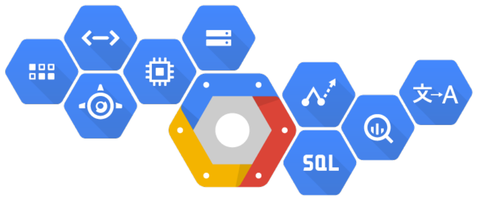
Big data in Earth science: Using Pangeo to work withOOI data in the Cloud

Tim Crone
Rich Signel, Dax Soule, Filipe Pires Alvarenga Fernandes

15 December 2021

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE



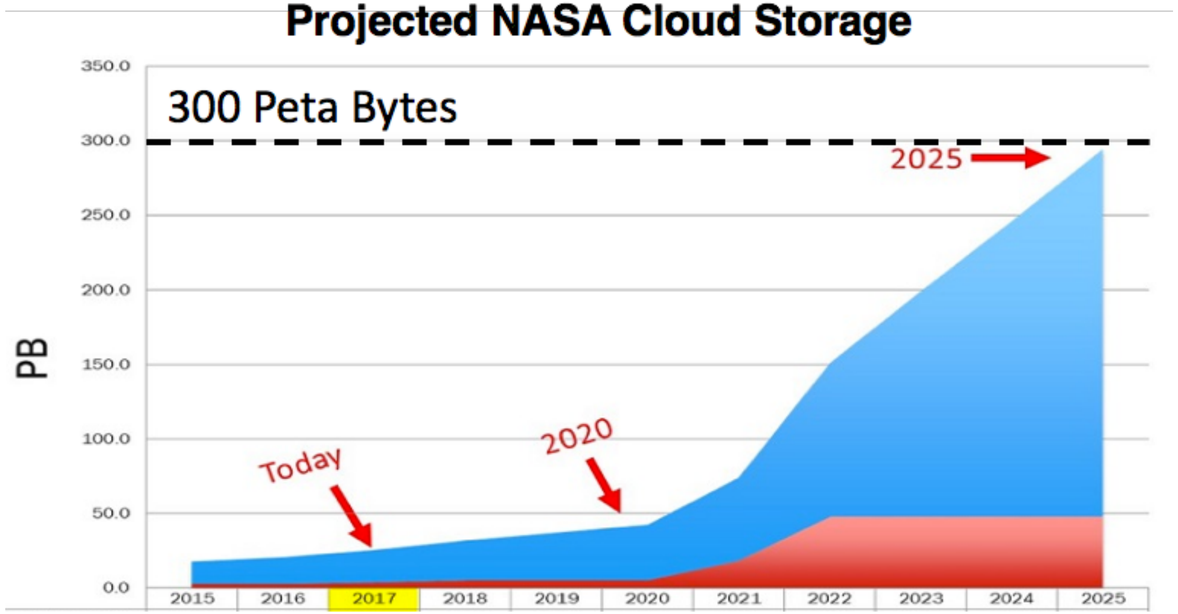
PANGEO

Big data in the geosciences

Increase in Earth Observations

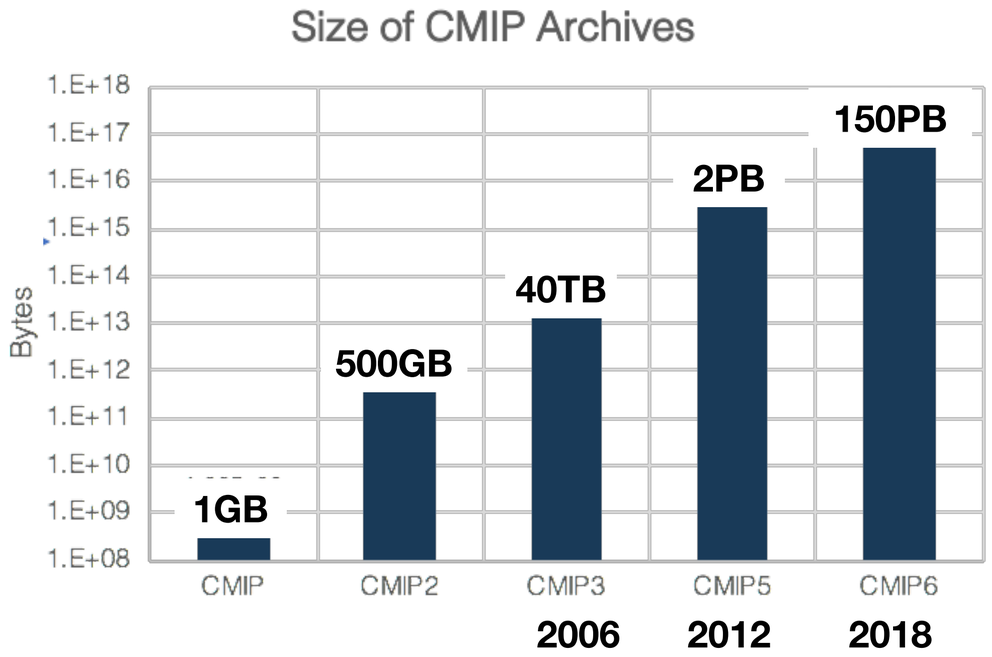
New sensors / platforms
Continuous observations
Multiple versions of derived datasets

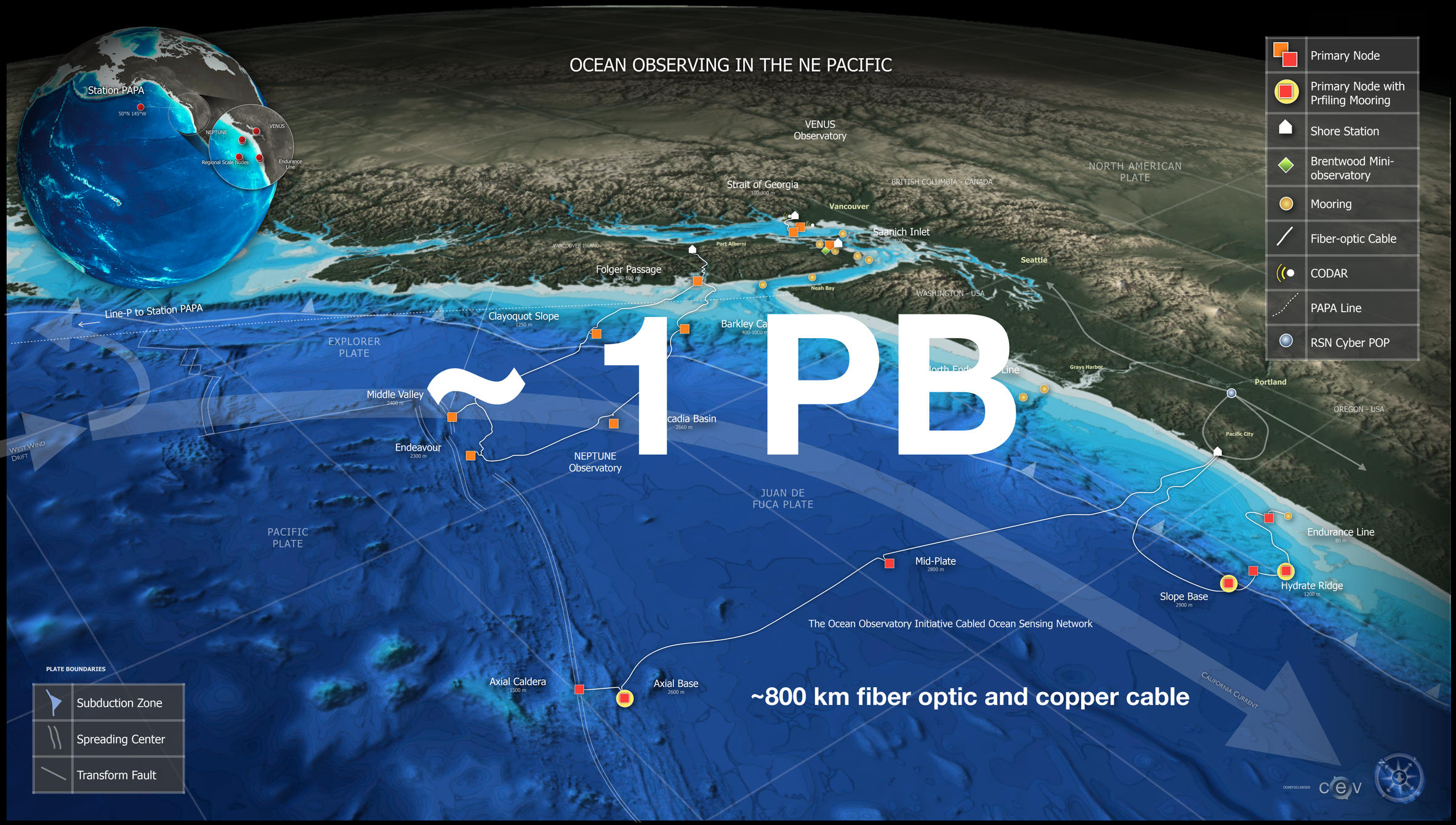
https://earthdata.nasa.gov/about/eosdis-cloud-evolution



