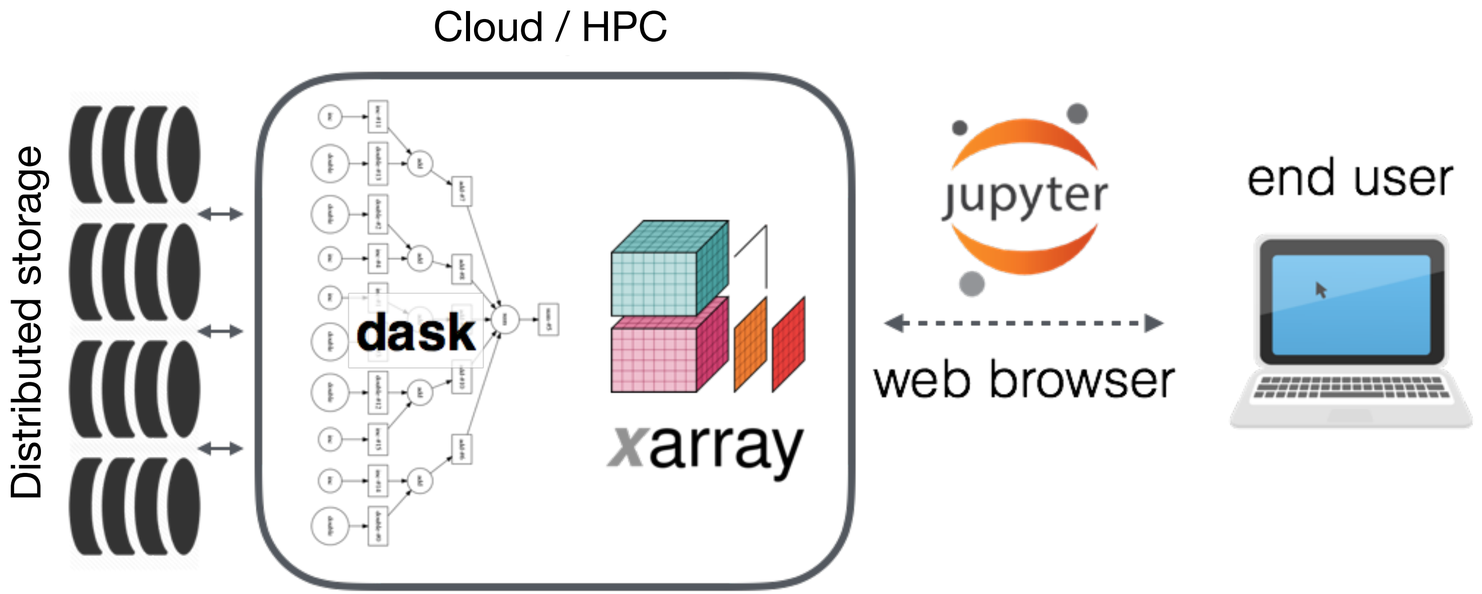



# Python for Big Data in Climate Science

Aaron Spring (MPI-M)

# Goals

- Demonstrate the data challenges of a climate scientists
- Show the analysis workflow of a climate scientist using python

# Outline

- What's a climate model? How does climate data output look like?
- Who to get climate data to shine (easily)?
- Live demo
- Reproducible science
- 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  climpred
- Loves ☀️, supports Eintracht Frankfurt, plays ⚽ 🎹, enjoys ⛺

