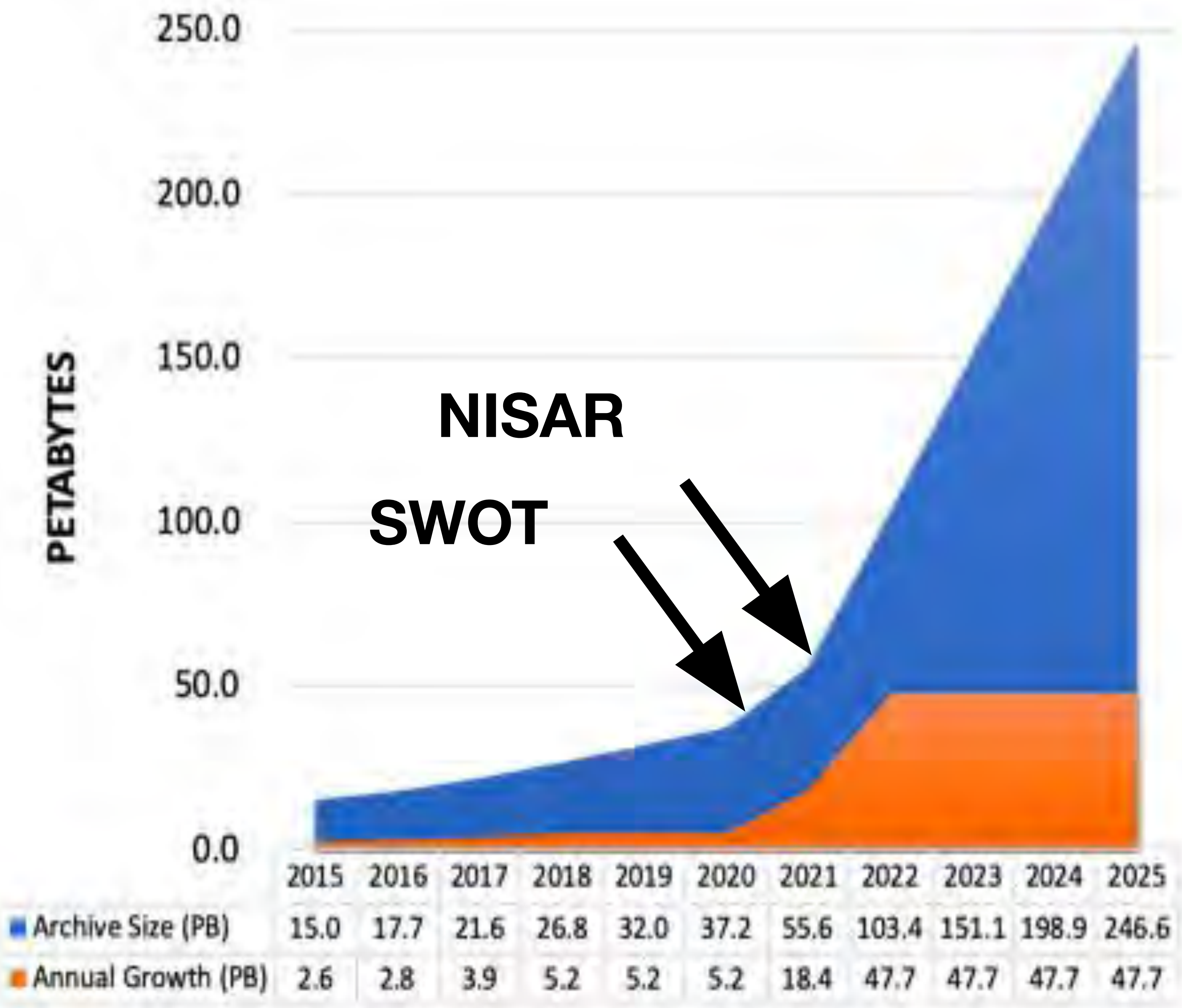
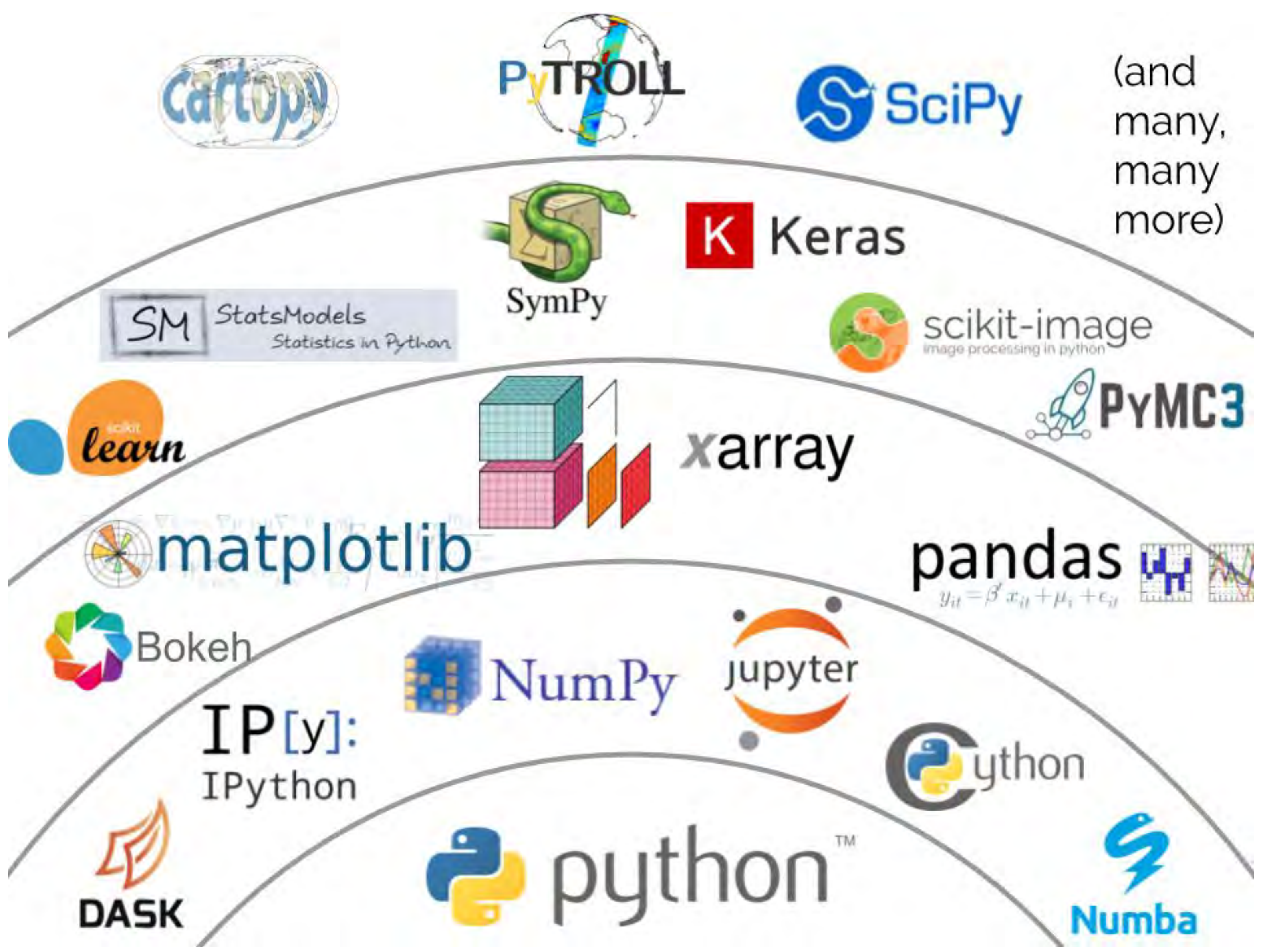




Google Cloud Platform

aws

Microsoft Azure

Empowering Transformational Science

Chelle Gentemann (Farallon Institute) twitter: [@ChelleGentemann]
Ryan Abernathey (Columbia / LDEO) twitter: [@rabernat]
Aimee Barciauskas (Development Seed) twitter: [@_aimeeb]

(there are lots of links in this presentation! click away!)