Increase in Model Data

Higher resolution
More process representation
Larger ensembles

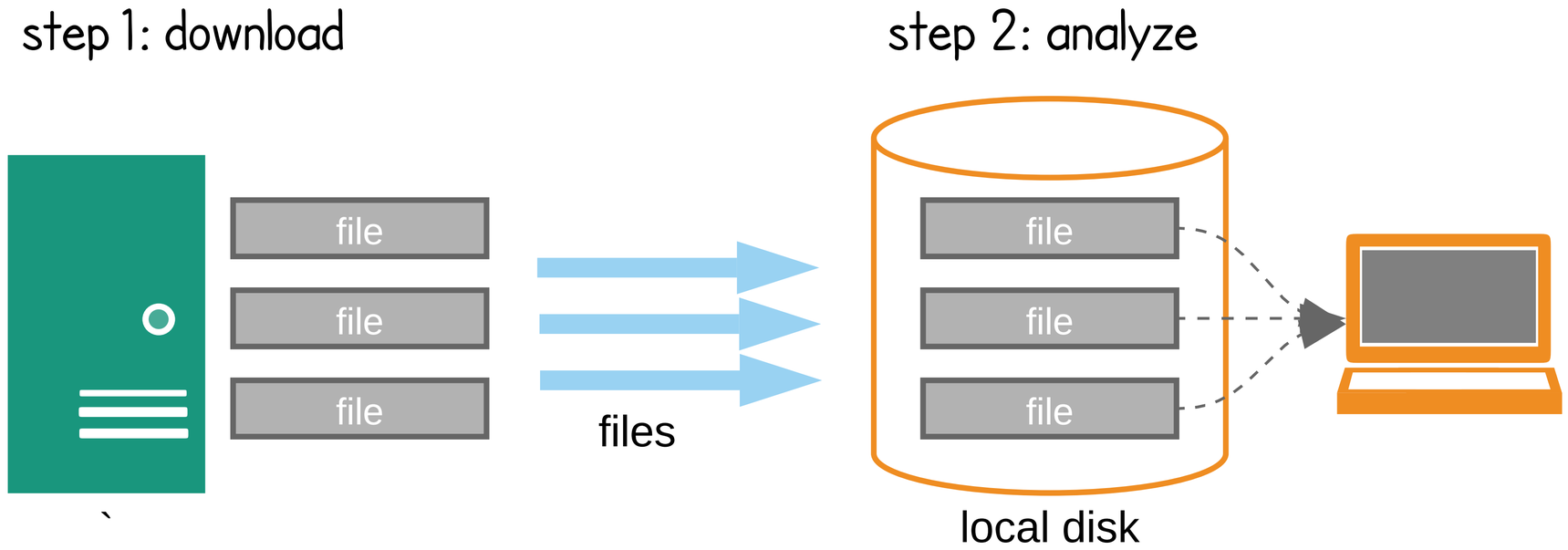



OCEAN OBSERVING IN THE NE PACIFIC
Station PAPA
50°N 145°W
NEPTUNE
VENUS
Regional Scale Nodes
Endurance Line
VENUS Observatory
Strait of Georgia
100-300 m
BRITISH COLUMBIA - CANADA
NORTH AMERICAN PLATE
Vancouver
Saanich Inlet
100 m
VANCOUVER ISLAND
Port Alberni
Folger Passage
20-100 m
Seattle
Neah Bay
WASHINGTON - USA
Line-P to Station PAPA
Clayoquot Slope
1250 m
Barkley Ca
400-1000 m
EXPLORER PLATE
~1 PB
Grays Harbor
Portland
OREGON - USA
Middle Valley
2400 m
cadia Basin
2660 m
Pacific City
Endeavour
2300 m
NEPTUNE Observatory
WEST WIND DRIFT
JUAN DE FUCA PLATE
PACIFIC PLATE
Endurance Line
80 m
Mid-Plate
2800 m
Hydrate Ridge
1200 m
Slope Base
2900 m
The Ocean Observatory Initiative Cabled Ocean Sensing Network
Axial Caldera
1500 m
Axial Base
2600 m
~800 km fiber optic and copper cable
CALIFORNIA CURRENT
Primary Node
Primary Node with Prfiling Mooring
Shore Station
Brentwood Mini-observatory
Mooring
Fiber-optic Cable
CODAR
PAPA Line
RSN Cyber POP
PLATE BOUNDARIES
Subduction Zone
Spreading Center
Transform Fault
OOINEP20130509
cev

How should scientific data analysis infrastructure be organized for the petabyte era?

PANGEO

File-based Approach



Data provider's responsibilities

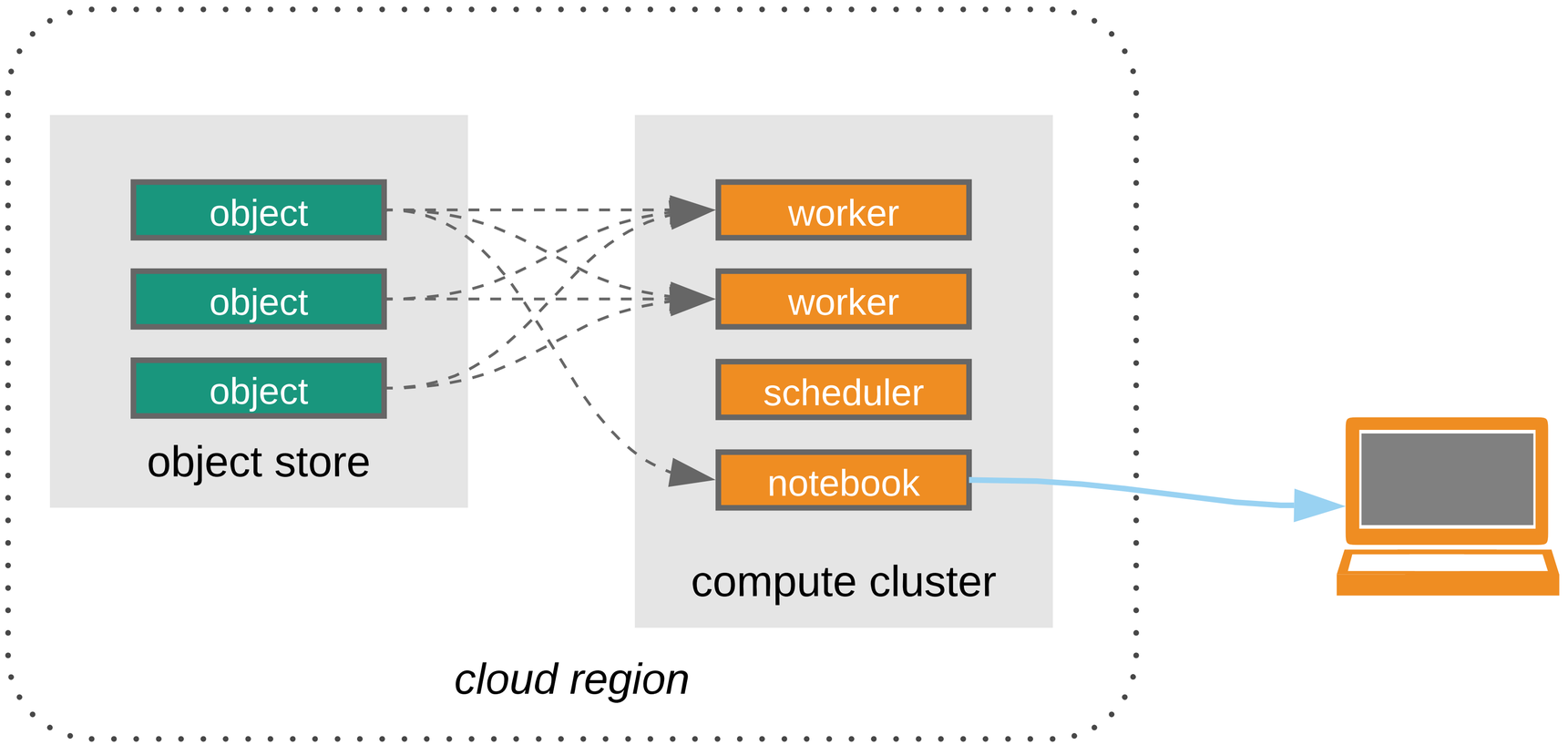
End user's responsibilities

Server-Side Database

Data provider's responsibilities

End user's responsibilities

Cloud-Native Approach



Data provider's responsibilities

End user's responsibilities

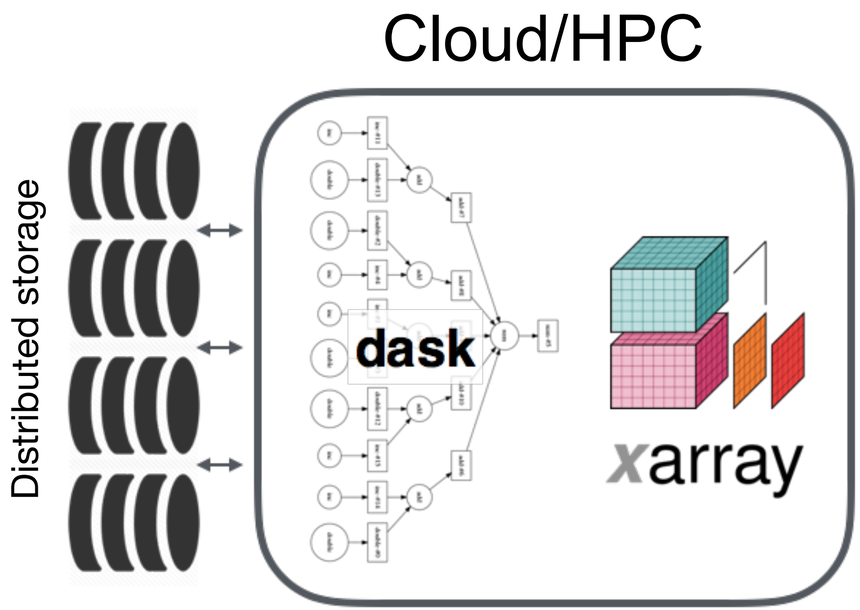
Pangeo

A *community-driven* effort for
Big Data geoscience

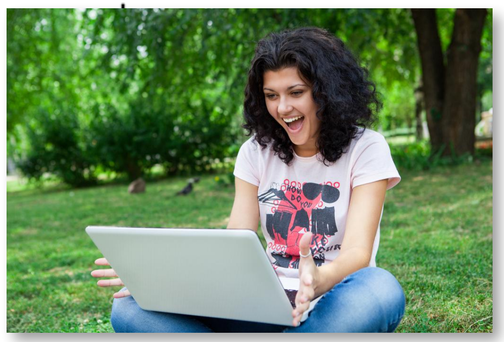
PANGEO

The Pangeo Architecture

Analysis Ready Data
Stored and cataloged on globally-available distributed storage (**e.g. S3, GCS**)



Jupyter for interactive access on remote systems

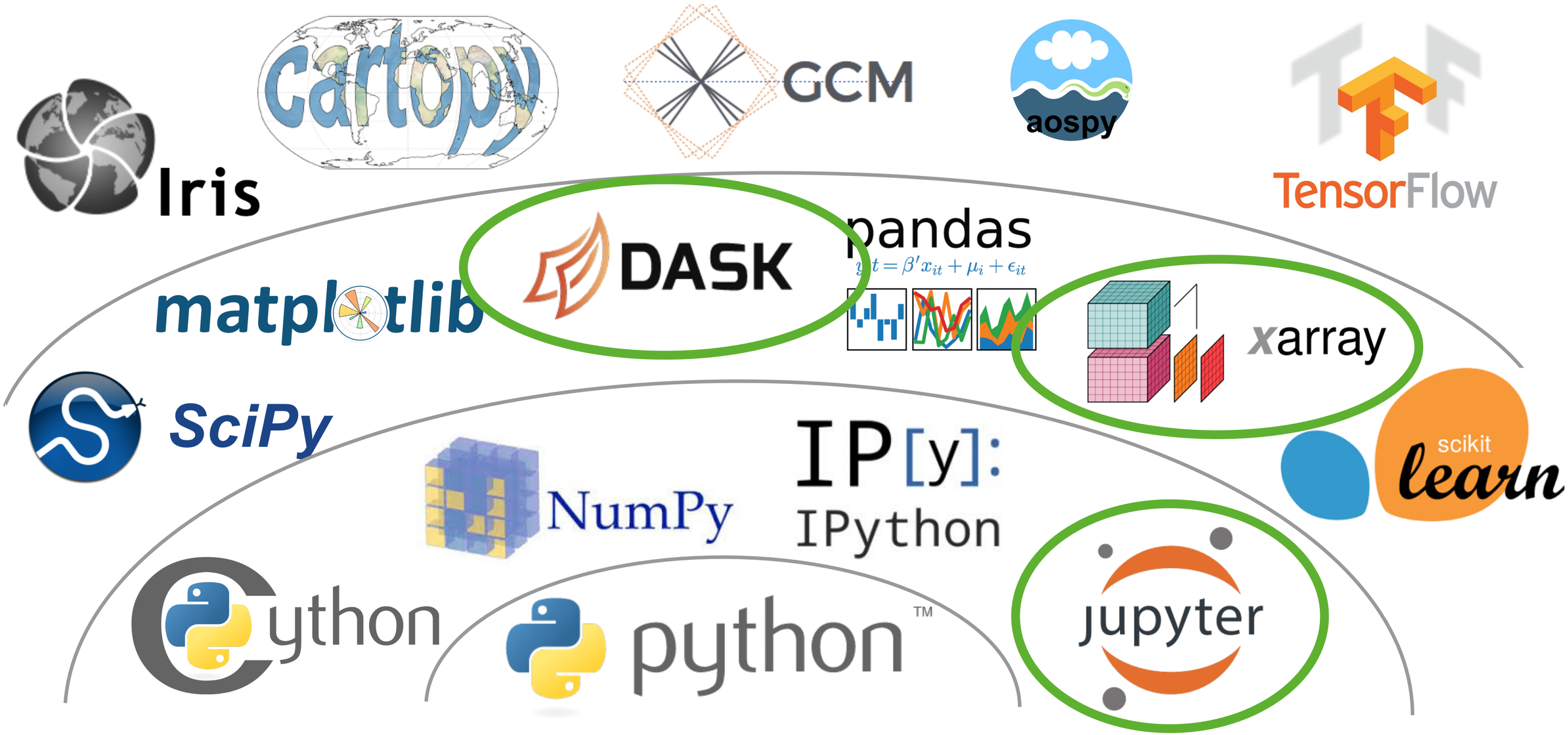


Parallel computing system built on top of **Kubernetes or HPC**.

