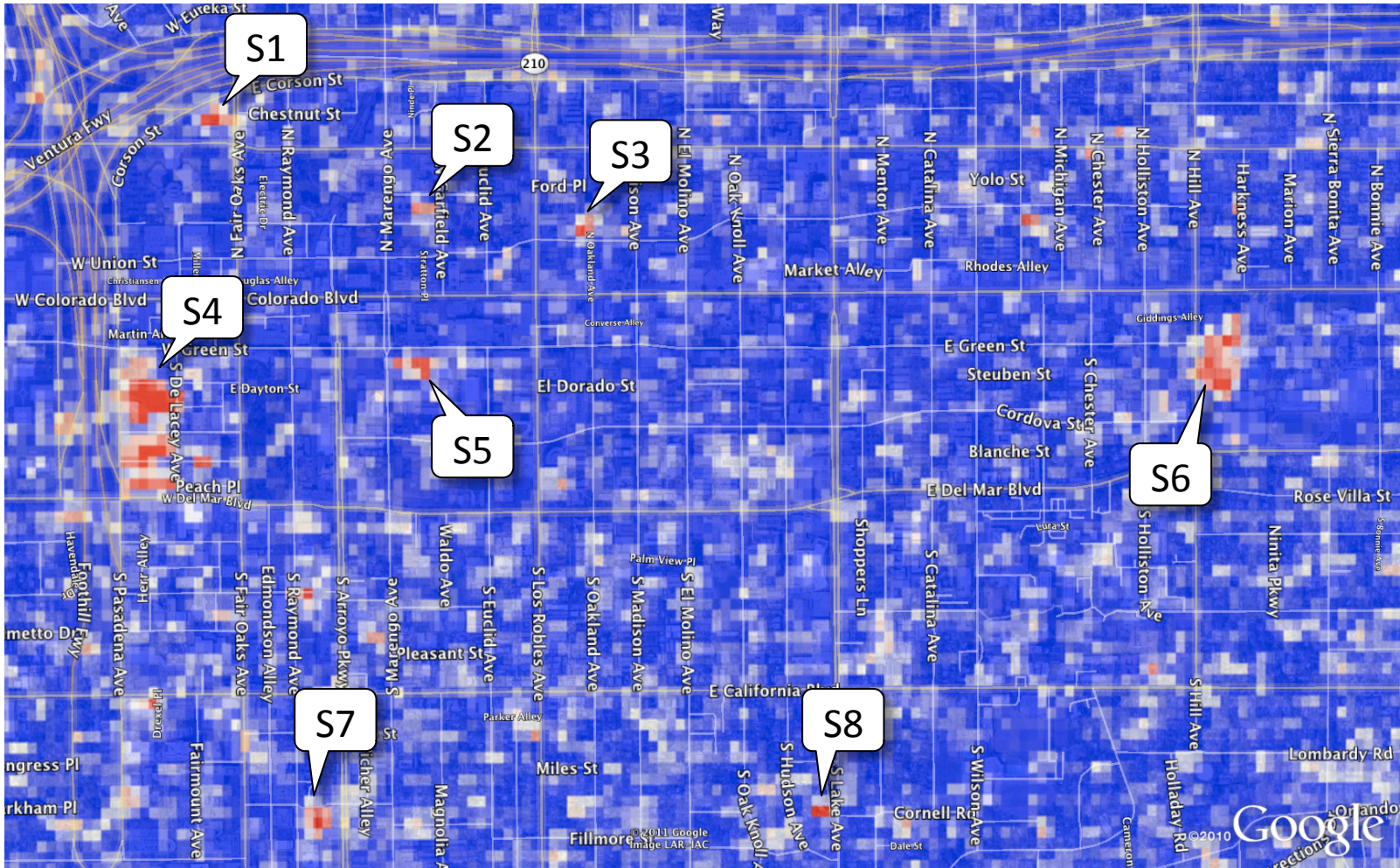
- Italian Space Agency (ASI) activated crisis response for this event
- First overflight of the area by COSMO-SkyMed X-band radar on 8/26/14
- Data delivered to ARIA data system on 8/26/14 at 11 pm
- ARIA generated map of deformation field (“interferogram”) on 8/27/14 by 4 am
- Each fringe represents 1.5 cm of motion
- Delivered to California earthquake Clearinghouse at 6 AM
- In use by Calif. Geological Survey field teams





# Damage Proxy Map

- Coherence change time series
  - Example: Pasadena 2006.12.31 - 2007.02.15 - 2008.02.18



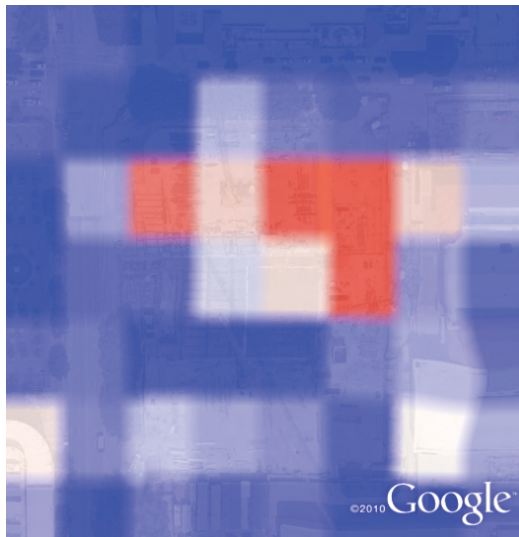
Source: Sang-Ho Yun (JPL)



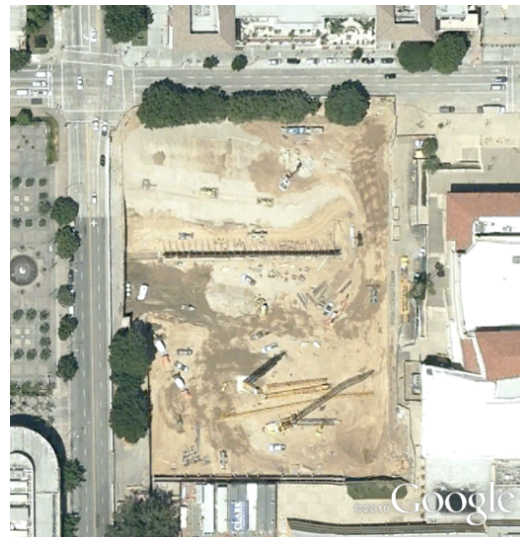
# Damage Proxy Map



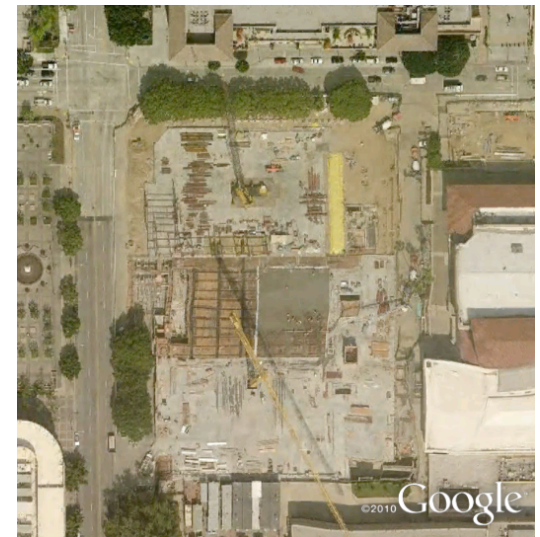
- Site 5: Pasadena Convention Center foundation



