National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Big Data Analytics and Sea Level Research

Thomas Huang
Data Scientist | Principal Investigator | Technologist | Architect
thomas.huang@jpl.nasa.gov

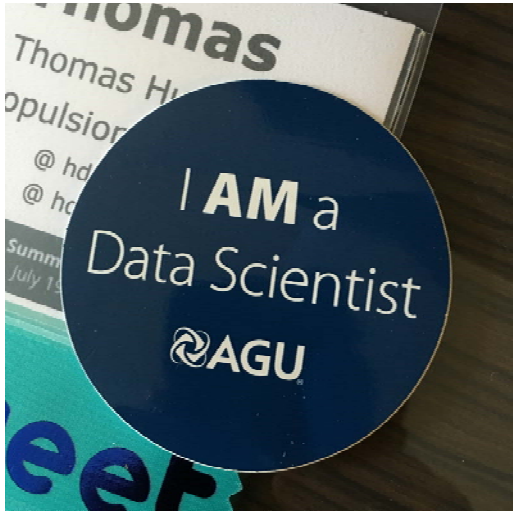
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data Scientist @NASA/JPL

- **Principal Investigator** for NASA AIST OceanWorks
- **Project Technologist** for the NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC) – http://podaac.jpl.nasa.gov
- **Co- Investigator and Architect** for the NASA Sea Level Change Portal – https://sealevel.nasa.gov
- **Architect** for Tactical Data Science Framework for Naval Research
- **Chair** for The Federation of Earth Science Information Partners (ESIP) Cloud Computing Cluster
- **Previously Principal Investigator / Co-Investigator** in several NASA-funded Big Data Analytic Projects
  - OceanXtremes: Oceanographic Data-Intensive Anomaly Detection and Analysis Portal – https://oceanxtremes.jpl.nasa.gov
  - Distributed Oceanographic Matchup Service (DOMS) – https://doms.jpl.nasa.gov
  - Mining and Utilizing Dataset Relevancy from Oceanographic Datasets (MUDROD)
  - Enhanced Quality Screening for Earth Science Data – https://vqss.jpl.nasa.gov
  - NEXUS - Big Data Analytic on the Cloud



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NASA Sea Level Change Portal – *https://sealevel.nasa.gov*

Goal for the NASA Sea Level Change Team

- Determine how much will sea level rise by [2100]?
- What are the key sensitivities?
- Where are the key uncertainties? Observables? Model Improvements

Goals for the NASA Sea Level Change Portal

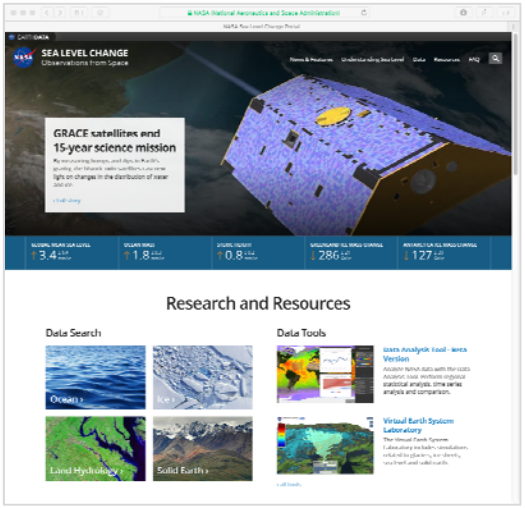
- Provide scientists and the general public with a “one-stop” source for current sea level change information and data
- Provide interactive tools for analyzing and viewing regional data
- Provide virtual dashboard for sea level indicators
- Provide latest news, quarterly report, and publications
- Provide ongoing updates through a suite of editorial products

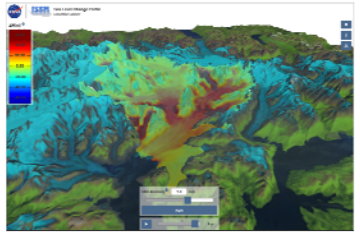
Requires

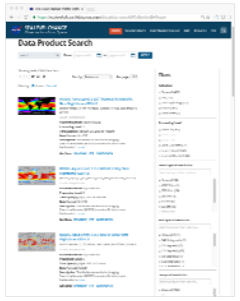
- Interdisciplinary collaboration
- Connect disciplines and evaluate dependencies

Sea Level Change Portal facilitates

- Easy interdisciplinary data comparison
- Access to latest news and information
- Collaboration (data and information exchange)



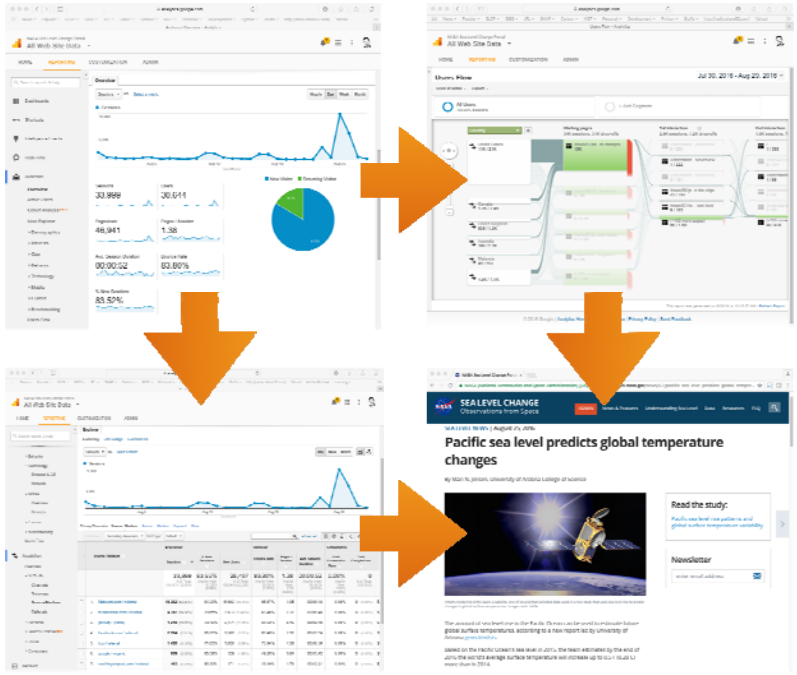




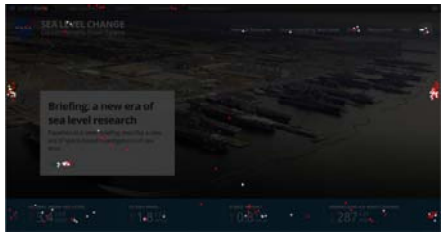
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Analyze User Interactions

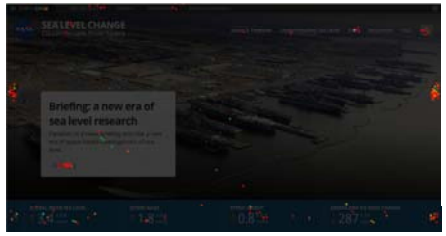
- Guide website layout
- Determine effectiveness of articles and contents
- Identify popular media outlines
- New and returning users

