INTAKE

intake - taking the pain out of data access

Mickaël Lalande (Institut des Géosciences de l’Environnement)

Original presentation from Aaron Spring (Max-Planck-Institute for Meteorology): https://www.youtube.com/watch?v=urL17kRUinE&feature=youtu.be

What is Pangeo?

“A community platform for Big Data geoscience”

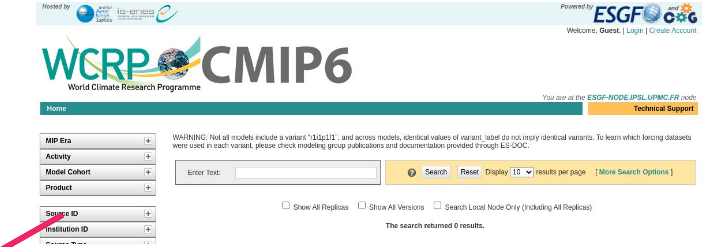
- Open Community
- Open Source Software
- Open Source Infrastructure

https://github.com/mickaellalande/MC-Toolkit/tree/master/Managing-big-data-in-geasciences-with-PANGEO (Aurélie)

What impacts the velocity of science?

Data, Software and Computation

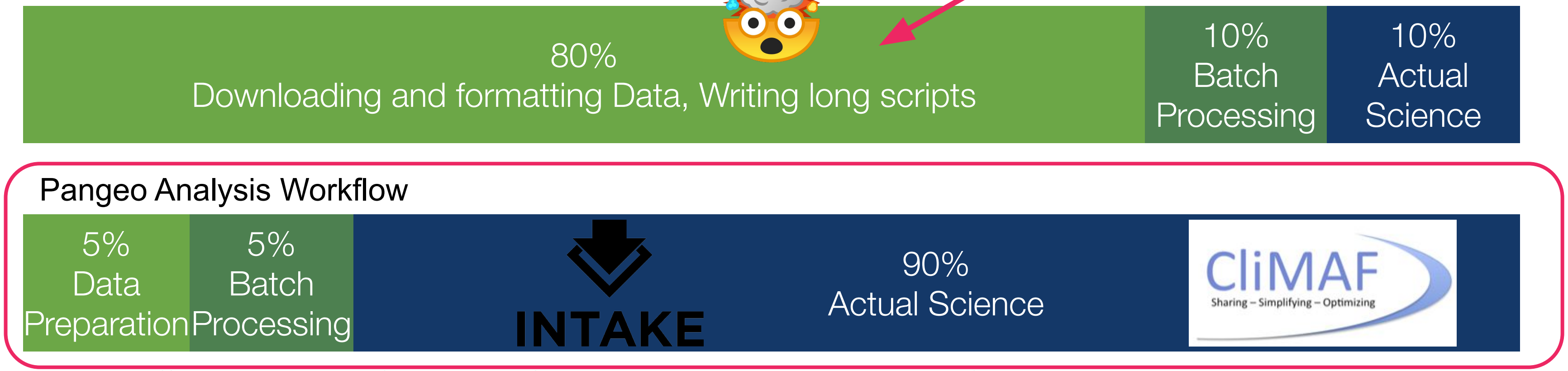
- Data: time to find, access, clean & format for analysis
- Software: easily available and combinable
- Computation: access and resources



WARNING: Not all models include a variant "r1i1p1f1", and across models, identical values of variant_label do not imply identical variants. To learn which forcing datasets were used in each variant, please check modeling group publications and documentation provided through ES-DOC.

The search returned 0 results.

Traditional Analysis Workflow



Adopted from https://speakerdeck.com/cgentemann/empowering-transformational-science?slide=4