PANGEO

# Climate Modelling

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

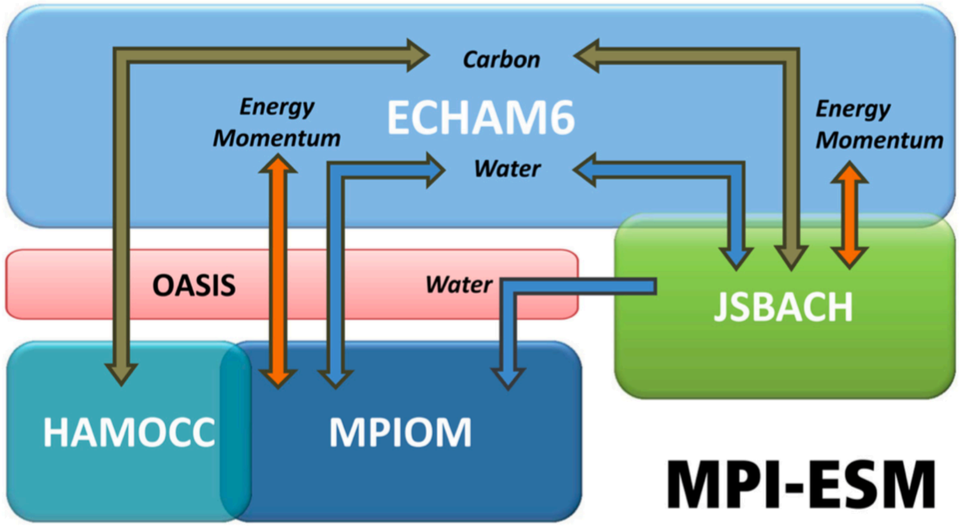



Photo by Franck V. on Unsplash

# Climate Model Output

- Multi-dimensional, labeled 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

