Jet Propulsion Laboratory
California Institute of Technology

Overview of JPL Data Science for Earth Science

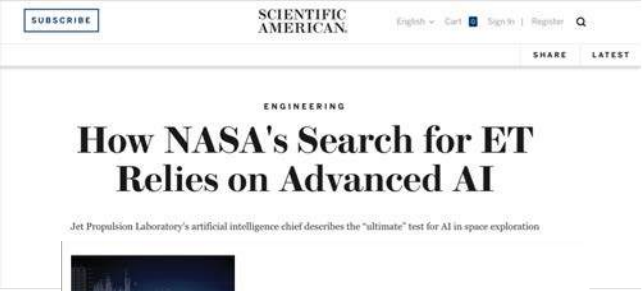
Thomas Huang
thomas.huang@jpl.nasa.gov
Group Supervisor - Computer Science for Data-Intensive Applications
Strategic Lead - Interactive Data Analytics

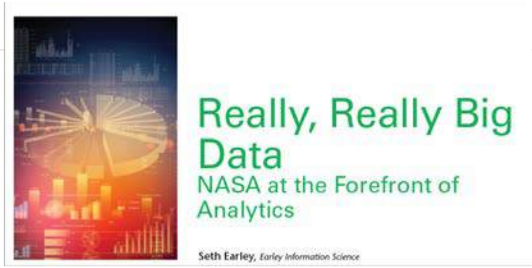
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.

[CL # 19-0887]

Tackling the Data Science and Data Challenges

- JPL is engaging data science and AI technologies and methodologies for science, mission operations, engineering applications
  - From onboard computing to scalable archives to analytics
  - Applying ML techniques with supporting infrastructure
- JPL has established a program focused on building and implementing an institution-wide strategy for data science and AI
  - Expanding from archives to enable data analytics as a first class activity
  - Methodology transfer across disciplines
  - Research partnerships with academia, government, and industry





Driving AI and Data Science into JPL Activities

- In 2017-2018, JPL launched 25 data science pilots
  - Spanning science, mission and Deep Space Network operations, and formulation
  - Building towards a data science vision of full utilization of data and agile application of analytics

Use Cases: Science

Use Cases: Mission Ops

Use Cases: Formulation

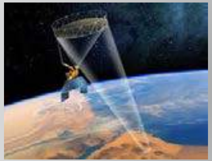
Use Cases: Institution

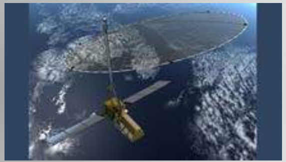
Applying Data Science Across the Mission-Science Data Lifecycle

End-to-End Data and Computational Architecture

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

Observational Platforms and Flight Computing



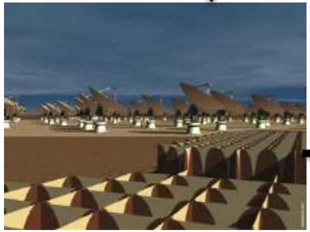


SMAP (Today): 485 GB/day **NI-SAR (2020): 86 TB/day**

(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*

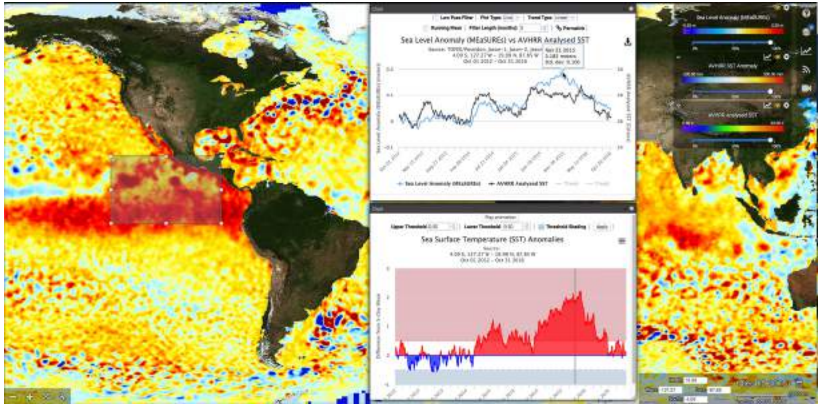
(2) Data collection capacity at the instrument continually outstrips data transport (downlink) capacity

Ground-based Mission Systems

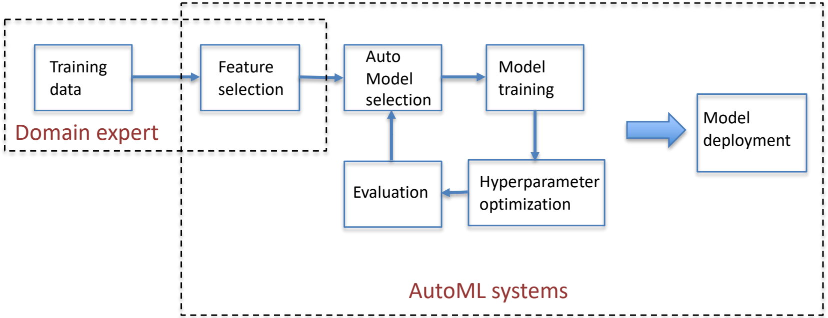
(3) Data distributed in massive archives; many different types of measurements and observations

