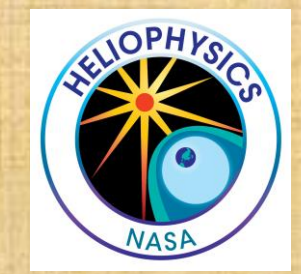
The PyHC Open Science Experiment: A PyHC session led by Rebecca Ringuette

Rebecca Ringuette[1]

[1]Affiliation not available

August 4, 2023

Ringuette, R., Niehof, J. T., Polson, S. A., Zheng, Y., Rastaetter, L., Antunes, A., ... Drozdov, A. (2023, August 3). Comparing Magnetopause Crossings using Open Science. https://doi.org/10.17605/OSF.IO/V4DRT



The PyHC Open Science Experiment

Rebecca Ringuette and many others

NASA GSFC Center for HelioAnalytics

Presented at the 2023 PyHC Spring Meeting at LASP/CU Boulder in Boulder, CO.

Session Outline

- Invited presentation on the **Open Science Framework** (10-15 min)
  Presented by Gretchen Geugeun

- **Project Introduction** (15-20 min)
  - The PyHC Open Science Experiment
  - Project Tour

- **Open Science and Heliophysics Infrastructure** (5-10 min)

- **Discussion** (45 min): What PyHC software changes are needed to better support this project and, more generally, open science? What funding is needed to complete these tasks?

The Open Science Framework

Presented by Gretchen Geugeun

Link to slides:
https://docs.google.com/presentation/d/1vtSmbsDweTLmS8wGgfwNu9aDQMZ3iMnPQeQTlKJje_M/edit?usp=sharing

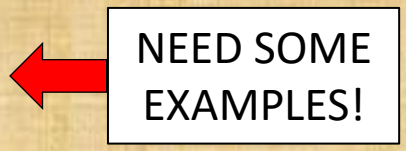
Project Introduction:
The PyHC Open Science Experiment

- The PyHC executable paper demonstrated:
  - How to **collaborate** between software developers/engineers and scientists,
  - How to **use multiple PyHC packages** to perform a science analysis,
  - How to produce an **executable paper** in Heliophysics, and
  - How such a collaboration **supports open science**.
- The goals for this work are to:
  - Apply the workflow developed to a **full-scale science** problem, specifically expanding the 2015 challenge with new data from MMS (https://ccmc.gsfc.nasa.gov/challenges/gem-magnetopause/),
  - Demonstrate how to **perform open science** in Heliophysics, and
  - Improve and **develop modern infrastructure** to streamline collaboration and contributions.

Project Introduction:
The PyHC Open Science Experiment

GEM Science Plan

- Expand to include multiple time ranges of MMS data where magnetopause crossings occurred (retrieved with ***pySPEDAS***),
- Generate the predictions using the empirical Shue model (using ***SpacePy***),
- Generate flythrough results for each contributed physics-based model output stored in s3 (via ***Kamodo***),
- Encourage the community to provide metrics calculation scripts using the flythrough results (built on ***PlasmaPy***), and
- Provide a platform where all contributors can search and reuse all components (on ***HelioCloud***).
- Multiple members of the community are expected to lead portions of the project and produce multiple papers, including a summary paper (coordinated on the ***Open Science Framework***).

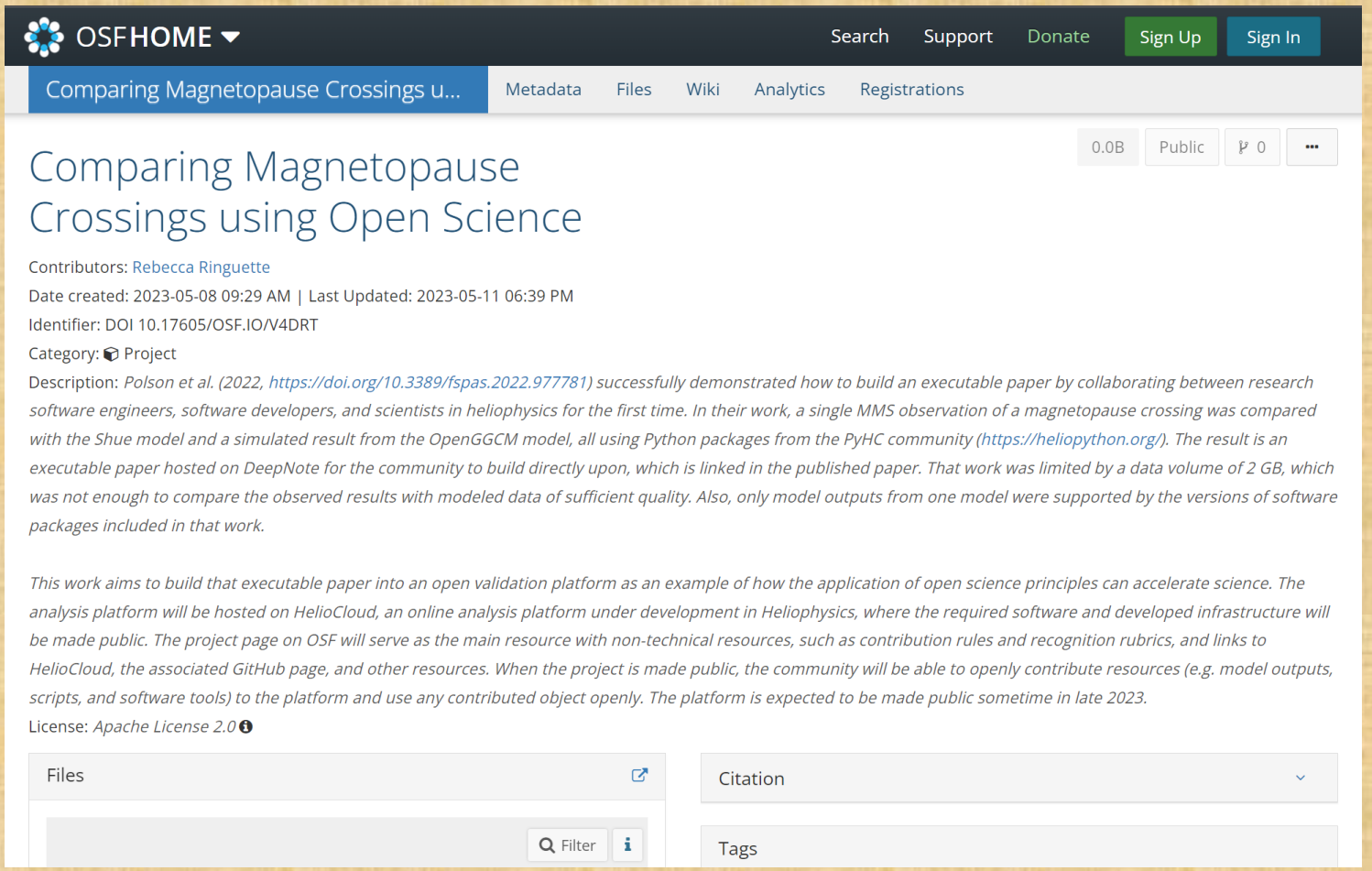


Project Introduction:
The PyHC Open Science Experiment

Open Science goal

- Perform the work **in the open** from the beginning,
- **Demonstrate** how to perform open science to the Heliophysics community and various agencies and nations,
- **Develop** any lacking infrastructure along the way (as reasonably possible),
- **Create** examples of rubrics for recognition/coauthorship and contribution/participation rules for open science, and
- **Publish** a paper describing the challenges discovered, lessons learned, advancements achieved, and how this work can be expanded upon.

Project Tour: OSF Project Page



Please make OSF/ORCiD accounts so I can add contributors!