DPM



2007.10.23



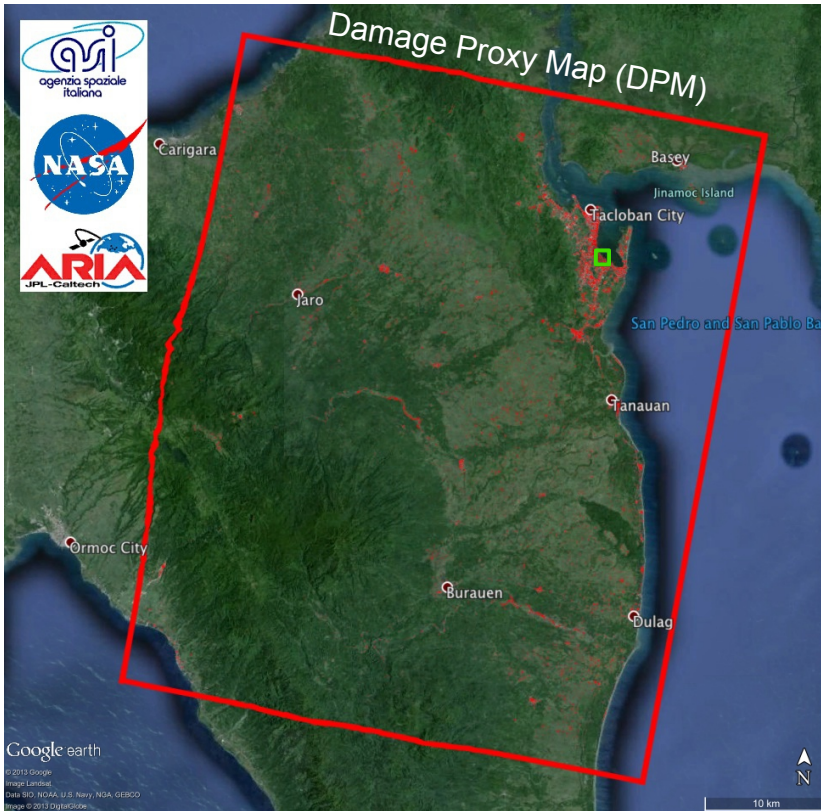
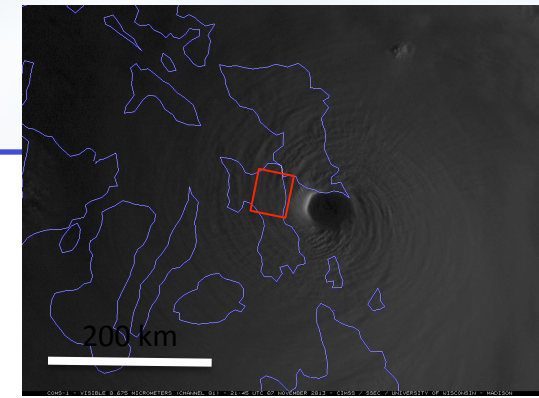
2008.01.09

Source: Sang-Ho Yun (JPL)

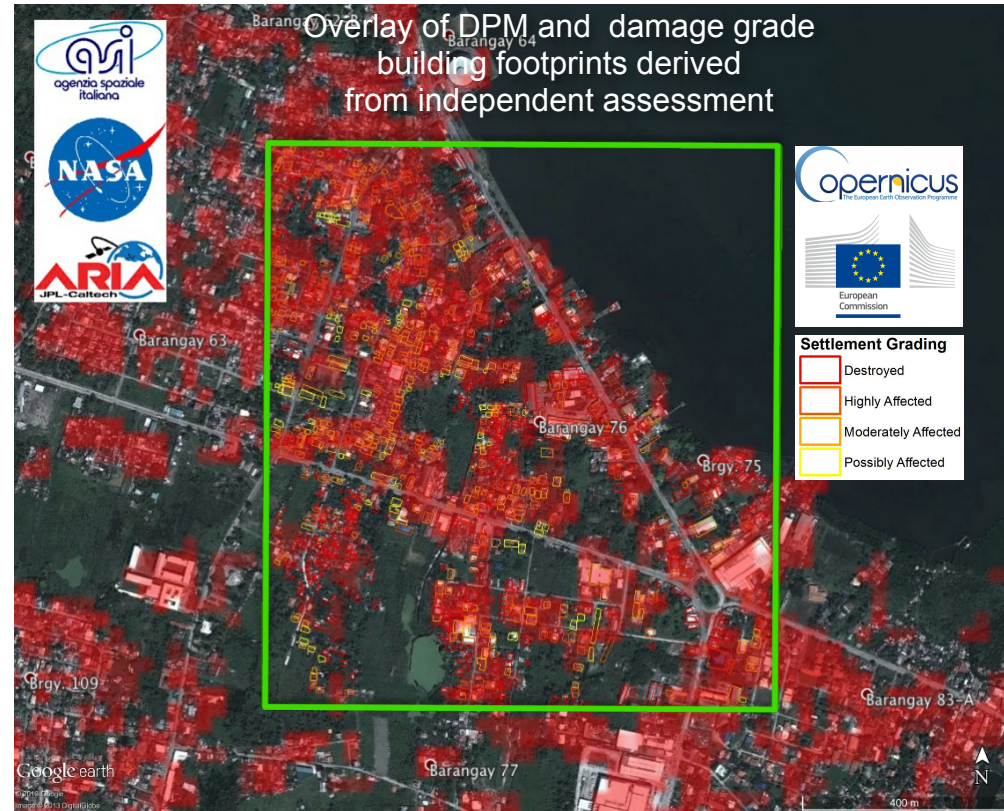


# Super Typhoon Haiyan Damage in Tacloban, Philippines Imaged with COSMO-SkyMed

- 2013-11-08 00:00 (UTC): Haiyan hit Philippines
- 2013-11-11 (Day 3): COSMO-SkyMed (X-band) data acquired
- 2013-11-11 (Day 3): Damage Proxy Map produced by ARIA
- Map distributed to organizations responding to the disaster.



COSMO-SkyMed © ASI (acquired on 2013/08/15, 2013/08/19, 2013/11/11) under a joint collaboration between JPL and ASI



Damage grade polygons derived by Copernicus Emergency Management Service from visual interpretation of pre-event and post-event optical images

