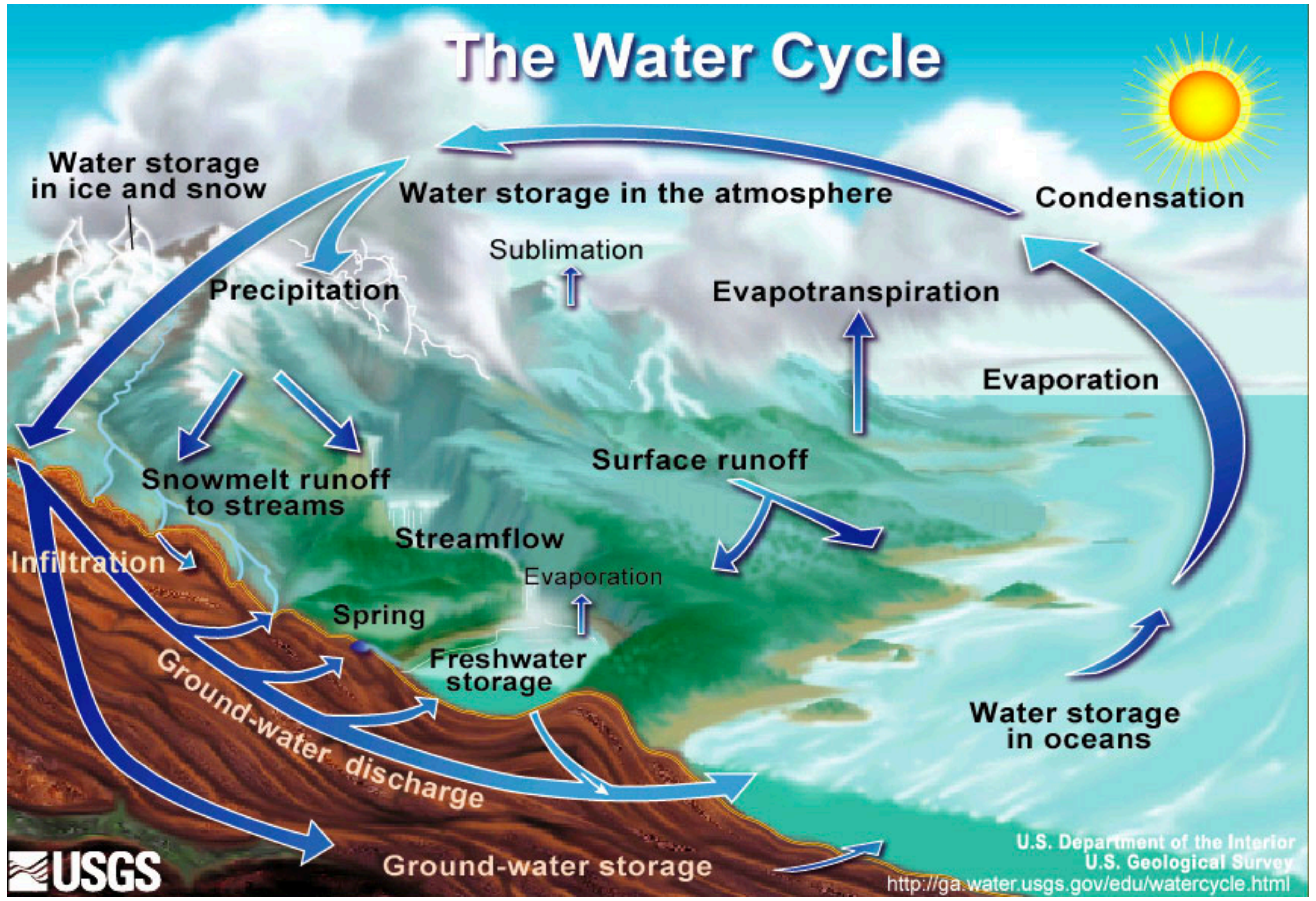
PANGEO

A COMMUNITY-DRIVEN EFFORT FOR
BIG DATA GEOSCIENCE

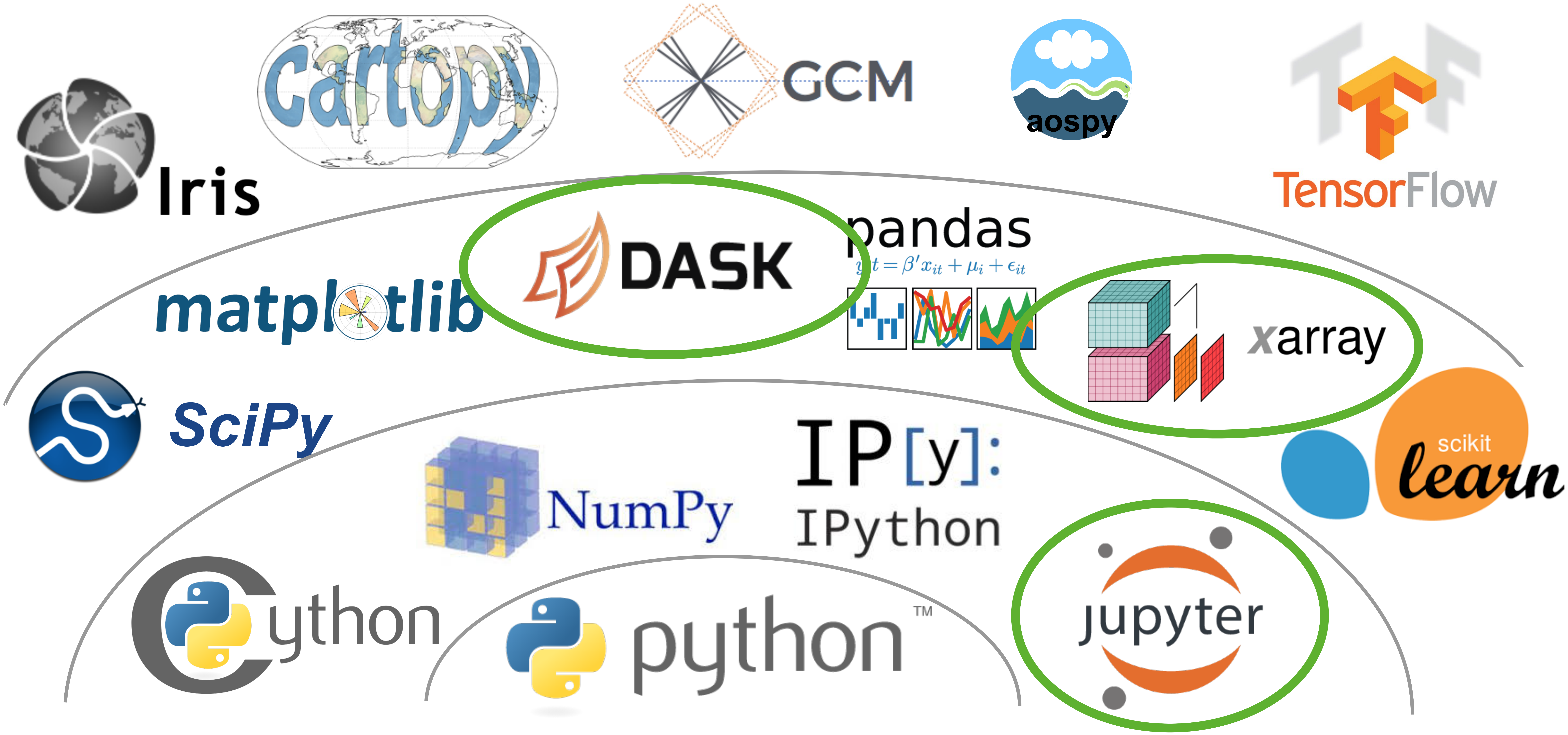
HELLO!

