When seconds matter – Big Data real-time streaming analytics and machine learning for geoscience and hazards research

Charles Meertens[1], David Mencin[1], Scott Baker[1], Kathleen Hodgkinson[1], Shelley Olds[1], Diego Melgar[2], Ivan Rodero[3], Anthony Simonet[3], J.J. Villalobos[3], Kristy Tiampo[4], Brie Corsa[1] UNAVCO, Boulder, Colorado; [2]Department of Earth Sciences, University of Oregon; [3]Rutgers Discovery Informatics Institute, Rutgers University; [4]CIRES and the Department of Geological Sciences, University of Colorado, Boulder

Presentation at October 2019 GeoSciFramework Annual Project Meeting
Alexandria, VA

Acknowledgements

Collaborative Research:NSCI:HDR:Framework:Data:GeoSCIFramework: Scalable Real-time Streaming Analytics and Machine Learning for Geoscience and Hazards Research
(Charles Meertens, UNAVCO [Award # 1835791, Ivan Rodero, The State University of New Jersey [Award # 1835692], Diego Melgar, University of Oregon [Award # 1835661], Kristy Tiampo, University of Colorado [Award # 1835566]

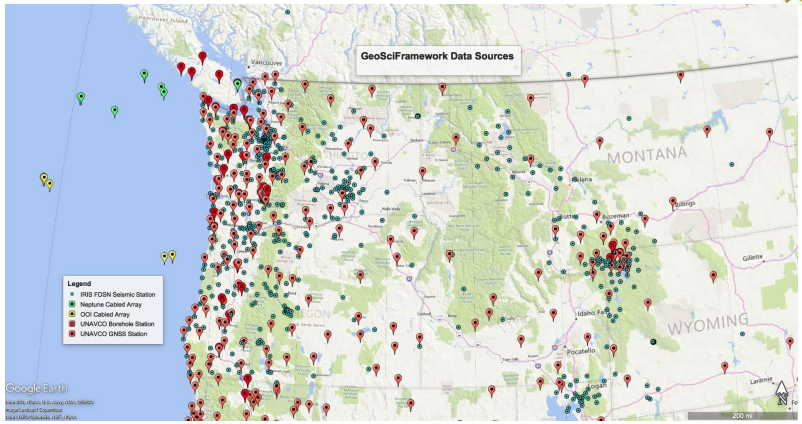
This 4-year award by the NSF Office of Advanced Cyberinfrastructure is jointly supported by the Cross-Cutting Program and Division of Earth Sciences within the NSF Directorate for Geosciences, the Big Data Science and Engineering Program within the Directorate for Computer and Information Science and Engineering, and the EarthCube Program jointly sponsored by the NSF Directorate for Geosciences and the Office of Advanced Cyberinfrastructure.

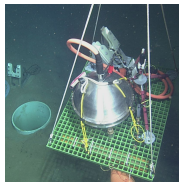
The project addresses goals of the National Strategic Computing Initiative (NSCI). a whole-of-nation effort to accelerate scientific discovery and economic competitiveness by maximizing the benefits of high-performance computing (HPC) research, development, and deployment. It also addresses the NSF *Harnessing the Data Revolution (HDR) Big Idea*. a national-scale activity to enable new modes of data-driven discovery that will allow new fundamental questions to be asked and answered at the frontiers of science and engineering.

GeoSciFramework

Project Overview: GeoSciFramework will provide an experimental computational framework that enables natural hazards research and enhanced earthquake, tsunami and volcano early warning systems.

Real-time streaming analytics and machine learning on continuous integrated data streams from thousands continental and oceanic high-rate sensors, when combined with satellite radar time series, will give a coherent high-resolution global-scale view of the motions of the earth over time periods of seconds to years.



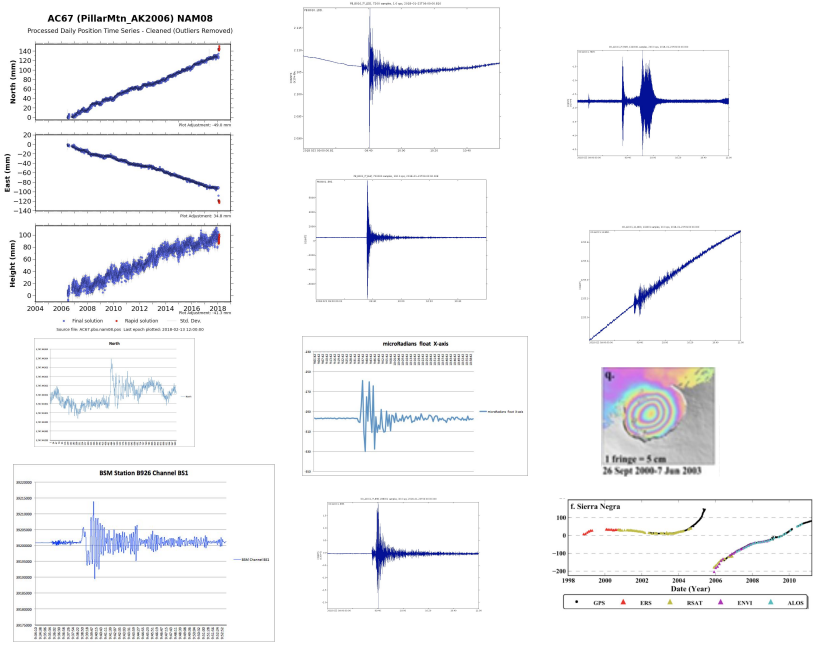


Map: Onshore and offshore sensor networks spanning Cascadia to Yellowstone. Photos: Ocean Bottom Seismometer (Rutgers/OOI), Seismograph Station (PNSN/IRIS), Borehole Strainmeter and GPS/GNSS Station at Mt. St. Helens (UNAVCO/GAGE)

- **Participating Institutions**: UNAVCO/GAGE, Rutgers University (Ocean Observatories Initiative - OOI), University of Colorado, University of Oregon
- **Collaborating Institutions**: IRIS/SAGE, University of Texas Arlington (TACC/XSEDE)

GeoSciFramework

- **Integrated data access**: The framework leverages and provides seamless access to considerable NSF investments in EarthScope (GAGE and SAGE) and OOI in situ sensor networks, internationally-operated space radar systems, and NSF XSEDE computational and data storage resources.



Sample data from GPS (1s/day), high-rate GPS (1s/sec), borehole strainmeters, pore pressure, borehole seismometer, tiltmeter, ocean bottom seismometer, ocean hydrophone, ocean bottom pressure, insar image and insar time series

EarthCube

GPS NETWORK

BINEX

BOULDER, NOC

PRIMARY SYSTEM
Data flow + processing

SECONDARY SYSTEM
Data flow + processing

CASTER 1

CASTER 2

BINEX

RTCM2.3

RTCM3.1

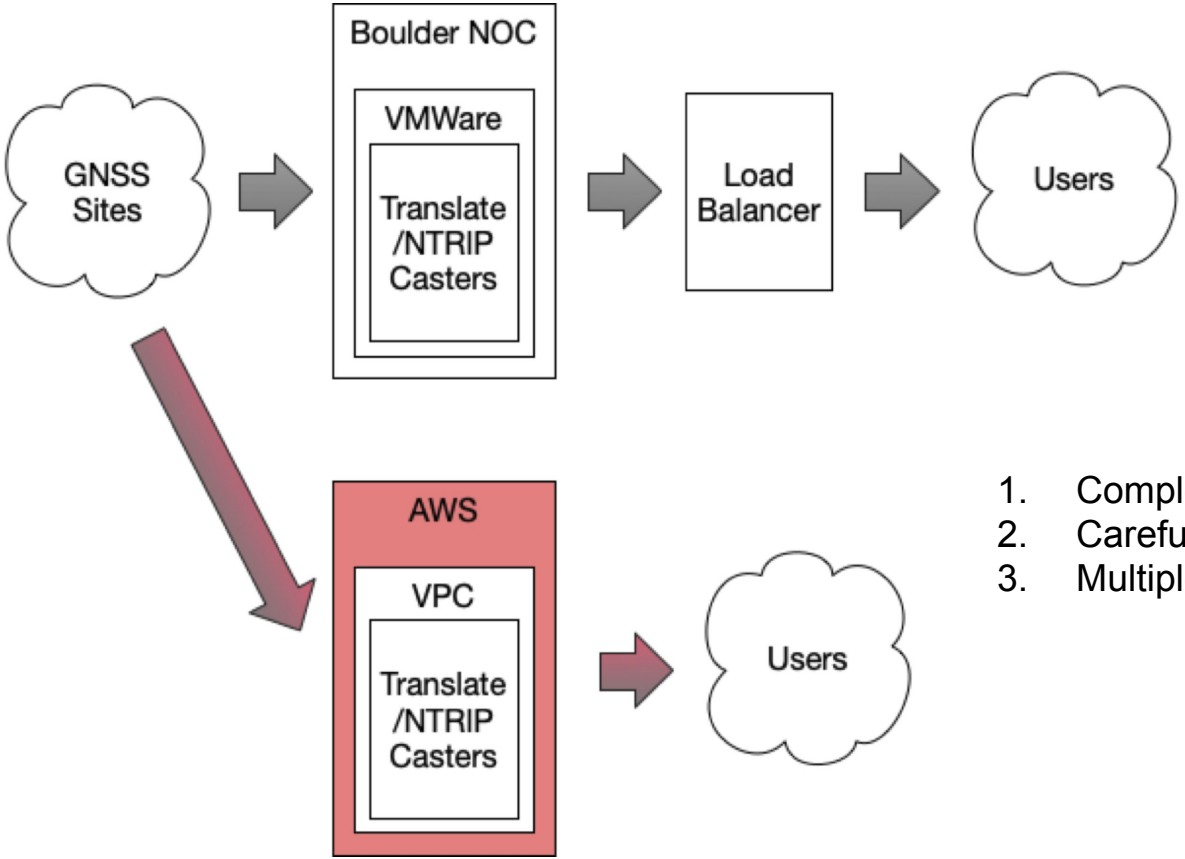
NMEA (RTX)

LOADBALANCER

SURVEYORS

RESEARCH

GOV



1. Complete overhaul of the VPN.
2. Careful cost analysis.
3. Multiple user communities.

Broader Impacts

- **Broader Impacts Activities:** Resources for internal and external capacity building are integral to the project including support for students and technical workshops, development of supportive materials such as online notebooks, and access to open software development platforms and computational resources.

- Two UNAVCO USIP student interns for Summer 2019.
- Working in-reach/out-reach material for GEOSciCloud and GEOSciFramework developing Jupyter noterbooks demonstrating and teaching access to UNAVCO data (e.g. how to access and use real-time GNSS positions)

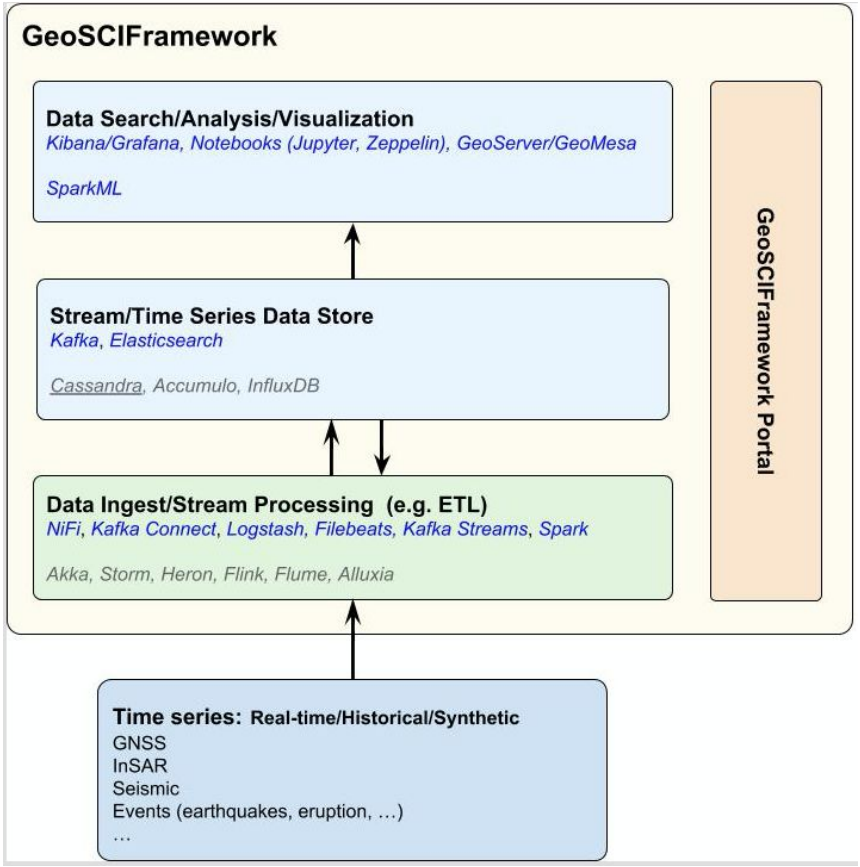
Rachel Terry

Lisa Knowles

EARTHCUBE

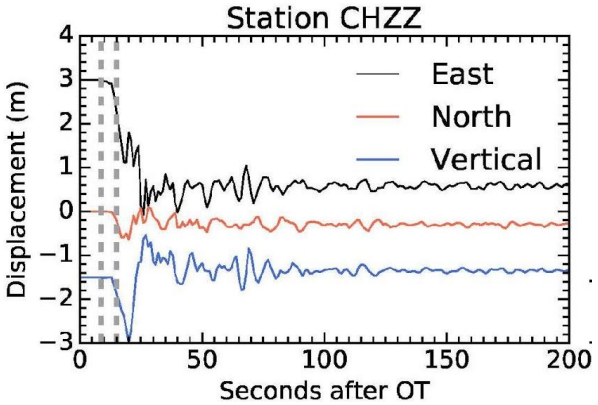
The project architecture provides generalized, scalable (laptops to cloud computing), fault-tolerant, real-time, event-based data processing and analysis capabilities for time series data from distributed sources.