Opportunities Enabled by Data Science

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

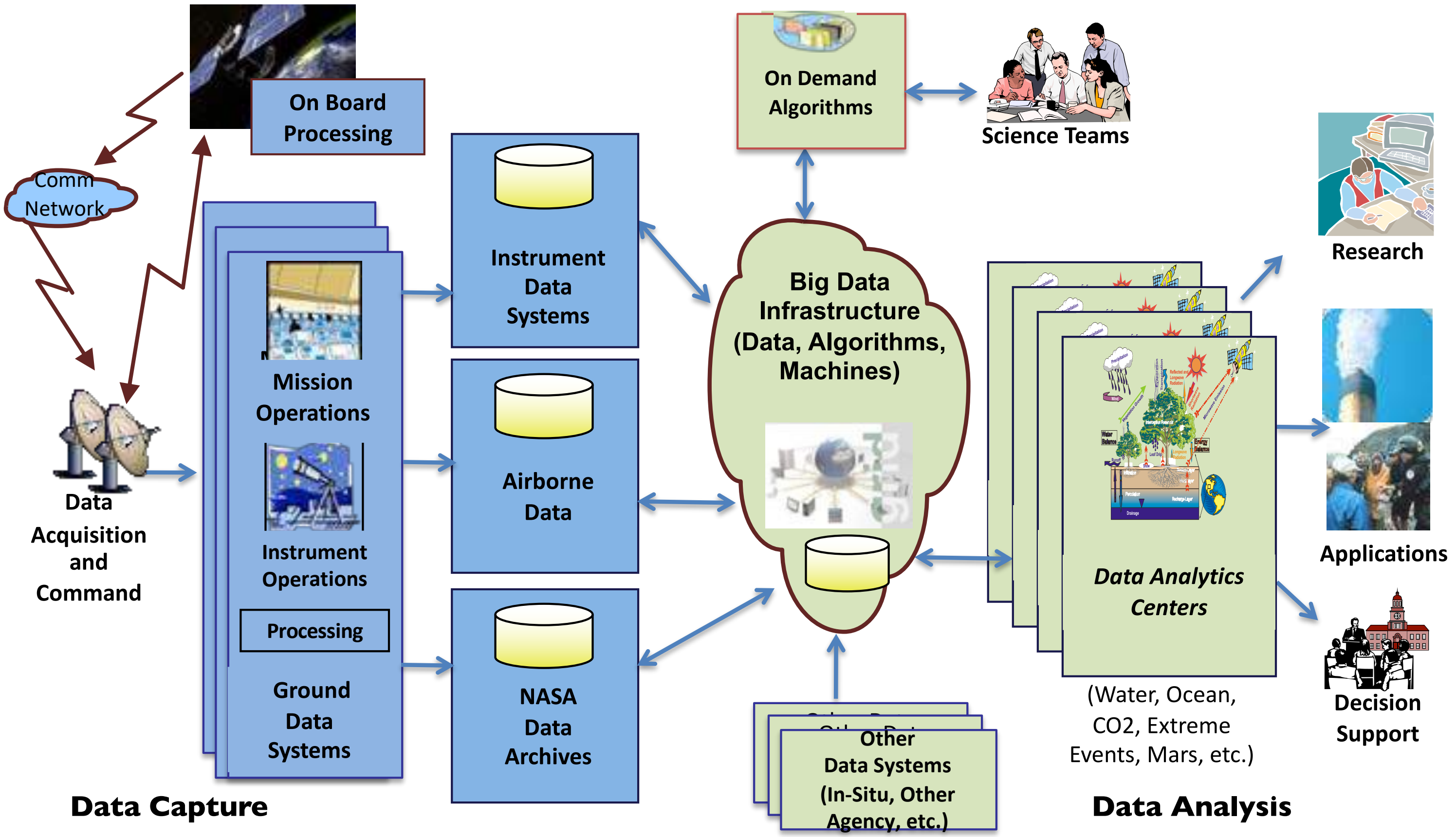


NASA AIST: OceanXtremes - Anomaly Detection Solution



Automate Machine Learning

Shift Toward Data Analytics



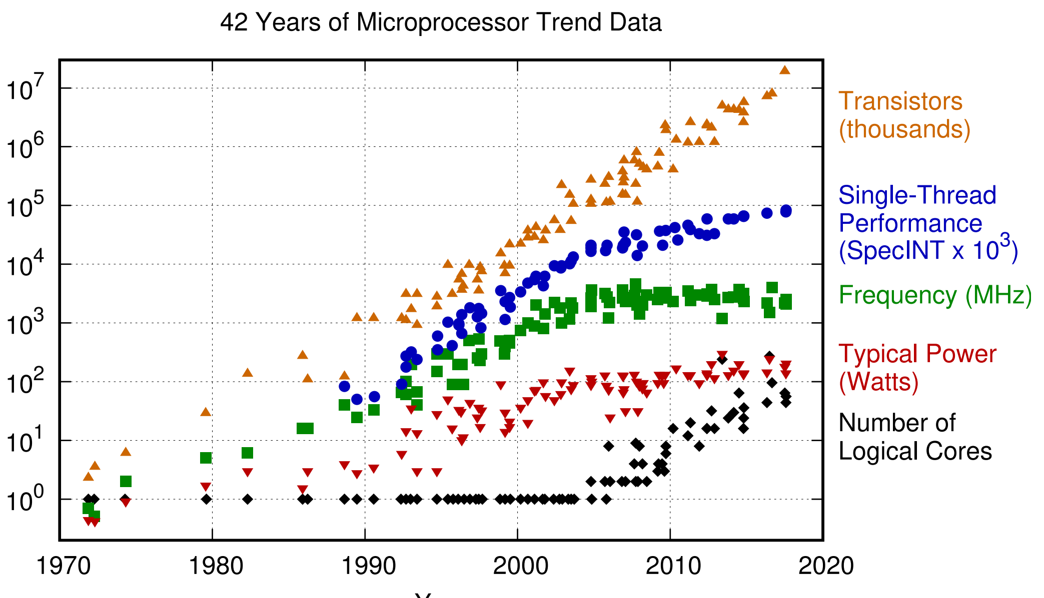
Processors are not Getting Faster

2004: First Pentium 4 processor with 3.0GHz clock speed

2018: Apple’s MacBook Pro has clock speed of 2.7GHz

14 years later, not much has gain in raw processing power

Modern big data architects are required to “think outside of the box”. Literally!



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

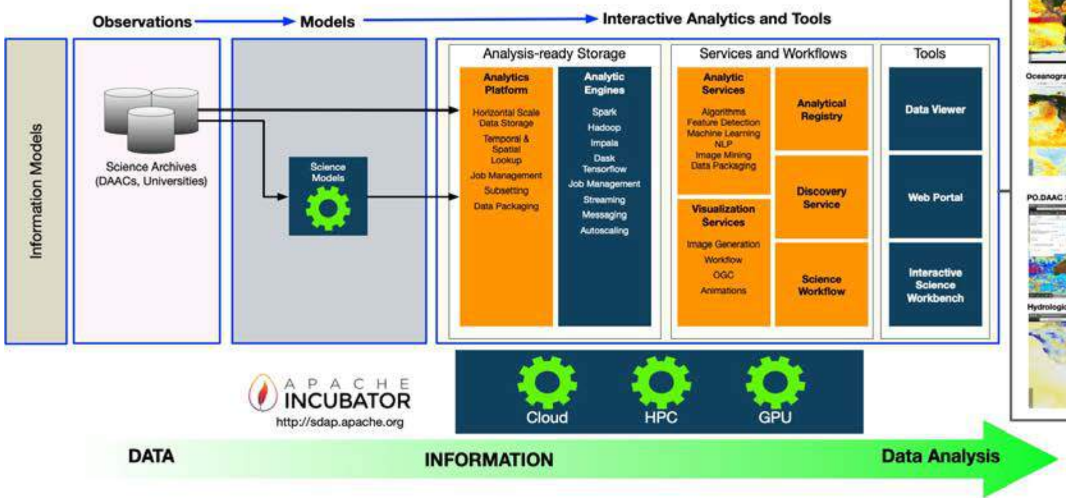
Jet Propulsion Laboratory
California Institute of Technology

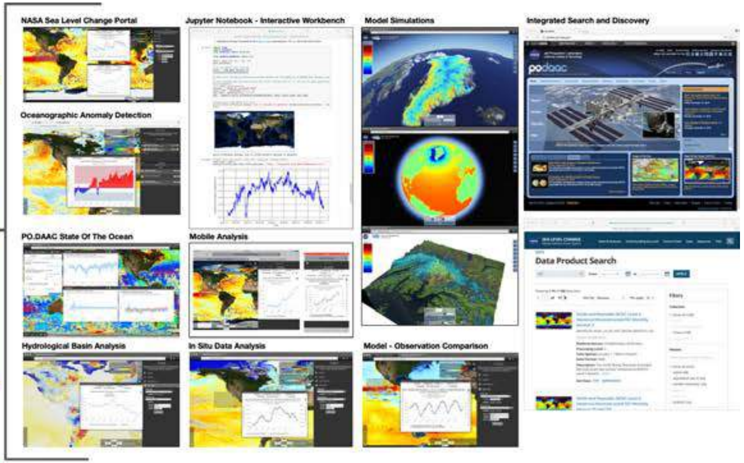
Some Background Info

- **Agencies are historically focused on systematic capture and stewardship of data for observational Systems**
- **With large amount of observational and modeling data,**
  - The overall cost for data stewardship is expecting to rise significantly
  - Finding and downloading is becoming inefficient
- **Reality with large amount of observational and modeling data**
  - Downloading to local machine is becoming inefficient
  - Search has gotten a lot faster, but finding the relevant measurement has becoming a very time consuming process
  - Analyze decades of regional measurement is labor-intensive and costly
- **Increasing "big data" era is driving needs to**
  - Scale computational and data infrastructures
  - Support new methods for deriving scientific inferences and **shift towards integrated data analytics**
  - Apply computational and data science across the lifecycle
- **Scalable Data Management**
  - Capture well-architected and curated data repositories based on well-defined data/information architectures
  - Architecting automated pipelines for data capture
- **Scalable Data Analytics**
  - Access and integration of highly distributed, heterogeneous data
  - Novel statistical approaches for data integration and fusion
  - Computation applied at the data sources
  - Algorithms for identifying and extracting interesting features and patterns