Dask tells the nodes what to do.

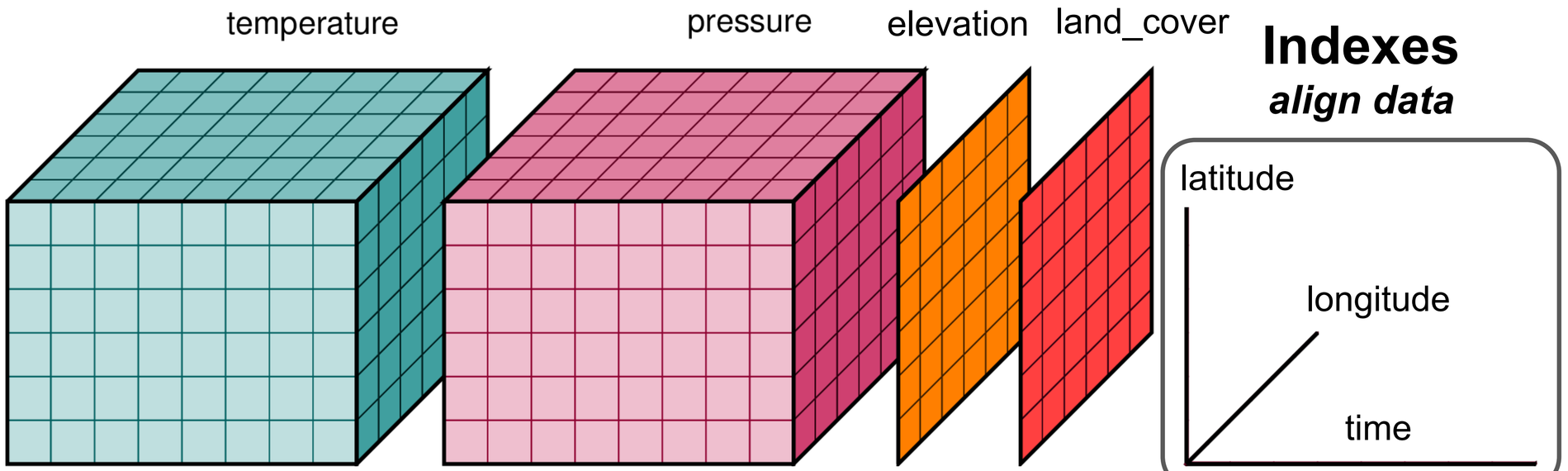
Xarray provides data structures and intuitive interface for interacting with datasets

Scientific Python for Geoscience



Credit: Stephan Hoyer, Jake Vanderplas (SciPy 2015)

Xarray: Multidimensional variables with coordinates and metadata



Indexes
align data

+ Attributes
metadata ignored by operations

Data variables
used for computation

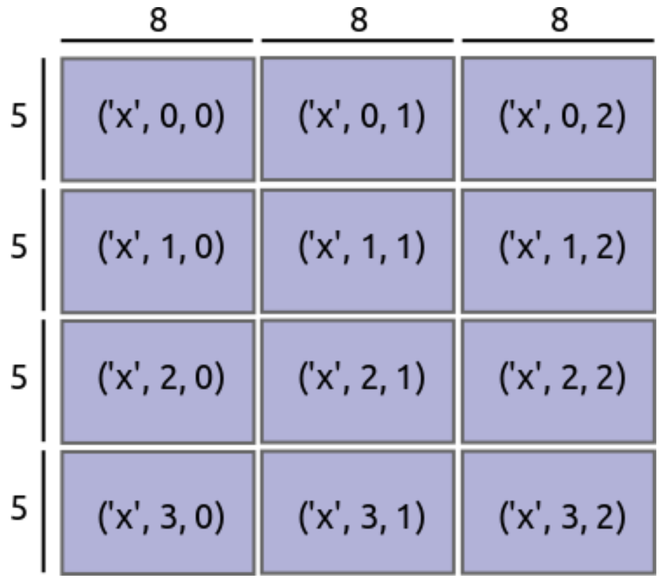
Coordinates
describe data

"netCDF meets pandas.DataFrame" plus! Dask integration

Credit: Stephan Hoyer

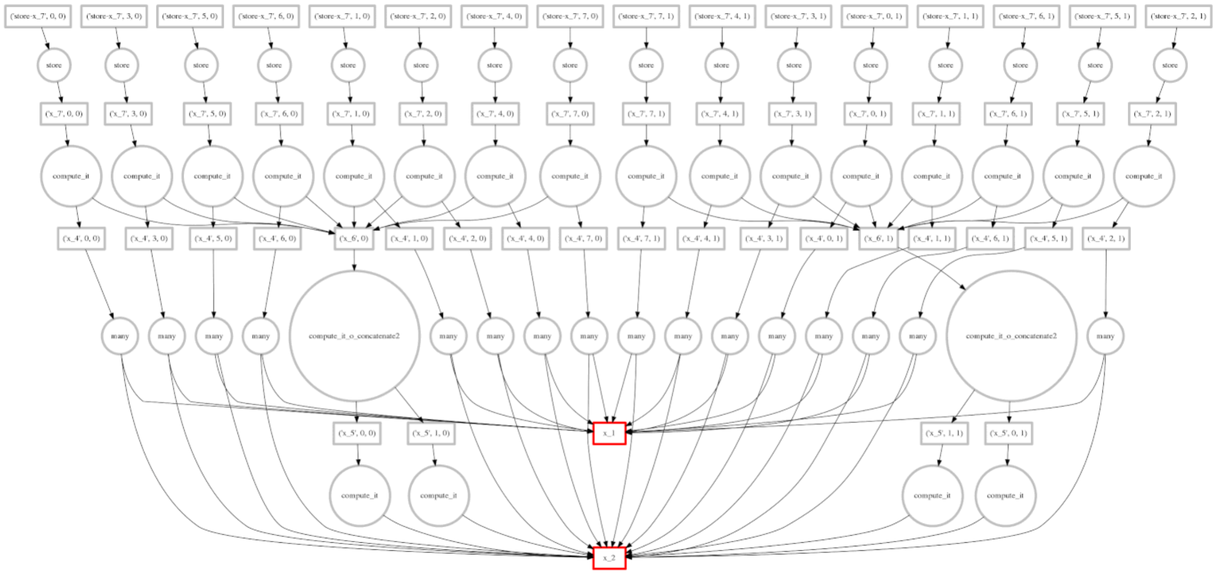
Dask: Simplified parallel and distributed processing in Python

https://github.com/dask/dask/



ND-Arrays are split into chunks that comfortably fit in memory

Dask.delayed allows the “lazy” loading of chunks which are not pulled into memory until required



Complex computations represented as a graph of individual tasks.

Scheduler optimizes execution of graph.

Pangeo JupyterLab Interface

PANGEO

hub.binder.pangeo.io

scottyhq/esip-tech-dive: ESIP Te... | JupyterLab | gee-houston-flood/1-houston-flo... | grfn_pangeo_demo/0-GRFN-de... | asp-pangeo-binder/0-skysat-vid... | stac-intake-landsat/0-intake-xar... | pangeo-tutorial-agu-2018/lands...

File Edit View Run Kernel Tabs Settings Help

Name | Last Mo

- binder — an hou
- data — an hou
- images — 20 hours
- notebooks — 20 hours
- demo-a... — 2 minutes
- abstract.... — 20 hours
- intake-la... — 23 minutes
- items.json — a minute
- LICENSE — 20 hours
- LICENSE... — 20 hours
- outline.md — 20 hours
- README... — 20 hours
- wa-bbox... — an hou
- Welcom... — 20 hours

demo-aws-2.ipynb

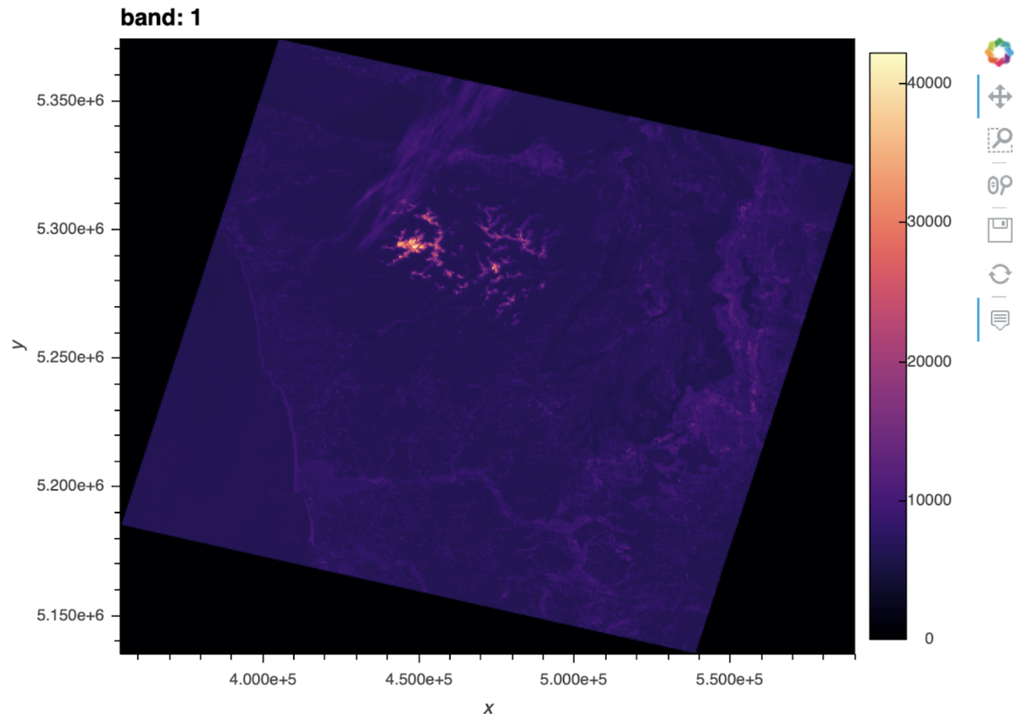
Code | Python 3

```
crs:          EPSG:32610
res:          (30.0, 30.0)
is_tiled:     1
nodatavals:   (nan,)
```