# Workflow without python

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

# Telling a story based on data

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

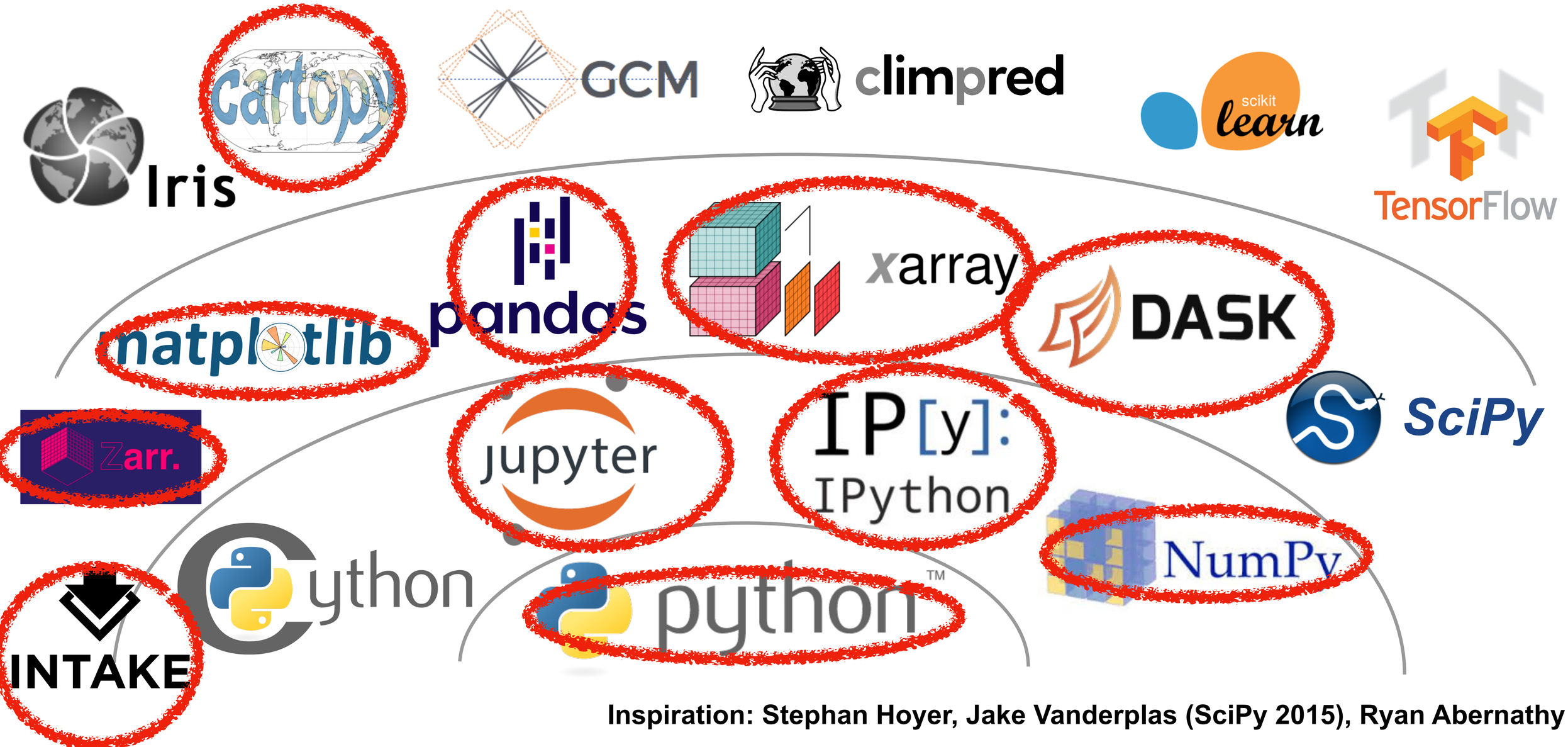
➡ Python ecosystem and 

# What is Pangeo?

*"A community platform for Big Data geoscience"*

- Open Community
- Open Source Software
- Open Source Infrastructure

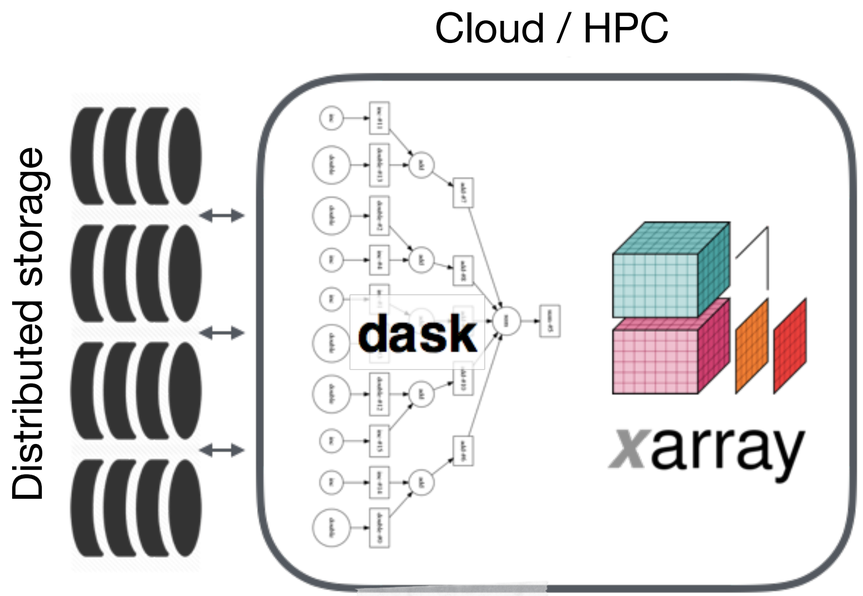
# Pangeo Software Ecosystem




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

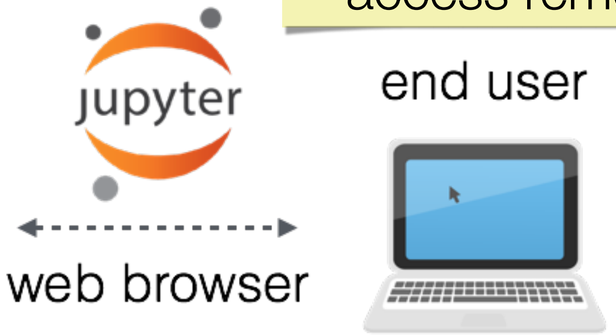
# HPC Architecture

"Analysis Ready Data" stored on distributed storage.




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

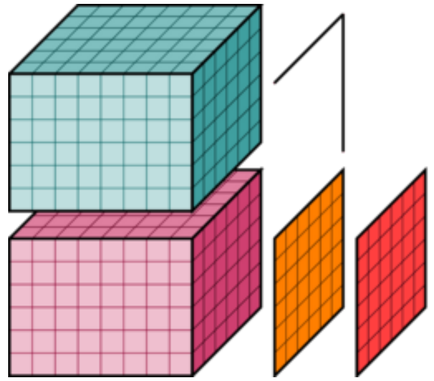
Dask tells the nodes what to do.




