Jet Propulsion Laboratory
California Institute of Technology

Advancing Data Science Technology Through Open Source

Thomas Huang
thomas.huang@jpl.nasa.gov

Group Supervisor - Computer Science for Data-Intensive Applications
Strategic Lead - Interactive Data Analytics

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.

Jet Propulsion Laboratory
California Institute of Technology

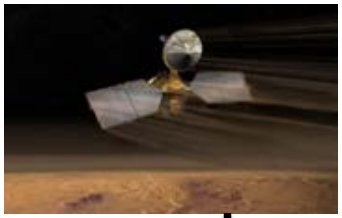
JPL is part of NASA and Caltech

- Federally-funded (NASA-owned) Research and Development Center (FFRDC)
- University Operated (Caltech)

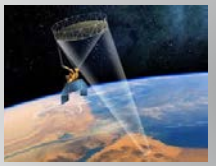
Data Lifecycle Model for NASA Space Missions

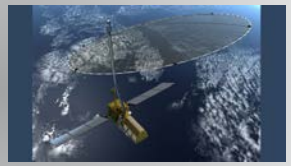
From Onboard Computing to Scalable Data Analytics

Emerging Solutions
- *Onboard Data Analytics*
- *Onboard Data Prioritization*
- *Flight Computing*



Observational Platforms and Flight Computing



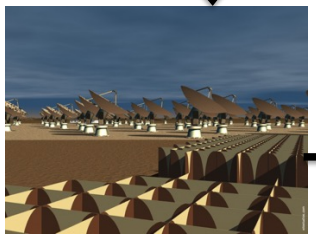


SMAP (Today): 485 GB/day **NI-SAR (2020): 86 TB/day**

(1) Too much data, too fast; cannot transport data efficiently enough to store

Massive Data Archives and Big Data Analytics

Emerging Solutions
- *Intelligent Ground Stations*
- *Agile MOS-GDS*



Emerging Solutions
- *Data Discovery from Archives*
- *Distributed Data Analytics*
- *Advanced Data Science Methods*
- *Scalable Computation and Storage*

(2) Data collection capacity at the instrument continually outstrips data transport (downlink) capacity

Ground-based Mission Systems

(3) Data distributed in massive archives; many different types of measurements and observations

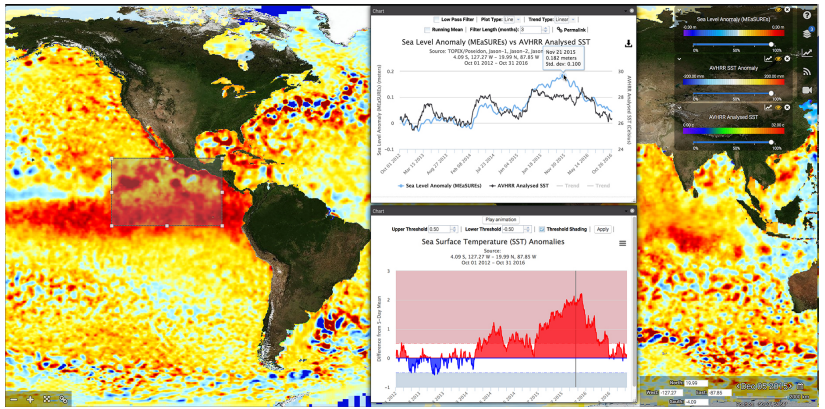
JPL's National Space Technology Applications Program Office

- Develop a sustaining business base through expanded relations with non-NASA sponsors
- Focus on leveraging NASA technology to solve problems of national significance for other agencies
- Out mission is to apply JPL's unique skills and assets to solve problems of national importance in a manner that is synergistic with our NASA mission
- Collaborations with elements of the three national space sectors: Military, Civil, and Commercial
- Development of partnerships that expand and enhance the JPL/NASA technology base

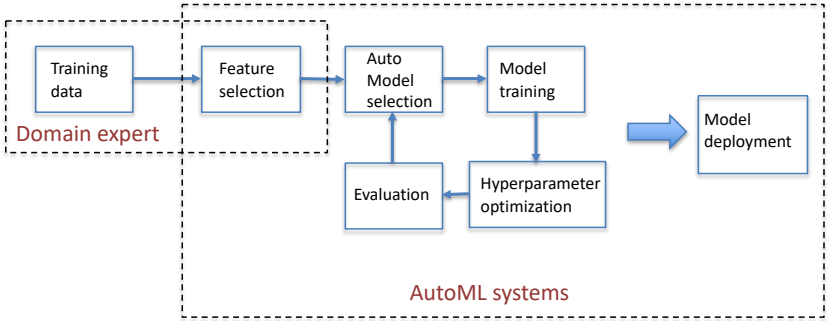
https://nsta.jpl.nasa.gov

Opportunities Enabled by Data Science and Open Source

1. Support scalability to capture and analyze NASA observational data
2. Apply data-driven approaches across the entire data lifecycle
3. Increase access, integration and use of highly distributed archival data
4. Increased data science services for on-demand, interactive visualization and analytics
5. Making software and tools freely available to empower the research community

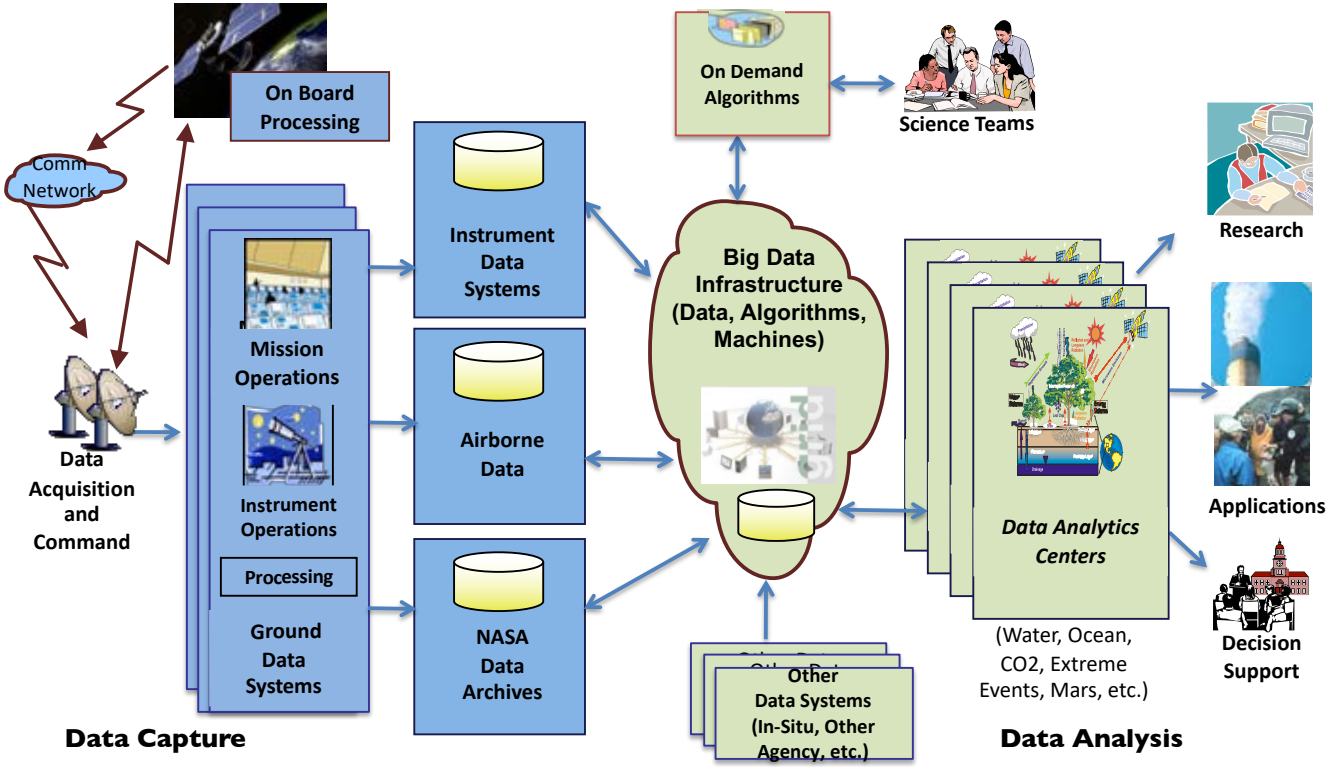


NASA AIST: OceanWorks - Anomaly Detection Solution



Automate Machine Learning

From Data Capturing to Applied Sciences



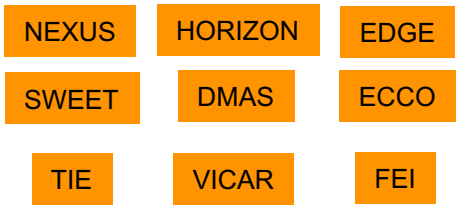
Jet Propulsion Laboratory
California Institute of Technology