Project web page: osf.io/v4drt/ DOI: 10.17605/OSF.IO/V4DRT

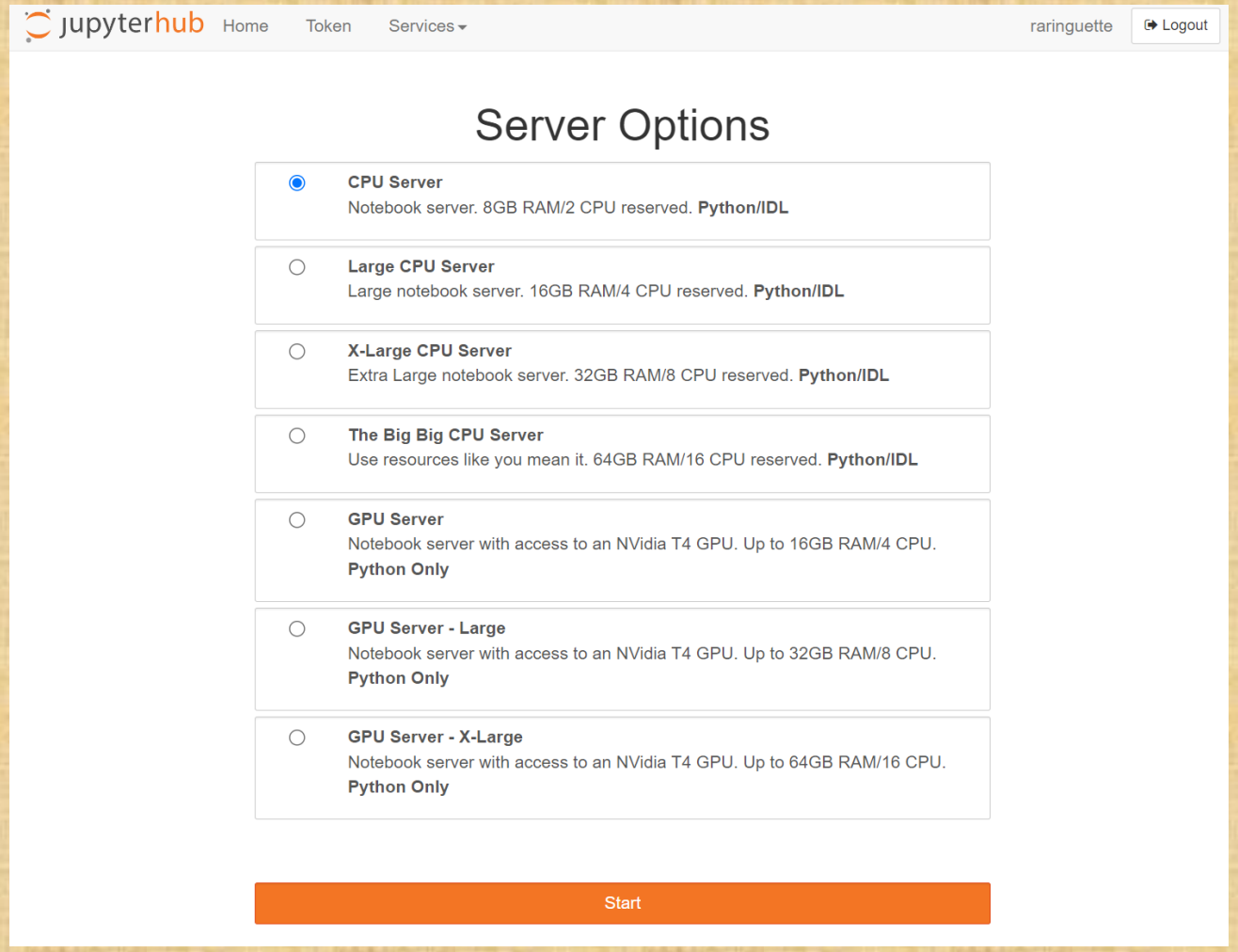
Project Tour: HDRL's HelioCloud

HelioCloud

- Cloud computing environment
- Executable and shareable notebooks
- Large file storage supported via public s3 buckets
- Initial compute and storage costs funded by HDRL

https://daskhub.hsdcloud.org

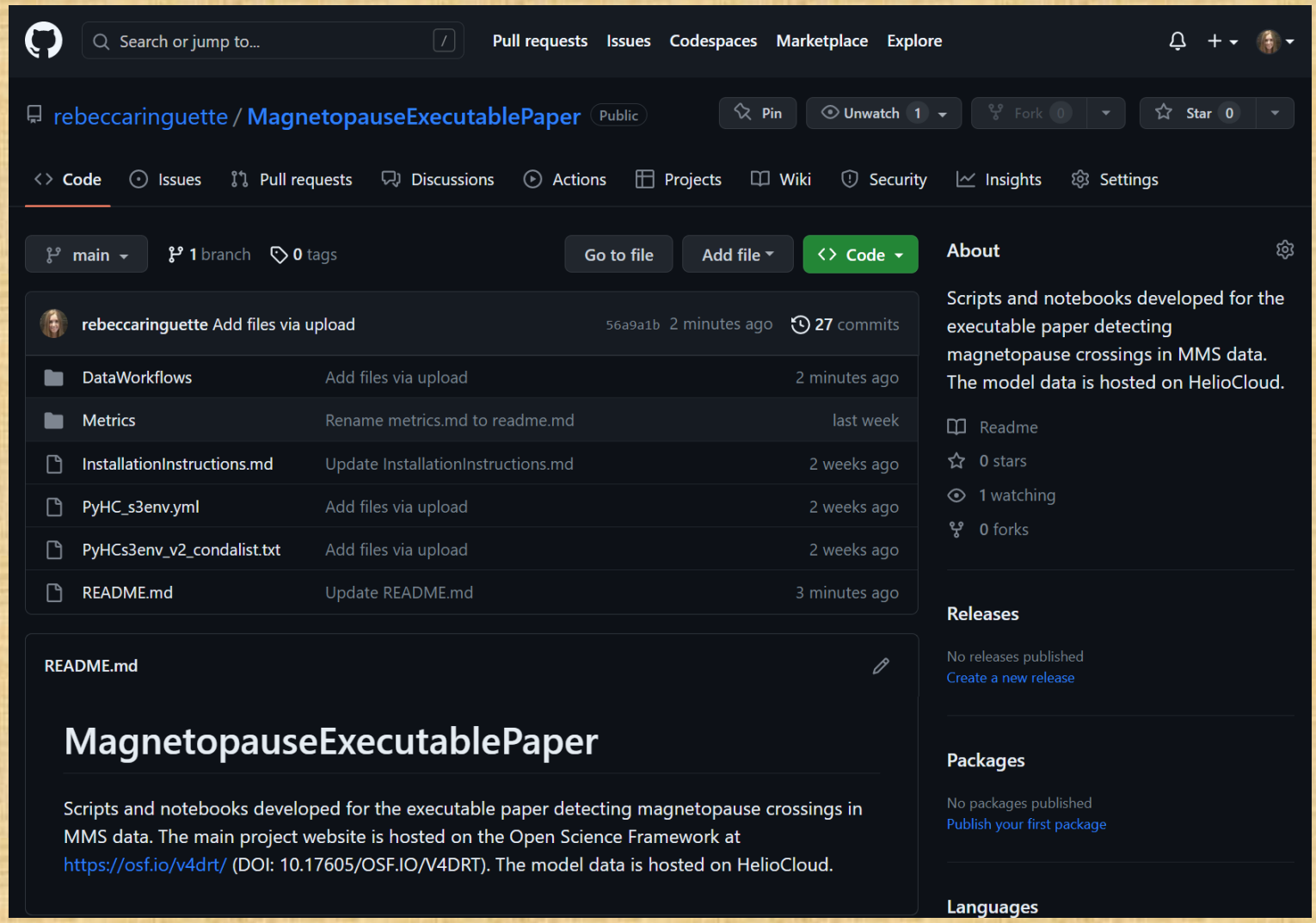
…more tomorrow in S. Antunes' presentation.