- Who am I?
  - Joe Hamman, Ph.D., P.E.
  - I am a scientist at the National Center for Atmospheric Research (RAL & CGD)
  - I study the impacts of climate change on the water cycle.
  - I am a core developer of Xarray
  - I am a founding member of the Pangeo project



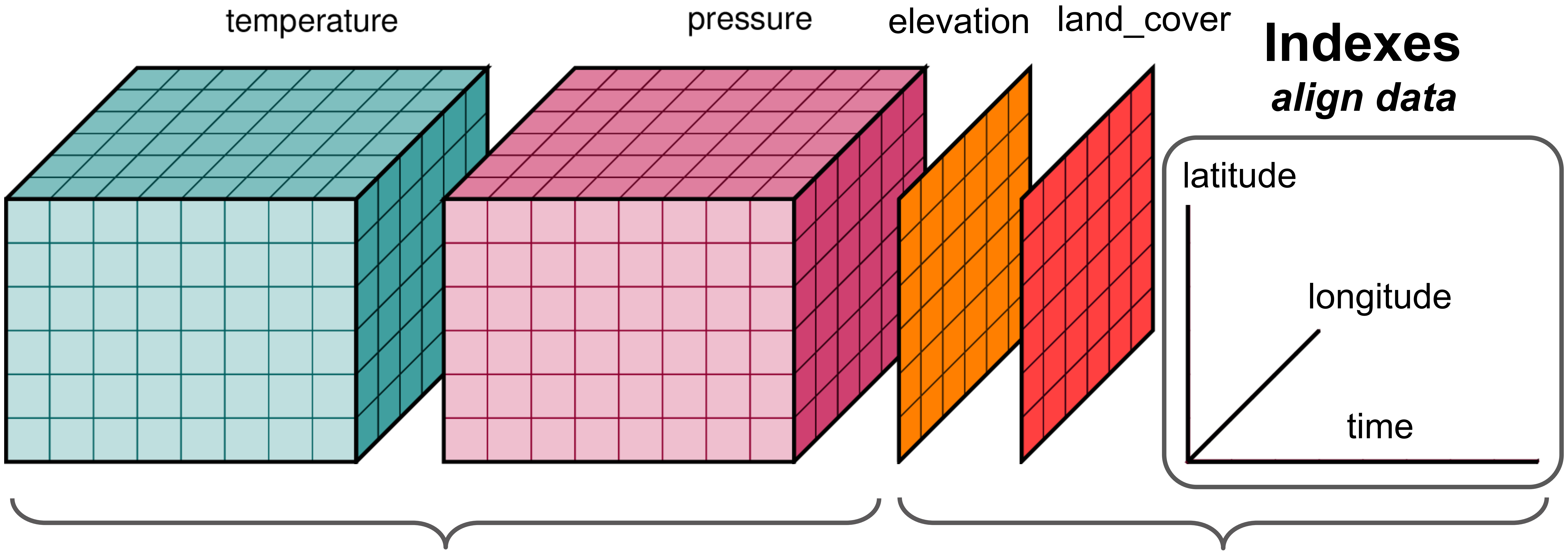
Github: @jhamman
Twitter: @HammanHydro
Web: joehamman.com

SCIENTIFIC PYTHON FOR DATA SCIENCE



Credit: Stephan Hoyer, Jake Vanderplas (SciPy 2015)

XARRAY DATASET: MULTIDIMENSIONAL VARIABLES WITH COORDINATES AND METADATA



Indexes
align data

+ Attributes
metadata ignored by operations

Data variables
used for computation

Coordinates
describe data

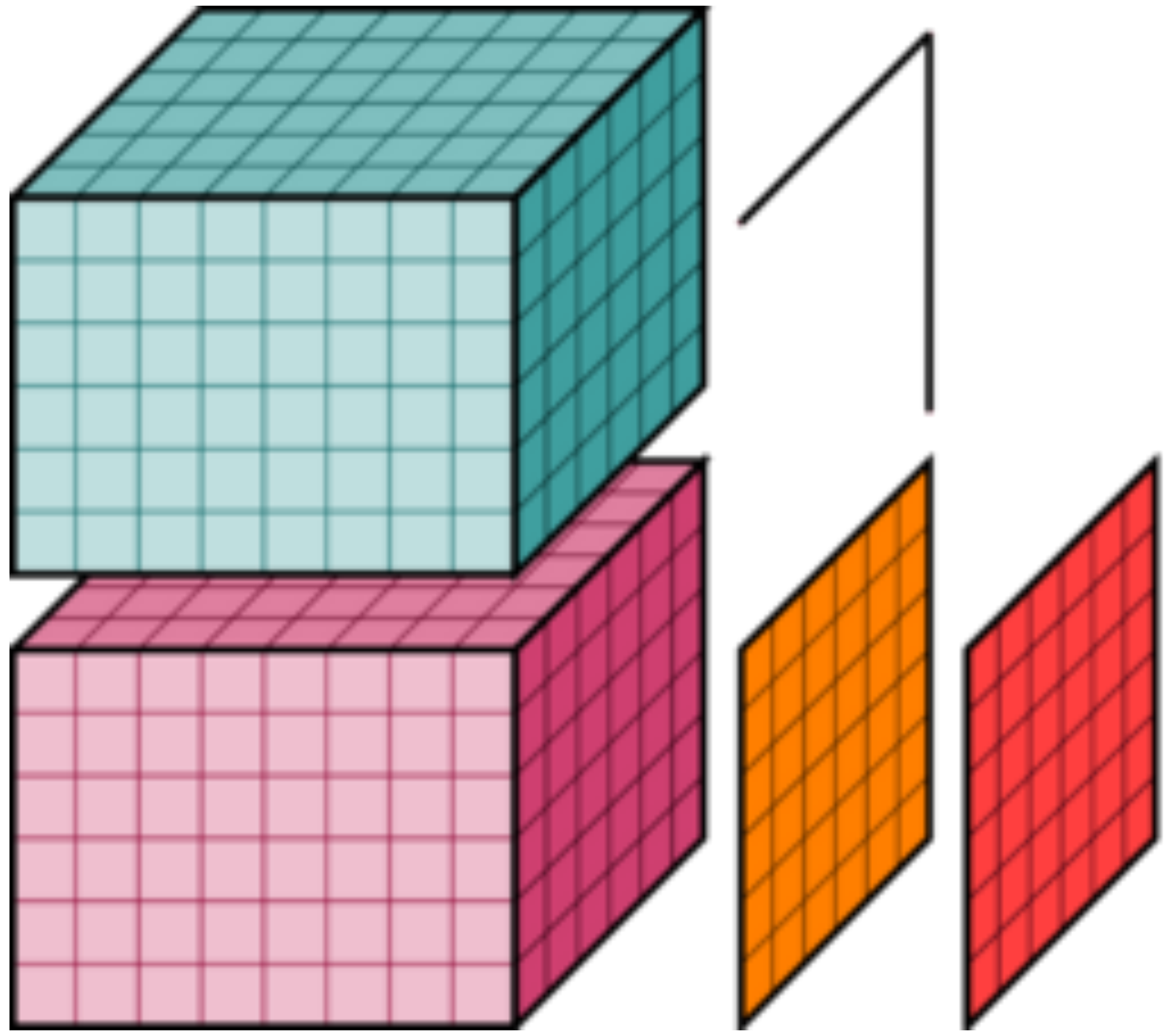
"netCDF meets pandas.DataFrame"

Credit: Stephan Hoyer

XARRAY

http://xarray.pydata.org

- label-based indexing and arithmetic
- interoperability with the core scientific Python packages (e.g., pandas, NumPy, Matplotlib)
- out-of-core computation on datasets that don’t fit into memory (thanks dask!)