Integrated Science Data Analytics Platform

Creating SaaS and PaaS for Science Tools and Services



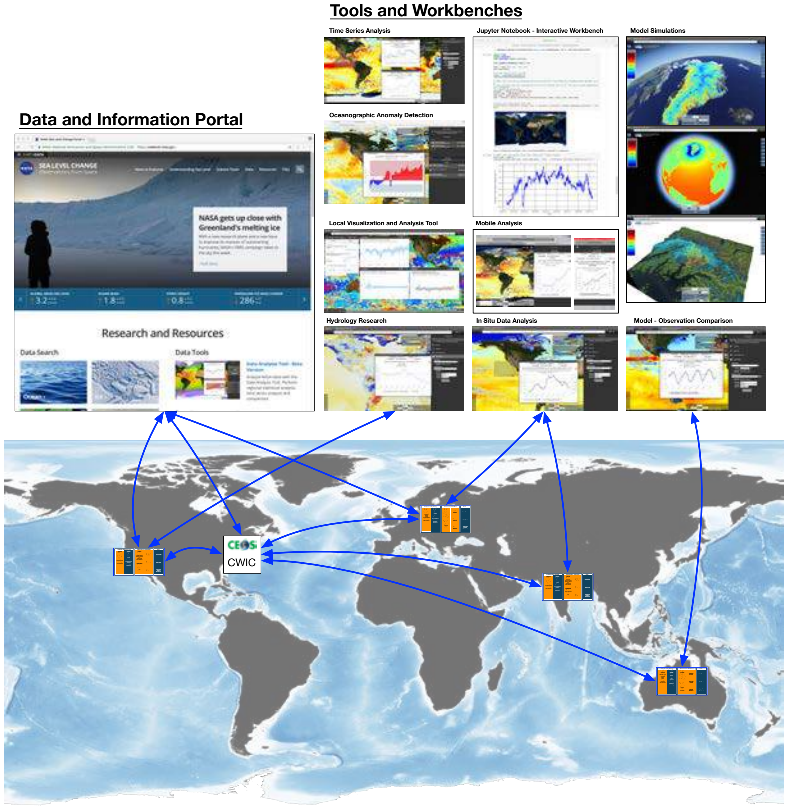


Systematic analysis of massive data

- **Integrated Science Data Analytics Platform**: an analytic center framework 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

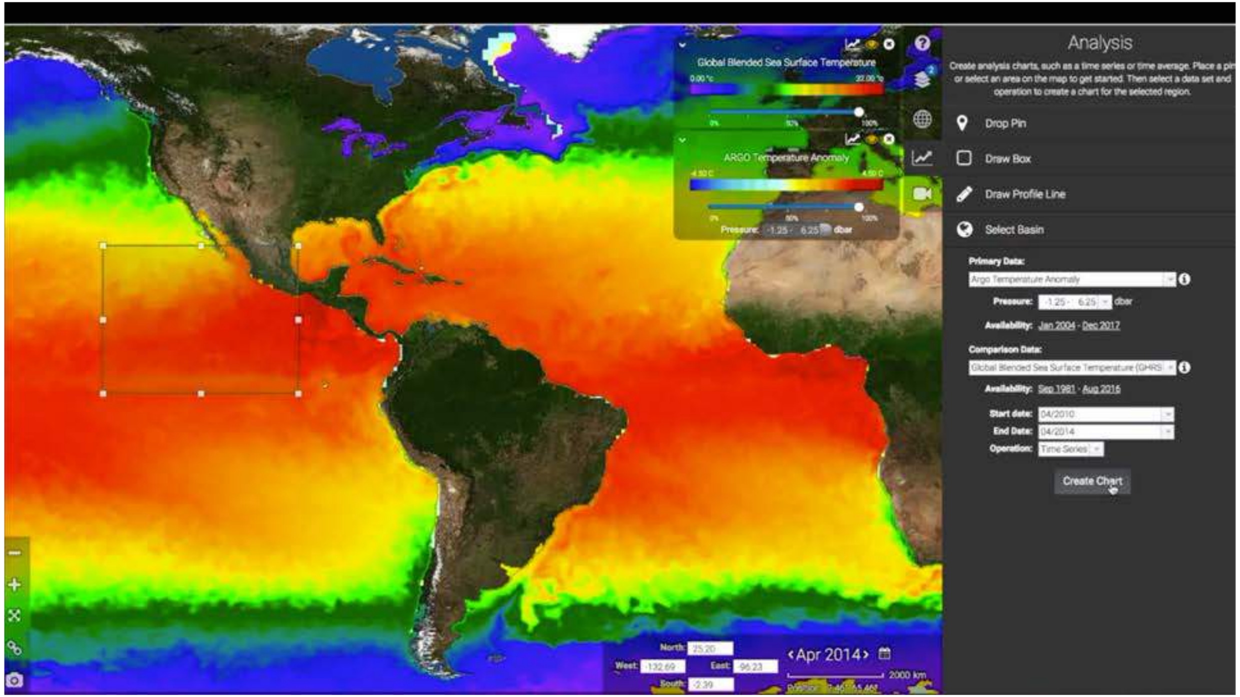
Architecture for Distributed Data System and Analysis

- **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
- Public accessible RESTful analytic APIs where computation is next to the data
- 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



Visualize and Analyze Sea Level

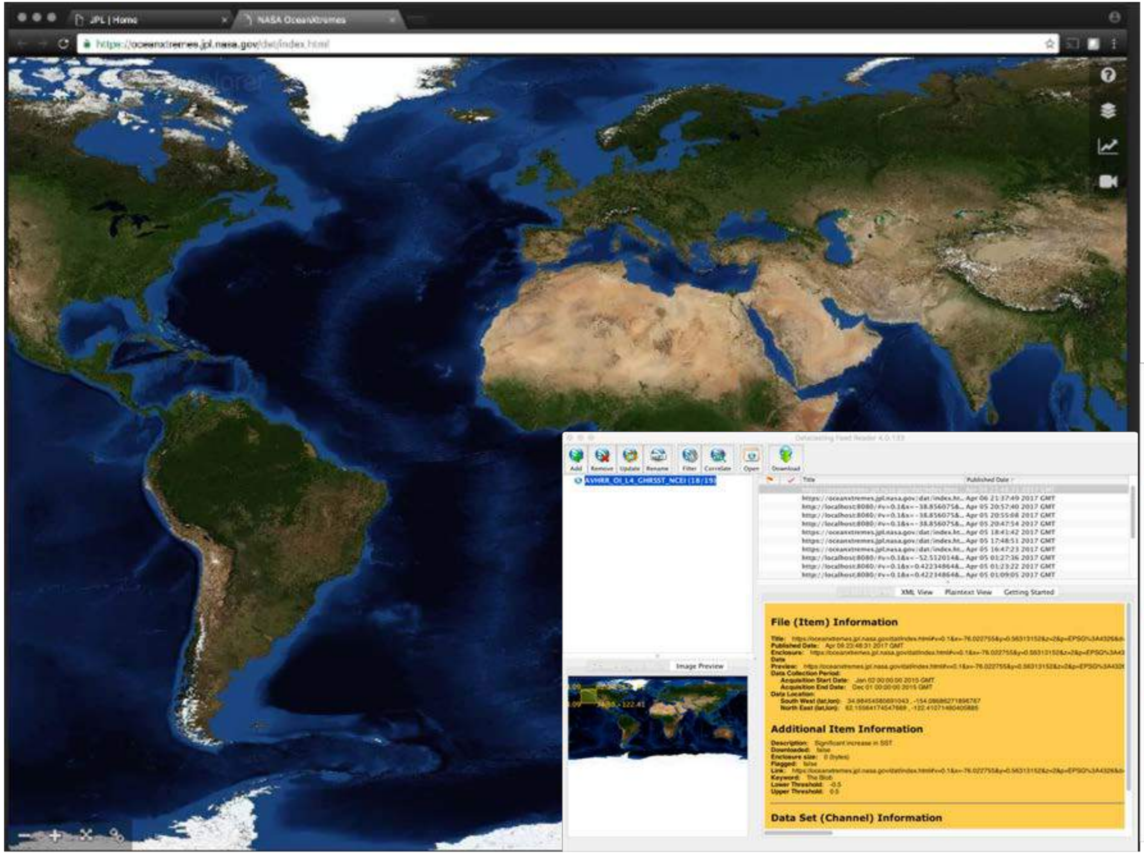
https://sealevel.nasa.gov



Analyze in situ and satellite observations

Analyze Sea Level mobiles

Analyze Ocean Anomaly – “The Blob”