NASA

EARTH SCIENCE DATA SYSTEMS

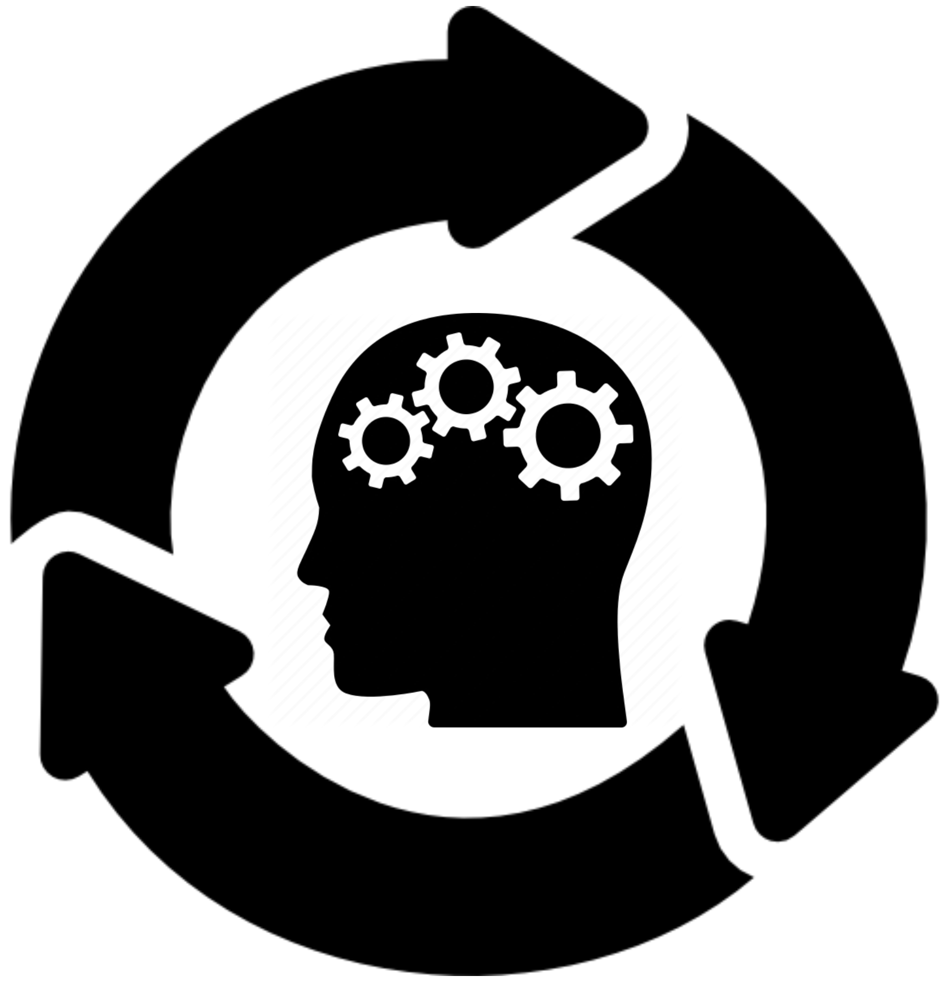
NASA Physical Oceanography Program

FARALLON INSTITUTE

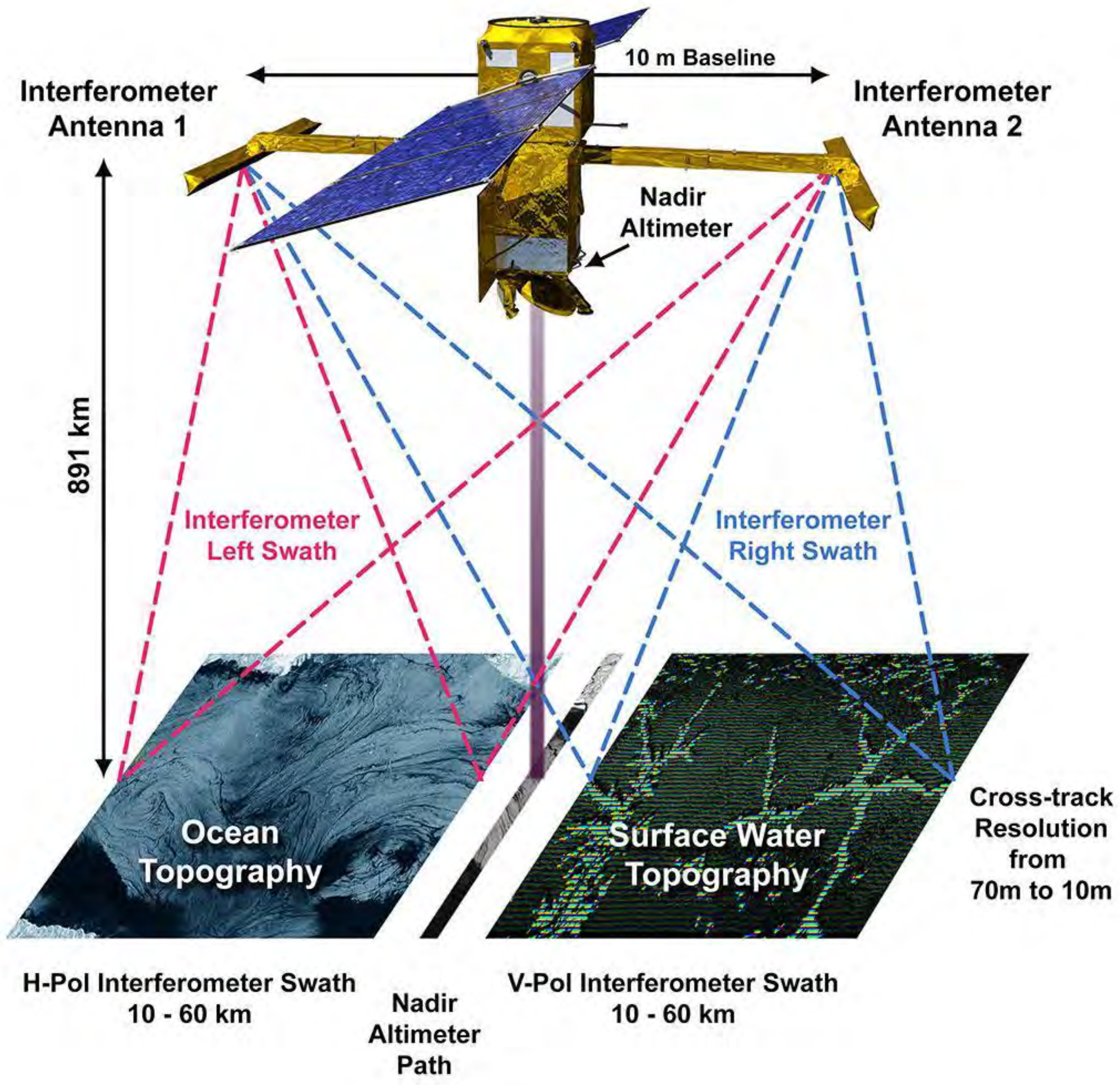
What Drives Progress in Earth System Science?

New Ideas / Hypotheses

$$\rho \frac{D\mathbf{u}}{Dt} = -\nabla p + \nabla \cdot \boldsymbol{\tau} + \rho \mathbf{g}$$



New Observations



New Simulations

What impacts the velocity of progress?

Data, Software, & Compute

Data: time to find, access, clean, & format data for analysis
Software: what tools are easily available?
Compute: access to compute == speed of results

80% Data Preparation (download, clean, & organize files)	10% Batch Processing	10% Think about science

Traditional methods of data access cannot leverage large volumes of data

US$1.5 billion

With a total cost estimated at US**$1.5 billion**, NISAR is likely to be the world's most expensive Earth-imaging satellite.

en.wikipedia.org › wiki › NISAR_(satellite)

NISAR (satellite) - Wikipedia

140 petabytes

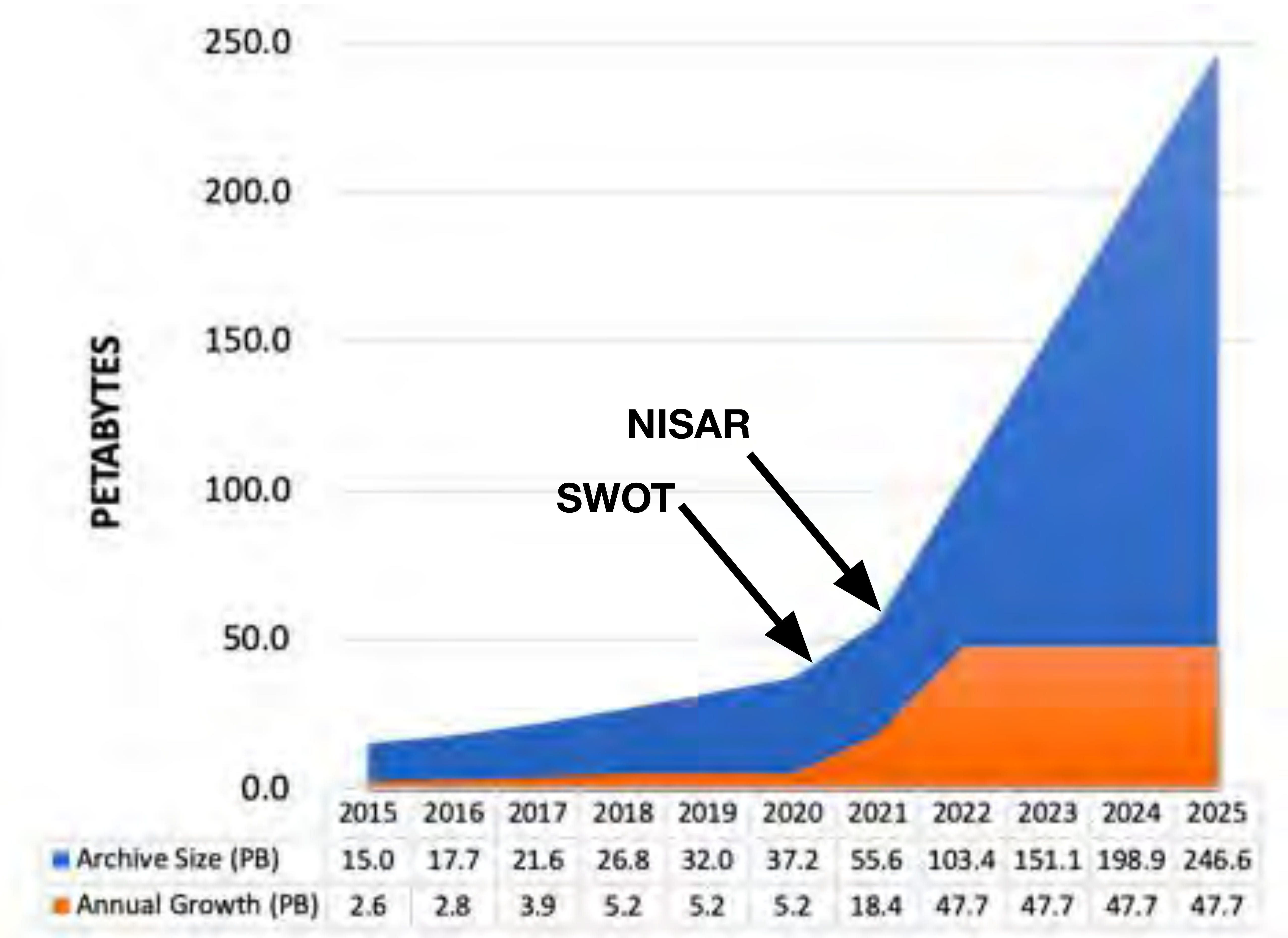
As Dobson notes, **NISAR** is expected to generate a tremendous **volume** of data over its scheduled three-year **mission** — as much as 140 petabytes (PB).

asf.alaska.edu › wp-content › uploads › 2019/06 › 201... PDF

Getting Ready for NISAR - Alaska Satellite Facility

About Featured Snippets Feedback

Data, Software, Compute



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

DATA ANALYTICS

New climate model data now in Google Public Datasets

Registry of Open Data on AWS

aws

Multi-Scale Ultra High Resolution (MUR) Sea Surface Temperature (SST)

climate earth observation environmental natural resource oceans satellite imagery sustainability water weather

Description

A global, gap-free, gridded, daily 1 km Sea Surface Temperature (SST) dataset created by 2 satellite SST datasets. Those input datasets include the NASA Advanced Microwave Sca (AMSR-E), the JAXA Advanced Microwave Scanning Radiometer 2 (AMSR-2) on GCOM-W Resolution Imaging Spectroradiometers (MODIS) on the NASA Aqua and Terra platforms, WindSat radiometer, the Advanced Very High Resolution Radiometer (AVHRR) on several situ SST observations from the NOAA iQuam project. Data are available from 2002 to pre original source of the MUR data is the NASA JPL Physical Oceanography DAAC.

Update Frequency

Daily

License

There are no restrictions on the use of these data.

Documentation

https://podaac.jpl.nasa.gov/dataset/MUR-JPL-L4-GLOB-v4.1

Managed By

https://faralloninstitute.org

See all datasets managed by https://faralloninstitute.org.

Contact

podaac@podaac.jpl.nasa.gov

Usage Examples

Tutorials

- Python Jupyter Notebooks by Chelle Gentemann, Rich Signell

Microsoft Azure | Open Datasets

Product

GOES-16

SatelliteImagery EarthObservation AIforEarth NOAA

Overview Data access

Weather imagery from the GOES-16 satellite.

The GOES-R (Geostationary Operational Environmental Satellite) program images weather phenomena from a set of satellites in geostationary orbits. The GOES-16 satellite is the first of four planned GOES-R satellites; GOES-16's orbit provides a view of the Americas.

This dataset currently includes the ABI-L2-MCMIPF product (Advanced Baseline Imager, Level 2, Cloud and Moisture Imagery, Full-disk). We may on-board other GOES-16 and GOES-17 products on request; please contact aiforearthdatasets@microsoft.com if you are interested in using additional GOES data on Azure.

This dataset is available on Azure thanks to the NOAA Big Data Program.