- wide range of input/output (I/O) options: netCDF, HDF, geoTIFF, zarr
- advanced multi-dimensional data manipulation tools such as group-by and resampling

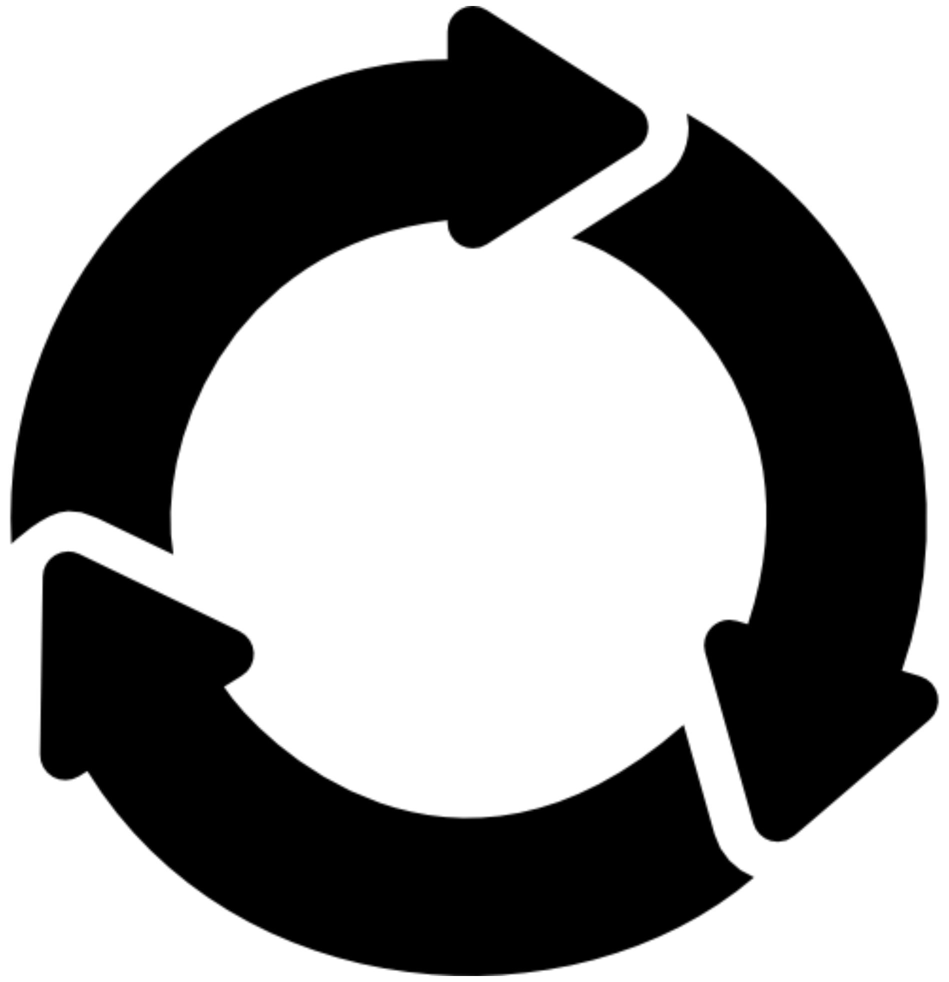
XARRAY UPDATES

- **Development Roadmap** (http://xarray.pydata.org/en/stable/roadmap.html)
  - More flexible grids/indexing
  - More flexible arrays/computing
  - More flexible storage backends
- **NumFOCUS Sponsorship**
  - https://numfocus.org/project/xarray
- **New Contributors**
  - New core devs: Spencer Clark and Deepak Cherian

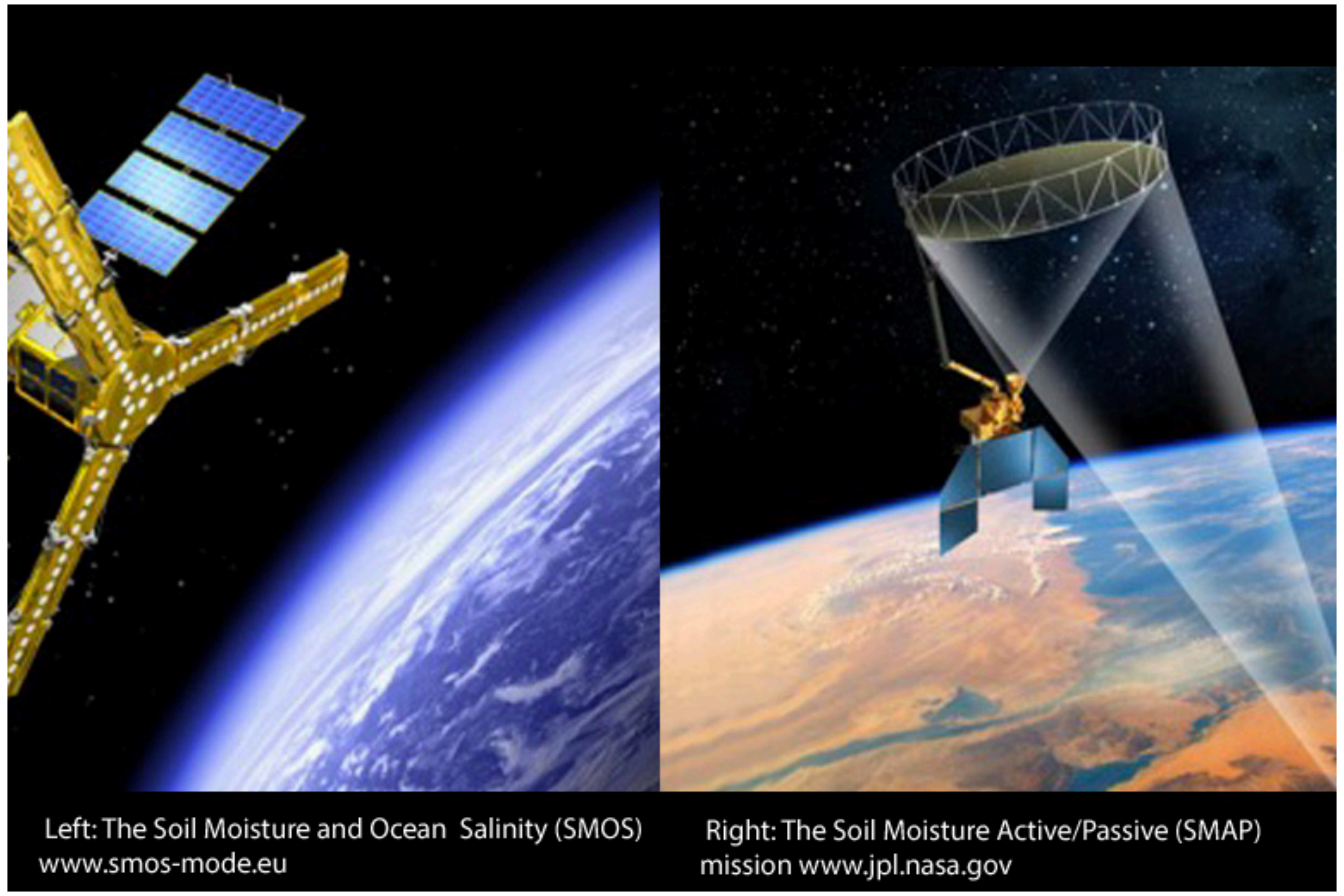
WHAT DRIVES PROGRESS IN GEOSCIENCE?