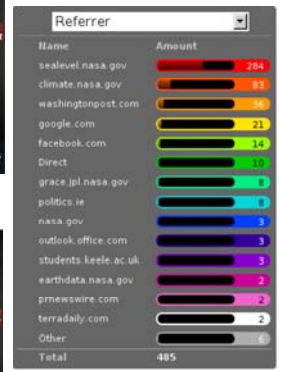


New vs. Returning

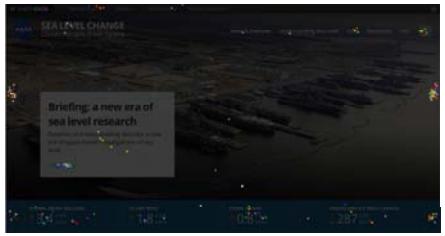


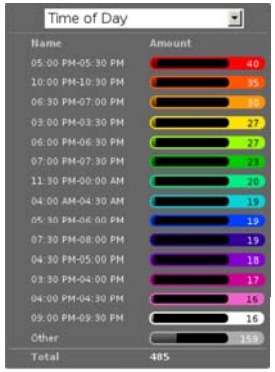
Referrer (Media Outlets)





Time to Click





National Aeronautics and Space Administration

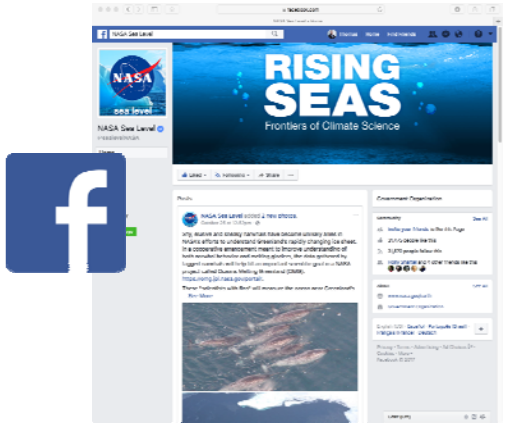
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Web, Social Medias, and Headliners

- 373K monthly page views
- 172K sessions
- 143K users
- **Social Medias**

  Twitter: @NASASeaLevel has over 23K followers

  Facebook: over 31K followers



TECH HEADLINES

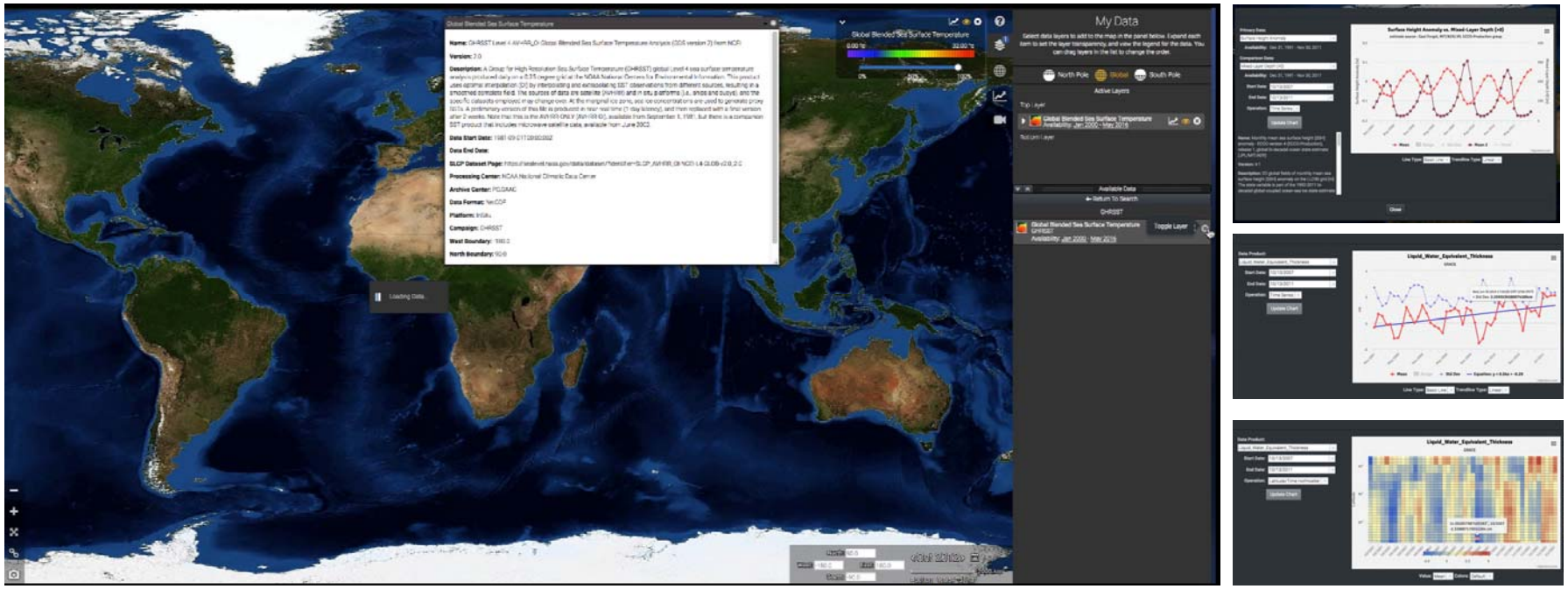
"NASA Sea Level Change Website Offers Everything You Need to Know About Climate Change"
http://www.techtimes.com/articles/147210/20160405/nasa-sea-level-change-website-offers-everything-need-know-climate.htm

"NASA's New Sea Level Site Puts Climate Change Papers, Data, and Tools Online"
http://techcrunch.com/2016/04/04/nasas-new-sea-level-site-puts-climate-change-papers-data-and-tools-online/

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Analyze Sea Level On-The-Fly *https://sealevel.nasa.gov*



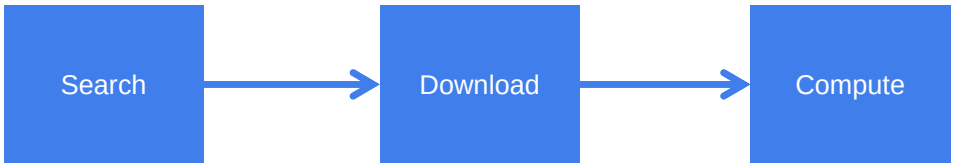
Sea Level Change - Data Analysis Tool

Visualizations | Hydrological Basins | Time Series | Deseason | Data Comparison | Scatter Plot | Latitude/Time Hovmöller | Etc.

Big Data and Data Centers

- **Increasing "big data" era is driving needs to**
  - Scale computational and data infrastructures
  - Support new methods for deriving scientific inferences
  - Shift towards integrated data analytics
  - Apply computation and data science across the lifecycle
- **For NASA Data Centers, with large amount of observational and modeling data, downloading to local machine 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. Too many matches
  - Finding the relevant measurement has becoming a very time consuming process "*Which SST dataset I should use?*"
  - Analyze decades of regional measurement is labor-intensive and costly
- **Limitations**
  - Little to no interoperability between tools and services: metadata standard, keyword, spatial coverage (0-360 or -180..180), temporal representation, etc.
  - Making sure the most relevant measurements return first
  - Visualization is nice, but it doesn't provide enough information about the event/phenomenon captured in the image.
  - With large amount of observational data, data centers need to do more than just storing bits

Traditional Data Analysis



- 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

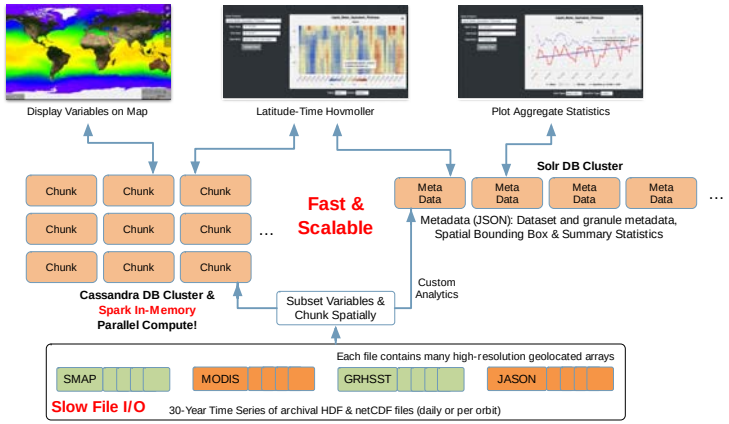
- Traditional methods for data analysis (time-series, distribution, climatology generation) can't scale to handle large volume, high-resolution data. They perform poorly
- Performance suffers when involve large files and/or large collection of files
- A high-performance data analysis solution must be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

