http://pangeo.io

https://discourse.pangeo.io/

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

https://medium.com/pangeo

@pangeo_data

PANGEO

AN OPEN SOURCE ECOSYSTEM FOR DATA-INTENSIVE SCIENCE

MAELSTROM Meeting
March 28, 2022

WHO AM I?

Physical Oceanographer

Ph.D. From MIT, 2012

Associate Prof. at Columbia / LDEO

https://ocean-transport.github.io/

Core developer of *Xarray*

Core developer of *Zarr*

Co-founder of *Pangeo*

Open Source Advocate

TAKE HOME MESSAGES

Data is at the center of AI/ML. Let's value data production, curation, and infrastructure as a core part of our discipline.

🤓

Don't build a highly customized data platform.
Leverage modular, open-source components and cloud-style computing.
Make your users happy.

😎

PANGEO

TWO PAPERS

THEME ARTICLE: JUPYTER IN COMPUTATIONAL SCIENCE

Cloud-Native Repositories for Big Scientific Data

Ryan P. Abernathey, Charles C. Blackmon-Luca, Timothy J. Crone, Naomi Henderson, Chiara Lepore, *Lamont–Doherty Earth Observatory of Columbia University, Palisades, NY, 10964, USA*

Tom Augspurger, *Anaconda, Des Moines, IA, 50313, USA*

Anderson Banihirwe and Joseph J. Hamman, *National Center for Atmospheric Research, Boulder, CO, 80305, USA*

Chelle L. Gentemann, *Farallon Institute, Petaluma, CA, 94952, USA*

Theo A. McCaie and Niall H. Robinson, *Met Office, University of Exeter, Exeter EX4 4PY, U.K.*

Richard P. Signell, *US Geological Survey, Woods Hole, MA, 02543, USA*

https://doi.org/10.1109/MCSE.2021.3059437

https://doi.org/10.1029/2020AV000354

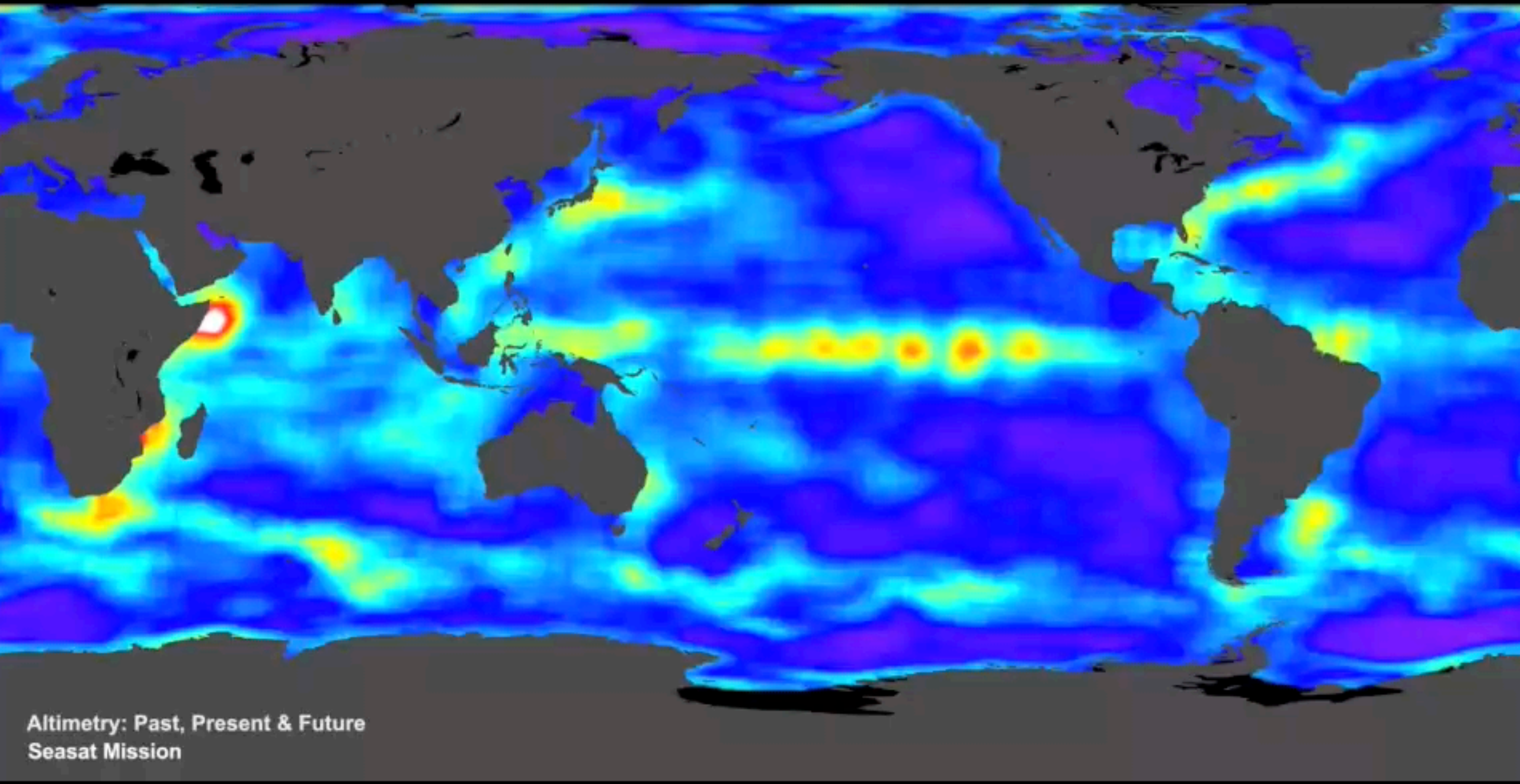
AGU Advances

Research Article | Open Access

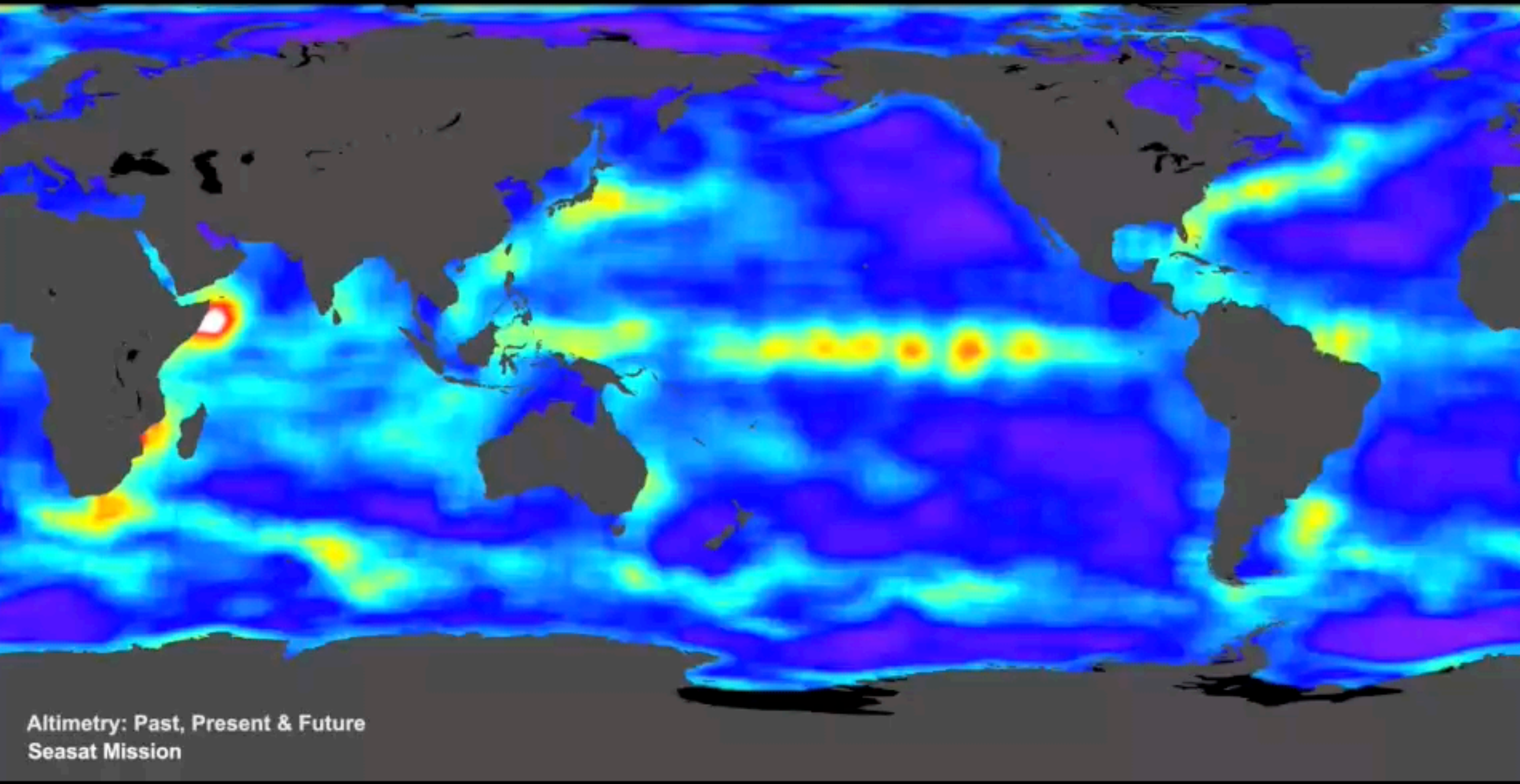
Science Storms the Cloud

C. L. Gentemann, C. Holdgraf, R. Abernathey, D. Crichton, J. Colliander, E. J. Kearns, Y. Panda, R. P. Signell,

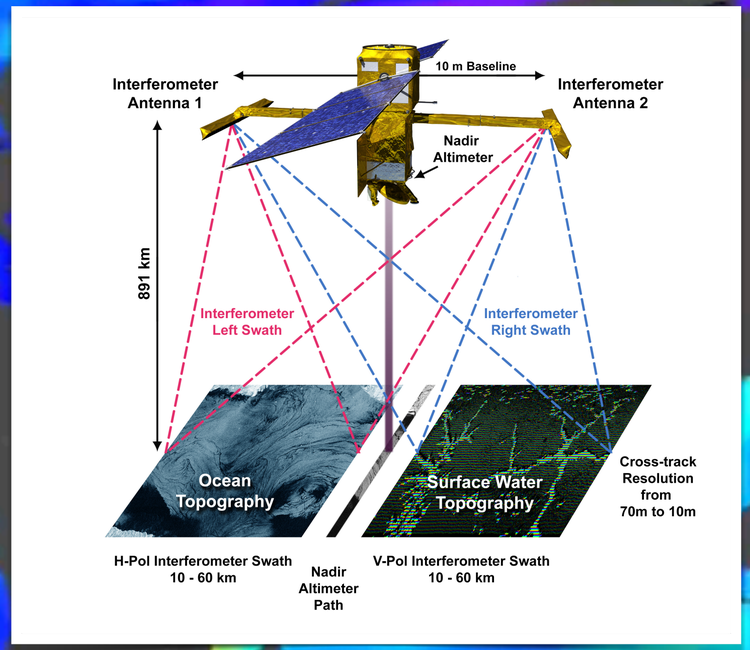
First published: 05 May 2021 | https://doi.org/10.1029/2020AV000354



Credit: JPL / NASA PO.DAAC

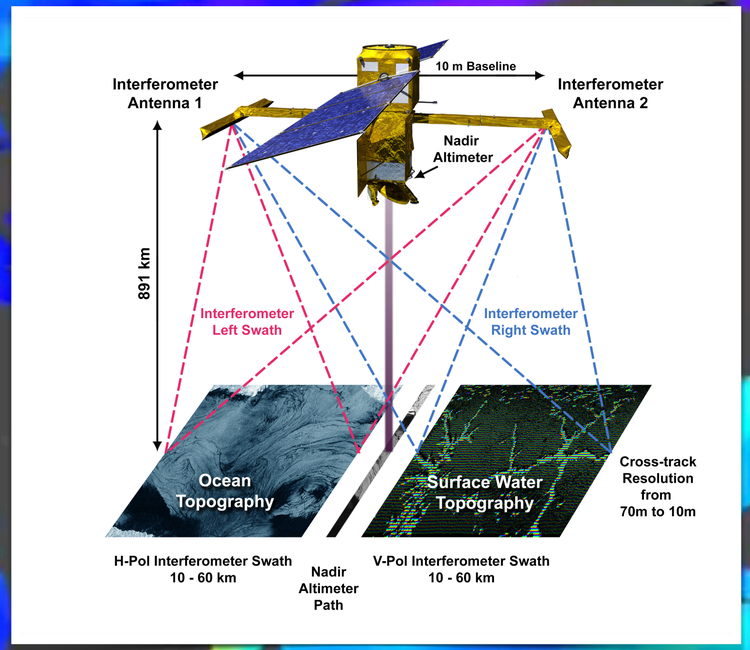


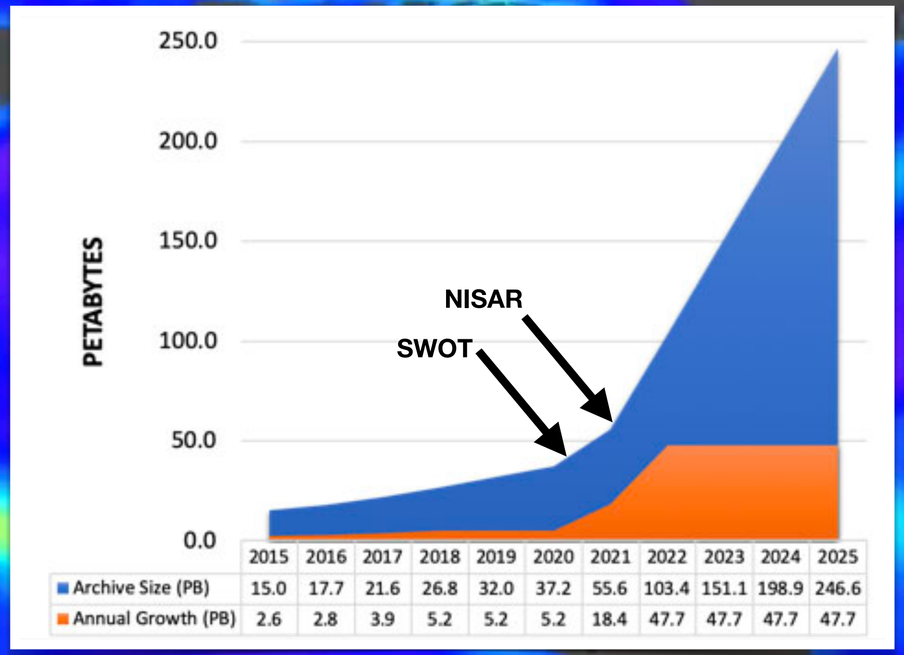
Credit: JPL / NASA PO.DAAC



Altimetry: Past, Present & Future
Seasat Mission

Credit: JPL / NASA PO.DAAC





	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

Altimetry: Past, Present & Future
Seasat Mission

Credit: JPL / NASA PO.DAAC

Global Eddy Resolving Simulations

Chris Hill (MIT) & Dimitris Menemenlis (JPL)

Global Eddy Resolving Simulations

Chris Hill (MIT) & Dimitris Menemenlis (JPL)

Global Eddy Resolving Simulations

Chris Hill (MIT) & Dimitris Menemenlis (JPL)

Ocean deformation radius: 10 km $\frac{NH}{f}$

Required spatial resolution: 1 km
Number of horizontal grid points: 10^8
Number of vertical levels: 100
Size of a single scalar field: 40 GB

