PANGEO

Enabling Open, Reproducible, and scalable big data geoscience for everyone

Anne Fouilloux, *Simula Research Laboratory, Oslo, Norway*
Tina Erica Odaka, *LOPS UMR 6523, CNRS-IFREMER-IRD-Univ.Brest-IUEM, Brest, France*
Benjamin Ragan-Kelley, *Simula Research Laboratory, Oslo, Norway*
Francesco Nattino, *Netherlands eScience Center, Netherlands*
Meiert W. Grootes, *Netherlands eScience Center, Netherlands*
Ou Ku, *Netherlands eScience Center, Netherlands*
Raymond Oonk, *SURF, Netherlands*

https://pangeo.io/

https://eosc.eu/

What is Pangeo?

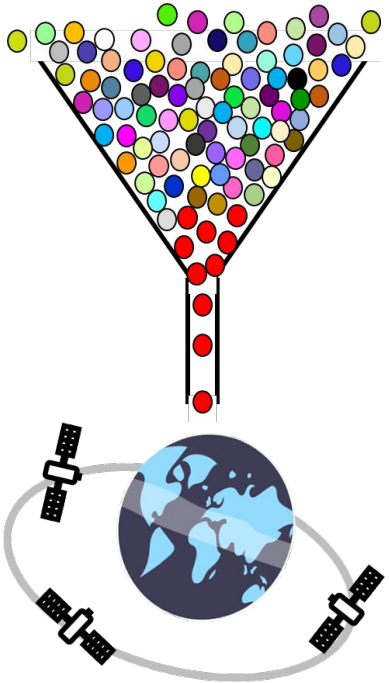
A global community initiative for Big Geoscience Data

- Based on
  - Open Community
  - Open Platform
  - Open deployments
- Try it yourself!