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

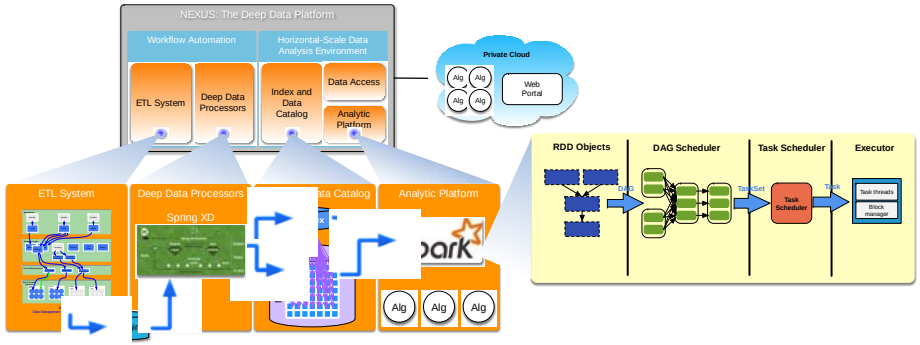
NEXUS: Scalable Data Analytic Solution

- NEXUS is a data-intensive analysis solution using a new approach for handling science data to enable large-scale data analysis
- Streaming architecture for horizontal scale data ingestion
- Scales horizontally to handle massive amount of data in parallel
- Provides high-performance geospatial and indexed search solution
- Provides tiled data storage architecture to eliminate file I/O overhead
- A growing collection of science analysis webservices using Apache Spark: parallel compute, in-memory map-reduce framework
- Pre-Chunk and Summarize Key Variables
  - Easy statistics instantly (milliseconds)
  - Harder statistics on-demand using Spark (in seconds)
  - Visualize original data (layers) on a map quickly (Cassandra store)
- **Algorithms** – Time Series | Latitude/Time Hovmöller| Longitude/Time Hovmöller| Latitude/Longitude Time Average | Area Averaged Time Series | Time Averaged Map | Climatological Map | Correlation Map | Daily Difference Average

Open Source: Apache License 2
https://github.com/dataplumber/nexus



Two-Database Architecture



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

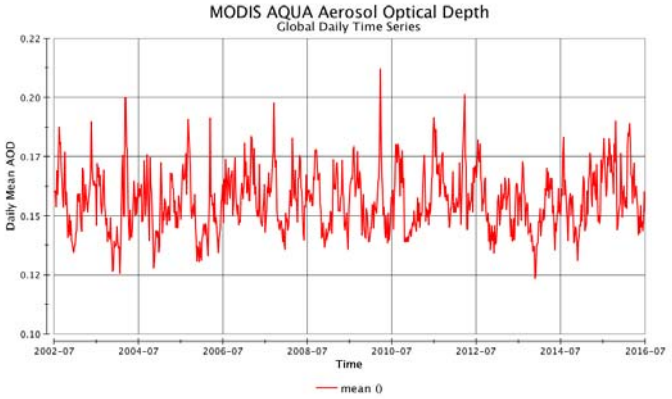
NEXUS Performance: Custom Spark vs. AWS EMR

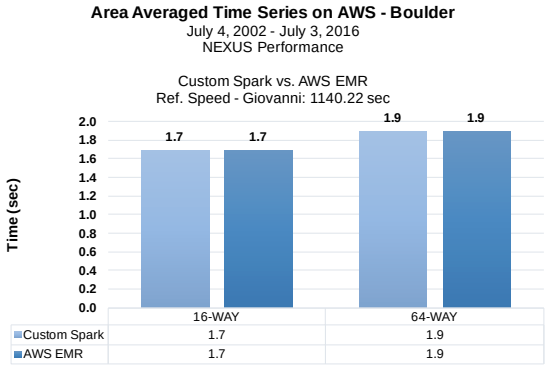
Dataset: MODIS AQUA Daily
Name: Aerosol Optical Depth 550 nm (Dark Target) (MYD08_D3v6)
File Count: 5106
Volume: 2.6GB
Time Coverage: July 4, 2002 – July 3, 2016

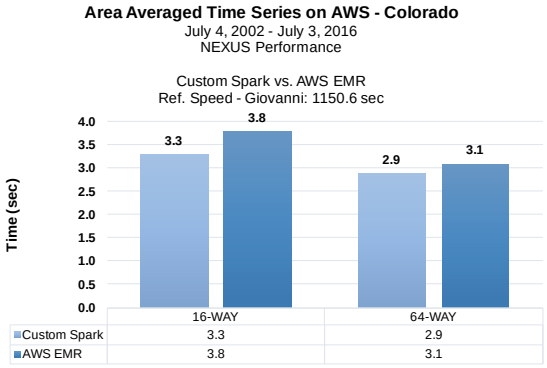
Giovanni: A web-based application for visualize, analyze, and access vast amounts of Earth science remote sensing data without having to download the data.

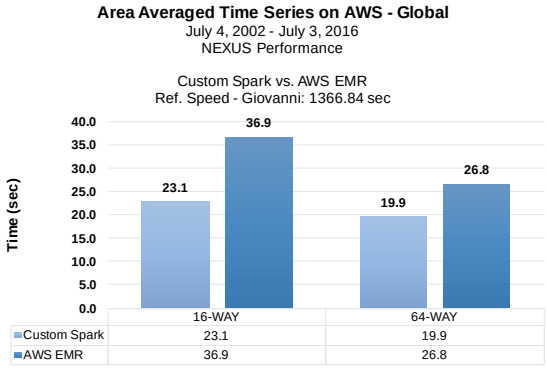
- Represents current state of data analysis technology, by processing one file at a time
- Backed by the popular NCO library. Highly optimized C/C++ library