New Ideas

$$q^{soil}_{liq,z} = \begin{cases} q_{rain} - q_{ix} - q_{sx} & z=0 \\ -K^{soil}\dfrac{\partial\psi}{\partial z} + K^{soil} & z>0 \end{cases}$$



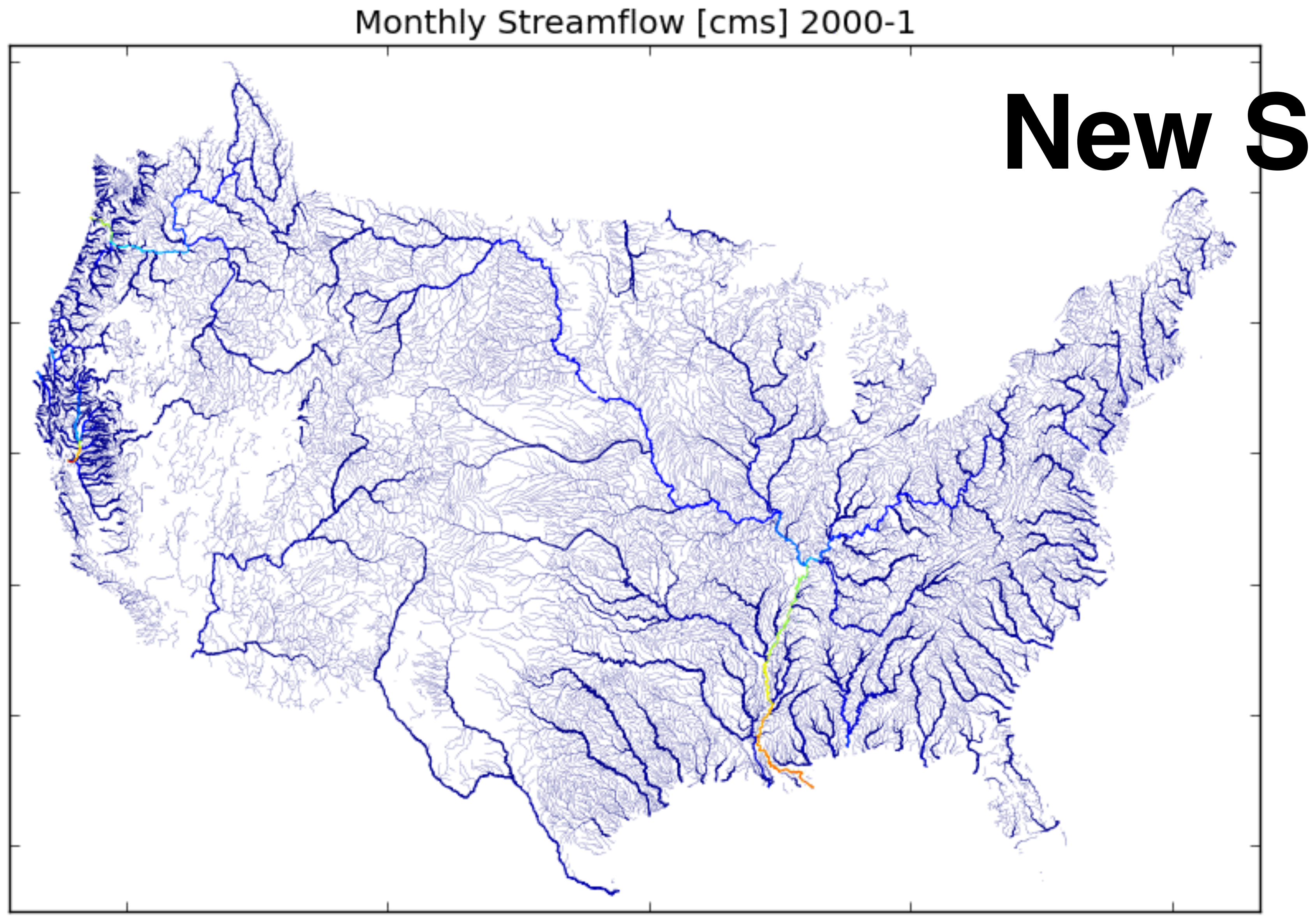
New Observations



Left: The Soil Moisture and Ocean Salinity (SMOS) www.smos-mode.eu

Right: The Soil Moisture Active/Passive (SMAP) mission www.jpl.nasa.gov

New Simulations



FRAGMENTATION PROBLEMS

1. **Software**
   - Few tangible incentives to share source code (funding agencies, journals)
   - Lack of extensible development patterns; often it is easier to “home grow” your own solution, rather than using someone else’s.
   - Result is that most geoscientific research is effectively unreproducible and prone to failure.
2. **Data sprawl**
   - Inefficiencies of many copies of the same datasets (“dark replicas”)
   - Lessons learned from the CMIP archives (CMIP3 was duplicated > 30x)
3. **Local vs. High-performance vs. Cloud Computing**
   - Traditional scientific computing workflows are difficult to port from a laptop, to HPC, to the cloud

PANGEO PROJECT GOALS

- Foster collaboration around the open source scientific Python ecosystem for ocean / atmosphere / land / climate science.
- Support the development with domain-specific geoscience packages.
- Improve scalability of these tools to to handle petabyte-scale datasets on HPC and cloud platforms.