JPL's Office of Technology Transfer

https://ott.jpl.nasa.gov

- Promote and facilitate the transfer of useful technologies to the commercial sector
- Focus on JPL intellectual property management and licensing, and commercialization support to apply JPL space technologies to non-NASA needs
- Goal is to infuse JPL-invented innovations into the private sector as quickly as possible so that taxpayers may benefit from NASA and JPL research
- Responsible for new technology reporting, NASA Tech Briefs, software release, patens, licensing, and commercialization

Over 100 JPL developed software have been approved for open source in recent years



Open Source Software Policy at JPL

- Embrace the open source paradigm for developing and disseminating software
- Licenses: Apache 2.0, Eclipse Public License (EPL) and BSD etc.
- Benefits of an open source policy
  - Facilitates exchange of ideas in research thereby foster exploration and experimentation
  - Facilitates productivity and efficiency in a collaborative development setting based on ease of sharing software artifacts
  - Facilitates ease of interaction and timeliness of software development support in addition to traditional vendors
  - Better positions the institution to attract software engineering fresh-outs who are fully engaged in the open source development paradigm and its application as a modern research and development practice
  - Increases institutional productivity when consuming Open Source software for appropriate uses and applications

NASA's Software Catalog

**** APPROVED ****

The software known as NEXUS: Deep Data Platform (NTR-50157) has been approved for release as open source. You are authorized to upload the software to open source repository when you are ready to do so.

**** APPROVED ****

The HORIZON code is approved for release as open source. You may upload the code at anytime to an open source repository.

**** APPROVED ****

The DMAS code is approved for release as open source. You may upload the code at anytime to an open source repository.

**** APPROVED ****

The EDGE code is approved for release as open source. You may upload the code at anytime to an open source repository.

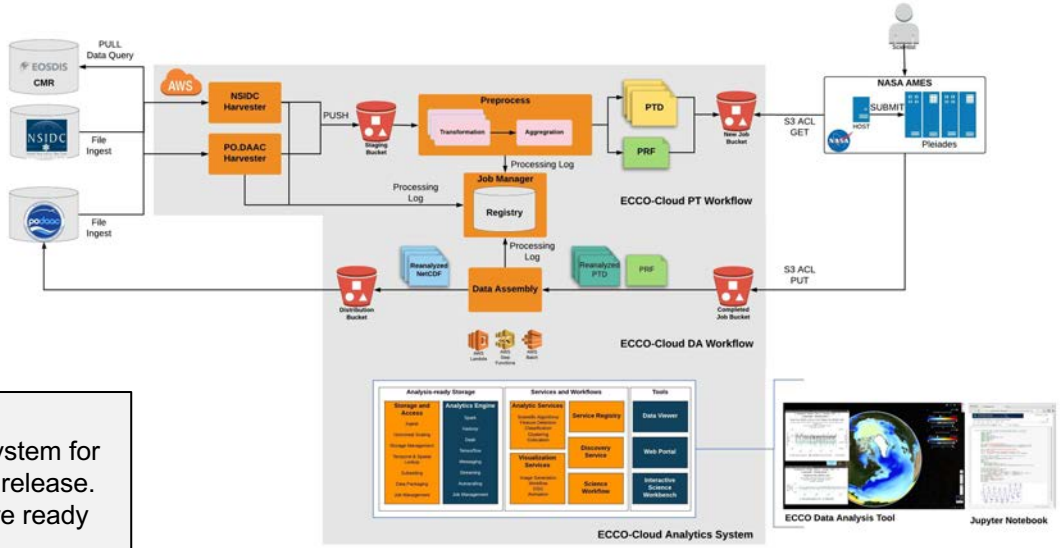
**** APPROVED ****

The TIE code is approved for release as open source. You may upload the code at anytime to an open source repository.

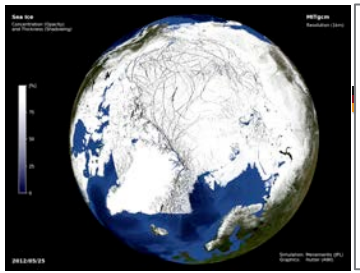
Data Access and the ECCO Ocean and Ice State Estimate

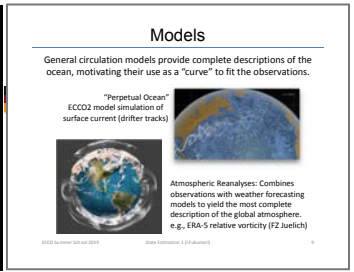
NASA ACCESS Program | PI: Patrick Heimbach; Co-Is: Ian Fenty and Thomas Huang

- **Estimating the Circulation and Climate of the Ocean** (ECCO) is a consortium endeavors to produce the best possible estimates of ocean circulation and its role in climate
- Combining state-of-the-art ocean circulation models with global ocean and sea-ice data in a physically and statistically consistent manner
- ECCO products are being used in studies on ocean variability, biological cycles, coastal physics, water cycle, ocean-cryosphere interactions, and geodesy



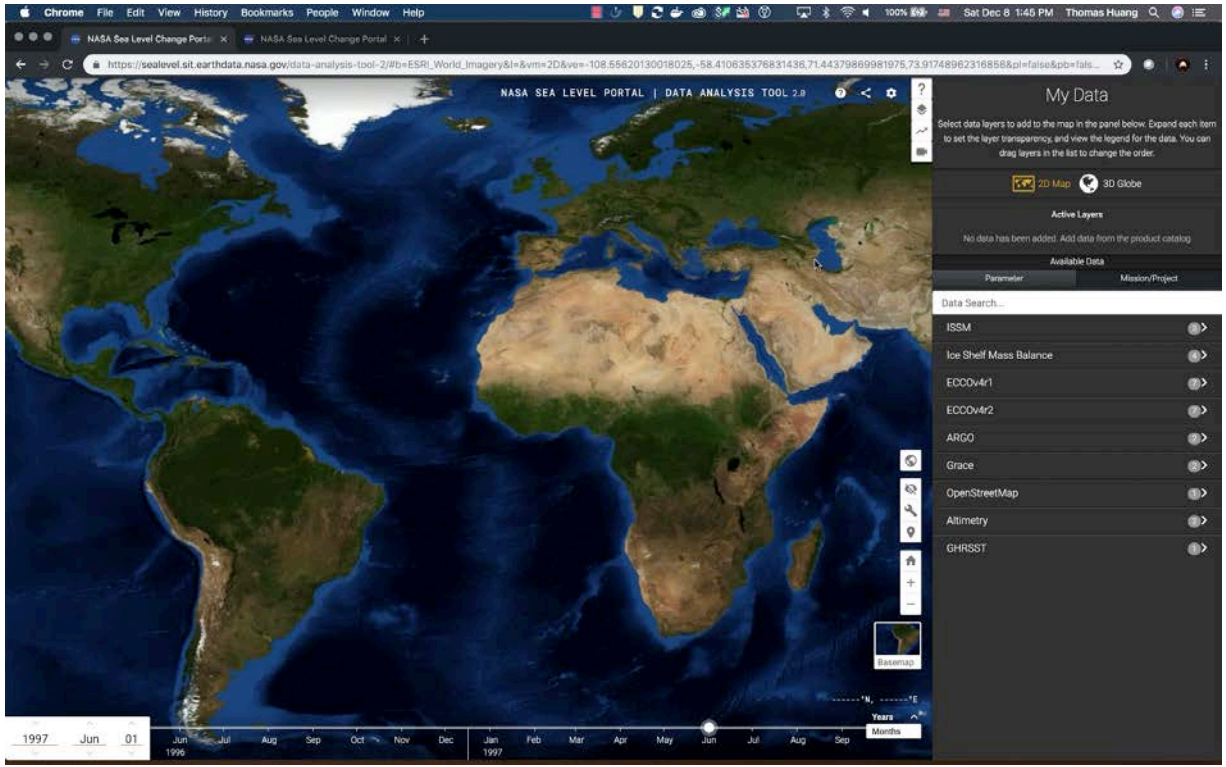
**** APPROVED ****
The software known as the Cloud-based Data Processing System for ECCO (NTR-51406) V1 has been approved for open source release. You may upload the code to a known repository when you are ready to do so.