- **Visualize** parameter
- **Compute** daily differences against climatology
- **Analyze** time series area averaged differences
- **Replay** the anomaly and visualize with other measurements
- **Document** the anomaly
- **Publish** the anomaly

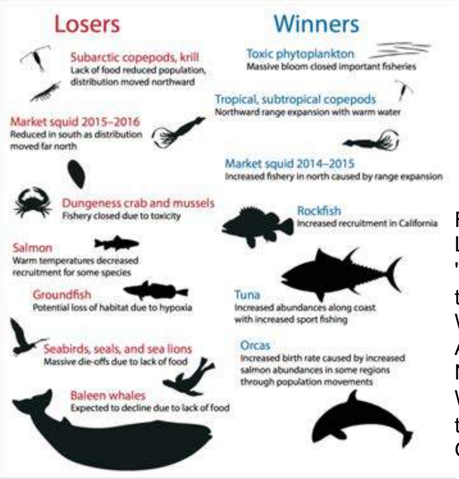
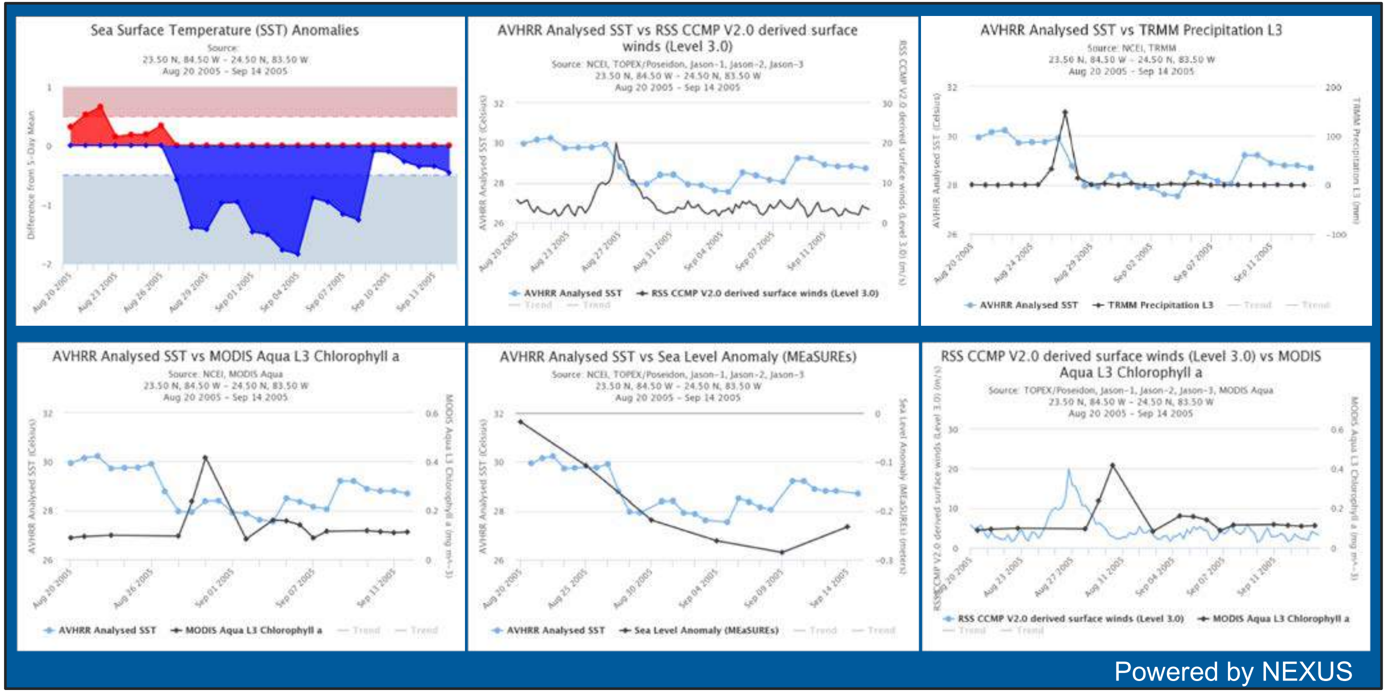


Figure from Cavole, L. M., et al. (2016). "Biological Impacts of the 2013–2015 Warm-Water Anomaly in the Northeast Pacific: Winners, Losers, and the Future." Oceanography 29.

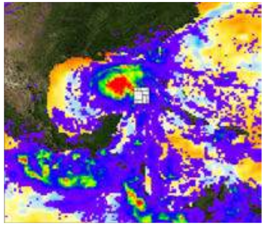
Hurricane Katrina Study



A study of a Hurricane Katrina–induced phytoplankton bloom using satellite observations and model simulations
Xiaoming Liu, Menghua Wang, and Wei Shi
JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 114, C03023, doi:10.1029/2008JC004934, 2009

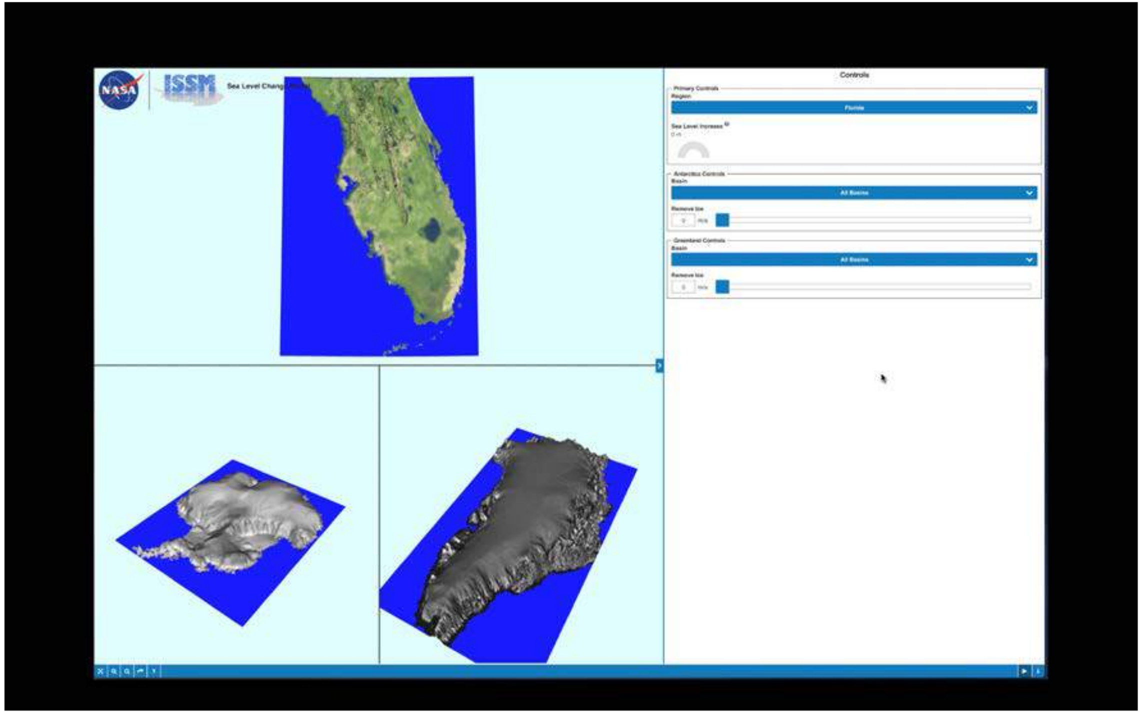
Hurricane Katrina passed to the southwest of Florida on Aug 27, 2005. The ocean response in a 1 x 1 deg region is captured by a number of satellites. The initial ocean response was an immediate cooling of the surface waters by 2 ℃ that lingers for several days. Following this was a short intense ocean chlorophyll bloom a few days later. The ocean may have been "preconditioned' by a cool core eddy and low sea surface height.

