NS15C-0386

Pangeo and Project Pythia: Helping Geoscientists Navigate the Scientific Python Ecosystem

Authors

Kevin Paul, Anderson Banihirwe, John Clyne, Orhan Eroglu, Maxwell Grover, Julia Kent
National Center for Atmospheric Research

Drew Camron, Ryan May
Unidata

Brian E J Rose, Kevin Tyle, Robert Ford
University at Albany (SUNY)

Abstract

Pangeo is a community of academic scientists and industry software engineers partnered to "foster collaboration around the open source scientific Python ecosystem for ocean, atmosphere, land, and climate science." While Pangeo is probably best known for its scalable software stack built on Jupyter, Xarray, and Dask, the Pangeo community's (arguably) greatest contribution is its support for and advocacy of open development practices and reproducible science. To aid in Pangeo's efforts to educate scientists and advance open science, Project Pythia was formed to create a "go-to" resource to help geoscientists navigate and learn the complex scientific Python ecosystem and to help train scientists in the best practices necessary to foster open science and open development. One of Project Pythia's most recent efforts is to build a Foundations Book, covering the basic technology and knowledge necessary to wade into the complex scientific Python ecosystem. In this presentation, I will introduce the audience to Pangeo, its software stack, and the current state of Project Pythia. I will attempt to cover how Pangeo can be useful to other communities, as well as how you can get involved.

Funding for Pangeo and Project Pythia both come from NSF EarthCube awards 1740633, 1740648, 2026863, and 2026899.

Poster Lightning Talk

- Hello! My name is Kevin Paul, and I am one of the many members of the Pangeo community, and I am also a member of Project Pythia.
- Pangeo is an open community of both scientists and software engineers, spanning academia and industry, working together to develop a scalable open source solution for reproducible science.

- Pangeo started in the oceanic and atmospheric modeling communities, but the technology growing from Pangeo is widely useful for many scientific fields.
- Pangeo's scalable software architecture depends on the interoperability of a number of different packages, all designed to fit into the PyData ecosystem. These include, but are not solely limited to, Xarray, Dask and Jupyter.
    - Xarray provides an interface for dealing with "labelled multidimensional data." That is, if your data can be represented by multidimensional arrays, such as NumPy NDArrays, then Xarray gives you the ability to access array values with coordinates instead of just array indices.
    - Additionally, Xarray provides an interface to your data that looks and feels like the Pandas interface to a DataFrame. Common Pandas-like operations like Groupby can be used to make operations like the calculation of seasonal averages, or moving-window averages, simple one-liners.
    - Think of Xarray as a mash-up of Pandas and NumPy! If you have unstructured data (like a table of values) then use Pandas DataFrames. But if you have structured data (like multidimensional arrays with coordinates along those array axes), then use Xarray.
    - Xarray was developed to make it possible to easily replace the NumPy NDarrays used "under the hood" with any NumPy "drop-in replacement." That means Xarray was easily modified to use distributed Dask Arrays (out of memory arrays) and not just NumPy arrays!
    - While it's never true to say that anything provides distributed parallelism "without any effort," Dask makes parallelizing your Xarray workflows much easier and much more flexible than something like MPI. (And if you've ever used Spark, Spark is great for unstructured data like DataFrames, but it struggles with structured data like arrays.)
    - Finally, Jupyter provides a means of sharing data analyses and analysis workflows in a reproducible way. Jupyter Notebooks can be developed to encode a sharable scientific analysis, like an "executable paper." And Jupyter Lab (together with Jupyter Hub) can be used to serve those Notebooks to others on hardware capable of running the analysis itself.
    - All of these technologies are designed to work together, in both HPC or cloud environments.
- As I mentioned, however, the Pangeo software ecosystem is not limited solely to Xarray, Dask and Jupyter. There are a growing number of third-party packages that provide unique and powerful capabilities to the core Pangeo technologies. And navigating these numerous Python packages can be daunting. For this reason, we started Project Pythia.
- Project Pythia's goal is to create a "go-to" resource for visitors to go from knowing nothing about Python to being an expert in the Pangeo stack...AND going from being a scientist who "just writes scripts" to a scientist who contributes to the open source Python community! Thus, Project Pythia is focused on not just teaching best practices around how to use the specific technologies, but how to contribute to those technologies and become better software developers, themselves.

- If you have time, please take a moment to look through the poster for this session. Click on the links contained in the poster, and explore the different components of Pangeo, the Pangeo software stack, and Project Pythia. If you would like to learn more, please reach out to me through the AGU iPoster platform, or to Pangeo community via Discourse or Twitter, or to folks behind Project Pythia, on Discourse or Twitter or via the provided email. We hope to hear from you and we hope to get your feedback on our work!