Open Community

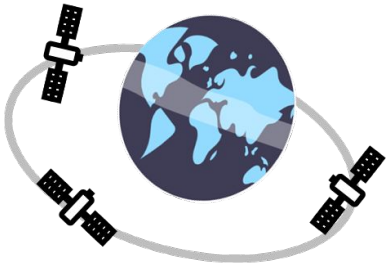
The Pangeo community

***Vision**: more equitable access to data, software and hardware to facilitate innovation*



Earth Observation market

Open Earth Observation data



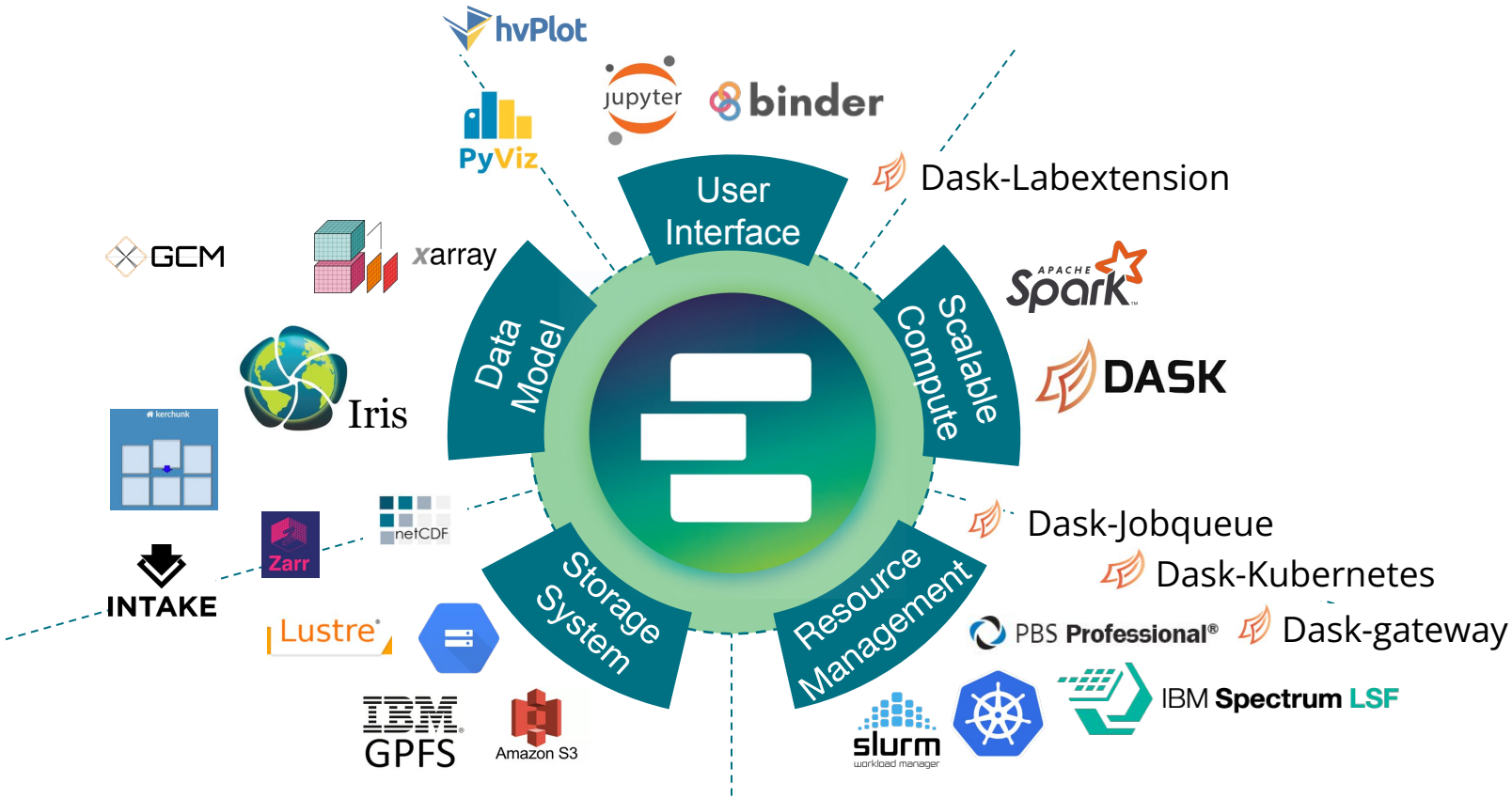
Facilitate take-up of the technology in the real world

A culture of collaboration

Open Platform

Modular, open and scalable platform

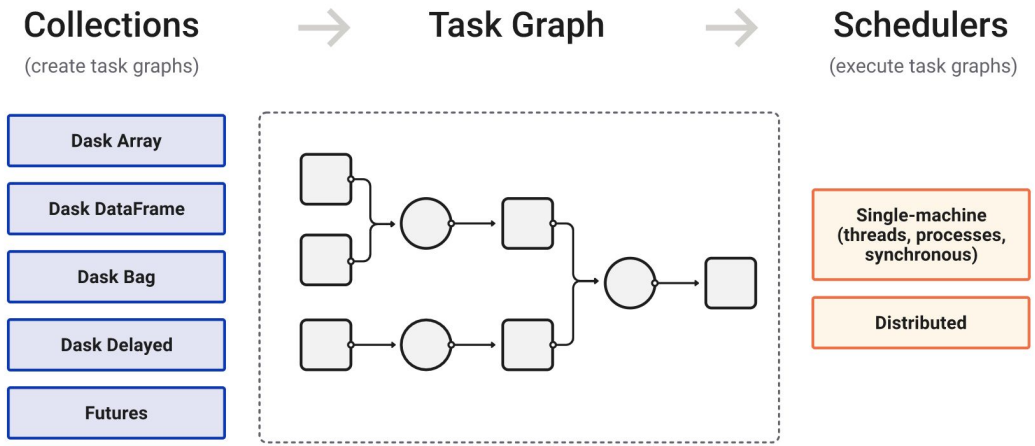
PANGEO



from laptop to cloud and HPC

Flexible library for parallel and distributed computing on "big (larger than memory) data".

