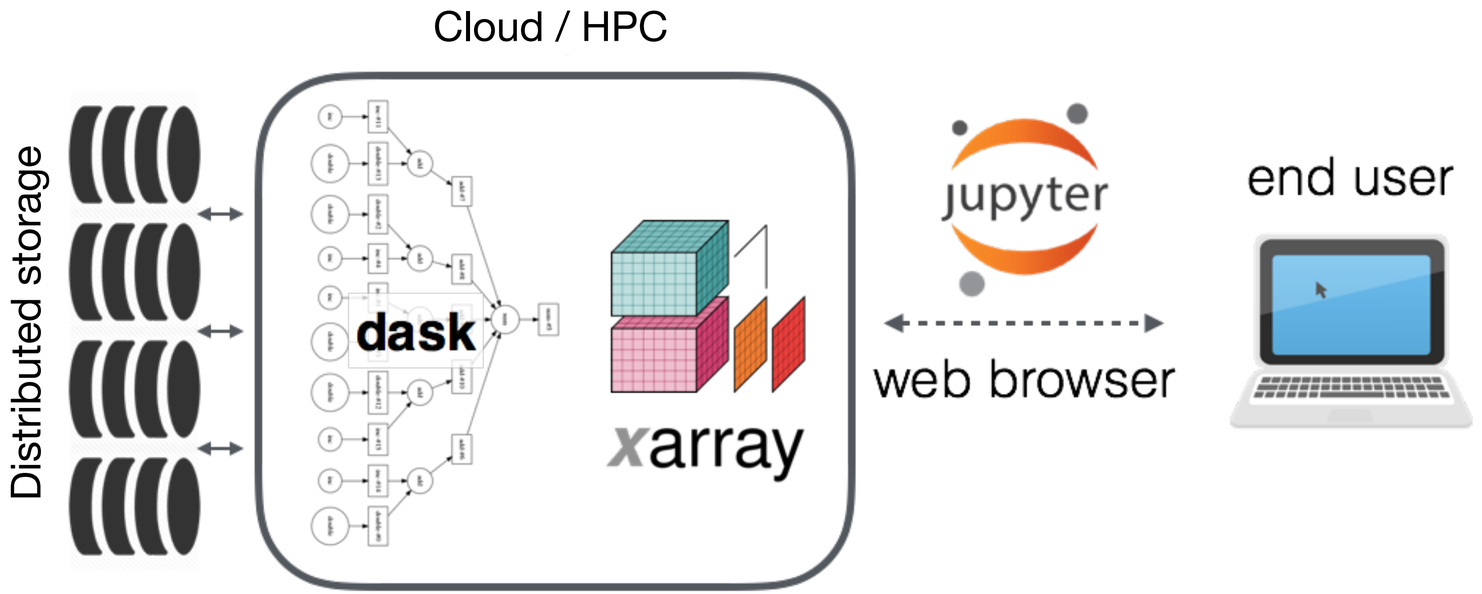


Python for Big Data in Climate Science

Aaron Spring (Max-Planck-Institute for Meteorology)

Goal

- Show capabilities of `xarray` with `dask`
- Glimpse into analysis workflow of a climate scientist

Outline

- What's a climate model? How does climate data output look like? Our toolbox
- Who to get climate data to shine (easily)? Live demo
- Science in the cloud

This is Aaron

- Background in physics
- Currently PhD candidate
  - Max-Planck-Institute for Meteorology in Hamburg
  - Topic: Variability and Predictability in the Global Carbon Cycle
  - Graduate early 2021
- Co-founder of `climpred`: ensemble forecast verification python package
- Loves ☀️, plays ⚽ 🎹, enjoys ⛺

climpred

Climate Data

- Multi-dimensional, labeled and homogeneously formatted data
- Geosciences: (dimensions `time`, `longitude`, `latitude`, `depth/height`, ...)
  - Weather forecast
  - Climate simulations
  - Satellite product
- Finance: Stock prices
- Format:
  - `netcdf`: Network Common Data Form builds upon `HDF5`
  - `zarr`: optimized for chunked data in the cloud

Climate Modelling

- Earth System Model (ESM)
- Simulating one year with MPI-ESM-LR:
  - 14 GB output
  - 5 Node hours (1 Node = 48 CPU)