Number of state variables: 10

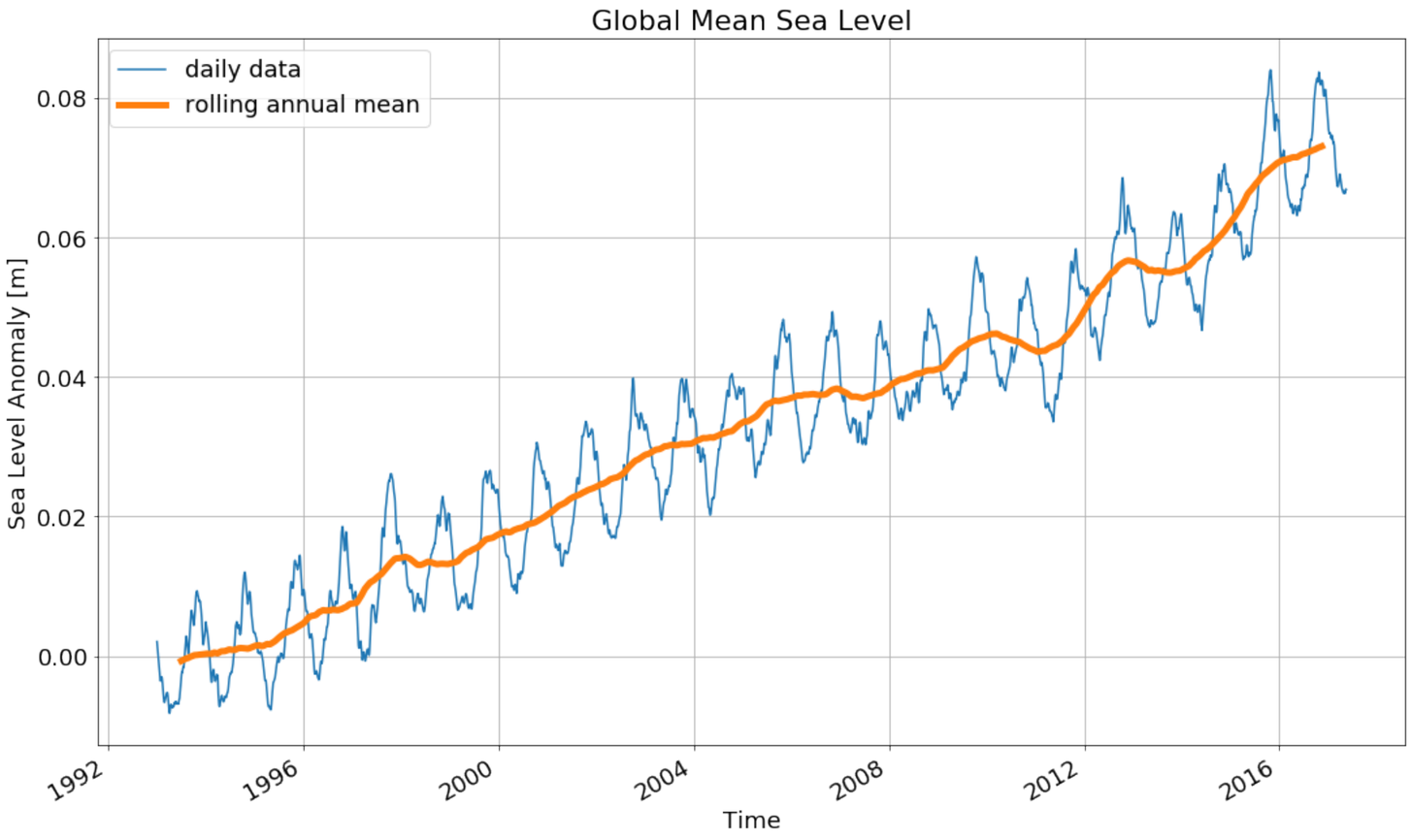
Required temporal resolution: 1 hour
Number of outputs per year: 10^5

Raw dataset size: 40 PB
Compressed dataset size: 1 PB

WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

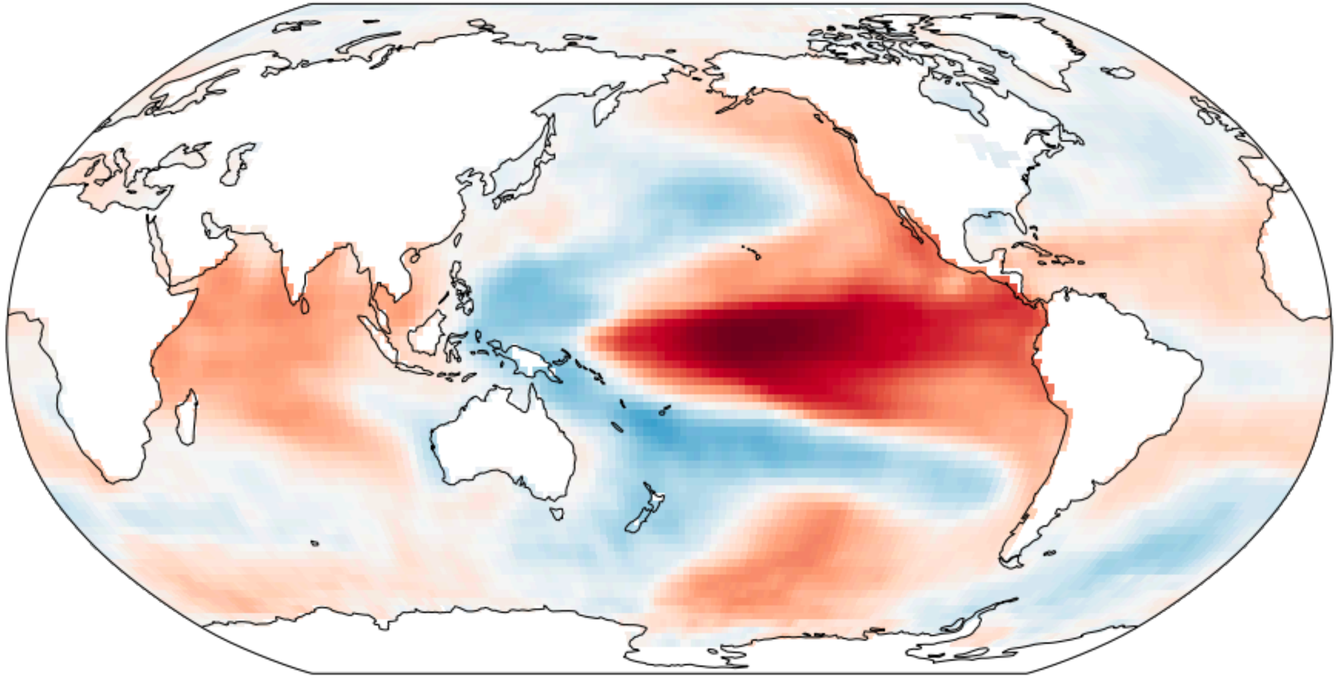
WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

Take the mean!



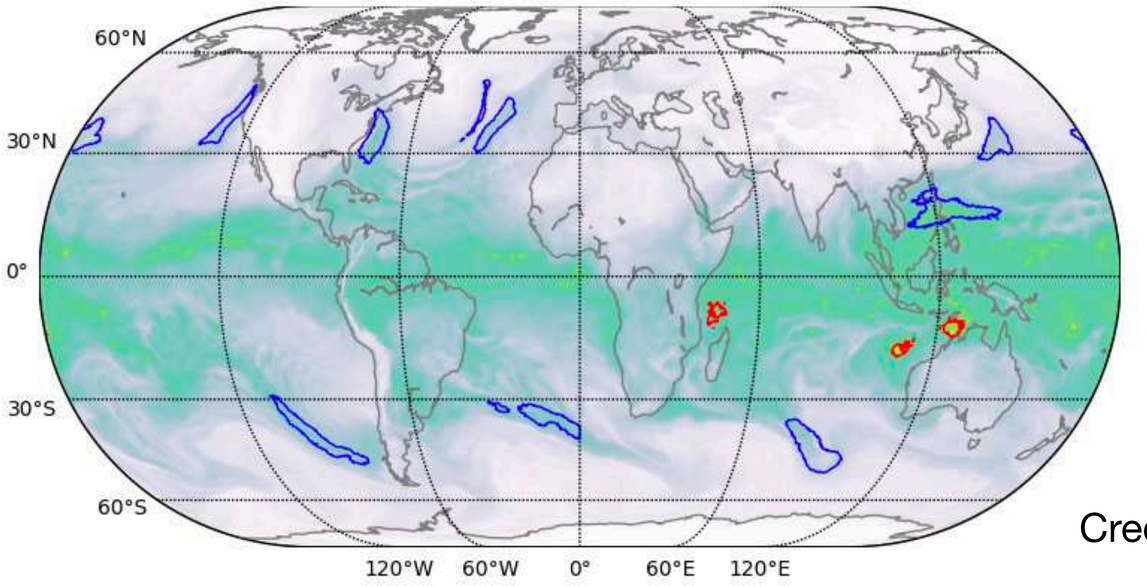
WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

Analyze spatiotemporal variability

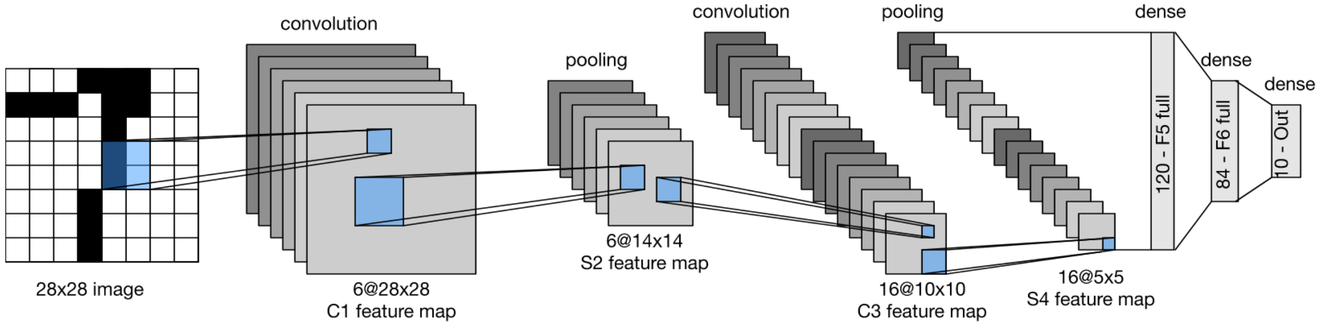


WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

Machine learning!

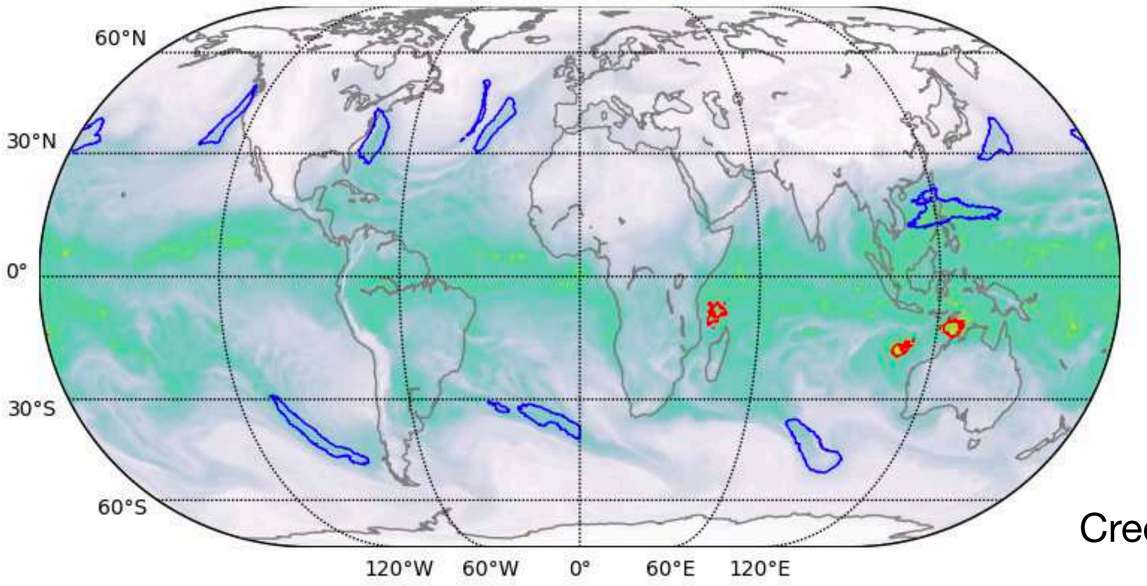


Credit: Berkeley Lab

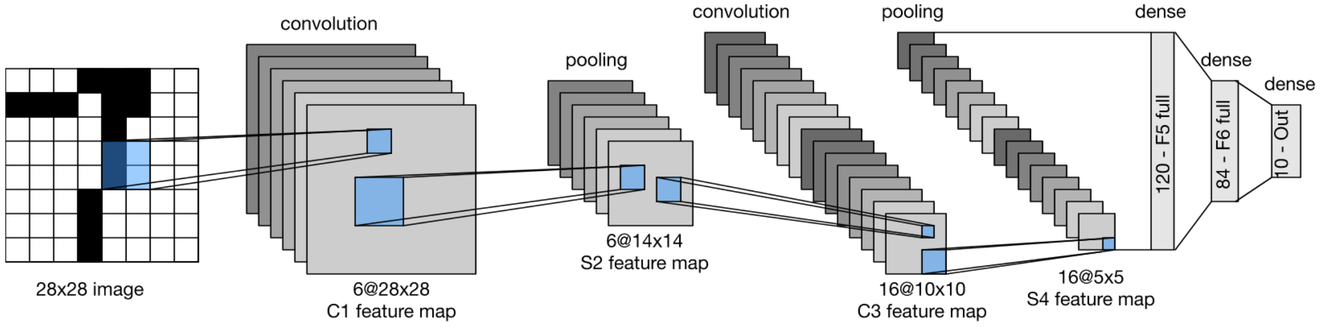


WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

Machine learning!

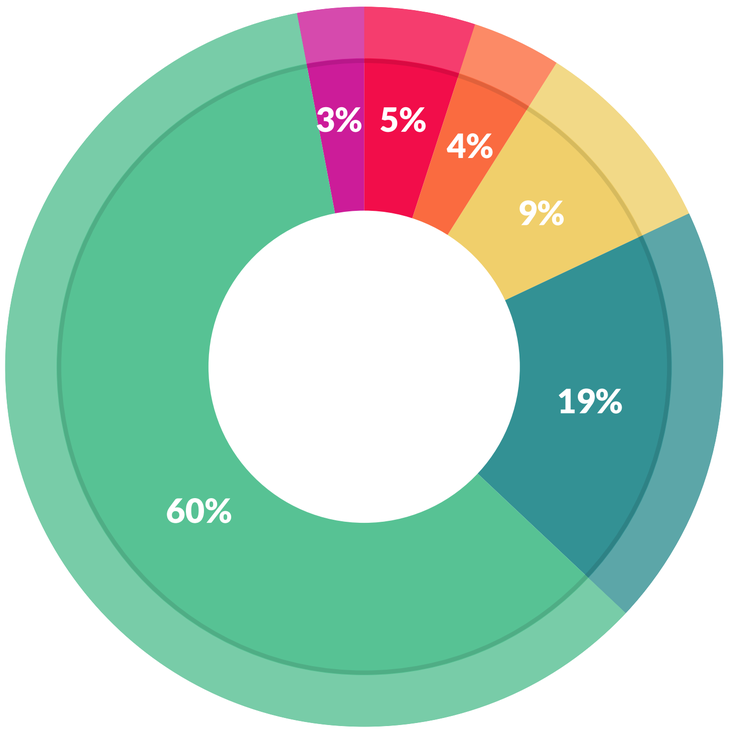


Credit: Berkeley Lab



80/20 RULE OF DATA SCIENCE

How do data scientists spend their time?



- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Crowdflower Data Science Report (2016)

PANGEO

"DATA WORK" IN AI

PUBLICATIONS ›

"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, Lora Mois Aroyo

SIGCHI, ACM (2021)

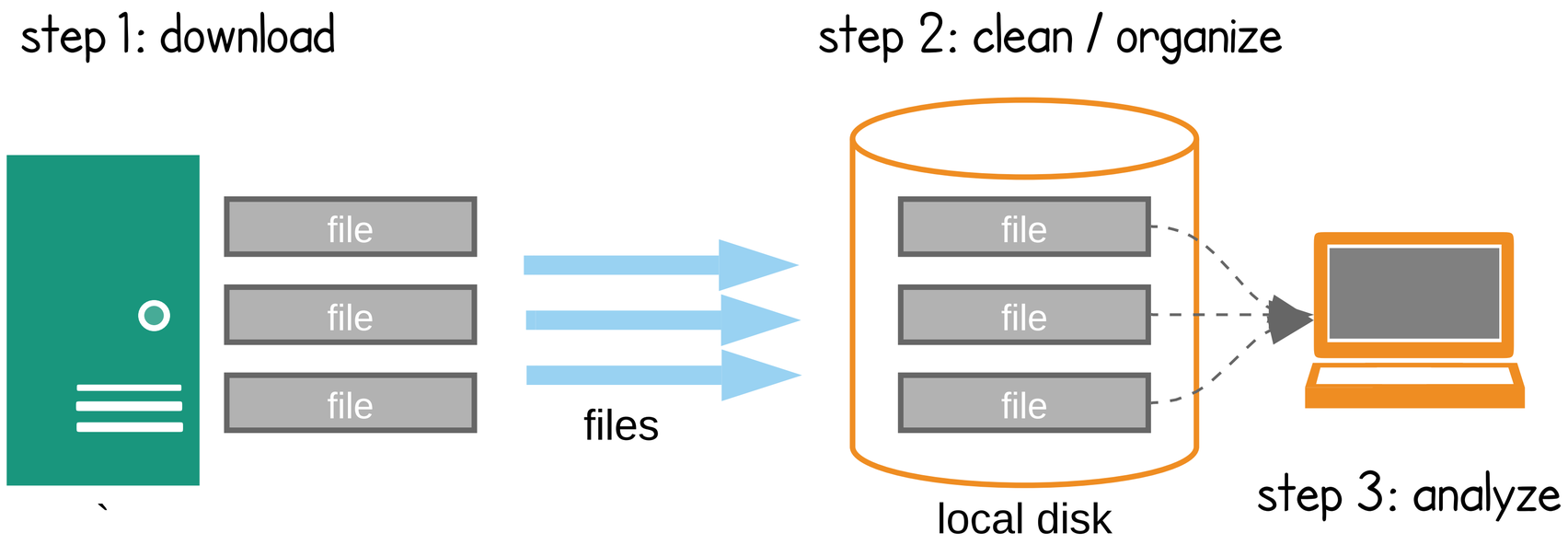
Download | Google Scholar | Copy Bibtex

Abstract

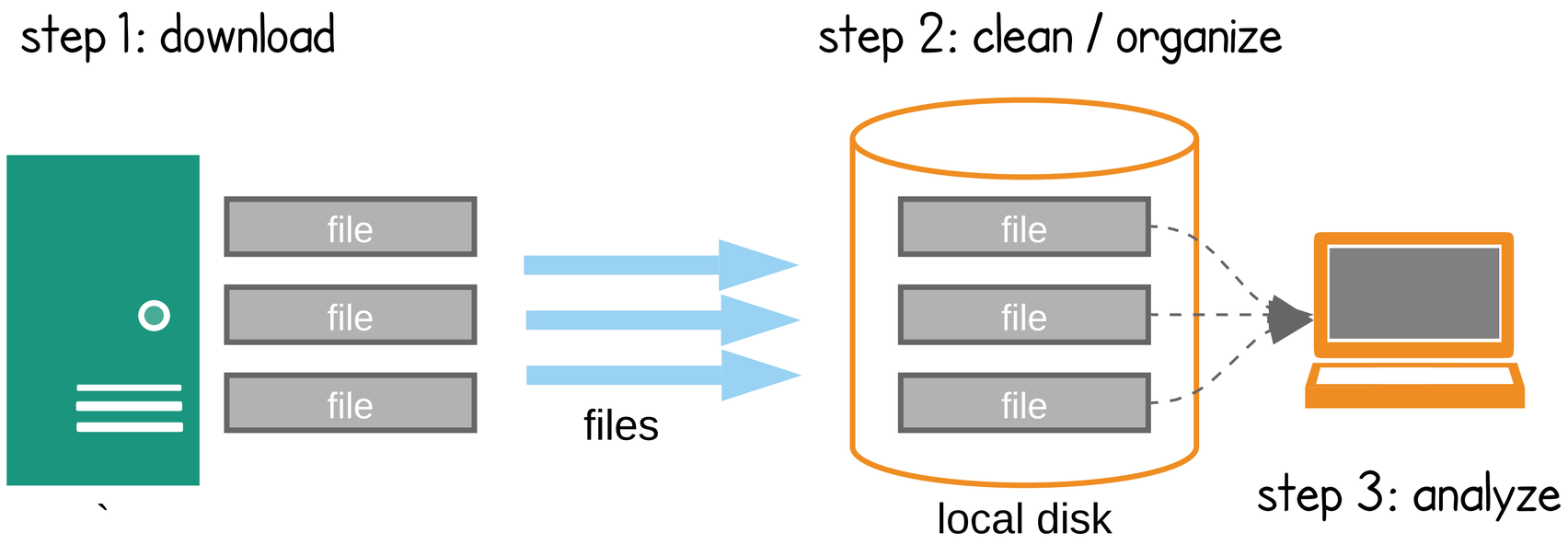
AI models are increasingly applied in high-stakes domains like health and conservation. Data quality carries an elevated significance in high-stakes AI due to its heightened downstream impact, impacting predictions like cancer detection, wildlife poaching, and loan allocations. Paradoxically, data is the most under-valued and de-glamorised aspect of AI. In this paper, we report on data practices in high-stakes AI, from interviews with 53 AI practitioners in India, East and West African countries, and USA. We define, identify, and present empirical evidence on Data Cascades---compounding events causing negative, downstream effects from data issues---triggered by conventional AI/ML practices that undervalue data quality. Data cascades are pervasive (92% prevalence), invisible, delayed, but often avoidable. We discuss HCI opportunities in designing and incentivizing data excellence as a first-class citizen of AI, resulting in safer and more robust systems for all.

"…incentivizing data excellence as a first-class citizen of AI…"

THE "DOWNLOAD" MODEL

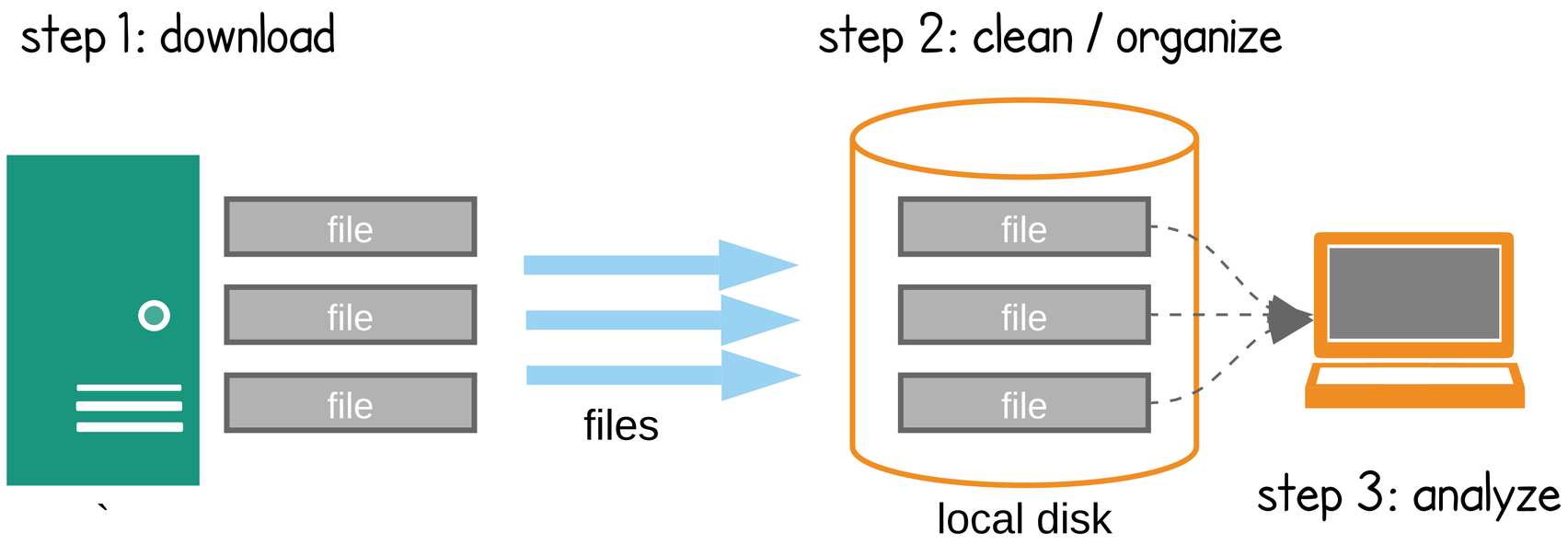


THE "DOWNLOAD" MODEL



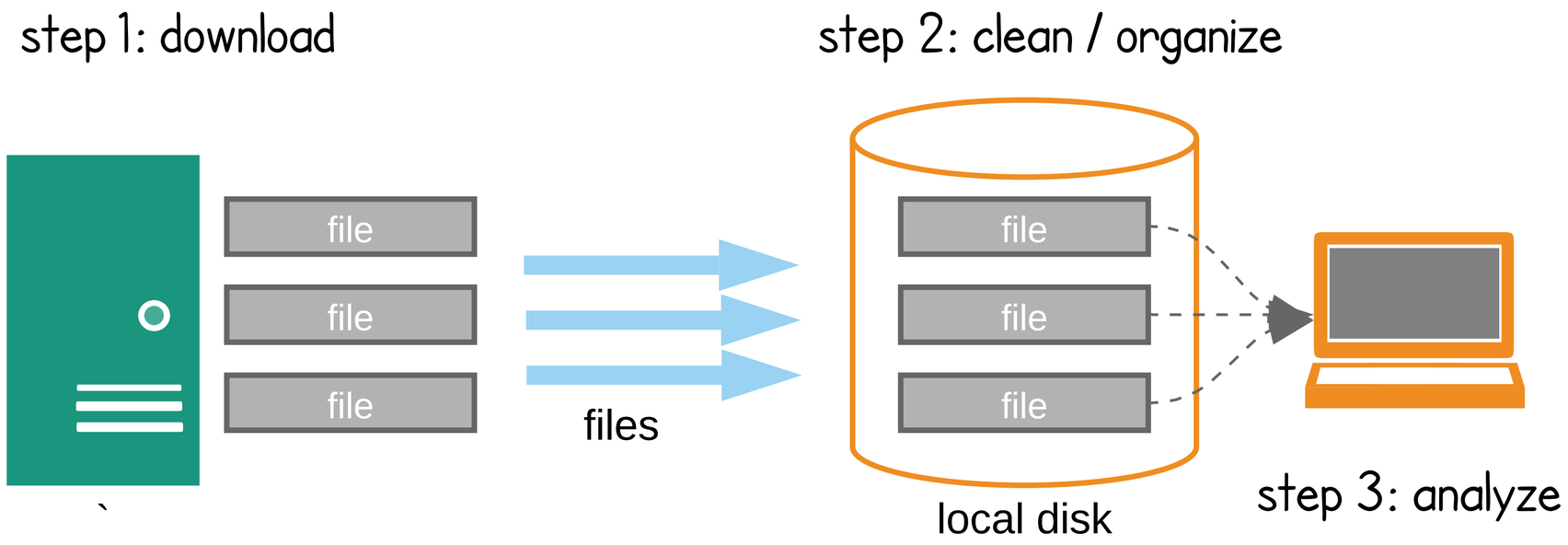
MB

THE "DOWNLOAD" MODEL



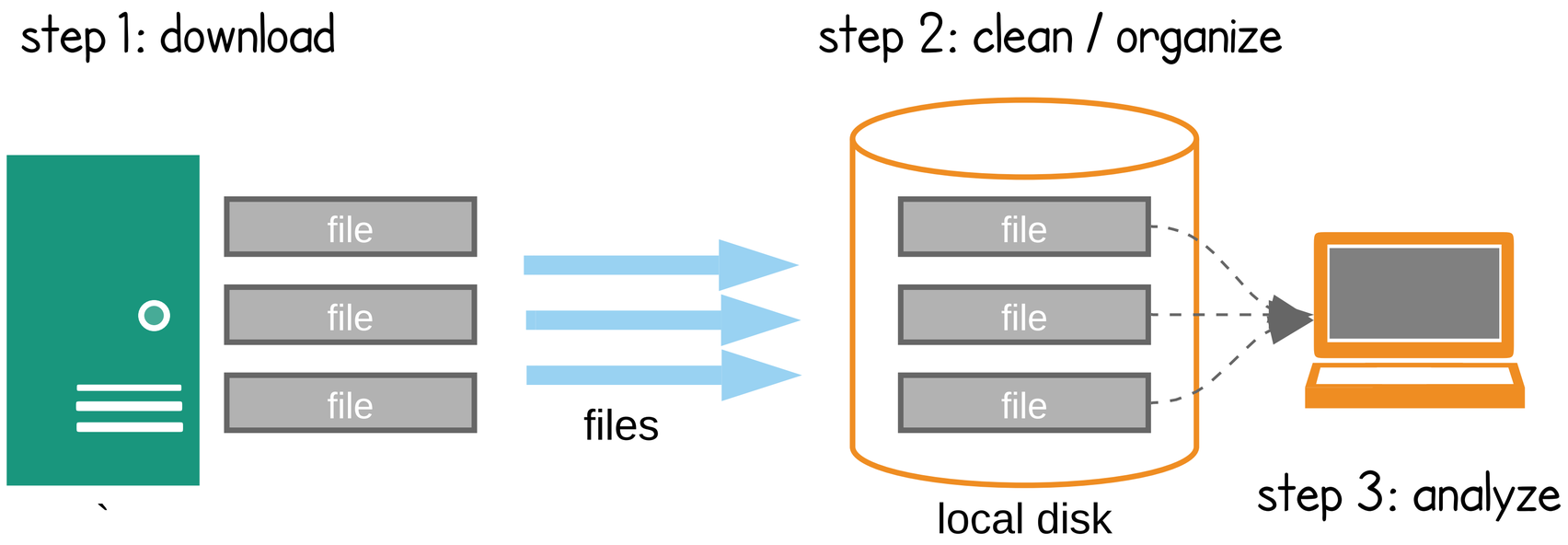
GB 😐

THE "DOWNLOAD" MODEL



TB

THE "DOWNLOAD" MODEL



PB 😱

NEVER MIND...

Let's "bring the compute to the data"!

HOW?

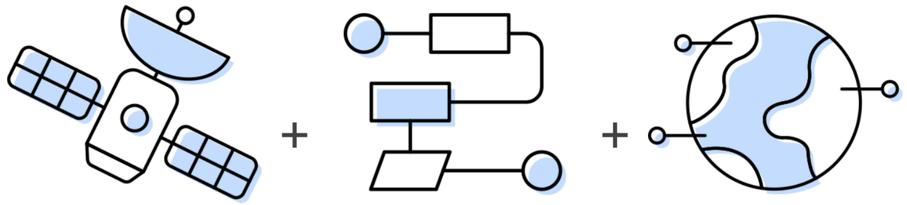
USE A "PLATFORM"

Google Earth Engine

Datasets FAQ Timelapse Case Studies Platform Blog Sign Up

Meet Earth Engine

Google Earth Engine combines a multi-petabyte catalog of satellite imagery and geospatial datasets with planetary-scale analysis capabilities and makes it available for scientists, researchers, and developers to detect changes, map trends, and quantify differences on the Earth's surface.