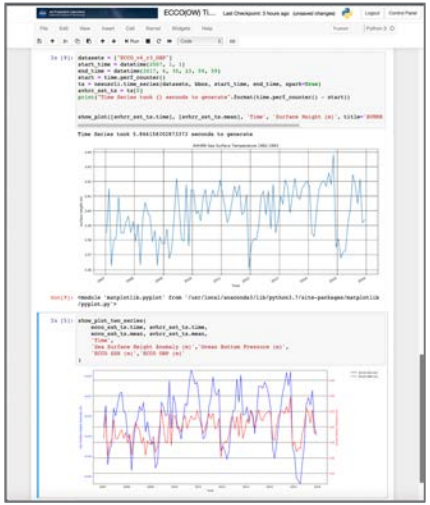


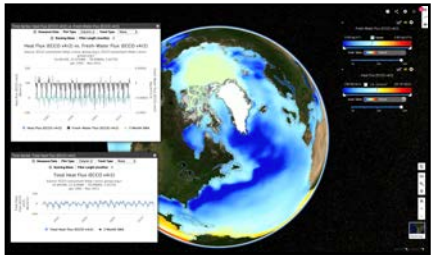


Jet Propulsion Laboratory
California Institute of Technology

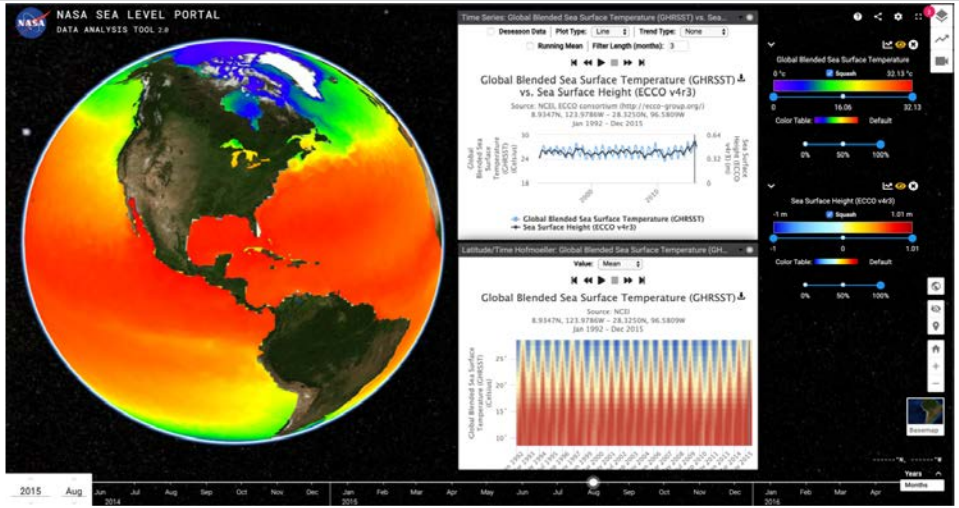
Interactive Analysis of ECCO Products

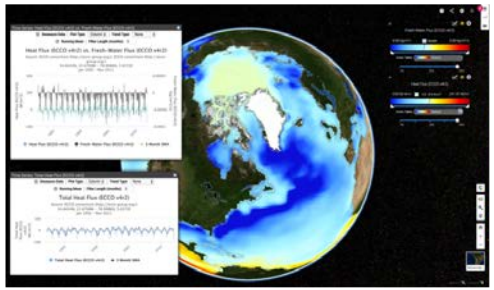


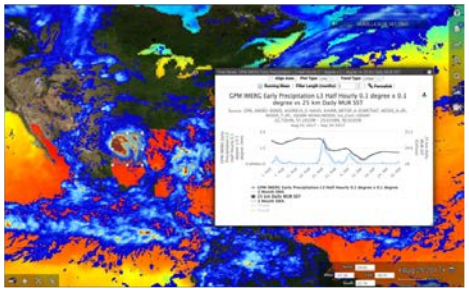




Professional Open Source Technologies

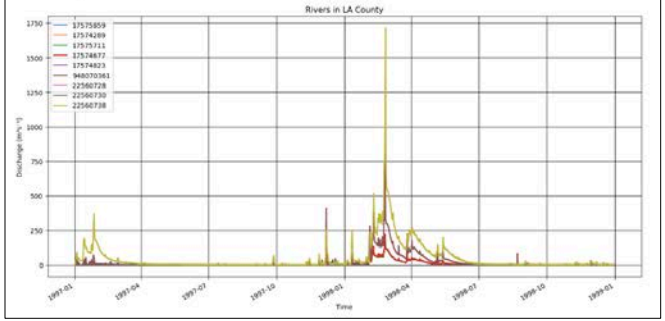






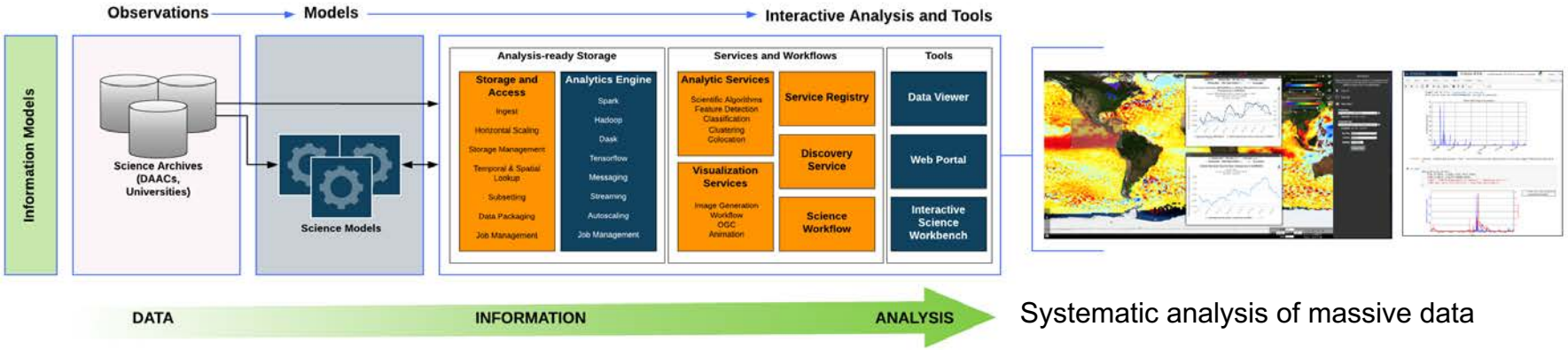
```
import requests
import json
import time
import nexuscli
from datetime import datetime

nexuscli.set_target("https://doms.jpl.nasa.gov", use_session=False)

# River IDs for the 10 largest (by max discharge rate) Rivers in LA County
#la_county_river_ids = [
#    20351643, 20357290, 20357300, 20357284, 20357292,
#    948070444, 20351637, 20357240, 20357296, 20351677]
la_county_river_ids = [17575859, 17574289, 17575711, 17574677, 17574823,
                       948070361, 22560728, 22560730, 22560738]

ds = "RAPID_WSWM"
start_time = datetime(1997, 1, 1)
end_time = datetime(1998, 12, 31, 23, 59, 59)
la_county_river_data = list()

start = time.perf_counter()
for river_id in la_county_river_ids:
    metadataFilter = "rivid_i:{}".format(river_id)
    result = nexuscli.subset(ds, None, start_time, end_time, None, metadataFilter)
    la_county_river_data.append(result)
print("Subsetting took {} seconds".format(time.perf_counter() - start))

show_plot([[point.time for point in river] for river in la_county_river_data], # x values
          [[point.variable['variable'] for point in river] for river in la_county_river_da
          'Time', # x axis label
          'Discharge (m³s⁻¹)', # y axis label
          legend=[str(r) for r in la_county_river_ids],
          title='Rivers in LA County'
         )
```

```
Target set to https://doms.jpl.nasa.gov
Subsetting took 4.413320103660226 seconds
```