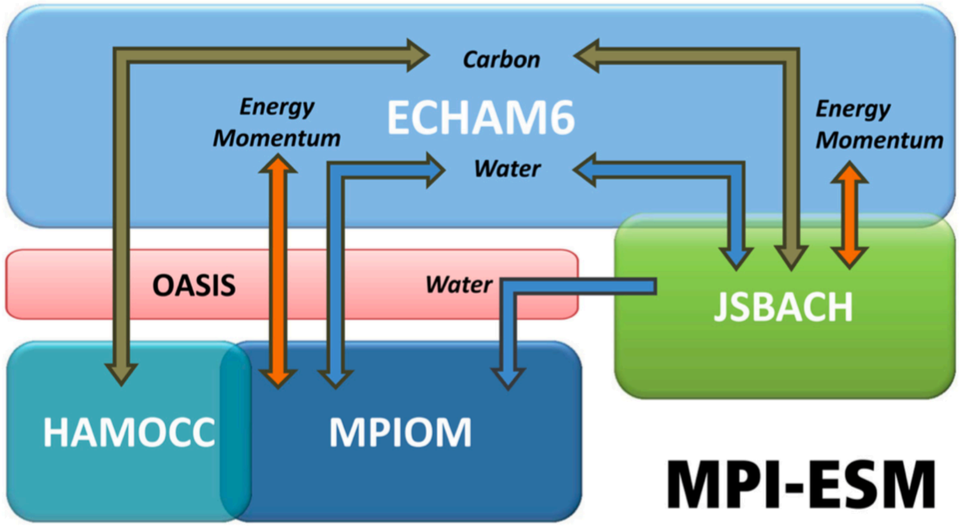


Photo by Franck V. on Unsplash

Workflow without python

- Domain specific languages and tools:
  - `NCL`: NCAR Command Language
  - `CDO`: Climate Data Operators
- Proprietary software:
  - `MATLAB`
- Disadvantages:
  - Long scripts, copy&paste, little software engineering, different software for computation and visualisation
  - Little reproducibility: intermediate files, no code sharing culture

Telling a story based on data

- Analyse available data and visualise results
- Danger of “Computing Too Much and Thinking Too Little” [Emanuel, 2020]
- Ideally:
  - technically: intuitive, easy, no need to understand every detail under the hood
  - conceptually: demanding and genius

➡ Python ecosystem and 

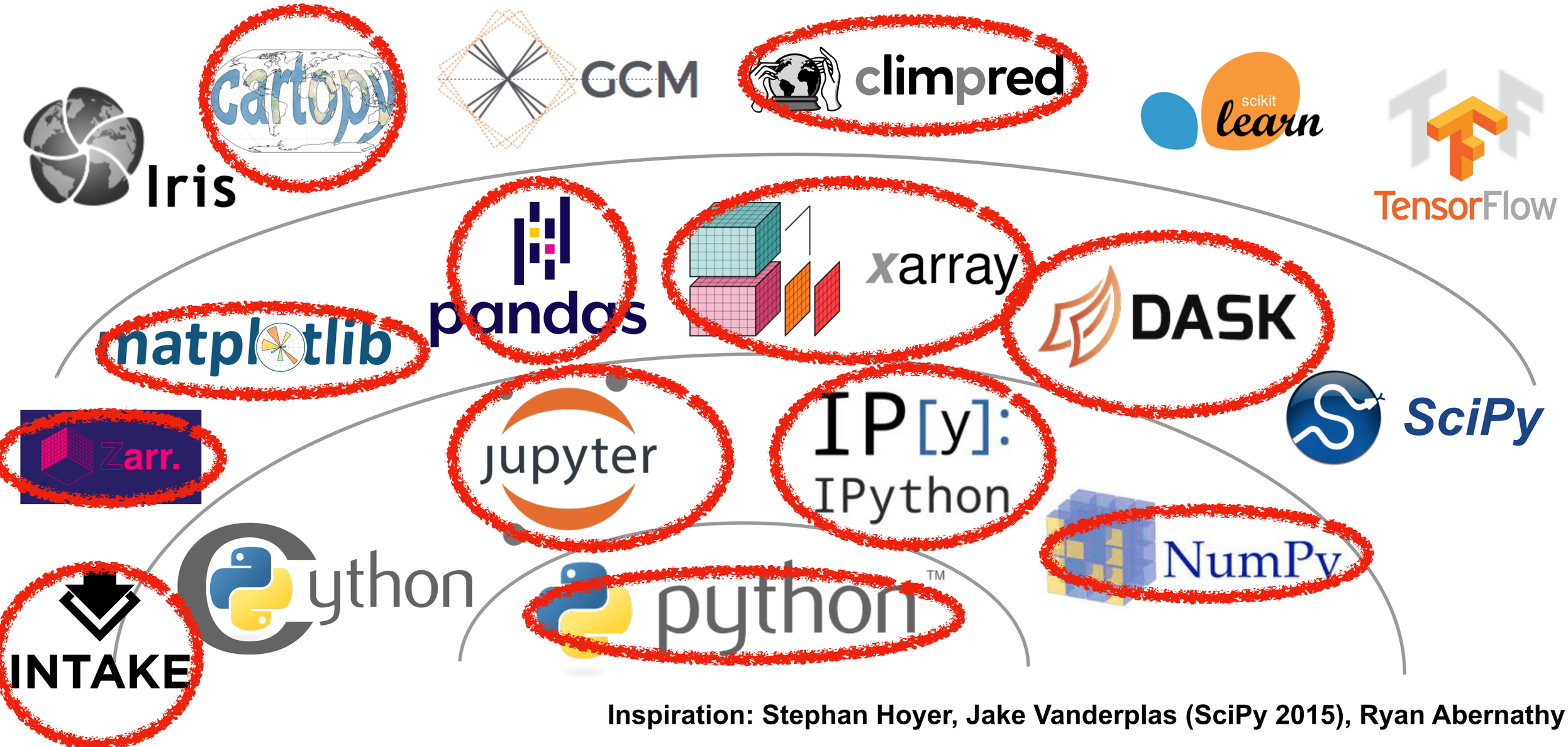
What is Pangeo?

