Advancing Open and Reproducible Water Data Science by Integrating Data Analytics with an Online Data Repository

Jeffery Horsburgh[1], Scott Black[2], and Anthony Castronova[3]

[1]Utah State University
[2]Consortium of Universities for the Advancement of Hydrologic Sciences, Inc.
[3]Consortium of Universities for the Advancement of Hydrologic Sciences, Inc

November 21, 2022

Abstract

Scientific and related management challenges in the water domain require synthesis of data from multiple domains. Many data analysis tasks are difficult because datasets are large and complex; standard formats for data types are not always agreed upon nor mapped to an efficient structure for analysis; water scientists may lack training in methods needed to efficiently tackle large and complex datasets; and available tools can make it difficult to share, collaborate around, and reproduce scientific work. Overcoming these barriers to accessing, organizing, and preparing datasets for analyses will be an enabler for transforming scientific inquiries. Building on the HydroShare repository's established cyberinfrastructure, we have advanced two packages for the Python language that make data loading, organization, and curation for analysis easier, reducing time spent in choosing appropriate data structures and writing code to ingest data. These packages enable automated retrieval of data from HydroShare and the USGS's National Water Information System (NWIS), loading of data into performant structures keyed to specific scientific data types and that integrate with existing visualization, analysis, and data science capabilities available in Python, and then writing analysis results back to HydroShare for sharing and eventual publication. These capabilities reduce the technical burden for scientists associated with creating a computational environment for executing analyses by installing and maintaining the packages within CUAHSI's HydroShare-linked JupyterHub server. HydroShare users can leverage these tools to build, share, and publish more reproducible scientific workflows. The HydroShare Python Client and USGS NWIS Data Retrieval packages can be installed within a Python environment on any computer running Microsoft Windows, Apple MacOS, or Linux from the Python Package Index using the PIP utility. They can also be used online via the CUAHSI JupyterHub server (https://jupyterhub.cuahsi.org/) or other Python notebook environments like Google Collaboratory (https://colab.research.google.com/). Source code, documentation, and examples for the software are freely available in GitHub at https://github.com/hydroshare/hsclient/ and https://github.com/USGS-python/dataretrieval.

Advancing Open and Reproducible Water Data Science by Integrating Data Analytics with an Online Data Repository

Jeffery S. Horsburgh

Utah State University

Scott Black, Anthony Castronova

Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI)

Utah Water Research Laboratory
UtahStateUniversity

Reproducibility is key

“If I have seen further it is by standing on the shoulders of Giants.”

Isaac Newton, 1625

Building trust in scientific research requires transparency and reproducibility

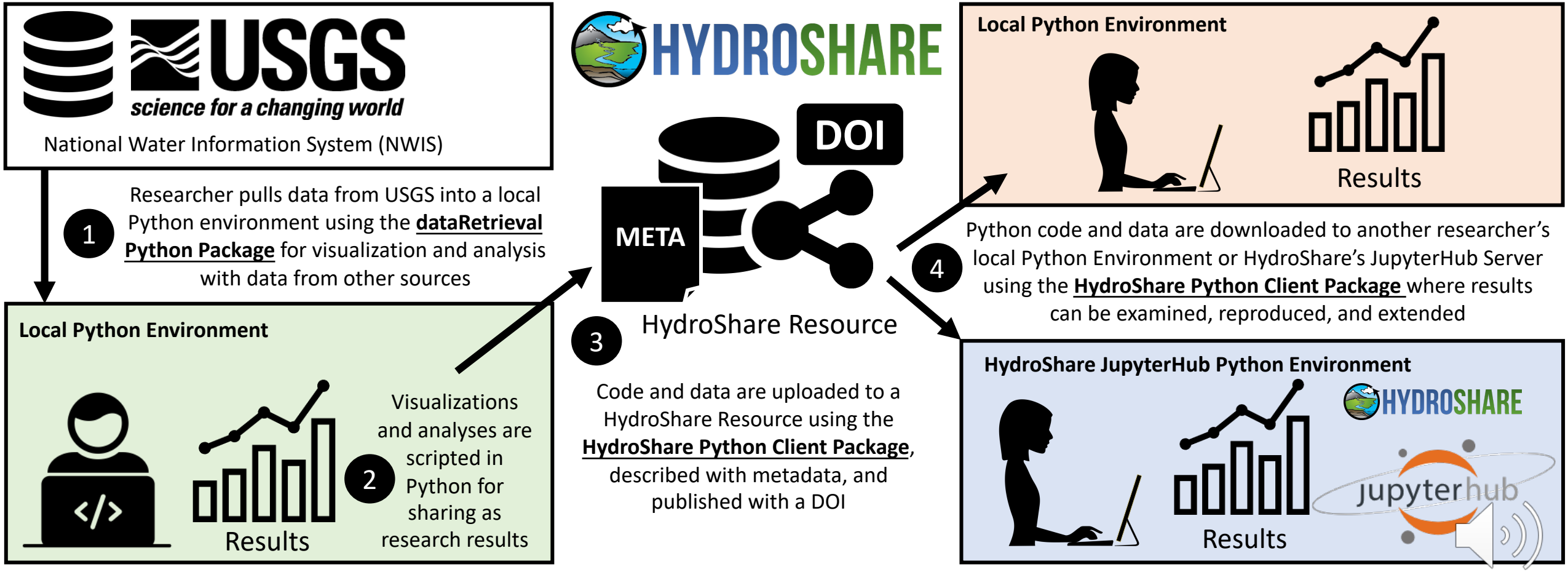
Collaborative (Reproducible) Data Science Workflow