PANGEO COLLABORATORS

NSF

EarthCube

NASA

Google Cloud Platform

aws

Lamont-Doherty Earth Observatory
Columbia University | Earth Institute

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Met Office

ANACONDA
Powered by Continuum Analytics

UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

USGS

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

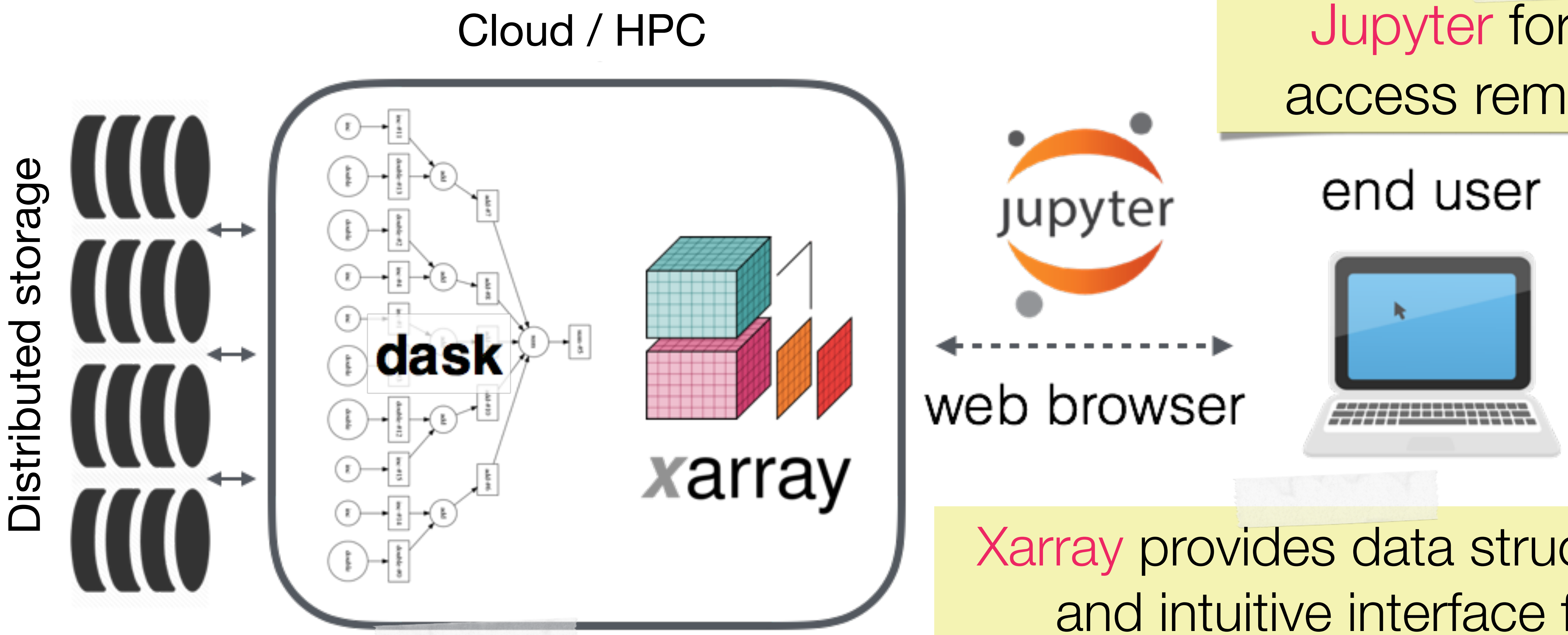
Rh g
RHODIUM
GROUP

JOHNS HOPKINS
UNIVERSITY

ECMWF

And many more…

PANGEO ARCHITECTURE



Jupyter for interactive access remote systems

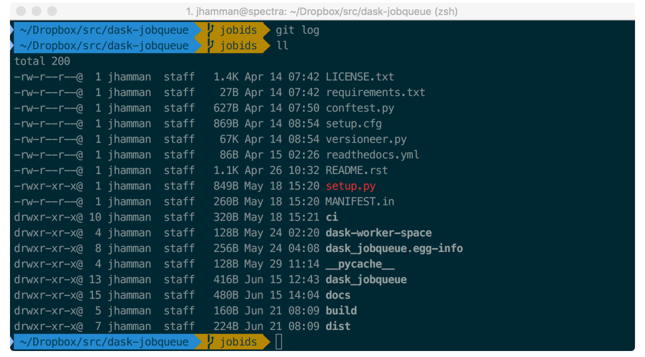
"Analysis Ready Data" stored on globally-available distributed storage.

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.