"A community platform for Big Data geoscience"

- Open Community
- Open Source Software
- Open Source Infrastructure

PANGEO

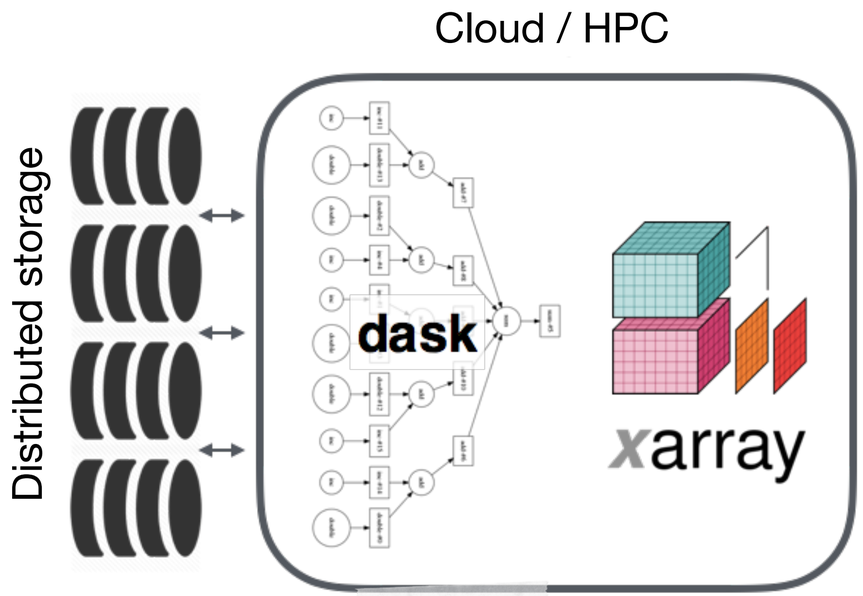
Pangeo Software Ecosystem



Inspiration: Stephan Hoyer, Jake Vanderplas (SciPy 2015), Ryan Abernathy

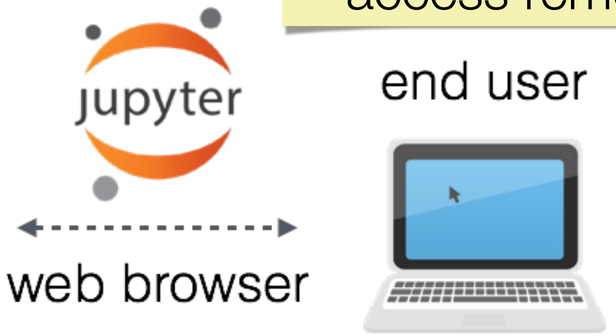
HPC Architecture

"Analysis Ready Data" stored on distributed storage.



Parallel computing system allows users deploy clusters of compute nodes for data processing.