**Settlement Grading**

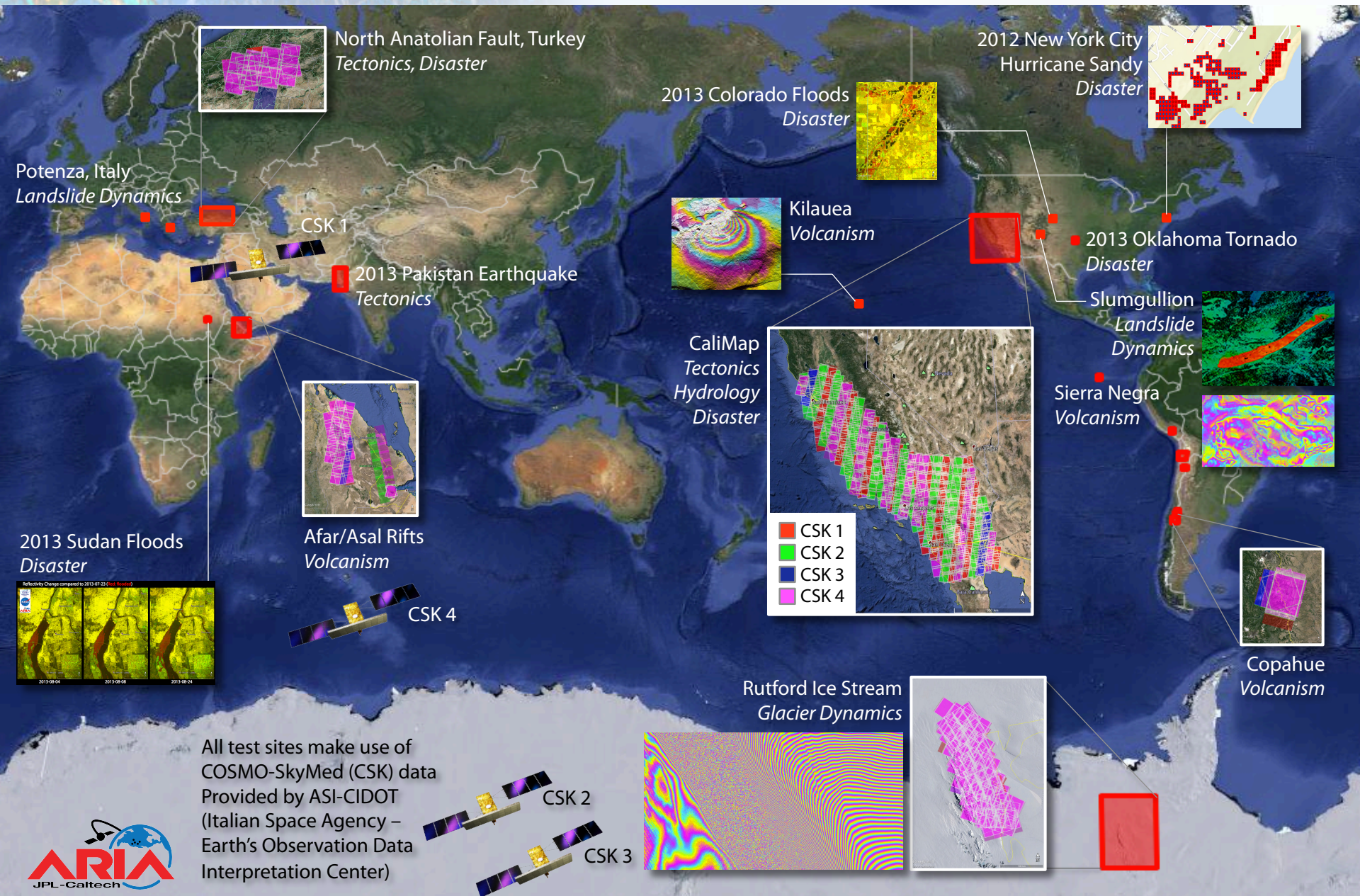
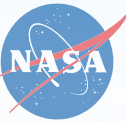
- Destroyed
- Highly Affected
- Moderately Affected
- Possibly Affected





- CaliMap (JPL/Caltech/ASI)
  - California coverage since 2013-05-24
  - 12 complete 10,000+ possible pairs
  - CSK data -25% of pairs meeting interferable criteria
    - cycles ascending and descending
- Dynamic Processes (AGI)
- Active Volcanoes (NASA ESI)
- Antarctica Monitoring (NASA Cryosphere Science)
- Asal Rift (NASA ESI)
- Other analysis projects benefiting from ARIA-MH web services
  - Slumgullion landslide
  - San Francisco Bay Faults
  - Baja California post-seismic analysis
  - La Habra earthquake analysis
- *Note: CSK data stream is used as a testbed to design and verify the ARIA-MH infrastructure. There is no plan to provide a long-term "service" to the general community on behalf of the Italians for access to CSK data and products, but we can make interferogram products available in the long-term.*

# Projects Supported by ARIA-MH (AIST 2011)







# Motivation for Cloud Computing

- Incoming flood of data volume
  - Nominally 1.2GB/scene
  - 100Ks of scenes
  - 10GBs-100GBs temp storage per data product processing
  - PBs-scale data products
  - Example InSAR satellites
    - COSMO-SkyMed (CSK) and CSK second generation data from ASI
    - Sentinel 1A/1B
    - ALOS-2
    - Decadal Survey: proposed NI-SAR mission (US L-band SAR)
- “Embarrassingly parallel” data product generation
- Monitoring, Subscriptions, and Actions
  - User definable bounding box regions of interest for nominal background monitoring/processing.
- Elasticity of computing when responding to events
- Process migration to geographically disperse data centers
  - ESDIS DAACs (e.g. ASF)
  - UNAVCO SAR Archive
  - ASI for CSK
  - DLR for TerraSAR-X
  - JAXA
  - Various GEO Supersites



# Compute, Data, & Cost Estimates



- Notional analysis comparing local hardware purchase versus AWS GovCloud usage
  - Process 16-days of data in at most 8-days.
    - 26K compute hours on 8-core nodes (3 years wall-clock processing)
  - EC2 instances with
    - Persistent EBS for cached data
    - Ephemeral local VM disk for scratch disk
  - Use AWS Glacier for cheaper long term storage (with lower data access latency)
  - Break-even point before 1-year mark
  - AWS market prices frequently changes (*estimate already outdated*)
  - Costs based on today's dollars. Does not account for inflation. Does not account for future AWS price drops.
  - On-premise costs do not consider overhead costs of cooling and electrical