Storage resources

Data are stored in blobs in NetCDF format (one blob per image) in the East US data center, in the following blob container:

https://goes.blob.core.windows.net/noaa-goes16

Within that container, data are named as:

```
[product]/[year]/[day]/[hour]/[filename]
```

Analytics Optimized Data Store (AODS)

a few examples of AODS formats

Parquet

COG
CLOUD OPTIMIZED
GEOTIFF

Zarr

Current method -
NetCDF files - organized into 'reasonable' data sizes per file, usually by orbit, granule, or day. Filename has information about date, sensor, version. Reading usually involved calculating the filename, opening, reading, processing, closing.

Analytics Optimized Data Store (one example of many different formats)
Zarr - makes large datasets easily accessible to distributed computing. Original data is stored in directories each having chunked data corresponding to dataset dimensions. Metadata is read by zarr libraries to read only the chunks necessary to complete a subsetting request.

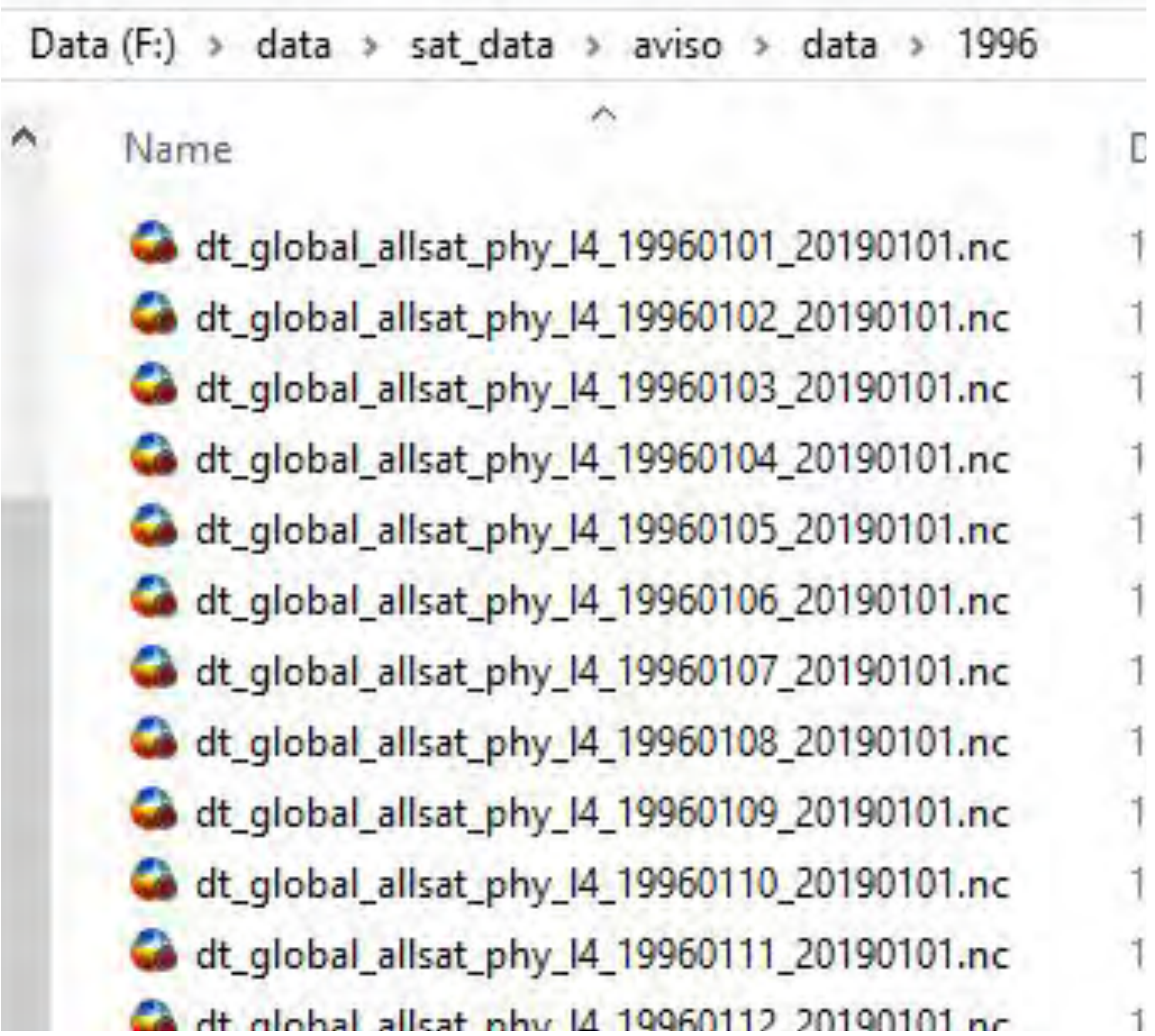
Technology advances -
Lazy loading - also known as asynchronous loading - defer initialization of an object until the point at which it is needed. Developed for webpages. Delays reading data until needed for compute.

Advanced OSS libraries:
Xarray - library for analyzing multi-dimensional arrays, lazy loading.
Dask - able to break a large computational problems into a network of smaller problems for distribution across multiple processors
Intake - lightweight set of tools for loading and sharing data in data science projects

What does a data store look like?

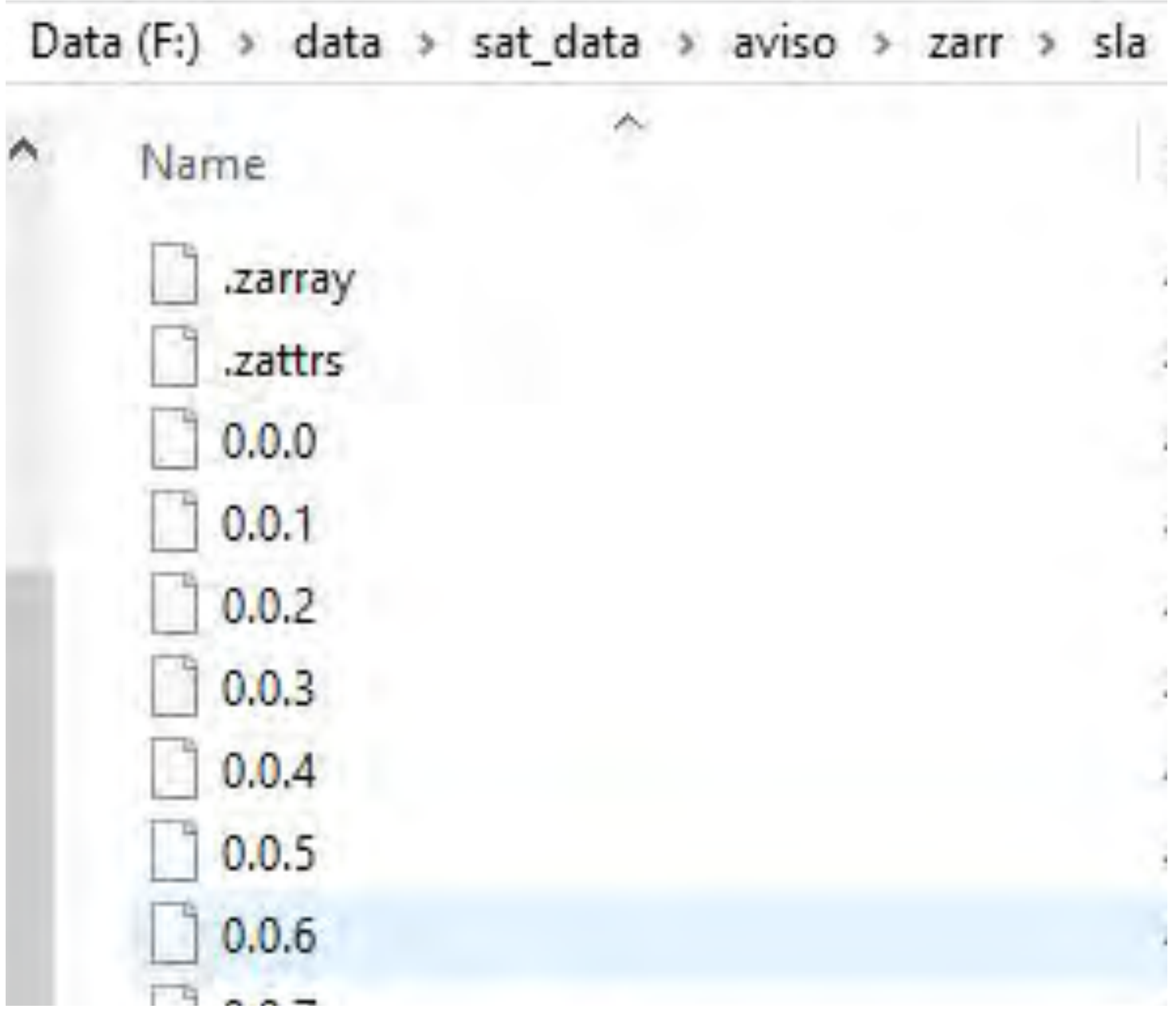
NetCDF

Organized so that each file can fit into RAM, usually by day, orbit, or granules



Zarr

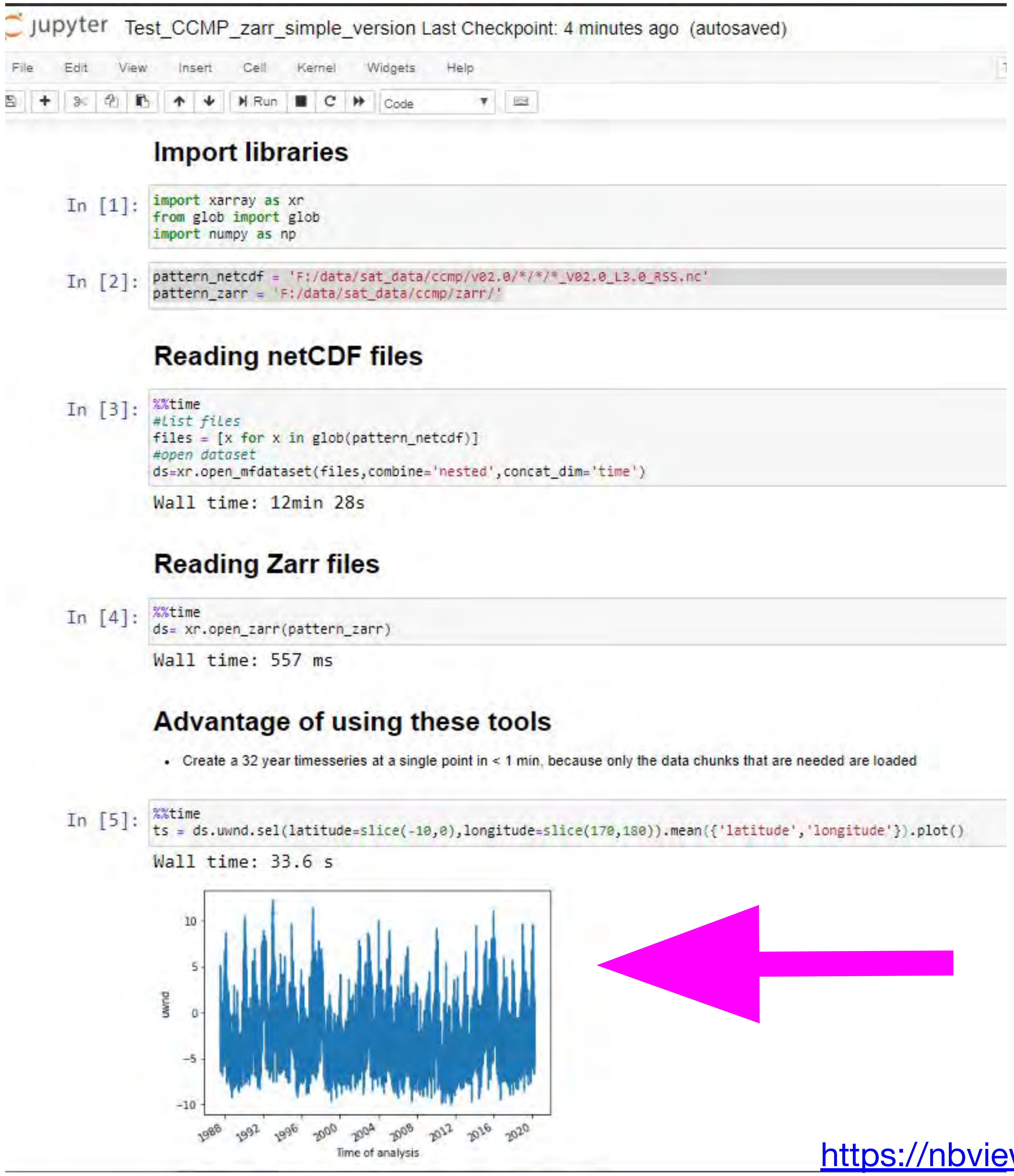
organization and format invisible to user, data accessed by metadata