[16]:
```
print(item.date)
img = da.hvplot.image(rasterize=True, width=700, height=500, cmap='magma')
img
```

```
2013-07-26
```

[16]:



[15]:
```
# Display with reprojection to lat/lon
print(item.date)
from cartopy import crs as ccrs
crs = ccrs.UTM(zone='10n') #hvplot needs to know original CRS
img = da.hvplot.image(crs=crs,rasterize=True, width=700, height=500, cmap='magma')
```

Dask Workers

CPU Use (%)

Memory Use (%)

worker	ncores	c
tcp://10.49.45.1(	2	4
tcp://10.49.45.1(	2	2
tcp://10.49.45.1(	2	2
tcp://10.49.45.1(	2	4

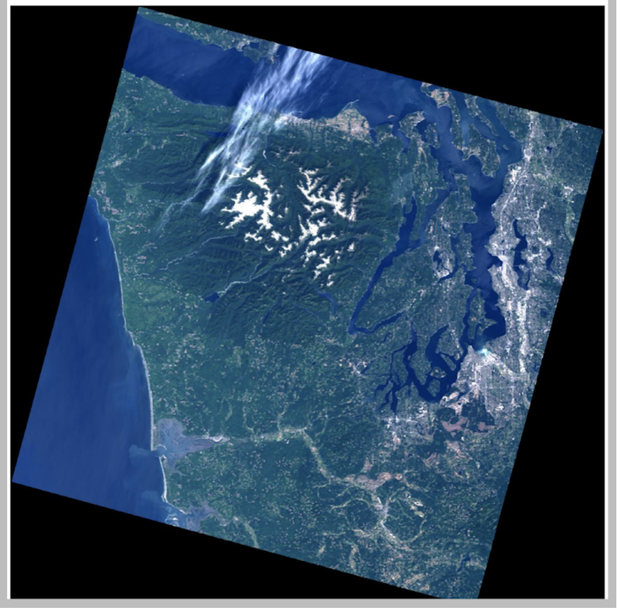
Dask Progress

Progress -- total: 0, in-memor

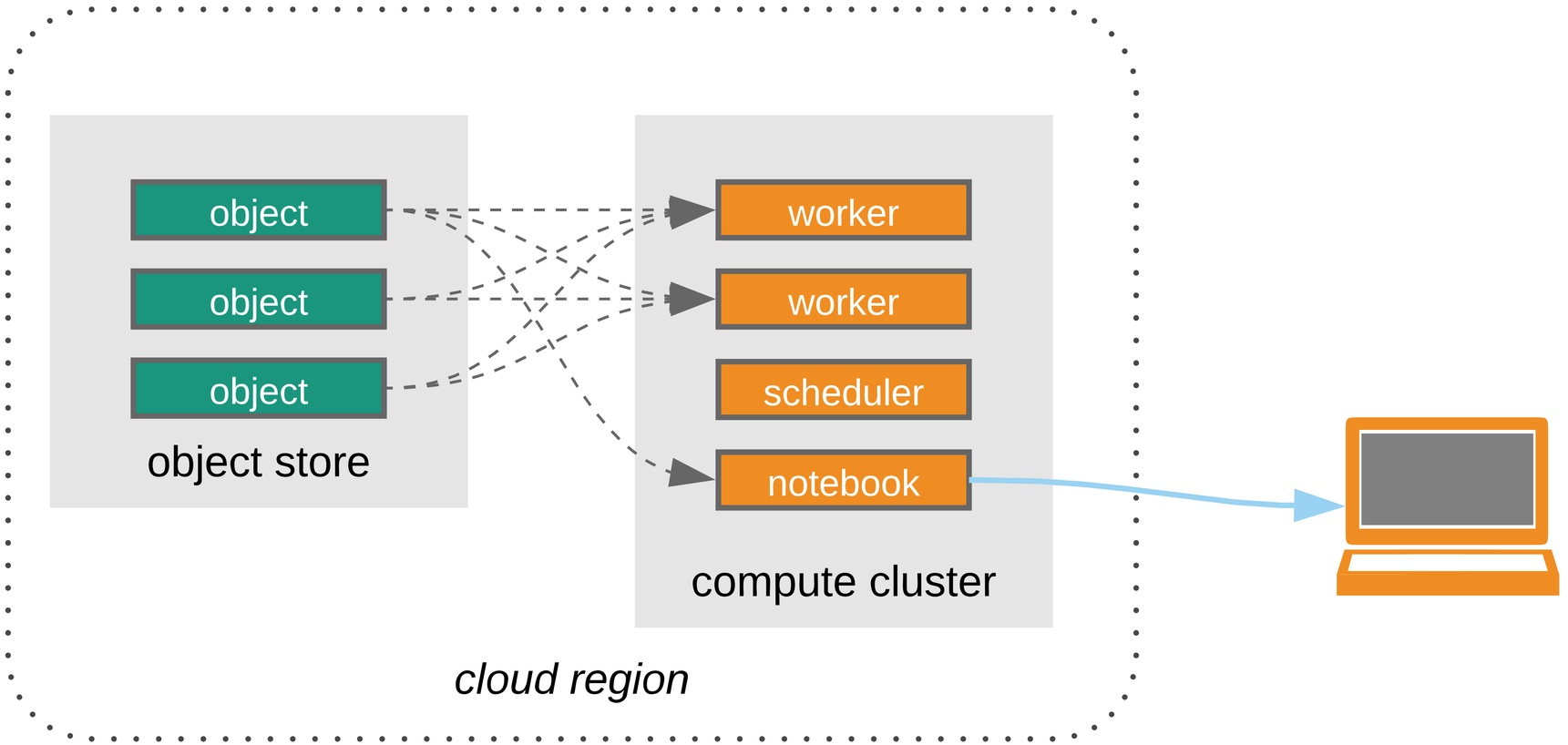
Output View

i 5

id=LC80470272013207LGN01 date=2013-07-26 19:03:48.916916+00:00 cloud%=1



Cloud-Native Approach

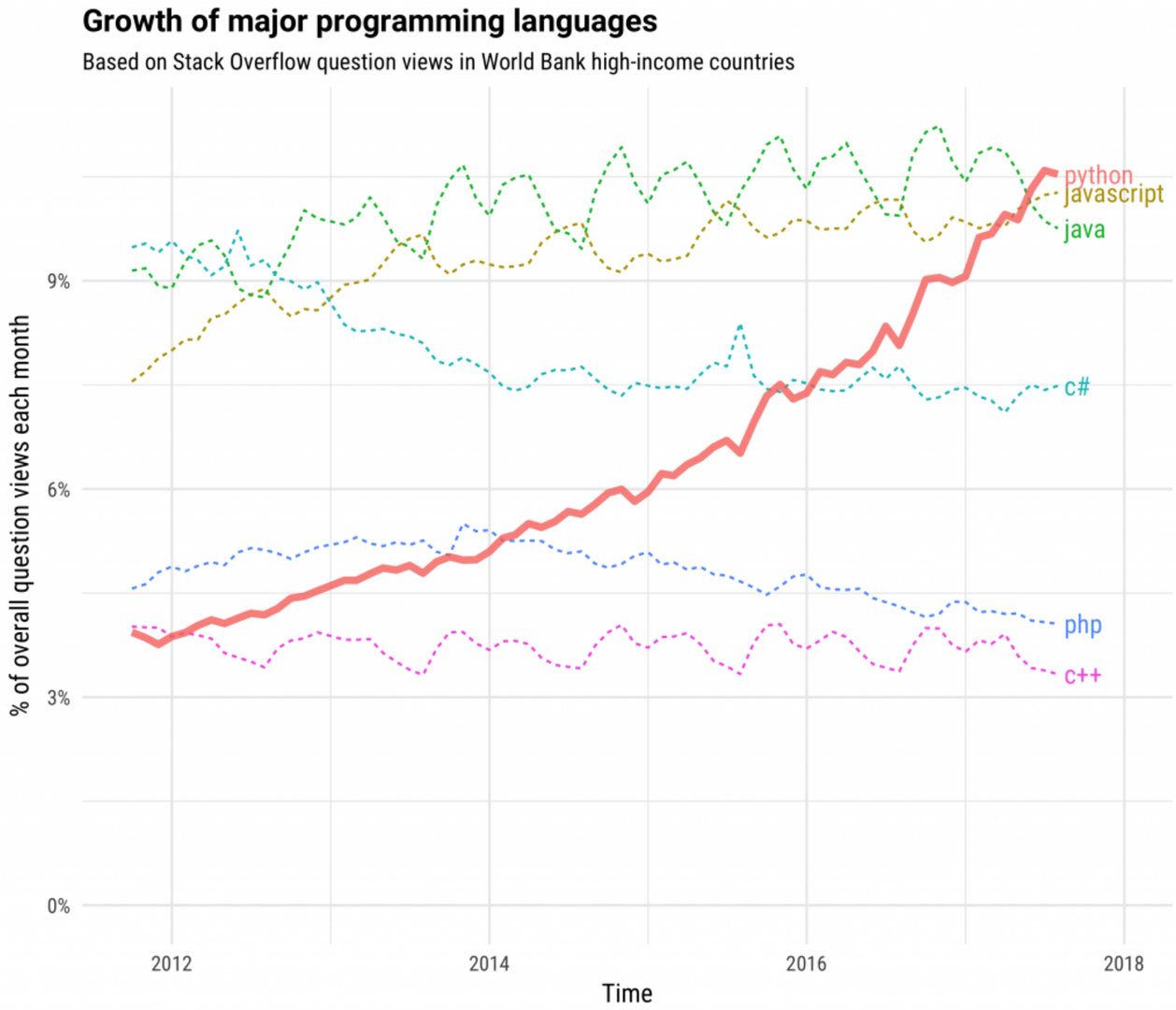


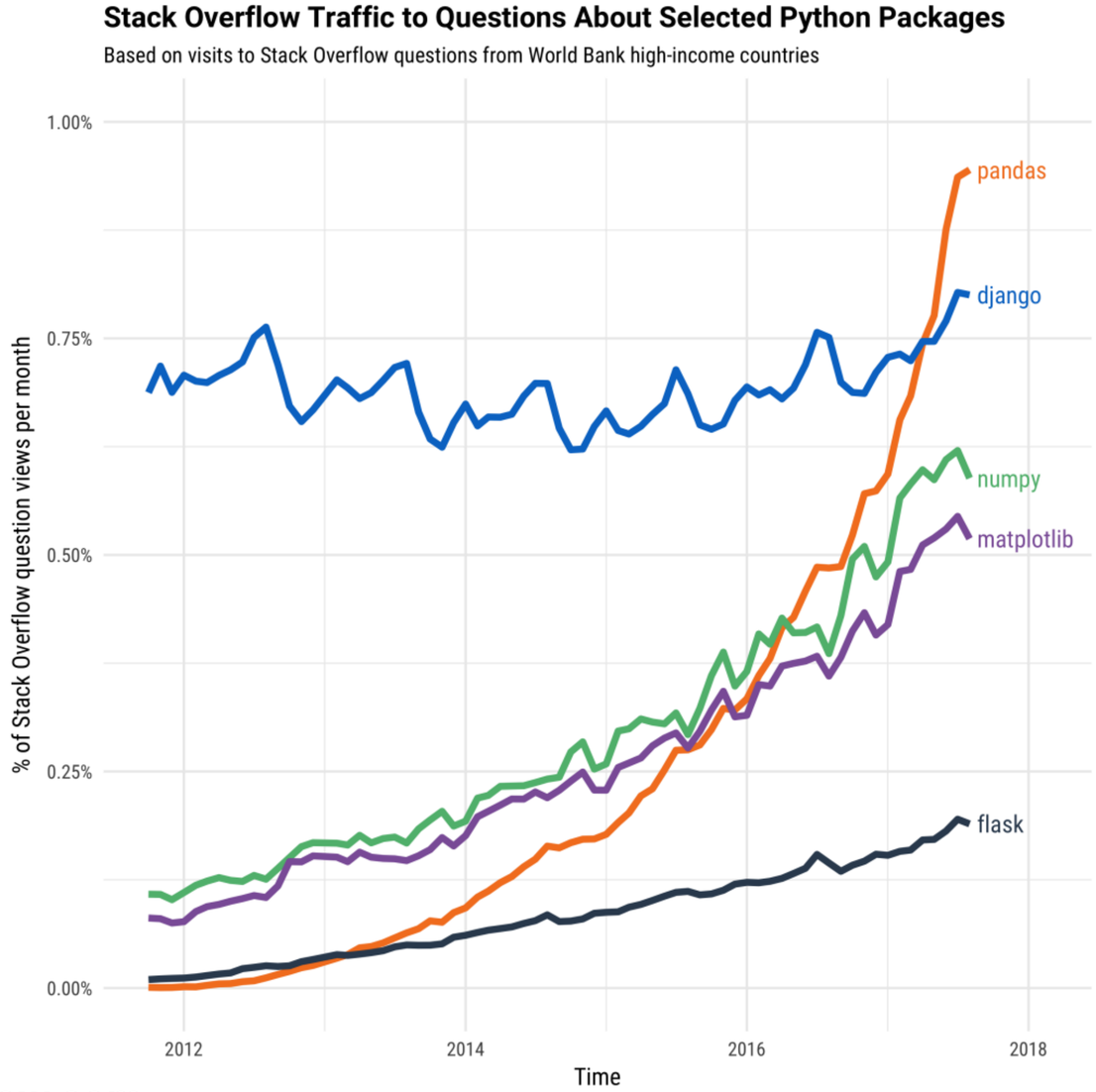
Data provider's responsibilities

End user's responsibilities

PANGEO

Scientific Python for Data Science





source: stackoverflow.com

PANGEO

Sign Up | Log In

all categories | Latest | Top | Categories

Topic	Replies	Views	Activity
Welcome to Pangeo Discourse Meta Pangeo is a community of scientists and software developers working together to improve the way we do scientific research. We work on software tools, such as the python packages Xarray and Dask, as well as customized env… read more	5	215	Oct 18
Call for speakers @ ESIP Pangeo session Science	4	34	20h
Access to Pangeo GCS Bucket to push model output from pre-CMIP6 experiments? Cloud	6	19	21h