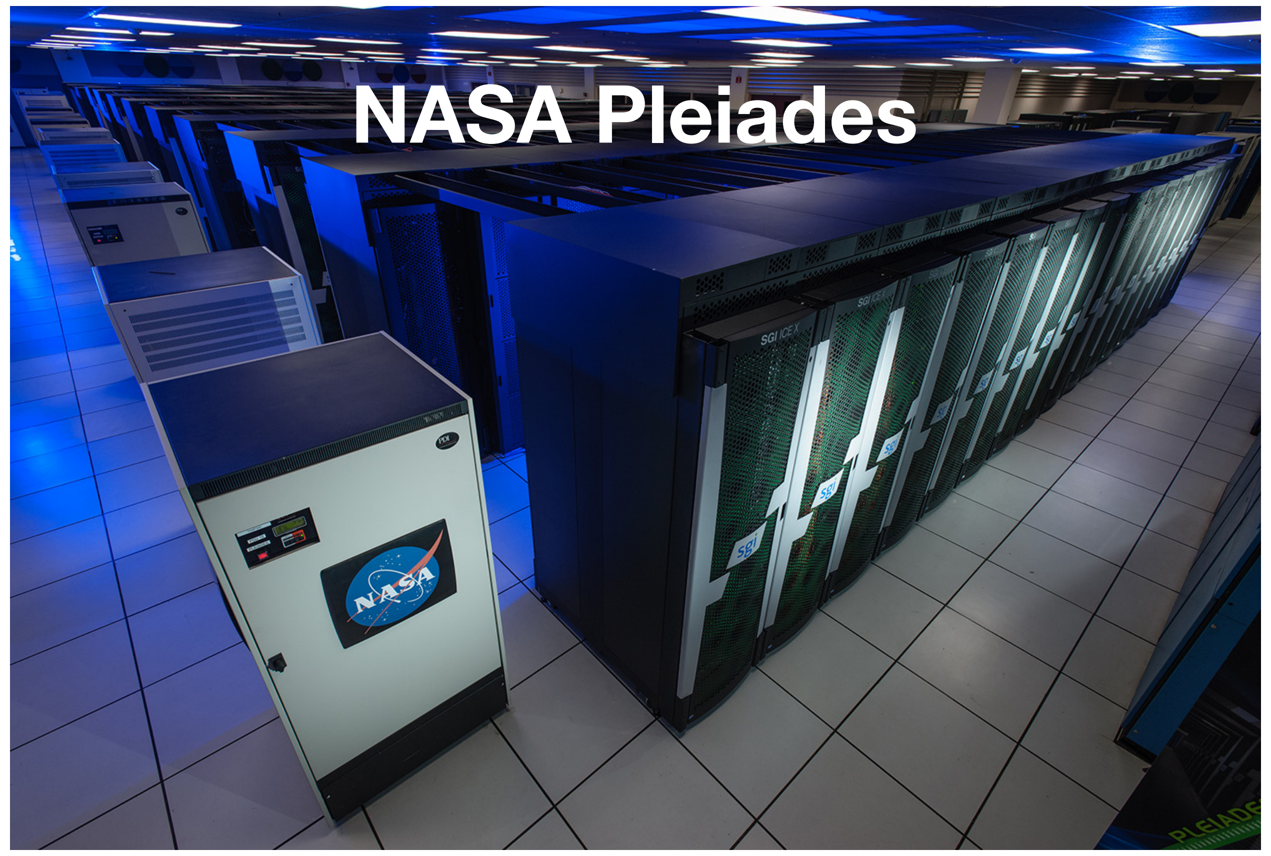
BUILD YOUR OWN PANGEO

Storage Formats	HDF	OPeNDAP	Cloud Optimized COG/Zarr/Parquet/etc.
ND-Arrays	NumPy	DASK	More coming…
Data Models	xarray	Iris	pandas $y_it = \beta' x_{it} + \mu_i + \epsilon_{it}$
Processing Mode	jupyter Interactive	Batch 	Serverless λ
Compute Platform	HPC	Cloud Google Cloud Platform	Local

```
1. jhamman@spectra: ~/Dropbox/src/dask-jobqueue (zsh)
 ~/Dropbox/src/dask-jobqueue  jobids  git log
 ~/Dropbox/src/dask-jobqueue  jobids  ll
total 200
-rw-r--r--@  1 jhamman  staff   1.4K Apr 14 07:42 LICENSE.txt
-rw-r--r--@  1 jhamman  staff    27B Apr 14 07:42 requirements.txt
-rw-r--r--@  1 jhamman  staff   627B Apr 14 07:50 conftest.py
-rw-r--r--@  1 jhamman  staff   869B Apr 14 08:54 setup.cfg
-rw-r--r--@  1 jhamman  staff    67K Apr 14 08:54 versioneer.py
-rw-r--r--@  1 jhamman  staff    86B Apr 15 02:26 readthedocs.yml
-rw-r--r--@  1 jhamman  staff   1.1K Apr 26 10:32 README.rst
-rwxr-xr-x@  1 jhamman  staff   849B May 18 15:20 setup.py
-rw-r--r--@  1 jhamman  staff   260B May 18 15:20 MANIFEST.in
drwxr-xr-x@ 10 jhamman  staff   320B May 18 15:21 ci
drwxr-xr-x@  4 jhamman  staff   128B May 24 02:20 dask-worker-space
drwxr-xr-x@  8 jhamman  staff   256B May 24 04:08 dask_jobqueue.egg-info
drwxr-xr-x@  4 jhamman  staff   128B May 29 11:14 __pycache__
drwxr-xr-x@ 13 jhamman  staff   416B Jun 15 12:43 dask_jobqueue
drwxr-xr-x@ 15 jhamman  staff   480B Jun 15 14:04 docs
drwxr-xr-x@  5 jhamman  staff   160B Jun 21 08:09 build
drwxr-xr-x@  7 jhamman  staff   224B Jun 21 08:09 dist
 ~/Dropbox/src/dask-jobqueue  jobids
```

PANGEO DEPLOYMENTS

HTTP://PANGEO.IO/DEPLOYMENTS.HTML


NASA Pleiades

NCAR Cheyenne

(SCALE USING JOB QUEUE SYSTEM)

PANGEO.PYDATA.ORG

BINDER.PANGEO.IO

Over 1000 unique users since March!

Google Cloud Platform

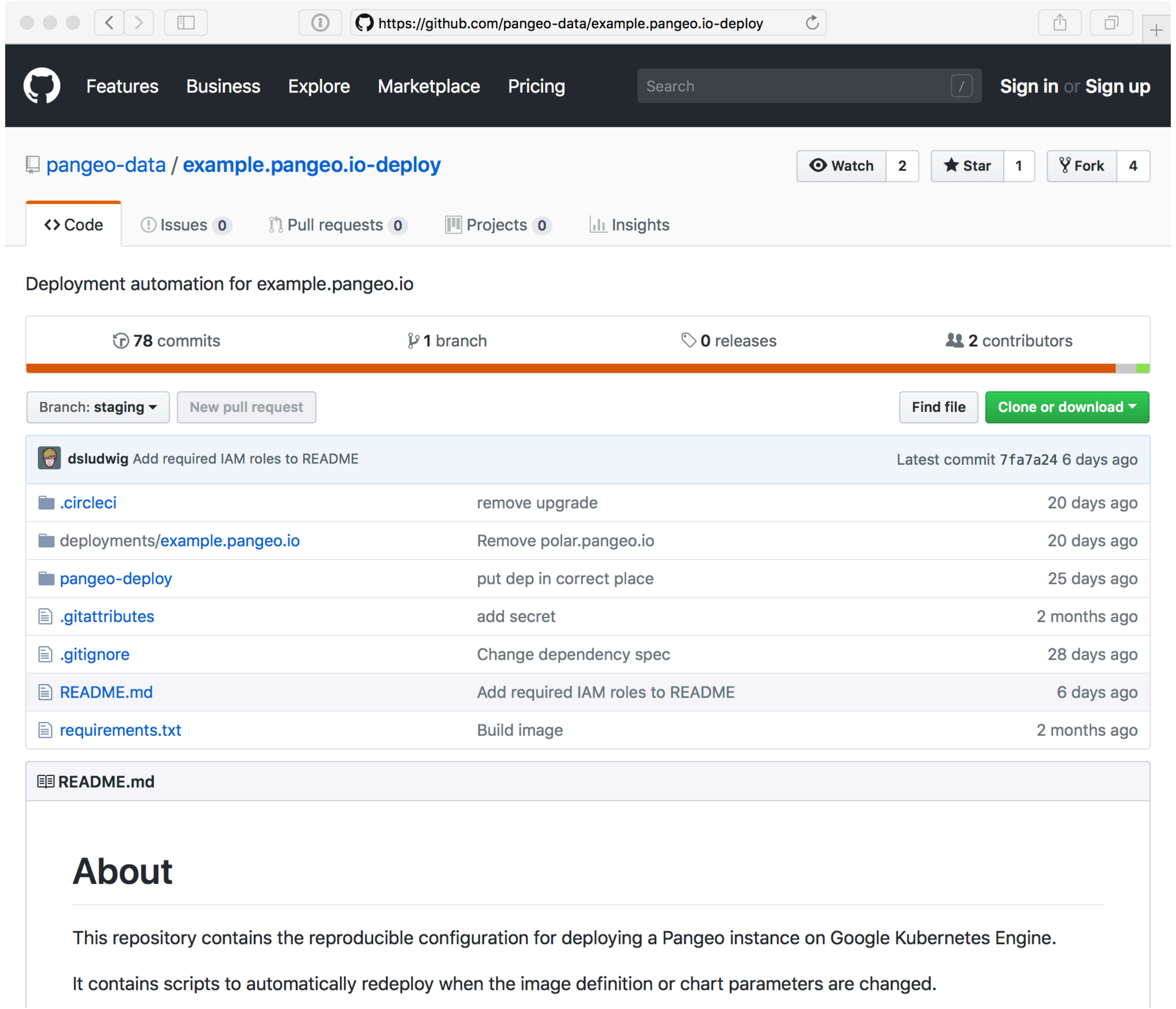
Microsoft Azure

aws

(SCALE USING KUBERNETES)

FOO.PANGEO.IO
DEPLOY YOUR OWN PANGEO

- What's in a typical Pangeo?
  - JupyterHub interface
  - Tools to deploy dask clusters
  - Customizable software/hardware environment
- Current effort to federate pangeo deployments for problem specific uses (e.g. cds.pangeo.io?)
- Custom deployments:
  - polar.pangeo.io
  - solar.pangeo.io
  - ocean.pangeo.io
  - hydroshare.pangeo.io
  - And more coming…



https://github.com/pangeo-data/example.pangeo.io-deploy

pangeo-data / example.pangeo.io-deploy

Deployment automation for example.pangeo.io

dsludwig Add required IAM roles to README — Latest commit 7fa7a24 6 days ago

.circleci	remove upgrade	20 days ago
deployments/example.pangeo.io	Remove polar.pangeo.io	20 days ago
pangeo-deploy	put dep in correct place	25 days ago
.gitattributes	add secret	2 months ago
.gitignore	Change dependency spec	28 days ago
README.md	Add required IAM roles to README	6 days ago
requirements.txt	Build image	2 months ago