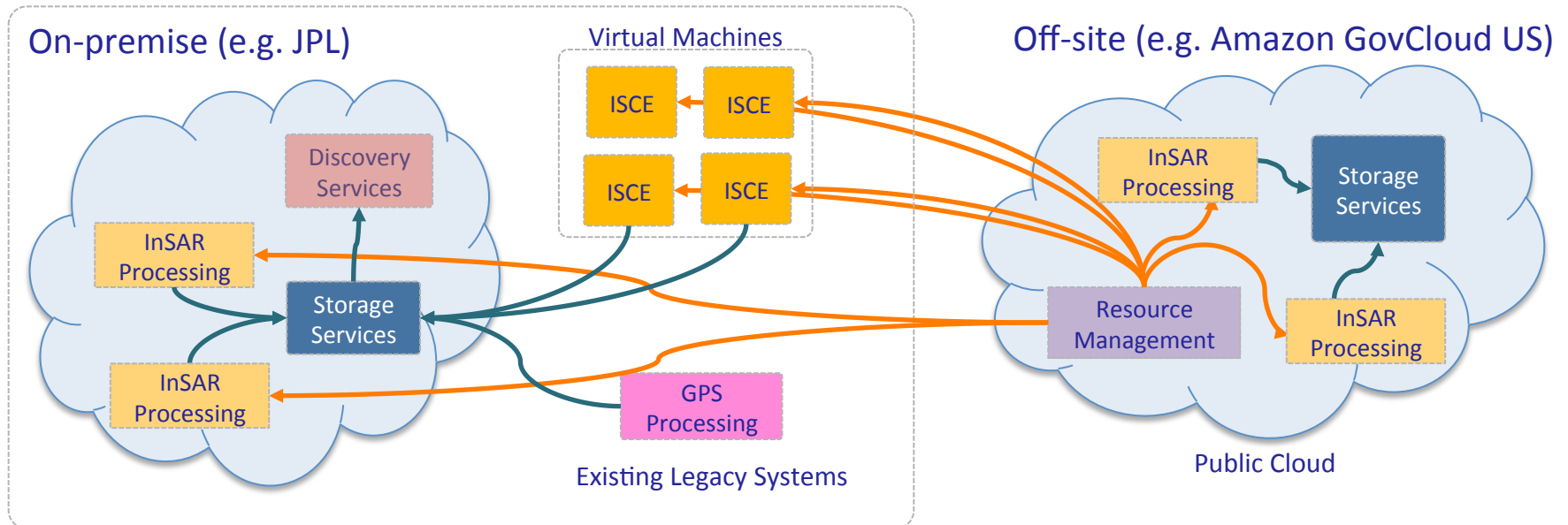
Cumulative months

		3	6	9	12	15	18	21	24	27	30	33	36
<b>Local Cluster (Cumulative)</b>													
Long-Term Storage	\$460.1	\$2.7	\$11.9	\$27.5	\$49.7	\$78.3	\$113.4	\$155.0	\$203.0	\$257.6	\$318.6	\$386.1	\$460.1
Transfer Fee for LTS data	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0
Transfer Fee for results	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0
CPU	\$135.0	\$11.3	\$22.5	\$33.8	\$45.0	\$56.3	\$67.5	\$78.8	\$90.0	\$101.3	\$112.5	\$123.8	\$135.0
Total		\$14.0	\$34.4	\$61.3	\$94.7	\$134.6	\$180.9	\$233.7	\$293.0	\$358.8	\$431.1	\$509.9	\$595.1
<b>Amazon GovCloud (Cumulative)</b>													
Long-Term Storage		\$0.21	\$1.1	\$3.2	\$7.0	\$13.0	\$21.6	\$33.4	\$48.9	\$68.5	\$92.8	\$122.2	\$157.3
Transfer Fee for LTS data		\$0.97	\$5.2	\$15.1	\$32.9	\$61.0	\$101.7	\$157.3	\$230.2	\$322.6	\$436.9	\$575.4	\$740.5
Transfer Fee for results		\$0.00	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0	\$0.0
CPU		\$6.22	\$18.7	\$37.3	\$62.2	\$93.3	\$130.6	\$174.2	\$223.9	\$279.9	\$342.1	\$410.6	\$485.2
Total		\$7.40	\$25.01	\$55.65	\$102.14	\$167.31	\$253.96	\$364.92	\$503.01	\$671.05	\$871.85	\$1,108.23	\$1,383.01

# Hybrid Cloud Computing Science Data System (HySDS)



- Utilizes both **on-premise** and **off-site** infrastructure
  - Leverage existing infrastructure investment
  - PB-scale processing and storage in public cloud currently too expensive
- Hybrid Cloud data system architecture
  - **Burst out** to public cloud when demand exceeds on-premise resources
  - Deploy AWS-compatible Eucalyptus cloud stack **on-premise**
- **Heterogeneous** computing nodes
- Resource management and data discovery can run anywhere
- Deploy **localized data repositories** closer to processing VMs
- Leverage **Amazon GovCloud US** to address export control and firewall security issues