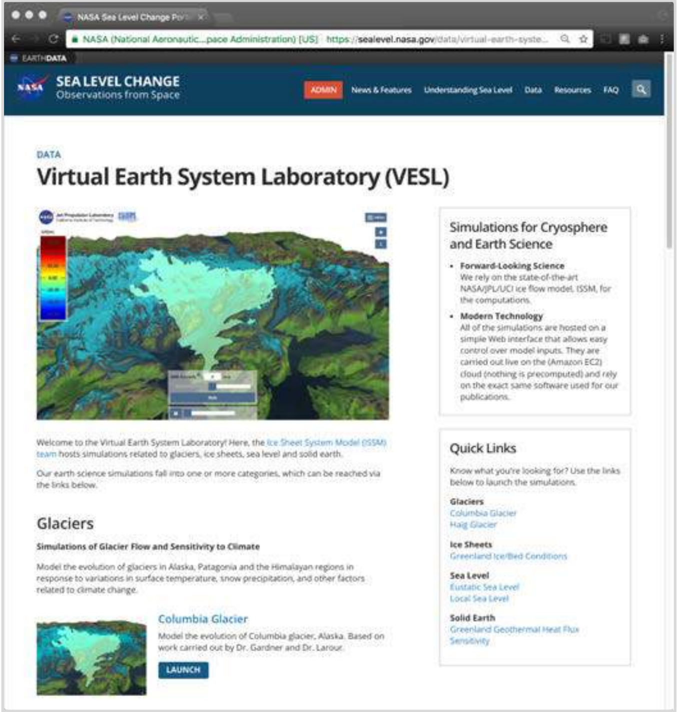
The SST drop is correlated to both wind and precipitation data. The Chl-A data is lagged by about 3 days to the other observations like SST, wind and precipitation.



Hurricane Katrina TRMM overlay SST Anomaly

Virtual Earth System Laboratory (VESL)

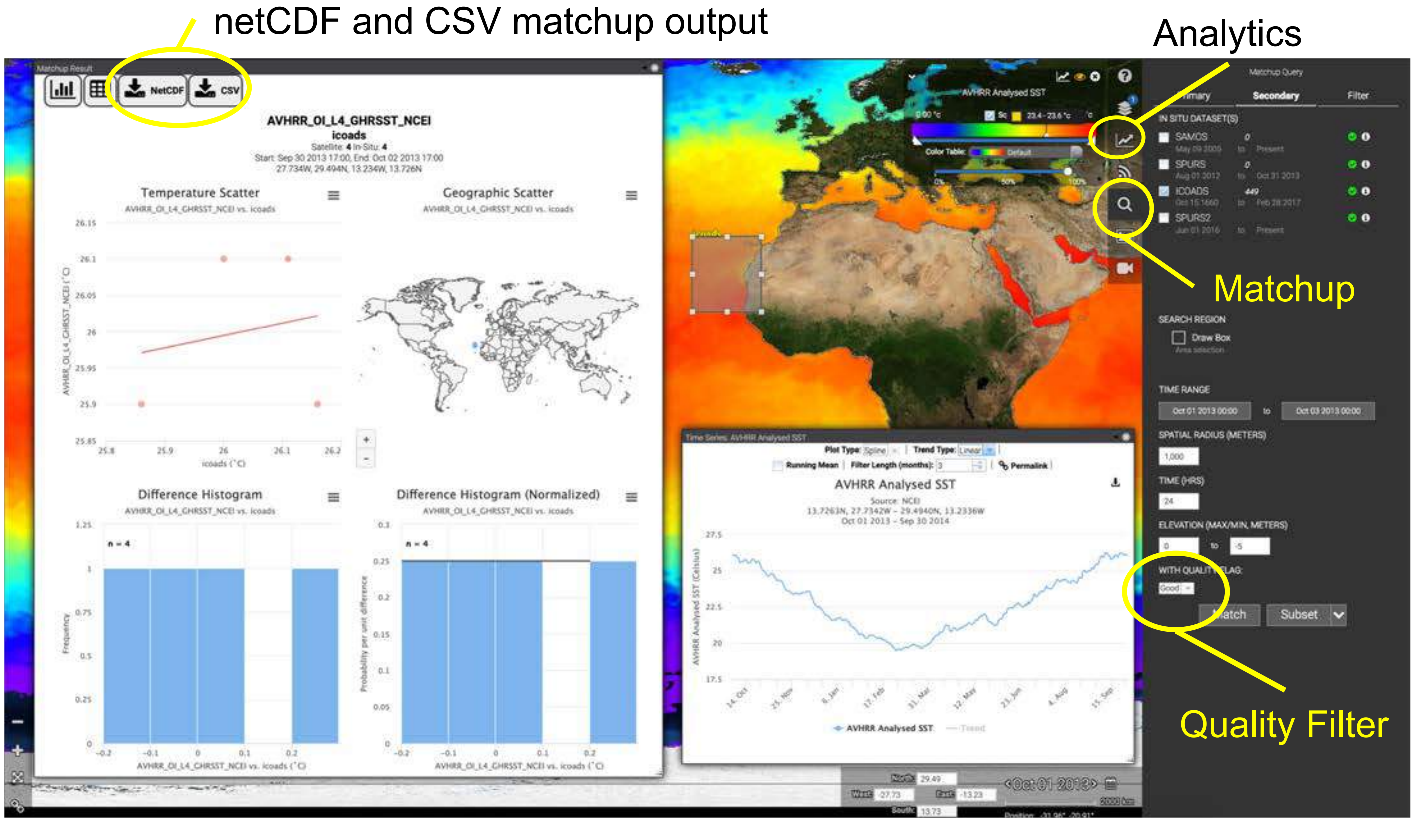




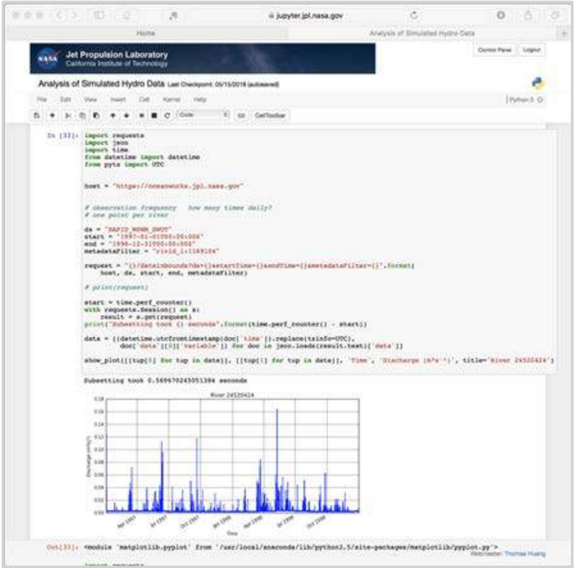
- Web-based 3D Simulations
- Computation on Amazon Cloud