- **Machine Learning**: an advanced convolutional neural network method is employed in an integrative multi-data environment. Machine learning algorithms and spatio-temporal analyses are trained using past events and informed by physics-based models.
- **Computational Resources**: The computationally intensive attempt to blindly correlate a large number of variables and large volumes of data will be performed on local clusters and NSF-funded cloud resources managed by XSEDE, such as Jetstream.
- **Application**: This method supports the automatic detection and characterization of rapid events such as earthquakes and tsunamis as well as slow-slip events or magmatic intrusions that evolve over a longer period of time expanding the potential for new scientific discoveries.
- **Broader Impacts Activities:** To facilitate discoveries, the system architecture will provide simplified access to tools, sophisticated workflows systems and training targeted at non-computer scientists (researchers and students).
- **Algorithm development**: An interactive environment allows users to test, modify, and implement their ideas as they integrate the large variety and volume of this data into new machine learning and analysis algorithms and products.



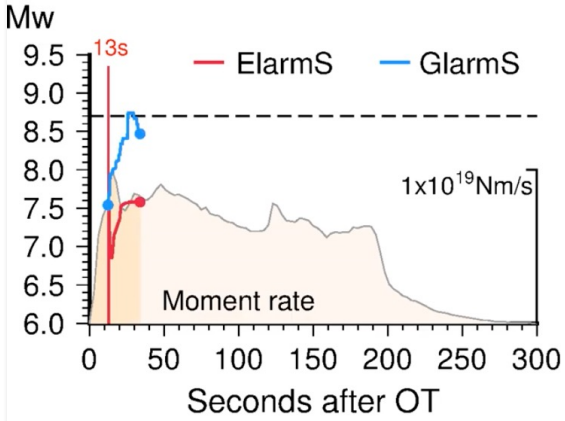
EarthCube

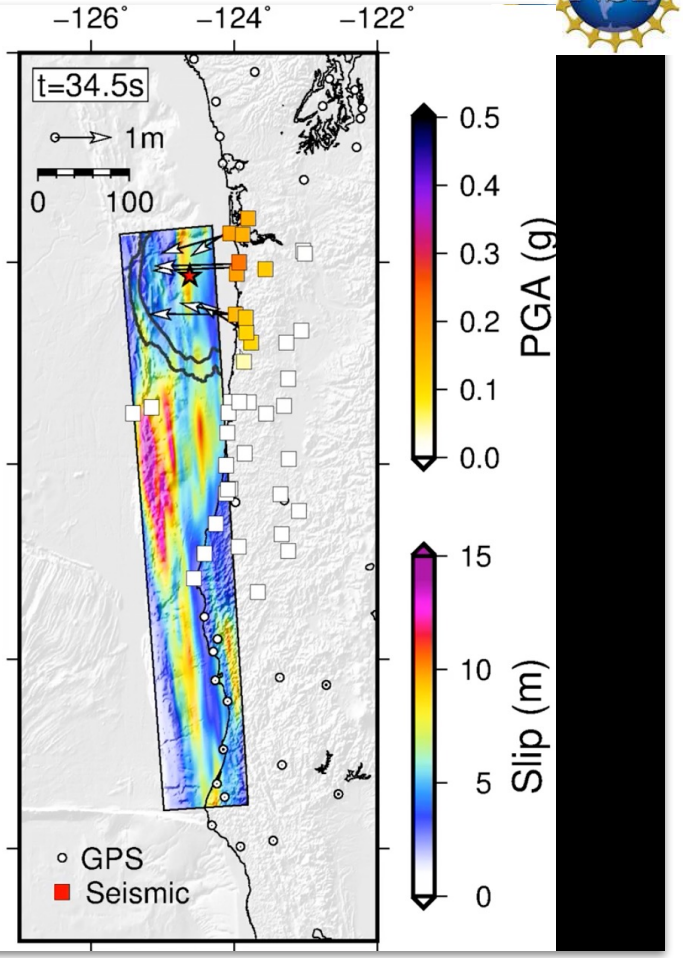
Real-time short-term forecasts



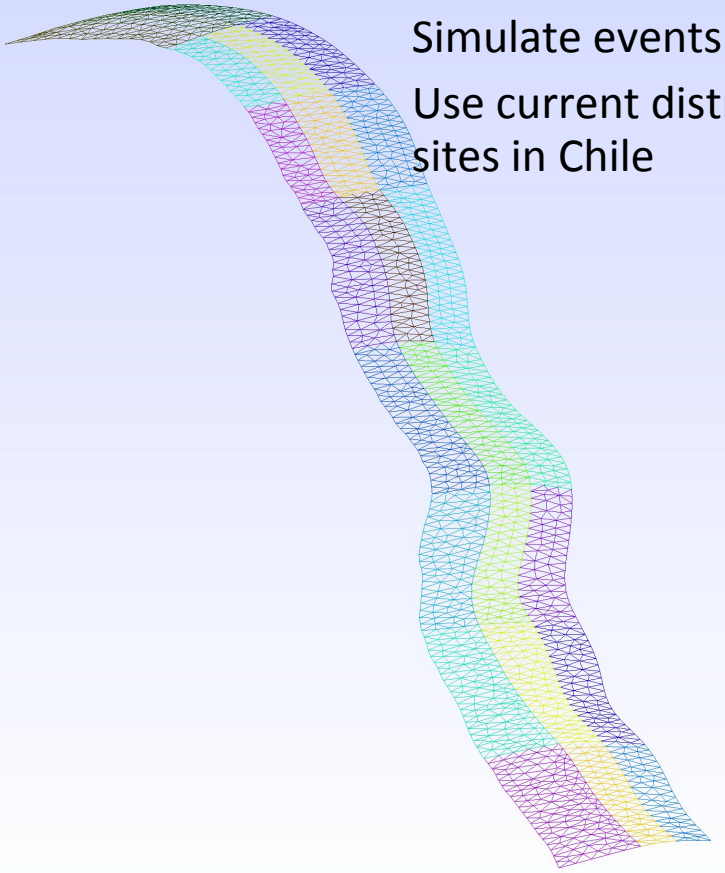
Goal:

Characterize a large earthquake and its hazards ASAP



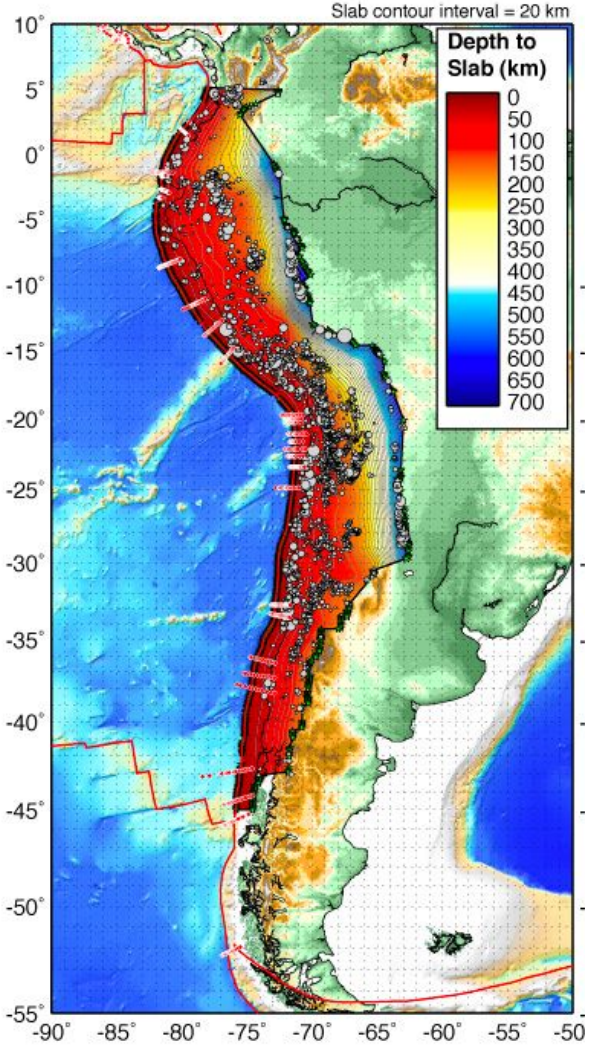


EARTHCUBE



Simulate events M7-M9

Use current distribution of sites in Chile



3400 Fakequakes

+

3400 noise data

=

6800 data

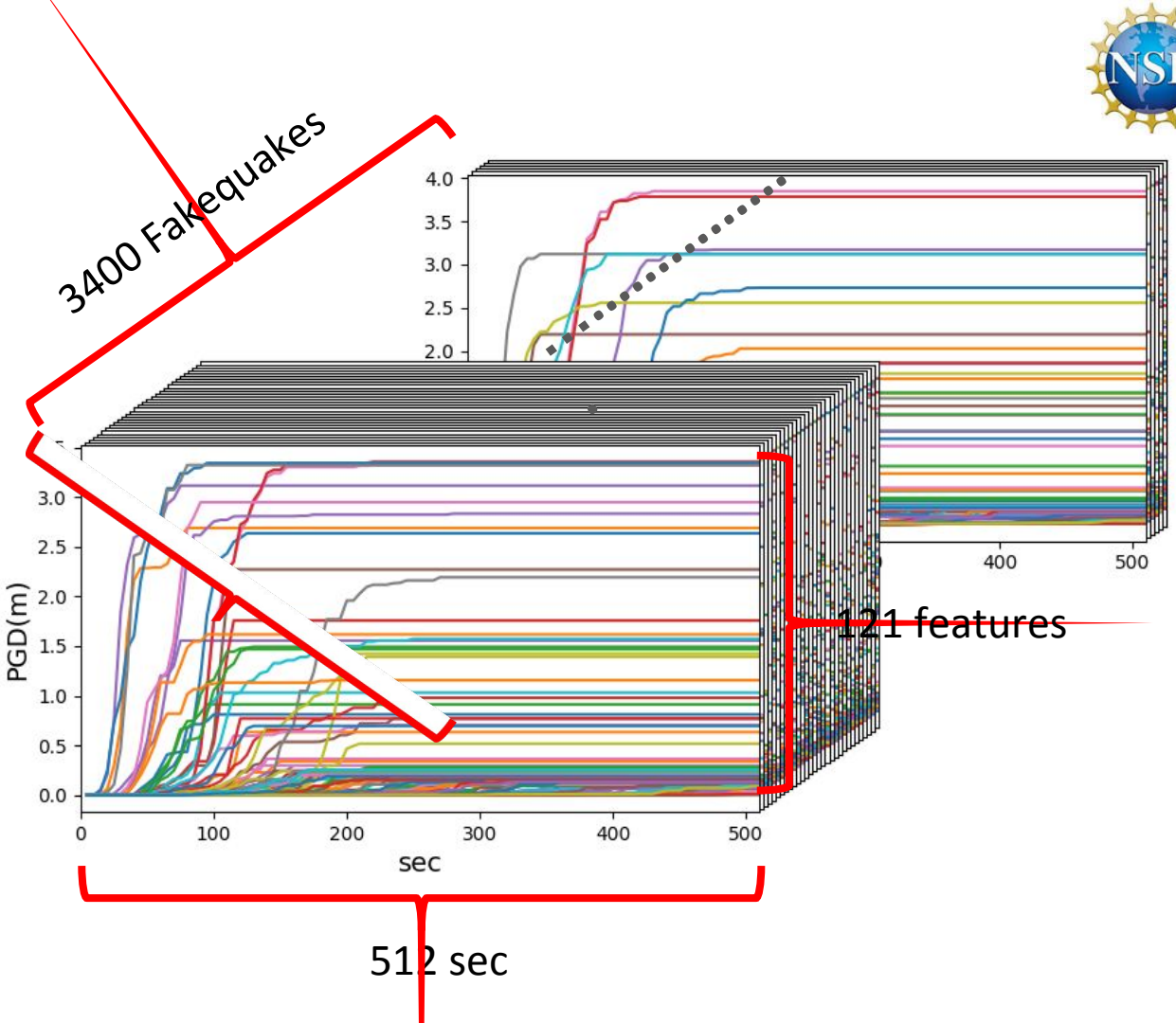
Training
5440 (80%)