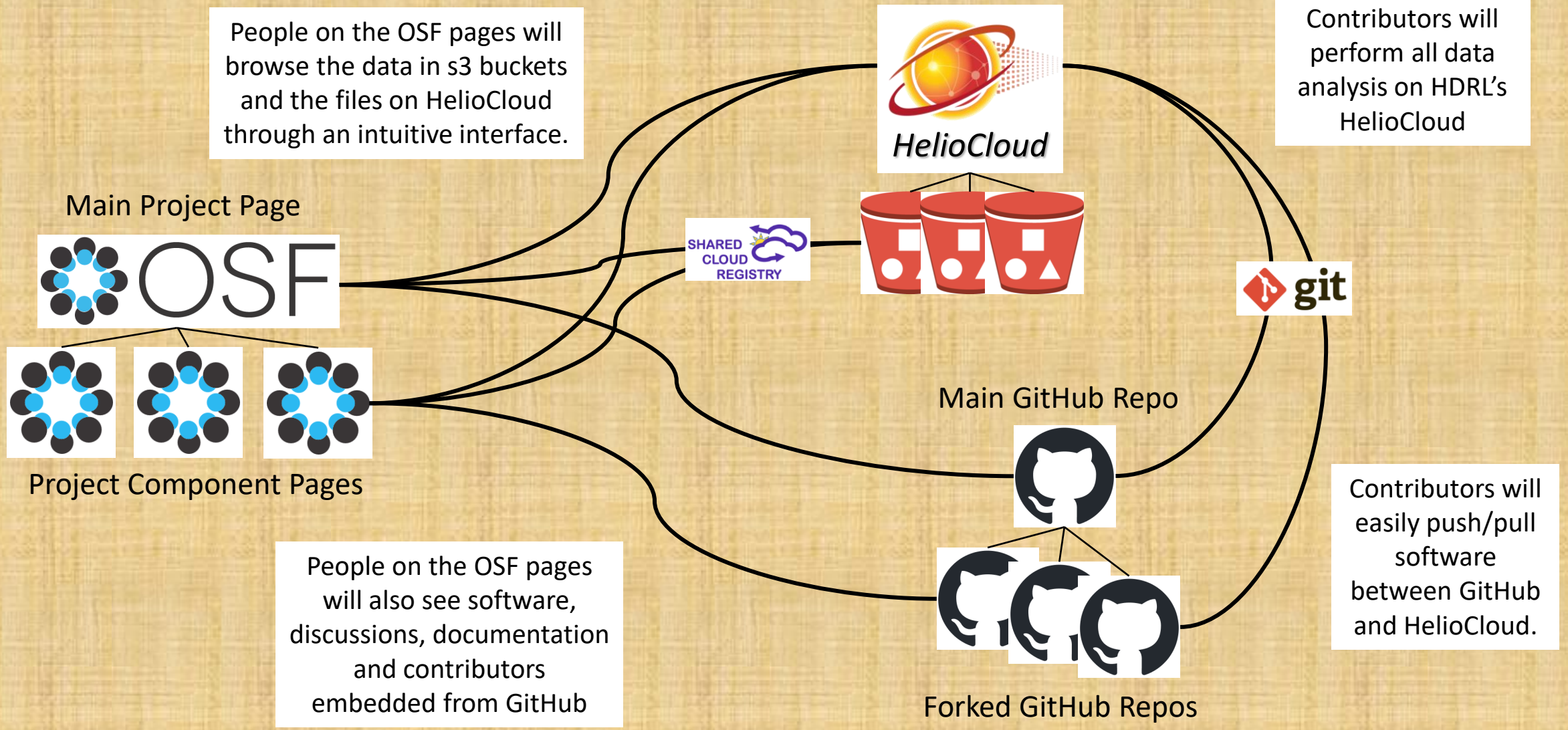
Project Tour: GitHub page

- Link to project webpage added to readme file.
- Scripts and notebooks stored in ‘DataWorkflows’ folder.
- Software environment information in top directory.

https://zenodo.org/badge/latestdoi/631044088

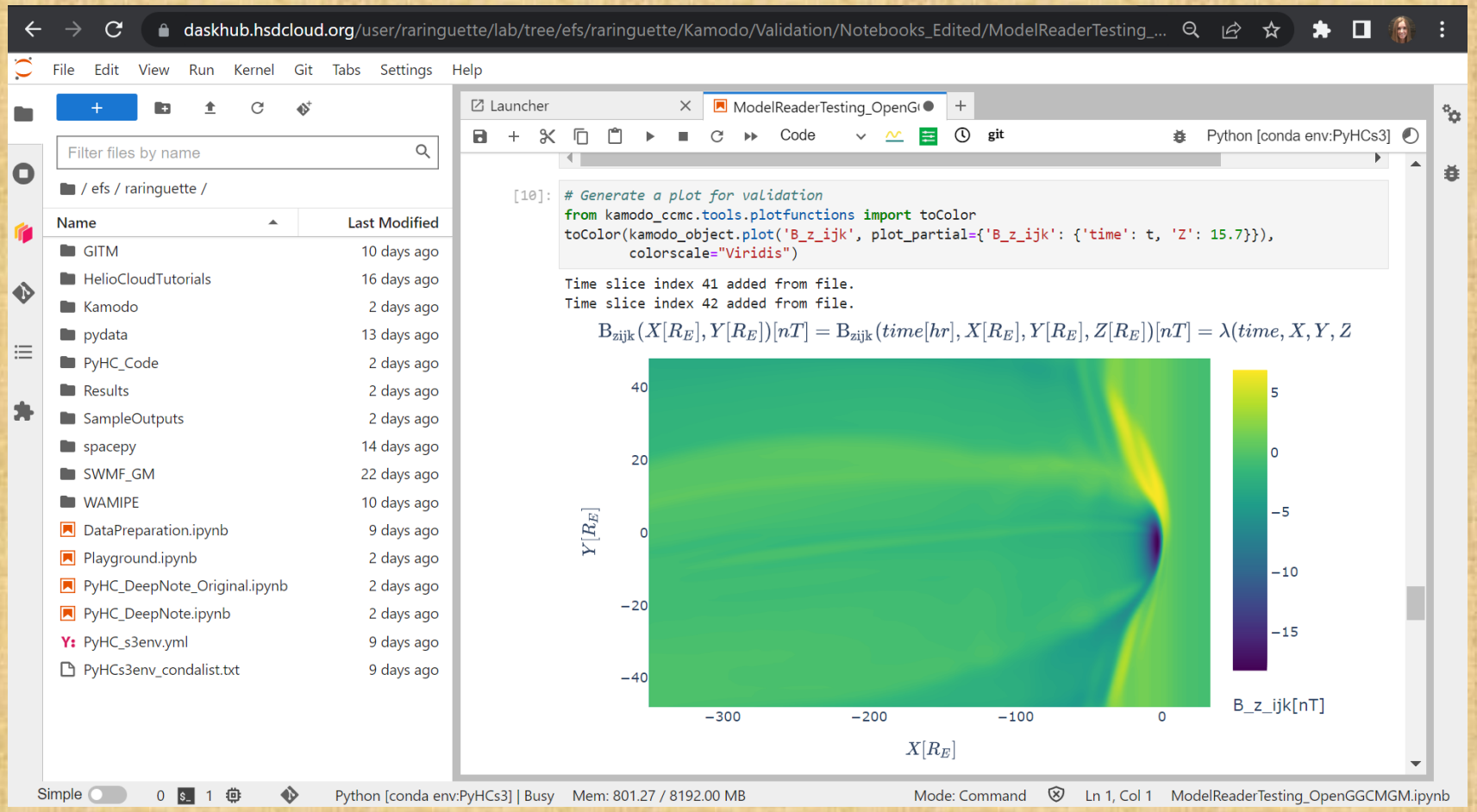


Project Tour: Linking It All Together



Project Tour: Current Status

- Project posted on OSF with a DOI
- Dependency conflicts resolved on HelioCloud
- ***Software runs on data in s3!*** (except for SWMF GM outputs)
- Workflows being planned and developed