Interpolation and regridding Science	7	169	1d
Pangeo Weekly Telecon News & Announcements	17	173	2d

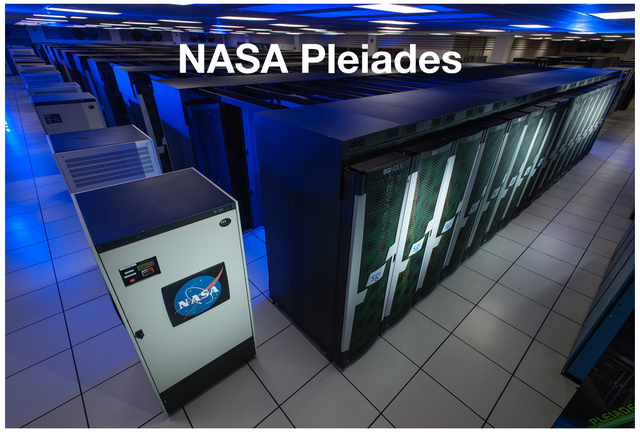
Pangeo Funding and Contributors

Pangeo Principles for Cloud-Native Science Infrastructure

- **Community-driven** - Our needs are no different from those of our peer institutions. By developing infrastructure collaboratively, we can accomplish much more than any one institution can alone.
- **Open source** - Because infrastructure is code, the code should be licensed in a way that enables the entire research community to reuse and build upon it.
- **Modular** - “all in one” solutions are impossible to maintain long term. Separation of concerns is a key principle of good software and systems engineering.
- **Vendor neutral** - Academic research infrastructure should use only vendor-neutral services APIs. If this principle is followed, it means we can redeploy our infrastructure anywhere.

PANGEO

Pangeo Deployments



NCAR Cheyenne

kubernetes

Google Cloud Platform

Jetstream

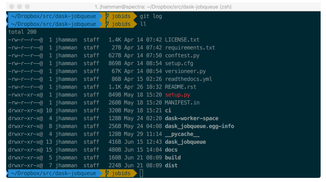
Microsoft Azure

aws

http://pangeo.io/deployments.html

PANGEO

Build your own pangeo

Storage Formats	HDF	OPeNDAP	Cloud Optimized Zarr/TileDB/Parquet/etc.
ND-Arrays	NumPy	DASK	More coming…
Data Models	xarray	Iris	pandas $y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$
Processing Mode	jupyter Interactive	Batch 	Serverless
Compute Platform	HPC 	Cloud Google Cloud Platform	Local