README.md

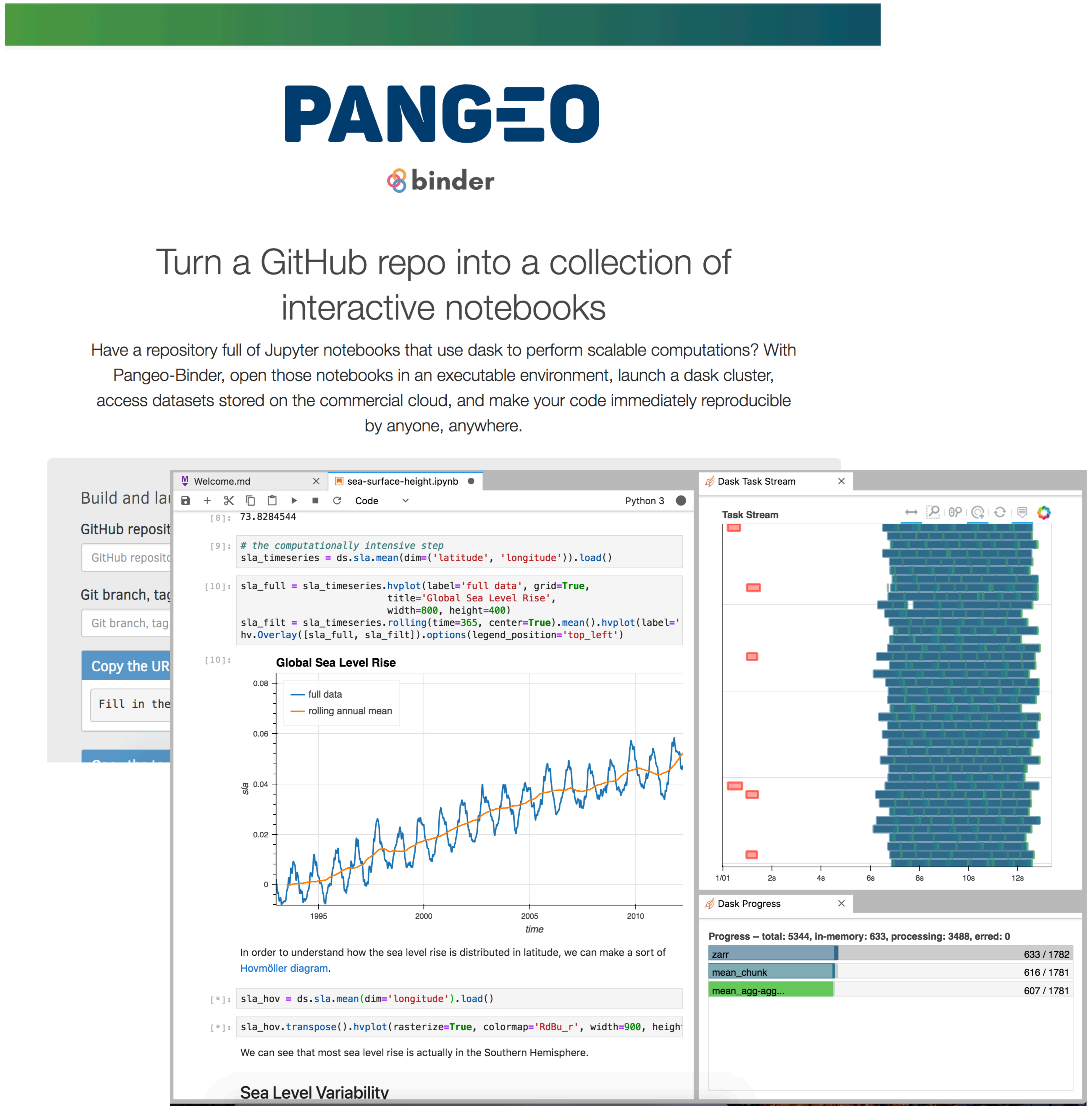
About

This repository contains the reproducible configuration for deploying a Pangeo instance on Google Kubernetes Engine.

It contains scripts to automatically redeploy when the image definition or chart parameters are changed.

BINDER.PANGEO.IO

- BinderHub
  - Highly customizable Jupyter environment
  - Automates Git repo -> docker image -> Jupyter notebook
  - Automates deployment of Dask clusters
- Easiest way to share Pangeo workflows
- Try it: https://bit.ly/2O9qJr3



PANGEO

binder

Turn a GitHub repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks that use dask to perform scalable computations? With Pangeo-Binder, open those notebooks in an executable environment, launch a dask cluster, access datasets stored on the commercial cloud, and make your code immediately reproducible by anyone, anywhere.

```
[8]: 73.8284544

[9]: # the computationally intensive step
     sla_timeseries = ds.sla.mean(dim=('latitude', 'longitude')).load()

[10]: sla_full = sla_timeseries.hvplot(label='full data', grid=True,
                                       title='Global Sea Level Rise',
                                       width=800, height=400)
      sla_filt = sla_timeseries.rolling(time=365, center=True).mean().hvplot(label='
      hv.Overlay([sla_full, sla_filt]).options(legend_position='top_left')
```

[10]:

In order to understand how the sea level rise is distributed in latitude, we can make a sort of Hovmöller diagram.

```
[*]: sla_hov = ds.sla.mean(dim='longitude').load()

[*]: sla_hov.transpose().hvplot(rasterize=True, colormap='RdBu_r', width=900, heigh
```

We can see that most sea level rise is actually in the Southern Hemisphere.

Sea Level Variability

```
Progress -- total: 5344, in-memory: 633, processing: 3488, erred: 0
zarr              633 / 1782
mean_chunk        616 / 1781
mean_agg-agg...   607 / 1781
```

PANGEO IN A NUTSHELL

- **Scientific Python ecosystem**
  - flexible, open-source, community driven
- **Interoperable**
  - integrates with existing/developing tools used by science community
- **Analysis ready data formats**
  - cloud optimized data (e.g. zarr)
- **Intuitive self-describing data models**
  - e.g. xarray, Iris
- **Scalable**
  - e.g. Dask, Kubernetes
- **Interactive**
  - e.g. Jupyter, JupyterHub, BinderHub
- **Cross platform**
  - HPC, Cloud, local computing

WHAT'S COMING FOR PANGEO

- Governance (https://github.com/pangeo-data/governance)
- Funding (new projects from NASA and NSF)
- AWS Open Datasets Program and Pangeo compute resources
- Science focus on remote sensing datasets
- Looking for new community partners