Integrated Analysis Tool

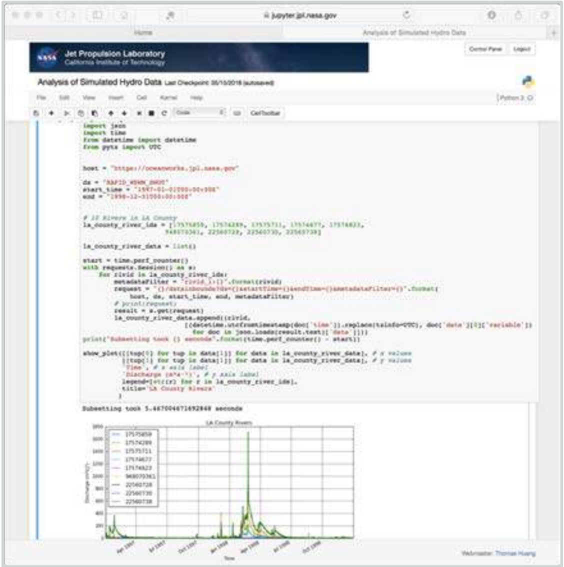
NASA AIST - OceanWorks



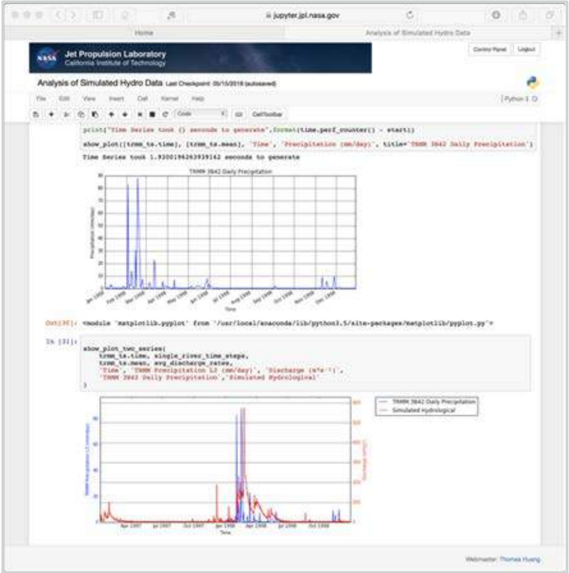
Support for Hydrology



Retrieval of a single river time series



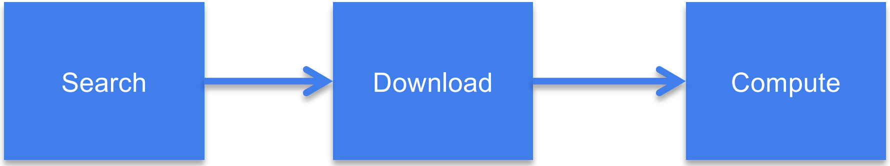
Retrieval of time series from 9 rivers

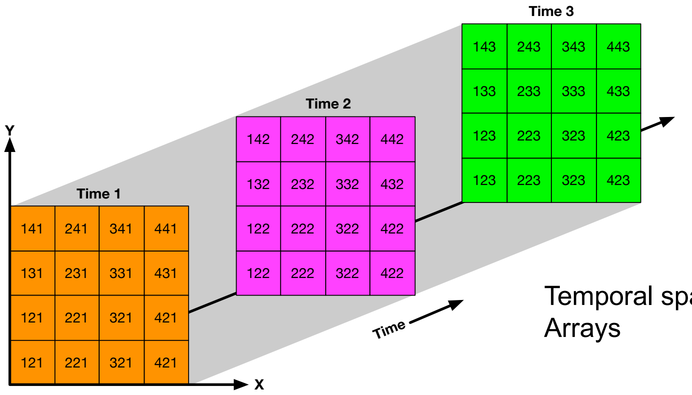


Time series coordination between TRMM and river

- 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

Traditional Method for Analyze Satellite Measurements



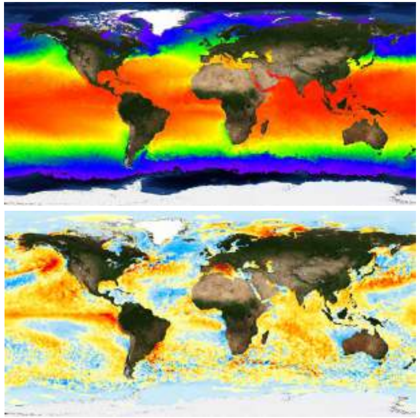


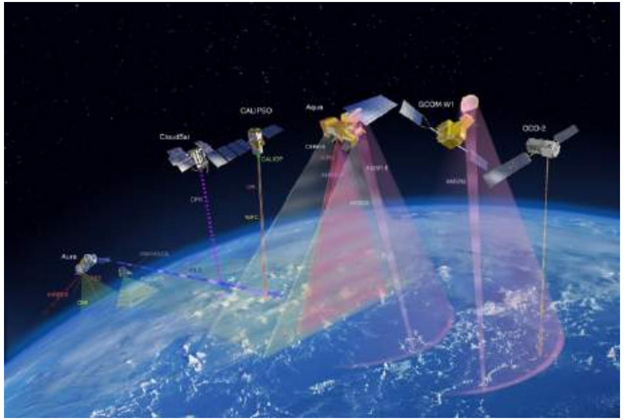
Temporal spatial Arrays

- Depending on the data volume (size and number of files)
- It could take many hours of download – (e.g. 10yr of observational data could yield thousands of files)
- It could take many hours of computation
- It requires expensive local computing resource (CPU + RAM + Storage)
- After result is produced, purge downloaded files

Observation

- Traditional methods for data analysis (time-series, distribution, climatology generation) can't scale to handle large volume, high-resolution data. They perform poorly
- Performance suffers when involve large files and/or large collection of files
- A high-performance data analysis solution must be free from file I/O bottleneck





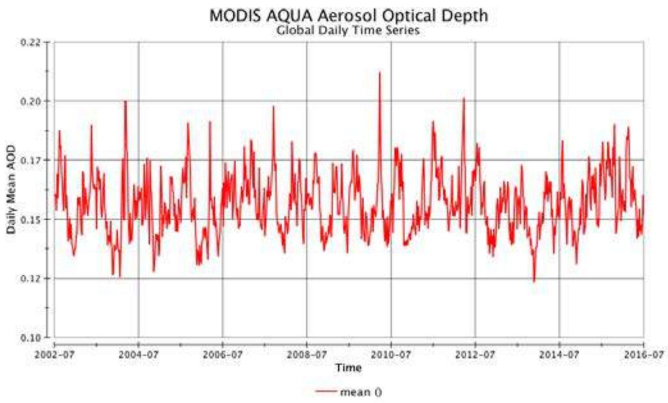
Parallel Analytics Performance

Dataset: MODIS AQUA Daily
Name: Aerosol Optical Depth 550 nm (Dark Target) (MYD08_D3v6)
File Count: 5106
Volume: 2.6GB
Time Coverage: July 4, 2002 – July 3, 2016

Giovanni: A web-based application for visualize, analyze, and access vast amounts of Earth science remote sensing data without having to download the data.

- Represents current state of data analysis technology, by processing one file at a time
- Backed by the popular NCO library. Highly optimized C/C++ library

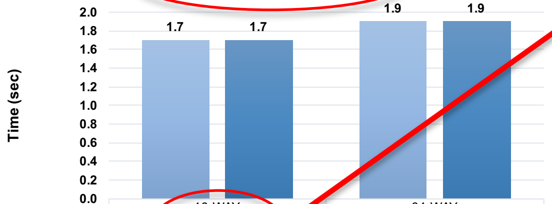
AWS EMR: Amazon's provisioned MapReduce cluster



Giovanni: 20 min
NEXUS: 1.7 sec

Area Averaged Time Series on AWS - Boulder
July 4, 2002 - July 3, 2016
NEXUS Performance

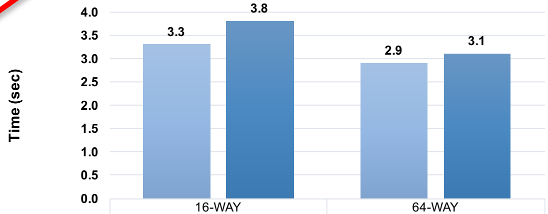
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1140.22 sec



	16-WAY	64-WAY
Custom Spark	1.7	1.9
AWS EMR	1.7	1.9

Area Averaged Time Series on AWS - Colorado
July 4, 2002 - July 3, 2016
NEXUS Performance