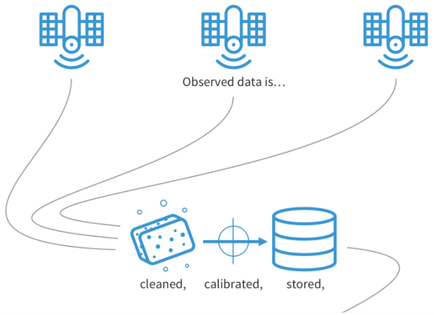
Satellite Imagery Your Algorithms Real World Applications

Descartes Labs

Platform Solutions Demos Company Contact Us

A platform for complex global systems

The Descartes Labs Platform is the missing link in making real world sensor data - from satellite imagery to weather - useful. We collect data daily from public and commercial sources, clean it, calibrate it, and store it in an easy-to-access catalog, ready for scientific analysis. We also give our customers the ability to throw huge amounts of computational power at that data and tools built on top of the data to make it easier to build models.



THE TROUBLE WITH "PLATFORMS"

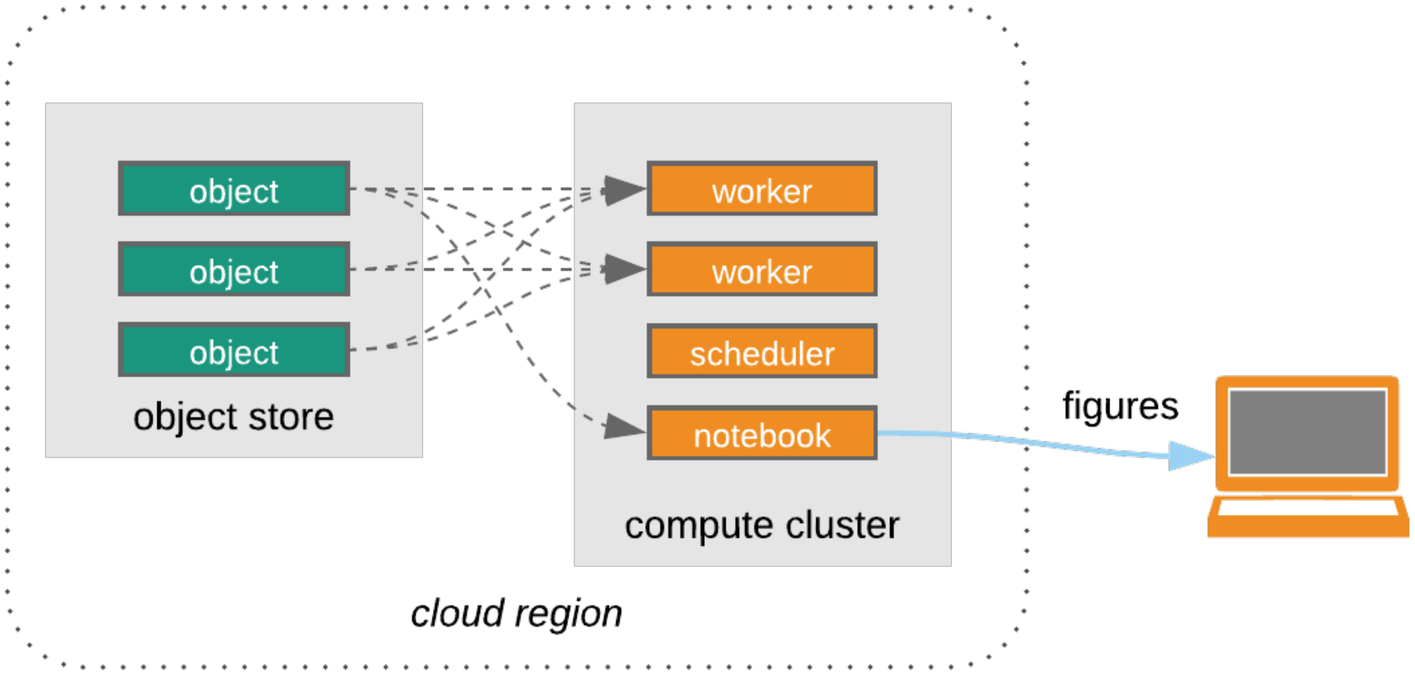
- Scientists' creativity often exceeds pre-baked capabilities.
  Desire to go under the hood

- What if you want to access data that isn't included?
  Data catalog is determined by provider, not users

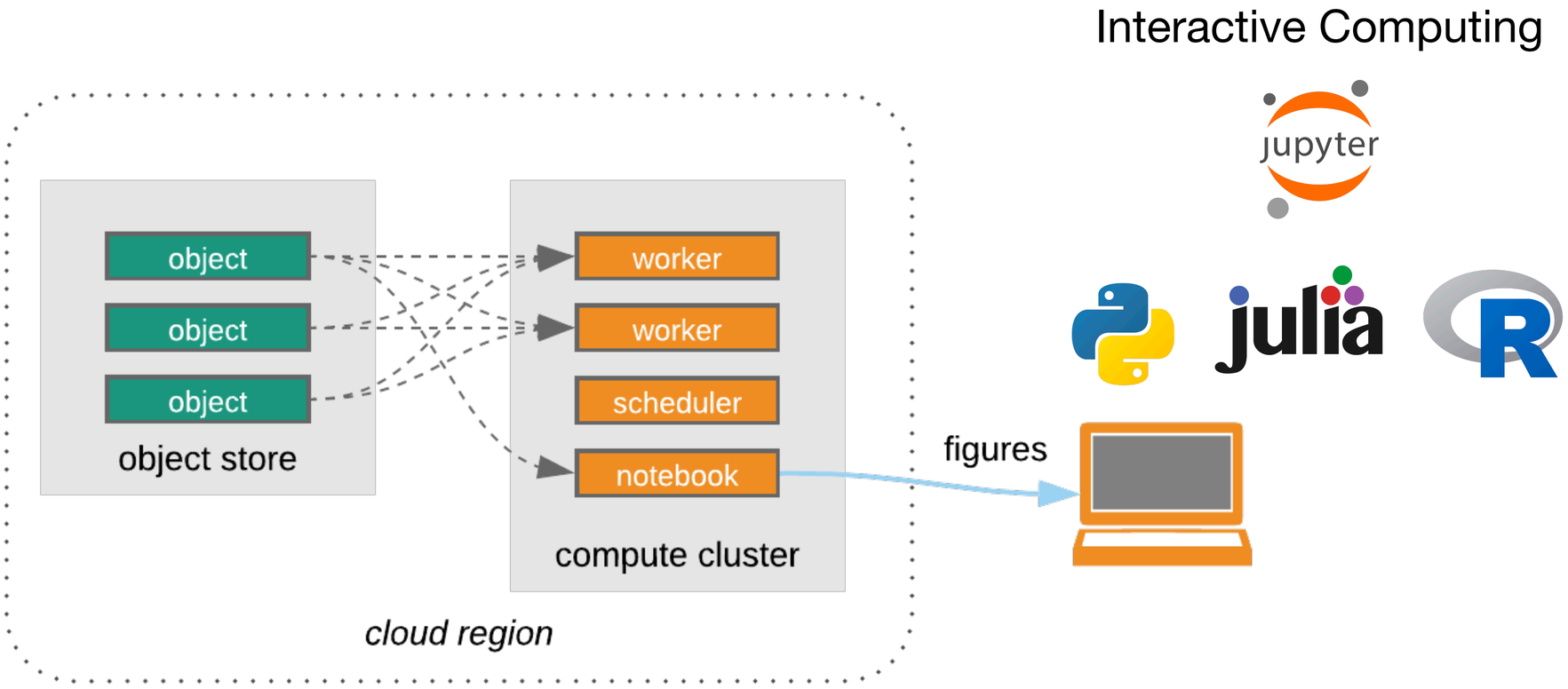
- Platforms are "single instance":
  Fear of lock-in, possibility platform will disappear

- Who pays? Do resource limitations constrain science?