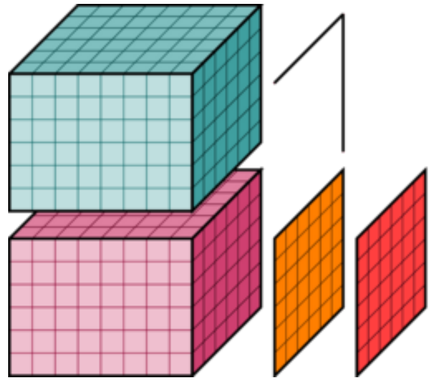
Dask tells the nodes what to do.



Jupyter for interactive access remote systems

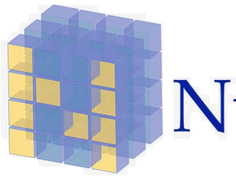
Xarray provides data structures and intuitive interface for interacting with datasets

xarray



xarray

- Analysis of multi-dimensional data
- Self-describing data
- Efficient: based on `numpy` and `dask`
- Simple: API inspired by `numpy` and `pandas`
- Stephan Hoyer and Joe Hamman (2017) “Xarray: N-D Labeled Arrays and Datasets in Python”

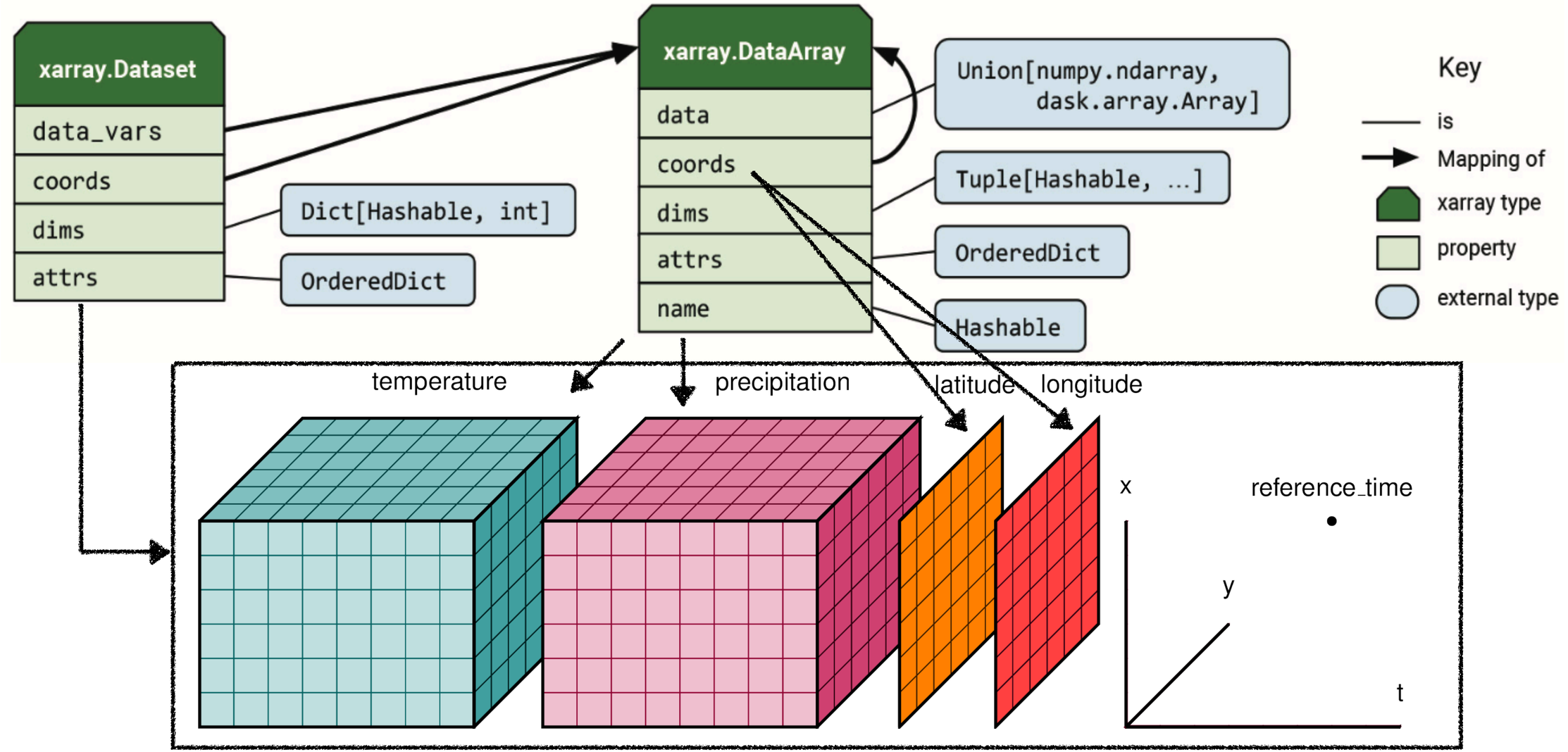


NumPy

DASK

PANGEO

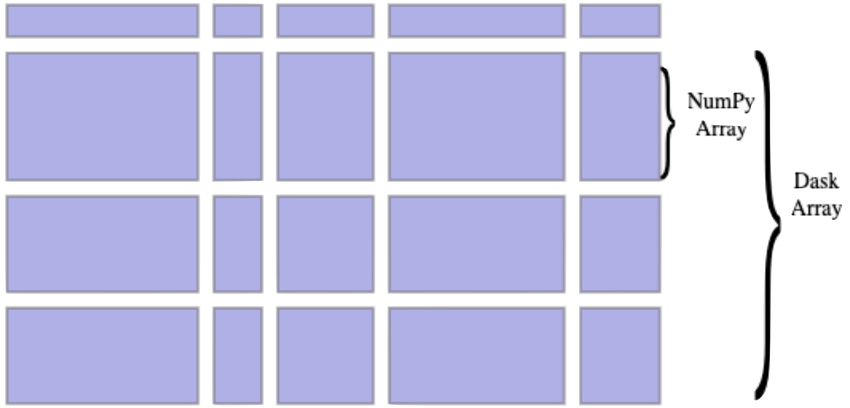
xarray data types



dask

DASK

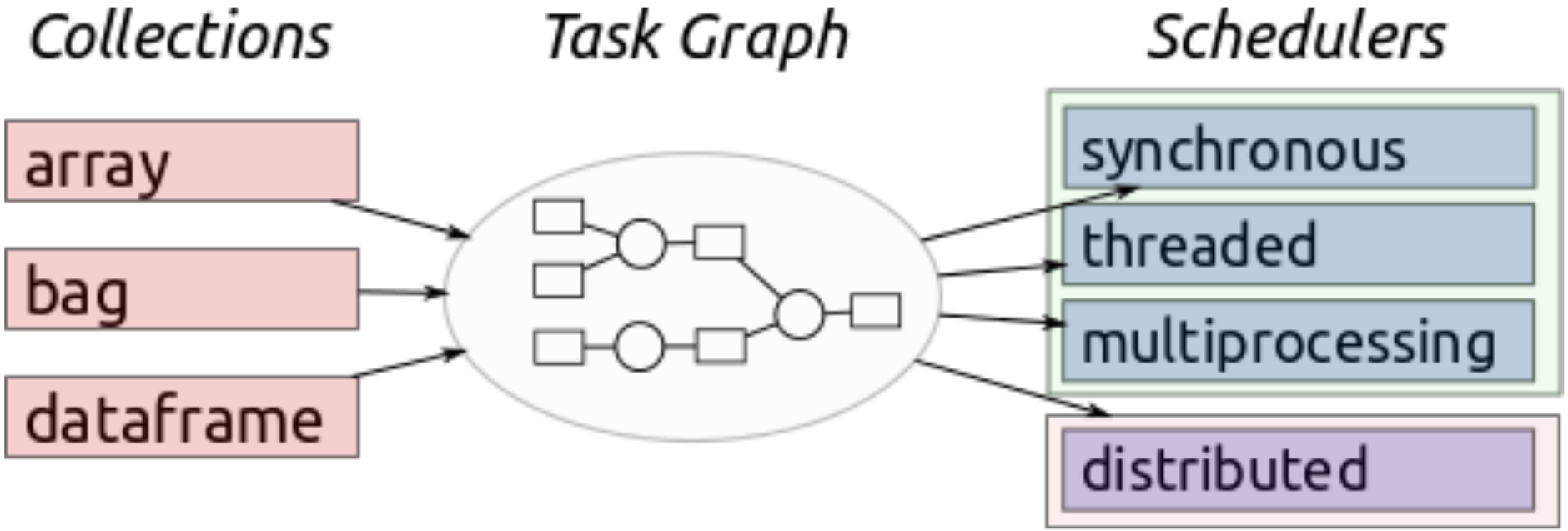
- Dynamic task scheduling
- Builds upon `multiprocessing`, `threading` and `concurrent`
- out-of-memory computation via chunking
- Scales from laptop to supercomputer
- Intuitive (known) API from `pandas` and `numpy`