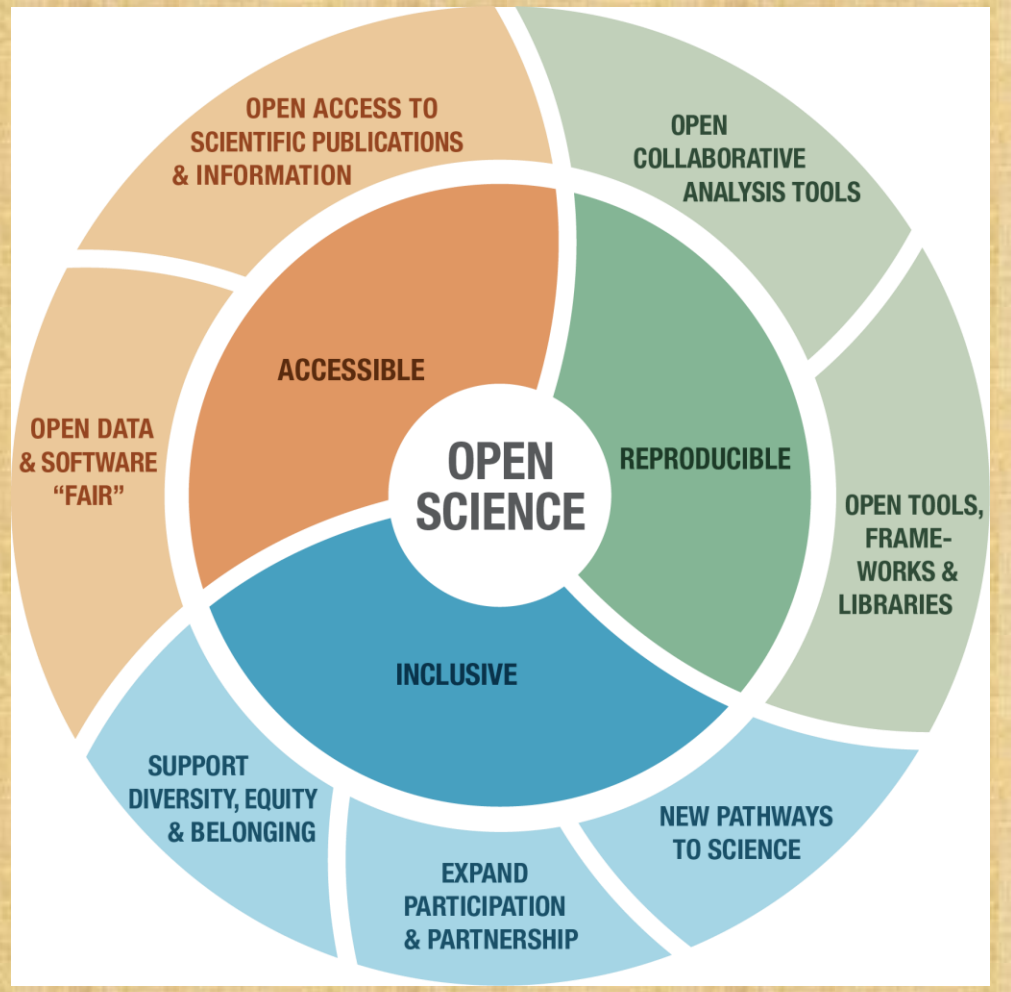
Project Tour: Path Forward

- **Work out kinks** in running PyHC software on data in s3 buckets.
- **Add Contributors!** *(Make an account on OSF/ORCiD so I can add you!)*
- **Link** HelioCloud, OSF and GitHub together (easier said than done!).
- **Streamline** workflows for contributors *(NEED TESTERS!).*
- **Draft** contribution/participation rules based on JWST example.
- **Finalize** contribution/participation rules at Fall PyHC meeting.
- **Present** at AGU 2023 a (hopefully) ready environment.

Any burning questions?

Open Science and Heliophysics Infrastructure

- ***Open Science*** is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity.
  (https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/)
- ***Why open science?***
  - Accelerates scientific discovery.
  - Greater collaboration and efficiency.
  - Enhanced transparency and reproducibility (NASEM, 2018, p. 3).
  - Mandated by the U.S. White House and NASA.
- Image Credit: NASA TOPS (https://zenodo.org/record/6565080#.ZFPvCnbMKUk)



Open Science and Heliophysics Infrastructure

- **FAIR** components (***F**indable, **A**ccessible, **I**nteroperable, **R**eusable*):
  - *Making good progress*: publications, observed data, metadata,
  - *Needs focused development*: modeled data, software,
  - *Exploration required*: model codes, software environments,
  - *The great unknown*: people, relationships, collaborations, …
- **Reproducible** results
  Executable papers? Analysis environments? How long to maintain and what depth of reproducibility?
- **Open** processes
  How to perform science in the open from the beginning?
- **Inclusive** collaborations
  How to make collaborations open?

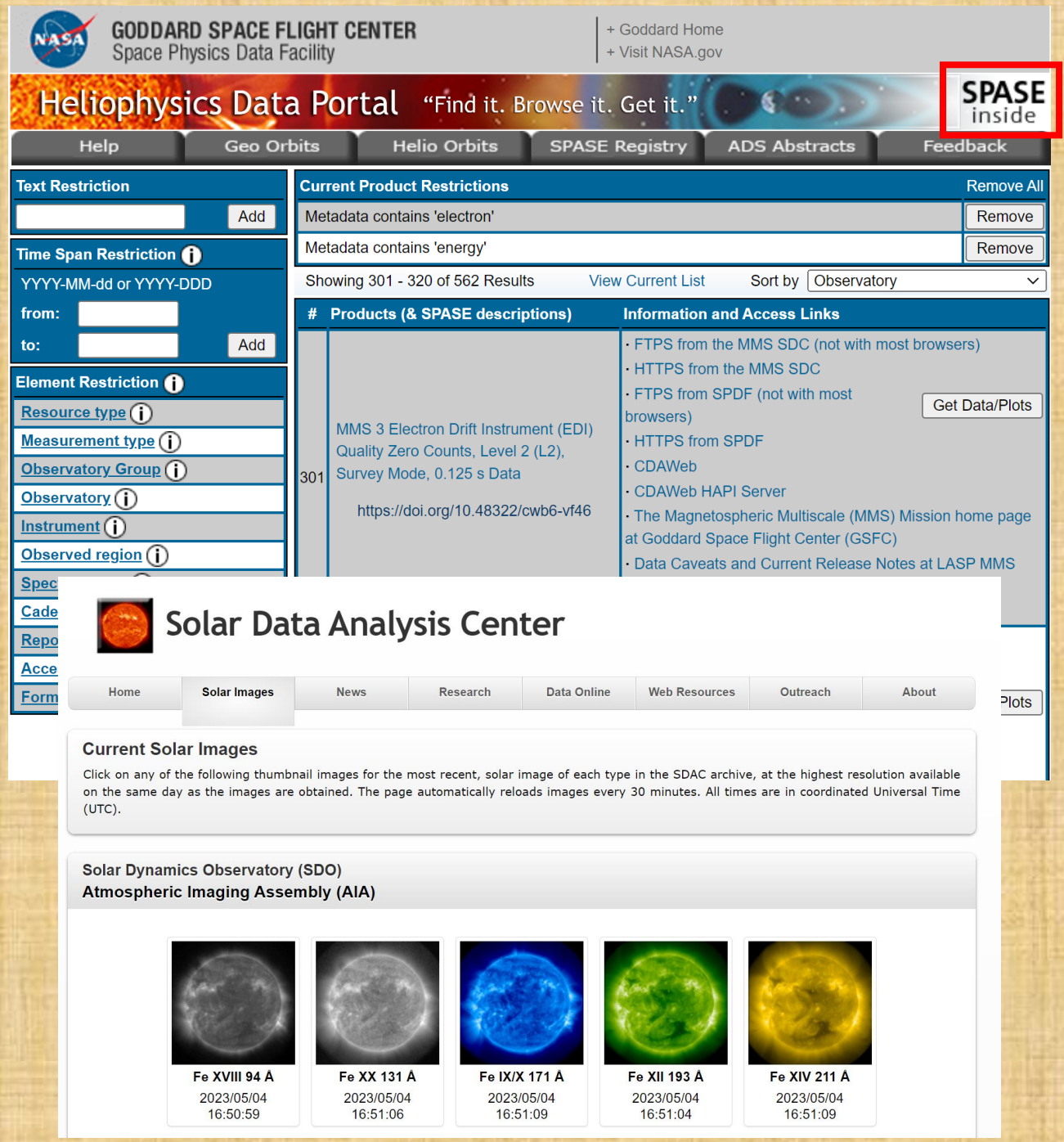
FAIR data and open-source software are NOT enough!