Testing
1360 (20%)

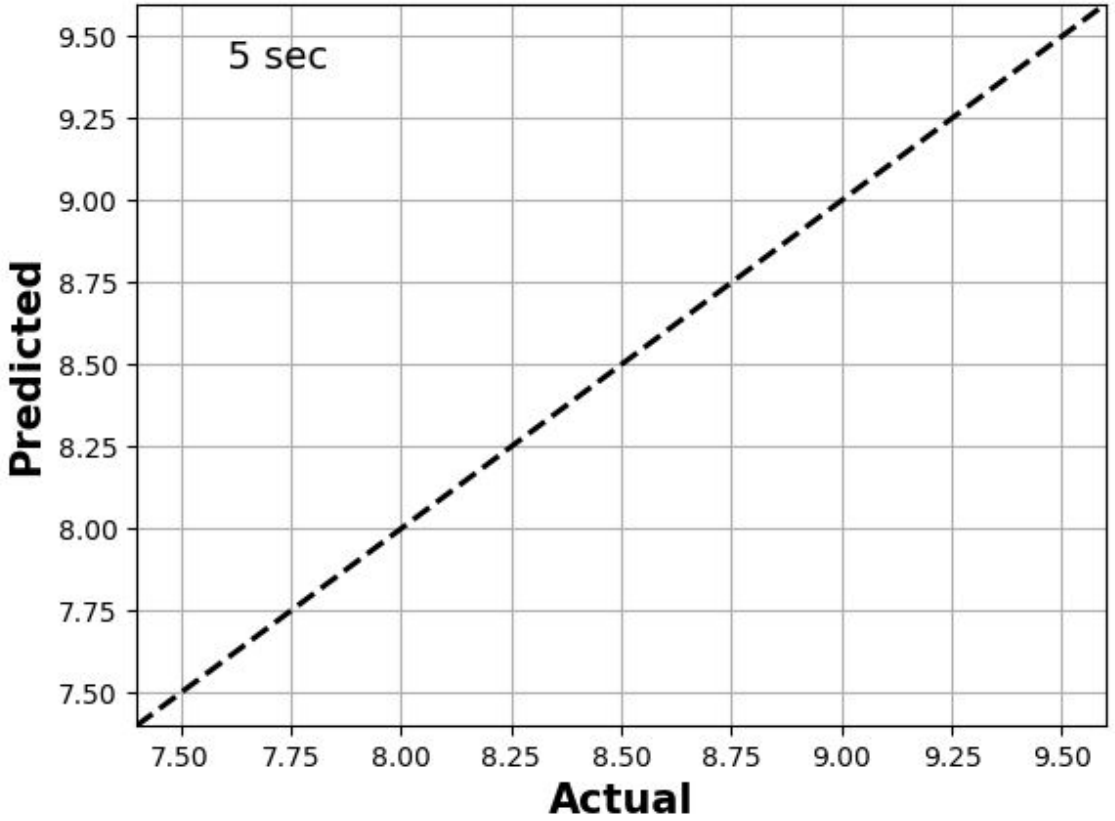
Training
4896 (90%)

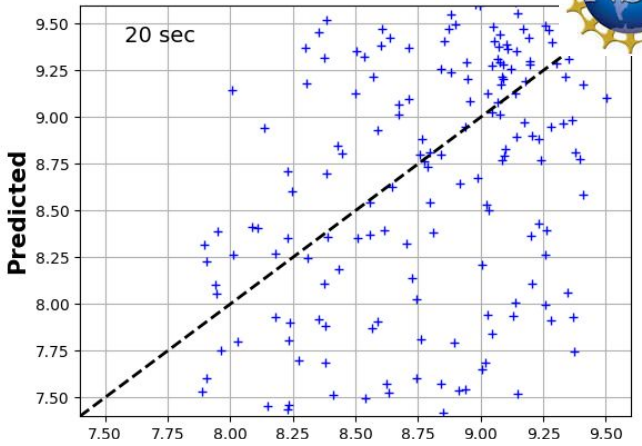
Validating
544 (10%)

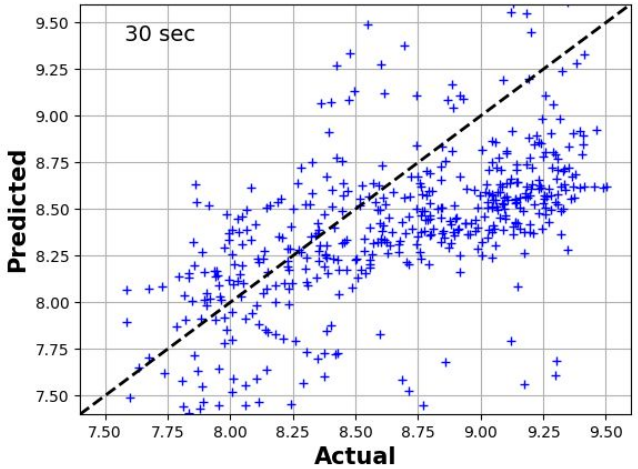
Best model!