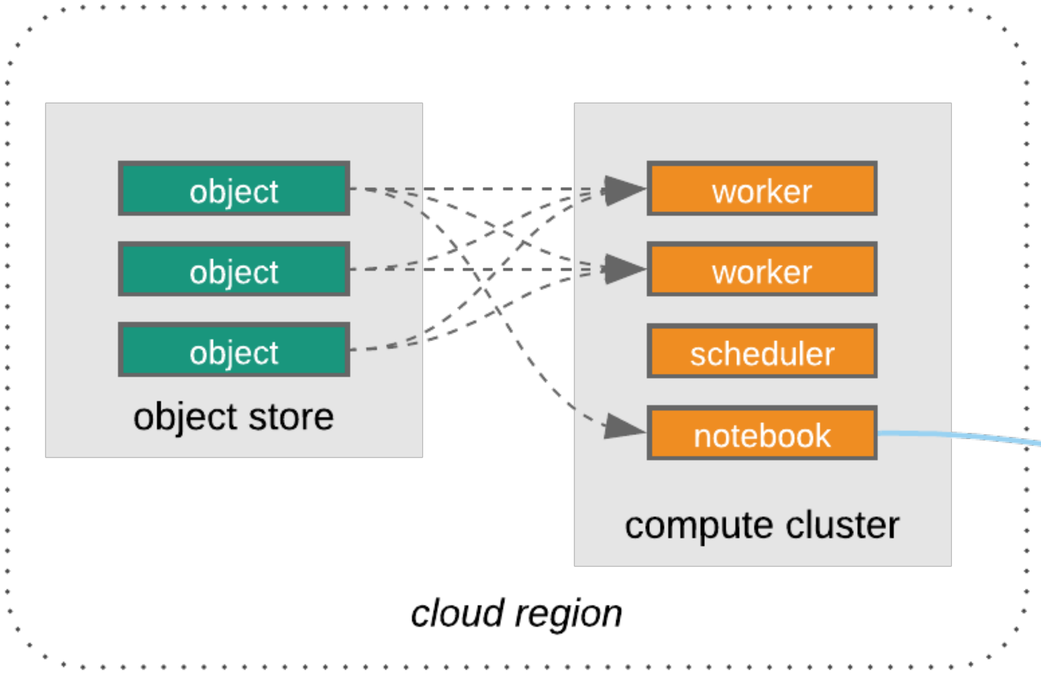
OPEN "CLOUD" ARCHITECTURE



OPEN "CLOUD" ARCHITECTURE



OPEN "CLOUD" ARCHITECTURE



Interactive Computing

jupyter

julia

R

figures

Machine Learning Frameworks

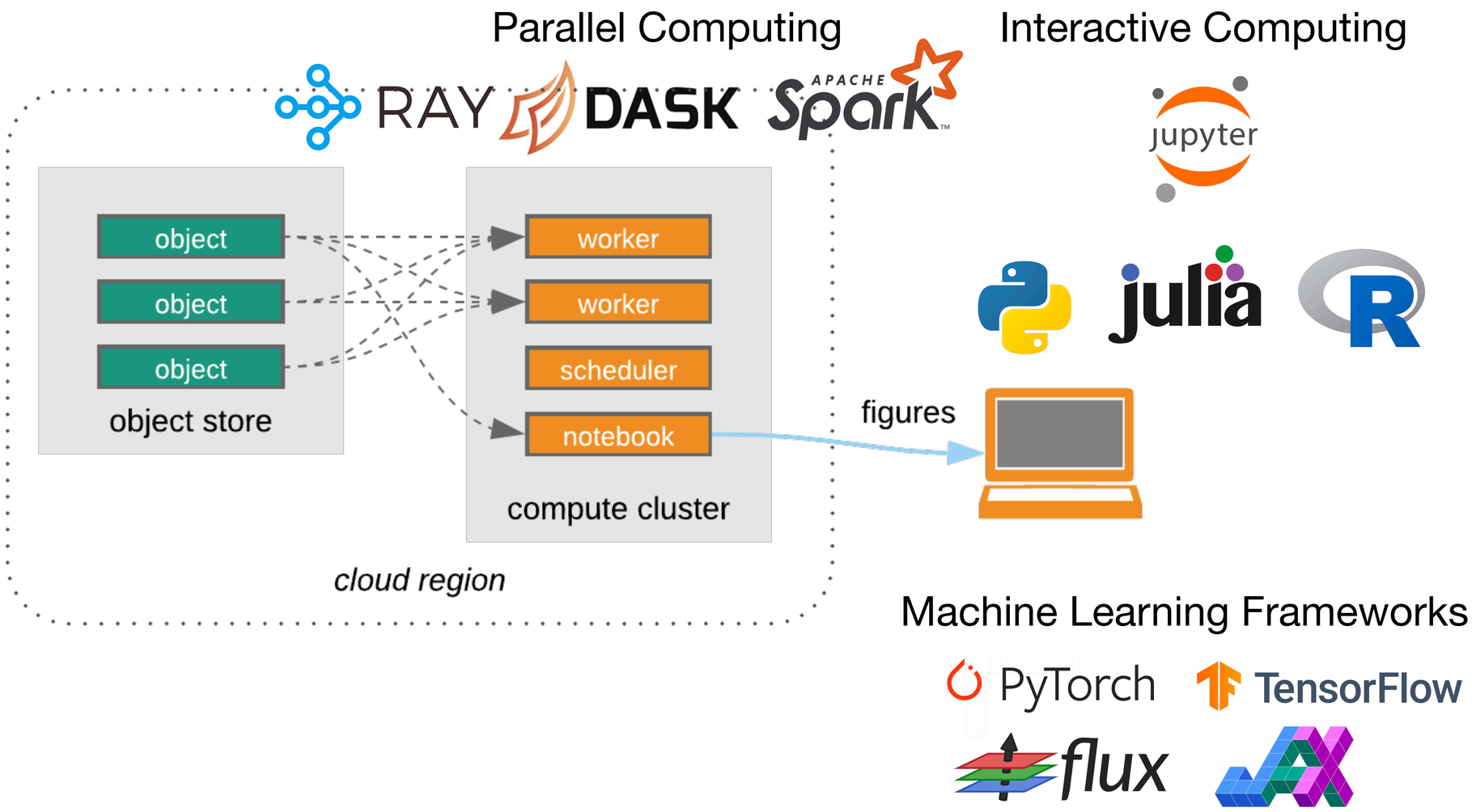
PyTorch

TensorFlow

flux

JAX

OPEN "CLOUD" ARCHITECTURE



OPEN "CLOUD" ARCHITECTURE

Analysis Ready Data
Cloud Optimized Formats

Parquet

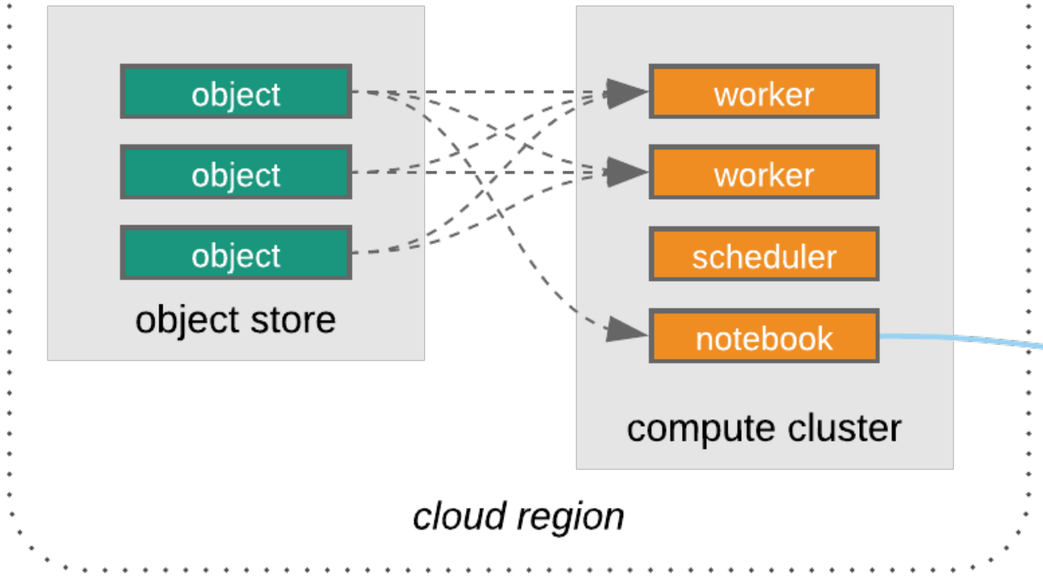
COG
CLOUD OPTIMIZED
GEOTIFF

Zarr

[tile]DB

Parallel Computing

RAY DASK APACHE Spark



Interactive Computing

jupyter

julia

R

figures

Machine Learning Frameworks

PyTorch TensorFlow

flux JAX

PANGEO

OPEN "CLOUD" ARCHITECTURE

Analysis Ready Data
Cloud Optimized Formats

Parquet

COG
CLOUD OPTIMIZED
GEOTIFF

Zarr

[tile]DB

Parallel Computing

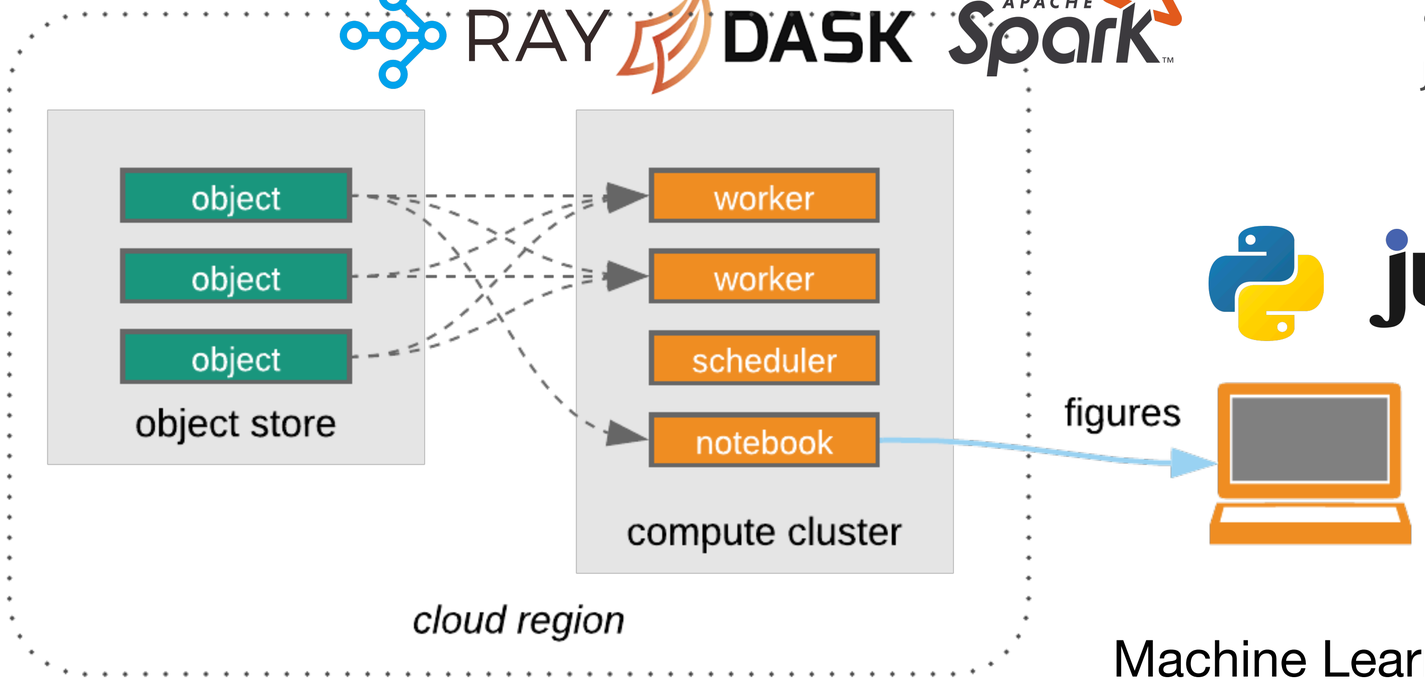
RAY DASK APACHE Spark

Interactive Computing

jupyter

julia

R



Machine Learning Frameworks

PyTorch TensorFlow

flux JAX

Data Provider's $

Google Cloud Platform

aws

Microsoft Azure

Data Consumer's $

WHAT IS PANGEO?

- Community obsessed with efficient data processing.

  Founded in 2017. Scientists and software developers coming together. http://pangeo.io/
  Weekly meeting / seminar. Discourse Forum. Annual meeting. Workshops at AGU / AMS / etc.

- Interoperable Software

  Foundation in Open Source Scientific Python: Jupyter, Xarray, Dask, Zarr. Broad ecosystem of interoperable packages for analysis, visualization, and machine learning.

- Data and Computing Infrastructure

  Deployment recipes for cloud and HPC. Open, public, cloud-based JupyterHubs and Binders for Data-proximate computing. PB of analysis-ready, cloud-optimized data stored in public cloud (GCS, AWS) and OpenStorageNetwork.

BRINGING TOGETHER ACADEMIA, GOV. AGENCIES & INDUSTRY

NSF

EarthCube

NASA

Lamont-Doherty Earth Observatory
Columbia University | Earth Institute

NCAR
National Center for Atmospheric Research

Anaconda
Powered by Continuum Analytics

Met Office

USGS

ECMWF

cnes
Centre National d'Études Spatiales

Coiled

GEOMAR

CSIRO

IRI

BIDS
Berkeley Institute for Data Science

University of Washington
eScience Institute

jupyter

Radiant Earth Foundation
Earth Imagery for Impact

JUPITER

developmentSEED

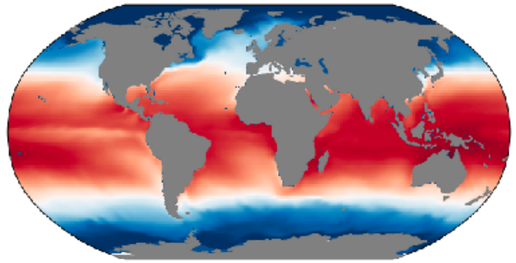
element84

Rhodium Group

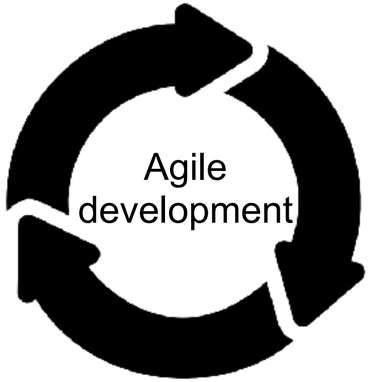
ClimaCell
Weather Revealed

THE PANGEO COMMUNITY PROCESS

Scientific users / use cases



- Define science questions
- Use software / infrastructure
- Identify bugs / bottlenecks
- Provide feedback to developers



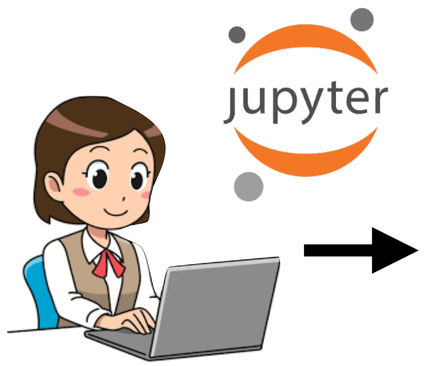
Open-source software libraries

- Contribute widely the the open source scientific python ecosystem
- Maintain / extend existing libraries, start new ones reluctantly
- Solve integration challenges

HPC and cloud infrastructure

- Deploy interactive analysis environments
- Curate analysis-ready datasets
- Platform agnostic

THE PANGEO OPEN-SOURCE STACK



Rich interactive computing environment in the web browser.

Domain specific packages

xgcm | xrft | xhistogram | climpred | gcm-filters | Etc.

xarray — High-level API for analysis of multidimensional labelled arrays.

DASK — Flexible, general-purpose parallel computing framework.

Zarr — Cloud-optimized storage for multidimensional arrays.

Cloud Services

Kubernetes

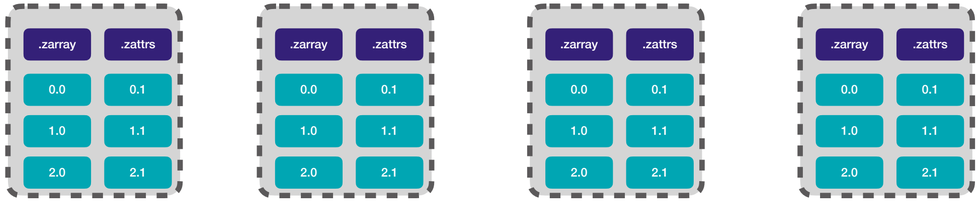
Object Storage

PANGEO-STYLE INFRASTRUCTURE

Data Lakes

http://catalog.pangeo.io

Zarr Datasets



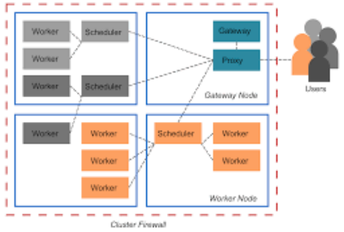
amazon S3

wasabi

open storage network

Compute Services

Dask Gateway



binder

jupyterhub

Node Pools (Autoscaling)

preemptible (spot instance)

normal

kubernetes

PANGEO

ANALYSIS-READY, CLOUD-OPTIMIZED (ARCO) DATA ACCELERATES SCIENCE

ARCO Data

xarray.Dataset

▸ Dimensions: (**latitude**: 720, **longitude**: 1440, **nv**: 2, **time**: 8901)

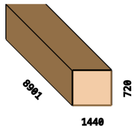
▾ Coordinates:

crs	()	int32	...
lat_bnds	(time, latitude, nv)	float32	dask.array<chunksize=(…
latitude	(latitude)	float32	-89.875 -89.625 ... 89.…
lon_bnds	(longitude, nv)	float32	dask.array<chunksize=(…
longitude	(longitude)	float32	0.125 0.375 ... 359.625…
nv	(nv)	int32	0 1
time	(time)	datetime64[ns]	1993-01-01 ... 2017-05…
axis :	T		
long_name :	Time		
standard_na…	time		

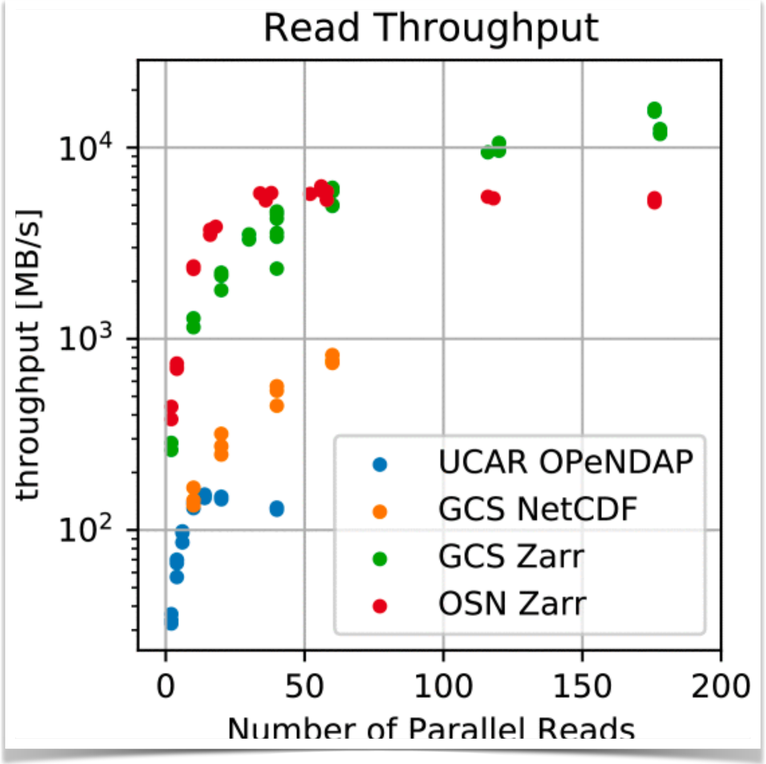
▾ Data variables:

adt	(time, latitude, longitude)	float64	dask.array<chunksize=(…

	Array	Chunk
Bytes	73.83 GB	41.47 MB
Shape	(8901, 720, 1440)	(5, 720, 1440)
Count	1782 Tasks	1781 Chunks
Type	float64	numpy.ndarray

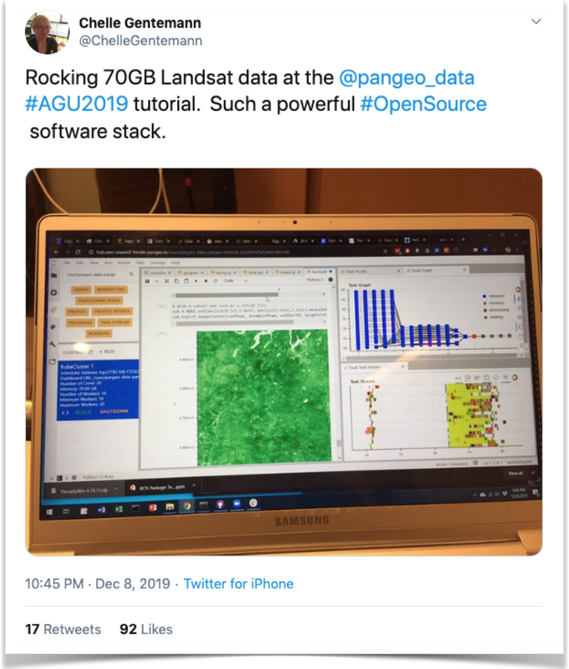


err	(time, latitude, longitude)	float64	dask.array<chunksize=(…
sla	(time, latitude, longitude)	float64	dask.array<chunksize=(…
ugos	(time, latitude, longitude)	float64	dask.array<chunksize=(…
ugosa	(time, latitude, longitude)	float64	dask.array<chunksize=(…
vgos	(time, latitude, longitude)	float64	dask.array<chunksize=(…
vgosa	(time, latitude, longitude)	float64	dask.array<chunksize=(…



- ARCO data production takes time and skill.
- Good ARCO data (COG, Zarr, TileDB, Parquet) + S3 obviates the need for some APIs / services.
- *We need venues for sharing / publishing ARCO data*

SCIENTISTS ❤️ PANGEO



- We didn't build very much new stuff; we just helped existing, community developed tools work together. Open and community-driven from day 1. *Sustainability*
- "Power users" always just want direct, data-proximate access to the raw data. *Simplicity*
- The same stack is an effective base-layer for apps / dashboards / APIs, etc. *Modularity*

AGENCIES USING PANGEO

Recent talks from the Pangeo Showcase Seminar: *https://zenodo.org/communities/pangeo/*

September 15, 2021 (v1) | Presentation | Open Access

View

The NWS Implementation of the OGC API - Environmental Data Retrieval

Mill, Shane; Olson, Steve;

An Environmental Data Retrieval (EDR) API provides a family of lightweight interfaces to access Environmental Data resources. Each resource addressed by an EDR API maps to a defined query pattern. This specification identifies resources, captures compliance classes, and specifies requirements which

Uploaded on September 29, 2021

AGENCIES USING PANGEO

Recent talks from the Pangeo Showcase Seminar: *https://zenodo.org/communities/pangeo/*

September 15, 2021 (v1) | Presentation | Open Access

The NWS Implementation of the OGC API - Environmental Data Retrieval

Mill, Shane; Olson, Steve;

An Envir
resource
and spec

Uploaded

View

April 29, 2021 (v1) | Other | Open Access

FAIR Climate Services using the Xarray ecosystem and OGC Standards

Huard, David;

Talk given at the Pangeo Showcase seminar series.