intake-esm

INTAKE

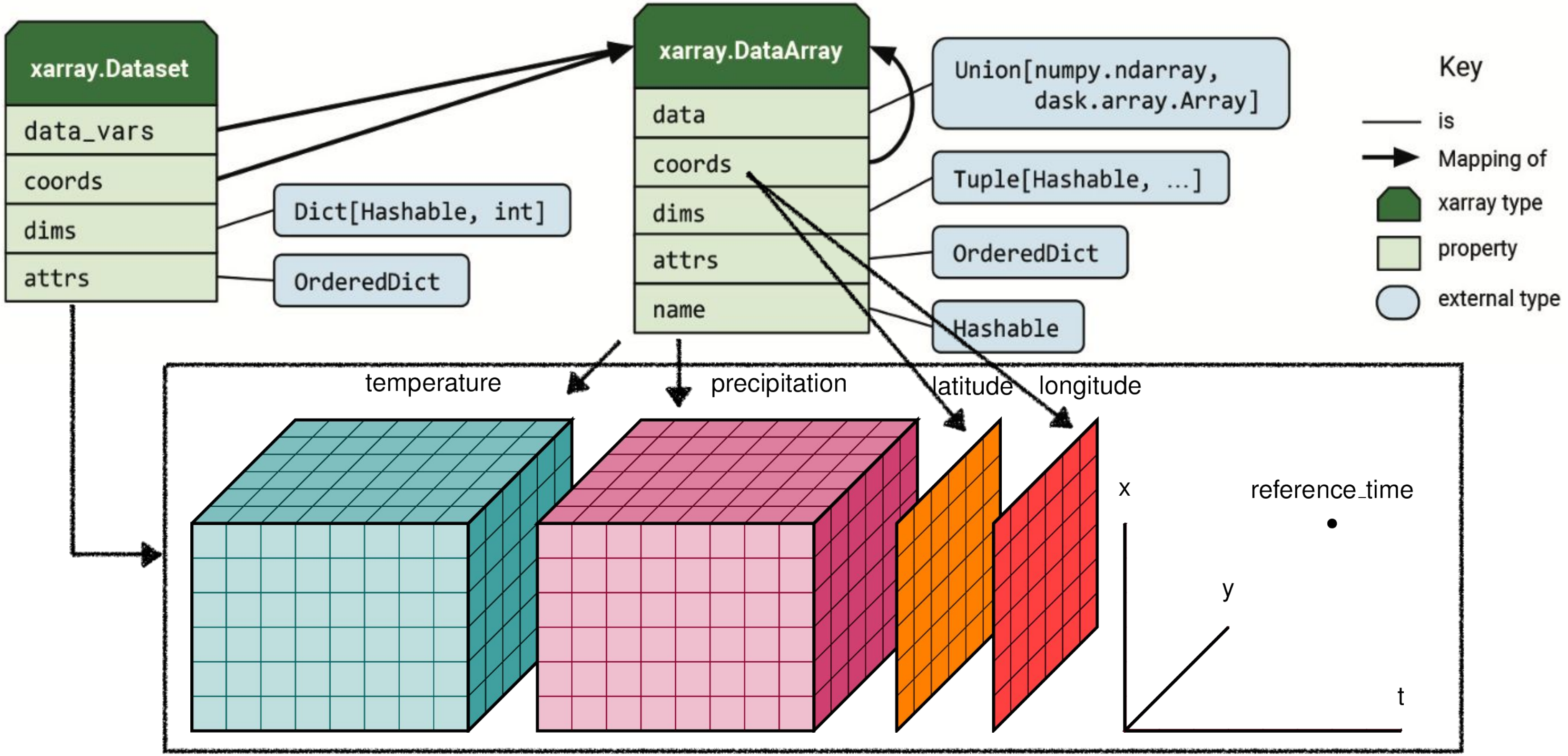
- Developed by Anderson Banihirwe at NCAR (**@andersy005**)
- Search and load ESM output
- Catalog builds easily from CMORized output
- Query in pandas.DataFrames
- Share/archive data sources used for your particular analysis
- Load data with dask into xarray

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”

https://github.com/mickaellalande/MC-Toolkit/tree/master/conda_environment_xarray_xesmf_proplot/xarray