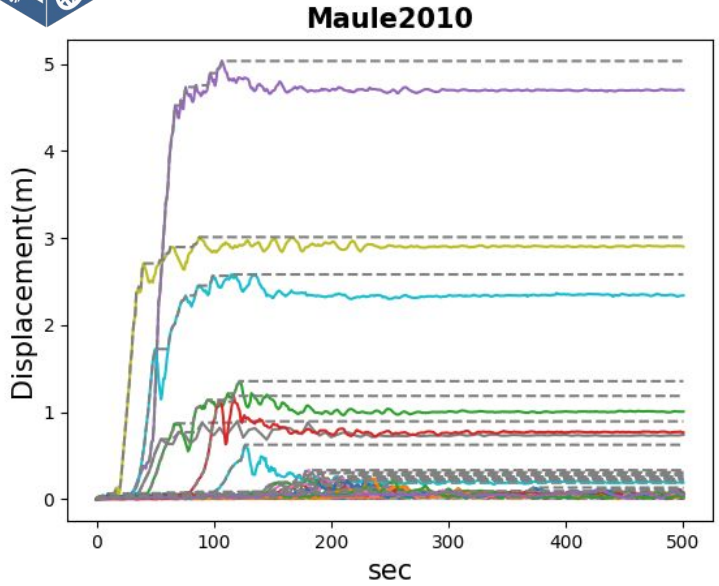
Trained by flat Mw

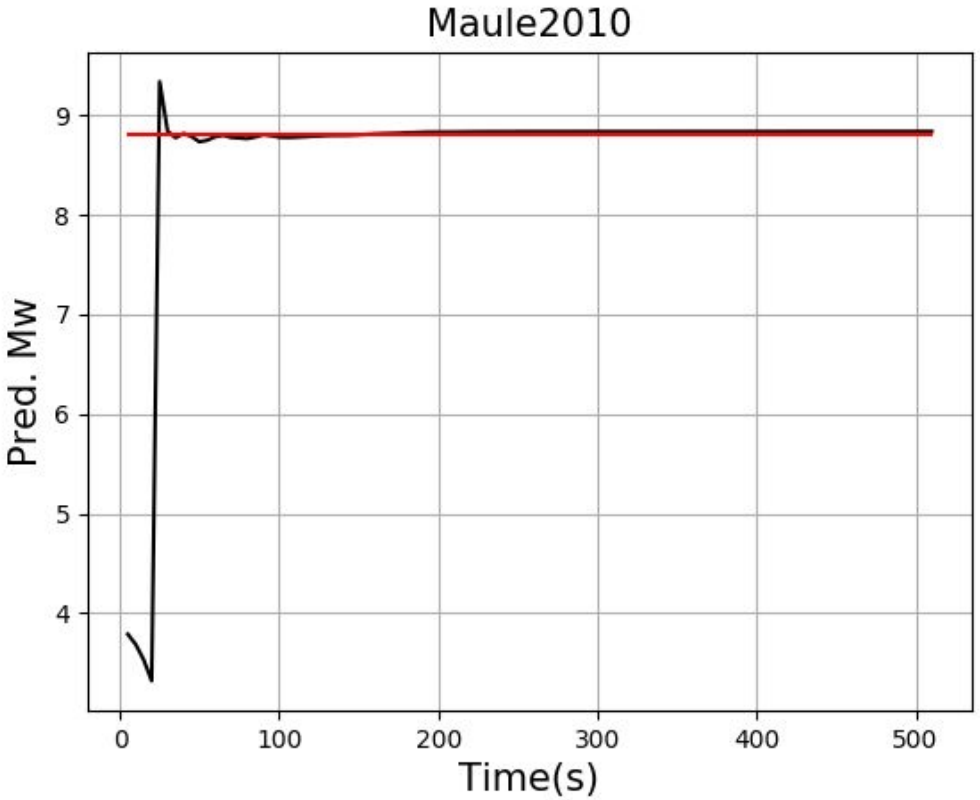






EarthCube

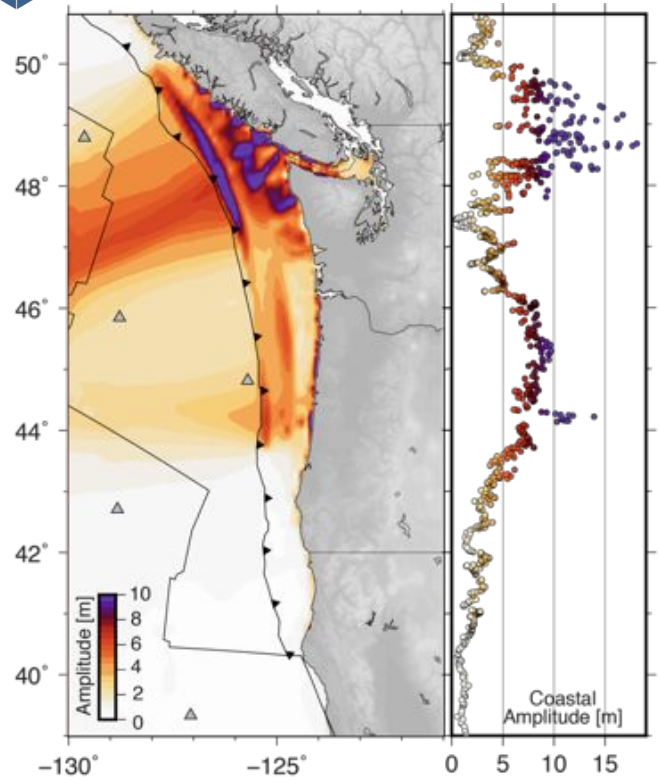


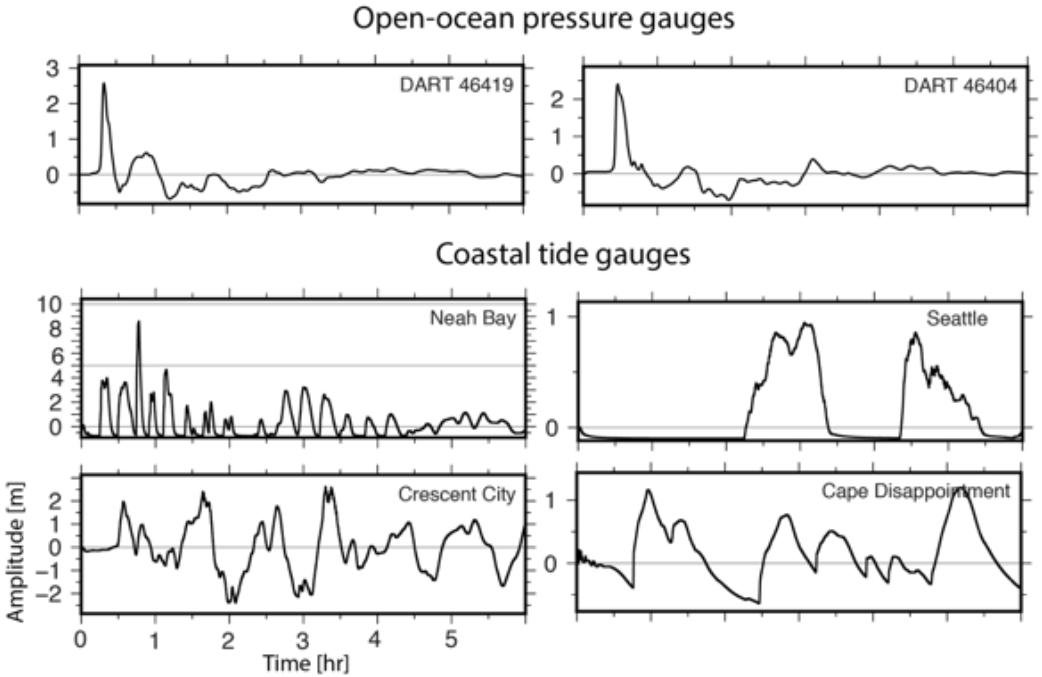


2010 M8.8 Maule earthquake

Magnitude convergence in ~25s

Offshore hazards simulations





- **Simulated Cascadia M8.7 earthquake**
- **Realistic offshore pressure and coastal tide gauge data**
- Modeled on-shore inundation
- Testing of algorithms, how well and how quickly can we forecast the induation?

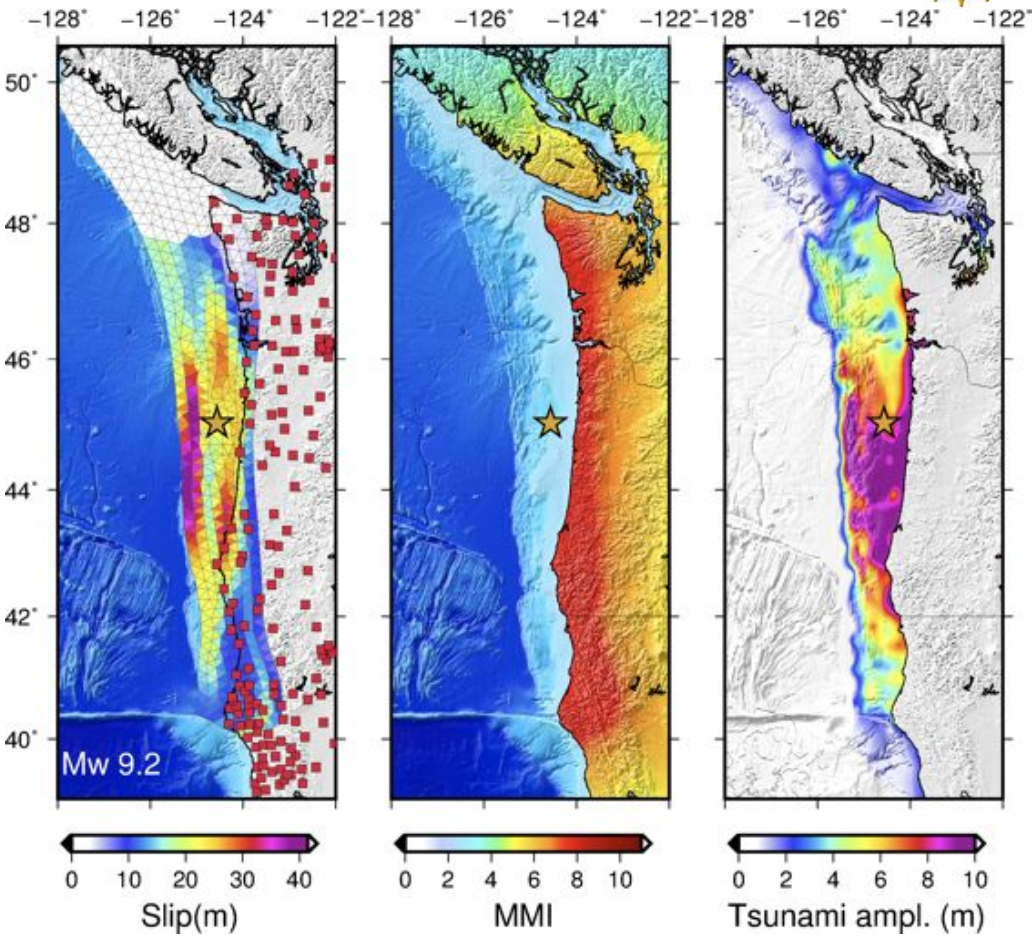
Williamson et al., in review

EarthCube

Next steps

Who cares about the earthquake?

Predict/forecast shaking intensity and tsunami heights

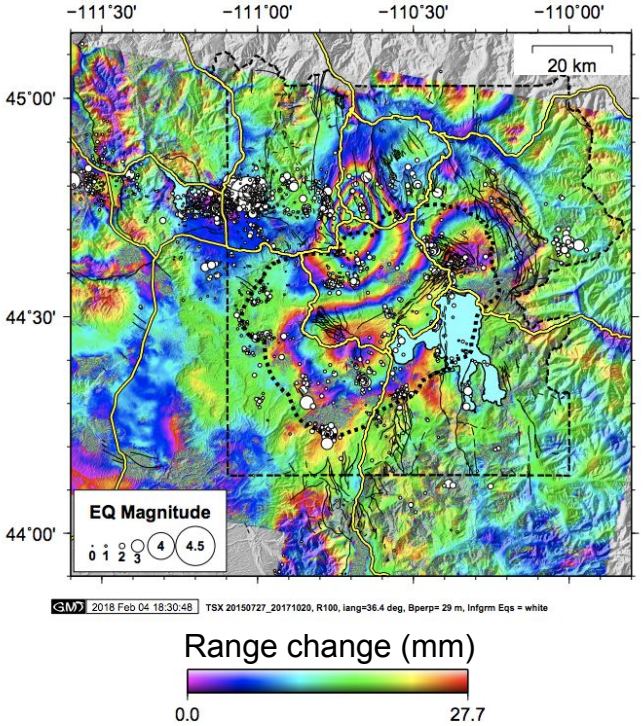


EARTHCUBE

Kristy Tiampo, University of Colorado

Intermediate-term Events. Natural catastrophes occur at a variety of spatial and temporal scales. In particular, solid earth hazards, such as large earthquakes and volcanic eruptions, often have very long interevent times and this makes it difficult to forecast their behavior. This part of the project pulls in multiple data sets to address the long- intermediate- and short-term forecasting of these types of events. Test sites include the Yellowstone magmatic center and the Hawaiian island volcanoes.

- **Data Sources**: Data types include in situ seismic, strain, GNSS, gas and thermal sensors and remote-sensing synthetic aperture radar (SAR) data.
- **Algorithms**: Repeatedly acquired SAR data from a single sensor can be used to obtain differential interferometric synthetic aperture radar (DInSAR) estimates of ground deformation and associated time series. In addition, a new technique, Multidimensional Small Baseline Subset, allows us to incorporate Interferometric SAR results from different satellites and wavelengths into one time series. Here we process DInSAR for Yellowstone and include them in the time series stream, integrated with GNSS data, providing 3-d surface motions of the caldera.
- **Broader Impacts Activities:** The analysis of SAR data on a global scale is a petabyte-scale Big Data problem that will be addressed using NSF XSEDE resources. Through the framework, researchers working on data integration of satellite radar and in situ ground deformation measurements will have easy and open access to multi-data real-time platforms on which to evaluate the latest results and to test data integration and inversion methodologies using these advanced data products, augmented and validated with additional monitoring data.



TerraSAR-X interferogram of ground motion at Yellowstone caldera (dashed black line). Each color contour represents a line of equal motion in the satellite line-of-site. Yellowstone lake is shown in blue; white circles are seismic events, July 2015-October 2017 (provided by Chuck Wicks, USGS).

TO DATE:

DInSAR time series processed for both Yellowstone and Hawaii.

Methodology designed, tested and implemented for automated time series processing of volcanoes using both GIAnT and MSBAS.

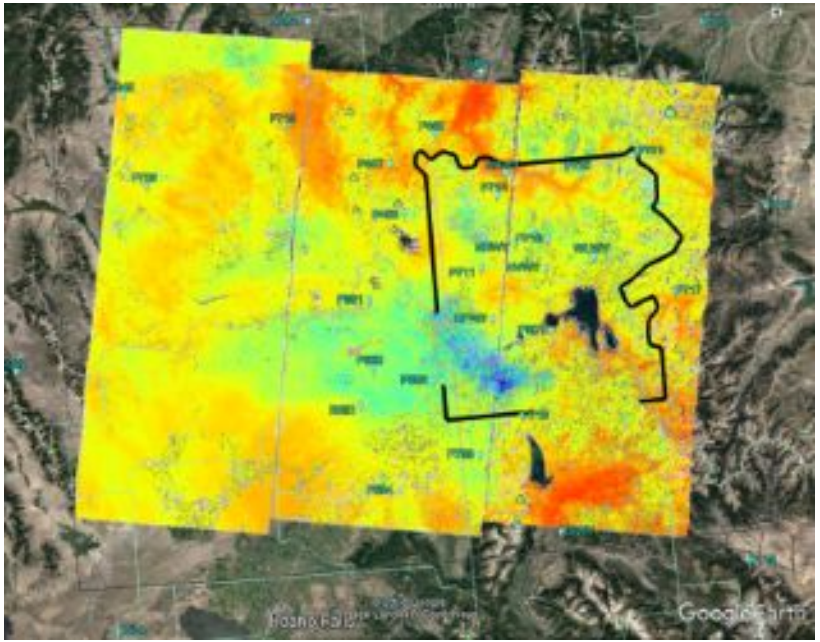
Algorithm for producing SAR data in SLC format subsampled to a regular grid, with topography removed and corrected for baselines and precise orbits prior to delivery (Zebker, 2017), facilitating rapid processing of interferograms and LOS displacement time series.

NEXT STEPS:

Integration of GPS and DInSAR time series.

Test combined DInSAR and GPS time series, as well as GPS time series and DInSAR time series as separate data streams, in machine learning for intermediate- and short-term forecasting .

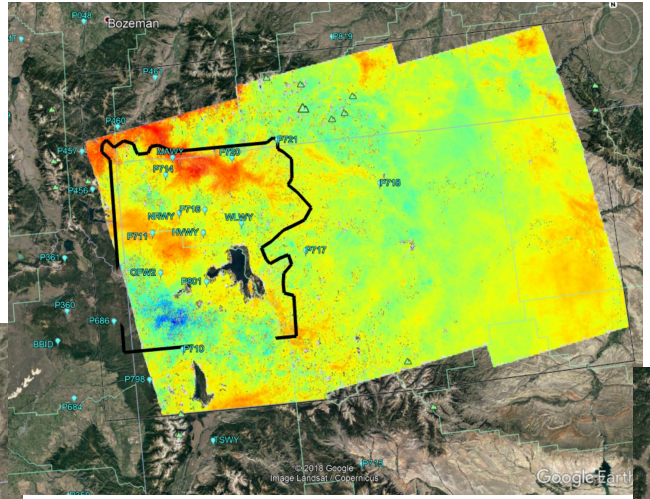
Acquisition of seismic data and integration into above methodologies.