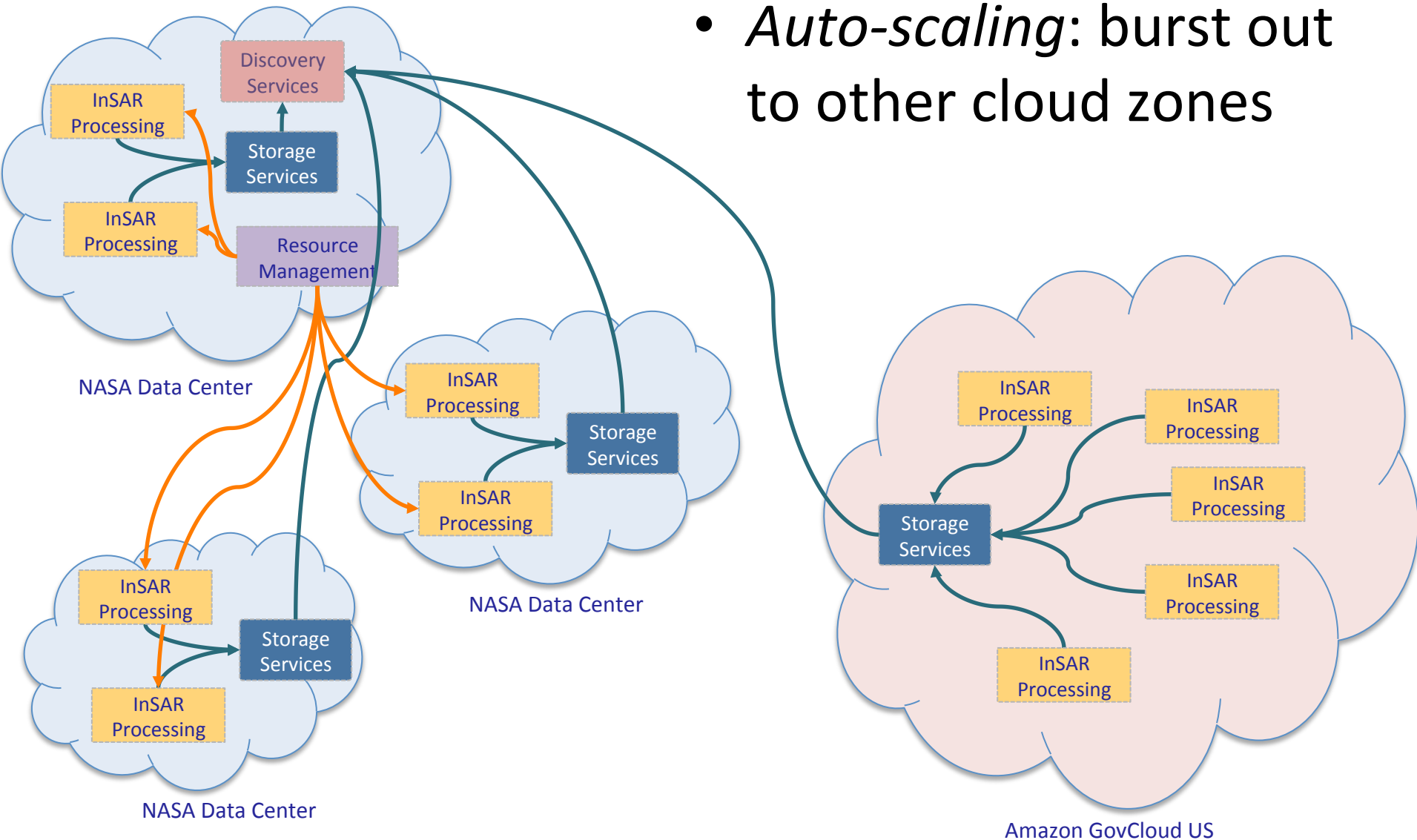




# Hybrid Cloud Auto-Scaling



- *Auto-scaling*: burst out to other cloud zones



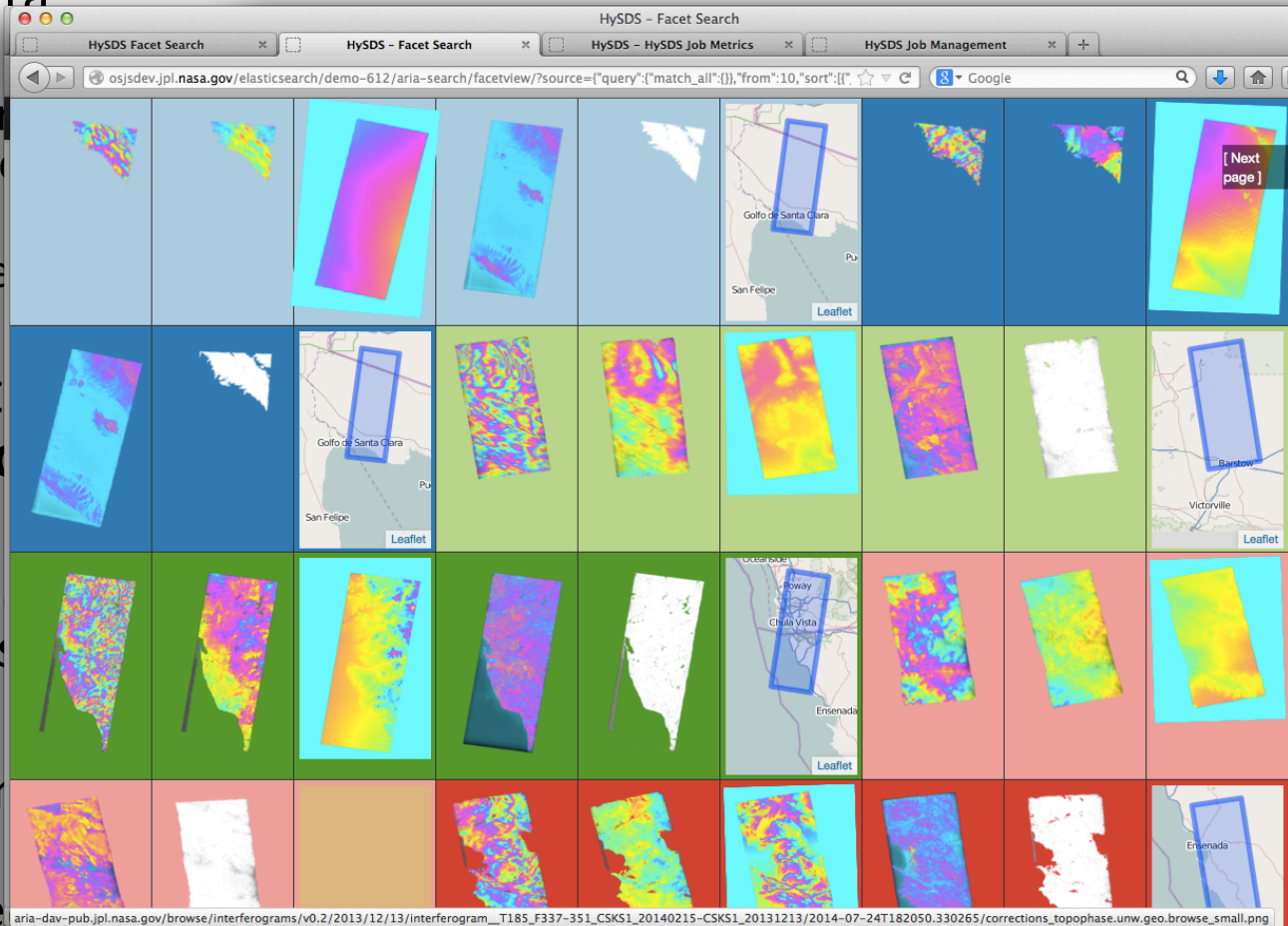
# Faceted Search of Data Products

- Faceted view of data products

- Enable users to “drill down” into multi-dimensional facets of data
- Sequentially apply constraints

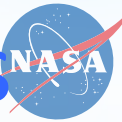
- Collaboration with crowd-sourced search tagging

- Facets for *reverse geocoding* for all data products
- Interactive *leaflets*
- REST web service endpoints also in use
  - Exposes same facets and constraints





# Empowering Users with Faceted Rules



HySDS Facet Search

https://aria-search.jpl.nasa.gov/?source=[{"query":{"match\_all":{}},"sort":[{"\_timestamp":{"order":"desc"}}],{"fields":{"\_timestamp":{}}]

Home Facet Search Repository KML Jobs Incoming

Logged in as: **ariaops** My Rules Logout

## HySDS Facet Search

faceted search interface for GeoRegionQuery

search term

**user tags**

count	OR	range
LC60 (46)		
HT4 (27)		
HT6 (25)		
LC50 (21)		
HT5 (20)		
LC40 (16)		
LC30 (15)		
chirp extension noise (13)		
HT9 (12)		
uncorrected (11)		

**system version**

count	OR	range
v0.4 (29588)		
v0.2 (23547)		
v0.3 (8123)		
v0.0 (3857)		
v0.1 (510)		

**dataset**

count	OR	range
CSK-RAW_B (28968)		

min latitude: -90  
max latitude: 90  
min longitude: -180  
max longitude: 180  
 Round values?

1 - 10 of 65625 next » Monitor Results Process Results

**CSKS2\_RAW\_HI\_12\_HH\_RD\_20141029015000\_20141029015007 (CSK-RAW\_B)**

region

Tuolumne Merced Madera Fresno Soledad

# Geospatial Faceted Search



Facet Search faceted search interface for GeoRegionQuery

search term

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]]]

Facet Search

HT9 (9)  
HT3 (7)

system version  
v0.0 (2951)  
v0.1 (261)

spacecraft  
CSKS4 (1542)  
CSKS1 (1418)  
CSKS2 (1072)  
CSKS3 (106)

orbit direction  
dsc (2399)  
asc (1739)

orbit number  
16124 (80)

next 10

https://ojsdev.jpl.nasa.gov:elasticsearch/demo/faceview/?source='query'&bool='must'&term='dataset:interferogram'&filter='geo\_shape:location'

- Spatial extents (polygon, bbox, circle, and reverse-geocoded region name)
- Temporal extents
- **Faceting with ESDIS GIBS** near real-time (NRT) map view overlays



# Facet Search with ESDIS GIBS Near Real-Time Basemaps



HySDS - Facet Search

osjstdev.jpl.nasa.gov/elasticsearch/demo-612/aria-search/facetview/?source={\"query\":{\"match\_all\":{}},\"size\":20,\"sort\":{\"\_timestamp\":{\"order\":

## HySDS - Facet Search

user tags

? 10 count ↓ OR

range

LC60 (46)  
HT4 (27)  
HT6 (25)  
LC50 (21)  
HT5 (20)  
LC40 (16)  
LC30 (15)  
chirp extension noise (13)  
HT9 (12)  
uncorrected (11)

system version

? 10 count ↓ OR

range

v0.4 (29588)  
v0.2 (23547)  
v0.3 (8123)  
v0.0 (3857)  
v0.1 (510)

dataset

? 10 count ↓ OR



range

CSK-RAW\_B (28968)  
CSK (26469)

Search term: [input field]