NASA is Investing in Analytics Center Framework

Creating SaaS and PaaS for Science Tools and Services

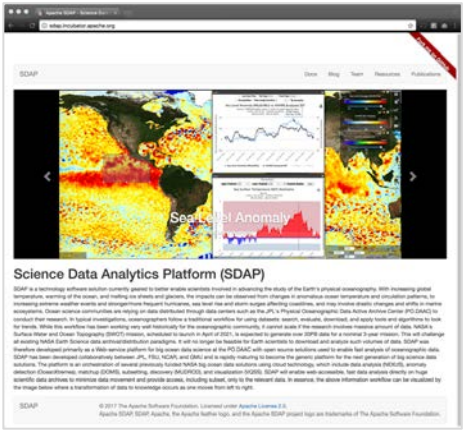


Systematic analysis of massive data

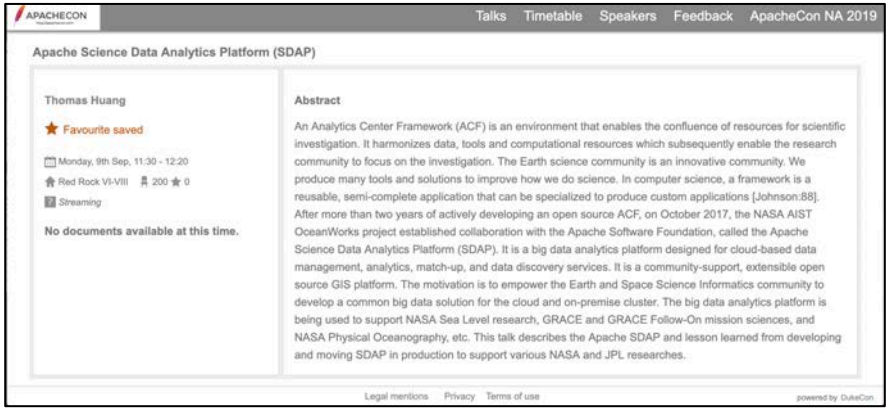
- An Analytics Center Framework (ACF) to provide an environment for conducting a science investigation
  - Enables the confluence of resources for that investigation
  - Tailored to the individual study area (physical ocean, sea level, etc.)
- Harmonizes data, tools and computational resources to permit the research community to focus on the investigation
- Scale computational and data infrastructures
- Shift towards integrated data analytics
- Algorithms for identifying and extracting interesting features and patterns

Managing Open Source

- After more than two years of active development, on October 2017 the **NASA ESOT/AIST OceanWorks** team established Apache Software Foundation and established the **Science Data Analytics Platform (SDAP)** in the **Apache Incubator**
- Technology sharing through Free and Open Source Software (FOSS)
- Why? Further technology evolution that is restricted by projects / missions
- It is more than GitHub
  - Quarterly reporting
  - Reports are open for community review by over 6000 committers
  - SDAP has a group of appointed international mentors
- **SDAP and many of its affiliated projects are now being developed in the open**
  - Support local cluster and cloud computing platform support
  - Fully containerized using Docker and Kubernetes
  - Infrastructure orchestration using Amazon CloudFormation
  - Satellite and model data analysis: time series, correlation map,
  - In situ data analysis and collocation with satellite measurements
  - Fast data subsetting
  - Upload and execute custom parallel analytic algorithms
  - Data services integration architecture
  - OpenSearch and dynamic metadata translation
  - Mining of user interaction and data to enable discovery and recommendations

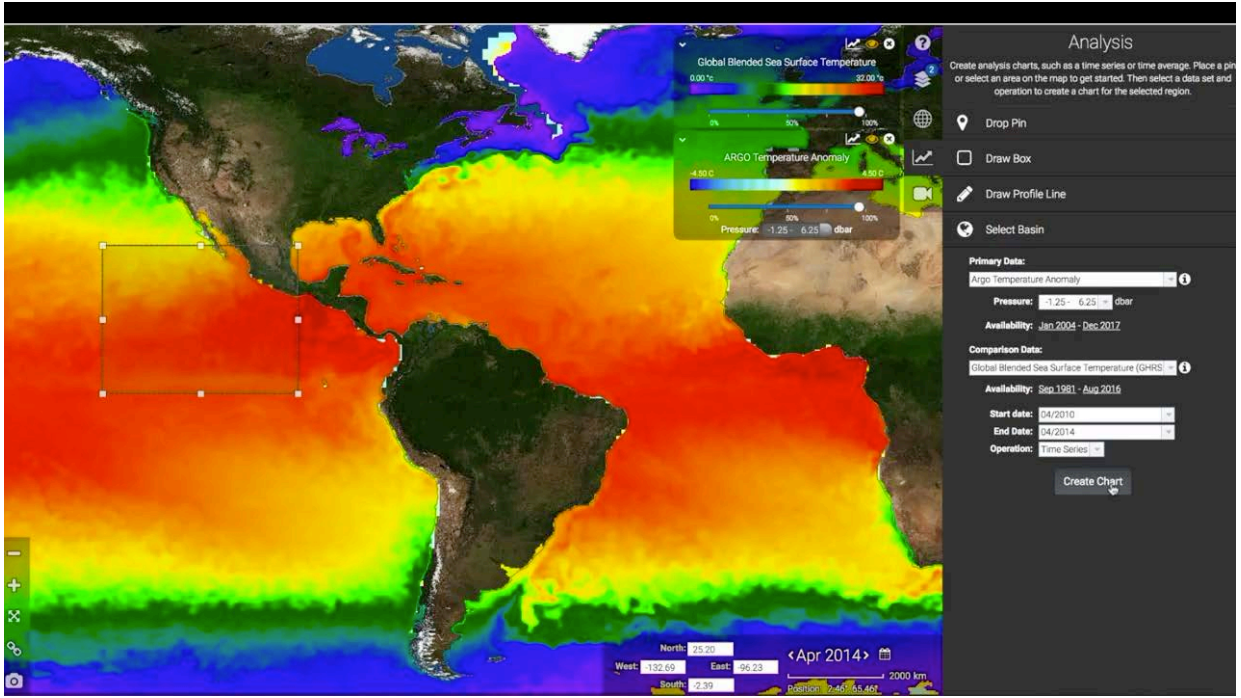


http://sdap.apache.org



The NASA Sea Level Data Analysis Tool (DAT)