- Dynamic task scheduling optimized for computation
- Extends pandas and NumPy APIs to larger-than-memory data (chunks)
- Fully customizable
- Scales from laptop to cluster (and back)
- Designed for interactive use
- Native python
- Almost like writing serial code (some idiosyncracies)

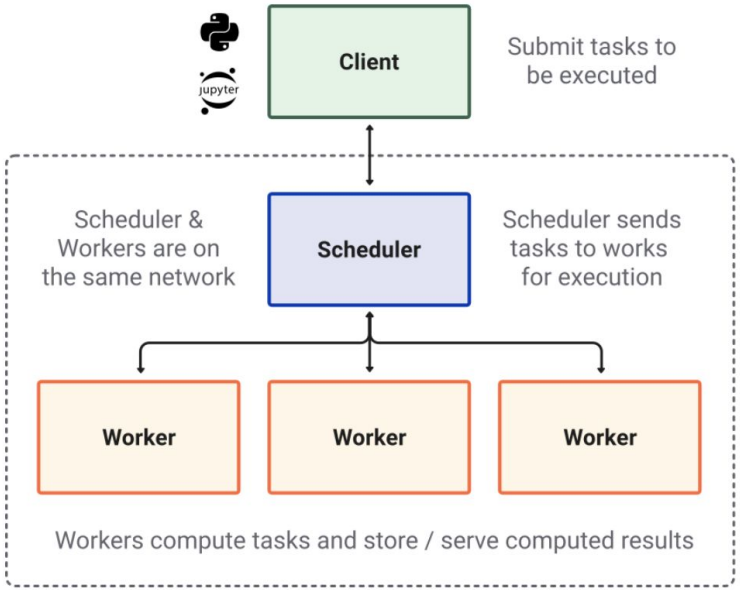


High level collections are used to generate task graphs which can be executed by schedulers on a single machine or a cluster.

Dask distributed scheduler

Supports fully asynchronous workflows and provides sophisticated data locality.

Seamlessly integrates with many/most HPC and cloud systems



Deploy anywhere and for everyone

Pangeo deployments

- Laptop, e.g. https://github.com/pangeo-data/pangeo-docker-images
- Cloud
  - Public Clouds (EOSC), e.g. Pangeo@EOSC https://pangeo-data.github.io/pangeo-eosc/
  - Commercial clouds (AWS, Microsoft, Google)
- EO Platforms
  - Formerly DIAS (CREODIAS …), e.g. https://dataspace.copernicus.eu/
- HPC
  - CNES, IFREMER, SURF, ..

Pangeo@EOSC: Why?

- Pangeo is **infrastructure-agnostic**, making it a great use case for EOSC (and other federated clouds).
- When you use Xarray, Dask and Jupyter Notebooks, you are essentially using Pangeo.
- Pangeo serves as a community and **facilitator** for deployment on various infrastructures.
- It is **open source**, eliminating vendor lock-in risks.
- At times, end-users are limited to using EU infrastructure.
- A **federation** of infrastructure **can't be** a single **centralized** infrastructure.

Pangeo eosc: How?

Custom environments

Online content

EOSC Community Hub

Cloud infrastructure

Reuse existing 2i2c Pangeo deployment configuration

EOSC Authentication for students, researchers, data scientists, etc.

https://pangeo-data.github.io/pangeo-eosc/

Pangeo Training Infrastructure as a Service

FOSS4G workshop

Overview

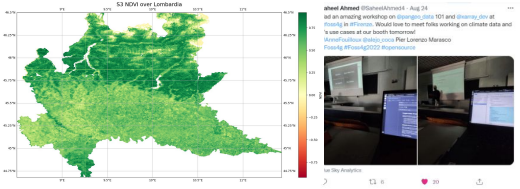
JB

Questions

- Why do chunking matter?
- How can I read datasets by chunks to optimize memory usage?