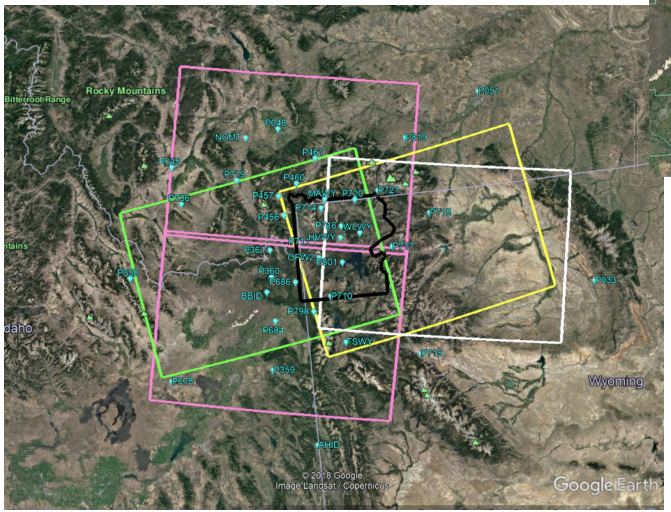
Automated processing Yellowstone caldera

Automated DInSAR processing and time series generation at Yellowstone

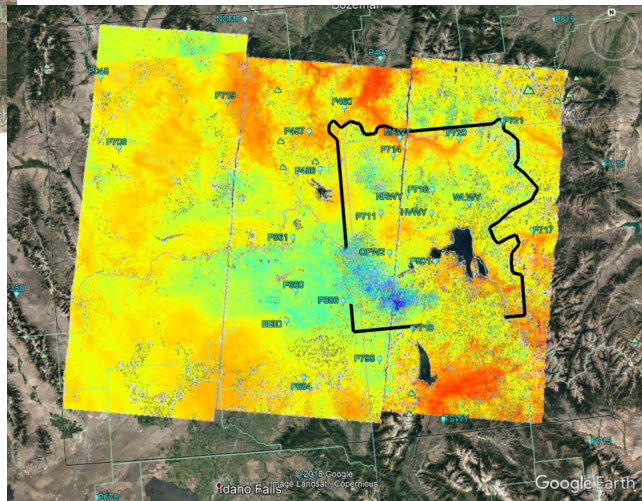
- **Available frames, below**
- **Final velocity map, ascending images (upper right)**
- **Final velocity map, descending images (lower right)**



Left: LOS velocity map, ascending track 49, Frame 142, February 2016 to December 2018.
Below: LOS velocity map, descending track 100, frame 146, February 2017 to January 2019.



Left: Yellowstone park outlined in black; separate frames, colored squares (2 ascending, 3 descending).

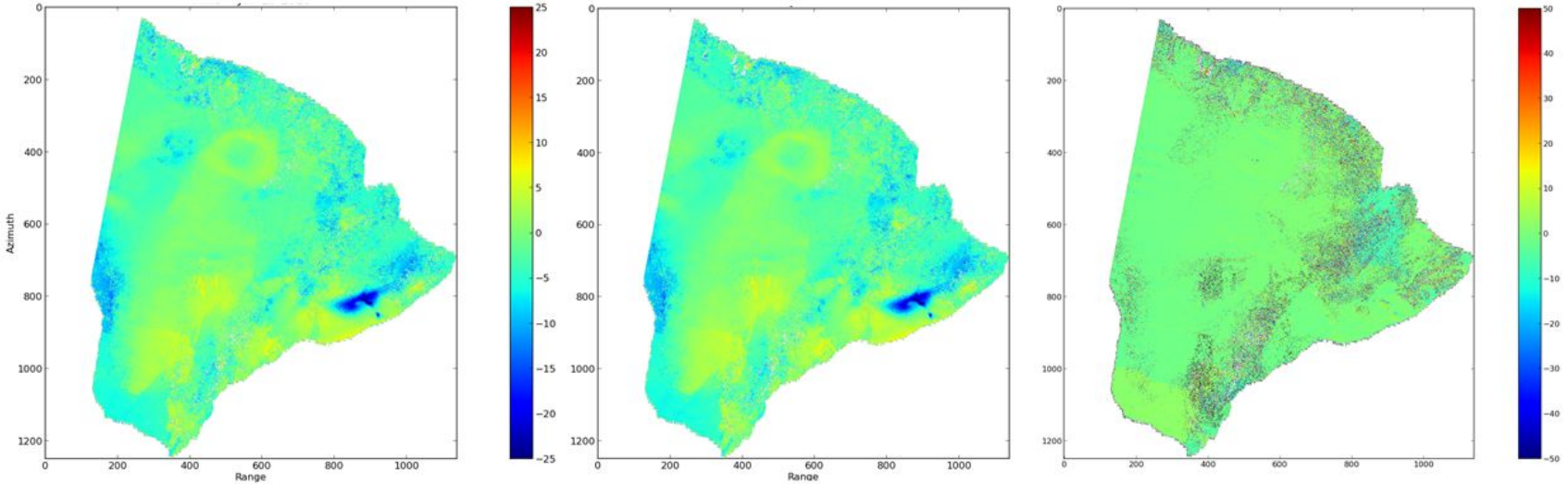


Automated processing Yellowstone and Hawaii

Impact of not using precise orbit corrections in real-time processing, automated time series generation

- **Simulated processing using precise orbits for processing older images (timesteps 1-39) and real-time orbits for the last six timesteps**

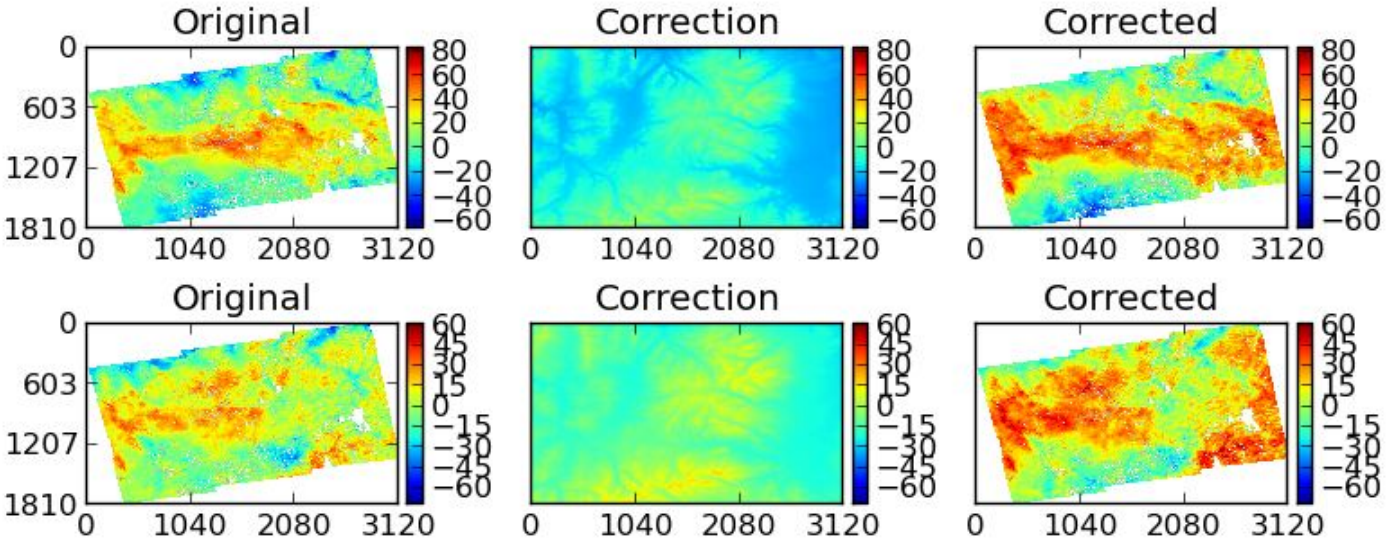
Hawaii time series processing, October 2017 through June 2018. Left: Results using all precise orbits. Center: Results using 39 images with precise orbits, 6 real-time orbits. Right: Difference between processing using all precise orbits (left) and a mixture (center). Note change in scales. LOS change in cm/year.



Automated processing Yellowstone and Hawaii

Assess the impact of atmospheric corrections on individual DInSAR image correction at Yellowstone

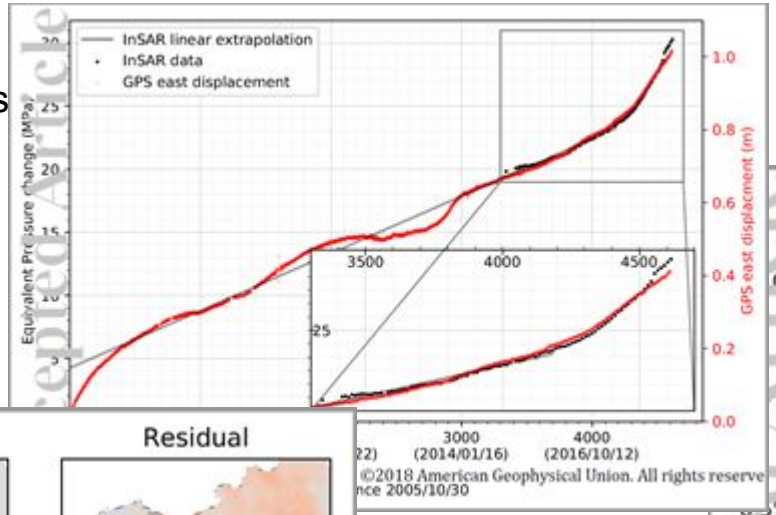
- **Here we use the Generic Atmospheric Correction Online Service for InSAR (GACOS) developed by COMET (Centre for the Observation and Modelling of Earthquakes, Volcanoes and Tectonics) for three individual time periods (ceg-research.ncl.ac.uk/v2/gacos/).**

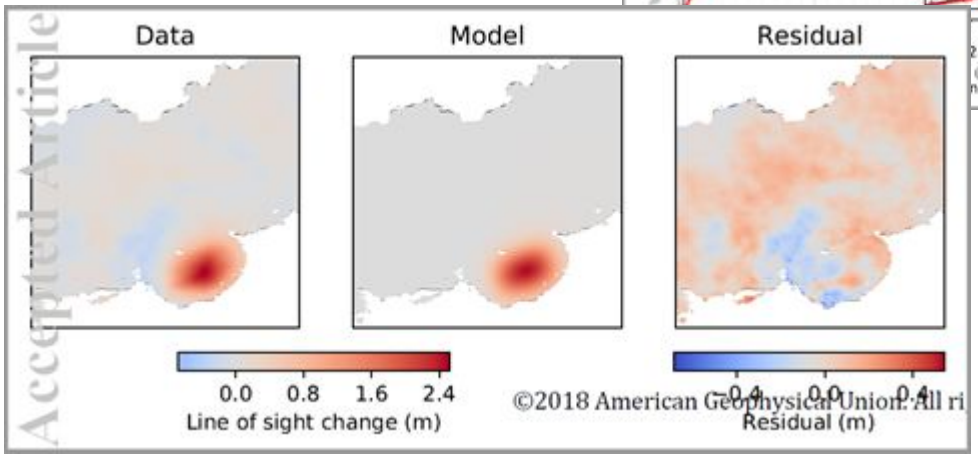


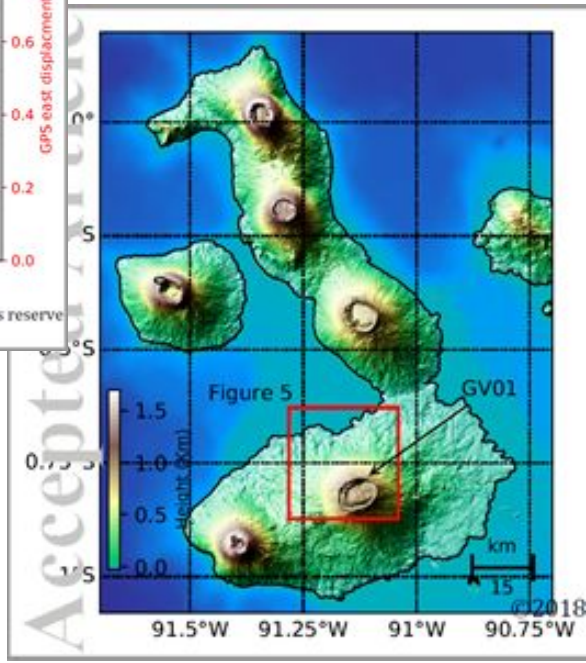
Top: Sentinel-1A DInSAR pair, 2017-12-10 to 2017-12-22. Left shows originally processed pair, the middle is the downloaded GACOS correction, and right is the corrected image. Bottom: Same as for the top, except that the time period is 2017-11-28 to 2017-12-10. Scale is LOS change in cm.