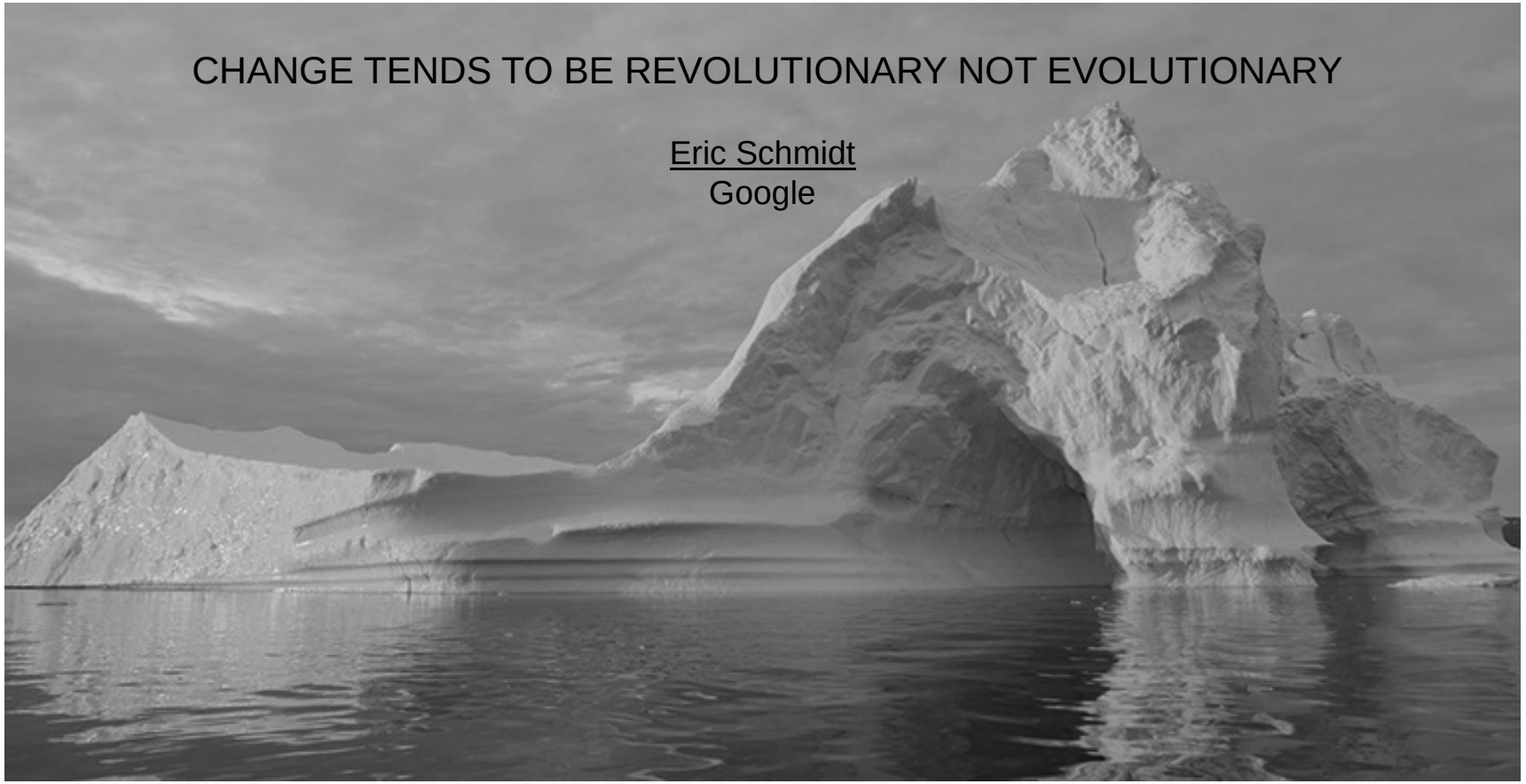
AWS EMR: Amazon's provisioned MapReduce cluster