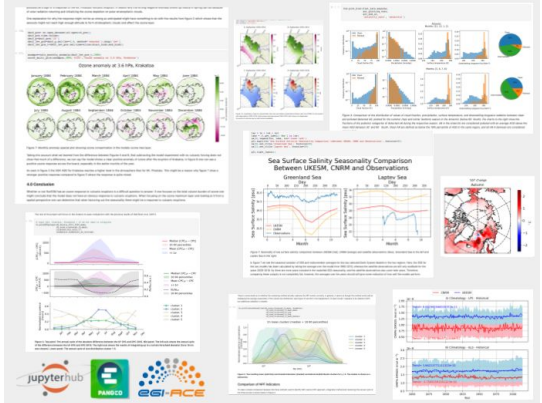
Objectives

- Learn about chunking
- Learn about zarr
- Use kerchunk to consolidate chunk metadata and prepare single ensemble datasets for parallel computing



Pangeo 101 workshop at the FOSS4G conference

eScience Course

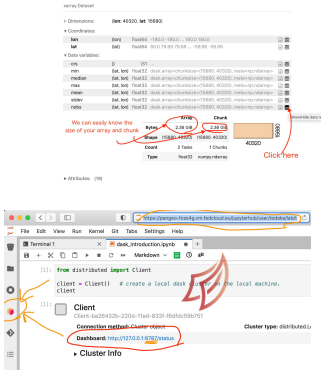


Integrate EOSC into master students' courses

CLIVAR Arctic Bootcamp

Supporting researchers in becoming competent practitioners in their scientific domain

dask



Pangeo & OpenEO at BiDS

PANGEO

openEO Platform

esa

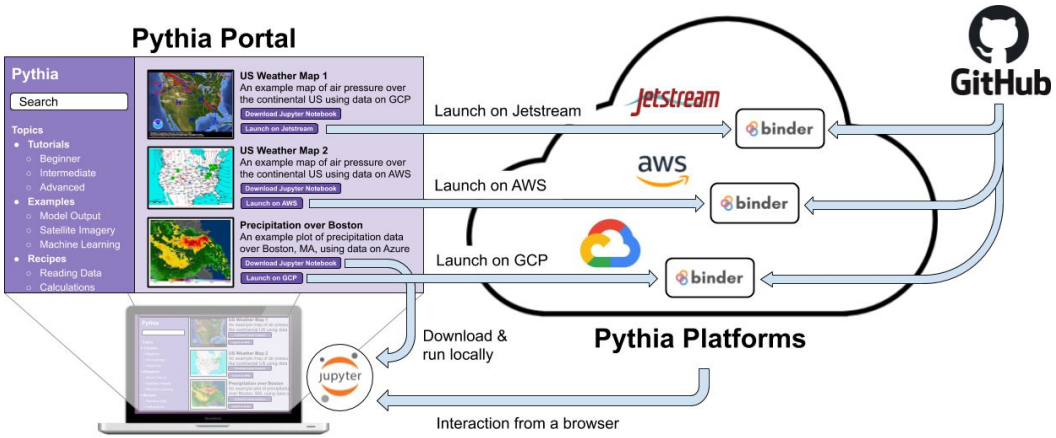
Request PTIaaS to train and teach!

Fill the form

Team-up with other initiatives to onboard new members and increase diversity

Team-up with Project Pythia

A Community Learning Resource for Geoscientists



https://projectpythia.org/

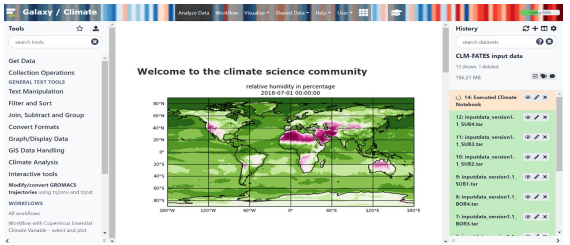
This work supported through the National Science Foundation Award #2026899.

Aspiration goal: Be the goto resource for learning the *Scientific Python Ecosystem*

- ★ Geoscience focused
- ★ From beginner to the power user
- ★ Tutorials, videos, examples, on-line courses, and sample data
- ★ Community owned!