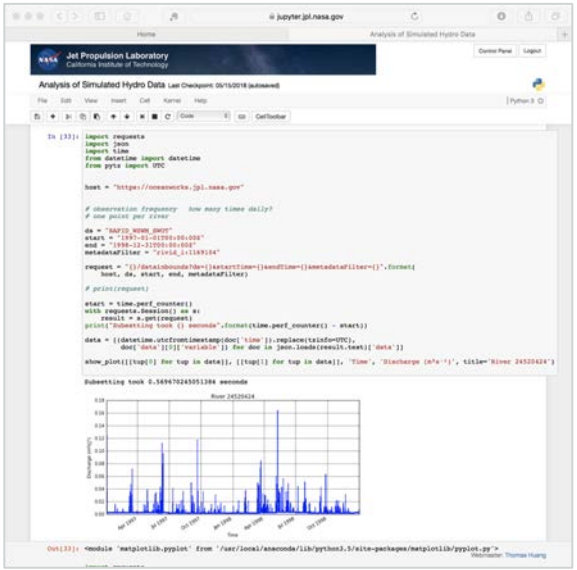
Developed using open source technologies and released as an open source solution



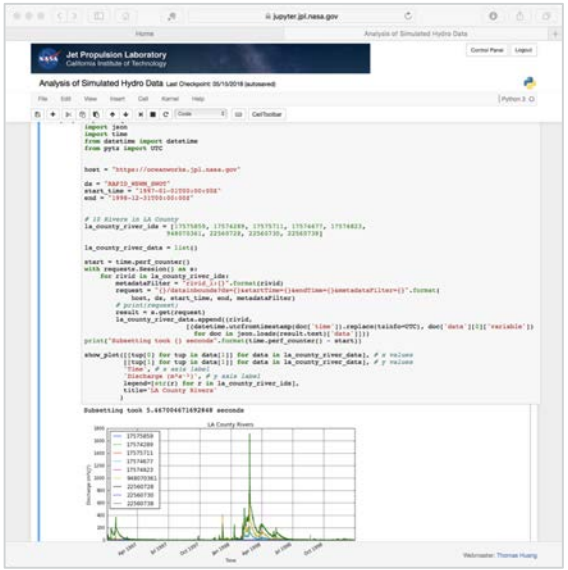
Analyze *in situ* and satellite observations

Analyze Sea Level on mobiles

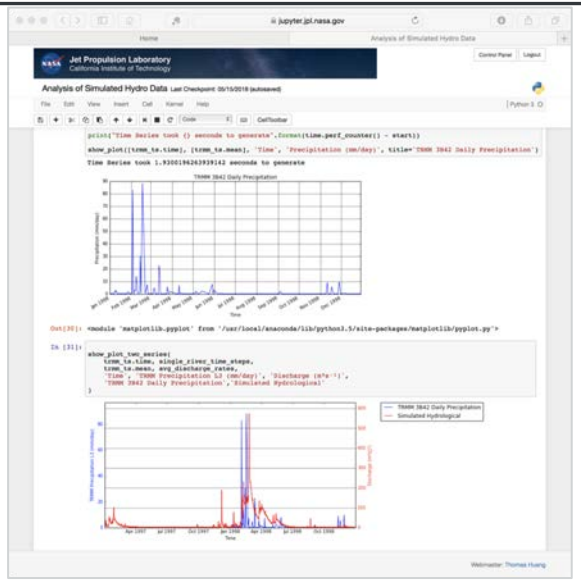
Analyze Large Collection of Observational Data Interactively
… across the ocean



Retrieval of a single river time series



Retrieval of time series from 9 rivers

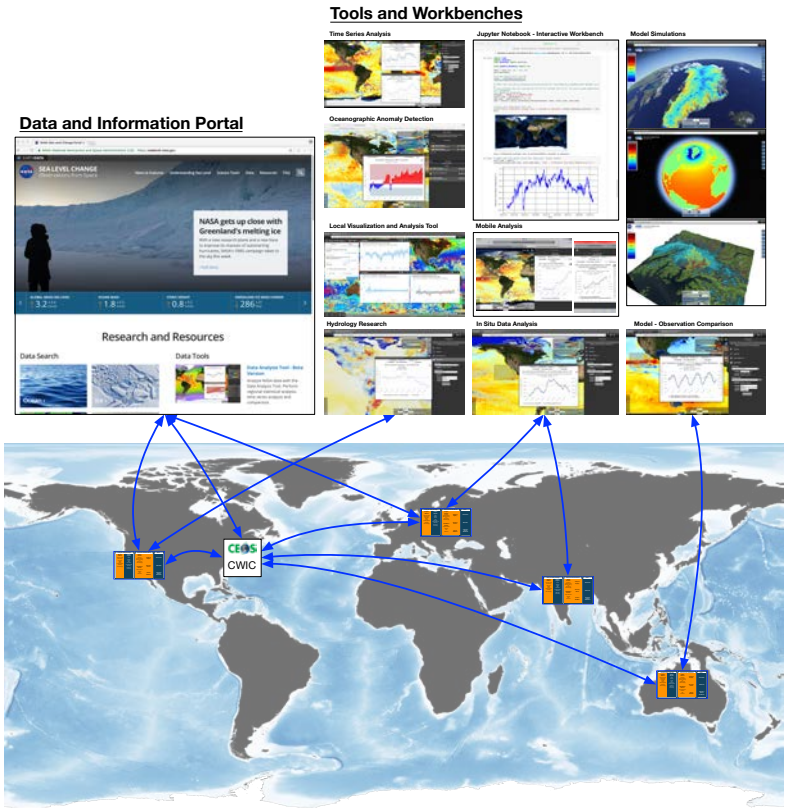


Coordination between TRMM and river

- Running Jupyter from Germany and interacts with analytics services hosted on Amazon and at JPL
- Simulated hydrology data in preparation for SWOT hydrology
- **River data**: ~3.6 billion data points. 3-hour sample rate. Consists of measurements from ~600,000 rivers
- **TRMM data**: 17 years, .25deg, 1.5 billion data points
- Sub-second retrieval of river measurements
- On-the-fly computation of time series and generate coordination plot

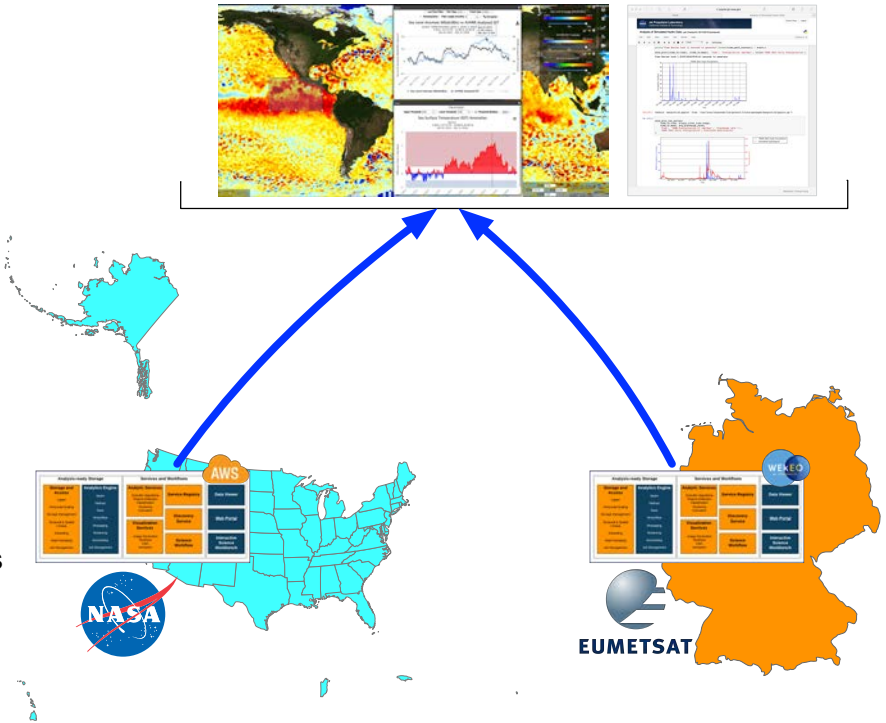
Distributed Analytics Center Architecture