- Matthew Rocklin (2015: “Dask: Parallel Computation with Blocked Algorithms and Task Scheduling”

dask workflow

- `dask.array –> numpy.array`
- `dask.bag –> iterable`
- `dask.dataframe –> pandas.DataFrame`



intake

INTAKE

- Taking the pain out of data access
- Finding, investigating, load and disseminating data with catalogs
- version control data sources
- distinction between data curator and analyst roles
- `intake-xarray` plugin: Catalogs pointing to `netcdf` files
- `intake-esm` plugin : Search and load ESM output
  - Query in `pandas.DataFrames`
  - Load data with `dask` into `xarray`
  - Developed by Anderson Banihirwe (**andersy005**)

Live demo

OR

launch binder

DKRZ mistral: https://www.dkrz.de/up/systems/mistral

My best practices

- Aim to `concat` all your data into one `xr.Dataset` and save to disk (preferably in `zarr`)
- Understand chunking and use `dask` incl. `dashboard`
- Leverage `intake`

PANGEO

Extensions to xarray

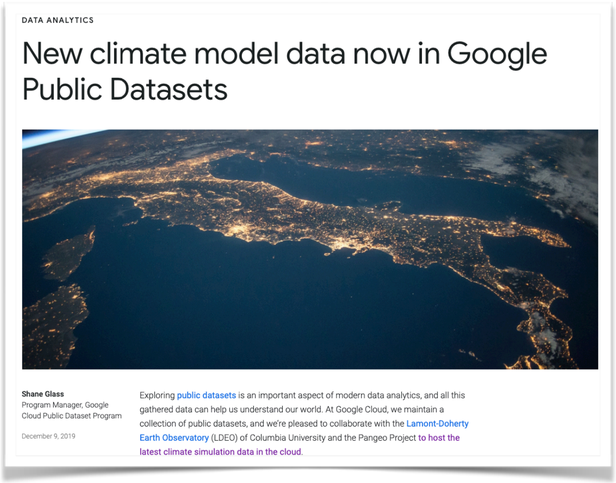
- `scipy` : (nearly) all functions callable with `xr.apply_ufunc`
- `dask_jobqueue` : parallelise `dask` across nodes
- `xskillscore` : verification metrics
- `cartopy` : projections of maps
- `geoviews` : dynamic visualisation of geo data
- `regionmask` : spatial aggregation based on shapefiles
- `xesmf` : regridding
- `xgcm` : grid aware operations
- `cmip6_preprocessing` : data cleaning for CMIP6 output
- `climpred` : verification of multi-dim ensemble forecasts
- `intake-xarray` : intake for netcdf files

... http://xarray.pydata.org/en/stable/related-projects.html

PANGEO

Challenges in Big Climate Data

- Dataset size & I/O performance —> solution proposal: pangeo cloud 💵
- Filesystems (`lustre`) not optimised to load many small files to load fast
- Little knowledge of climate scientists in software engineering, testing, continuous integration, parallelism