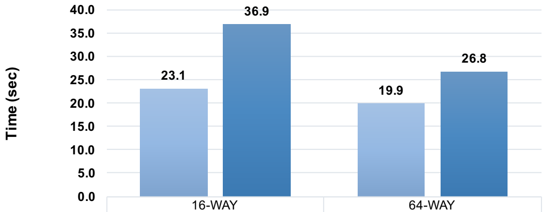
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1150.6 sec



	16-WAY	64-WAY
Custom Spark	3.3	2.9
AWS EMR	3.8	3.1

Area Averaged Time Series on AWS - Global
July 4, 2002 - July 3, 2016
NEXUS Performance

Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1366.84 sec

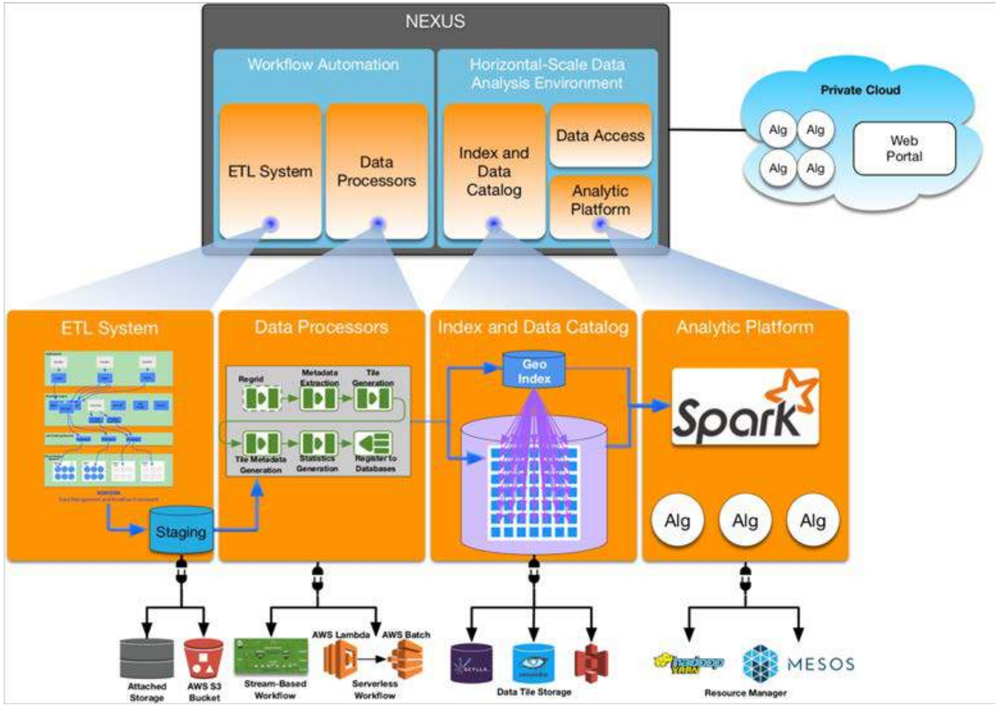


	16-WAY	64-WAY
Custom Spark	23.1	19.9
AWS EMR	36.9	26.8

Algorithm execution time. Excludes Giovanni's data scrubbing processing time

Evolve the Parallel Analytics Architecture

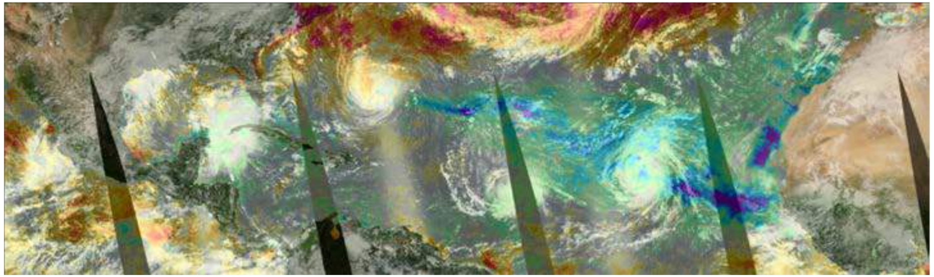
- **Several container-based deployment options**
  - Local on-premise cluster
  - Private Cloud
  - Amazon Web Service
- **Automate Data Ingestion with Image Generation**
  - Cluster based
  - Serverless (Amazon Lambda and Batch)
- **Data Store Options**
  - Apache Cassandra
  - ScyllaDB
  - Amazon Simple Storage Service (S3)
- **Resource Management Options**
  - Apache YARN
  - Apache MESOS
- **Analytic Engine Options**
  - Custom Apache Spark Cluster
  - Amazon Elastic MapReduce (EMR)
  - Amazon Athena (work-in-progress)



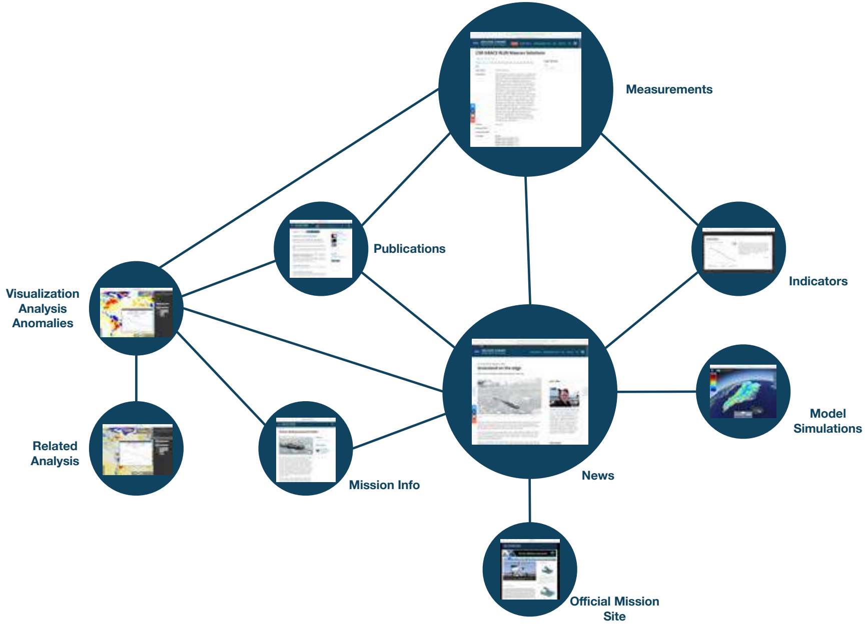
Apache SDAP's NEXUS supports public/private Cloud and local cluster deployments

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

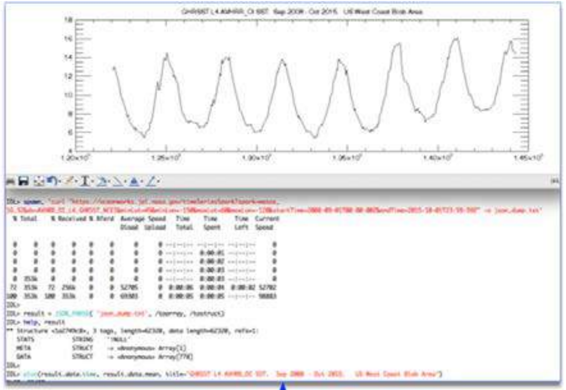


Air-sea Interaction during Hurricanes Florence, Joyce, and Helene in the Atlantic Ocean