Uploaded on April 29, 2021

View

Environment and Climate Change Canada
Environnement et Changement climatique Canada

AGENCIES USING PANGEO

Recent talks from the Pangeo Showcase Seminar: *https://zenodo.org/communities/pangeo/*

September 15, 2021 (v1) | Presentation | Open Access

The NWS Implementation of the OGC API - Environmental Data Retrieval

Mill, Shane; Olson, Steve;

An Envir
resource
and spec

Uploaded

View

April 29, 2021 (v1) | Other | Open Access

FAIR Climate Services using the Xarray ecosystem and OGC Standards

Huard, Davi

Talk given at t

Uploaded on Apr

View

October 20, 2021 (v1) | Presentation | Open Access

Scalable Sustainability with the Planetary Computer

Augspurger, Tom; Morris, Dan; Emanuele, Rob; McFarland, Matt; Sanchez-Andrade Nuno, Bruno;

We'll see how Microsoft's Planetary Computer team has embraced the ideas and techniques pioneered by pangeo, putting them to work for sustainability. We'll see the use of STAC for cataloging and querying data, analysis-ready data in blob storage, and JupyterHub-based environments for sca

Uploaded on October 25, 2021

View

PANGEO

AGENCIES USING PANGEO

Recent talks from the Pangeo Showcase Seminar: *https://zenodo.org/communities/pangeo/*

September 15, 2021 (v1) | Presentation | Open Access

The NWS Implementation of the OGC API - Environmental Data Retrieval

Mill, Shane; Olson, Steve;

An Enviro
resource
and spec

Uploaded

View

April 29, 2021 (v1) | Other | Open Access

FAIR Climate Services using the Xarray ecosystem and OGC Standards

Huard, Davi

Talk given at t

Uploaded on Apr

View

October 20, 2021 (v1) | Presentation | Open Access

Scalable Sustainability with the Planetary Computer

Augspurger, Tom;

We'll see how Micro
sustainability. We'll s
environments for sc

Uploaded on October 25

Microsoft

View

November 3, 2021 (v1) | Presentation | Open Access

xcube - Python package for Earth Observation data cubes

Brandt, Gunnar; Norman Fomferra;

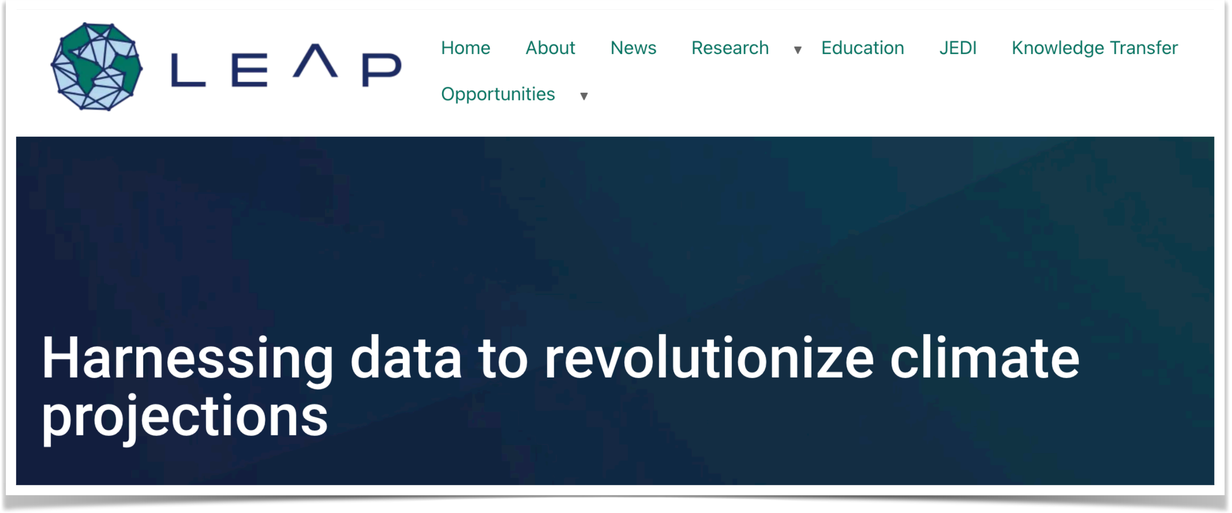
xcube is a software package and toolkit for generation and exploitation of analysis-ready data cubes from Earth Observation and other geographical data. It is based on Python's popular data science technology stack, particularly on xarray, dask, and zarr, and is freely available as open-source s

Uploaded on November 11, 2021

esa
European Space Agency

View

PANGEO EMPOWERS AI/ML RESEARCH



- Software for data processing, exploration, visualization
- ARCO data lakes for building training datasets
- Infrastructure for collaborative, data-proximate analytics and model development / training

Limitation: Pangeo today is all about *offline* data analytics.
It does not address the needs of *online* machine learning and direct integration with ESMs.

🤔

SUMMARY

- AI/ML don't work without good data. Exploratory analysis and visualization are prerequisites to more advanced methods. *We need to support this!*
- Pangeo provides open-source building blocks for data-intensive science; use these however makes sense for MAELSTROM.
- Let's build a federation of interoperable data and computing services on top of generic cloud-style infrastructure.

Thanks to our funders!

NSF

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

NASA

GORDON AND BETTY MOORE FOUNDATION

LEARN MORE

http://pangeo.io

https://discourse.pangeo.io/

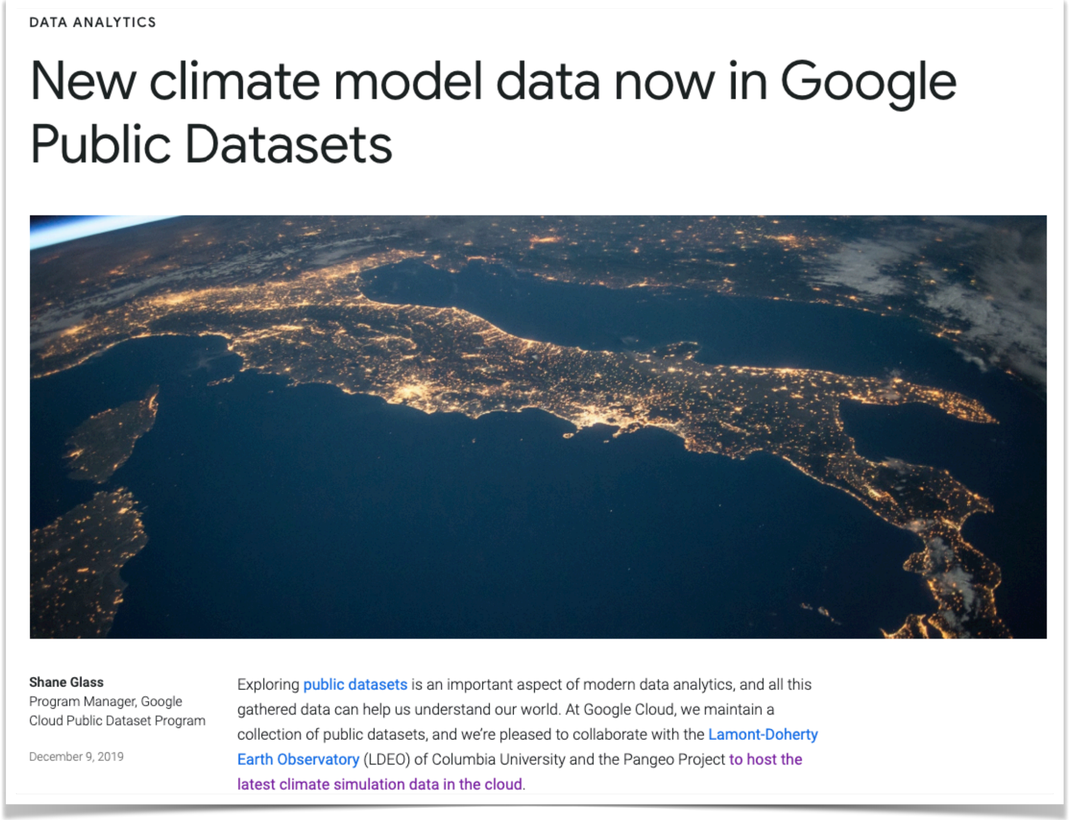
https://github.com/pangeo-data/

https://medium.com/pangeo

@pangeo_data

Extra Slides

CMIP6 CLOUD DATASET



DATA ANALYTICS

New climate model data now in Google Public Datasets

Shane Glass
Program Manager, Google Cloud Public Dataset Program

December 9, 2019

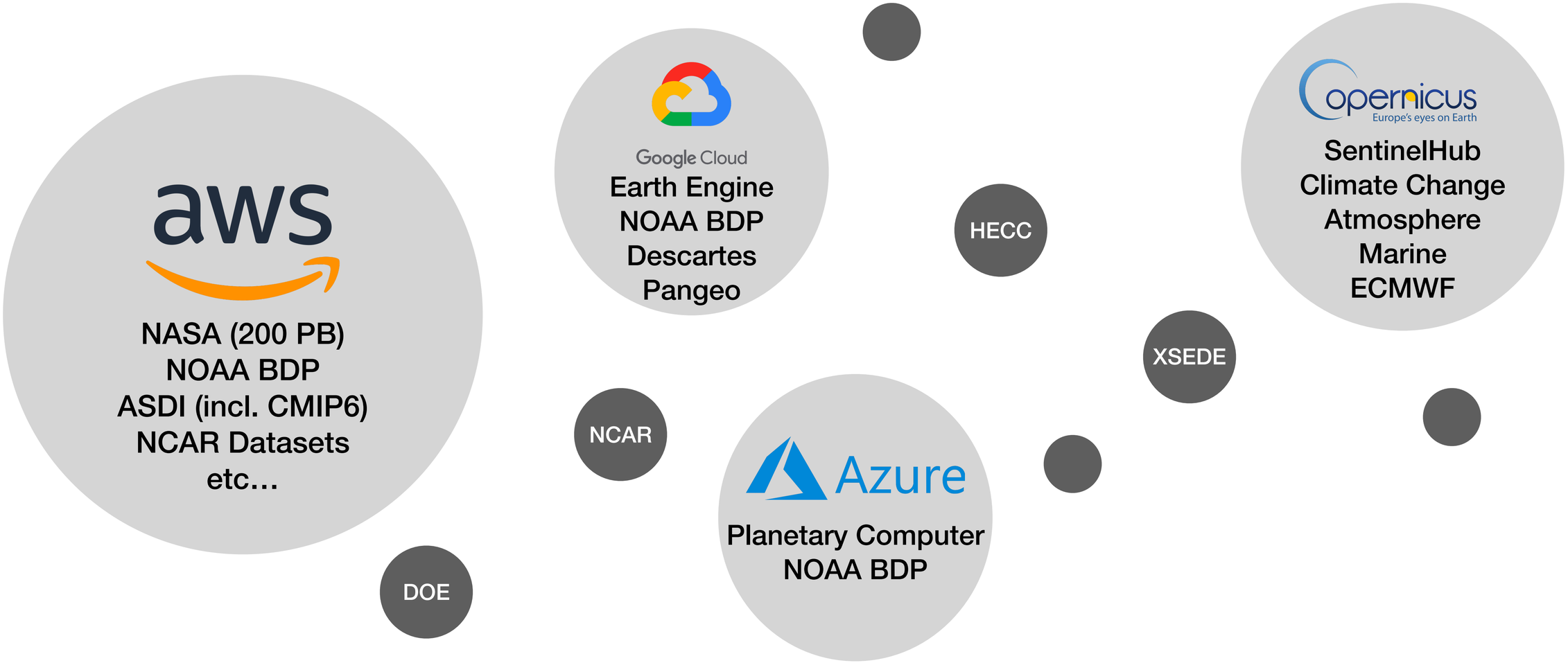
Exploring public datasets is an important aspect of modern data analytics, and all this gathered data can help us understand our world. At Google Cloud, we maintain a collection of public datasets, and we're pleased to collaborate with the Lamont-Doherty Earth Observatory (LDEO) of Columbia University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

- Pangeo partnered with ESGF and Google Cloud to provide a new public dataset
- > 1 PB and counting
- Data stored in Zarr format
- Google provides free hosting in GCS
- Mirrored on AWS

https://pangeo-data.github.io/pangeo-cmip6-cloud/

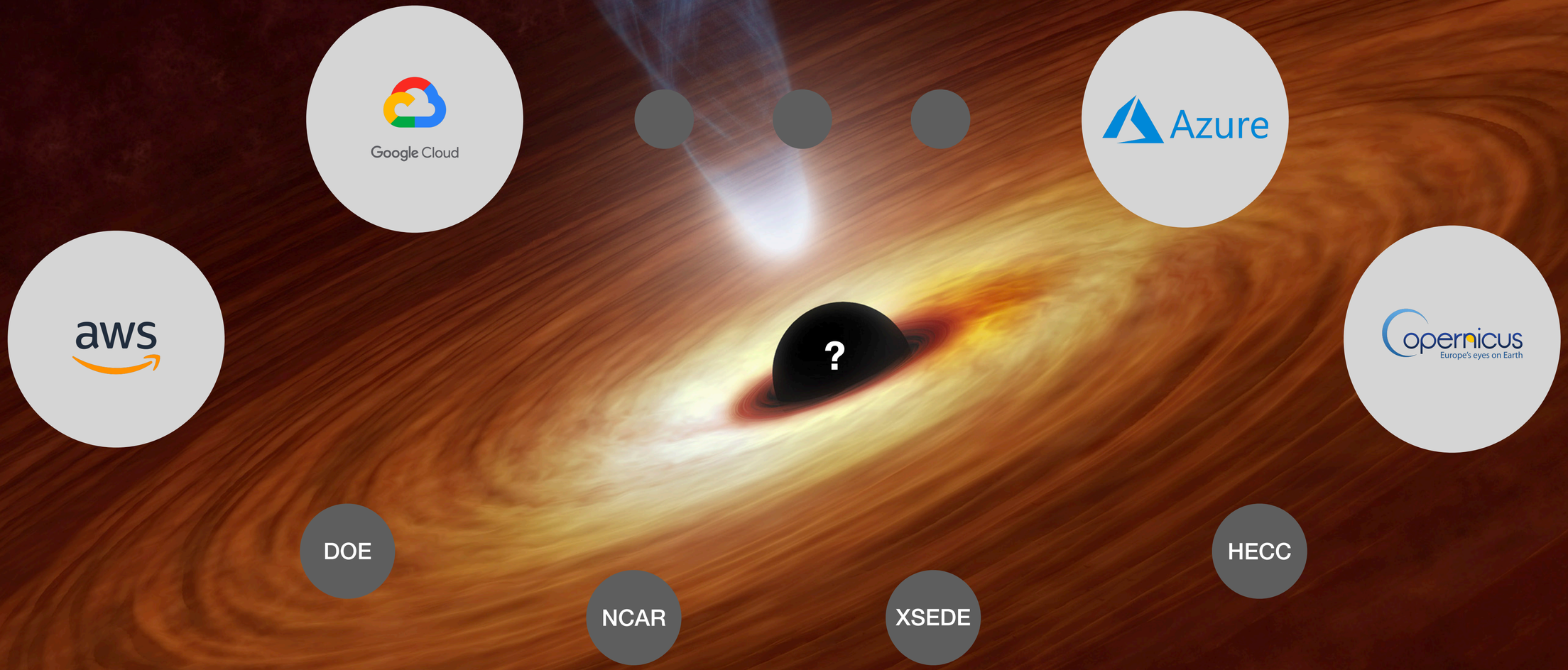
FUTURE CHALLENGE: DATA GRAVITY

“Data gravity is the ability of a body of data to attract applications, services and other data." - Dave McCrory