1 - 20 of 65625 next » WGET Script Download All

CSKS2\_RAW\_HI\_12\_HH\_RD\_20141029015000\_20141029015007 (CSK-RAW\_B)



tags: jpl-temp  
version: B-SF  
system version: v0.4  
continent: North America  
location: Fresno County, California, United States  
cities: Three Rocks  
bbox: 36.67,-120.28,36.74,-120.85,36.26,-120.37,36.33,-120.93  
sensing start: 2014-10-29T01:50:00.013716 | sensing stop: 2014-10-29T01:50:06.713362  
spacecraft name: CSKS2 | track number: 59 | beam number: 12 | lat index min: 362 | lat index max: 368  
orbit direction: dsc | look direction: right | orbit number: 37268 | orbit repeat: 237  
vertical baseline: -27.73891021443932  
horizontal baseline: -287.10422526537036  
perpendicular baseline: 207.33613284293753  
product type: RAW\_B  
requestor: MAPCALIFORNIA\_CIDOT\_JPL\_2013

# Faceted Resource Management



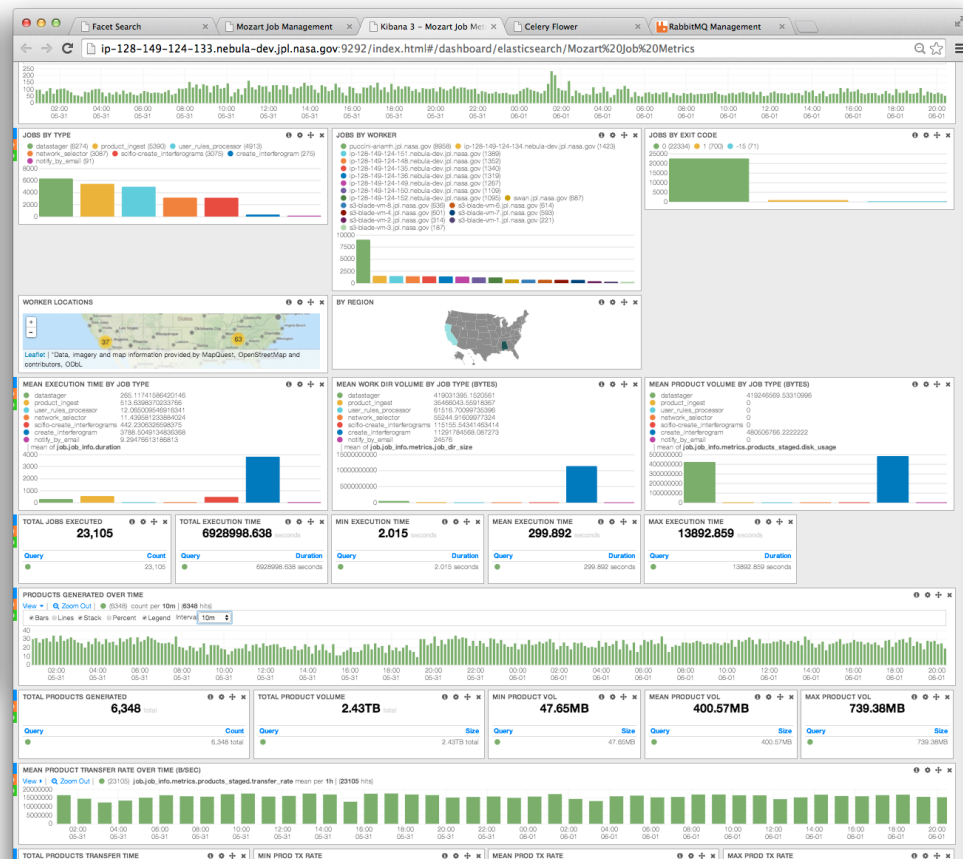
- Facetted real-time view of distributed cloud compute jobs
  - Status
  - VM types
  - Compute nodes
  - Cloud zones
  - Hardware resource utilization
  - By hardware specs
  - DAV view
  - Etc.

The screenshot displays the Mozart Job Management web interface. The browser address bar shows the URL: `mozart-ariamh.jpl.nasa.gov:8888/?source={"query":{"bool":{"must":{"t`. The page title is "Mozart Job Management". The main navigation bar includes buttons for "Total", "Quued", "Running", "Completed", and "Failed". A modal window is open, showing details for a specific job. The job ID is `create_interferogram-CSKS2_RAW_HI_05_HH_RD_20130525020436_20130525020443.interferogram.json_2-20131213T064`. The modal window has tabs for "Job Info", "Context", "STDOUT", "STDERR", and "Work Dir", with "Work Dir" selected. The "Work Dir" tab shows a file listing for the location `/jobs/2013/12/13/create_interferogram-CSKS2_RAW_HI_05_HH_RD_20130525020436_20130525020443.interferogram.json_2-20131213T064634.670636Z/`. The file listing table has columns for "Name", "Last modified", and "Size".

Name	Last modified	Size
Parent Directory	-	-
checkInterferogramByInputHash.log	13-Dec-2013 08:28	153
context.json	13-Dec-2013 08:28	444
createInterferogram_0.log	13-Dec-2013 08:40	35K
CSKS2_RAW_B_HI_05_HH_RD_SF_20130525020432_20130525020439.h5	25-May-2013 04:36	1.0G
CSKS2_RAW_B_HI_05_HH_RD_SF_20130525020436_20130525020443.h5	25-May-2013 04:41	1.0G
CSKS2_RAW_B_HI_05_HH_RD_SF_20130525020441_20130525020448.h5	25-May-2013 04:49	1.0G
CSKS2_RAW_B_HI_05_HH_RD_SF_20130525020446_20130525020452.h5	25-May-2013 04:37	1.0G
CSKS2_RAW_B_HI_05_HH_RD_SF_20130813020359_20130813020406.h5	13-Aug-2013 06:28	1.0G
CSKS2_RAW_B_HI_05_HH_RD_SF_20130813020403_20130813020410.h5	13-Aug-2013 06:44	1.0G
CSKS2_RAW_B_HI_05_HH_RD_SF_20130813020408_20130813020415.h5	13-Aug-2013 06:28	1.0G
CSKS2_RAW_B_HI_05_HH_RD_SF_20130813020413_20130813020419.h5	13-Aug-2013 06:29	1.2G
CSKS2_RAW_HI_05_HH_RD_20130525020436_20130525020443.interferogram.json_2	13-Dec-2013 08:28	54K
file0D8FJl	13-Dec-2013 08:41	1.0G
filezOTW1u	13-Dec-2013 08:41	232M
getInputHash.log	13-Dec-2013 08:28	0
insar.log	13-Dec-2013 08:28	0
insarMH.xml	13-Dec-2013 08:37	1.5K
interferograms_found.txt	13-Dec-2013 08:28	7
isce.log	13-Dec-2013 08:40	622
job.json	13-Dec-2013 08:28	4.9K
netset_hash.txt	13-Dec-2013 08:28	32
output_0.raw_0	13-Dec-2013 08:38	1.0G
output_0.raw_0.aux	13-Dec-2013 08:38	350K
output_0.raw_1	13-Dec-2013 08:39	1.0G
output_0.raw_1.aux	13-Dec-2013 08:39	350K
output_0.raw_2	13-Dec-2013 08:40	1.0G
output_0.raw_2.aux	13-Dec-2013 08:40	349K
output_0.raw_3	13-Dec-2013 08:40	1.0G
output_0.raw_3.aux	13-Dec-2013 08:40	350K
stderr.txt	13-Dec-2013 08:28	490



# Dashboard for Real-Time Faceted Metrics for Distributed Hybrid Cloud Data System



- Dashboard for **faceted metrics**
- Real-time metrics of distributed hybrid cloud computing infrastructure
- Integrated metrics across on-premise and off-site cloud compute nodes

