Pangeo for Plasma

Lessons for plasma software from the climate data analytics community

Thomas Nicholas
(Columbia University / Lamont-Doherty Earth Observatory)

thomas.nicholas@columbia.edu

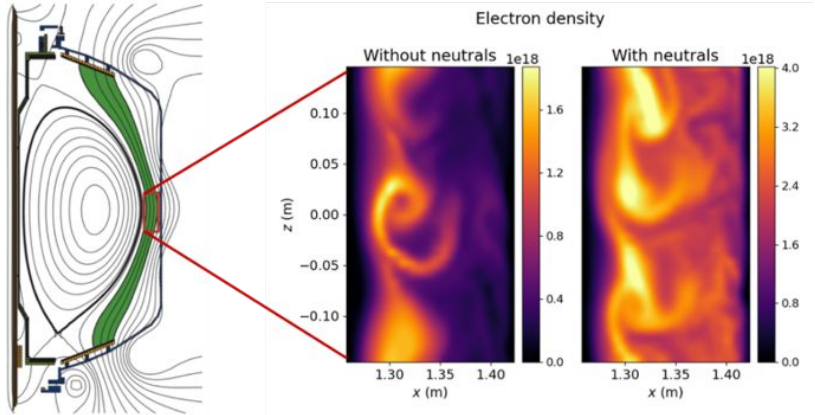
Who am I?

Who am I?

CCFE
CULHAM CENTRE FOR FUSION ENERGY

FUSION CDT

UNIVERSITY of York



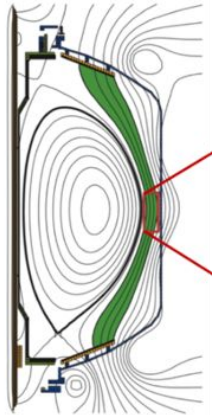
PhD with Ben Dudson, Fulvio Militello, BOUT++

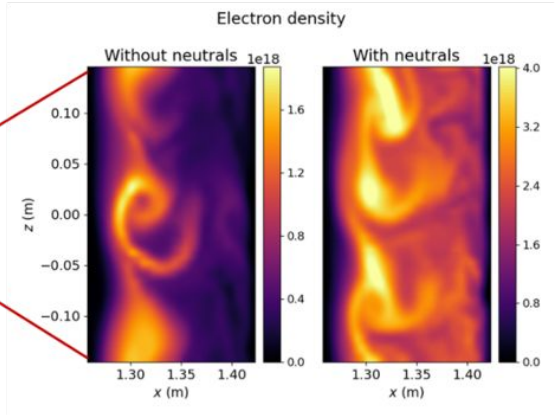
Who am I?

CCFE
CULHAM CENTRE FOR FUSION ENERGY

FUSION CDT

UNIVERSITY of York

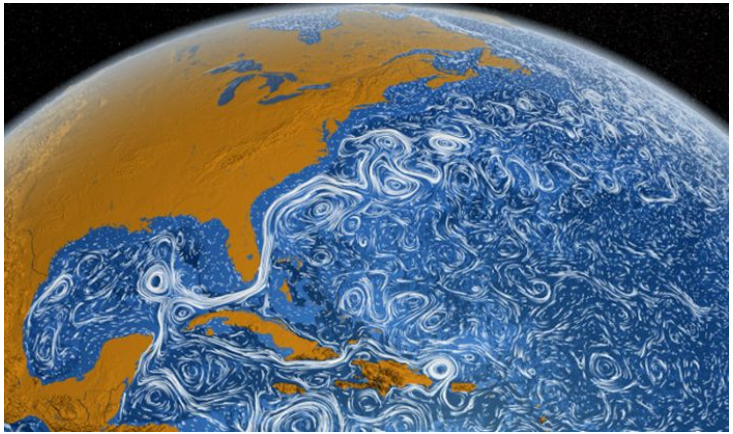




PhD with Ben Dudson, Fulvio Militello, BOUT++

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE

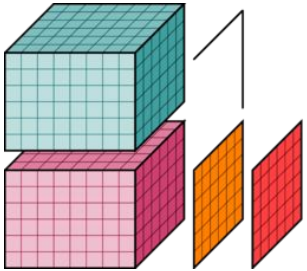
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



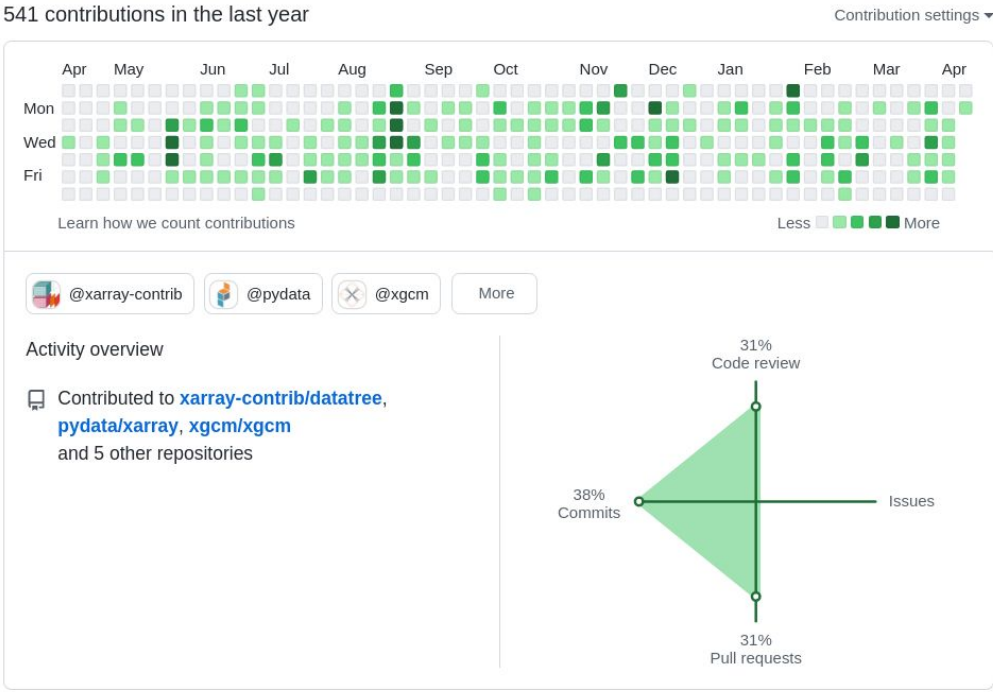
RSE with Ryan Abernathey, various projects

What do I do now?