Poster Sections

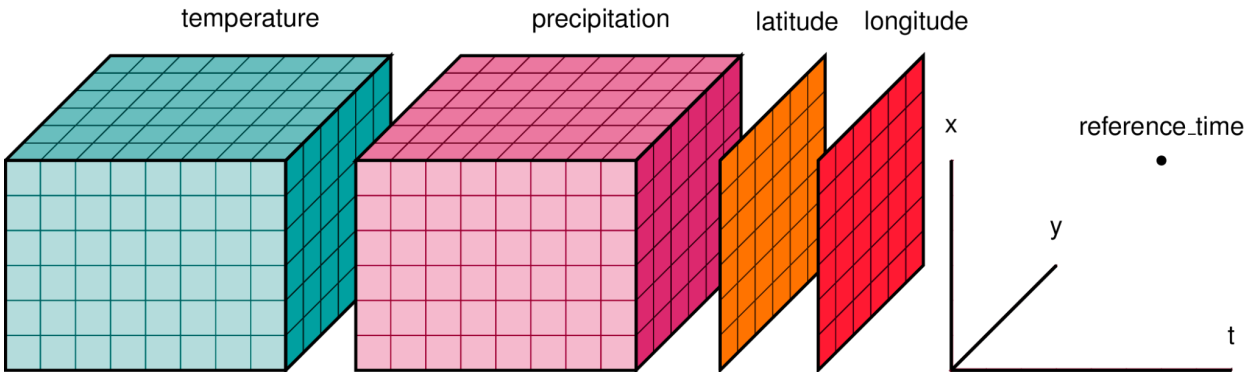
Pangeo

Pangeo (https://pangeo.io) is a community of scientists and software engineers, in academia and industry, working together to develop scalable open source software solutions that enable open, reproducible, and scalable science. Pangeo encourages the development of an ecosystem of software solutions that are interoperable, based on the PyData scientific software stack. Pangeo fosters collaboration around these software solutions and attempts to improve the scalability of these solutions on HPC as well as cloud systems.

The Pangeo software ecosystem is centered around **Xarray**, **Dask**, and **Jupyter**.

Xarray

Xarray (https://xarray.pydata.org) is an open source project and Python package that combines the ease with which Pandas deals with labeled data with the power of NumPy for multidimensional arrays.

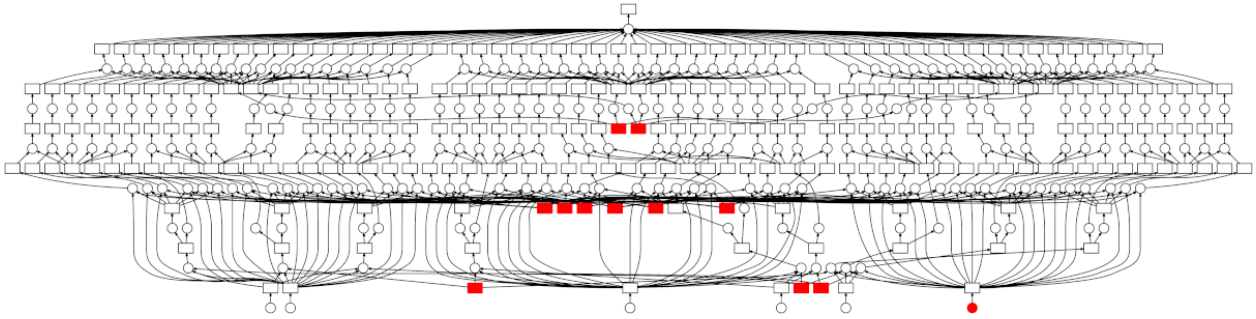


Labelling geoscientific data makes it possible for users to request values from multidimensional arrays using geospatial coordinates, rather than just array indices. This simple feature vastly improves the ease with which users interact with the data as well as the readability of their analysis code. Combined with some of the most powerful features of Pandas, Xarray can reduce complicated operations on gridded geospatial data to simple one-liners! Xarray uses

Dask “under the hood,” making it possible to perform these complex operations in parallel with no added complexity passed to the user.