EARTH SCIENCE DATA SYSTEMS

FARALLON INSTITUTE

How to access data?



My version of lazy loading before I knew python - on bedrest, pregnant with twins

Modern software tools use lazy loading to access large datasets

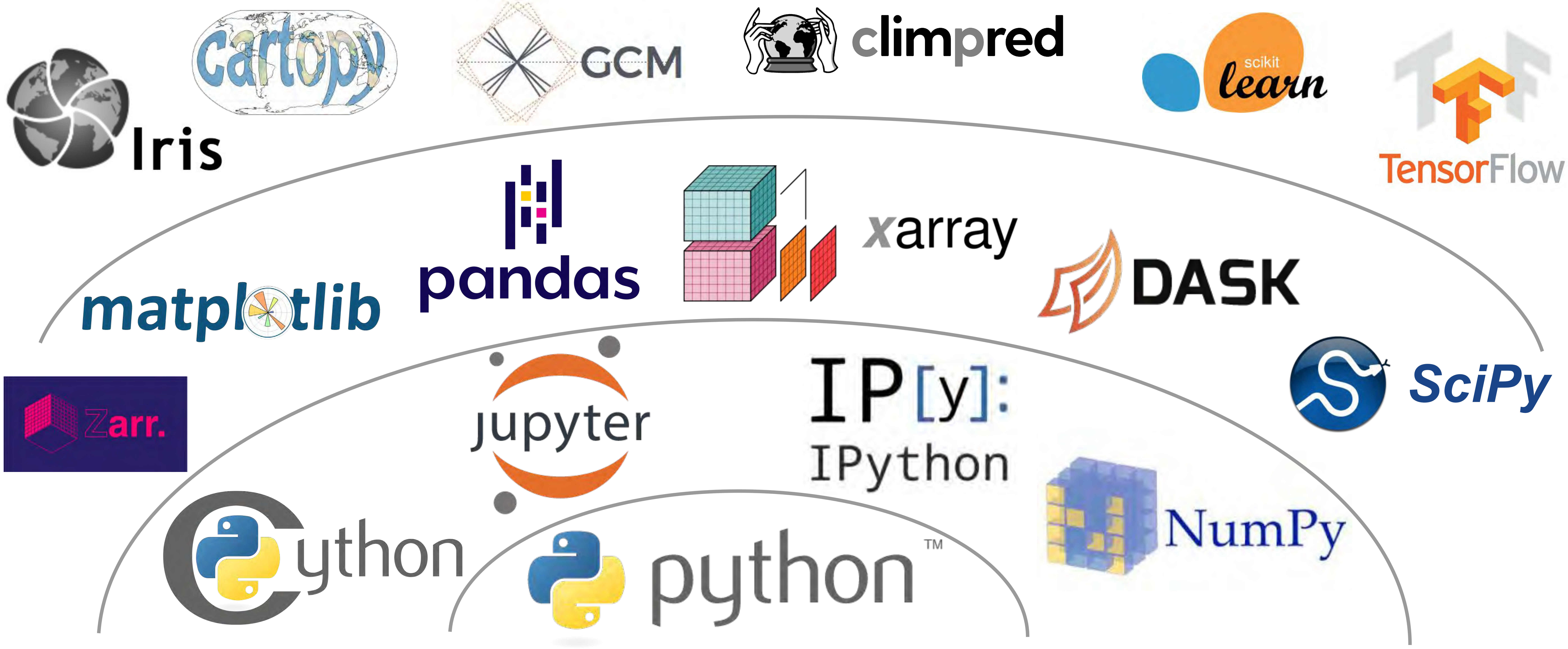
Reading in netCDF data: 13 minutes (depends on computer)
1 - user creates list of filenames
2 - access dataset by reading the metadata distributed through files

Reading in Zarr data: 0.1 seconds (metadata consolidated)
1 - access dataset by reading the consolidated metadata

STOP -------------- THIS IS DIFFERENT -------------------
1 line of code to access a 32-year, global, 25km dataset
1 line of code to select a region, calculate mean, & plot time series
in LESS than 1 minute

https://nbviewer.jupyter.org/github/cgentemann/Biophysical/blob/master/Test_CCMP_zarr_simple_version.ipynb

Data, Software, Compute



Inspiration: Stephan Hoyer, Jake Vanderplas (SciPy 2015)

Intake

Read CMIP6 data from Google Cloud using intake

500TB!!!

The CMIP6 data is a huge collection of different experiments. Access to these data uses the intake-esm library which you then use the catalog to select specific variables, experiments, or activities. (**Note: intake-esm is quite news and experimental.**) There are some great tutorials here and here.

More information on CMIP6 is here and variable names here

A nice introduction is here

The Pangeo catalog listing is here

```
%%time
col = intake.open_esm_datastore("https://raw.githubusercontent.com/NCAR/intake-esm-datastore/master/catalogs/pangeo-cmip6.json")

col
```

```
CPU times: user 898 ms, sys: 156 ms, total: 1.05 s
Wall time: 1.76 s
pangeo-cmip6-ESM Collection with 267459 entries:
    > 15 activity_id(s)
    > 33 institution_id(s)
    > 70 source_id(s)
    > 102 experiment_id(s)
    > 140 member_id(s)
    > 29 table_id(s)
    > 369 variable_id(s)
    > 10 grid_label(s)
    > 267459 zstore(s)
    > 60 dcpp_init_year(s)
```

Search the collection for historical, monthly, air temperature, for one realization

You can use the `variable id` (link given above) to search for different parameters, change the `table id` from atmospheric monthly to ocean monthly, 3hrly data etc. More information on what you can search for is in this tutorial

```
cat_cmip = col.search(experiment_id=['ssp585','historical'],  # pick the `historical` forcing experiment
               table_id='Amon',          # choose to look at atmospheric variables (A) saved at monthly resolution (mon)
               variable_id='tas',        # choose to look at near-surface air temperature (tas) as our variable
               member_id = 'r1i1p1f1')   # arbitrarily pick one realization for each model (i.e. just one set of initial conditions)
cat_cmip
```

```
pangeo-cmip6-ESM Collection with 67 entries:
    > 2 activity_id(s)
    > 27 institution_id(s)
    > 42 source_id(s)
    > 2 experiment_id(s)
```

https://github.com/pangeo-gallery/osm2020tutorial

Intake catalogues can significantly reduce friction (or barriers to entry) caused by figuring out how to access data stores

A powerful way to provide access for non-experts

Maintains all metadata

Users can easily build OSS libraries to harmonize data

Anaconda Cloud 0.1.4 | conda-forge 557 | pypi v0.1.4 | build passing | build passing | codecov 45% | License MIT | DOI 10.5281/zenodo.3743397

cmip6_preprocessing

Frustrated with how 'dirty' CMIP6 data still is? Do you just want to run a simple (or complicated) analysis on various models and end up having to write logic for each seperate case? Then this package is for you.

Developed during the cmip6-hackathon this package provides utility functions that play nicely with intake-esm.