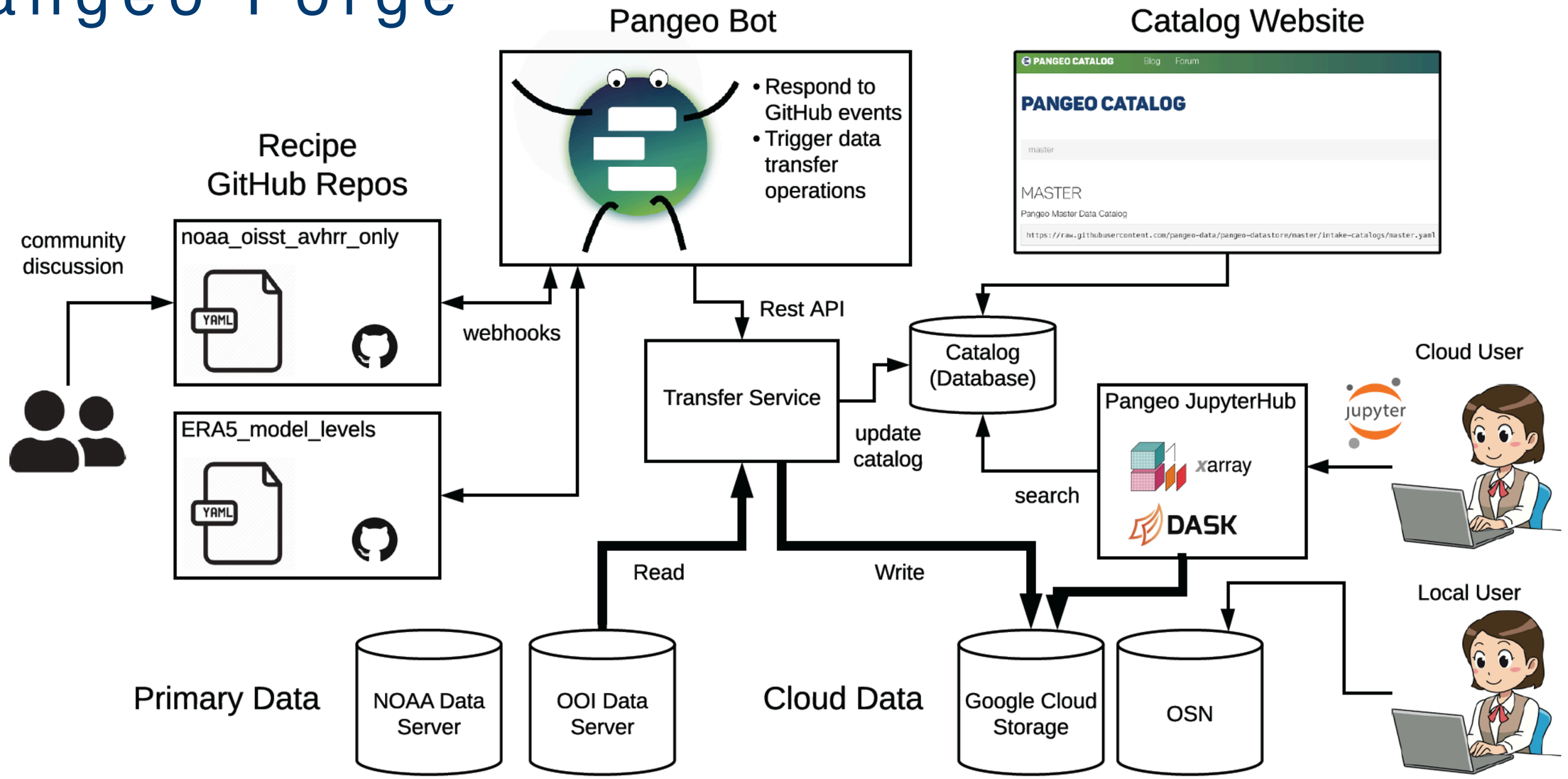
Toil

Adapted from the Google Site Reliability Engineering Guide

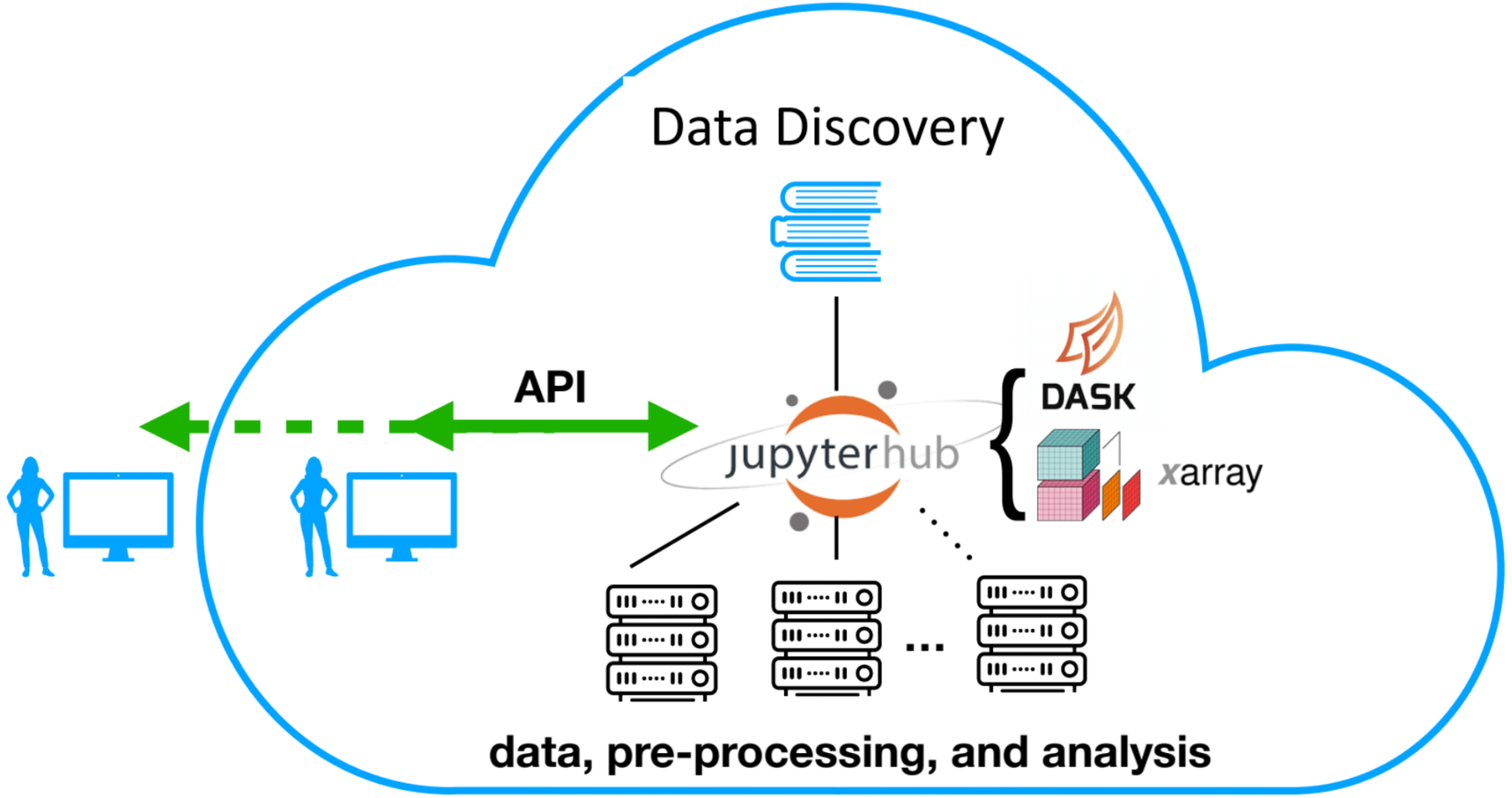
"Toil is the kind of work tied to scientific research that tends to be manual, repetitive, automatable, devoid of enduring value, and that scales linearly with the number of papers published."

B.Beyer, C. Jones, J. Petoff, and N. Murphy (2016)
Site Reliability Engineering: How GoogleRuns
Production Systems, O'Reilly Media, Incorporated.

Pangeo Forge



The Future is Now



Robinson et al., 2019

Some Things about the cloud that we have come to understand in the last couple of years

1. Compute is actually really inexpensive in the cloud
2. Storage is expensive, if you have to pay retail, but its possibly worth it
3. There is magic in the data centers. The providers typically don’t say much about how their data centers are organized, but the data in the cloud can move really fast around the data centers.

Learn More and Get Involved!

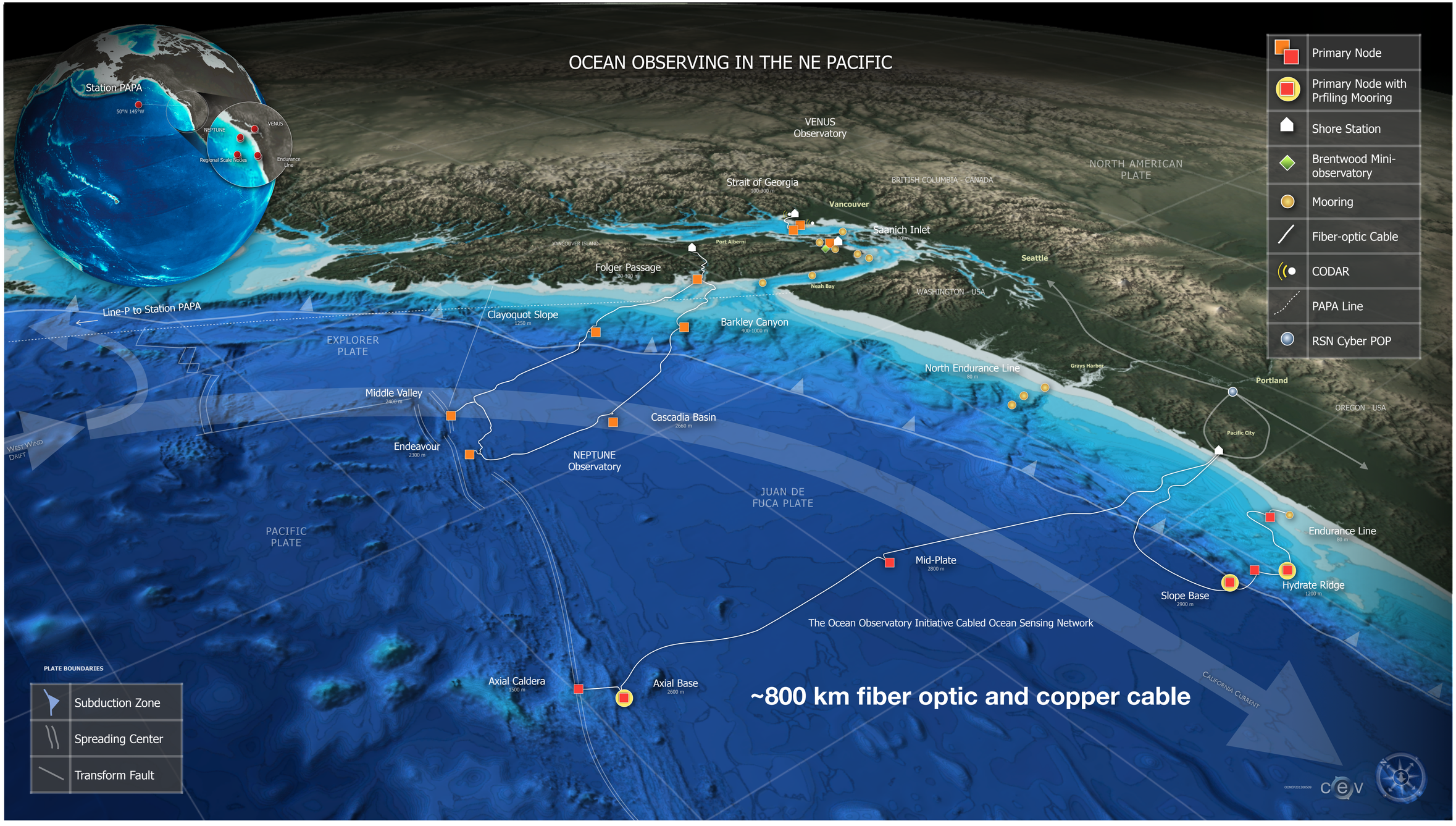
- Presentation: ESIP Tech Dive (01/2018) by Ryan Abernathy & Matthew Rocklin
- Website: http://pangeo.io/
- Github: https://github.com/pangeo-data
- Blog: https://medium.com/pangeo
- Gitter: https://gitter.im/pangeo-data
- Discourse: https://discourse.pangeo.io/

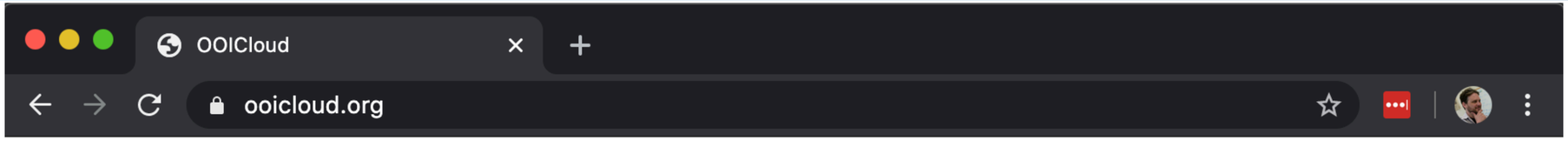
The Ocean Observatories Initiative (OOI) and OOI Data

https://oceanobservatories.org

OOI Organizations

OOI CamHD


OCEAN OBSERVING IN THE NE PACIFIC
Station PAPA
50°N 145°W
NEPTUNE
VENUS
Regional Scale Nodes
Endurance Line
VENUS Observatory
Strait of Georgia
100-300 m
BRITISH COLUMBIA - CANADA
NORTH AMERICAN PLATE
Vancouver
Saanich Inlet
100 m
VANCOUVER ISLAND
Port Alberni
Seattle
Folger Passage
30-100 m
Neah Bay
WASHINGTON - USA
Line-P to Station PAPA
Clayoquot Slope
1250 m
Barkley Canyon
400-1000 m
EXPLORER PLATE
North Endurance Line
80 m
Grays Harbor
Portland
OREGON - USA
Middle Valley
2400 m
Cascadia Basin
2660 m
Pacific City
WEST WIND DRIFT
Endeavour
2300 m
NEPTUNE Observatory
JUAN DE FUCA PLATE
PACIFIC PLATE
Endurance Line
80 m
Mid-Plate
2800 m
Hydrate Ridge
1200 m
Slope Base
2900 m
The Ocean Observatory Initiative Cabled Ocean Sensing Network
Axial Caldera
1500 m
Axial Base
2600 m
~800 km fiber optic and copper cable
CALIFORNIA CURRENT
Primary Node
Primary Node with Prfiling Mooring
Shore Station
Brentwood Mini-observatory
Mooring
Fiber-optic Cable
CODAR
PAPA Line
RSN Cyber POP
PLATE BOUNDARIES
Subduction Zone
Spreading Center
Transform Fault