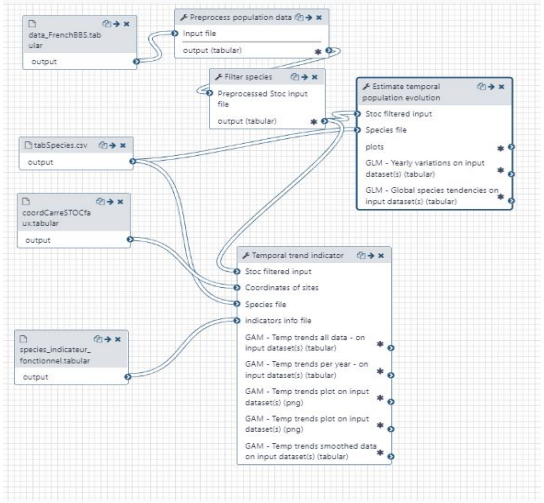
Source: Clyne et al (2022). ARM/ASR Open Science Virtual Workshop May 10, 2022

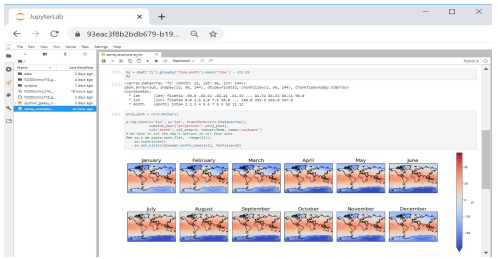
Team-up with Galaxy Europe



Galaxy is an open-source community and platform for FAIR data analysis. It offers:

- **Graphical User Interface (GUI) for users with no programming skills**
- Workflow editor to create and run fully reproducible data analysis
- Compute & Storage to everyone (free registration)
- Self Paced Learning material with the Galaxy Training Network
- Training Infrastructure as a Service (TiaaS) free and ready to use with private queues where only training's jobs run





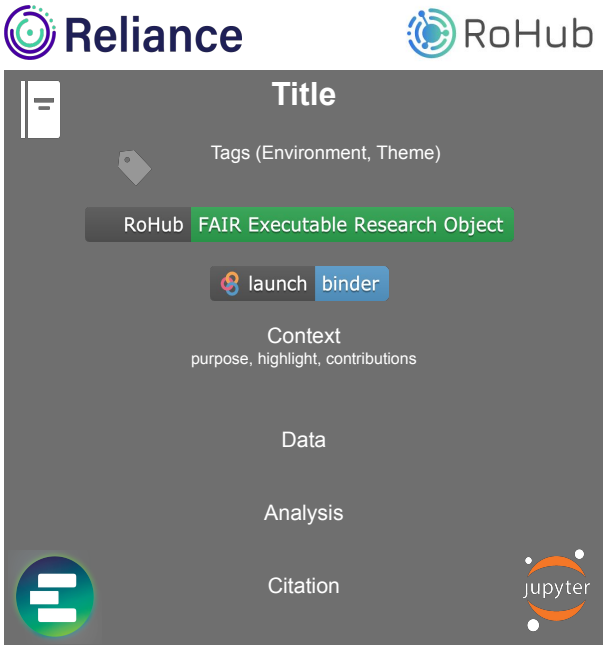
Galaxy PROJECT

TIaaS

Team-up with the Environmental Data Science Book

https://edsbook.org/welcome.html

*Living, open and community-driven online resource to **showcase and support the publication** of data, research and open-source tools.*



Initiative led by Alejandro Coca-Castro

GeoPython community

- Geopandas
- GDAL
- Pysal

Digital open books, building on Jupyter notebooks etc.