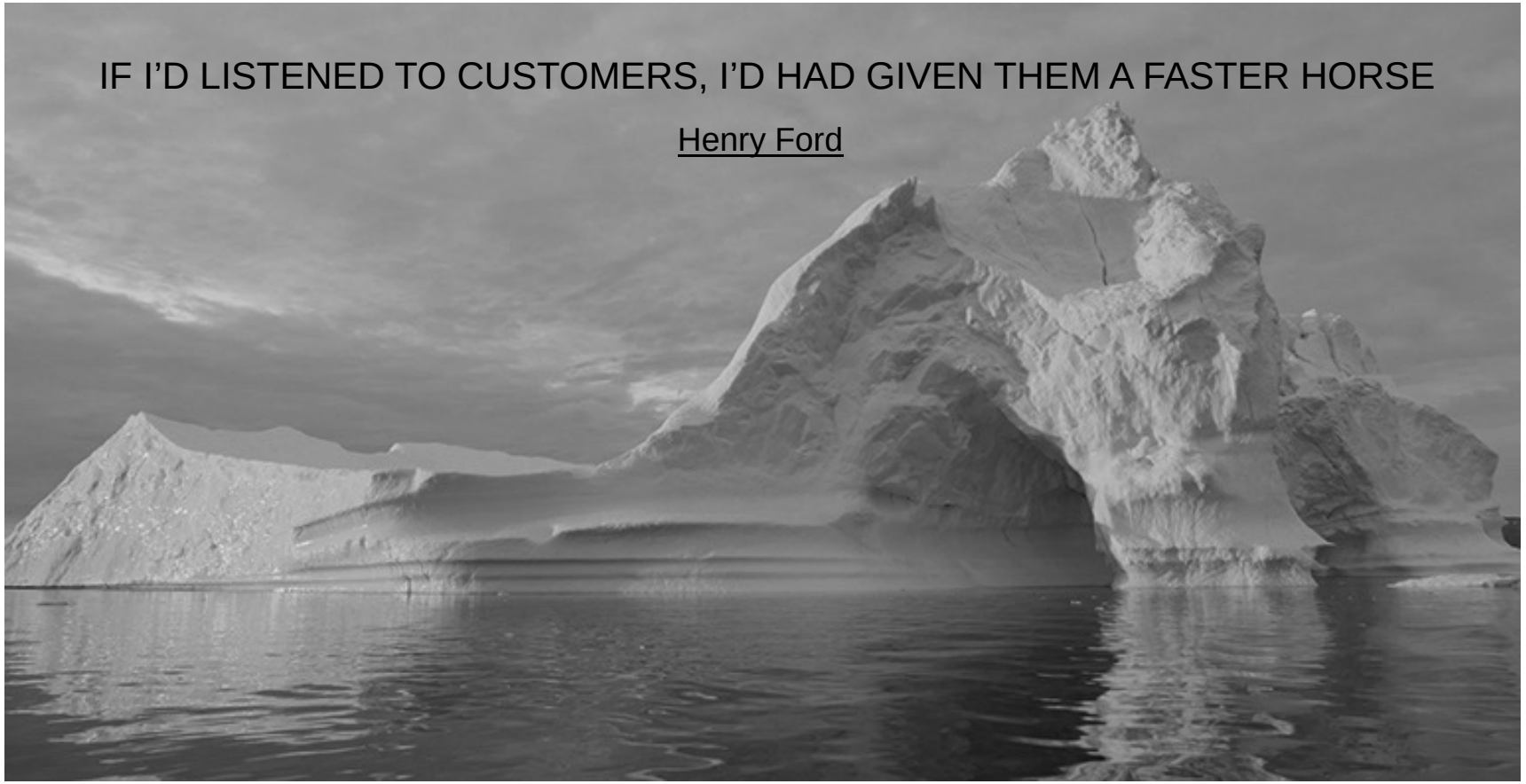








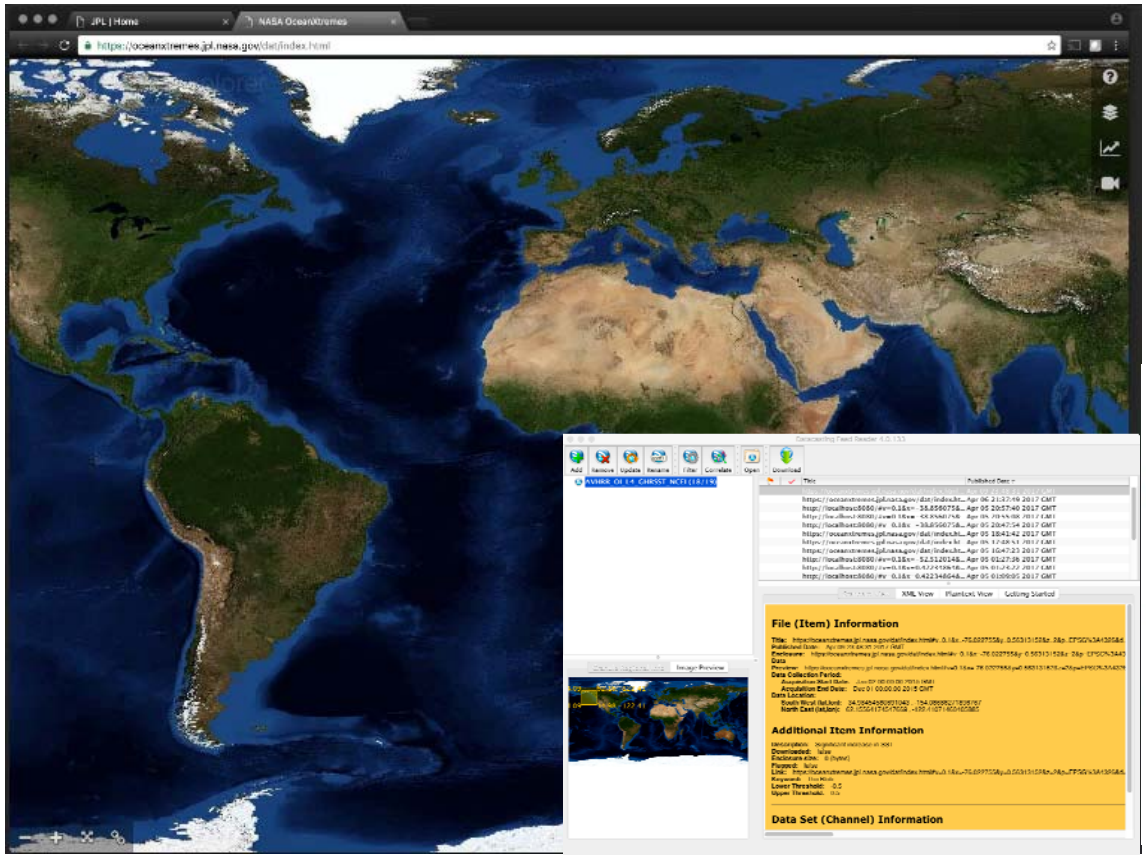




National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

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

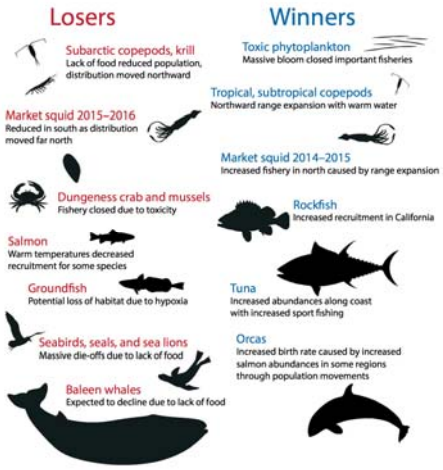
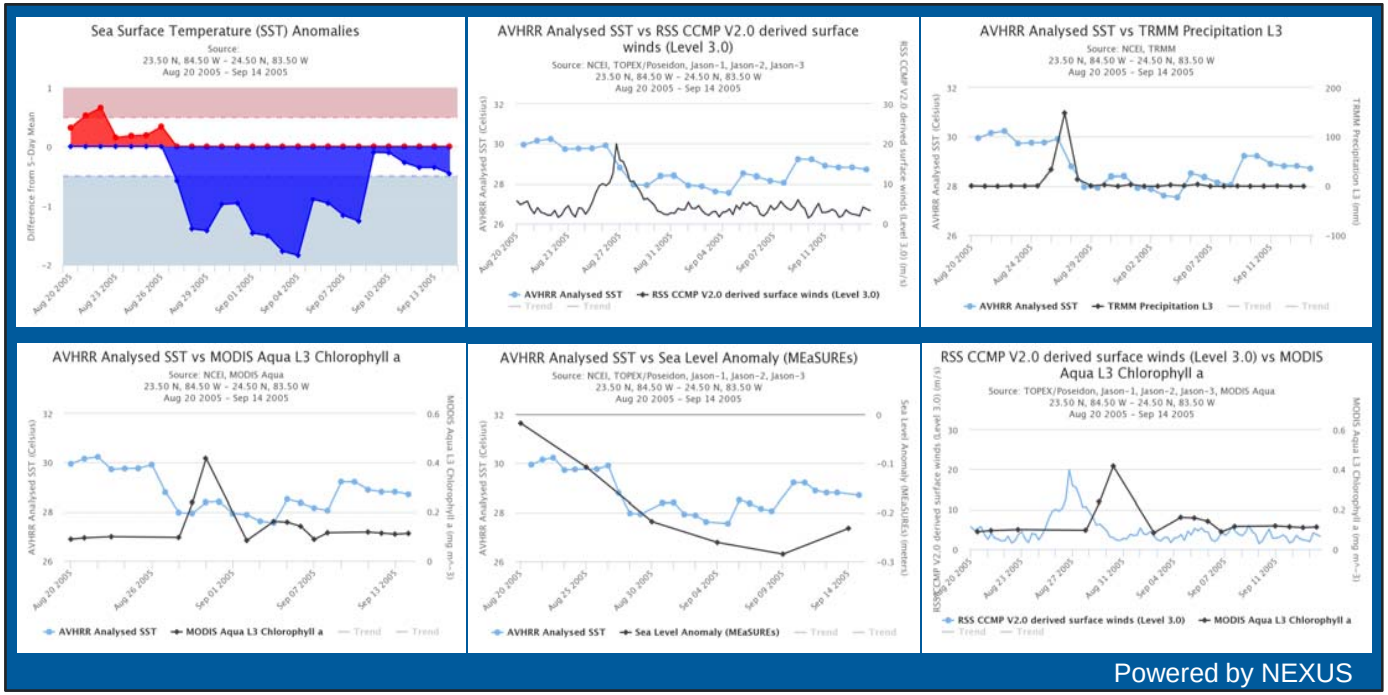


Figure from Cavole, L. M., et al. (2016). "Biological Impacts of the 2013–2015 Warm-Water Anomaly in the Northeast Pacific: Winners, Losers, and the Future." Oceanography 29.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

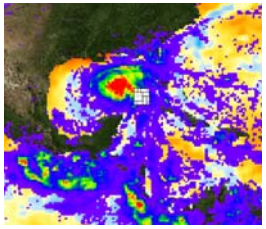
Hurricane Katrina Study



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.