# Computing Resources Used



- On-Premise Cloud Computing Resources

- Eucalyptus Availability Zone at JPL OCIO 600 Data Center
- Eucalyptus Availability Zone at JPL OCIO 230 Data Center
- Eucalyptus Availability Zone at JPL 202 Data Center
- Nebula cloud compute nodes at JPL
- OpenStack Zone at JPL OCIO 600 Data Center

- Public Cloud Computing Resources

- Amazon AWS – GovCloud US
  - Consolidated costing
  - JPL network IP address space in GovCloud instances

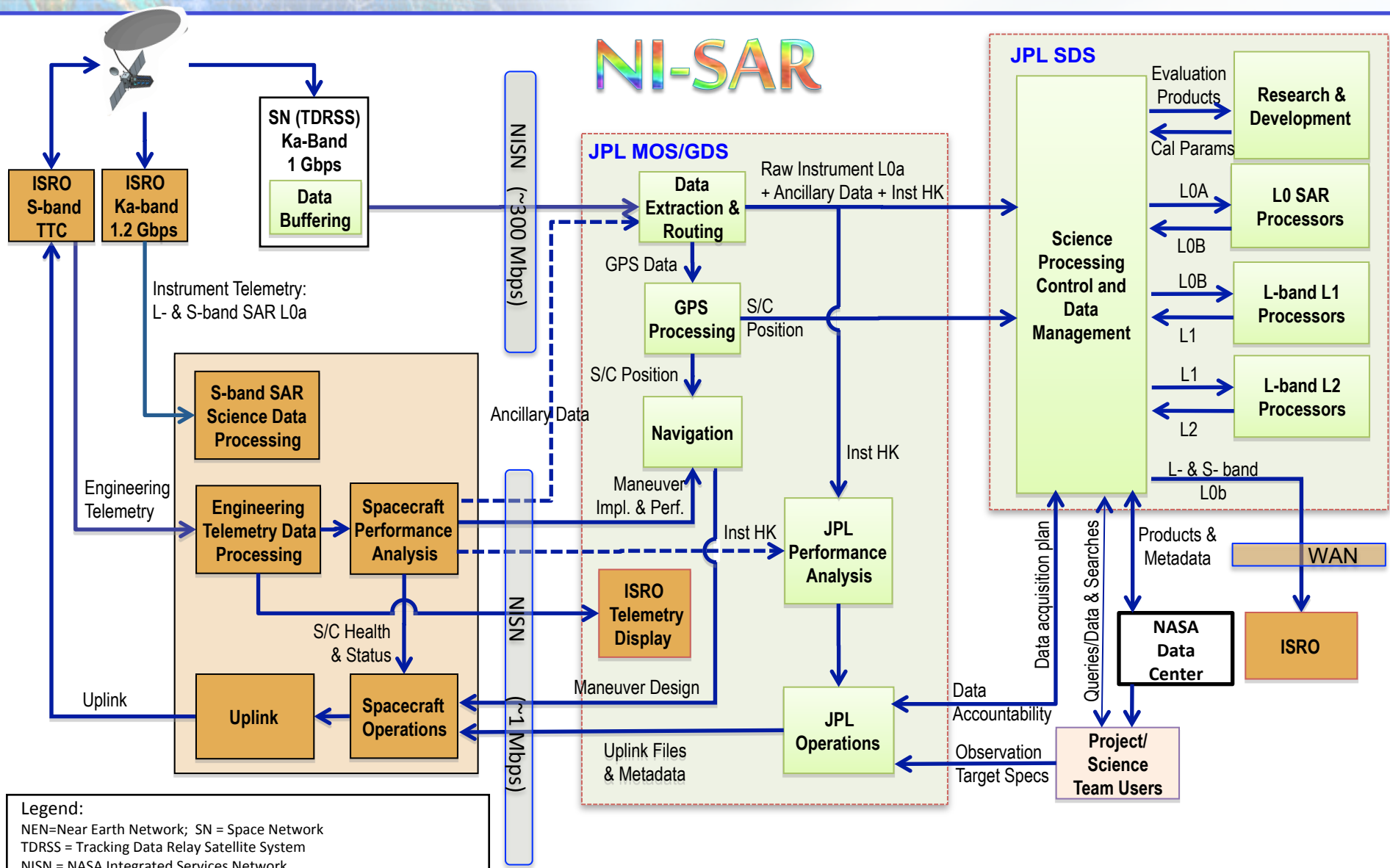
- *Auto-scaling* of computing resources

- Able to *burst* out to AWS when compute demand exceeds at of on-premise resources
- Seamless transitioning between on-premise and public compute nodes
- “*keep up*” processing of near real-time data stream
- “*bulk*” processing





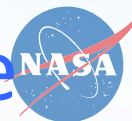
# NISAR Mission System End-to-End Data Flow



**Legend:**  
 NEN=Near Earth Network; SN = Space Network  
 TDRSS = Tracking Data Relay Satellite System  
 NISN = NASA Integrated Services Network  
 GPS = Global Positioning System;  
 WAN=Wide Area Network  
 HK=HouseKeeping; Inst=Instrument

2014-10-30T08:30-04:00

# Estimated NISAR Daily Scene Data Volume



Data Volume Loading (GB)		NI-SAR - L/S-Band			
Processors	Input Data	Input	Intermediate	Output	To DAAC
LOA (catalog incoming raw data)	from s/c	15		15	
LOB	LOA	15		17.25	17.25
Subtotal				17.25	17.25
L1					
SLC	LOB	17.25		159	159
MLC	SLC		40		
MLD Browse	MLC			2.5	2.5
Subtotal				161.5	161.5
L2 (all modes)	pair LOB	34.5			
SLC pairs, internal			318		
Interferogram				40	40
Pwr Images (Mst/Slv)				40	40
Coherence				20	20
Unwrapped itfm (desired)				20	
Geocoded itfm				40	40
Geocoded Coherence				20	20
Geocoded Amplitude				40	40
Geocoded Unwrapped itfm (desired)				20	
Subtotal				240	200
L2 Biosphere (Quad)	LOB	13.8			
SLC, internal			127.2		
Stoke's matrix				23	23
Polarimetric Coherence (desired)				23	
Geocoded Stoke's matrix				23	23
Geocoded Polarimetric Coherence (desired)				23	
Subtotal				92	46
Total				510.75	424.75

Assumption:  
 240x240 km data takes  
 15 GB/scene or  
 200 scenes/day and  
 scale all numbers  
 proportionally from  
 daily numbers



# Big Data Handling



- 3TB/day LOA from ground data system
  - Nominally 200 scenes/day @ 15GB/scene (240kmx240lm)
- 85TB/day derived data products (LOA to L2 interferograms) to DAAC. (~31 PB/year)
- *Estimates of sustained ~1GB/sec data transfer from SDS to DAAC*
- These are upper-bound estimates
  - NISAR L-band uses lower resolution (10m) as compared to existing higher-resolution CSK processing (3m)
  - Estimates are for storage in historic raw (e.g. BIP, BIL, BSQ) format
  - Use of HDF5 with chunked compression will reduce volume requirements
- Handling NISAR scenes
  - New ISCE processing code needs to be developed.
    - Currently no concrete ISCE estimates on LOA to L2 processing time of NISAR 240km x 240km data takes.
  - 240km x 240km, 10m resolution
  - Phase unwrapping expected to be less of an issue as NISAR L-band has less decorrelation
- ARIA-MH's CSK processing as reference
  - CSK's use of HDF5 saw 3-4X in size reduction from raw storage. Compression ratio is also scene-dependent.
  - CSK scene processing from LOA to L2 unwrapped interferogram
    - Under 1-hour on 8-core compute nodes for 40km x 100km, 3m resolution. ~4 scene swath with 5% overlap.
- New architectures needed to support data handling and throughput
  - SDS processing at JPL *on-premise*
    - *Bottlenecks with data transfer out to DAAC*
  - SDS processing **at** DAAC on-premise
    - NASA center politics and policy updates needed
  - SDS processing at public cloud
  - Hybrid cloud: on-premise and public cloud for processing and data storage