PANGEO



xarray



What I hope to convince you of

- Our computational infrastructure **needs to change a lot**
- Can use **solutions from climate science** community
- **Modular approach** makes everyone's work easier
- **Opportunities exist** for plasma coders...

The White House announces

The Federal Year of Open Science

NASA ✦ NSF ✦ NOAA ✦ DOA ✦ DOC ✦ DOE ✦ GSA ✦ NEH ✦ NIH ✦ NIST ✦ USDA ✦ US

Along with other organizations, including CENDI group, voluntary collaboration among Federal managers, and HELIOS, a coalition of 80+ universities

A multi-agency initiative across the federal government to spark change and inspire open science engagement through events and activities that will advance adoption of open science.

Website: https://open.science.gov/

WH: https://www.whitehouse.gov/ostp/news-updates/

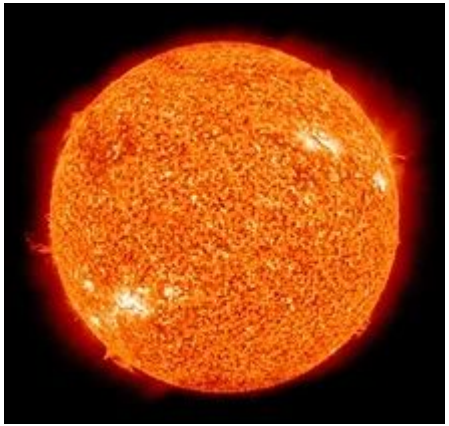
Nature: https://doi.org/10.1038/d41586-023-00019-y

Climate Science == Plasma Physics

- **Multidimensional** (often fluid turbulent)
- **Large** (bigger than local RAM)
- On regular but warped **grids**
- Often pulled from **central** servers
- From multiple sources but with **common structure** (e.g. experimental and simulation data for same device).

Climate Science == Plasma Physics

- **Multidimensional** (often fluid turbulent)
- **Large** (bigger than local RAM)
- On regular but warped **grids**
- Often pulled from **central** servers



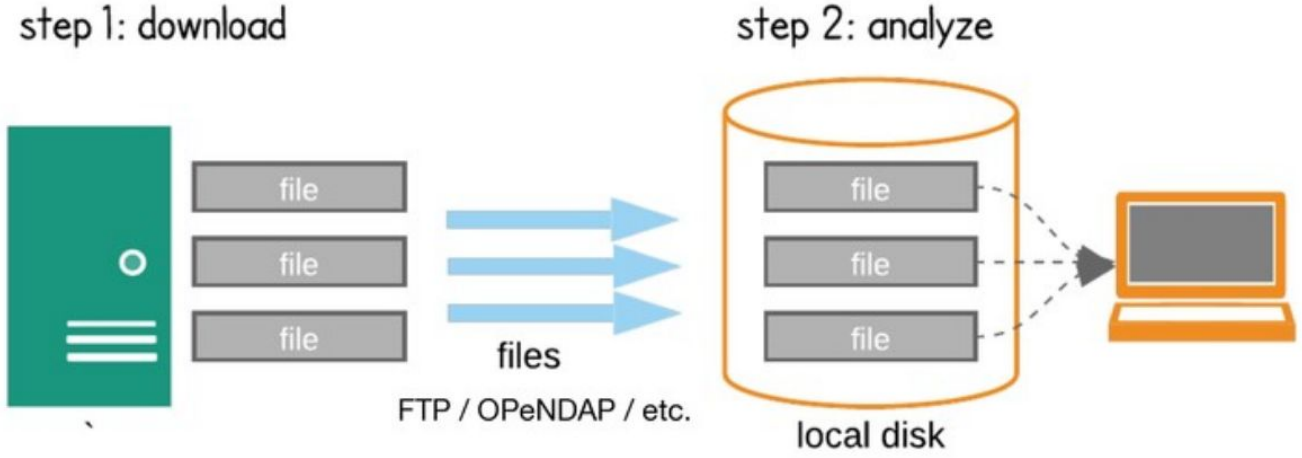
=

- From multiple sources but with **common structure** (e.g. experimental and simulation data for same device).