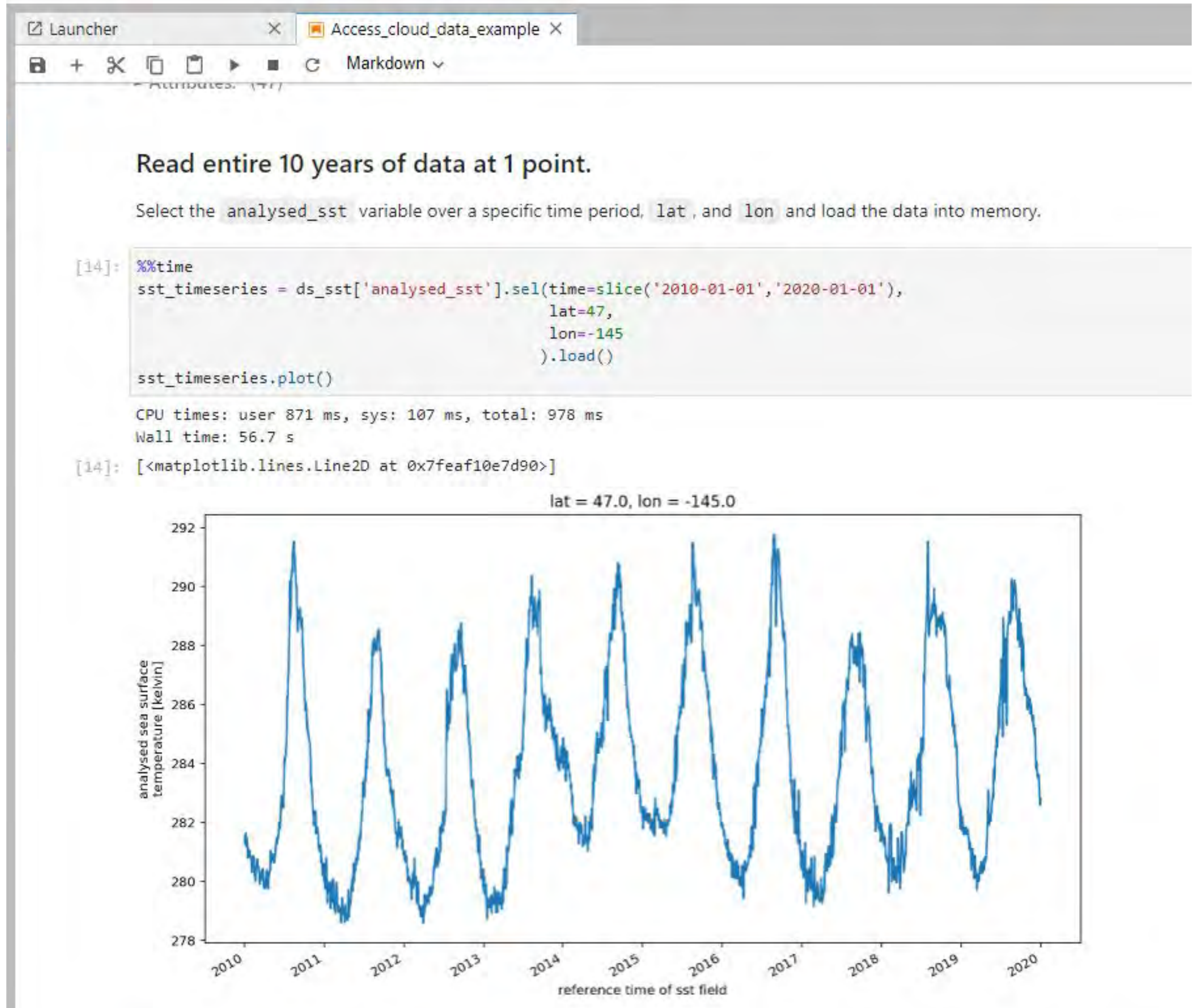
@JuliusBusecke

Julius Busecke
@JuliusBusecke Follows you

Xarray

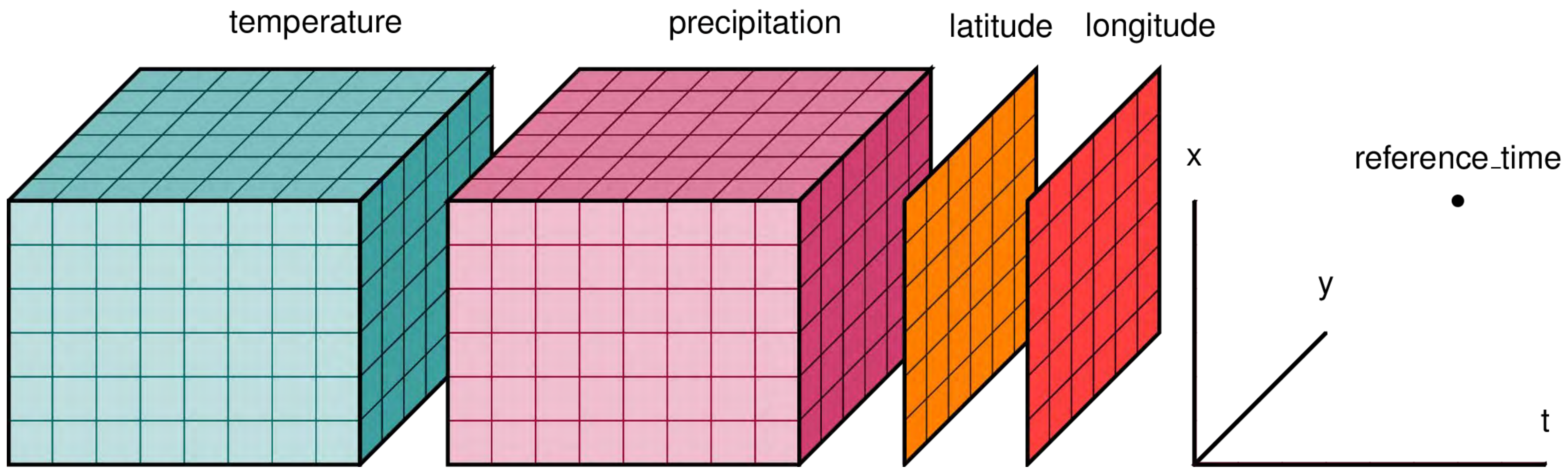
@xarray_dev

NUMFOCUS
OPEN CODE = BETTER SCIENCE



Xarray introduces labels in the form of dimensions, coordinates and attributes on top of raw NumPy-like multidimensional arrays, which allows for a more intuitive, more concise, and **less error-prone** developer experience.

Coding with coordinates == CLARITY



Make a timeseries:
data.sel(latitude=10,longitude=0,method='nearest').plot()

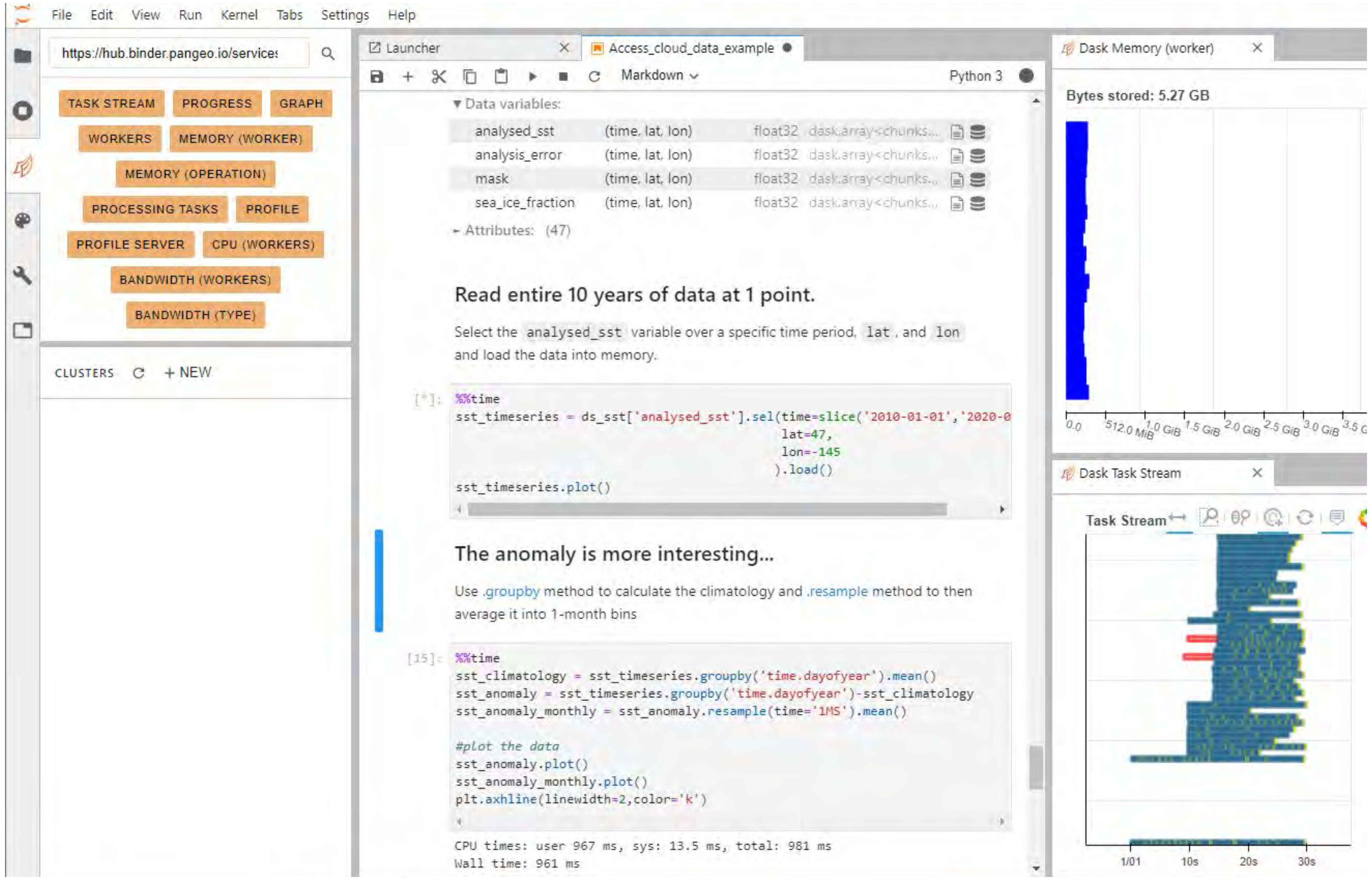
Regrid data for multivariate analysis:
data_regrid = data.interp_like(other_data)

https://github.com/pangeo-gallery/osm2020tutorial

Dask

@dask_dev

NUMFOCUS
OPEN CODE = BETTER SCIENCE



Dask is a flexible library for parallel computing in Python.

Xarray integrates with Dask to support parallel computations and streaming computation on datasets that don't fit into memory. When you are using Xarray, you are likely using Dask without even realizing it.

The Pangeo binder and jupyterhub use Dask Gateway to manage access to the Dask clusters & kubernetes.

Cluster performance visualization using the Dask viewers: memory use, profile, CPU, etc.

https://github.com/pangeo-gallery/osm2020tutorial

NASA EARTH SCIENCE DATA SYSTEMS

FARALLON INSTITUTE

Data, Software, Compute

Analytics Optimized Data Store (AODS)