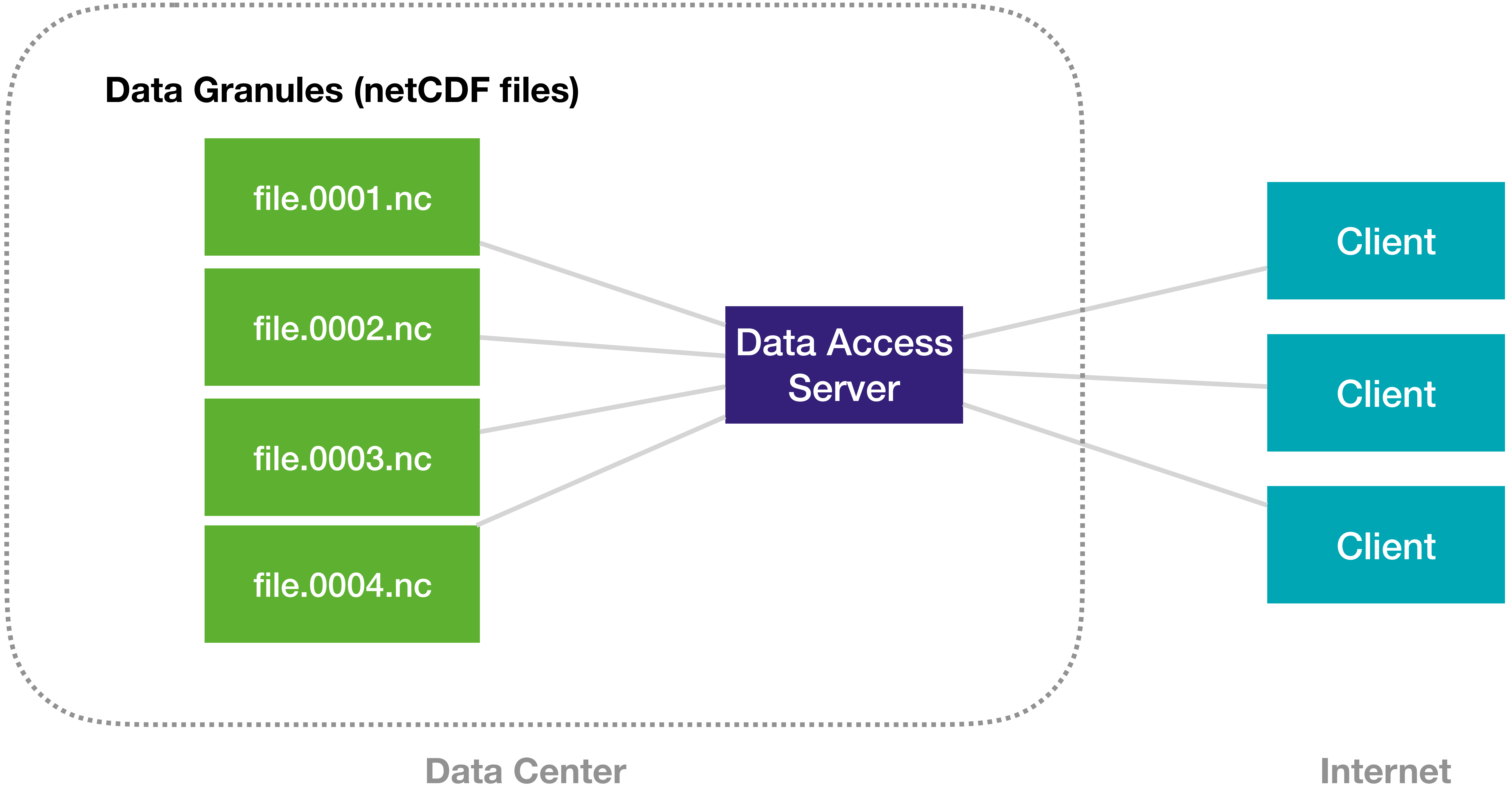
HOW TO GET INVOLVED

HTTP://PANGEO.IO

- Access and existing Pangeo deployment on an HPC cluster, or cloud resources (eg. binder.pangeo.io)
- Adapt Pangeo elements to meet your projects needs (data portals, etc.) and give feedback via GitHub: github.com/pangeo-data/pangeo
- Participate in open-source software development!

SHARING DATA IN THE CLOUD

Traditional Approach: A Data Access Portal

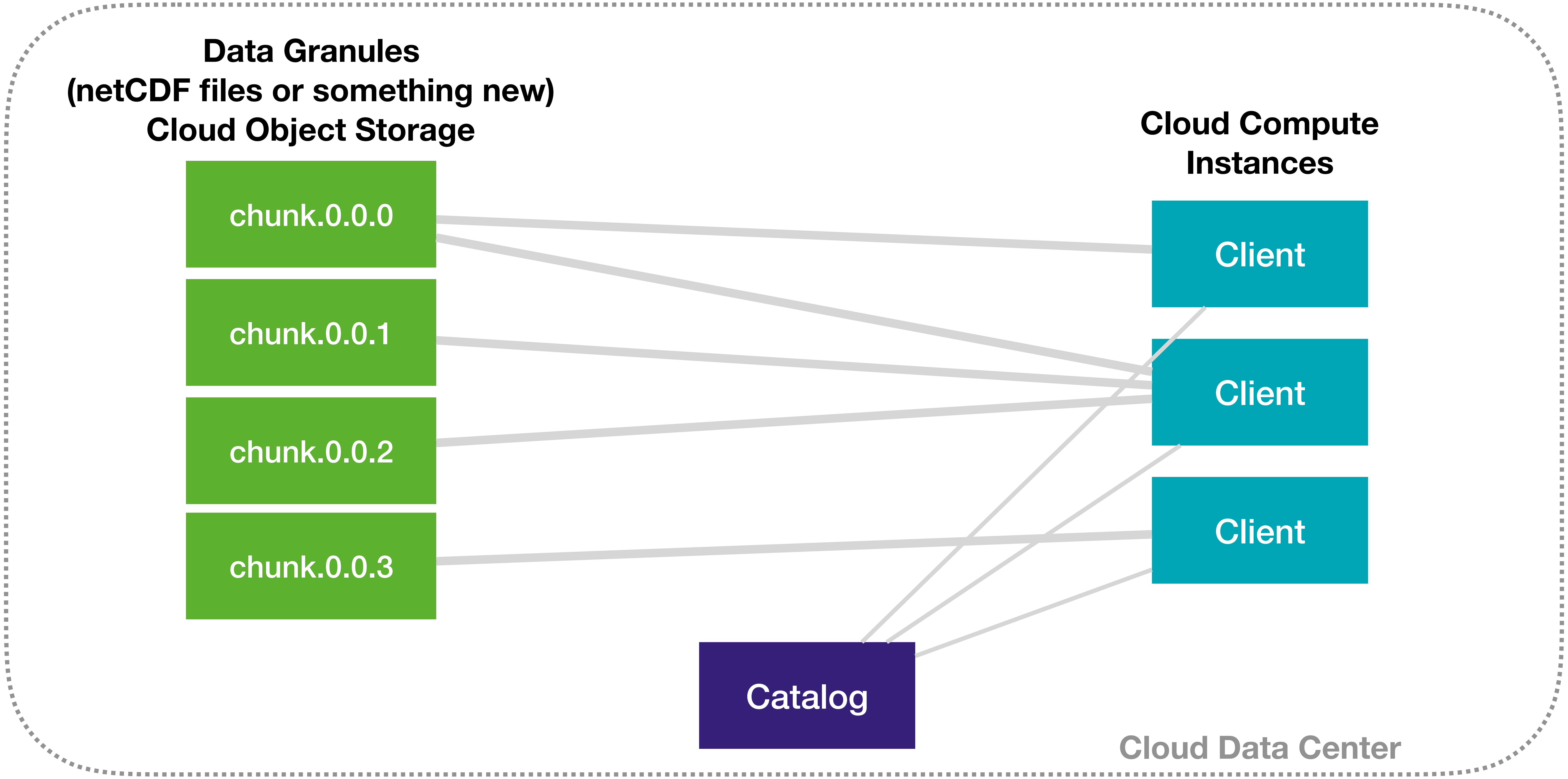


ON-DEMAND ANALYSIS-READY DATA

- **Too big to move**: assume data is to be used but not copied
- **Self-describing**: data and metadata packaged together
- **On-demand**: data can be read/used in its current form from anywhere
- **Analysis-ready**: no pre-processing required

SHARING DATA IN THE CLOUD

Direct Access to Cloud Object Storage



Zarr

?