Typical scientific workflow

Typical scientific workflow

DOWNLOAD

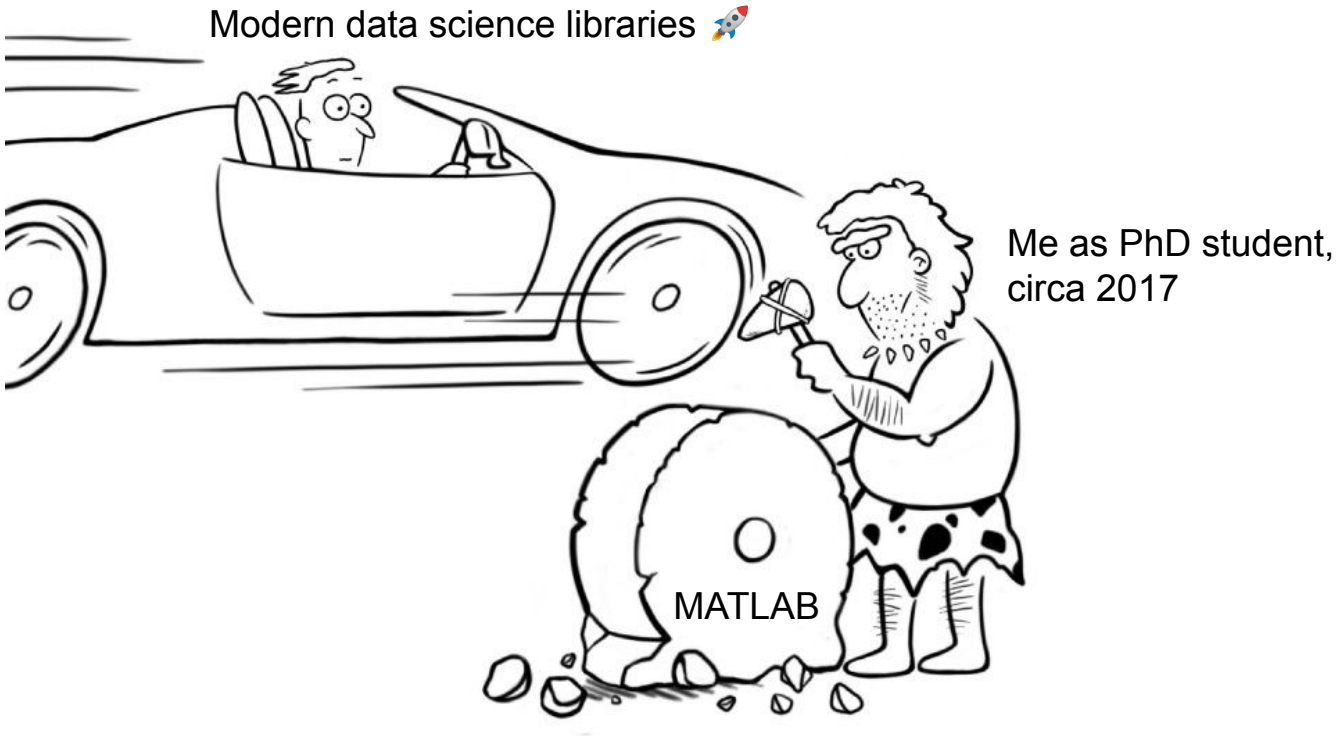


step 3: debug

Because you likely rolled-your-own code…

Problem 1: Code not reused

Problem 1: Code not reused



Problem 2: Data accessibility

Problem 2: Data accessibility

PRIVILEGED INSTITUTIONS CREATE
"DATA FORTRESSES*"

Image credit: Moahim, CC BY-SA 4.0, via Wikimedia Commons

*Coined by Chelle Gentemann

Problem 2: Data accessibility

PRIVILEGED INSTITUTIONS CREATE
"DATA FORTRESSES*"

Image credit: Moahim, CC BY-SA 4.0, via Wikimedia Commons

*Coined by Chelle Gentemann

Problem 2: Data accessibility

PRIVILEGED INSTITUTIONS CREATE
"DATA FORTRESSES*"

NIF

GENERAL ATOMICS

CCFE
CULHAM CENTRE FOR FUSION ENERGY

Image credit: Moahim, CC BY-SA 4.0, via Wikimedia Commons

*Coined by Chelle Gentemann

Problem 3: Scale

“Brb, let me just go download the data to my laptop…”

Problem 3: Scale

“Brb, let me just go download the data to my laptop…”

MB

Problem 3: Scale

“Brb, let me just go download the data to my laptop…”

GB

Problem 3: Scale

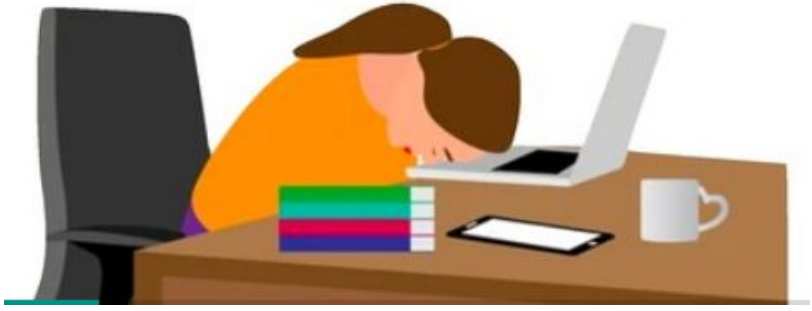
“Brb, let me just go download the data to my laptop…”

TB

Problem 3: Scale

“Brb, let me just go download the data to my laptop…”

PB 😱



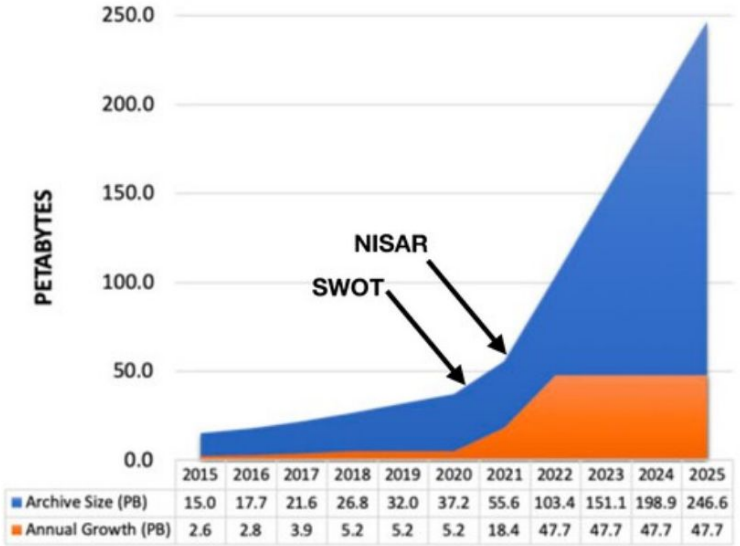
Problem 3: Scale

ITER NEWSLINE -

30 NOV, 2020 Print | Read the latest published articles

ITER Scientific Data Centre

HOW TO MANAGE 2 PETABYTES OF NEW DATA EVERY DAY



	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Archive Size (PB)	15.0	17.7	21.6	26.8	32.0	37.2	55.6	103.4	151.1	198.9	246.6
Annual Growth (PB)	2.6	2.8	3.9	5.2	5.2	5.2	18.4	47.7	47.7	47.7	47.7