It is okay to start there, but we *must* look beyond for guidance on infrastructure design.

Observational Data

- A growing number of datasets…
  - Are **searchable** through a modern interface (using SPASE),
  - Have **citable** DOIs independent of publications,
  - Are **downloadable** both through web pages and APIs,
  - Are **browsable** via quick-look plots, and
  - Are available on the **cloud**.

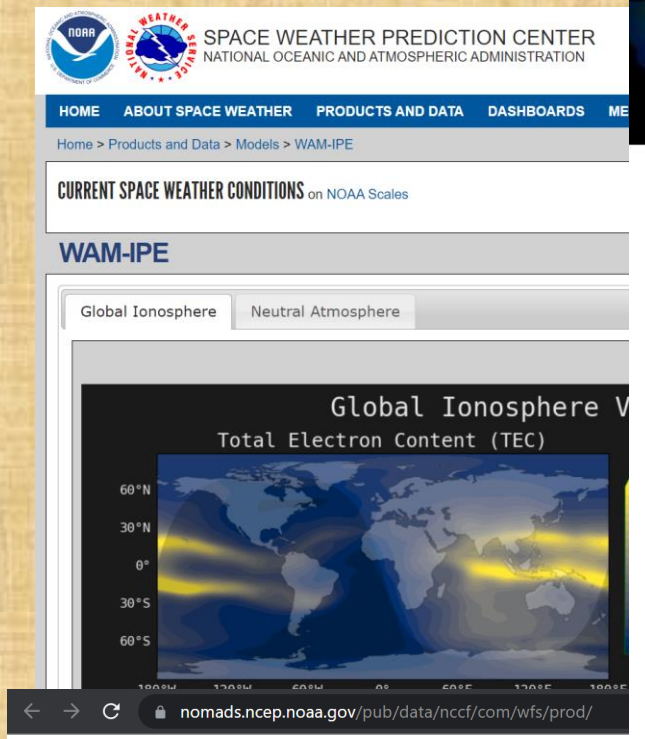
How can PyHC better advertise data access and analysis support in PyHC packages?

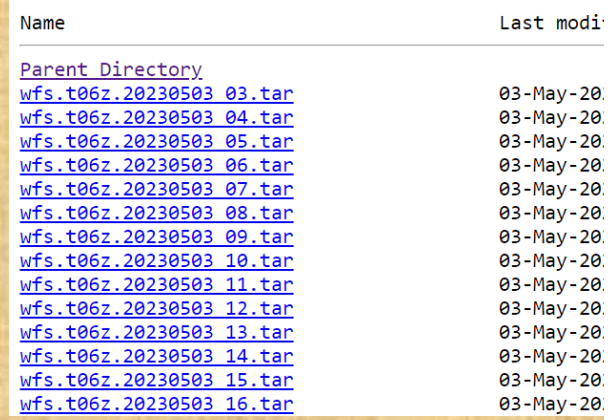


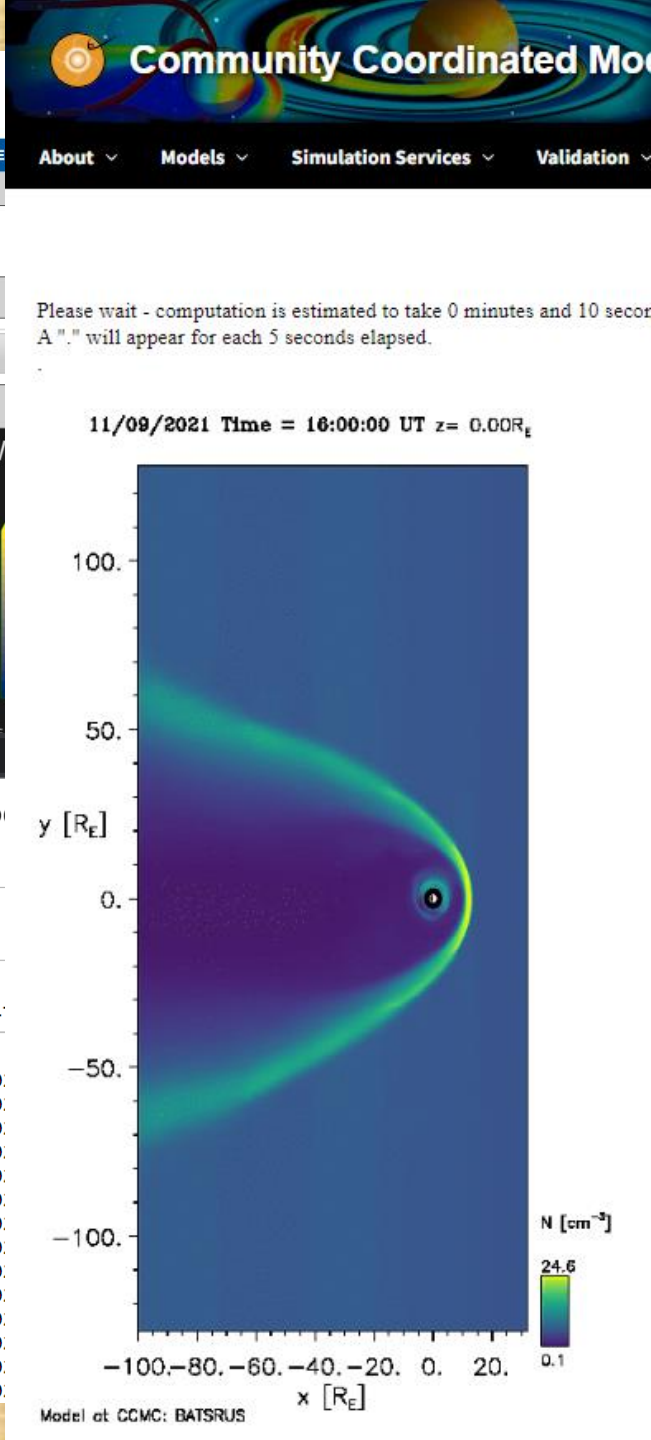
Modeled Data

- Infrastructure supporting modeled data is **far less developed**.
  - No modern **search** interfaces,
  - **DOIs** are not assigned,
  - Few modeled datasets are **downloadable** through a website,
  - Only reduced versions are **available** through an API, and
  - Some quick-look **plotting** capabilities are available, but are not easily accessible.

How can PyHC help with these issues?