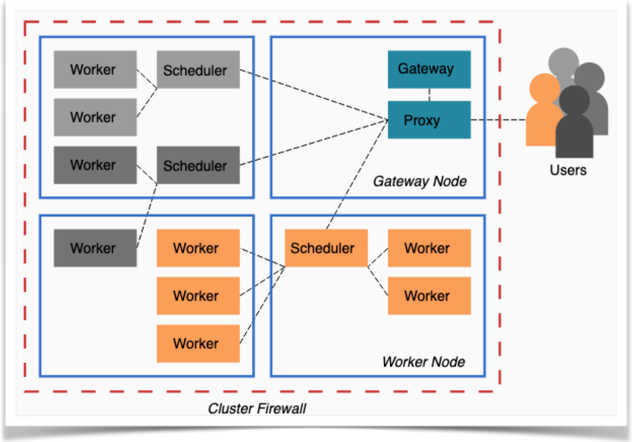
DATA ANALYTICS

New climate model data now in Google Public Datasets

Shane Glass
Program Manager, Google Cloud Public Dataset Program

December 9, 2019

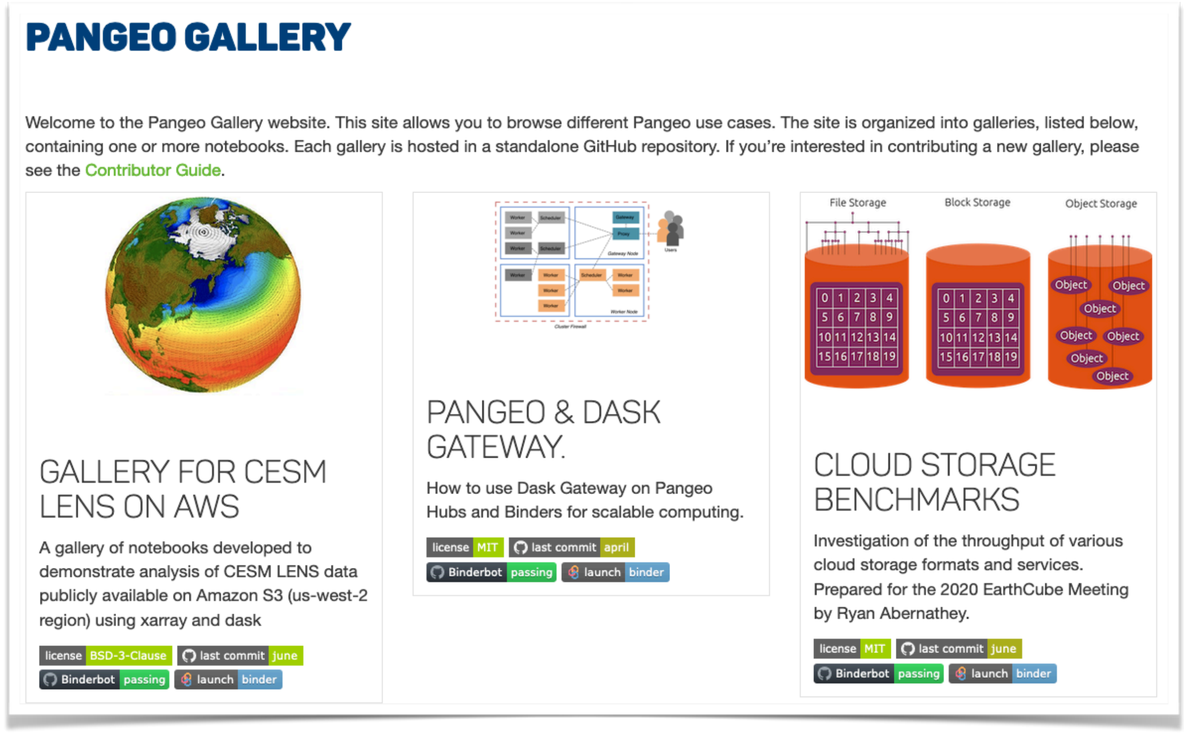
Exploring public datasets is an important aspect of modern data analytics, and all this gathered data can help us understand our world. At Google Cloud, we maintain a collection of public datasets, and we're pleased to collaborate with the Lamont-Doherty Earth Observatory (LDEO) of Columbia University and the Pangeo Project to host the latest climate simulation data in the cloud.



PANGEO

Reproducible science

- Science in a GitHub repo:
  - http://gallery.pangeo.io/
  - Data in the cloud
  - nbgitpuller to execute and render notebook
  - reproducible with binder



➡Play with climate data yourself: launch binder

PANGEO

References

- Papers:
  - Rocklin, M. (2015). Dask: Parallel Computation with Blocked algorithms and Task Scheduling. 126–132. doi: 10/gfz6s5
  - Hoyer, S., & Hamman, J. (2017). xarray: N-D labeled Arrays and Datasets in Python. Journal of Open Research Software, 5(1). doi: 10/gdqdmw
  - Emanuel, K. (2020). The Relevance of Theory for Contemporary Research in Atmospheres, Oceans, and Climate. *AGU Advances*, *1*(2), e2019AV000129. doi: 10/gg3dzt
- Pictures:
  - xarray website, dask website, MPIM, DKRZ, pangeo
- Tutorials:
  - xarray: https://xarray-contrib.github.io/xarray-tutorial/scipy-tutorial/00_overview.html
  - dask: https://tutorial.dask.org/03_array.html
  - pangeo: http://gallery.pangeo.io/