Parquet

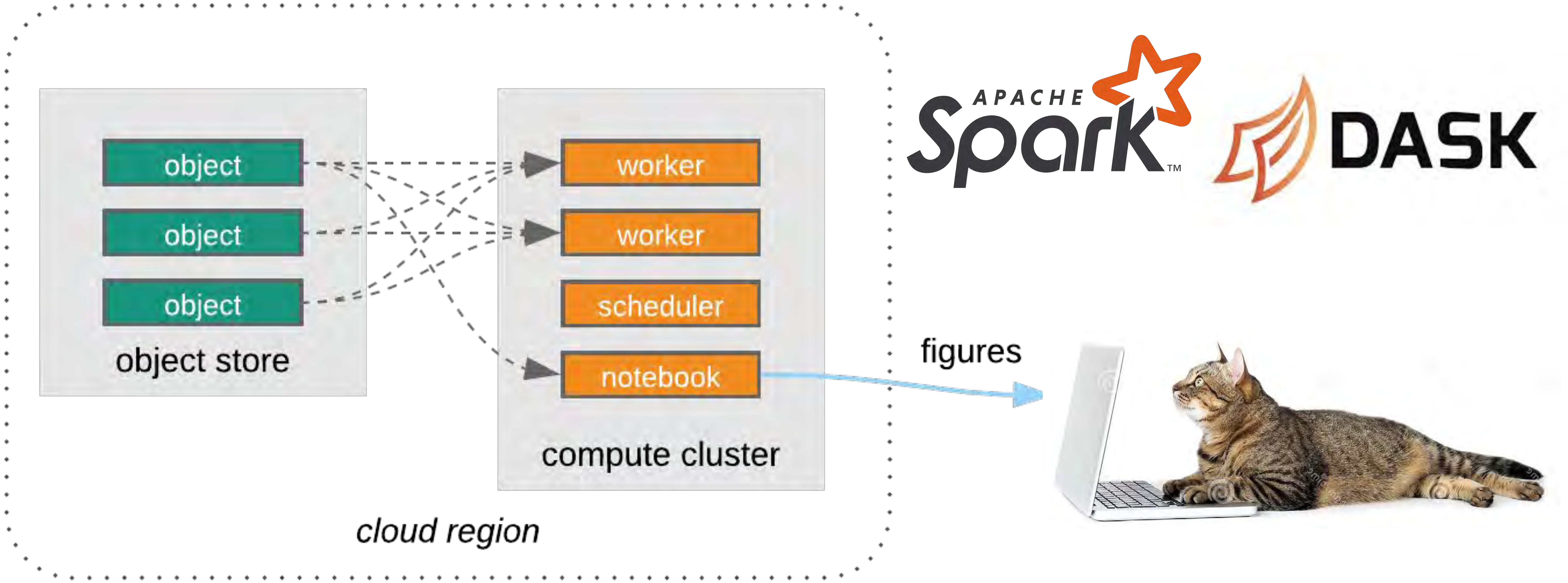
COG CLOUD OPTIMIZED GEOTIFF

Zarr

Scalable Parallel Computing Frameworks

APACHE Spark

DASK



Data Provider's
$

Data Consumer's
$

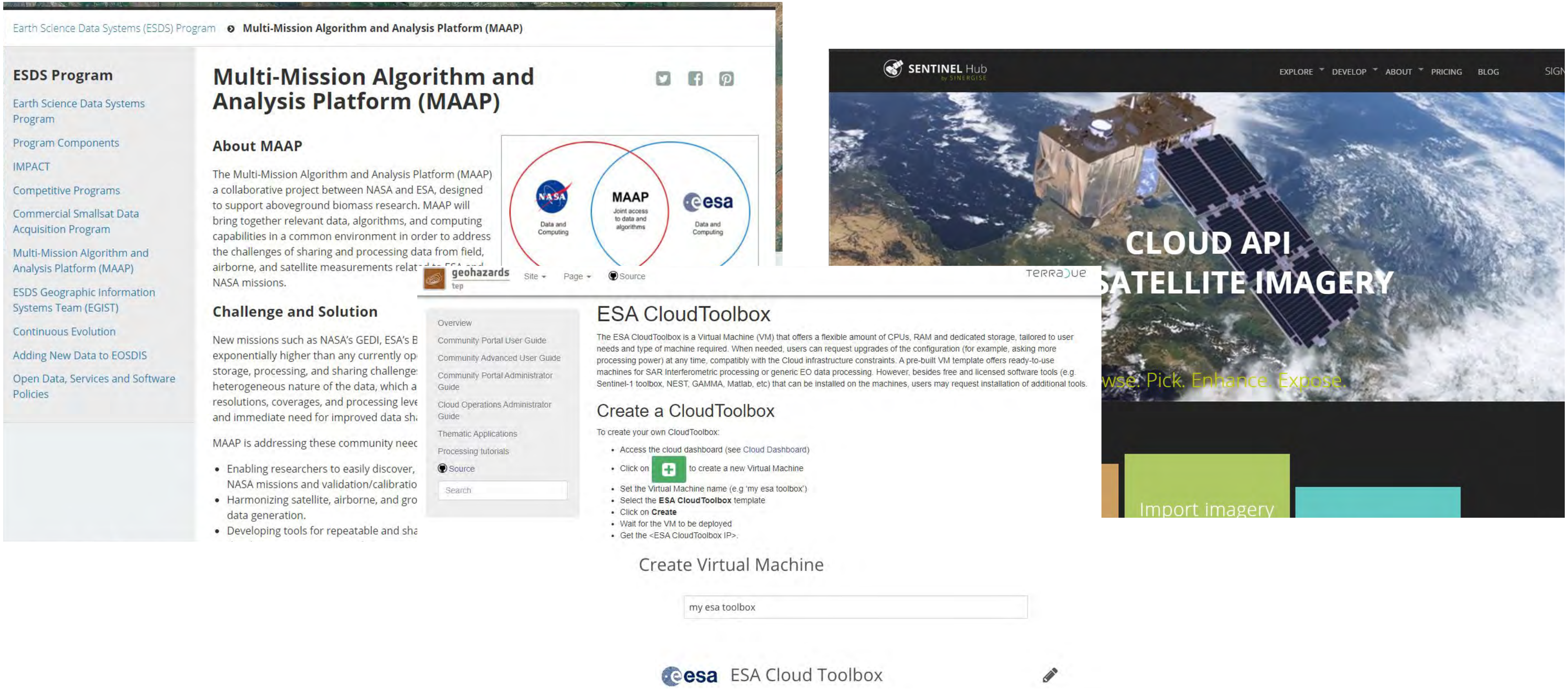
Google Cloud Platform

aws

Microsoft Azure

Jetstream

Agency driven solutions



Grass-Roots Solutions

Erin Dougherty
@edougherty_

My #dayofscience: paper revisions 🙃 and using @pangeo_data to analyze massive amounts of high-res climate data to understand floods in a current and future climate over the U.S. When not doing this, I ❤️ observing #wx directly via field work and watching storms. ⛈️

2:38 PM · Oct 15, 2019 · Twitter for iPhone

5 Retweets **51** Likes

---

pangeo Retweeted

Patrick Gray
@clifgray

One of the highlights of #AGU19 for me was a workshop on @pangeo_data. If you're interested in earth sci + geospatial analysis at scale I can't recommend their tutorial enough. Find it at github.com/pangeo-data/pa.... Clear and concise intro to #xarray, #dask, #geopandas, and #intake.

PANGEO
pangeo-data/pangeo-tutorial
Interactive jupyter notebooks for pangeo tutorial events - pangeo-data/pangeo-tutorial
github.com

2:03 PM · Dec 16, 2019 · Twitter Web App

15 Retweets **64** Likes

Andrew Williams
@AndrewIWilliams

It's taken a while, but I think that the whole xarray/Dask/cloud thing has finally clicked !

I've been working with CESM Large Ensemble data for a few months now - moved onto a Pangeo server and managed to speed up my workflow massively !

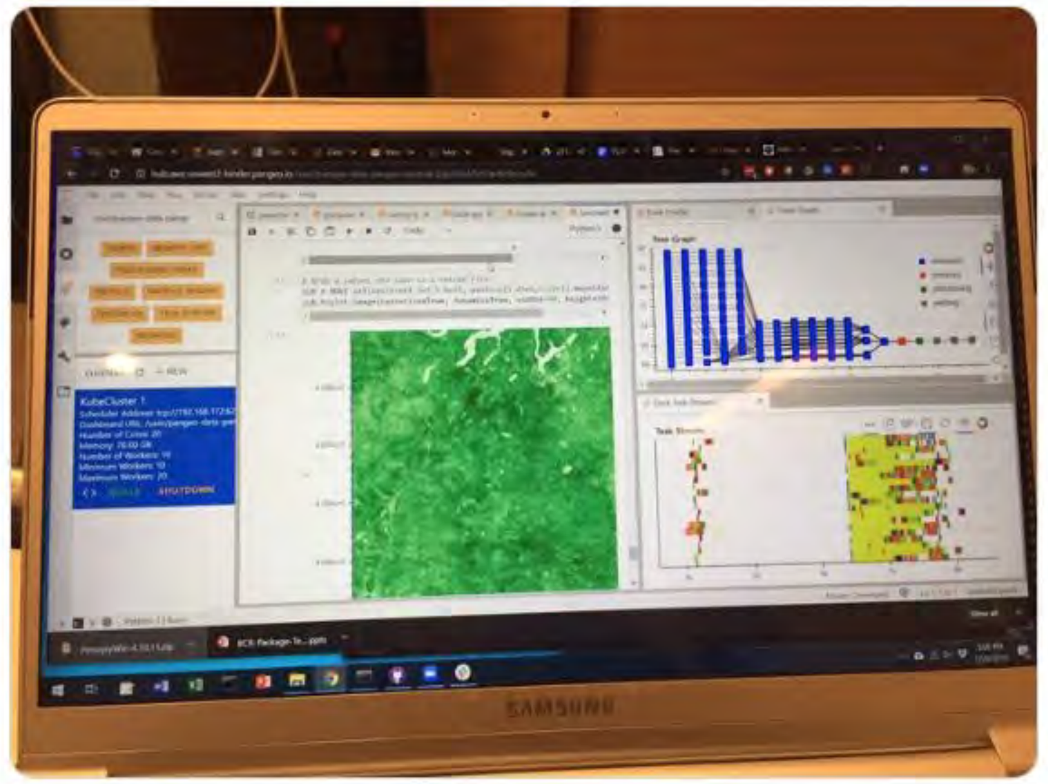
@xarray_dev, @dask_dev, @pangeo_data

10:23 AM · Jan 12, 2020 · Twitter Web App

6 Retweets **43** Likes

Chelle Gentemann
@ChelleGentemann

Rocking 70GB Landsat data at the @pangeo_data #AGU2019 tutorial. Such a powerful #OpenSource software stack.



10:45 PM · Dec 8, 2019 · Twitter for iPhone

17 Retweets **92** Likes

Scott Collis @Cyclogenesis_au · Nov 11, 2019
Teaching a course on #@ONScience #OpenScience with @Shobenase at @MonashUni with 20 very diverse attendees @environmentca @argonne @armnewsteam

1 3 13

pangeo Retweeted

Scott Collis
@Cyclogenesis_au

Replying to @Cyclogenesis_au @Shobenase and 4 others

Forgot to mention all done on @pangeo_data ! Making future Pangeans!

9:15 PM · Nov 11, 2019 · Twitter Web App

1 Retweet **5** Likes

---

pangeo Retweeted

Andrew Pauling
@andrewp109

#cmip6hack is just wrapping up, and has changed the way I will think about, and hopefully do, climate model analysis in the future. The @pangeo_data infrastructure makes it all so easy.

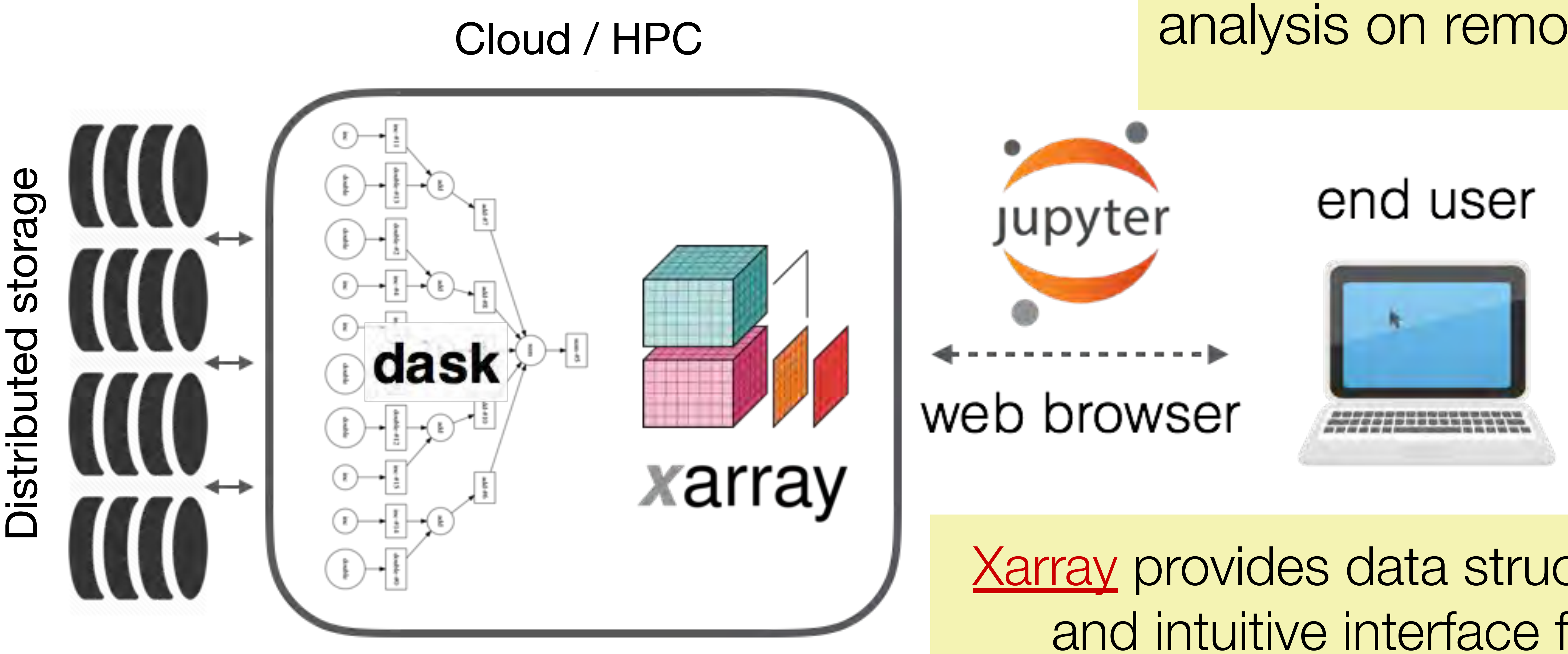
5:29 PM · Oct 18, 2019 · Twitter Web App