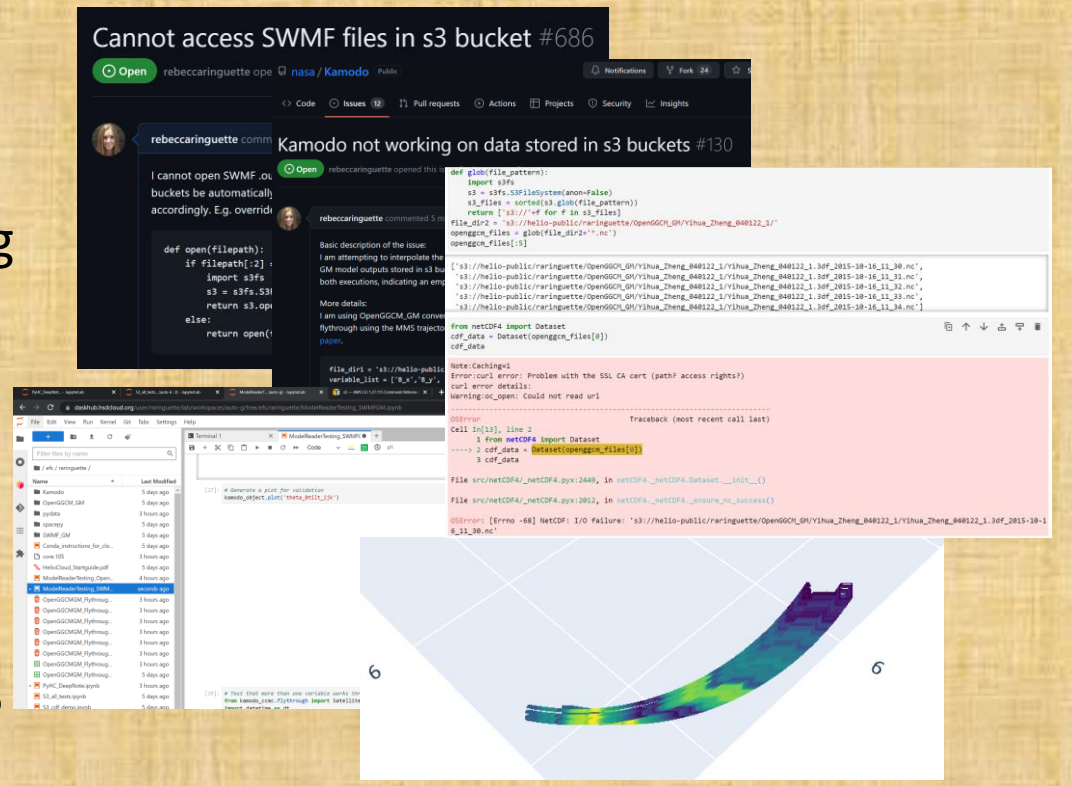






Software

- Sustained push is underway to **open-source all software** (including modeling code) generated with taxpayer dollars.
- Open-sourcing software is **NOT enough**.
  - Dependency conflicts (!),
  - Conda/pip installability on multiple operating systems (e.g. Mac, Windows, Linux),
  - Lacking documentation,
  - Need examples and tutorials,
  - Capability to run on the cloud,
  - Maintenance for long-term reusability,
  - Support staff for questions/problems, and
  - Containerization for software environments?



PyHC should take the lead here. What paths forward have low-hanging fruit?

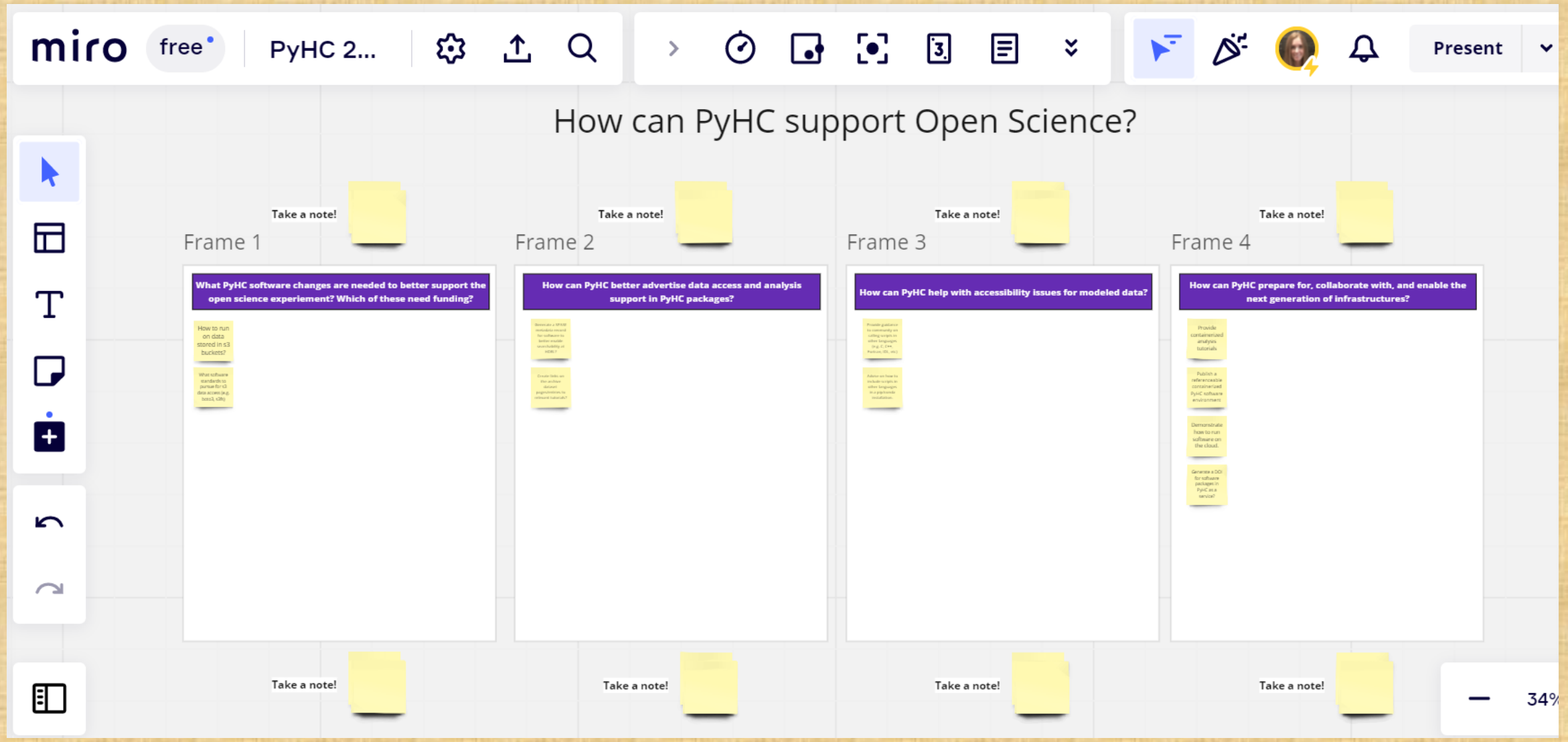
Where is this going?

- Build a **distributed data infrastructure** system:
  - **Observational and modeled data** hosted and served by multiple institutions,
  - **Containerized model codes** available on the cloud from multiple institutions,
  - All searchable from a united modern interface through **connected metadata**,
  - All **accessible** using multiple methods (e.g. file links, APIs, quick-look plots).
- Build a **collaborative analysis infrastructure** system:
  - Analysis environments with **software already installed** (and referenceable),
  - **Reusable executable** analysis tutorials for how to use the data,
  - Searchable through **connected metadata**,
  - **Accessible** through the cloud (e.g. downloadable containers or cloud platforms).

How can PyHC prepare for, collaborate with, and enable these infrastructures?

Discussion Time!

Discussion: PyHC support of Open Science



Discussion: PyHC support of Open Science

Scan the QR code or use the link to contribute to the discussion!

https://tinyurl.com/5n6tsxt6

What PyHC software changes are needed to better support the open science experiement? Which of these need funding?

How to run on SWMF GM data stored in s3 buckets?

Need a metrics script examples that uses PlasmaPy

mean error, RMS, root square mean, absolute error

acknowledgment of the code in papers

a common release schedule so that there is always a compatible / functioning, relatively recent and complete set of PyHC tools

understand/development of the metric of the code "usefulness"

We need funding for long-term maintenance of code!

1

standardized packaging
prereqs
install scripts
dir structure

DOI's for software - should PyHC offer DOI minting or should a DOI be required for software to be listed by PyHC?

Code Standards implementing best practices

PyHC packages should test against the main branch or release candidates of their dependencies

How can PyHC better advertise data access and analysis support in PyHC packages?

Generate a SPASE metadata record for software to better enable searchability at HDRL?

Improved keyword usage/search in PyHC Projects page

<-- Category that separates solar physics, magnetospheric, etc?

Out-of-the-box: A custom GPT-powered chatbot on pyhc.org that can explain what's available in conversation

booth with tutorials at meetings

Create links on the archive dataset pages/entries to relevant tutorials?

Also 👆, improve the PyHC gallery ;)

community crib sheets with examples for how to use a package - as someone learns to use a new package, the sample code they create (with some curation?) could be useful examples for the next newcomer

<-- include these in the PyHC gallery?

Webinars (similar to pySPEDAS)

Advertisement at ALL conferences

PyHC office hours?

have a standard set of tutorials that anyone going to a meeting can take so that we have lots of reporesentatives who do the training

<--- Can pull from the PyHC summer school tutorials, our gallery, etc, and create some kind of nice pre-packaged set of examples for training?