# AIST Impacts to NISAR and SWOT



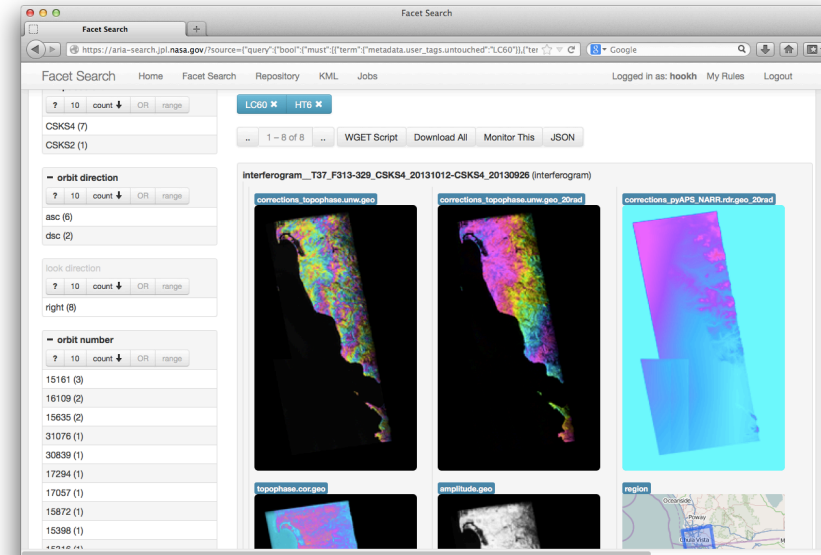
- AIST2011 funds ARIA-MH
- Developed NASA's first hybrid-cloud computing science data system (HySDS)
  - Elastic InSAR processing
  - Elastic data storage
    - Public cloud object store costs still an issue
    - On-premise cloud storage options
  - Process migration (moving processing closer to large data)
  - Multi-dimensional faceted browse
    - For data products
    - For situational awareness of science data system
    - Faceted user rules for conditional monitoring and processing
    - Collaboration
- Hybrid-cloud computing science data system architecture (HySDS) developed for ARIA-MH being assessed for **reuse for NISAR and SWOT**
  - HySDS can scale up to NISAR and SWOT needs
  - HySDS enables SDS architectures to process Big Data volumes on-premise, in public cloud, and at DAACs to minimize data movement.



# Big Data Analytics Needs



- **Real End-User Needs**
- Machine tags
  - Production-system auto-generated machine tags
- User-based social tagging
  - Social tagging by users
- Quality Assessment
  - Patterns of data from tagging
- Analysis
  - Improve understand through QA and tagging
- **Need for Big Data Analytics**
  - Too much data to manually tag
  - Automating large-scale analytics
  - Improve quality issues with the high-volume interferograms such as unwrapping problems, mis-registration, erroneous coherence, atmospheric noise and/or ionospheric artifacts.



# Infusion Points



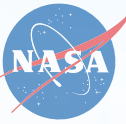
(MH)  
Coordinating multiple funded efforts  
General hybrid-Cloud Computing  
Space Data System to lower risks

The screenshot shows a web browser window with the URL <https://msas-search.jpl.nasa.gov/>. The page title is "Facet Search" and it includes navigation links for "CVO", "Repository", "Jobs", and "Contact". A user is logged in as "hookh". The main content area displays a map of the United States with a red and blue overlay. Below the map are search controls, including a "Round values?" checkbox and buttons for "WGET Script", "Download All", "Monitor This", "Process This", and "JSON". On the left, there are facet filters for "system version" (v6, 94) and "dataset" (AIRS3RET, 4162). A "resolution" filter is also present, showing options like "(1 day, 24 standard pressure levels, 1 degree, 1 degree) (4162)".

The screenshot shows a web browser window with the URL <https://gcis-search.jpl.nasa.gov/>. The page title is "GCIS Facet Search" and it includes a "Home" link. The main content area displays a search results page for "1700-years-of-global-temperature-from-proxy-data". The search term is "1700-years-of-global-temperature-from-proxy-data" and the results are sorted by relevance. The page shows 744 results. A prominent result is titled "1700-years-of-global-temperature-from-proxy-data (figure)" with a sub-title "title: 1700 years of Global Temperature from Proxy Data". Below the title is a line graph showing "1700 Years of Global Temperature Change from Proxy Data" with a y-axis labeled "Temperature (K)" and an x-axis labeled "Year". The graph shows a clear upward trend in temperature over the last 1700 years, with a significant increase in the last century. The graph includes a legend for "observed", "proxy-based records", and "thermometer-based records". Below the graph is a "caption" section that reads: "Changes in the temperature of the Northern Hemisphere from surface observations (in red) and from proxies (in black; uncertainty range represented by shading) relative to 1961-1990 average temperature. These analyses suggest that current temperatures are higher than seen globally in at least the last 1700 years, and that the last decade (2001 to 2010) was the warmest decade on record. (Figure source: adapted from Mann et al. 2008ade3fd09-603e-4fae-b252-1a4142392ea0).". Below the caption are links for "html:" and "json:". The "html:" link is <http://data.globalchange.gov/report/nca3/chapter/appendix-climate-science-supplement/figure/1700-years-of-global-temperature-from-proxy-data> and the "json:" link is <http://data.globalchange.gov/report/nca3/chapter/appendix-climate-science-supplement/figure/1700-years-of-global-temperature-from-proxy-data.json>. Below the graph is another result titled "1997-1998-el-nino-event (figure)" with a sub-title "title: The 1997 to 1998 El Nino Event".

- A Multi-Sensor Water Vapor,
  - MEaSUREs
  - Reuse of HySDS for MEaURE
- **Planned External Infusions**
  - Italian Space Agency (ASI)
  - USGS Hawaiian Volcano Observatory (HVO)

# Summary



- The global coverage offered by satellite-based SAR missions, and rapidly expanding GPS networks can provide orders of magnitude more observations and improve hazard response
  - ...if we have a data system that can efficiently monitor and analyze the voluminous data, and provide users the tools to access data products.
- Hybrid cloud may be more effective for these needs
  - Do “keep up” processing **on-premise**
  - **Off-site** elasticity (bursting) for bulk processing
- Hybrid cloud computing may be a more effective approach to addressing **lower latency** and **Big Data volume** processing needs
- Big Data Analytics needed to improve quality of data products and processing algorithms
- PB-scale data volumes in cloud computing and storage significant enough to **affect architectural design of data system**
- **Faceted browse** of data system and data products improve understanding
- Real-time dashboards for “**situational awareness**”
- **Monitoring**
  - events for automatic processing
  - data products for custom actions
- AIST-funded HySDS
  - Already infused into ACCESS, MEaSURES, and GCIS data systems
  - Will be infused into USGS and ASI
  - Being assessed for NISAR and SWOT science data systems