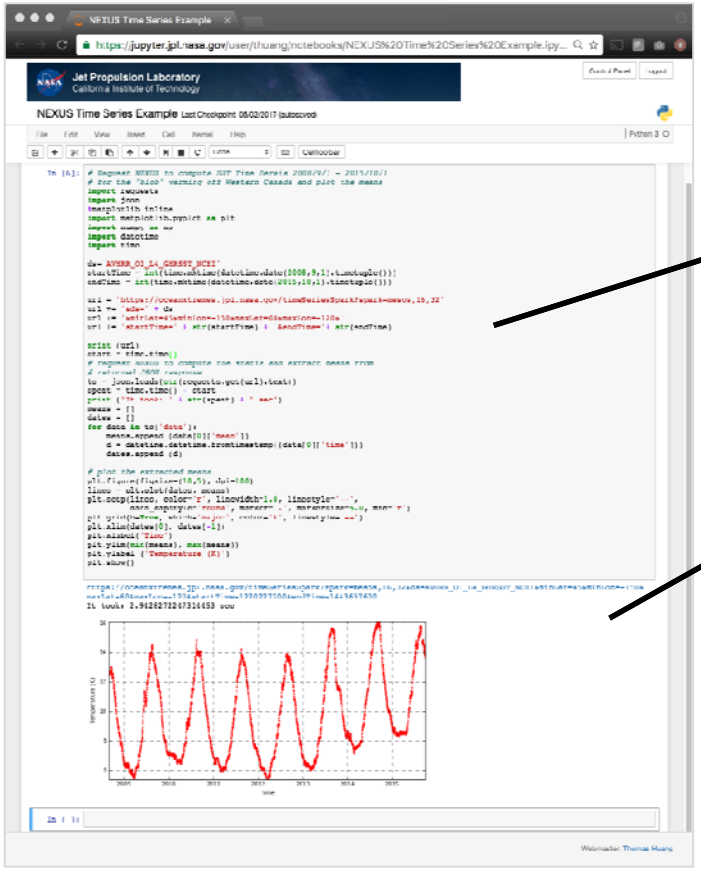


Hurricane Katrina TRMM overlay SST Anomaly

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Enable Science without File Download




```
# Request NEXUS to compute SST Time Series 2008/9/1 - 2015/10/1
# for the "blob" warming off Western Canada and plot the means
…
ds='AVHRR_OI_L4_GHRSST_NCEI'

url = … # construct the webservice URL request


# make request to NEXUS using URL request
# save JSON response in local variable
ts = json.loads(str(requests.get(url).text))

# extract dates and means from the response
means = []
dates = []
for data in ts['data']:
    means.append (data[0]['mean'])
    d = datetime.datetime.fromtimestamp((data[0]['time']))
    dates.append (d)

# plot the result
…
```

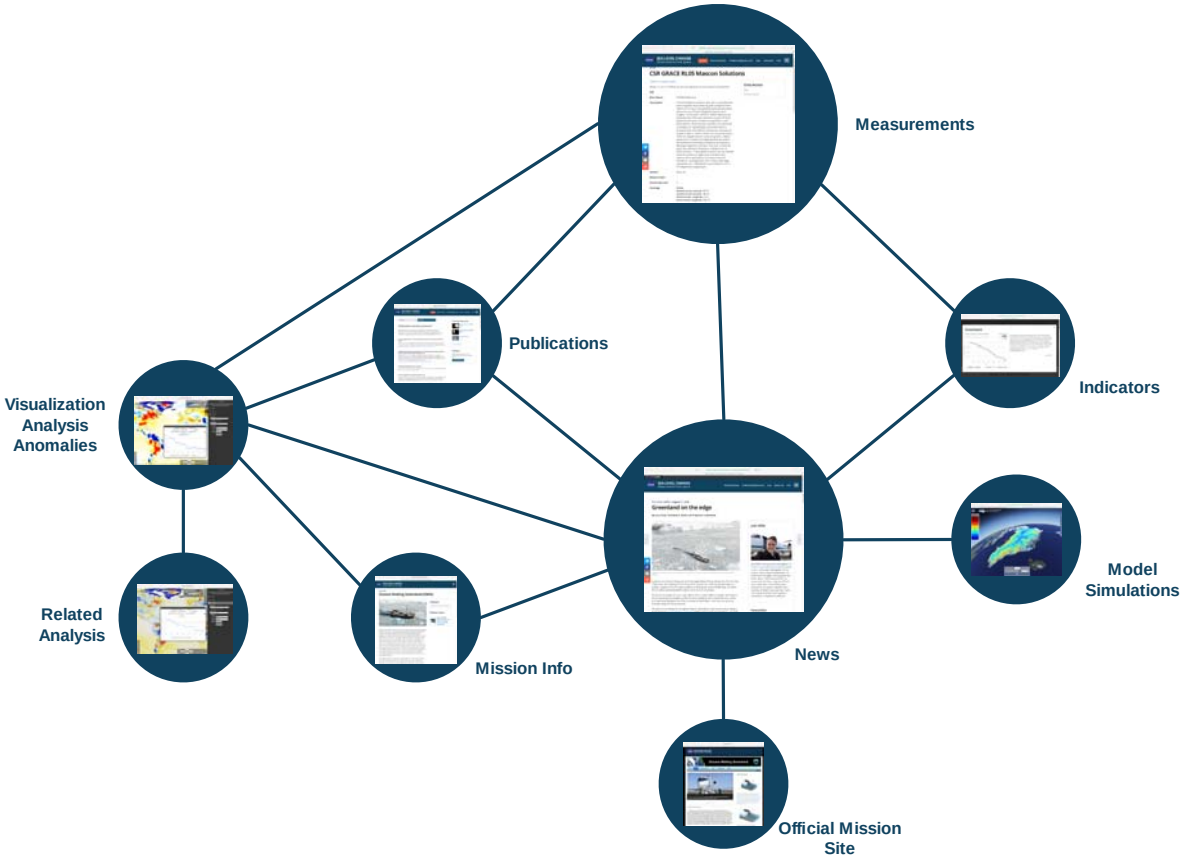
```
https://oceanxtremes.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GHRSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=1220227200&endTime=1443657600

It took: 2.9428272247314453 sec
```

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

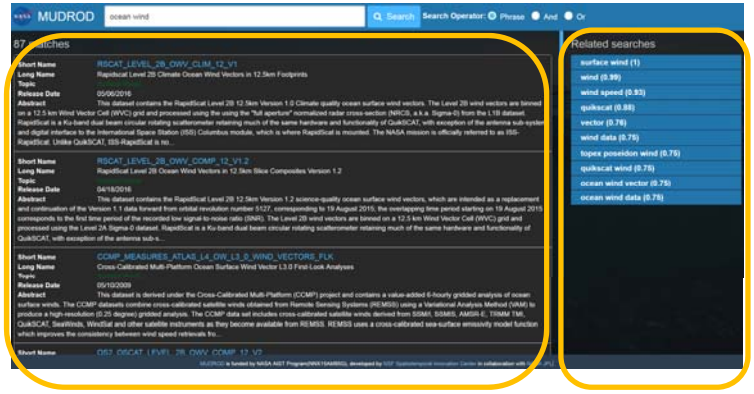
Developing Information Discovery Solutions



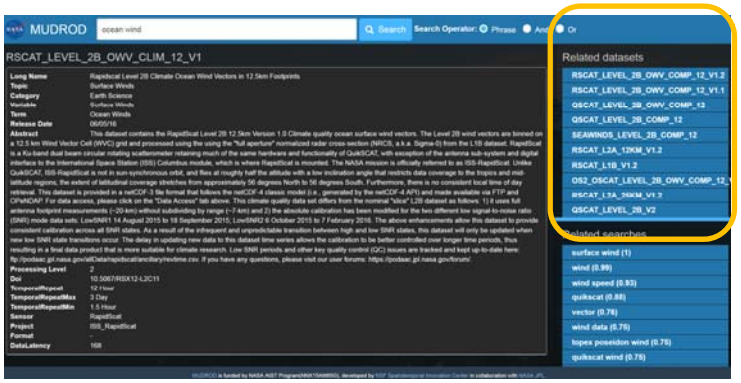
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Search and Discovery

- **Search** – look for something you expect to exist
  - Information tagging
  - Indexed search technologies like Apache Solr or ElasticSearch
  - The solution is pretty straightforward
- **Discovery** – find something new, or in a new way
  - This is non-trivial
  - Traditional ontological method doesn't quite add up
  - The strength of semantic web is in inference
  - What happen when we have a lot of `subClassOf`, `equivalentClassOf`, `sameAs`?
  - How wide and deep should we go?
- **Relevancy**
  - It is domain-specific
  - It is personal
  - It is temporal
  - It is dynamic



Search Ranking
Based on a machine learning model (RankSVM) which takes a number of features, such as vector space model, version, processing level, release date, all-time popularity, monthly-popularity, and user popularity.