- **Committee of Earth Observation Satellites (CEOS) Ocean Variables Enabling Research and Applications for GEO (COVERAGE) Initiative**
- Seeks to provide **improved access** to **multi-agency ocean remote sensing data** that are **better integrated with in-situ and biological observations**, in support of **oceanographic and decision support applications** for societal benefit.
- A community-support open specification with common taxonomies, information model, and API (maybe security)
- Putting value-added services next to the data to eliminate unnecessary data movement
- Avoid data replication. Reduce unnecessary data movement and egress charges
- Analytic engine infused and managed by the data centers perhaps on the Cloud
- Researchers can perform multi-variable analysis using any web-enabled devices without having to download files



Distributed Analytics Center Architecture

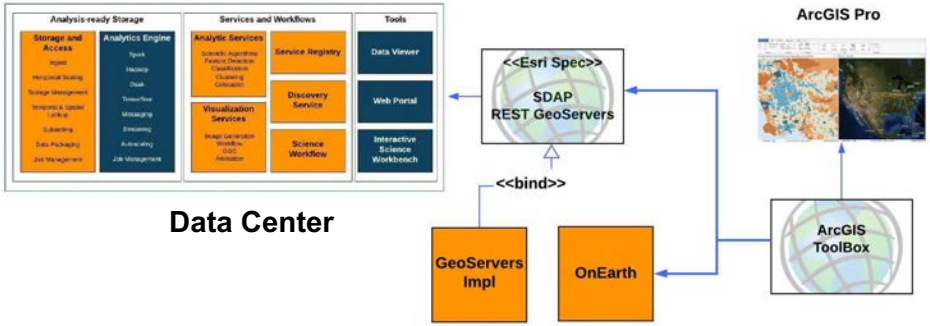
- **Committee of Earth Observation Satellites (CEOS) Ocean Variables Enabling Research and Applications for GEO (COVERAGE) Initiative**
- Seeks to provide **improved access** to **multi-agency ocean remote sensing data** that are **better integrated with in-situ and biological observations**, in support of **oceanographic and decision support applications** for societal benefit.
- A community-support open specification with common taxonomies, information model, and API (maybe security)
- Putting value-added services next to the data to eliminate unnecessary data movement
- Avoid data replication. Reduce unnecessary data movement and egress charges
- Analytic engine infused and managed by the data centers perhaps on the Cloud
- Researchers can perform multi-variable analysis using any web-enabled devices without having to download files



COVERAGE Phase B 2019-2020

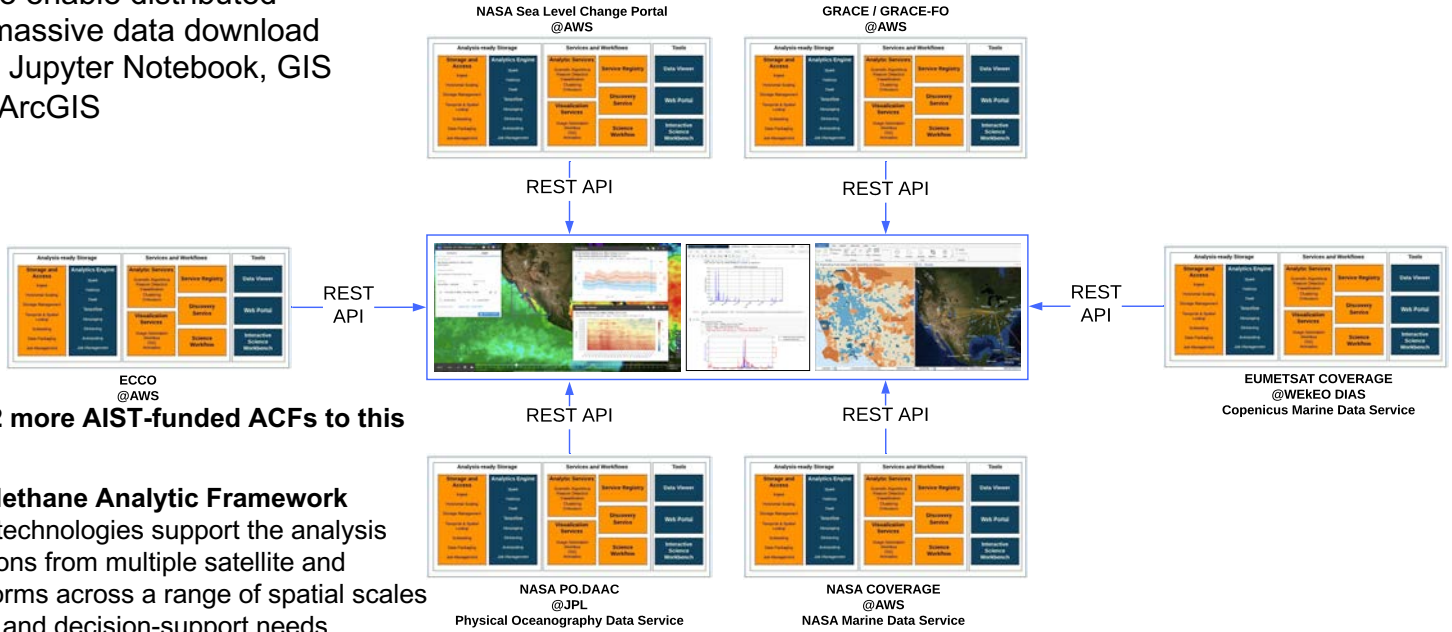
Enabling the Private Sector and Community

- Embrace open standards (ISO, OGC, etc.)
- Open source = empowering organizations and community
- Example: Building open source bridge between Esri's ArcGIS with Apache SDAP
  - Allowing data center to use SDAP, which is free and open source
  - Allowing Esri user community to directly access and analyze satellite observational data directly using Esri applications without having to download massive collection of data to their local computers



Federated Data Analytics

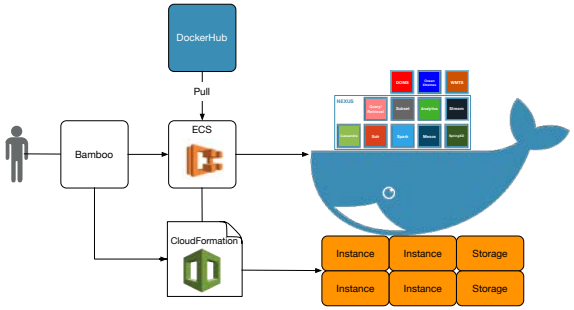
Federated ACF instances to enable distributed analytics without requiring massive data download and transfer. Clients can be Jupyter Notebook, GIS Web Applications, and Esri ArcGIS

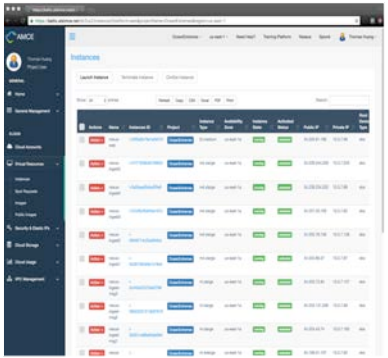


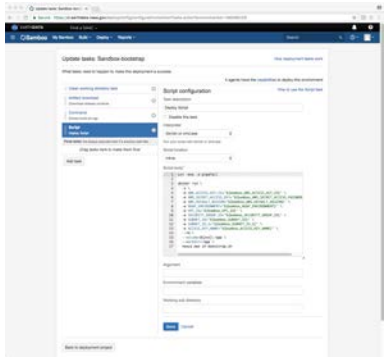
2019 – 2021, we will be added 2 more AIST-funded ACFs to this federation