Data Science Platform to Embrace Many Program Languages

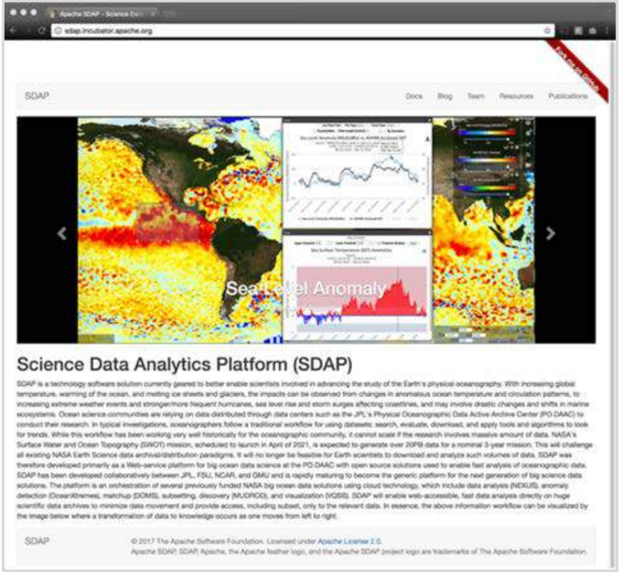
```
IDL> spawn, 'curl
"https://oceanworks.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_
GHRSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=2008-09-
01T00:00:00Z&endTime=2015-10-01T23:59:59Z" -o json_dump.txt'
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed

  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0     0    0     0    0     0      0      0 --:--:--  0:00:01 --:--:--     0
  0     0    0     0    0     0      0      0 --:--:--  0:00:02 --:--:--     0
  0     0    0     0    0     0      0      0 --:--:--  0:00:03 --:--:--     0
  0  353k    0     0    0     0      0      0 --:--:--  0:00:03 --:--:--     0
 72  353k   72  256k    0     0  52705      0  0:00:06  0:00:04  0:00:02 52702
100  353k  100  353k    0     0  69303      0  0:00:05  0:00:05 --:--:-- 98883
IDL>
IDL> result = JSON_PARSE( 'json_dump.txt', /toarray, /tostruct)
IDL> help, result
** Structure <1a2749c8>, 3 tags, length=62320, data length=62320, refs=1:
   STATS           STRING    '!NULL'
   META            STRUCT    -> <Anonymous> Array[1]
   DATA            STRUCT    -> <Anonymous> Array[778]
IDL>
IDL> plot(result.data.time, result.data.mean, title='GHRSST L4 AVHRR_OI SST.  Sep
2008 - Oct 2015.   US West Coast Blob Area')
PLOT <29457>
```



Credit: Ed Armstrong
Jun. 05, 2018

Free and Open Open Source Software (FOSS)

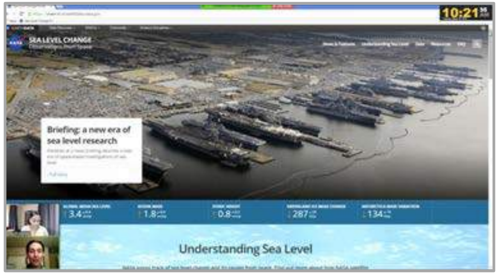
- October 2017, established Apache Software Foundation and established the **Science Data Analytics Platform (SDAP)** in the **Apache Incubator**
- Technology sharing through Free and Open Source Software (FOSS)
- Why? Further technology evolution that is restricted by projects / missions
- It is more than GitHub
    - Quarterly reporting
    - Reports are open for community review by over 6000 committers
    - SDAP has a group of appointed international Mentors: Jörn Rottmann, and Suneel Marthi
- SDAP and its affiliated projects are now being developed in the open
    - For local cluster and cloud computing platform
    - Fully containerized using Docker (multiple containers)
    - Infrastructure orchestration using Amazon CloudFormation
    - Analyzing satellite and model data
    - In situ data analysis and colocation with satellite measurements
    - Fast data subsetting
    - Data services integration architecture
    - OpenSearch and dynamic metadata translation
    - Mining of user interactions and data to enable discovery and recommendations
    - Streamline deployment through container technology

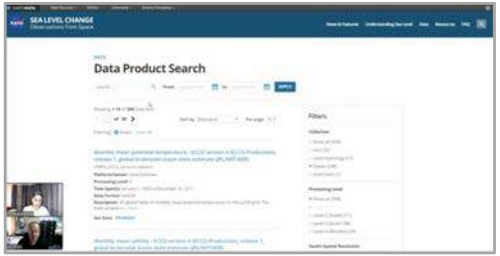


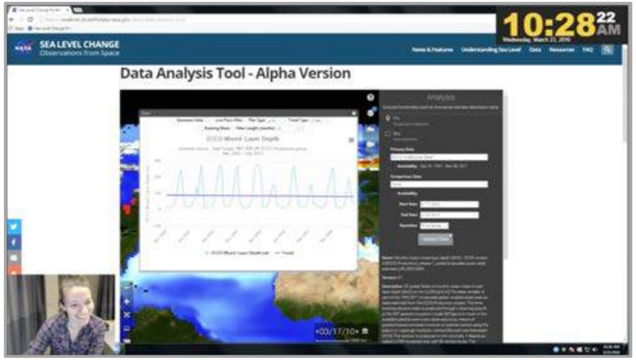
http://sdap.apache.org

Know The User's Real Needs

- **Work on improving communication - building bridge between IT and science**
  - **JPL's Data Science Program** is consists of technologists, project scientists, mission operations, etc.
  - Our science users tends get overwhelmed by tech jargons and cloud terminology
  - Learn to develop common language
- **Understand** how and for what purposes users obtain data and information
- **Describe** users' pain points and unmet needs for extracting, visualizing, comparing and analyzing science data
- **Identify** architectural approaches for tackling the real needs and identify opportunities for enhancing cross-disciplinary collaborative activities on the web portal.







Jet Propulsion Laboratory
California Institute of Technology

Building Community-Driven Open Source Solution

- Develop in the open, so every data provider can infuse the same software stack next to their data
- Establish or leverage an existing governance policy
- Community accessible issue tracking and documentations
- Community validation
- Evolve the technology through community contributions
- Share recipes and lessons learned
- Open source != less secure. Some open source technologies, Linux, Apache Webserver,, GNU, etc., have already been adopted by enterprises for years
- Host webinars, hands-on cloud analytics workshops and hackathons



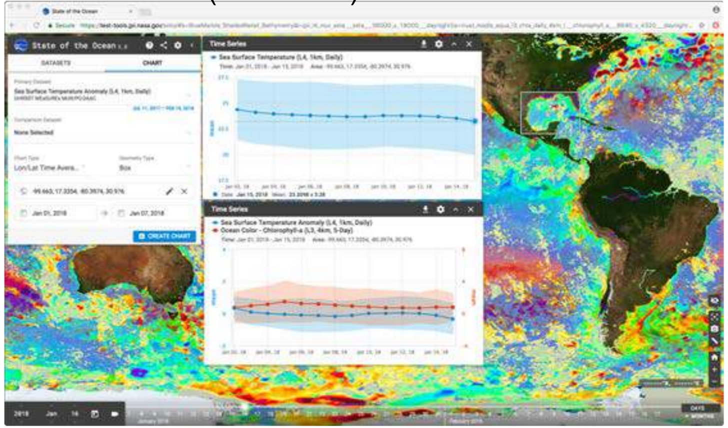
Big Data Analytics and Cloud Computing Workshop, 2017 ESIP Summer Meeting, Bloomington, IN

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

- The gap between visionary to pragmatists is significant. It must be the primary focus of any long-term high-tech marketing plan – Geoffrey Moore
- Become an expert in the production environment and devote resources in creating automations
- Give project engineering team early access to the PaaS
- Deliver all technical documents and work with project system engineering
- Provide user-focused trainings

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

NASA Sea Level Change Team



Jet Propulsion Laboratory
California Institute of Technology

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
- Focus on end-to-end data and computation architecture, and the total cost of ownership
- JPL Strategy is to drive Data Science into the fabric of JPL by
    - Launching cross-institution pilots
    - Building a trained workforce
    - Linking to the mission-science data lifecycle
- Invest in Interactive Analytics that simplifies the integration of *multiple* Earth observing remote sensing instruments; comparison against models
- Disruptive Innovations are products that require us to change our current mode of behavior or to modify other products and services – Geoffrey Moore
- AI and Data Science will be an essential part of NASA's future!

Thomas Huang
thomas.huang@jpl.nasa.gov
Jet Propulsion Laboratory
California Institute of Technology