Next Step: Machine learning

Model and time series produced from ICA/machine learning analysis of the 2018 eruption of Sierra Negra (Gaddes et al., 2019).





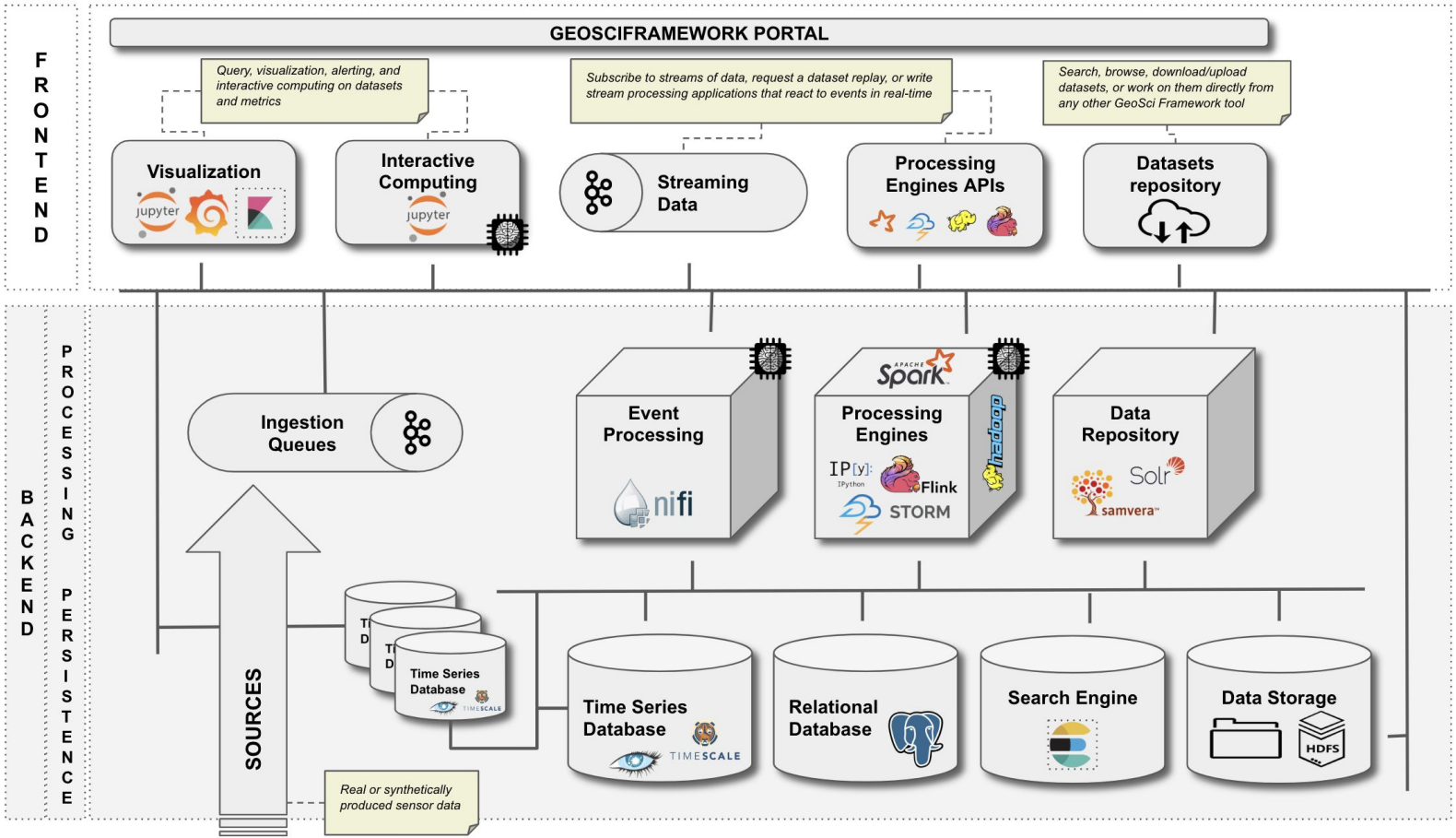


Framework Architecture Guidelines

- Based on independent, replaceable, modular components.
- Don’t blindly commit to a technology stack, don’t get locked in.
- Highly Available and Scalable.
    - Automatic Fail-Over and Self-Healing mechanisms.
    - Scalable by design and Auto-Scaling capabilities.
- Must adopt DevOps and Monitoring practices to help manage complexity and reduce overheads.

EARTHCUBE

NSF



EarthCube

NSF

Infrastructure, an example

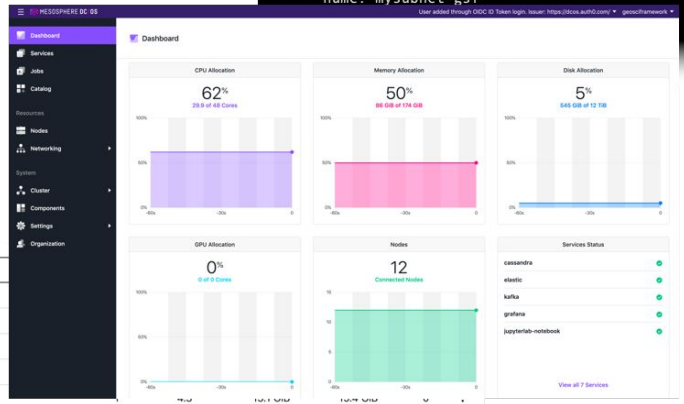
```
heat_template_version: rocky

description: >
  dcos cluster hosts

resources:

  mynetwork-gsf:
    type: OS::Neutron::Net
    properties:
      name: mynetwork-gsf

  mysubnet-gsf:
    type: OS::Neutron::Subnet
    properties:
      name: mysubnet-gsf
```



- Automation is a must to manage complexity, scalability, testing, and reproducibility.
- Tools like *Terraform*, *CloudFormation*, *Heat*, etc. allow to "codify" the details of an infrastructure.
- No need to go all the way with one tool, use the right tool for the job.

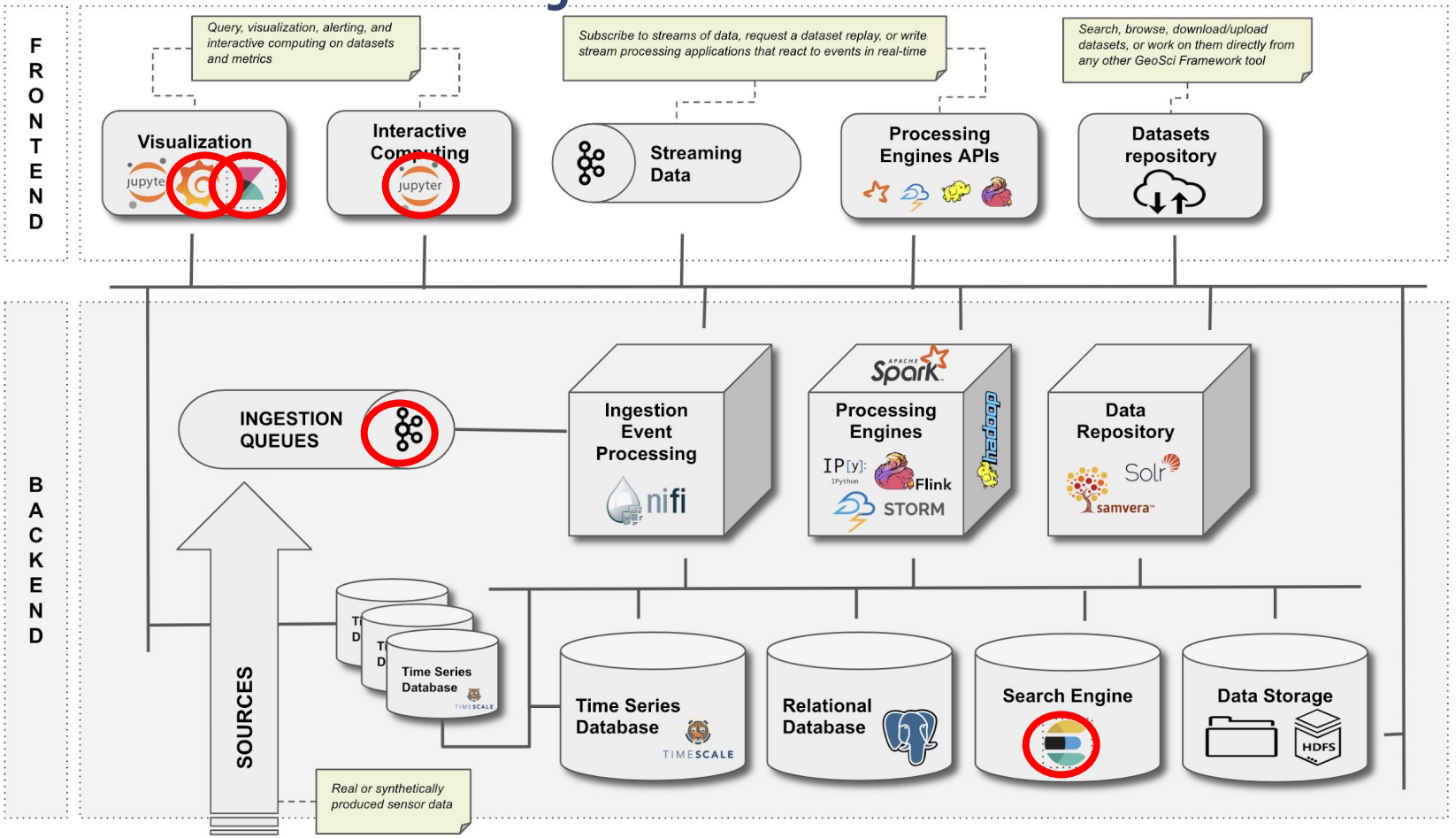
E.g. RDI² (Rutgers) testbed:

Name	Status	Version	Region					
cassandra	Running	3.11.4	N/A					
elastic	Running	6.8.1	N/A					
grafana	Running		N/A					
jupyterlab-notebook	Running		N/A					
kafka	Running	2.3.0	N/A					
kibana	Running	6.8.1	N/A	1	0.5	2 GiB	0 B	0
marathon-lb	Running		N/A	1	2	1 GiB	0 B	0

- *Spinning up the infrastructure hosts defined using a Heat template takes around **5** minutes.*
- *Deploying an Open Source DC/OS cluster using Ansible takes around **5** minutes.*
- *Takes **20** minutes to deploy 5-n Cassandra cluster, 3-n Kafka cluster, 6-n Elasticsearch cluster, Kibana, Grafana and Jupyterhub.*

In about 30 minutes we have deployed a **reproducible** full-fledged "modern" data architecture!