Data archives point to PyHC packages that work well with that data.

we need an army of tutors - offer specific PyHC grants for people to do promulgation - or this could be an add-on to another grant - a little extra money to attend one or two extra meetings and give totorials

Should PyHC hire a community manager like Astropy?

Hosting work parties to improve project documentation

More example notebooks in PyHC gallery that use multiple packages together

How can PyHC help with accessibility issues for modeled data?

Provide guidance to community on calling scripts in other languages (e.g. C, C++, Fortran, IDL, etc)

get CCMC to engage more with the community and find out more about what users need

many modelers are using Paraview, so PyHC could spend resources to support Paraview

Advise on how to include scripts in other languages as dependencies in a pip/conda installation.

Simplify the process to have SpacePy as a dependency

make a specific call in the HTM NASA AO for improving model access

Create metadata standard(s) for simulation inputs/outputs

standardized Data Storage format
access methods

modeling results hosting/server should exist

Partner with IHDEA

How can PyHC prepare for, collaborate with, and enable the next generation of infrastructures?

Provide containerized analysis tutorials

Improve culture by building a foundation of psychological safety

Continued outreach to the next generation of scientific software developers and scientists through engaging and informative PyHC Summer schools, presentations and booths at other meetings, etc.

yearly simultaneous release of all PyHC packages for compatibility

Publish a referenceable containerized PyHC software environment

Per Gretche's diagram where the last step before making open science required is making open science being rewarding, we should advertise how it's rewarding (and if we can't figure out how its rewarding to individuals, we reconsider making it more rewarding)

Funding for a software engineer who focuses on package/docs/testing infrastructure?

Hackathon or similar to try out a lot of different ways to containerize/compartmentalize a turnkey py?HC environment

Demonstrate how to run software on the cloud.

Generate a DOI for software packages in PyHC as a service

allow multiple options for obtaining a viable collection of PyHC tooling

What software standards to pursue for s3 data access (e.g. boto3, s3fs)?

is Podman/Docker a viable 'package system' for prereqs+code to work?

Docker alternative: https://podman.io/

use singularity instead of Docker

data access beyond local files of all kinds (HAPI, other companies' clouds)

https://docs.sylabs.io/guides/3.5/user-guide/introduction.html

(singularity)

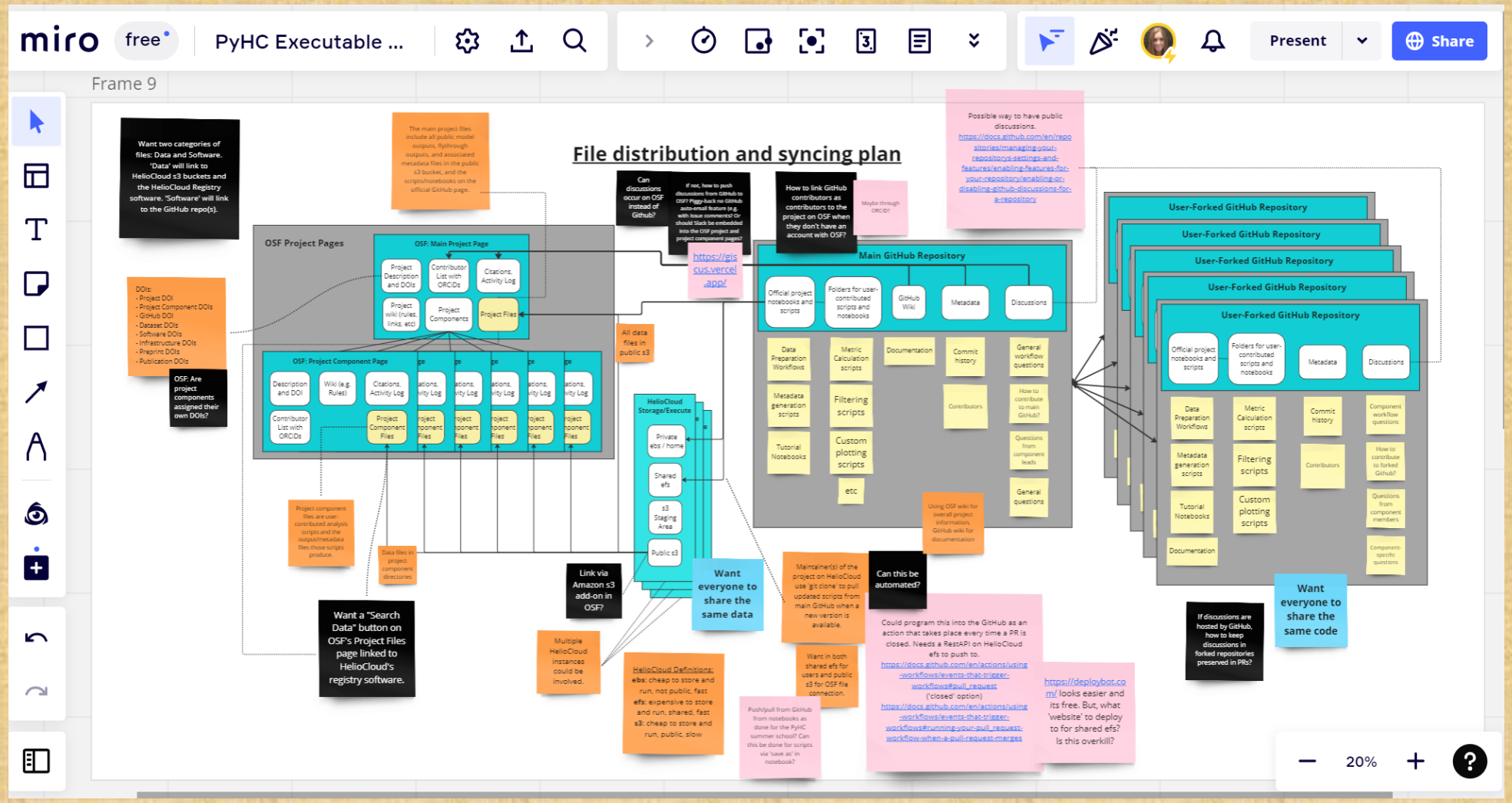
The Execution Model prevent arbitrary code execution (aka no bitcoin mining and spam)

Hardware dependence/support (large data, high core count, large memory, GPUs)