Jupyter for interactive access remote systems

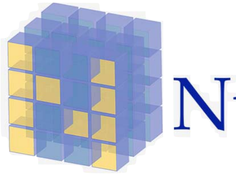
Xarray provides data structures and intuitive interface for interacting with datasets

# xarray



xarray

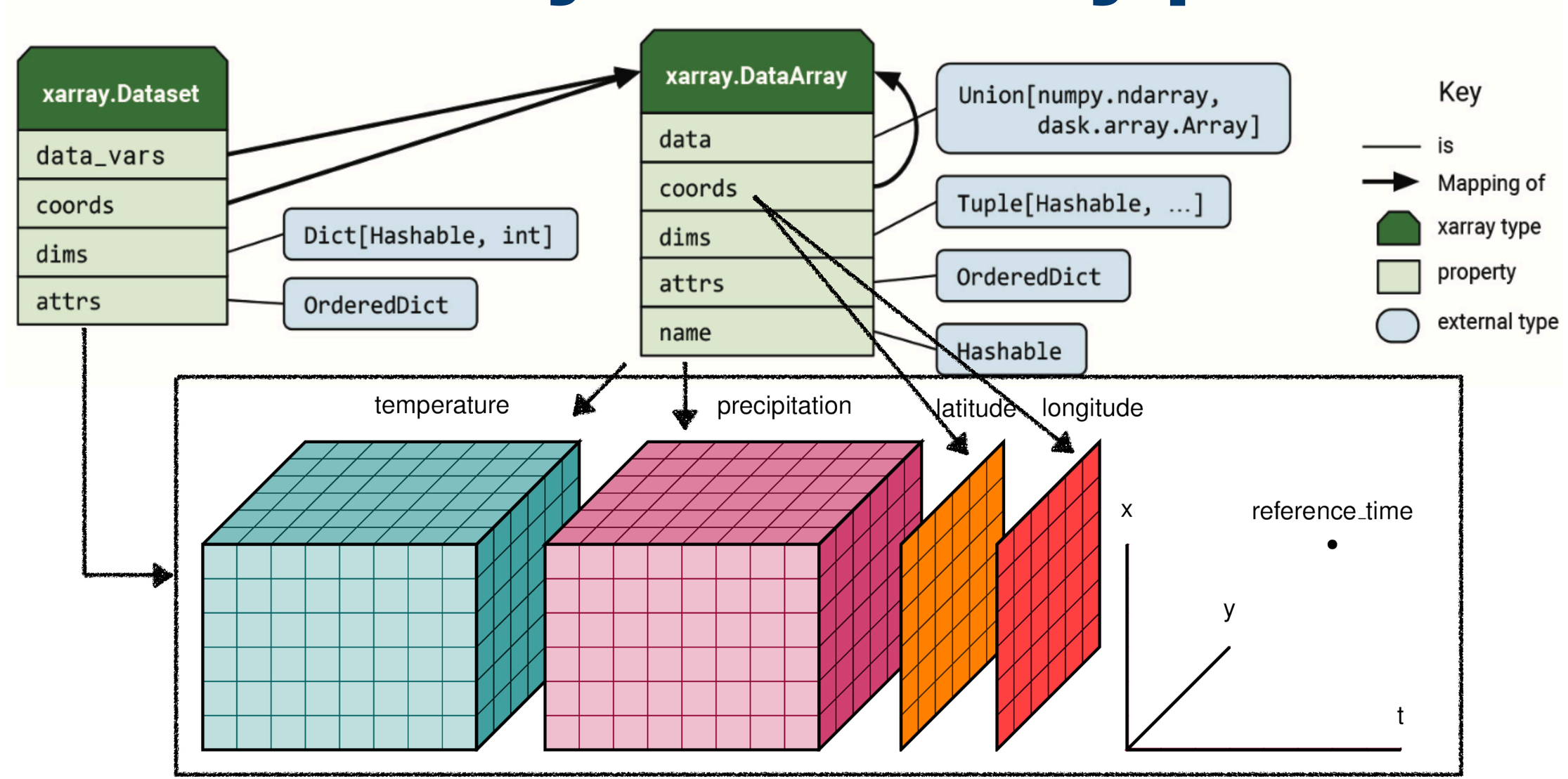
- Analysis of multi-dimensional data
- Self-describing data
- Simple: API inspired by `pandas`
- Efficient: based on `numpy` and `dask`
- Stephan Hoyer and Joe Hamman (2017) “Xarray: N-D Labeled Arrays and Datasets in Python”