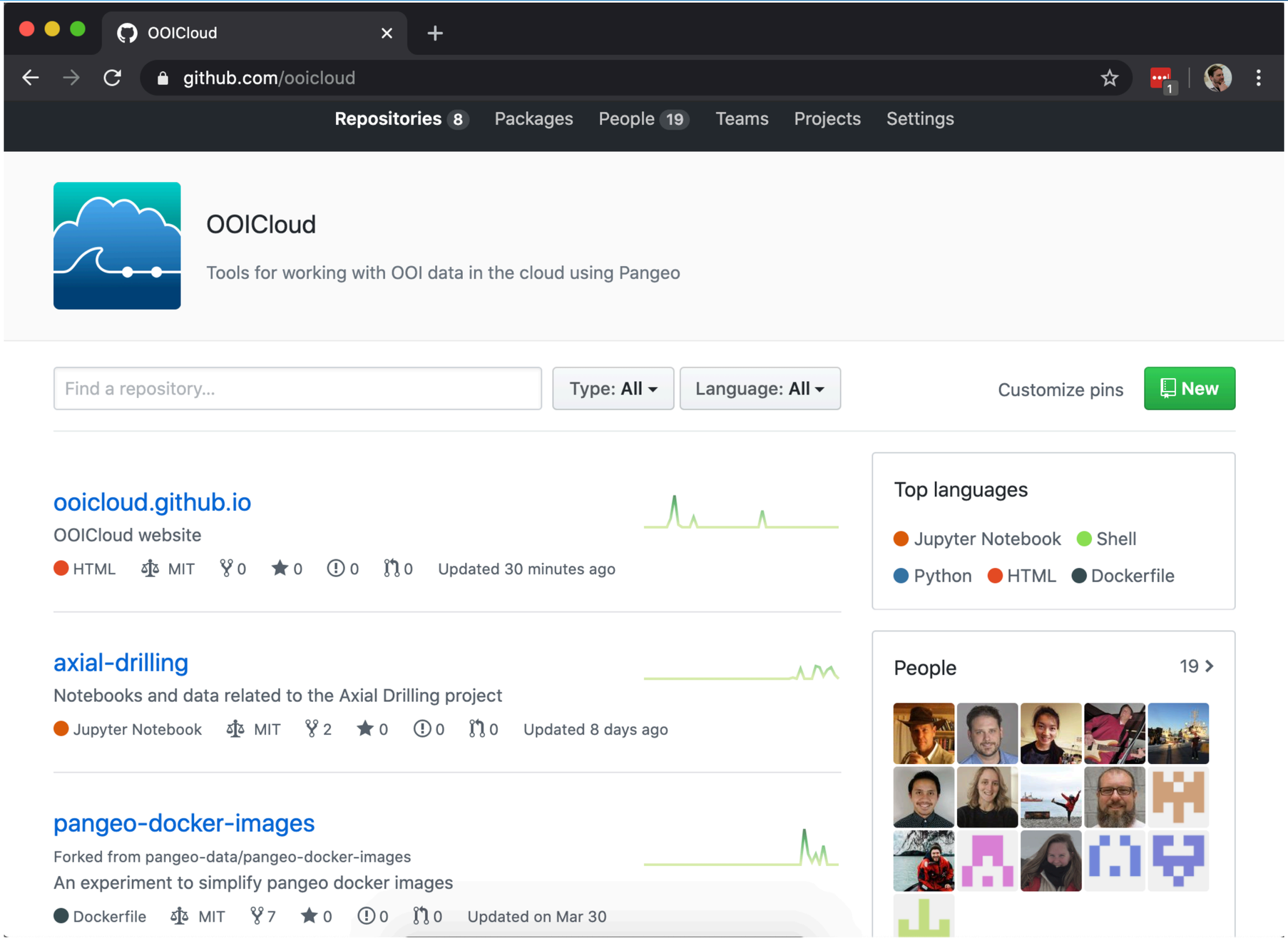
OOI CLOUD

About

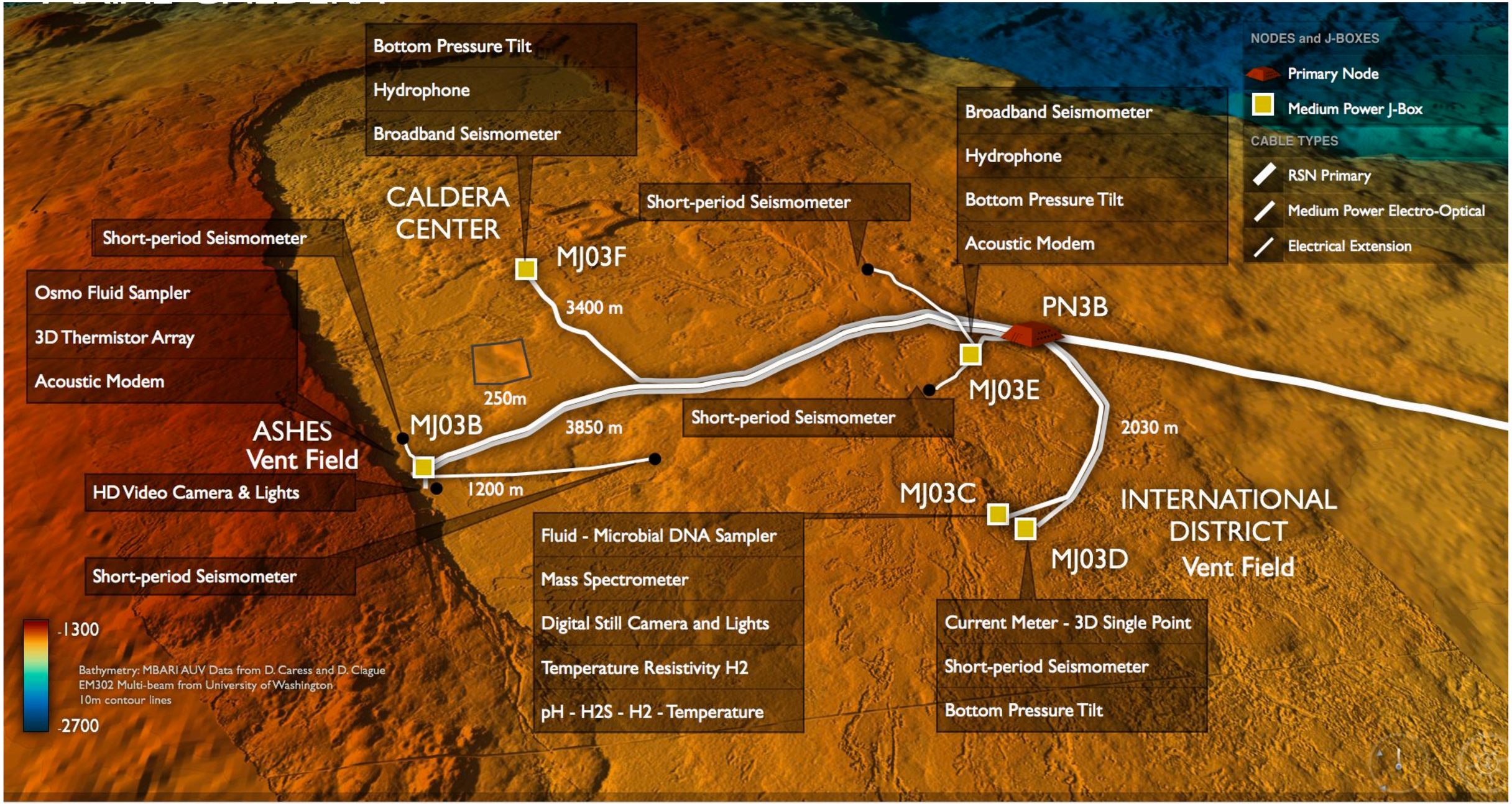
The OOICloud Project is working to make data from the Ocean Observatories Initiative (OOI) publically available in the cloud and accessible through a Pangeo interface. A primary goal is to provide these data to the scientific community using a cloud-performant object storage model, and to provide large-scale remote compute capabilities for research investigations.

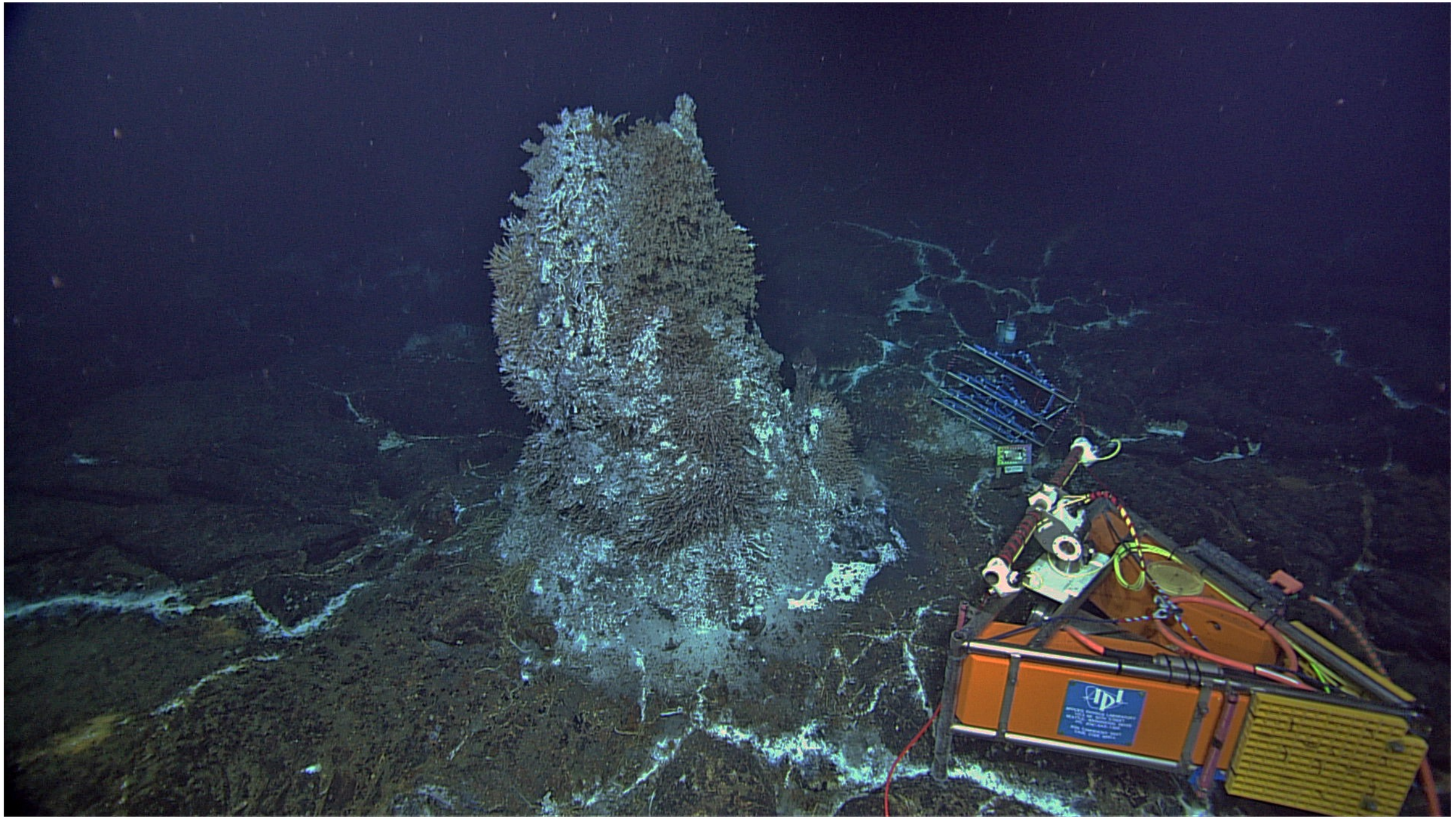
With a generous gift from the Microsoft AI for Earth program, OOICloud resides in the Azure Cloud, and currently contains all of the data from the OOI HD video camera deployed at a hydrothermal vent in the caldera of Axial Volcano, a submarine volcano on the Juan de Fuca Ridge.

OOICloud will begin incorporating other OOI datasets in the coming year, starting with data from the geodetic instruments and the echosounders.