https://earthdata.nasa.gov/eosdis/cloud-evolution

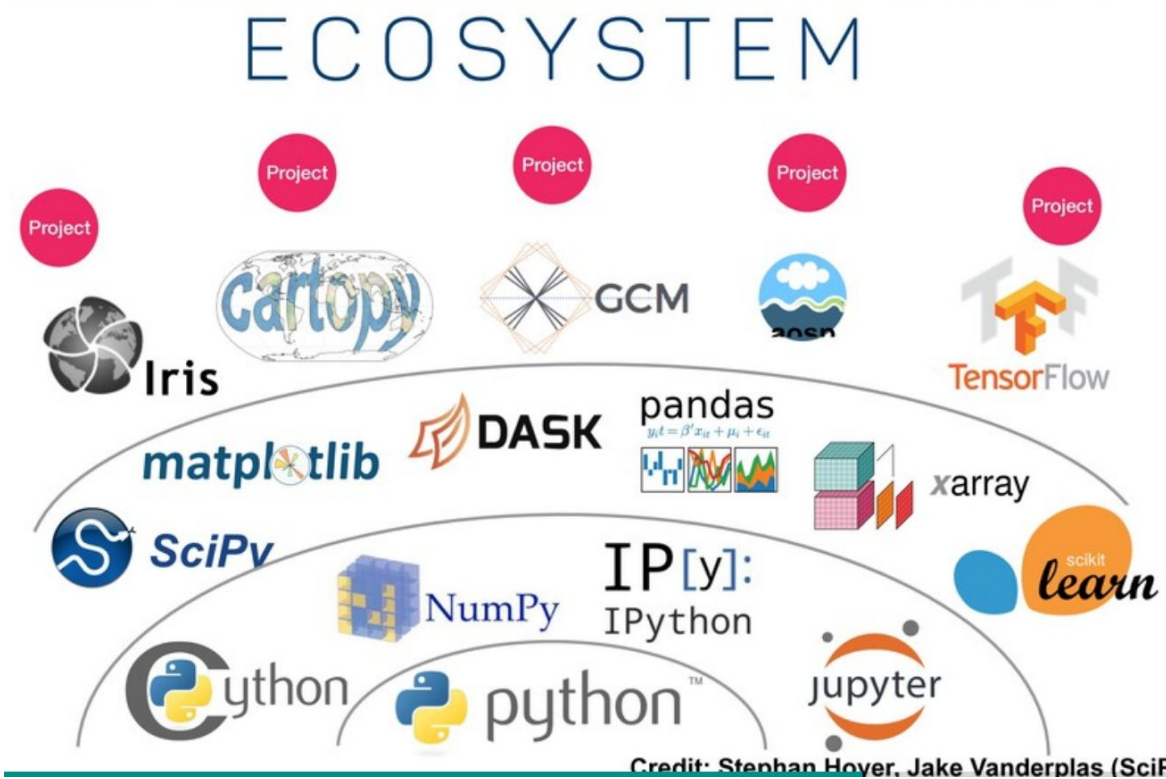
Geoscientists' solution:

Geoscientists' solution:

A community platform for Big Data geoscience

Solution 1: Modular, open ecosystem

Solution 1: Modular, open ecosystem



Science projects

Domain-specific packages

Domain-agnostic libraries

General-purpose tools

Solution 1: Modular, open ecosystem

PANGEO COMMUNITY

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE

NSF

EARTHCUBE

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Met Office

NASA

USGS

ECMWF

GEOMAR

UNIVERSITY of WASHINGTON
eScience Institute

BIDS
BERKELEY INSTITUTE FOR DATA SCIENCE

jupyter

cnes
CENTRE NATIONAL D'ÉTUDES SPATIALES

CSIRO

ANACONDA
Powered by Continuum Analytics

JUPITER

element84

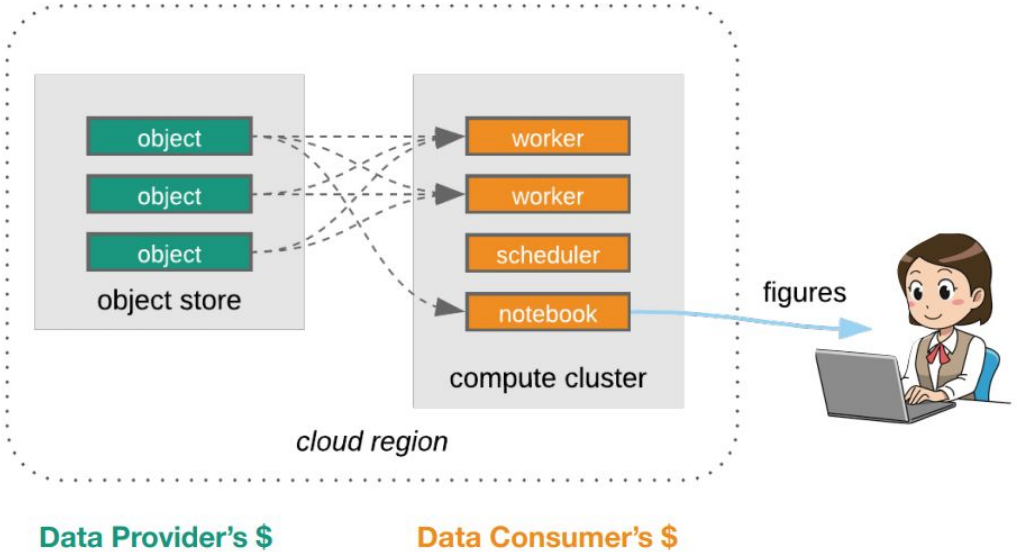
developmentSEED

RHODIUM GROUP

CLIMACELL
Weather Revealed

HTTP://PANGEO.IO

Solution 2: Cloud Computing



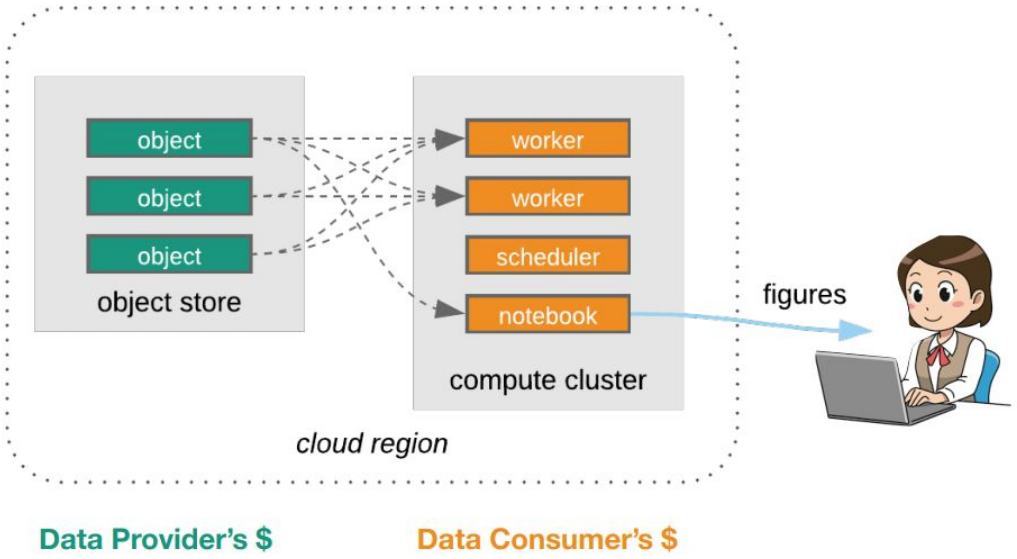
Solution 2: Cloud Computing

Analysis Ready Data
Cloud Optimized Formats

Parquet

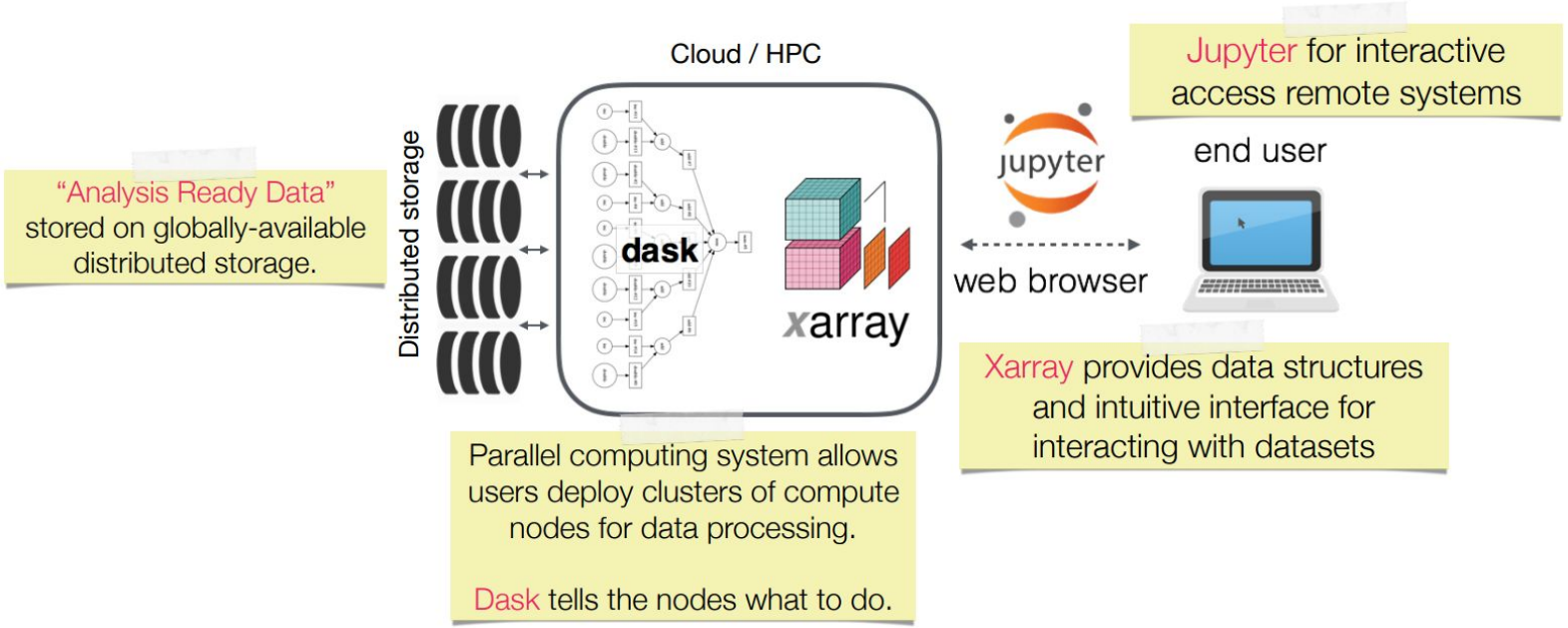
COG
CLOUD OPTIMIZED
GEOTIFF

Zarr



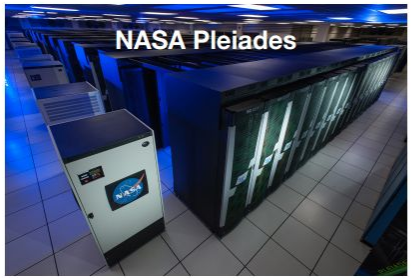
Solution 2: Cloud Computing

PANGEO ARCHITECTURE



Solution 2: Cloud Computing

PANGEO DEPLOYMENTS