NumPy

DASK

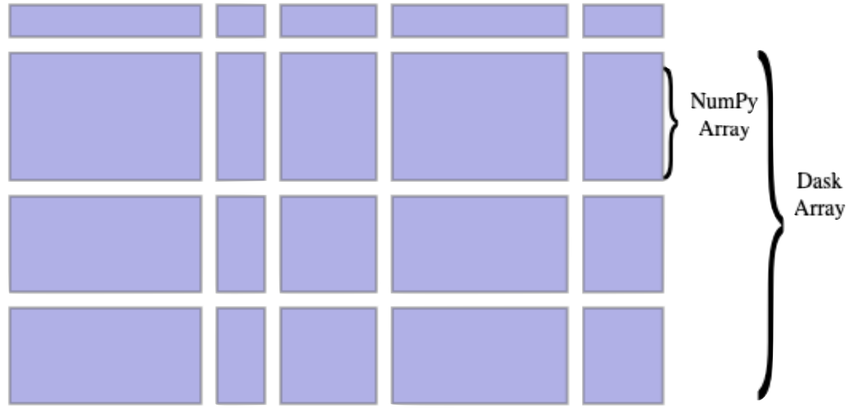
# xarray data types



# dask

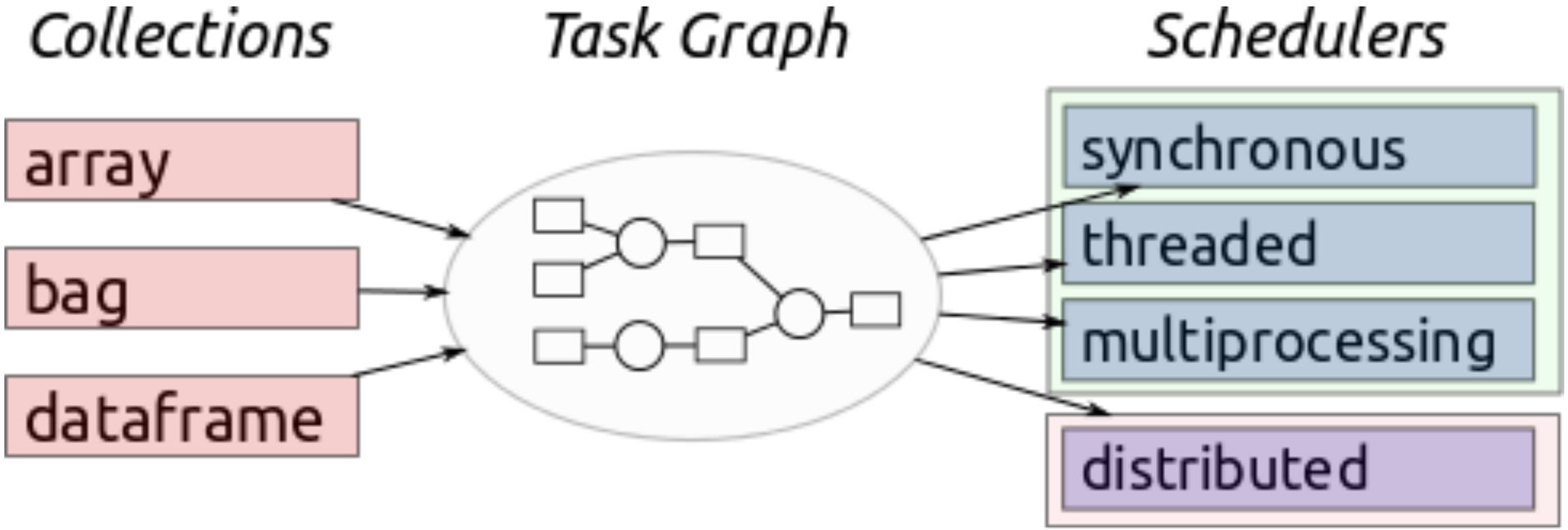
DASK

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

- Finding, investigating, load and disseminating data with catalogs
- Taking the pain out of data access
- `intake-esm plugin`: Search and load ESM output
  - Query in `pandas.DataFrames`
  - Load data with `dask` into `xarray`
  - Developed by Anderson Banihirwe (**andersy005**)

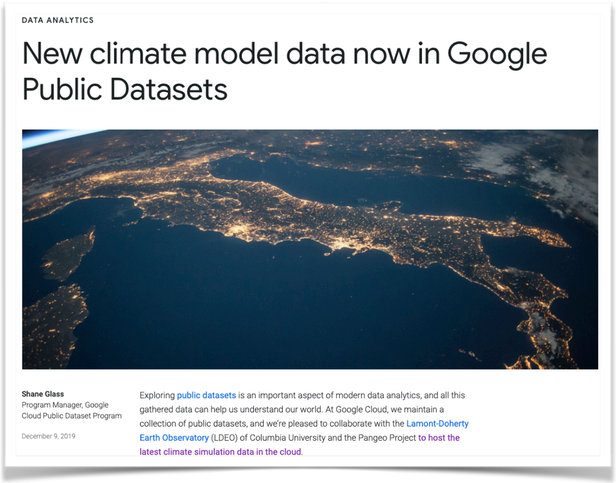
# Try climate data yourself!

OR

launch binder

# Challenges in Science Big Data

- Dataset size
- I/O performance —> solution Pangeo cloud 💵
- Filesystems (`lustre`) not optimised to load many small files fast
- Little / no knowledge of scientists in software engineering, testing, continuous integration, parallelism



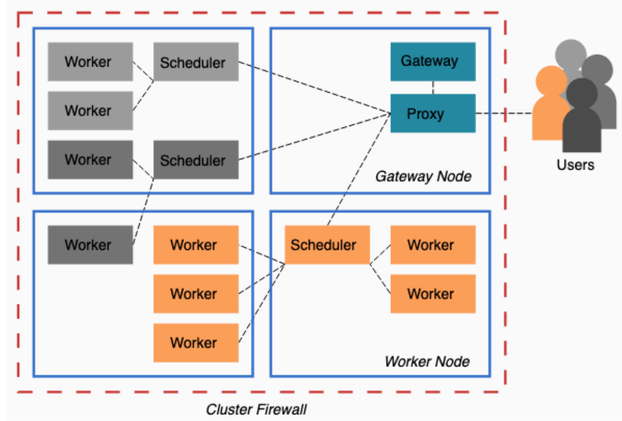
DATA ANALYTICS

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.



# Reproducible science

- GitHub repo, binderbot, nbgitpuller to execute notebook, rendered on GitHub http://gallery.pangeo.io/

# 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`
- Read the `xarray` and `dask` docs.
- Write re-usable code and share
- Pair programming (PRs on GitHub: `climpred`, `xskillscore`, `esmtools`, ...)
- Browse GitHub for smart `xarray` use-cases (start with `pangeo`)

# Extensions to xarray

- `scipy` : (nearly) all functions callable with `xr.apply_ufunc`
- `xskillscore` : verification metrics
- `cartopy` : projections of maps
- `geoviews` : dynamic visualisation of geo data
- `regionmask` : spatial aggregation based on shapefiles
- `xesmf` : regridding
- `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

# 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
- Pictures:
  - xarray website, dask website, MPIM, DKRZ, pangeo
- Tutorials:
  - xarray
  - dask
- Docs:
  - scipy, xskillscore, regionmask, intake-xarray, intake-esm, dask, xarray, intake, cartopy,