- “Geocomputation with Python” (or R) by Michael Dorman, Anita Graser, Jakub Nowosad, Robin Lovelace: https://py.geocompx.org/
- “Geographic Data Science with Python”, by Sergio J. Rey, Dani Arribas-Bel and Levi J. Wolf: https://geographicdata.science

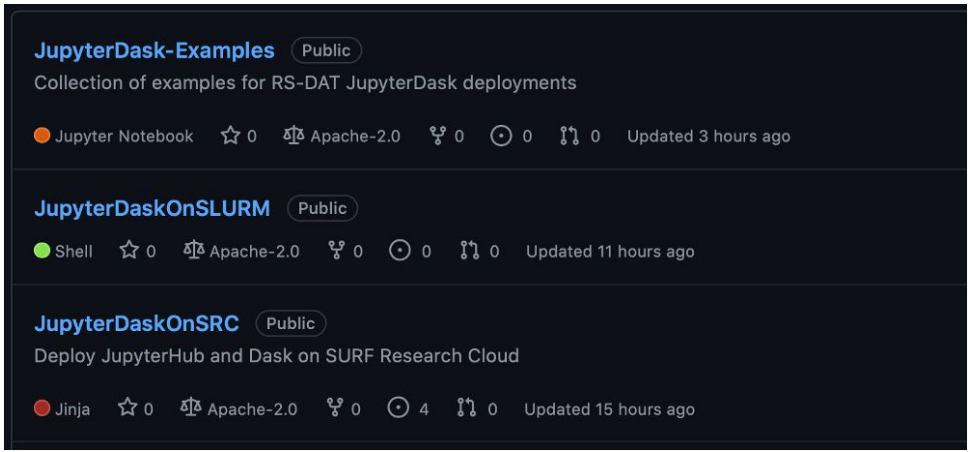
→ JupyterGIS: https://blog.jupyter.org/jupytergis-d63b7adf9d0c (new ESA Digital Innovation project)

Pangeo@SURF: scaling on HPC/HTC

Pangeo@SURF

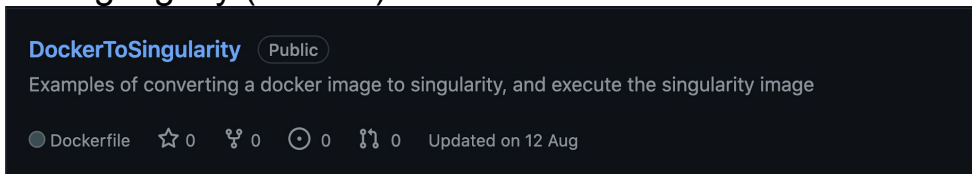
RS-DAT Components

https://github.com/RS-DAT/JupyterDaskOnSLURM



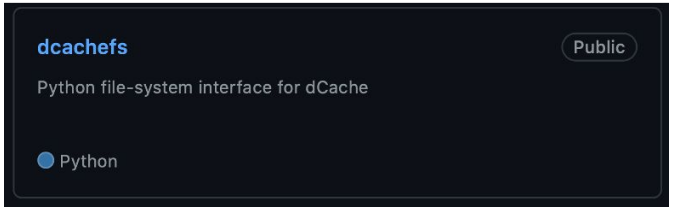
Scalable analysis with Jupyter

Using legacy (Docker) containers for HPC



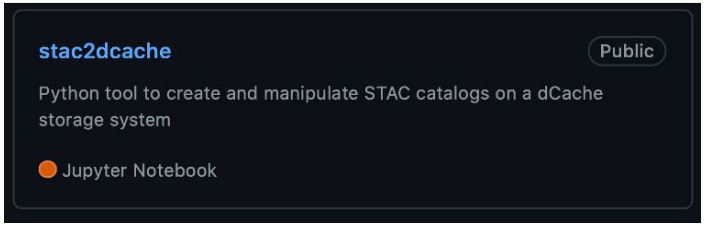
https://github.com/RS-DAT/DockerToSingularity

https://github.com/RS-DAT/dcachefs



Python interface to dCache

https://github.com/RS-DAT/stac2dcache



utility functions to manage STAC catalogs (and the underlying data) on dCache.