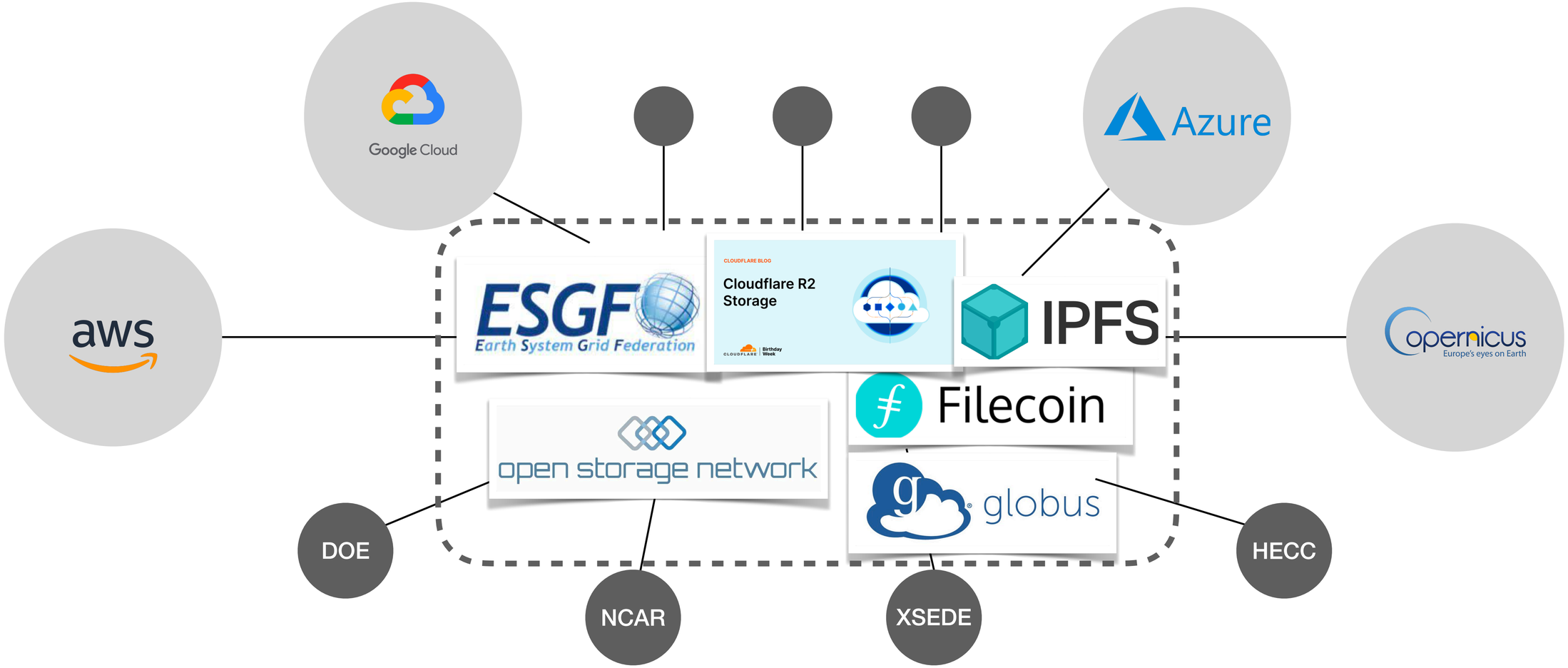
DATA GRAVITY

What is the stable steady-state solution?

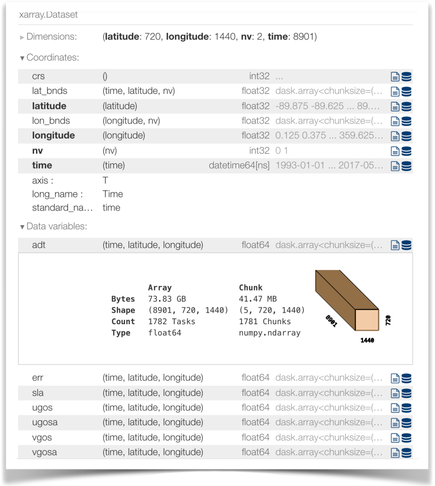


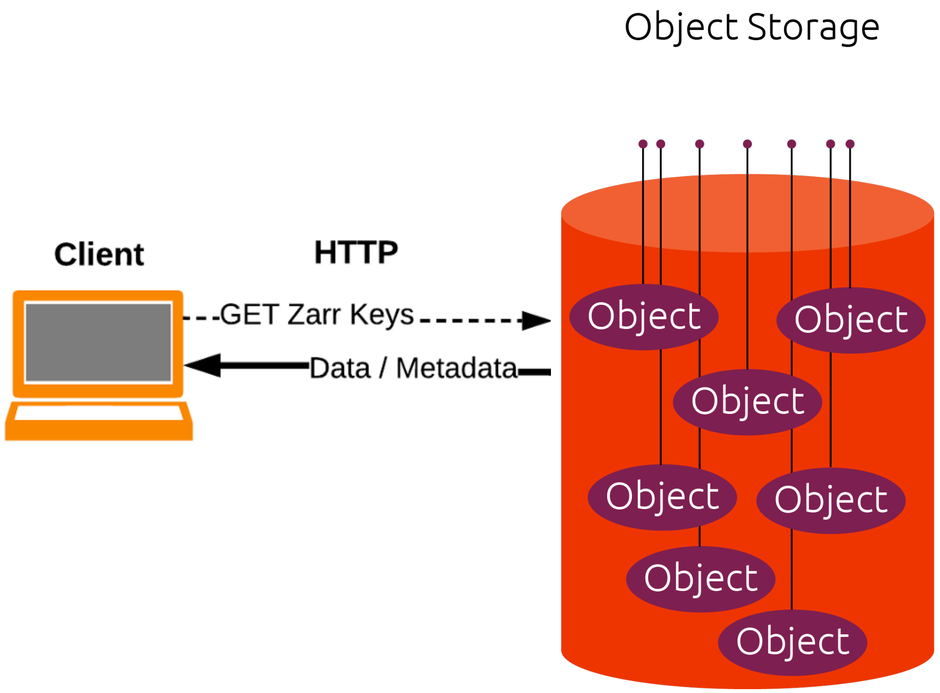
WE NEED A GLOBAL SCIENTIFIC DATA COMMONS

Edge storage, decentralized web, web3

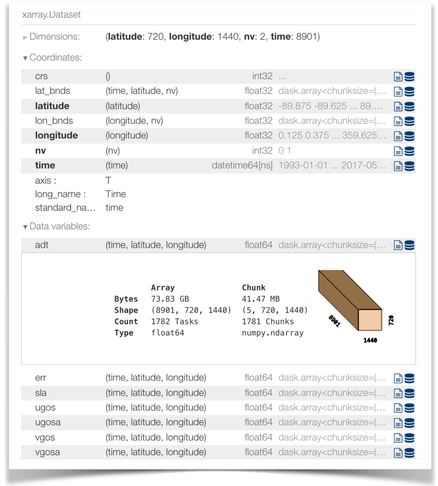


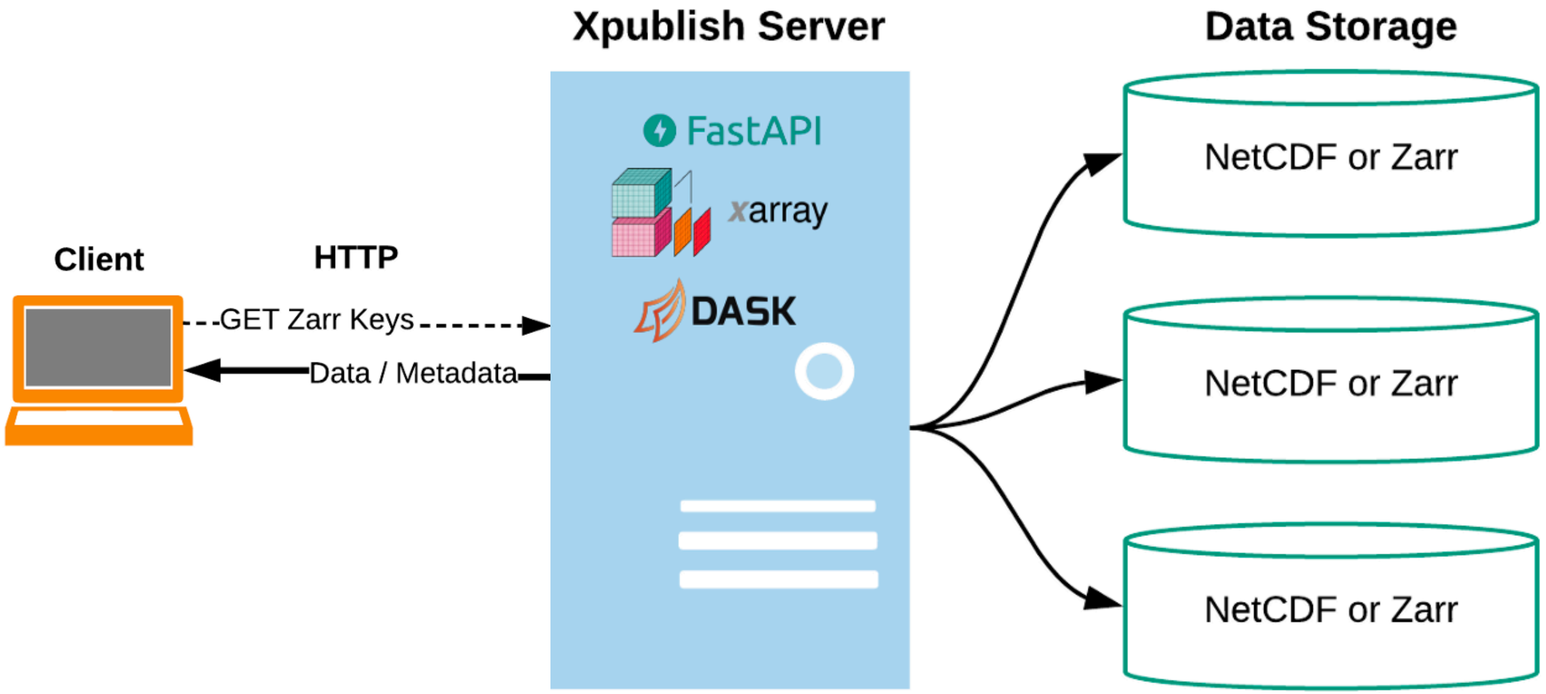
ARCO DATA + HTTP (S3) IS MORE PERFORMANT AND FLEXIBLE THAN A BESPOKE API





ARCO DATA + HTTP (S3) IS MORE PERFORMANT AND FLEXIBLE THAN A BESPOKE API





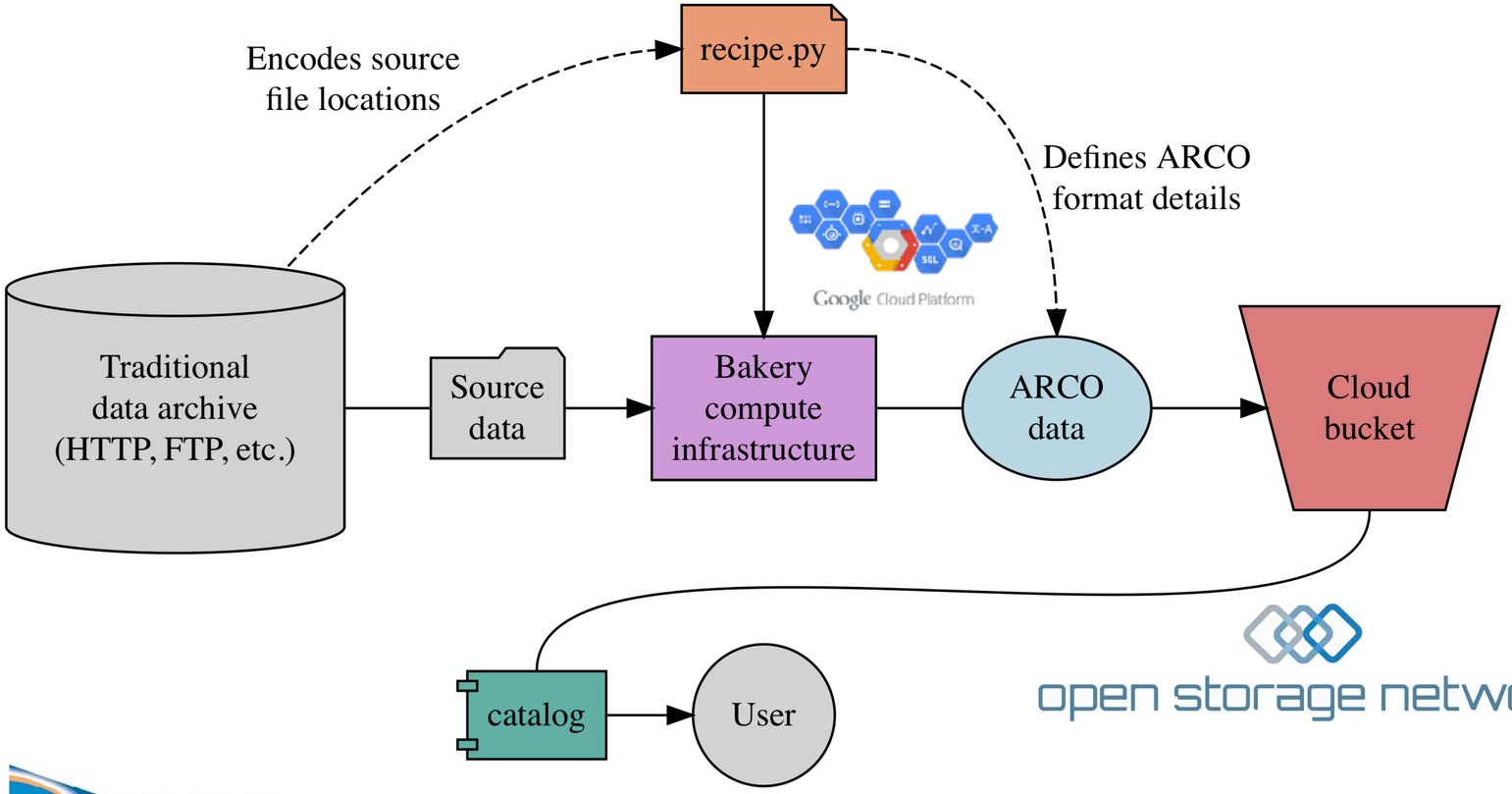
https://xpublish.readthedocs.io/
Serve dynamically generated Zarr data over HTTP.
Client can't tell the difference.

PANGEO FORGE: DEMOCRATIZING ARGO DATA PRODUCTION

https://pangeo-forge.org/

An open source platform for creating ARCO datasets.