6 Retweets **24** Likes

Pangeo Architecture

@pangeo_data

Jupyter for interactive data analysis on remote systems



"Analytics Optimized Data Stores" stored on globally-available distributed storage.

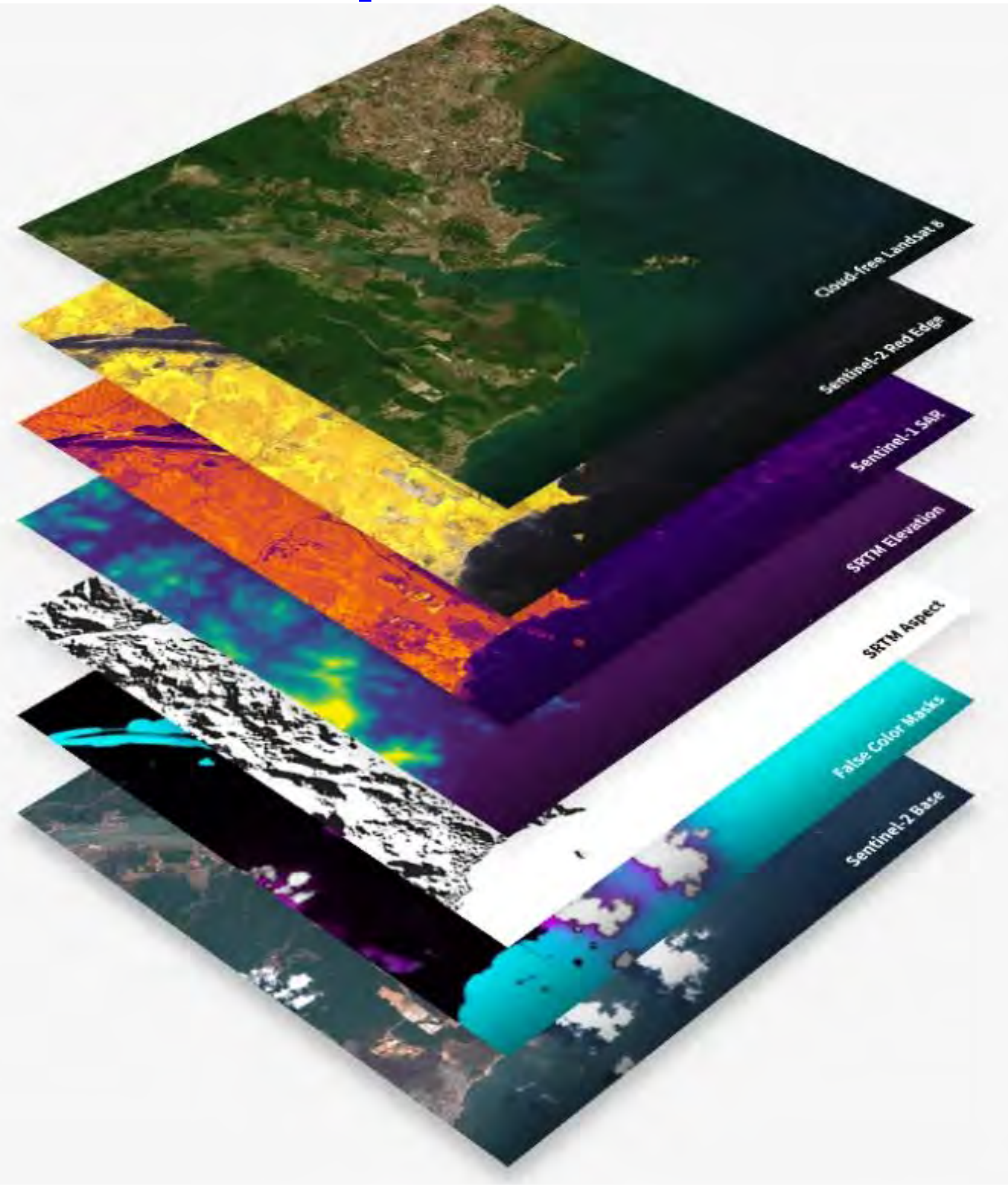
Xarray provides data structures and intuitive interface for interacting with datasets

Parallel computing system allows users deploy clusters of compute nodes for data processing.
Dask tells the nodes what to do.

How can data providers reduce barriers?

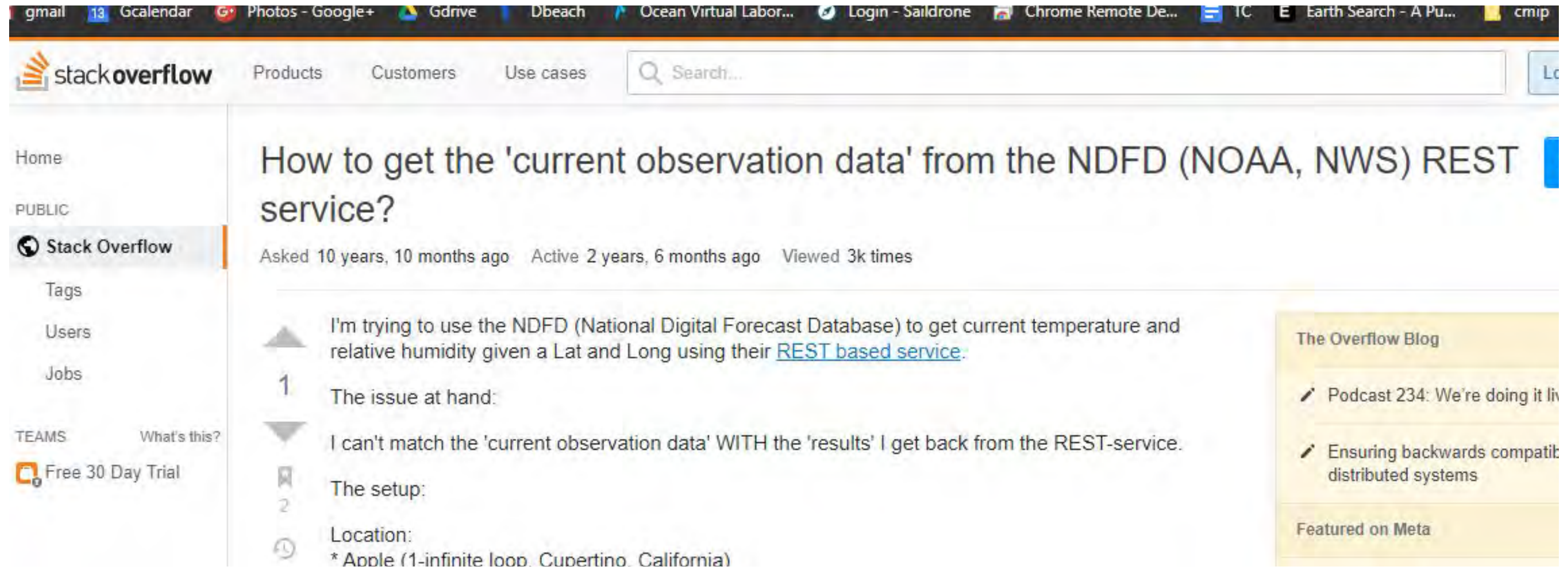
Reimagine how cloud data access and tools can enable transformational science

Publish cloud-optimized data



Interactive tutorials

Increase user interactions/feedback



Contribute to OSS tools

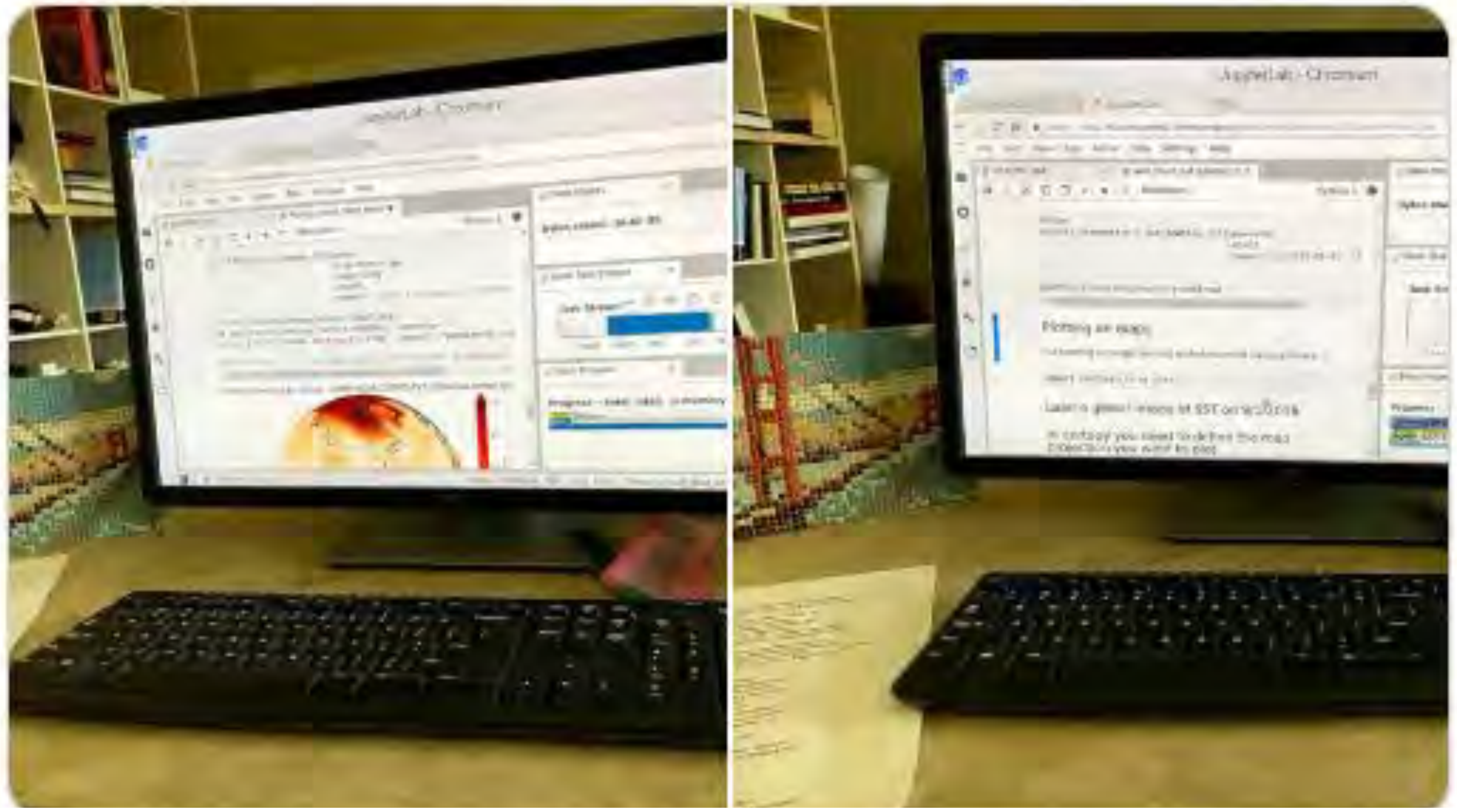
Julius Busecke
@JuliusBusecke Follows you