OCEAN.PANGEO.IO

Google Cloud Platform

NCAR Cheyenne

Microsoft Azure

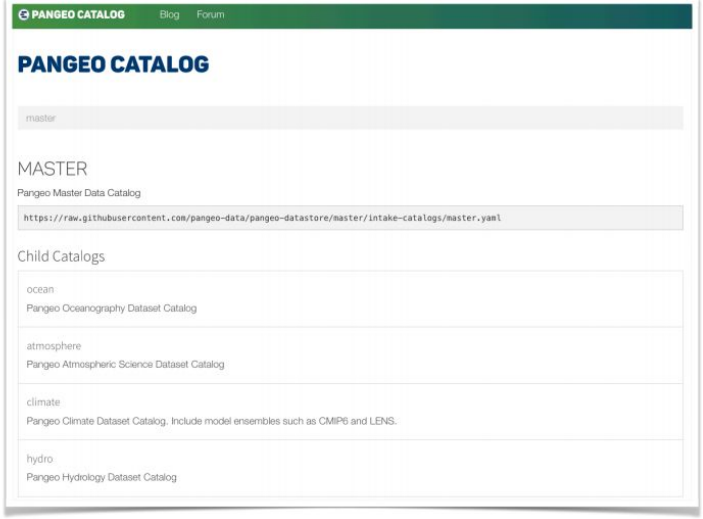
aws

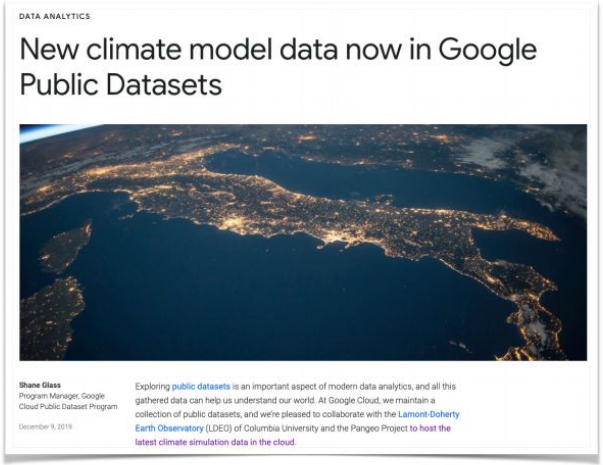
HTTP://PANGEO.IO/DEPLOYMENTS.HTML

Solution 2: Cloud Computing

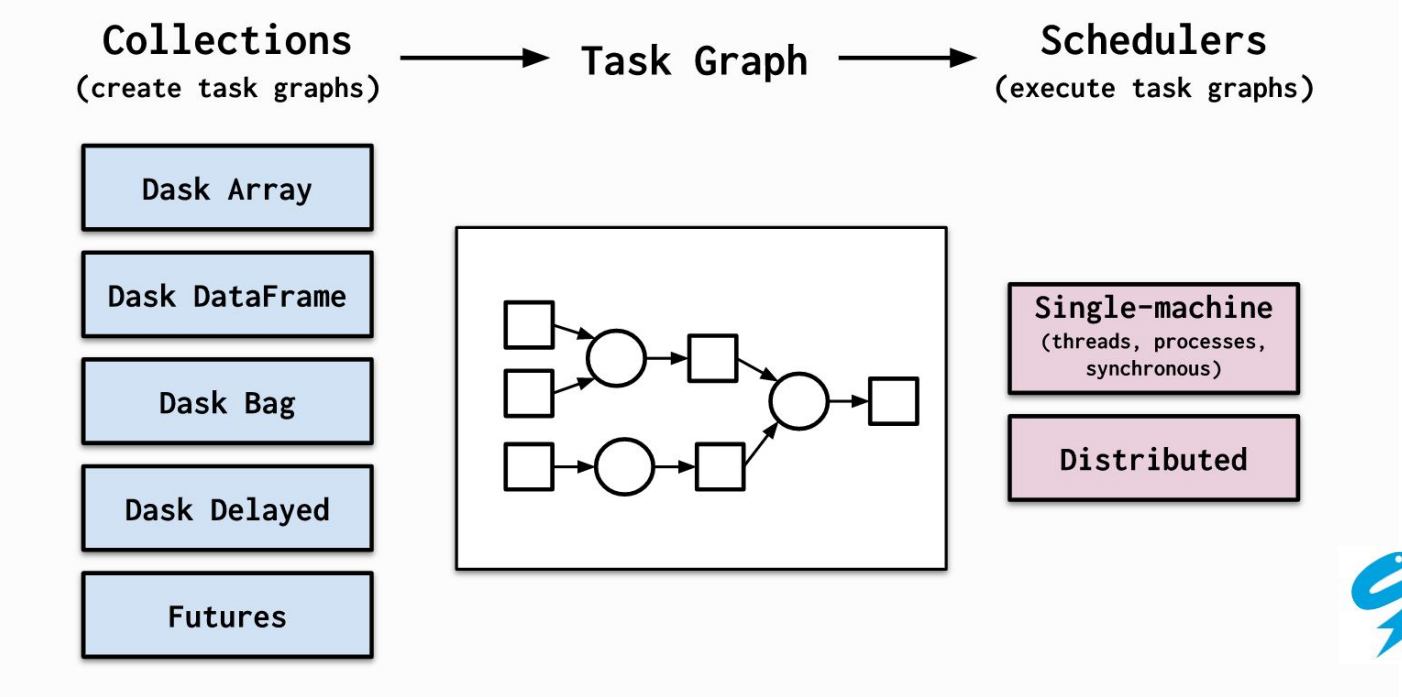
PANGEO CLOUD DATA CATALOG

CATALOG.PANGEO.IO





Solution 3: Parallel computing frameworks

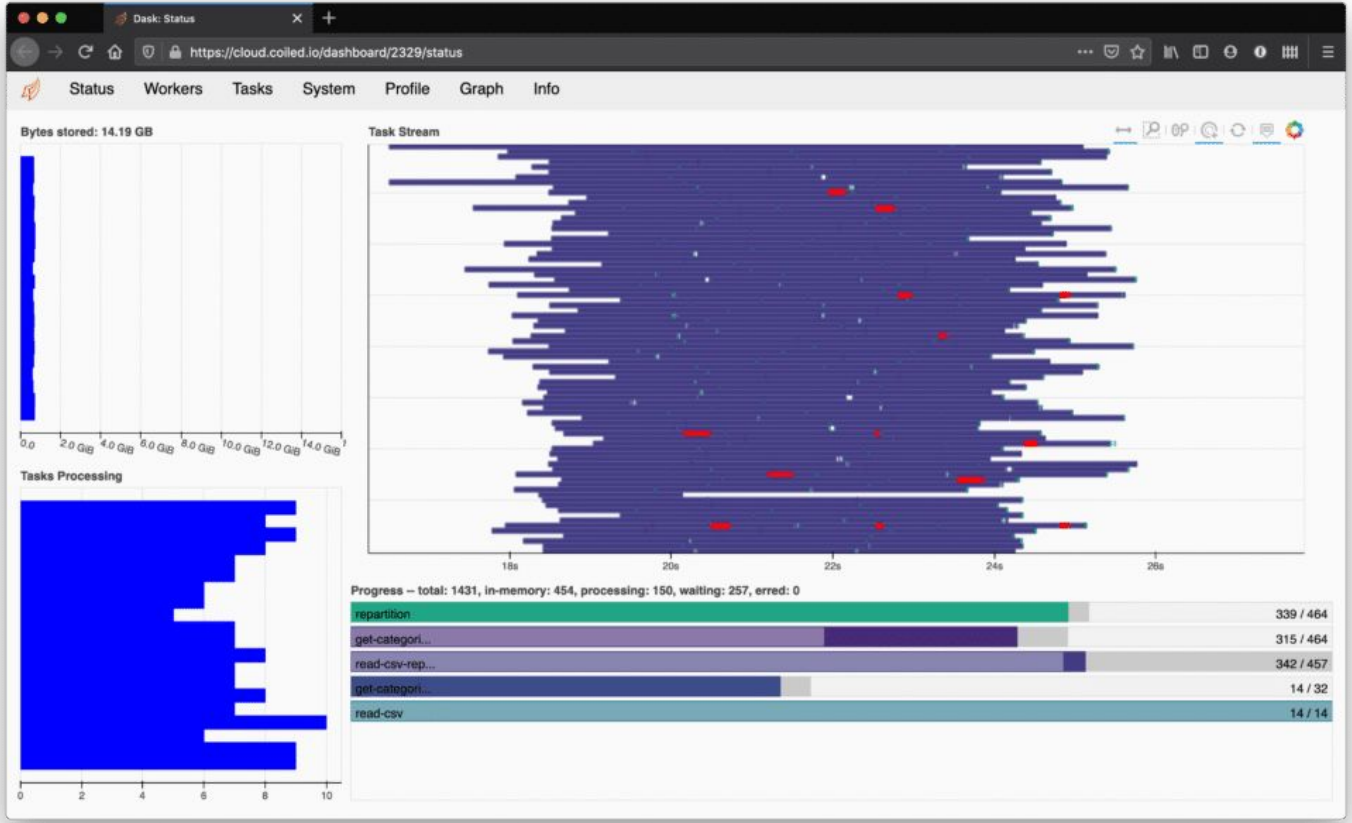


DASK

APACHE Spark™

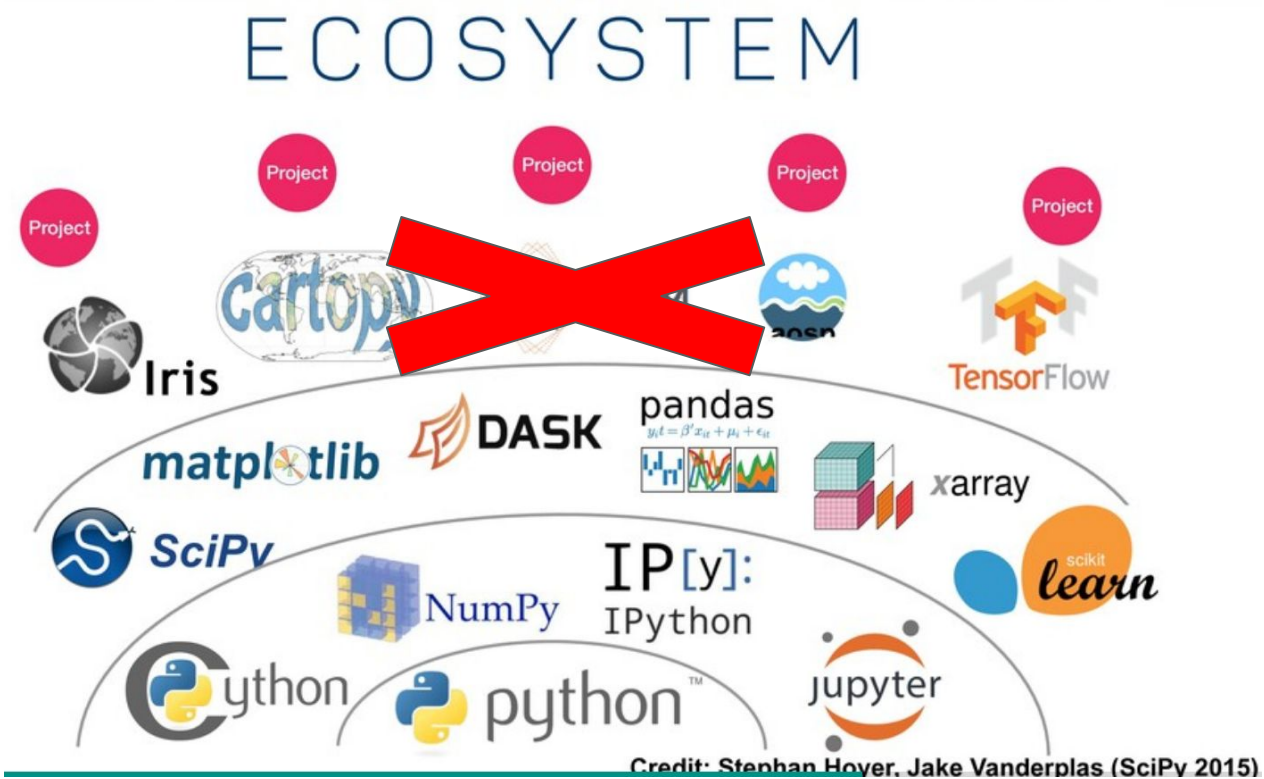
Numba

Solution 3: Parallel computing frameworks



How might this work for plasma?

How might this work for plasma?