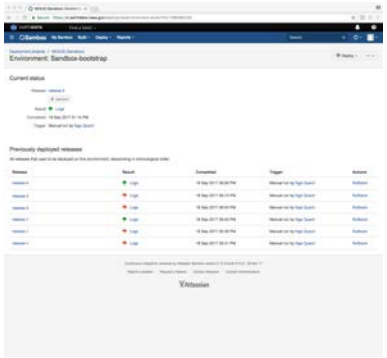
- **NASA AIST-18 Multi-scale Methane Analytic Framework (M^2AF)** to develop an mature technologies support the analysis and use of methane observations from multiple satellite and airborne remote sensing platforms across a range of spatial scales necessary to address science and decision-support needs.
- **NASA AIST-18 On-Demand Geospatial Spectroscopy Processing Environment on the Cloud (GeoSPEC)**. GeoSPEC is to establish an ACF for scientific community to visualization and analysis of high-level imaging spectroscopy products and to facilitate the SBG mission and the NEON project.

Automated, Container-based Deployment



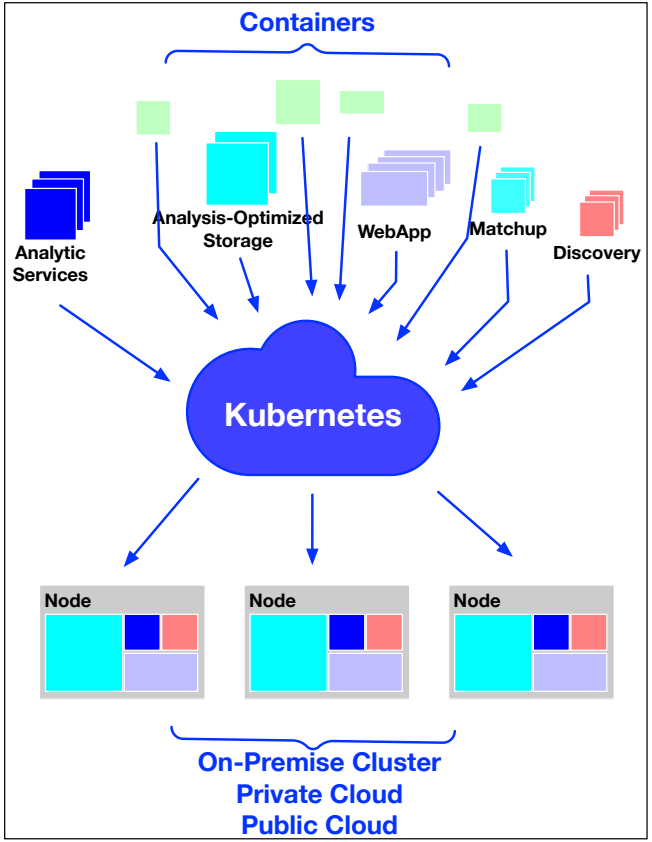






NASA AIST Managed Cloud Environment

NASA Next Generation Application Platform (NGAP)



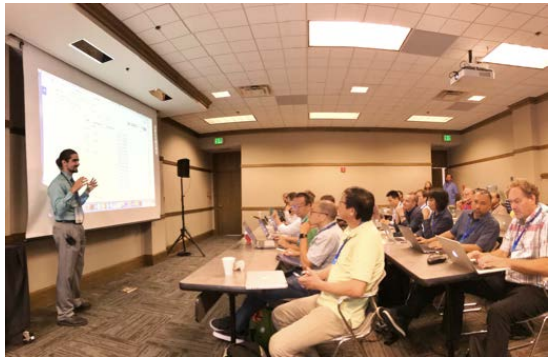
Jet Propulsion Laboratory
California Institute of Technology

Apache SDAP Acknowledgement

Ed Armstrong/JPL
Jason Barnett/LARC
Andrew Bingham/JPL
Carmen Boening/JPL
Mark Bourassa/FSU
Mike Chin/JPL
Marge Cole/NASA
Tom Cram/NCAR
Dan Crichton/JPL
Maya DeBellis/JPL
Rich Doyle/JPL
Jocelyn Elya/FSU
Ian Fenty/JPL
Eamon Ford/JPL
Kevin Gill/JPL
Frank Greguska/JPL
Patrick Heimbach/UT Austin
Ben Holt/JPL
Thomas Huang/JPL
Joe Jacob/JPL
Zaihua Ji/NCAR
Yongyao Jiang/GMU
Felix Landerer/JPL
Yun Li/GMU
Eric Lindstrom/NASA
Mike Little/NASA
Chris Lynnes/NASA
Lewis McGibbney/JPL
David Moroni/JPL
Kevin Murphy/NASA
Charles Norton/JPL
Jean-Francois Piolle/IFREMER
Nga Quach/JPL
Brandi Quam/NASA
Shawn Smith/FSU
Ben Smith/JPL
Adam Stallard/FSU
Rob Toaz/JPL
Vardis Tsontos/JPL
Suresh Vannan/JPL
Jorge Vazquez/JPL
Ou Wang/JPL
Brian Wilson/JPL
Steve Worley/NCAR
Elizabeth Yam/JPL
Phil Yang/GMU
Alice Yepremyan/JPL

Building Community-Driven Open Data and Open Source Solutions

- Deliver solutions to establish coherent platform solutions
- Embrace open source software
- Community validation
- Evolve the technology through community contributions
- Share recipes and lessons learned
- Technology demonstrations
- Host webinars, hands-on cloud analytics workshops and hackathons



Big Data Analytics and Cloud Computing Workshop, 2017 ESIP Summer Meeting, Bloomington, IN

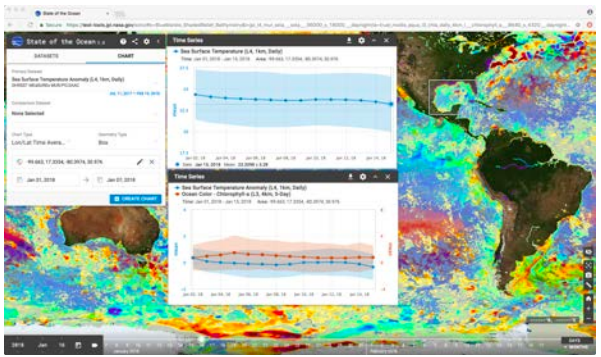
2019 JPL Data Science Showcase

Partner with NASA and non-NASA Projects - Deliver to Production

- **The gap between visionary to pragmatists is significant**. – Geoffrey Moore
- Become an expert in the production environment and devote resources in automations
- Give project engineering team early access to the PaaS
- Deliver all technical documents and work with project system engineering
- Provide project-focused trainings

NASA's Sea Level Change Team

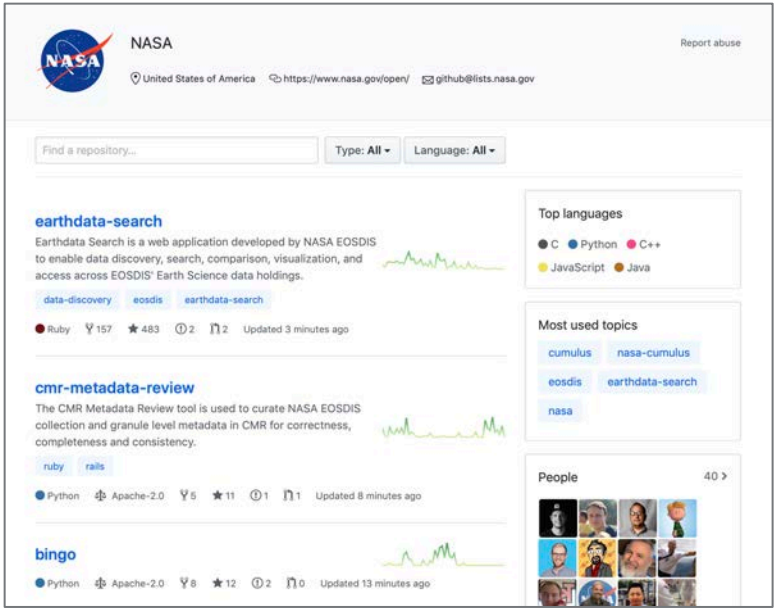
CEOS SIT Technical Workshop



NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)