Jupyter Dask on Slurm

SURF H*C Infrastructure: Spider (HTC), Snellius (HPC); SLURM scheduler

- JupyterLab instance with Dask and Git extensions, scalable Dask cluster, running on SLURM managed HTC/HPC system
- Basic idea
  - Launch JupyterLab server and Dask scheduler as long-running batch job
  - SSH port-forwarding to connect to JupyterLab server
  - Lauch workers as short-lived batch jobs (fast thru queue)

Ensure ease of use by abstracting details from user

A community of developers, scientists and users

America
Time zone +11-> -2

Every Wednesday, alternating between 12pm ET and 4pm ET

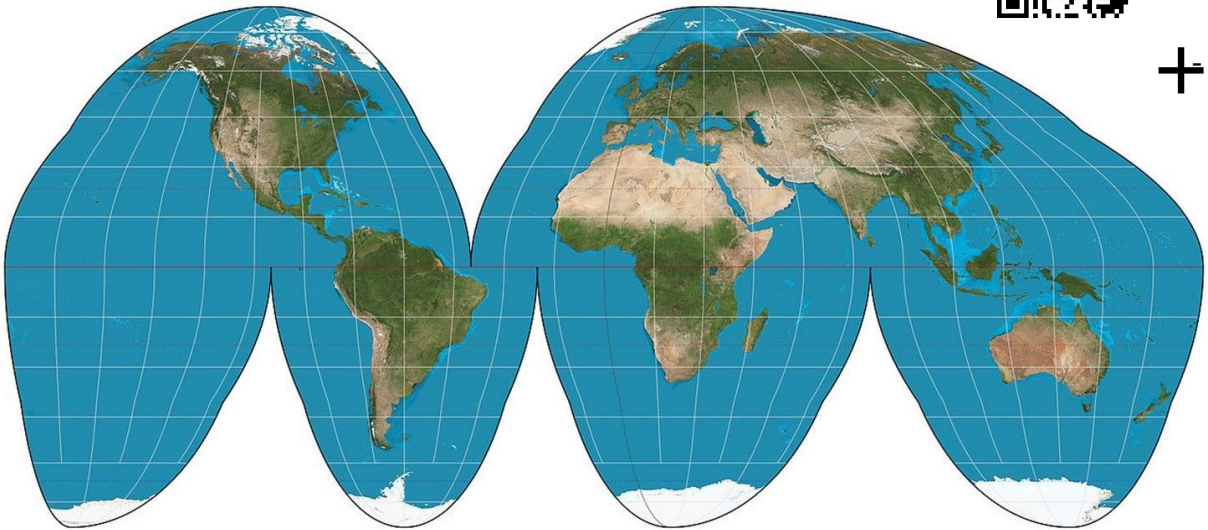
Europe, Africa, West Asia
Time zone -1-> +5

Every Tuesday at 9.30 a.m. CET/CEST here

Join us here

Australia,East Asia
Time zone -1-> +5

3rd Friday of every month at 1pm Australian Eastern Time



\+

Check out for different discussions ;

- software-production-visualization
- data-management
- software-geospatial-bridge
- cloud-partner-managed-infrastructure
- cloud-pangeo-managed-infrastructure
- democratization-science-community
- scientific-publishing
- software-massive-scale
- software-regridding

More info & links here - https://pangeo.io/meeting-notes.html

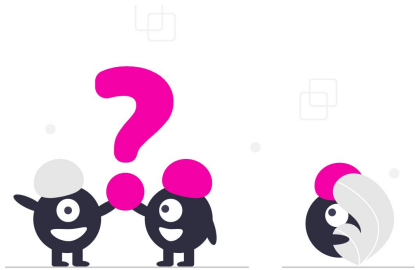
Pangeo: where can I find information?

- http://pangeo.io/index.html#what-is-pangeo
  - (https://github.com/pangeo-data/pangeo)
- Try some examples in the Pangeo Gallery
  - http://gallery.pangeo.io
- Subscribe in discourse, start discussions
  - https://discourse.pangeo.io/
- Follow pangeo-data from Github
- Discord channel
  - https://discord.gg/ex5qqEyyTz
- Interesting informations exist also in blogs
  - https://medium.com/pangeo
- Code of conduct.
  - https://github.com/pangeo-data/governance/blob/master/conduct/code_of_conduct.md

Hands on

https://pangeo-data.github.io/geo-open-hack-2024/intro.html

Thank you for your attention!



Pangeo eosc