OOICloud
github.com/ooicloud
Repositories 8
Packages
People 19
Teams
Projects
Settings
OOICloud
Tools for working with OOI data in the cloud using Pangeo
Find a repository...
Type: All
Language: All
Customize pins
New
ooicloud.github.io
OOICloud website
HTML MIT 0 0 0 0 Updated 30 minutes ago
axial-drilling
Notebooks and data related to the Axial Drilling project
Jupyter Notebook MIT 2 0 0 0 Updated 8 days ago
pangeo-docker-images
Forked from pangeo-data/pangeo-docker-images
An experiment to simplify pangeo docker images
Dockerfile MIT 7 0 0 0 Updated on Mar 30
Top languages
Jupyter Notebook
Shell
Python
HTML
Dockerfile
People
19

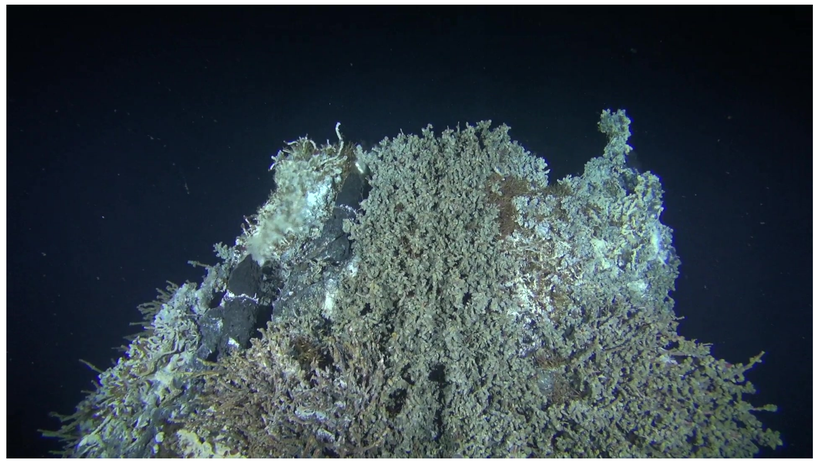
https://oceanobservatories.org
https://ooicloud.org
https://github.com/ooicloud


Bottom Pressure Tilt
Hydrophone
Broadband Seismometer
CALDERA CENTER
Short-period Seismometer
Broadband Seismometer
Hydrophone
Bottom Pressure Tilt
Acoustic Modem
NODES and J-BOXES
Primary Node
Medium Power J-Box
CABLE TYPES
RSN Primary
Medium Power Electro-Optical
Electrical Extension
Short-period Seismometer
Osmo Fluid Sampler
3D Thermistor Array
Acoustic Modem
MJ03F
3400 m
PN3B
250m
MJ03E
ASHES Vent Field
MJ03B
3850 m
Short-period Seismometer
2030 m
HD Video Camera & Lights
1200 m
MJ03C
INTERNATIONAL DISTRICT
Vent Field
MJ03D
Fluid - Microbial DNA Sampler
Mass Spectrometer
Digital Still Camera and Lights
Temperature Resistivity H2
pH - H2S - H2 - Temperature
Short-period Seismometer
Current Meter - 3D Single Point
Short-period Seismometer
Bottom Pressure Tilt
-1300
-2700
Bathymetry: MBARI AUV Data from D. Caress and D. Clague
EM302 Multi-beam from University of Washington
10m contour lines

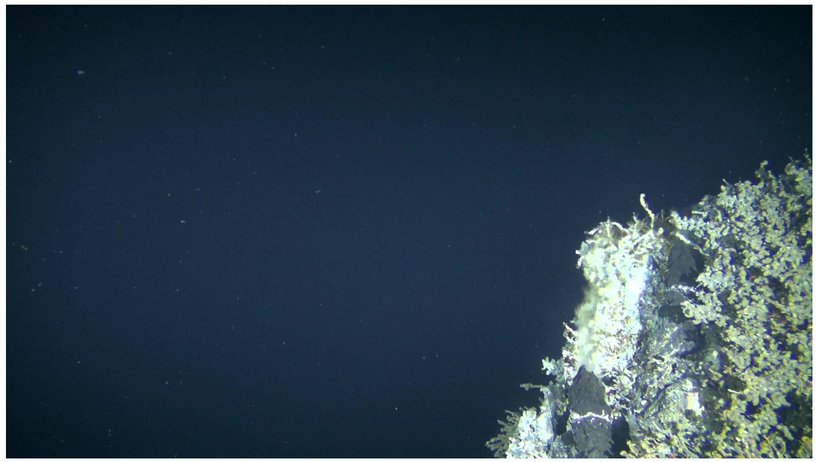


Science!

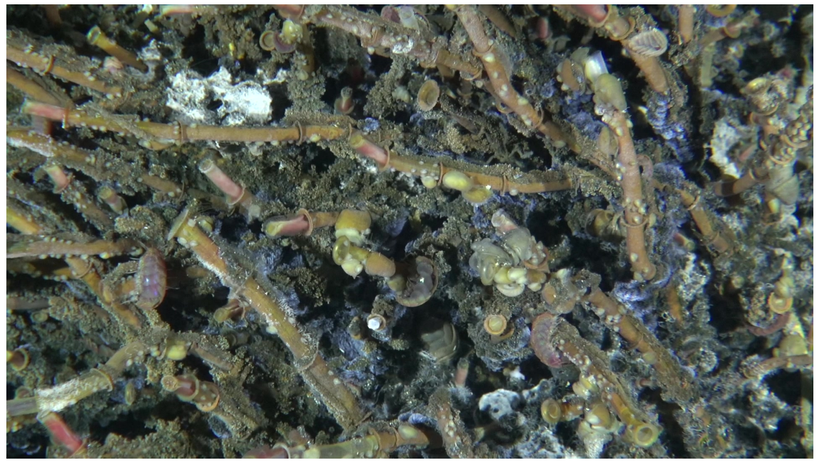
Geology



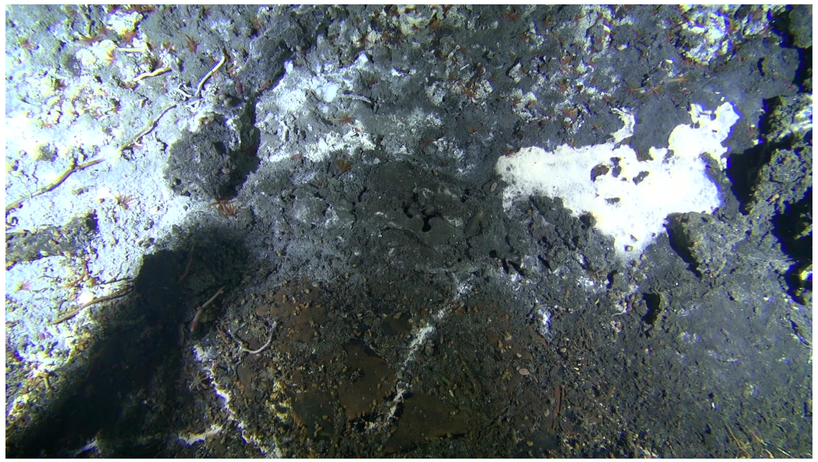
Oceanography and Hydrology



Macrofauna



Bacterial Mats



PyCamHD

https://github.com/tjcrone/pycamhd

- Python library for working with the CamHD raw data archive
- Image extraction functions
- File information functions
- Archive information functions

Archive Stats

camhd.get_stats() Return the total number of MOV files and the total size of the MOV files (in TB) in the data archive. Returns an integer and a float.

camhd.get_file_list() Return a list of all MOV files in the data archive as fully-qualified URLs. Returns a list of strings.

File Information

camhd.get_atom_sizes(filename) Return the sizes of the three top-level atoms in a remote file. Returns three integers.

camhd.get_frame_count(filename[, moov_atom]) Return the number of frames in a remote file. Returns an integer.

camhd.get_frame_sizes(filename[, moov_atom]) Return the sizes of all frames in a remote file. Returns a list of integers.

camhd.get_frame_offsets(filename[, moov_atom]) Return the offsets of all frames in a remote file. Returns a list of integers.

File Components

camhd.get_moov_atom(filename) Retrieve the moov atom from a remote file. Returns a string containing raw packed binary data.

camhd.get_frame_data(filename, frame_number[, moov_atom]) Retrieve the raw ProRes encoded frame data from a frame in a remote file. Returns a string containing raw packed binary data.

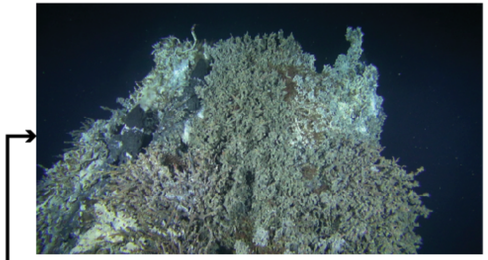
camhd.get_avi_file(frame_data) Adds an appropriately structured AVI header to frame_data. frame_data should be a string containing raw packed binary data as returned by get_frame_data(). Returns a string containing raw packed binary data.

Write Output

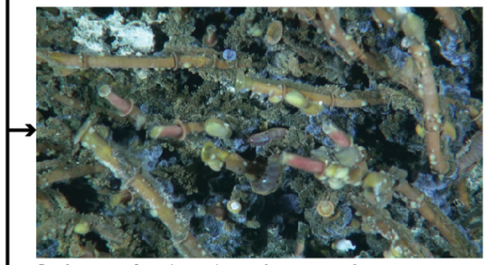
camhd.write_frame(filename, frame_number[, moov_atom]) Writes a single-frame AVI file. The resulting AVI file can be converted to a TIFF, PNG, YUV, or another image or movie format using ffmpeg. YUV conversions are lossless, as would be conversions to any valid container format using a video stream copy. All CamHD ProRes encoded video frames are key frames

Interactive Example

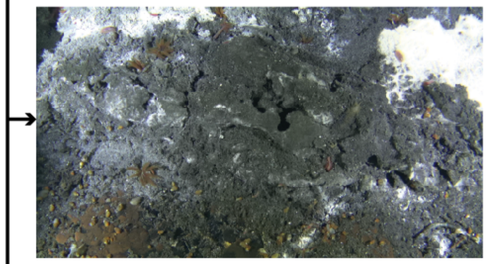
```
>>> import camhd
>>> filename = 'https://rawdata.oceanobservatories.org/files/RS03ASHS/
      PN03B/06-CAMHDA301/2016/07/29/CAMHDA301-20160729T210000Z.mov'
>>> moov_atom = camhd.get_moov_atom(filename)
>>> frame_set = [1980, 7845, 11460, 15930]
>>> for frame_number in frame_set:
...   camhd.write_frame(filename, frame_number, moov_atom)
```



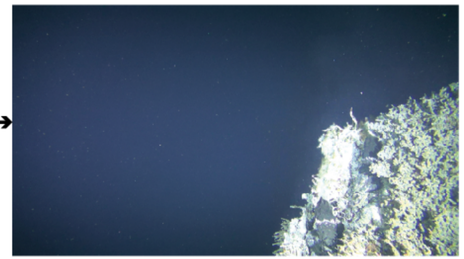
Study: vent structure growth, dynamics, fluid flow rates, heat and chemical fluxes.



Study: macrofaunal population dynamics and interactions.



Study: sulfide alteration, colonization, chemosynthesizing bacterial mats, things that snails do.



Study: suspended bacterial floc concentration, water column turbidity.