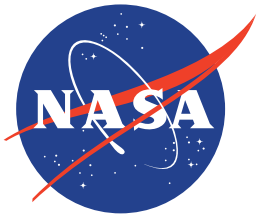
In Summary

- **You've got to think about big things while you're doing small things, so that all the small things go in the right direction** – Alvin Toffler
- Climate research requires Autonomously Sustainable Solutions
- Focus on delivering professional quality open source solutions
- Enables end-to-end data and computation architecture, and the total cost of ownership
- From generalization to specialization
- Open source should not be a destination, it should be in place from the beginning
- How a technology is being managed will determine how far it can go



https://github.com/nasa

If you want to go fast, go Alone. If you want to go far, go Together.



JPL Caltech

Thomas Huang

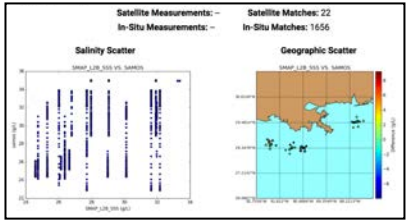
thomas.huang@jpl.nasa.gov

Jet Propulsion Laboratory

California Institute of Technology

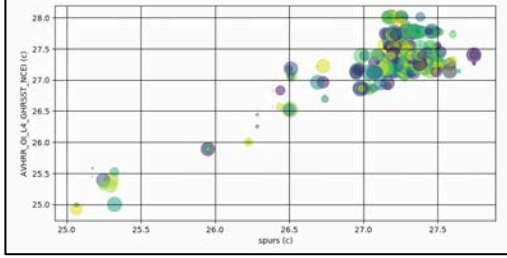
In Situ to Satellite Matchup

- Typically data matching is done using one-off programs developed at multiple institutions
- A primary advantage of SDAP's matchup service is the reduction in duplicate development and man hours required to match satellite/in situ data
  - Removes the need for satellite and in situ data to be collocated on a single server
  - Systematically recreate matchups if either in situ or satellite products are re-processed (new versions), i.e., matchup archives are always up-to-date.
- Provides data querying, subset creation, match-up services, and file delivery operational.
- Plugin architecture for in situ data source using EDGE, a open source implementation of Open Search



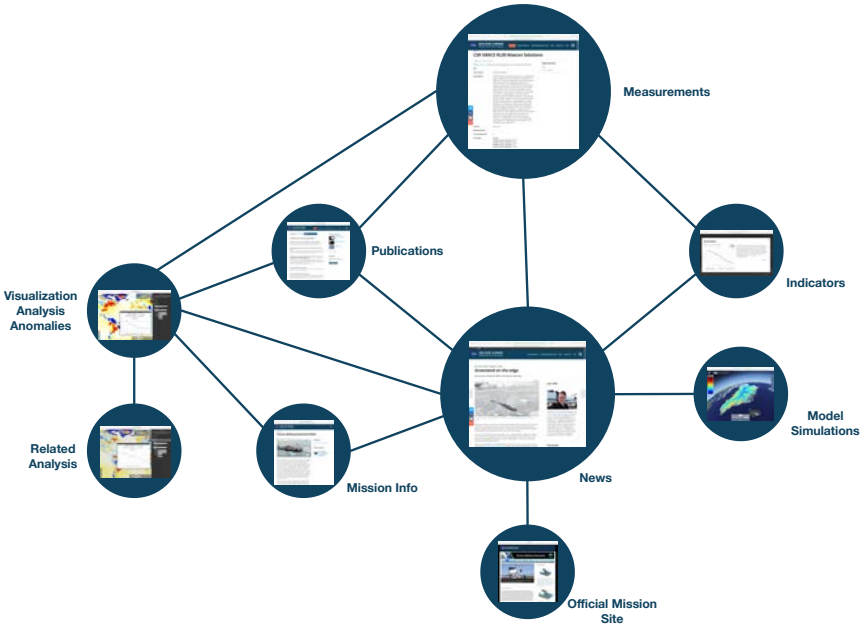

```
import requests
import json
import numpy as np
import datetime
import time

primary='AVHRR_OI_L4_GHRSST_NCEI'
secondary='spurs'
startTime='2013-10-01T00:00:00Z'
endTime='2013-10-30T23:59:59Z'
timeTolerance='86400'
radius='10000.0'
bbox='-30,15,-45,30'
platforms='1,2,3,4,5,6,7,8,9'
parameter='sst'
depthMin='0'
depthMax='5'

url = 'https://doms.jpl.nasa.gov/match_spark?'
url += 'primary=' + primary
url += '&matchup=' + secondary
url += '&startTime=' + startTime
url += '&endTime=' + endTime
url += '&tt=' + timeTolerance
url += '&rt=' + radius
url += '&b=' + bbox
url += '&platforms=' + platforms
url += '&parameter=' + parameter
url += '&matchOne=true'
url += '&depthMin=' + depthMin
url += '&depthMax=' + depthMax

print (url)
start = time.time()
ts = json.loads(str(requests.get(url).text))
spent = time.time() - start
print ("It took: " + str(spent) + " sec")

satellite = []
in_situ = []
for data in ts['data']:
    for matches in data['matches']:
        satellite.append(data['sea_water_temperature'])
        in_situ.append(matches['sea_water_temperature'])

show_plot(in_situ, satellite, secondary+' (c)', primary+' (c)')
```

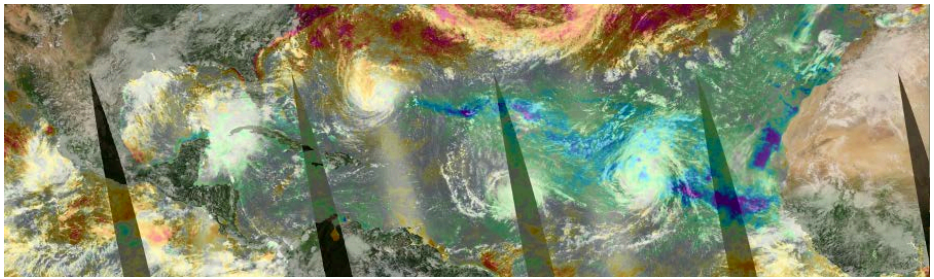
```
https://doms.jpl.nasa.gov/match_spark?primary=AVHRR_OI_L4_GHRSST_NCEI&matchup=spurs&startTime=2013-10-01T00:00:00Z&endTime=2013-10-30T23:59:59Z&tt=86400&rt=10000.0&b=-30,15,-45,30&platforms=1,2,3,4,5,6,7,8,9&parameter=sst&matchOne=true&depthMin=0&depthMax=5
It took: 4.915551424026489 sec
```



Tackling Information Discovery

- One of the big changes in Earth science is finding the relevant data and related online information
- We are developing smarter data search and discovery solution that is capable of adjusting search result according how user search, retrieval, and external events
- Use Machine Learning methods to adjust search ranking by taking a number of features into consideration
- Semantically mind dataset metadata to identify relationship
- Dynamically detect relationship between data, models, tools, publications, and news
- **Relevancy** is Domain-specific, Personal, Temporal, and Dynamic





Air-sea Interaction during Hurricanes Florence, Joyce, and Helene in the Atlantic Ocean