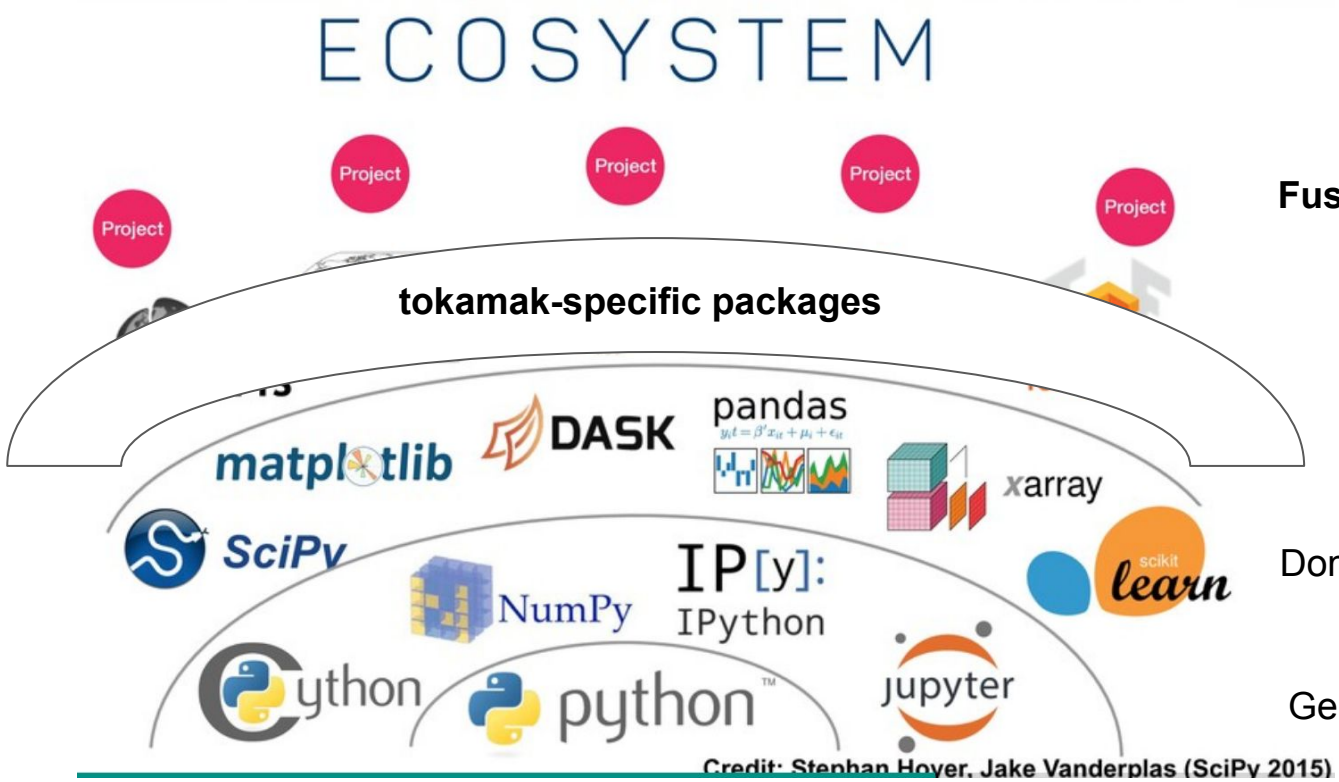


Fusion plasma projects

Domain-agnostic libraries

General-purpose tools

How might this work for plasma?



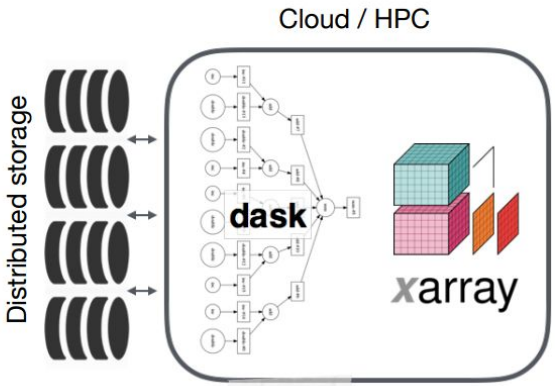
Fusion plasma projects

Domain-agnostic libraries

General-purpose tools

How might this work for plasma?

PANGEO ARCHITECTURE -PLASMA



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

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

Dask tells the nodes what to do.

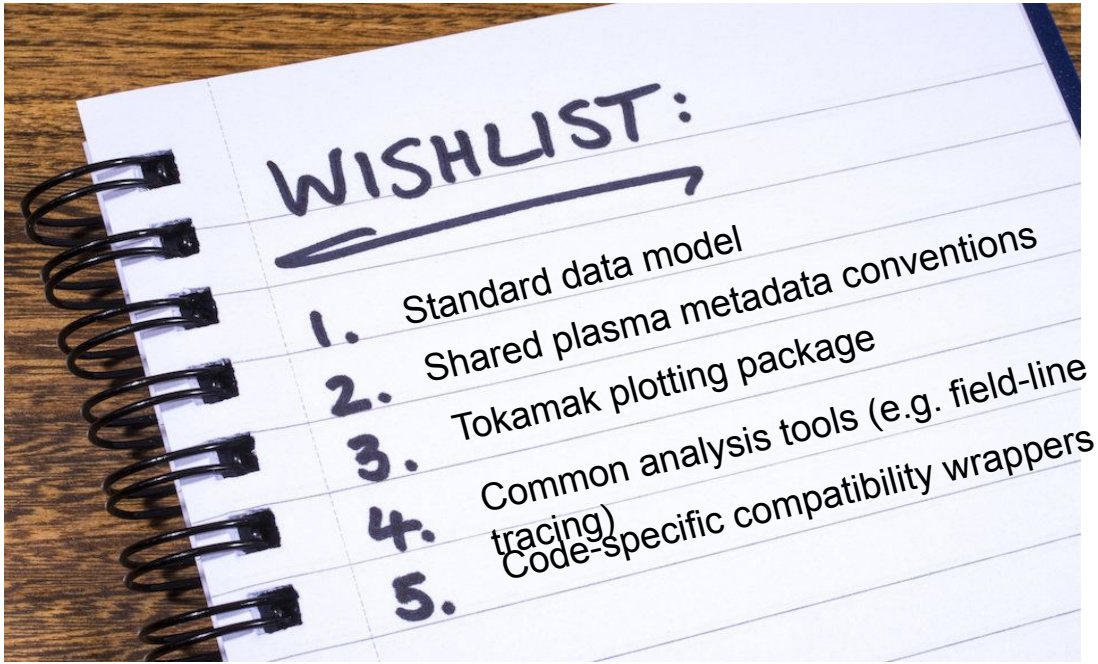
web browser

Jupyter for interactive access remote systems

end user

Xarray provides data structures and intuitive interface for interacting with datasets

How might this work for plasma?



Blog post: https://hackmd.io/@TomNicholas/rkyERwcoO#

Other bonuses of joining this ecosystem

- Parallel and out-of-core analysis
- Labelled dimensions
- Unit-aware arithmetic
- Easier reproducibility
- Plotting flexibility
- Machine Learning integration

Summary

- **Geoscience has same problems as plasma physics 🌍🤝🌞**
- **Being solved using:**
  - **Modular community software ecosystem 🔧**
  - **Cloud computing 🌩️**
  - **Parallel execution frameworks 🚀**
- **It's working for them - it could work for us! 🔬**

LEARN MORE

http://pangeo.io

https://github.com/pangeo-data/

https://medium.com/pangeo

@pangeo_data