xarray **data types**

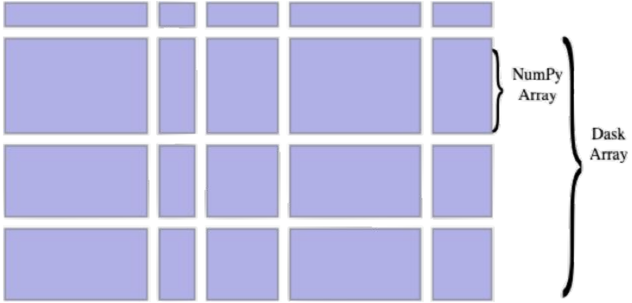


Extensions to xarray

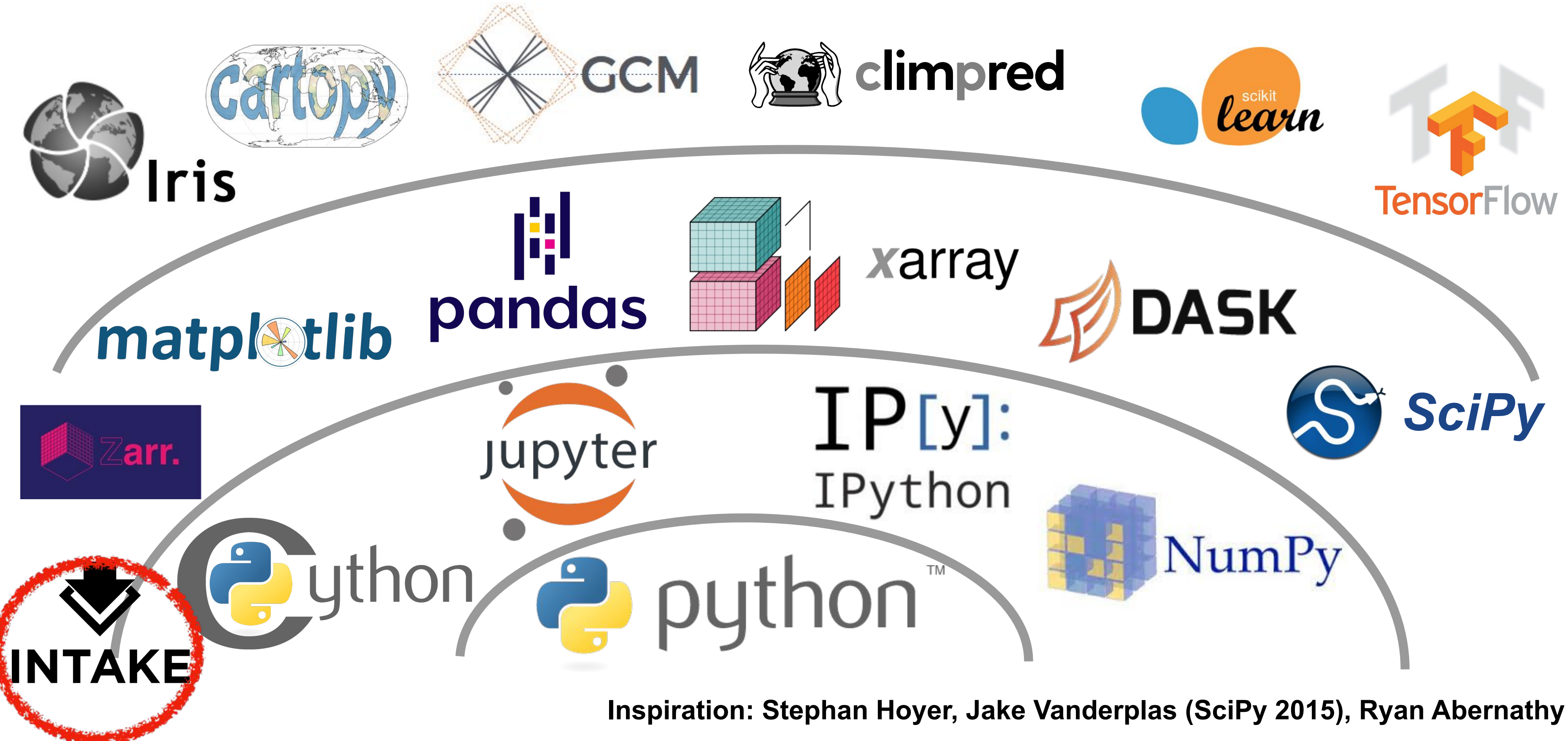
- scipy : (nearly) all functions callable with xr.apply_ufunc
- dask_jobqueue : parallelise dask across nodes
- xskillscore : verification metrics
- cartopy : projections of maps
- geoviews : dynamic visualisation of geo data
- regionmask : spatial aggregation based on shapefiles
- xesmf : regridding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

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”