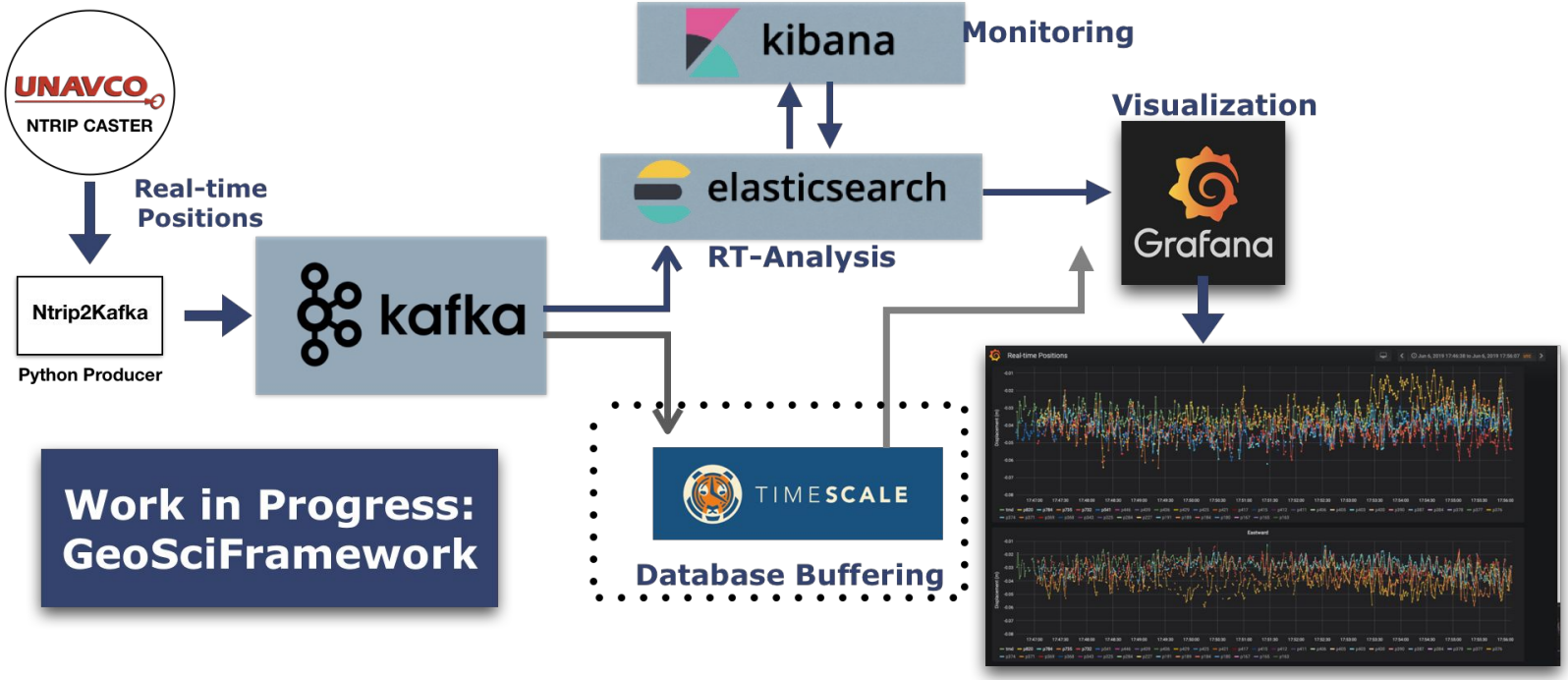
Project Architecture



EarthCube

Software component evaluation

Integration of RT data flow, analysis and archiving



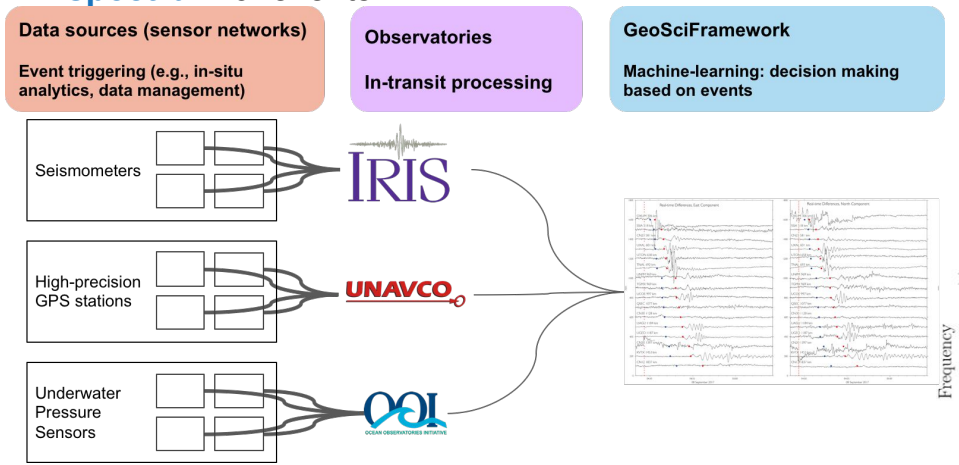
Use-case: Machine Learning@GeoSciFramework for Tsunami Early Warning

"Increase precision and delay for Tsunami warning by analyzing multiple geographically distributed data sources simultaneously"

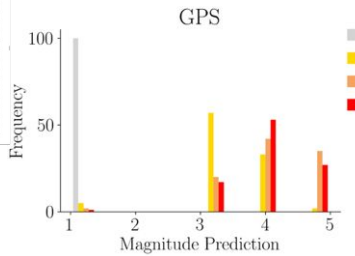
To issue Tsunami Early Warnings, earthquakes must first be characterized (magnitude, location, speed of displacement, etc.)

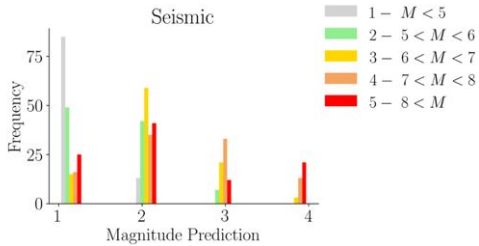
Seismometers are good for the ***smaller* earthquakes** (< 6.5), **high-precision GPS** are good for ***larger* earthquakes**

Goal: combining multiple data sources to improve the precision and delay to issue warnings by covering the **whole spectrum** of events



# 60s MTS	GPS (# Events)	Seismic (# Events)
Magnitude < 5	7,718 (170)	1,038 (349)
$5 \leq$ Magnitude < 6	3,859 (85)	None
$6 \leq$ Magnitude < 7	991 (4)	266 (4)
$7 \leq$ Magnitude < 8	432 (6)	249 (6)
Magnitude > 8	265 (4)	133 (4)
Total	13,265 (269)	1,686 (363)





EarthCube

Breakdown of tasks for Exemplar

1. **Data Producers (Kafka) for real-time data**
    a. UNAVCO
        i. Kathleen has this for streaming positions from GNSS
    b. IRIS – Kathleen Seedlink to Kafka
    c. OOI – JJ (optional for real-time from sensor not IRIS)

2. **Data Ingesters (not through Kafka)**
    a. Scott has daily .pos and UNR daily and 5 minute (ETL)
        i. Geopackage
    b. SAR scenes to HDF5
        i. Data from Sentinel on XSEDE (Scott)
        ii. Timeseries (Kristy and Brie)
            1. Create HDF5 Phase (GeoCoded)
            2. Hawaii timeseries (Scott and Kristy)
    c. Synthetic data and historic event data ingester to HDF5 (Tim visit Scott)
    d. Copy GNSS ppp files to XSEDE data directory for Diego to do noise analysis on (Kathleen, Scott, Diego)

Breakdown of tasks for Exemplar

3. **Metadata Management**
    a. Metadata source to TimescaleDB
        i. Vocabulary (Ivan)
        1. Lat, Long, Instrument, Sample rate create a GoogleDoc and suggest minimum vocabulary (see Scott's example)
        2. FAIR attribution e.g. provenance of data
        ii. GPS – Ingestdb from RT-GNSS system to TimescaleDB (Kathleen)
        iii. IRIS data - pull dataless seed to TimescaleDB (Kathleen)
        iv. SAR - mv_ssara (Scott has this)
        v. Synthetic Metadata (Tim)
4. **Data Consumers**
    a. Training synthetic timeseries machine learning (Tim and Diego)
    b. Inference / forecasting on actual event data (Tim and Diego)
    c. Volcano deformation source model (Brie and Kristy)
        i. Later add strain, GPS etc
    d. Inference / forecasting on actual event data (Brie and Kristy)

5. **Jupyter notebooks (GIT)**
    a. ETL in notebooks
        i. See GNSS positions (Scott)
        ii. Timeseries of Doppler from SAR
        iii. Data search and access for GeoServer. WFS accessor
        iv. Analysis for machine learning

6. **Scalability and other component testing**
    a. **Jetstream platform – install all components of framework (JJ and Scott)**
        i. DCOS and Components
        1. Kubernetes
        2. Kafka
        3. Jupiter Hub
            a. Tensor Flow
        4. TimescaleDB
        5. Elasticsearch/Kibana (AWS Opendistro)
        6. Grafana
        7. Geoserver
        8. Scott's collection of geodesy tools

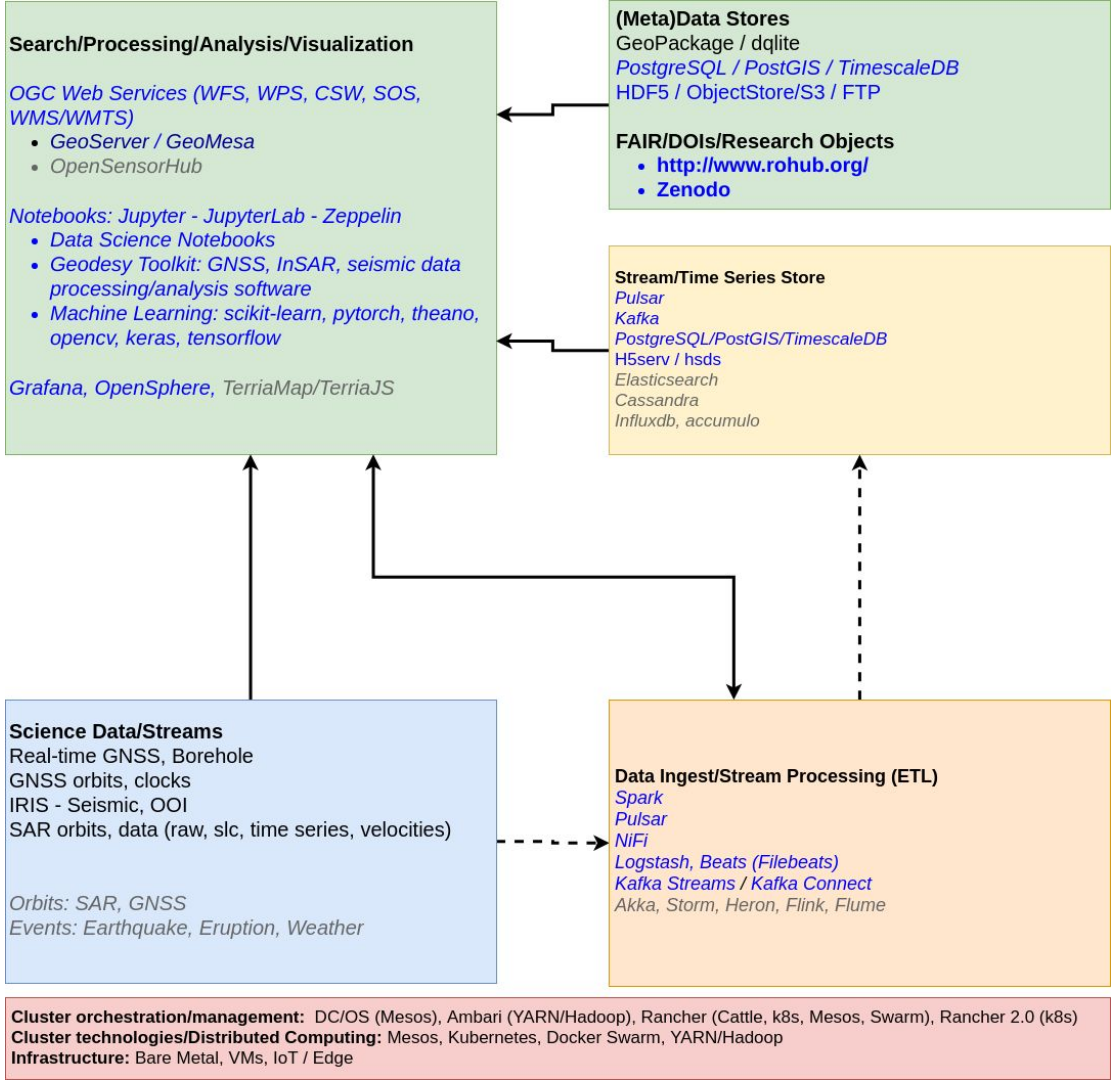
        ii. Docker Compose for local development. Launch entire stack on local machine
        1. Look at Scott's notebook development repo
            a. Tutorial for inreach for the framework
        2. Kathleen's Compose file for
    b. **Task of looking at persistency layer (file or database) scalability for timeseries**
        i. Take all OOI data and IRIS / UNAVCO timeseries and put into persistancy layer and test response, scalability
        ii. Autoscaling