Useful Links

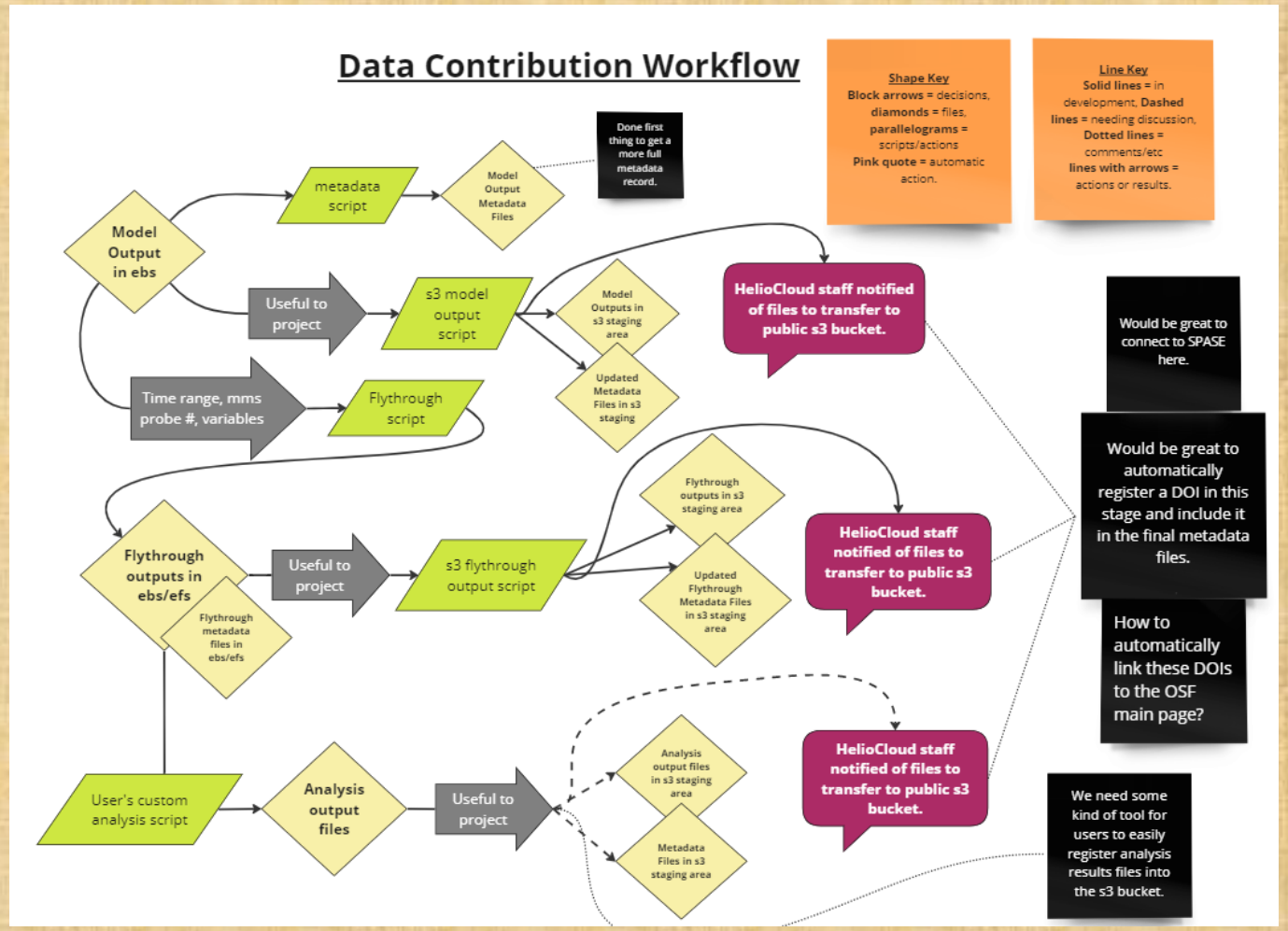
- OSF project page: https://osf.io/v4drt/
- GitHub project page: https://github.com/rebeccaringuette/MagnetopauseExecutablePaper or https://zenodo.org/badge/latestdoi/631044088
- HelioCloud: https://daskhub.hsdcloud.org
- CCMC 2015 Magnetopause challenge: https://ccmc.gsfc.nasa.gov/challenges/gem-magnetopause/
- Polson et al. 2022 journal article: https://doi.org/10.3389/fspas.2022.977781
- Polson et al. 2022 on Deep Note: https://deepnote.com/workspace/shawn-polson-c095a0fb-f02d-416d-9c94-c4a9c4e8e54d/project/PyHC-Paper-101b9646-3fd0-4978-a48e-a4f3e708a0ac/notebook/Making_an_Executable_Paper_with_the_Python_in_Heliophysics_Community_to_Foster_Open_Science_and_Improve_Reproducibility-c3a5772e5ce24ce1942b001696d52251
- This presentation's link on PyHC's google drive: https://docs.google.com/presentation/d/1c2bP0zDdiJWMCPZZxzm9U3zx80NPH_SC/edit?usp=share_link&ouid=118198339287841207428&rtpof=true&sd=true
- Miro board link with preliminary project workflows (view): https://miro.com/welcomeonboard/Q05NeGI2M0taVGtxekJjZXkzVzJsdzFud3R3SlF4RWJ4RGN0NXBmazZheThhd1d4aUNLclVDOFM1WHhXa01ZWXwzMDc0NDU3MzU3OTk0ODcyOTEyfDI=?share_link_id=953968385486
- Miro board link with PyHC 2023 spring session discussion (edit): https://miro.com/welcomeonboard/M2JRd2NNSXhPcXpPRXZJbDFHczcwRHNqYzUzSk9XelBzeVdkT2p3NThhQlVBRkQ4VEJzaXc5MWNGT3dhUkIxR3wzMDc0NDU3MzU3OTk0ODcyOTEyfDI=?share_link_id=522570131865

Project Tour: Preliminary Project Workflows



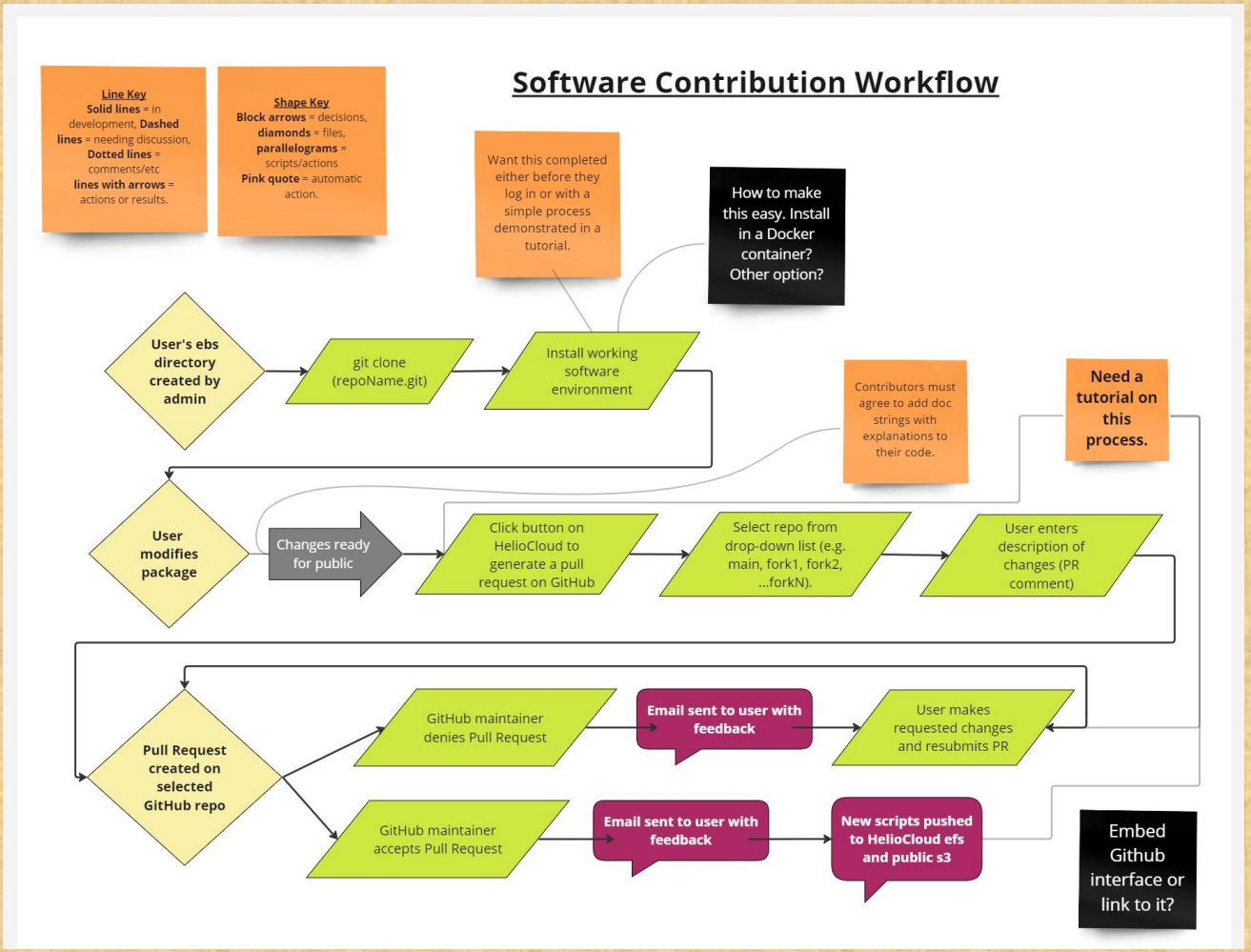
https://miro.com/app/board/uXjVMMaO61I=/?share_link_id=703942255605

Project Tour: Preliminary Project Workflows



https://miro.com/app/board/uXjVMMaO61I=/?share_link_id=703942255605

Project Tour: Preliminary Project Workflows



https://miro.com/app/board/uXjVMMaO61I=/?share_link_id=703942255605