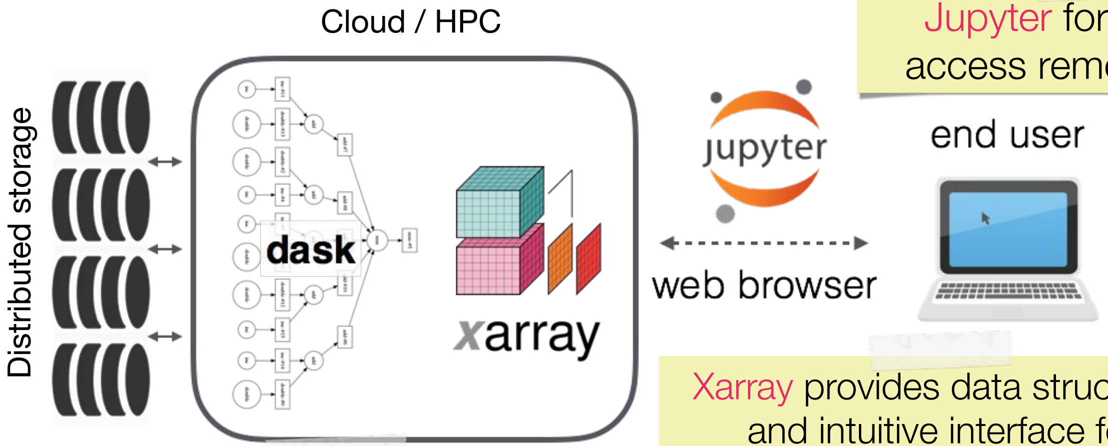


Pangeo Software Ecosystem



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

HPC Architecture



Jupyter for interactive access remote systems

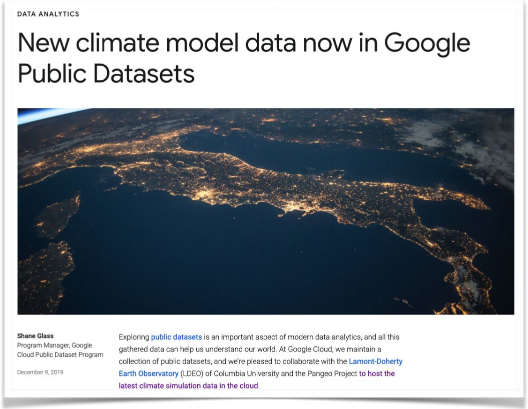
“Analysis Ready Data” stored on distributed storage.

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

Dask tells the nodes what to do.

Xarray provides data structures and intuitive interface for interacting with datasets

Pangeo in the cloud



DATA ANALYTICS

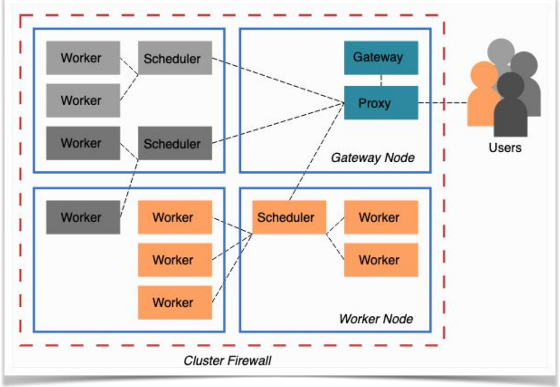
New climate model data now in Google Public Datasets

Shane Glass
Program Manager, Google Cloud Public Dataset Program

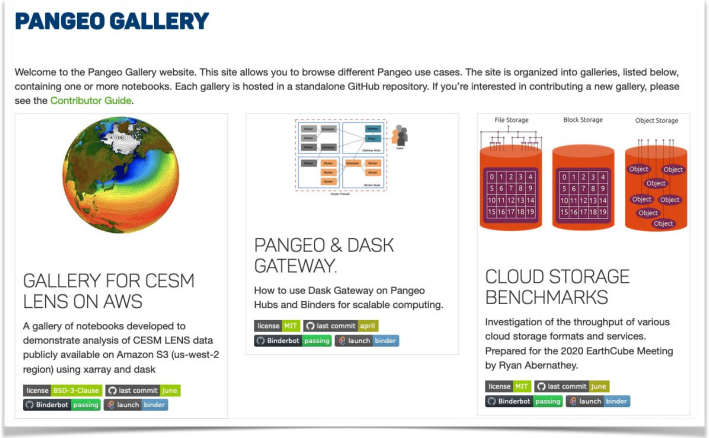
December 9, 2019

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.

https://cloud.google.com/blog/products/data-analytics/new-climate-model-data-now-google-public-datasets



- Server-side computing



PANGEO GALLERY

Welcome to the Pangeo Gallery website. This site allows you to browse different Pangeo use cases. The site is organized into galleries, listed below, containing one or more notebooks. Each gallery is hosted in a standalone GitHub repository. If you're interested in contributing a new gallery, please see the Contributor Guide.

GALLERY FOR CESM LENS ON AWS

A gallery of notebooks developed to demonstrate analysis of CESM LENS data publicly available on Amazon S3 (us-west-2 region) using xarray and dask

license BSD-3-Clause | last commit june | Binderbot passing | launch binder

PANGEO & DASK GATEWAY.

How to use Dask Gateway on Pangeo Hubs and Binders for scalable computing.

license MIT | last commit april | Binderbot passing | launch binder

CLOUD STORAGE BENCHMARKS

Investigation of the throughput of various cloud storage formats and services. Prepared for the 2020 EarthCube Meeting by Ryan Abernathey.

license MIT | last commit june | Binderbot passing | launch binder

- Science in a GitHub repo:
  - http://gallery.pangeo.io/
  - Data in the cloud
  - reproducible with binder

➡️Play with cloud data yourself: launch binder

GMST historical+obs with intake

- Demo in Jupyter
- Pangeo-binder: https://github.com/mickaellalande/intake_CMIP6/tree/pangeo-notebook
- Intake experimental on CICLAD!
  (voir avec Guillaume Levasseur)

References

- Papers:
  - Rocklin, M. (2015). Dask: Parallel Computation with Blocked algorithms and Task Scheduling. 126–132. doi: 10.25080/Majora-7b98e3ed-013
  - 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
  - https://authorea.com/users/372628/articles/490577-cloud-native-repositories-for-big-scientific-data
- 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/
- Similar talks: Empowering Transformational Science - https://speakerdeck.com/cgentemann/empowering-transformational-science?slide=19