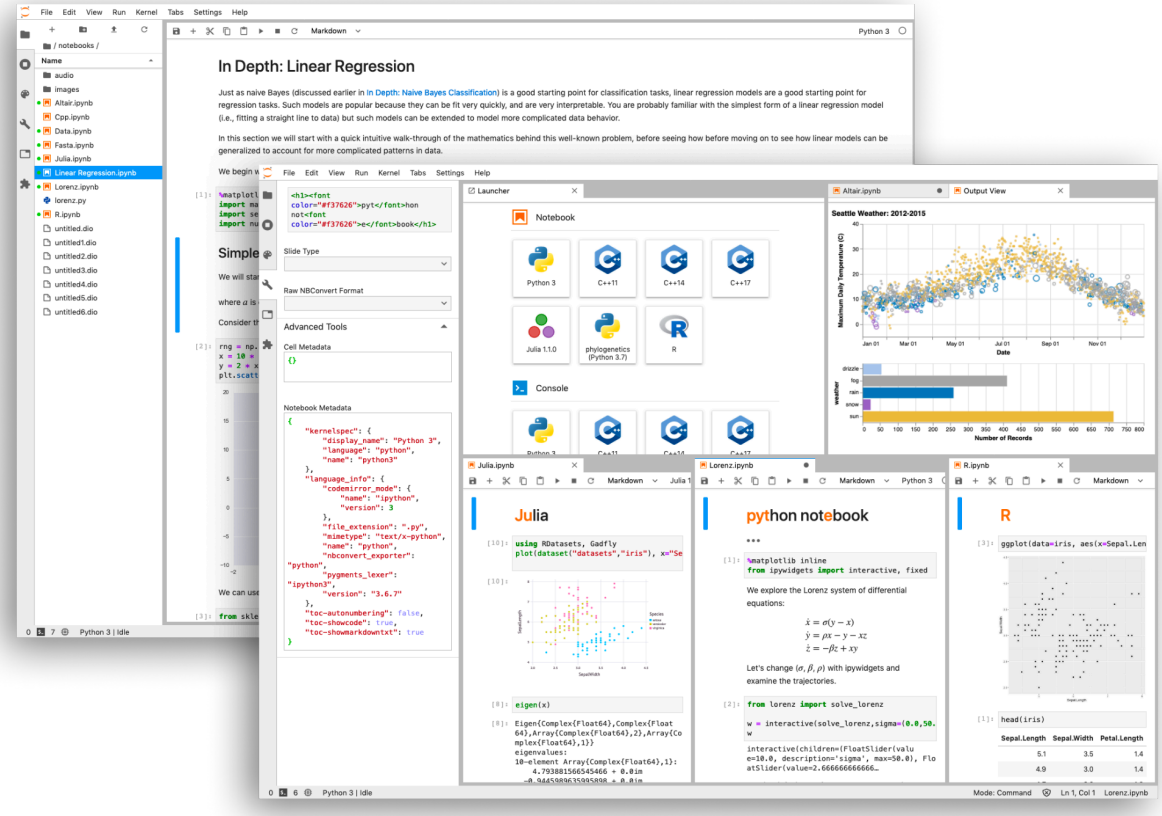
Dask

Dask (https://dask.org) is an open source Python package that provides advanced parallelism for tools in the PyData stack, including NumPy, Pandas, and scikit-learn. Dask makes parallelizing operations on large, multidimensional arrays or large DataFrames easy. Unlike MPI, Dask does the job of deciding how to parallelize large operations for you, rather than requiring that the user implement the parallel algorithm inline with the code.



Jupyter

Project Jupyter (https://jupyter.org) is responsible for a suite of open standards and open source software solutions that make reproducible data science possible. Jupyter Notebooks are an open document format based on JSON that combines narrative with executable code.



Jupyter Notebooks have become the de facto standard in the scientific community for writing and sharing data analysis. Jupyter Lab is a web-based server application that lets users edit and execute Jupyter Notebooks natively in their web browsers. Jupyter Hub is a web-based server application that allows users to authenticate and launch instances of Jupyter Lab on remote systems. With Jupyter Books, you can combine multiple Jupyter Notebooks (or markdown documents, or many other formats) together into publication-quality websites with interactive content.

Project Pythia

If navigating all of these tools and technologies sounds daunting, don't worry! That's why Project Pythia was created. For many who are new to the Python world, the open source ecosystem of packages is a new thing. It's hard to navigate, and it's hard to know where to even start! Project Pythia is an NSF EarthCube project to create a go-to resource for geoscientists to learn about how to use and how to navigate the scientific Python ecosystem. Project Pythia is not just about the technology, though, it is also about teaching scientists how to give back to the open source community.

PROJECT PYTHIA

Project Pythia has built four components that geoscientists can use to learn and navigate the scientific Python ecosystem:

- **The Pythia Portal** (https://projectpythia.org), a centralized location for everything related to Project Pythia and jumping off point to all of Project Pythia's educational content and resources,
- **The Foundations Book** (https://foundations.projectpythia.org), a Jupyter Book covering the foundational content needed to get started in the Pangeo ecosystem of packages and technologies,
- **The Resource Gallery** (https://projectpythia.org/gallery), a filterable list of tutorial videos, examples galleries, Jupyter Books, and more that geoscientists can use to learn, and
- **The Platform**, on which Project Pythia educational resources can run, providing students with an interactive environment within which they can learn by doing.

For more information about Project Pythia, see "Project Pythia: A Resource to Help Geoscientists Navigate the Scientific Python Ecosystem" (IN55C-05) here at AGU!

Get involved!

Both Pangeo and Project Pythia need your help!

You can reach out and engage with members of the Pangeo community, visit the Pangeo Discourse Forum (https://discourse.pangeo.io/) or engage directly with Pangeo developers on GitHub (https://github.com/pangeo-data/). You can also stay up to date with Pangeo projects and developments by following Pangeo's Twitter feed (https://twitter.com/pangeo_data) or Pangeo's Medium blog (https://medium.com/pangeo).

Project Pythia is also looking for your help on how to learn and navigate the growing number of scientific Python packages that are of use to the geoscience community. Reach out to the entire Project Pythia team and open source community via the projectpythia@ucar.edu email address, or contact us on GitHub via the Portal Discussion Board (https://github.com/ProjectPythia/projectpythia.github.io/discussions) or the Foundations Book Discussion Board (https://github.com/ProjectPythia/pythia-foundations/discussions). Or join us weekly at our public meetings every Thursday at 11:00AM Mountain (1:00PM Eastern) on Zoom! (See the Pythia Portal for more details.)