We crowdsource “recipes” for ARCO data from the global science community. Cloud automation builds the datasets in a scalable and reproducible way.



open storage network

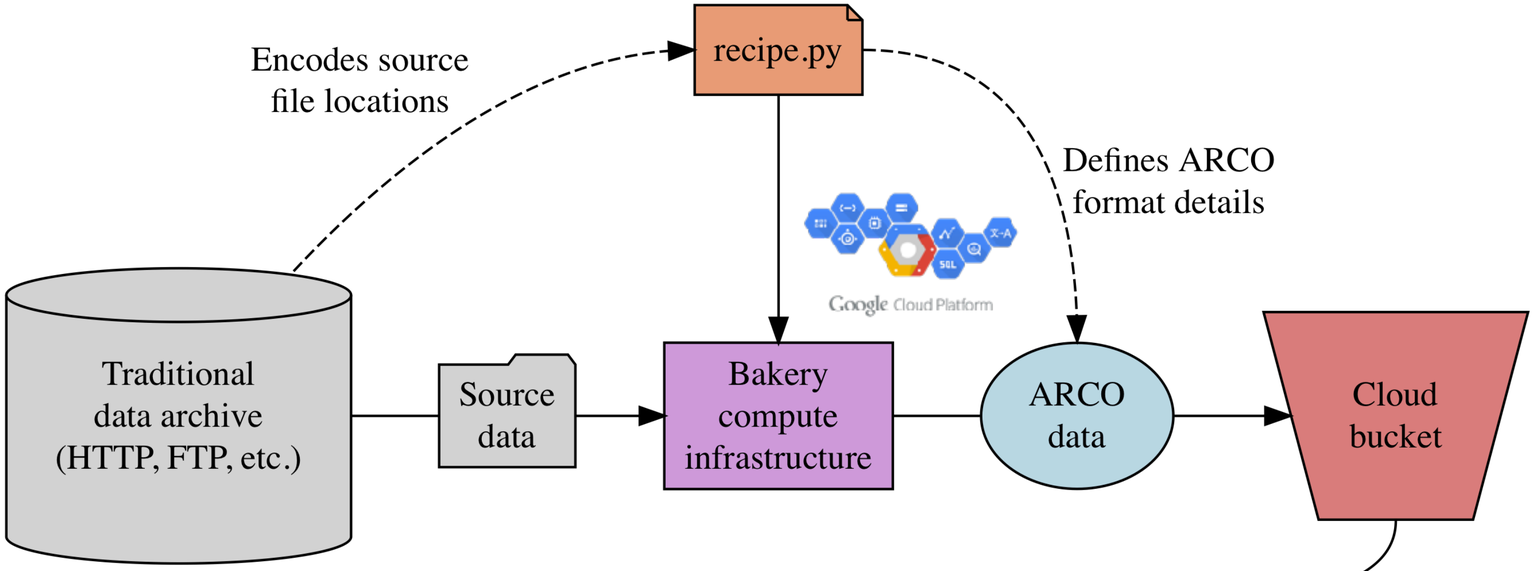
NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

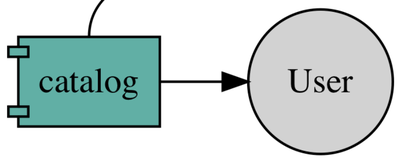
PANGEO FORGE: DEMOCRATIZING ARGO DATA PRODUCTION

https://pangeo-forge.org/

An open source platform for creating ARCO datasets.

We crowdsource “recipes” for ARCO data from the global science community. Cloud automation builds the datasets in a scalable and reproducible way.

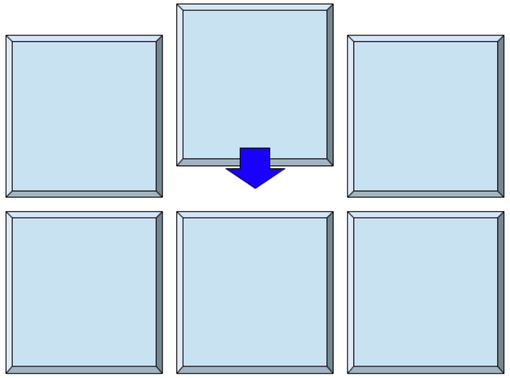


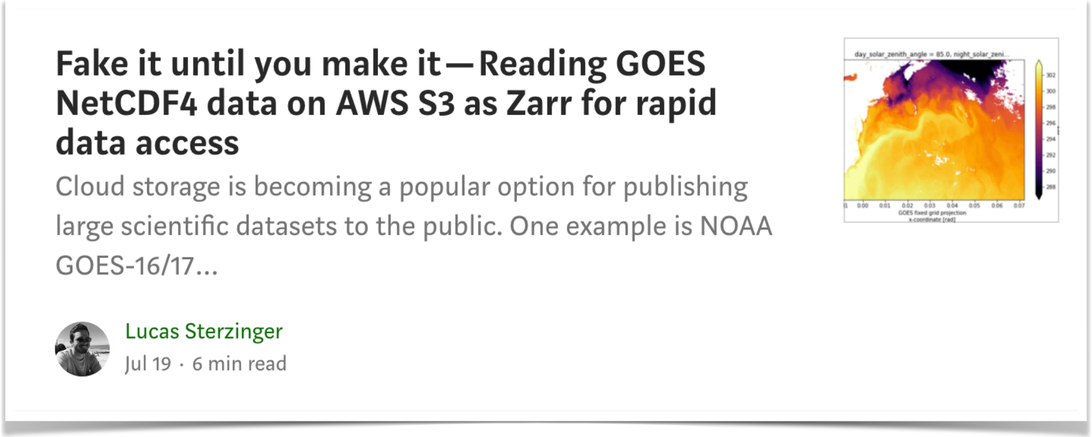


open storage network

KERCHUNK: MAKE YOUR LEGACY DATA LOOK AND FEEL LIKE ZARR

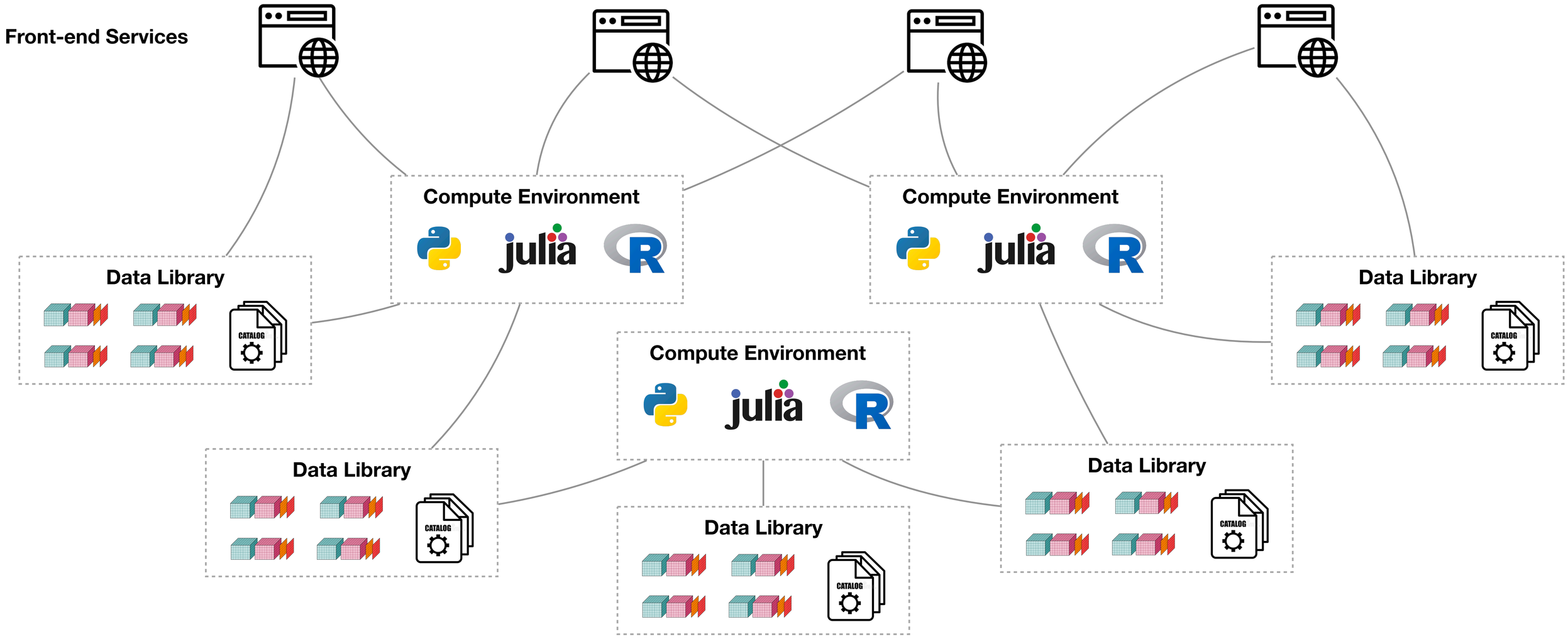
https://fsspec.github.io/kerchunk/



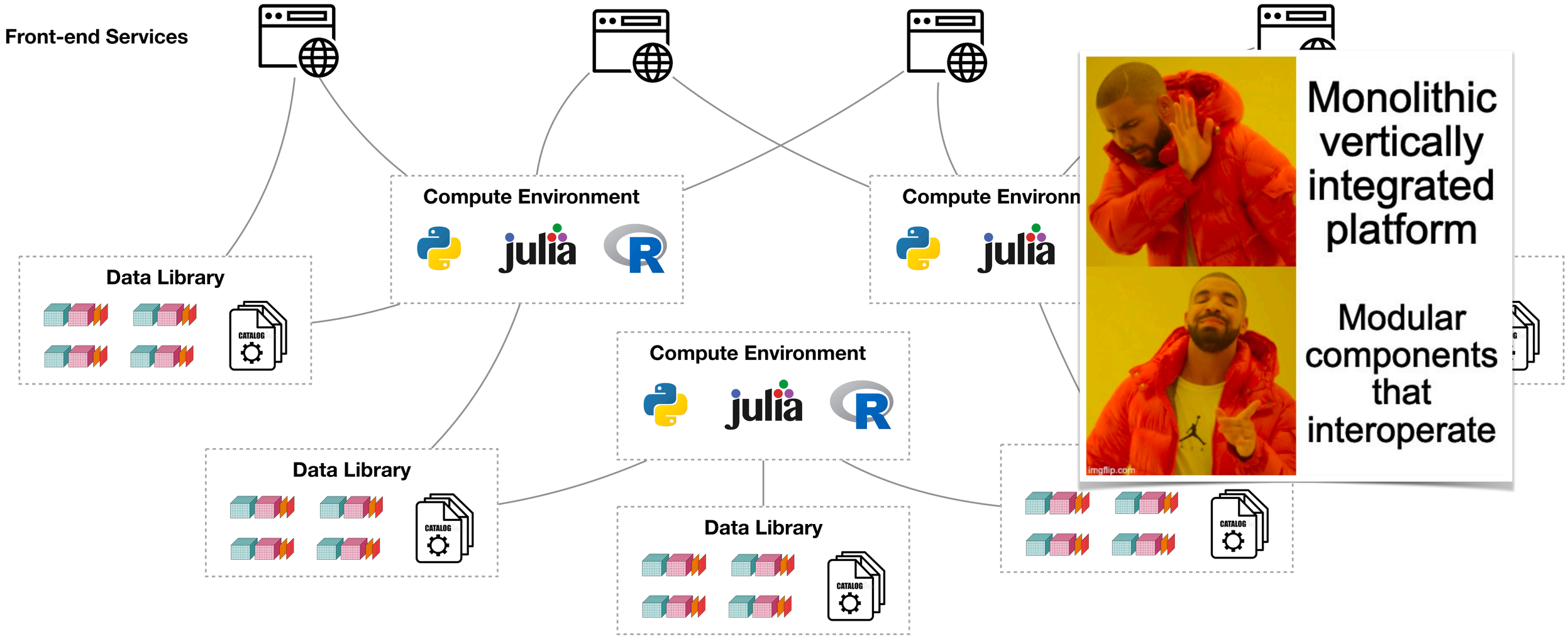


- Provides a unified way to represent a variety of chunked, compressed binary data formats (e.g. NetCDF/HDF5, GRIB2, TIFF, …)
- Allows efficient access to data from traditional file systems or cloud object storage.
- Create virtual datasets from multiple files by extracting the byte ranges, compression information etc. and storing this metadata in a new, separate object.
- Open Spec, python implementation.

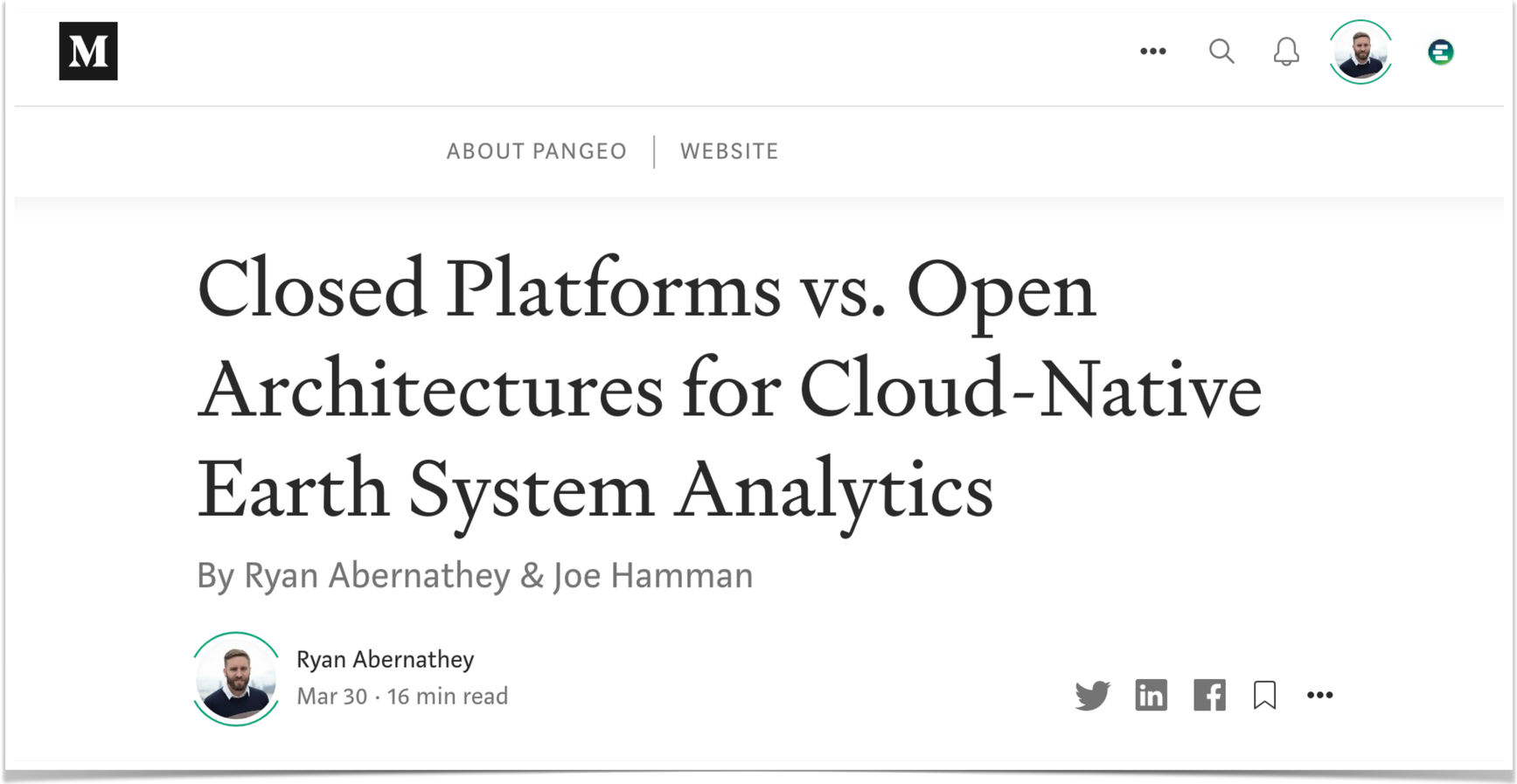
Vision: an extensible, interconnected federation of high-performance ESM data analytics environments



Vision: an extensible, interconnected federation of high-performance ESM data analytics environments



PANGEO



https://medium.com/pangeo