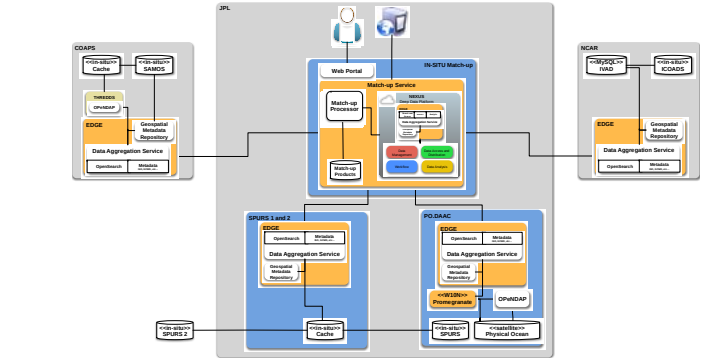
Search Recommendation
Based on dataset metadata content and web session co-occurrence

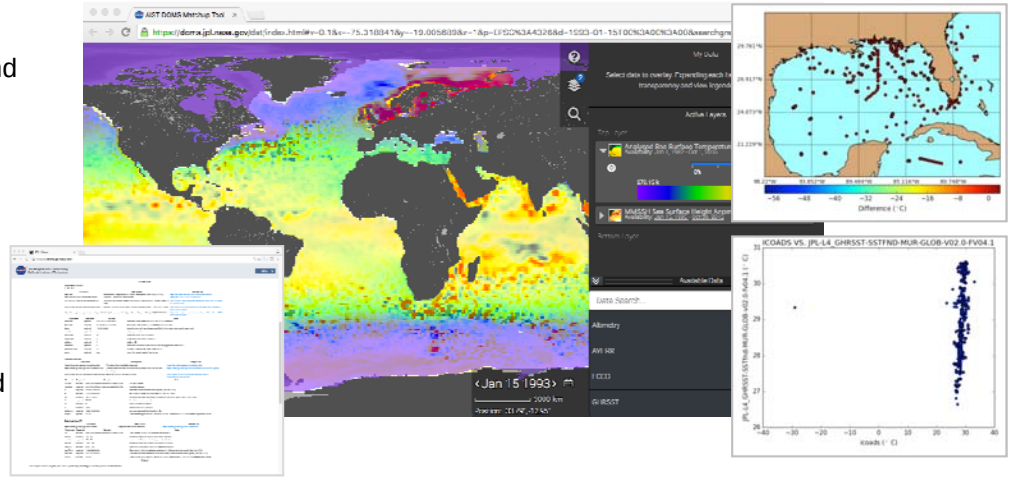
National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

In Situ to Satellite Matchup

- Distributed Oceanographic Matchup Service
- Typically data matching is done using one-off programs developed at multiple institutions
- A primary advantage of DOMS 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.
- In situ data nodes at JPL, NCAR, and FSU operational.
- Provides data querying, subset creation, match-up services, and file delivery operational.
- Prototype graphical user interface (UI) and APIs accessible for external users.
- Plugin architecture for in situ data source using EDGE
  - Extensible Data Gateway Environment is an Apache License 2 open source technology
  - https://github.com/dataplumber/edge
- Defined specification for packaging matchup results. Working with Unidata and ESDSWG's data interoperability and standard groups



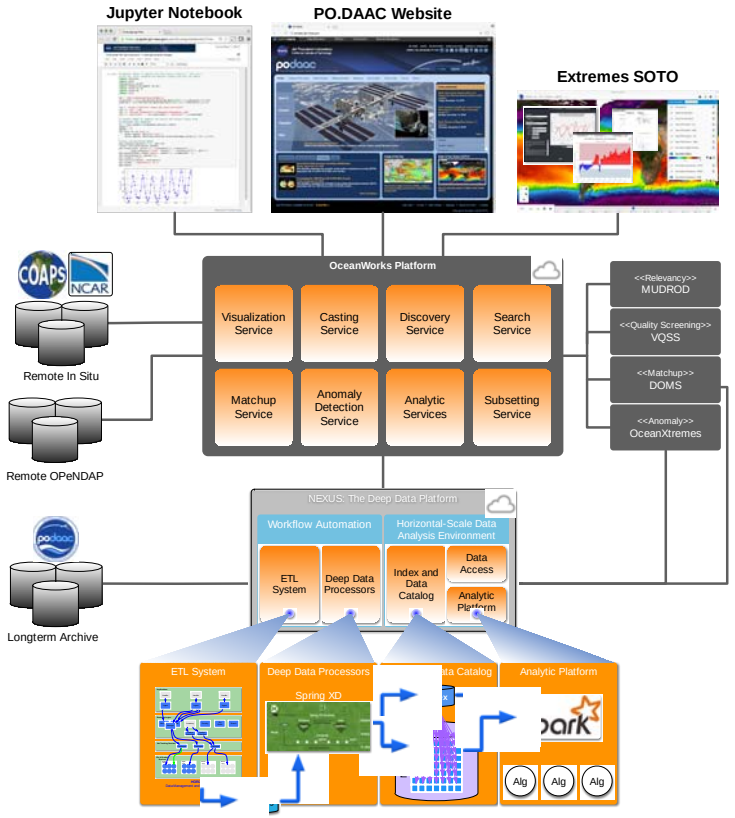


National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

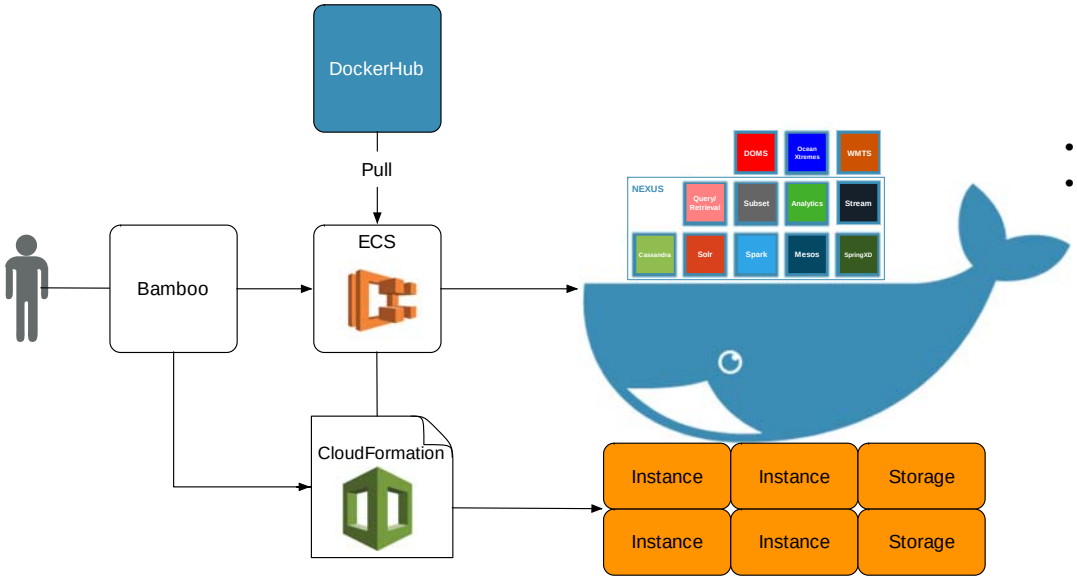
AIST OceanWorks

- **OceanWorks** is to establish an **Integrated Data Analytic Center** at the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) for Big Ocean Science
- Focuses on technology integration, advancement and maturity
- Collaboration between JPL, FSU, NCAR, and GMU
- Bringing together PO.DAAC-related big data technologies
  - **OceanXtremes –** Anomaly detection and ocean science
  - **NEXUS –** Big data analytic platform
  - **Data Container Studies**
  - **DOMS –** Distributed in-situ to satellite matchup
  - **MUDROD –** Search relevancy and discovery – linking datasets, services, and anomalies through recommendations
  - **VQSS –** Virtualized Quality Screening Service