EarthCube

Next Steps

- Settle on phase 1 system architecture and software stack
- Download and automatically process SAR data on XSEDE
- Finish event modeling software
- Finish initial data integration
- Publish Jupyter Notebooks
- Complete Summer “Sprint”

EARTHCUBE



NSF

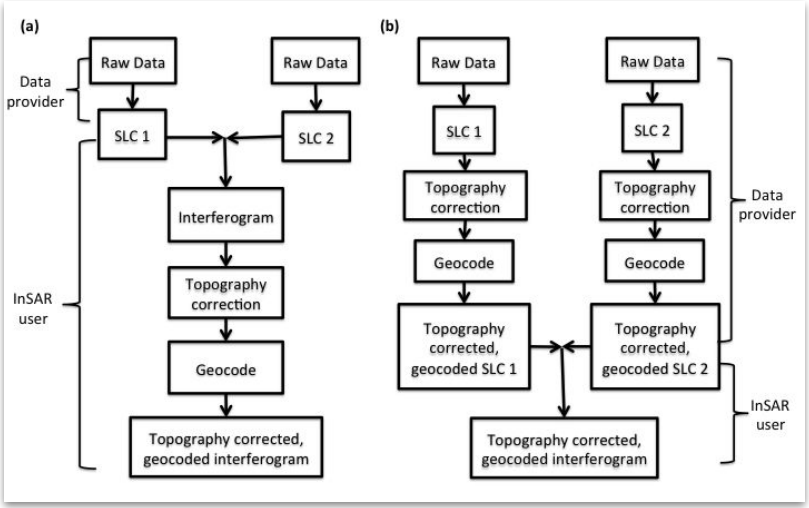
EarthCube

3 minute update:
ML and earthquake characterization

Diego Melgar, Jiun-ting Lin, Dara Goldberg
University of Oregon
Earth Sciences

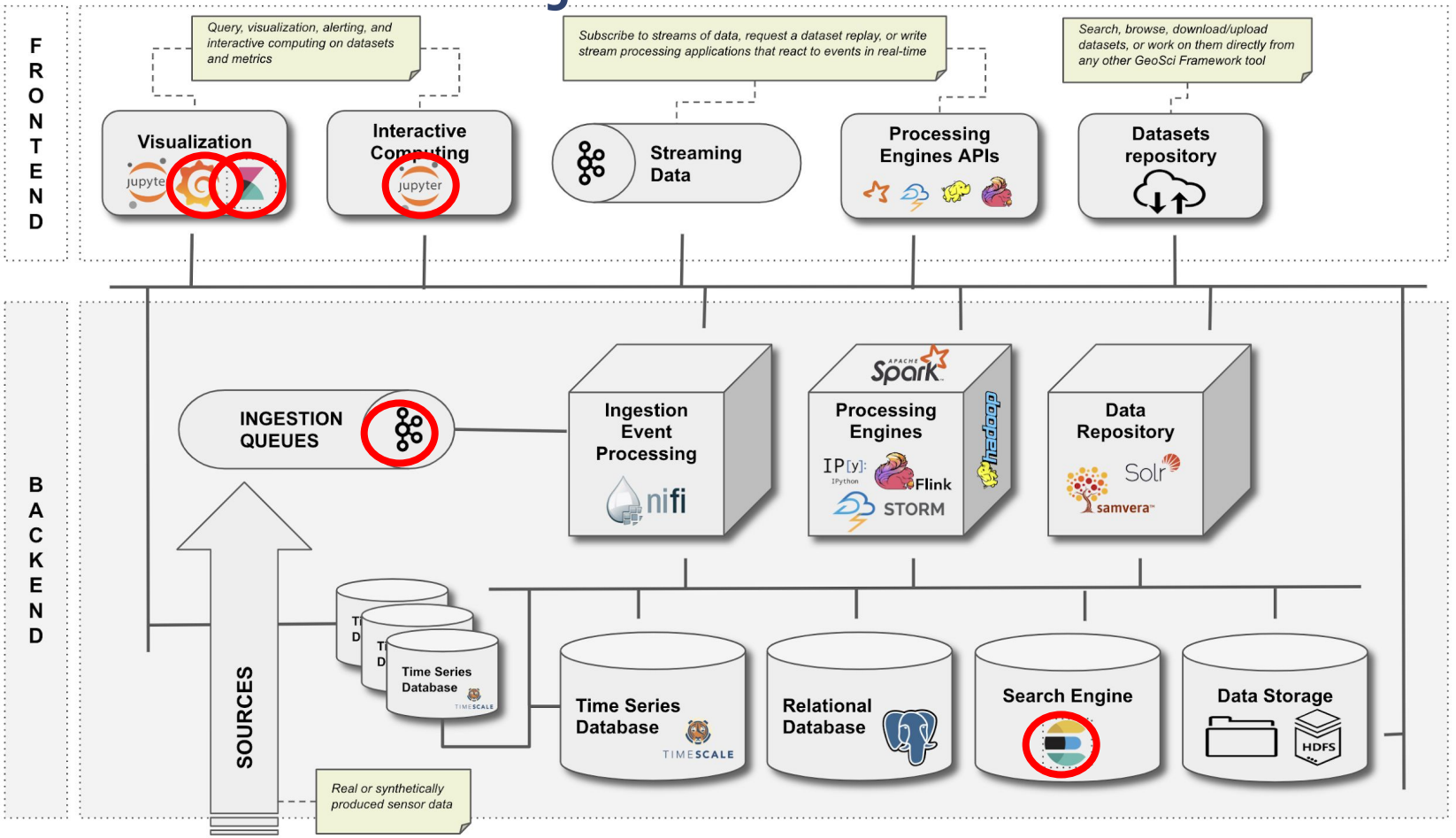
Update and Next Steps: InSAR Processing

- Finished time series for Hawaii and Yellowstone with Sentinel and have them set up to automatically update
- Applying new method of InSAR processing: compiling new scripts to work within GMT5SAR
- Accessing GPS data so we can start to implement machine learning on the time series, both separately and as an integrated product
- We also have started to source seismic data for Hawaii (IRISS and ANSS), to start to generate time series for machine learning inputs too



(a) Traditional InSAR processing workflow, (b) The proposed InSAR processing workflow. (Zheng, Y., & Zebker, H., 2017)

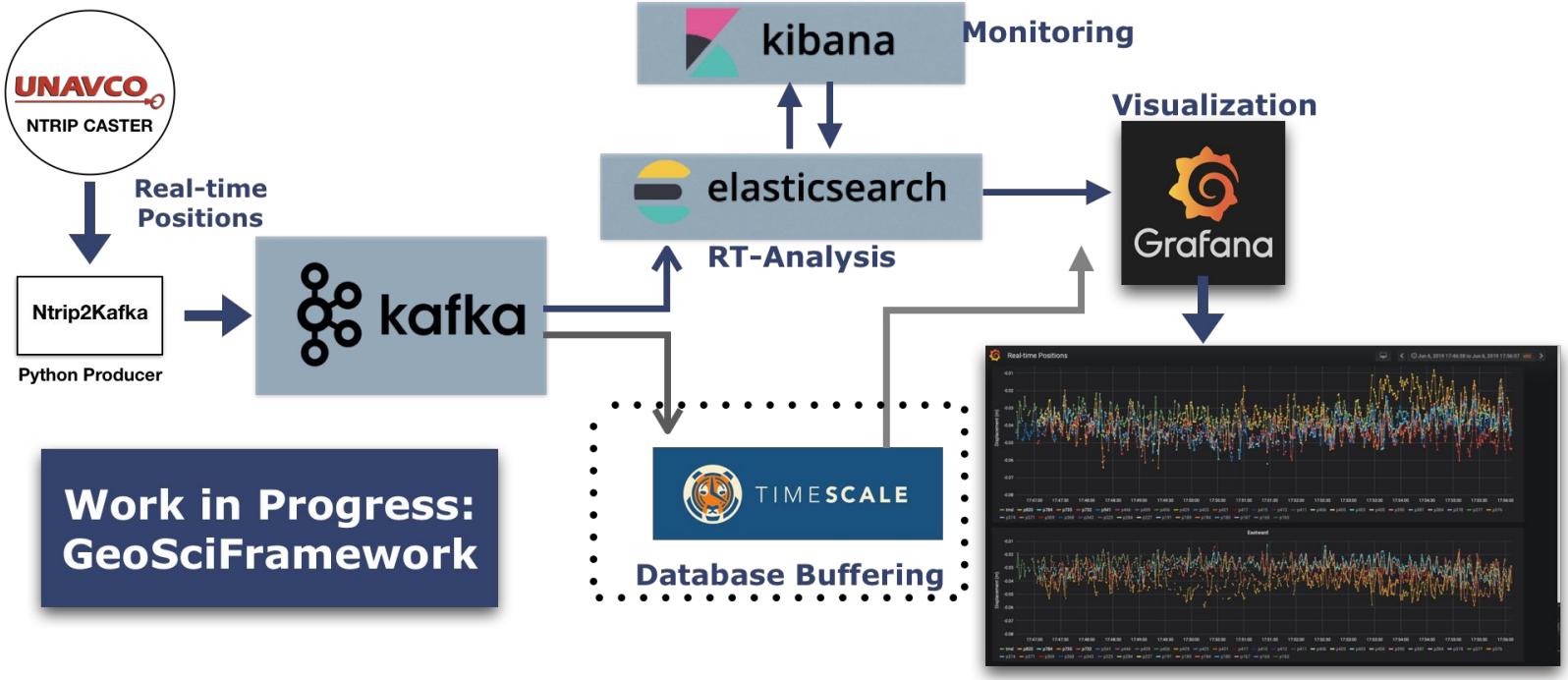
Project Architecture



EarthCube

Software component evaluation

Integration of RT data flow, analysis and archiving



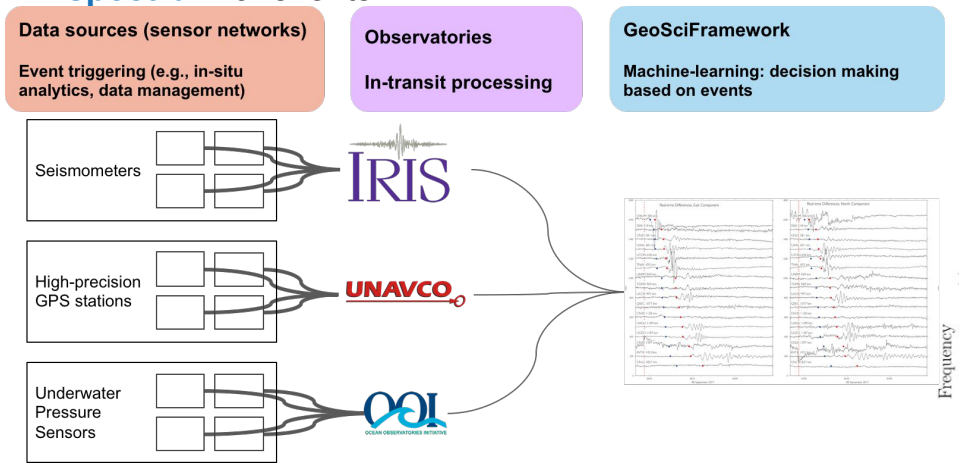
Use-case: Machine Learning@GeoSciFramework for Tsunami Early Warning

"Increase precision and delay for Tsunami warning by analyzing multiple geographically distributed data sources simultaneously"

To issue Tsunami Early Warnings, earthquakes must first be characterized (magnitude, location, speed of displacement, etc.)

Seismometers are good for the ***smaller* earthquakes** (< 6.5), **high-precision GPS** are good for ***larger* earthquakes**

Goal: combining multiple data sources to improve the precision and delay to issue warnings by covering the **whole spectrum** of events



# 60s MTS	GPS (# Events)	Seismic (# Events)
Magnitude < 5	7,718 (170)	1,038 (349)
$5 \leq$ Magnitude < 6	3,859 (85)	None
$6 \leq$ Magnitude < 7	991 (4)	266 (4)
$7 \leq$ Magnitude < 8	432 (6)	249 (6)
Magnitude > 8	265 (4)	133 (4)
Total	13,265 (269)	1,686 (363)