How does minimizing barriers to data change science?

3

Chelle Gentemann @ChelleGentemann · Mar 1

On my kids @Raspberry_Pi running @TeamKano #OpenSource OS I'm analyzing @GCPcloud #cmip6 climate data and @awscloud MUR SST from @podaac. A $36 computer running processes on both AWS and GCP with over 80 workers and 245GB. #openscience! @NASAEarth



Levels the playing field for all who want to contribute

EARTH SCIENCE
DATA SYSTEMS

FARALLON INSTITUTE

Impacts: Reduce Time to Science

Traditional Project Timeline

80% Data Preparation (download, clean, & organize files)	10% Batch Processing	10% Think about science

Cloud-based Project Timeline

5% Load AODS	5% Parallel Processing	90% Think about science

Impacts: Reproducibility

Reproducibility in data-driven science requires more than just code!

Traditional Project Code

```
# step 1: open data (stored on local hard drive)
>>> data = open_data("/path/to/private/files")
Error: files not found
```

Cloud-based Project Code

```
# step 1: open data (globally accessible)
>>> data = open_data("http://catalog.pangeo.io/path/to/dataset")
# step 2: process data
>>> process(data)
```

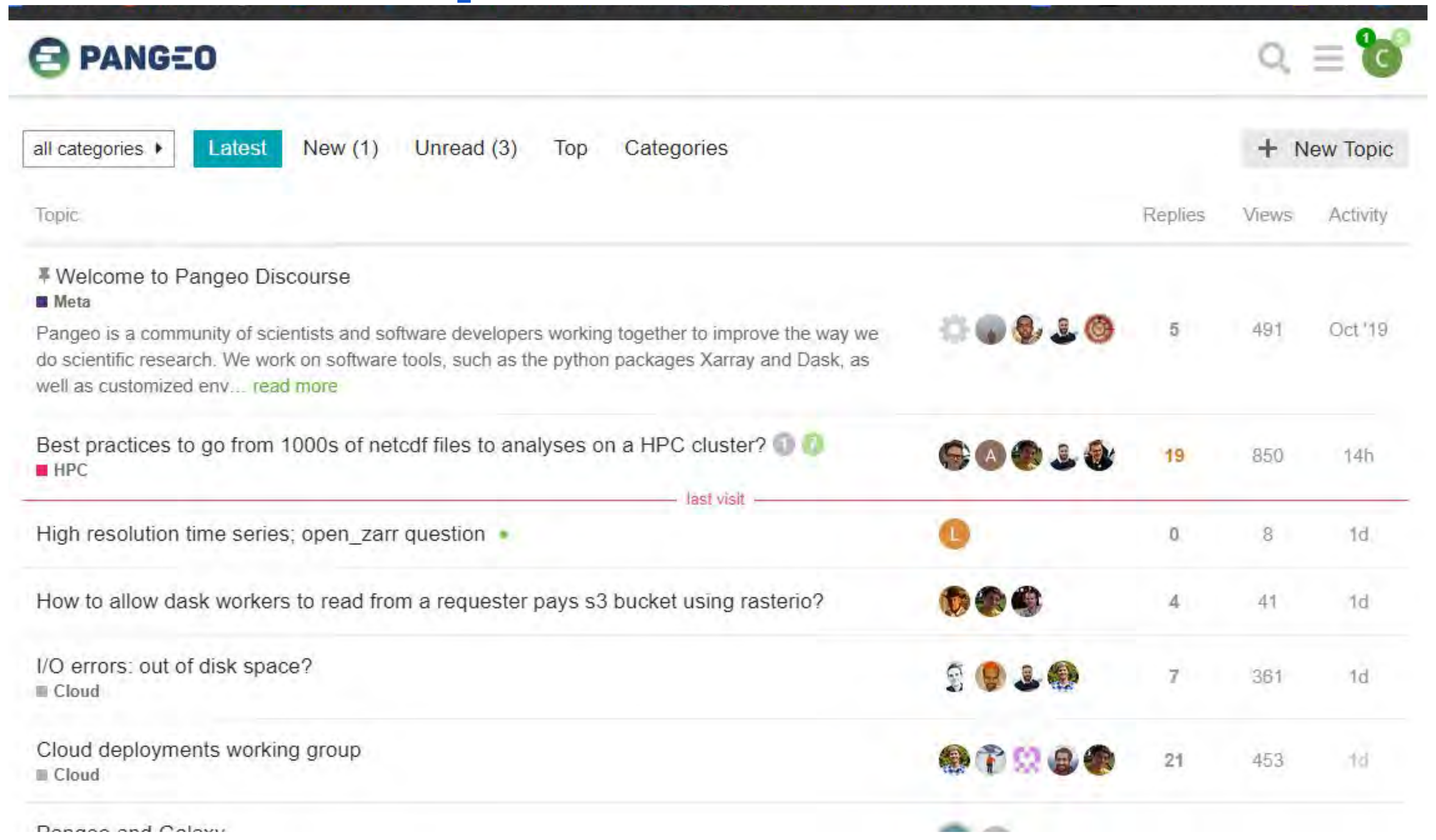
EARTH SCIENCE DATA SYSTEMS

FARALLON INSTITUTE

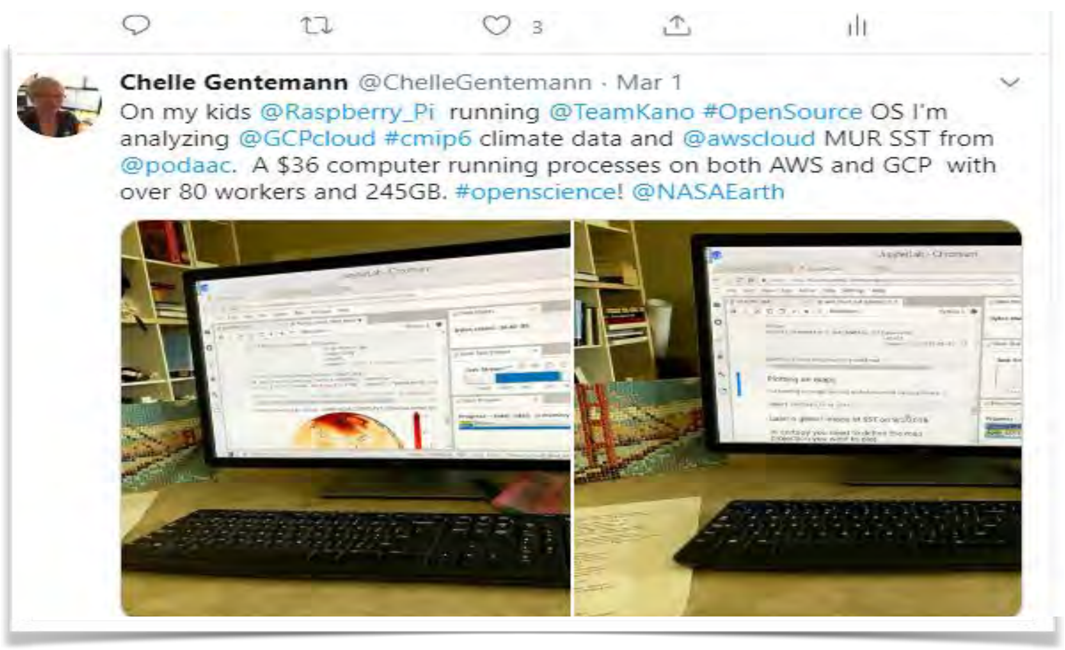
What impacts the velocity of progress?

Data, Software, & Compute

Open source science

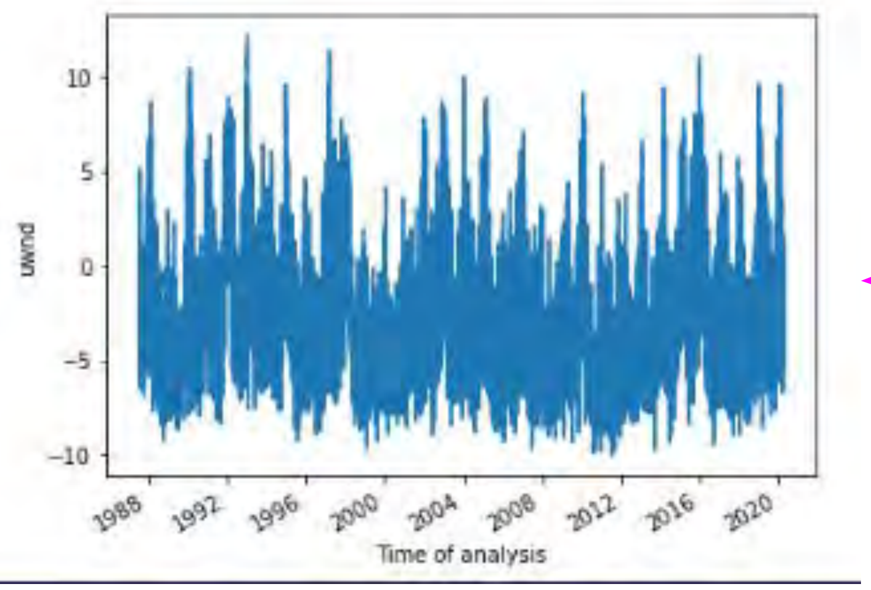


Thank you!



NUMFOCUS

OPEN CODE = BETTER SCIENCE



STOP ------------- THIS IS DIFFERENT -------------------

1 line of code to read in entire 32 year global 25km dataset

1 line of code to select a region, calculate & plot a mean time series

in LESS than 1 minute