Deployment Automation

- Template-based infrastructure deployment and provision
- Container-based system deployment
- Automate allocation of EC2 instances and storages
- Turnkey deployment of computing clusters
- Automate multi-container deployment
- Enable rapid Cloud deployment



- Docker all services
- **Why?**
  - Rapid application deployment
  - Portability across machines
  - Application-centric vs machine/server-centric
  - Version control and component reuse
  - Secure due to isolation and encapsulation
  - Sharing
  - Lightweight footprint
  - Minimal overhead
  - Simplified maintenance

Open Source

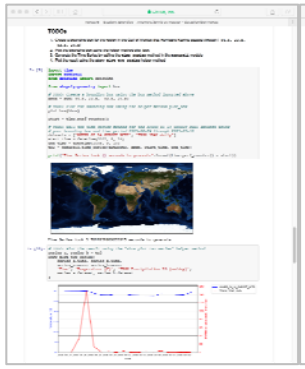
- Technology sharing through Free and Open Source Software (FOSS)
- Further technology evolution that is restricted by projects / missions
- **Science Data Analytic Platform (SDAP)**, the implementation of **OceanWorks**, in **Apache Incubator**
  - Cloud platform
  - Analyzing satellite and model data
  - In situ data analysis and coordination with satellite measurements
  - Fast data subsetting
  - Mining of user interactions and data to enable discovery and recommendations
  - Streamline deployment through container technology

http://sdap.incubator.apache.org

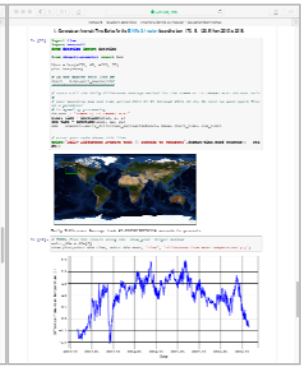
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Community Engagement and Support

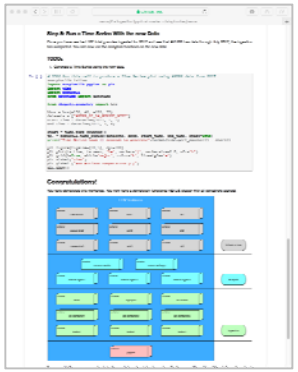
- Develop in the open
- Working with Apache Incubator
- Target Apache top-level project by 2019.
- Public hands-on workshops
- Organize technical sessions at conferences
- Invited speaker and panelist
- Lead Editor: 2018 Wiley Book on **Big Earth Data Analytics in Earth, Atmospheric and Ocean Sciences**



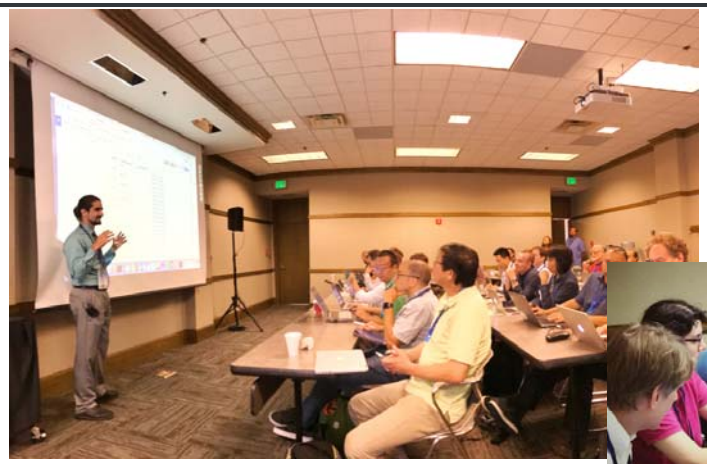
Analyze Hurricane Katrina by comparing SST and TRMM time series

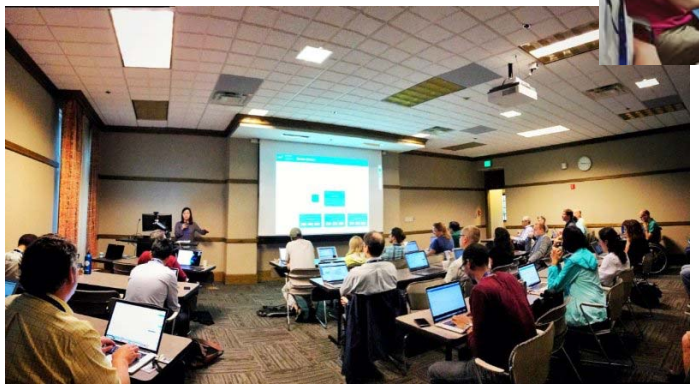


Generate daily difference average
"The Blob" is an oceanographic anomaly



Each participant deployed 3 computing clusters, a total of 24 containers on EC2

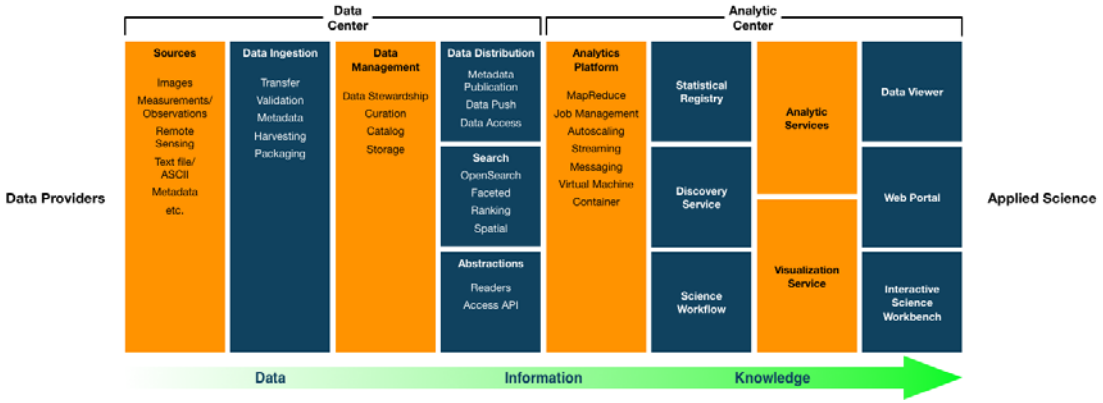


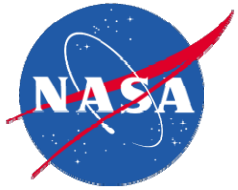


In Summary

- Traditional method for scientific research (search, download, local number crunching) is unable to keep up
- Think beyond the archive
- Connected information enables discovery
- Community developed solution through open sourcing
- Thanks to the NASA ESTO/AIST and Sea Level Rise programs, and the NASA ESDIS project
- Investment in data and computational sciences
- Data Centers need to be in the business of Enabling Science!
- OceanWorks infusion 2018 – 2019
  - Watch for changes to the Sea Level Change Portal
    - Even faster analysis capabilities
    - More variety of measurements – satellites, in situ, and models
    - Event more relevant recommendations
  - NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)

Transforming Data to Knowledge





National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Thomas Huang
Jet Propulsion Laboratory
California Institute of Technology

"Without counsel plans fail, but with many advisers they succeed." – Proverbs 15:22