- Easily create a digital instance of a dataset or data science workflow
- Quickly share it with colleagues (perhaps privately at first)
- Add value through collaboration, annotation, and iteration
- Describe with metadata
- Eventually…share publicly or formally Publish so others can reuse

What is the role of data repositories in this scientific workflow?

Connecting Visualization and Analysis with an Online Repository

- Better enabling collaborative data science workflows and reproducibility

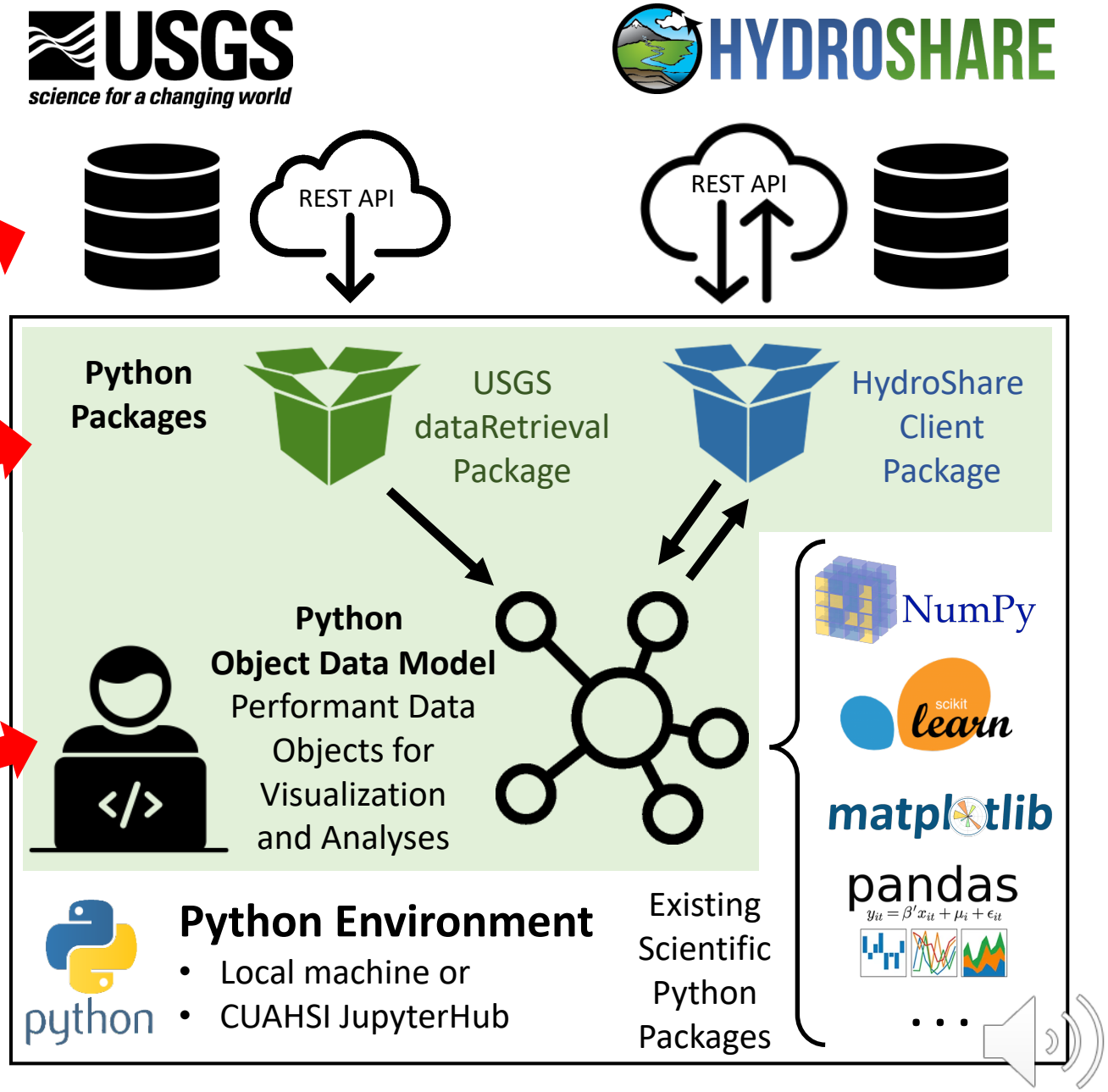


The tools needed to make this work

Data repositories

Tools for accessing and interacting with those repositories

A Python representation of the data retrieved that can be operated on using existing data science tools

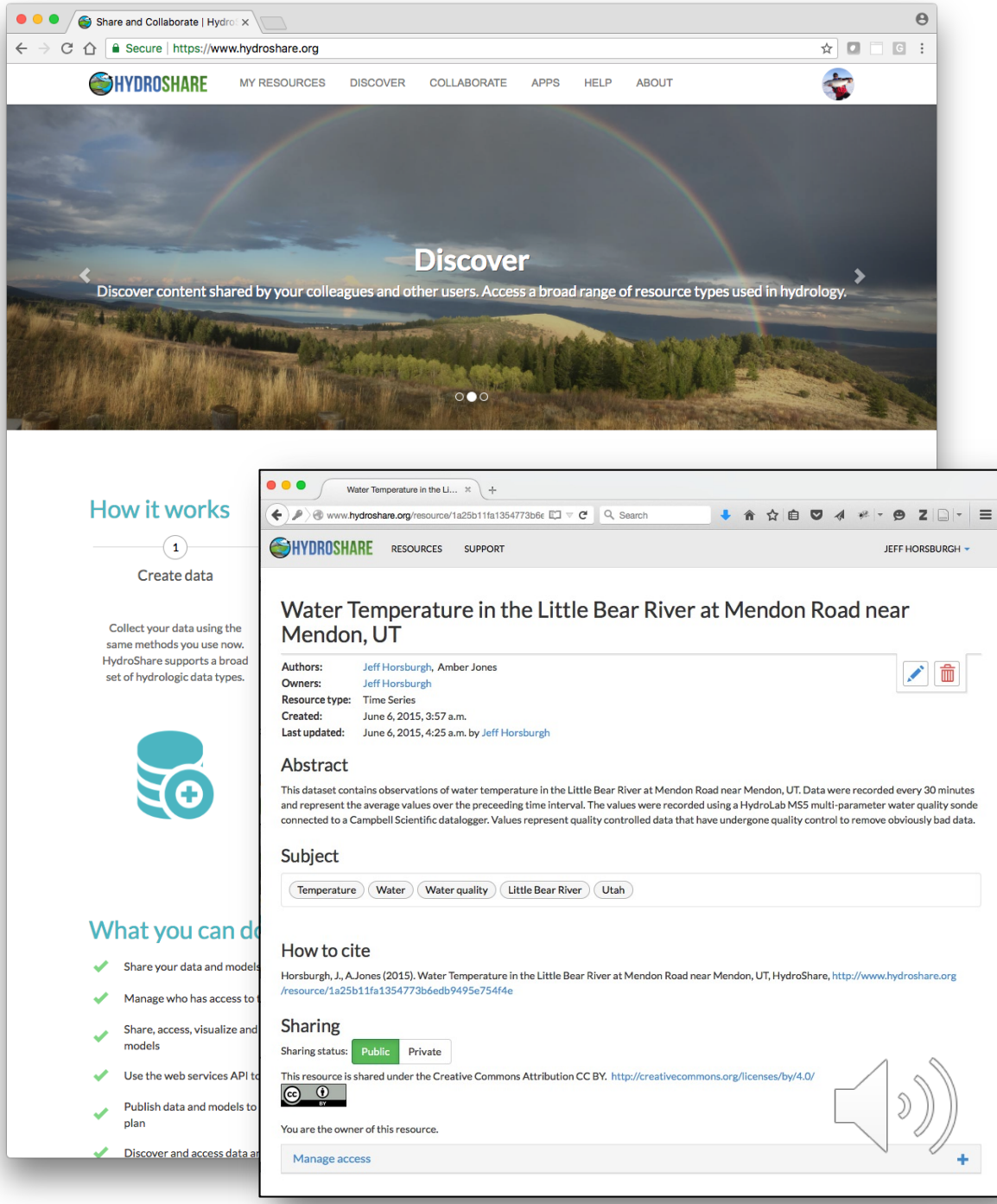


HYDROSHARE

http://www.hydroshare.org

- A repository for sharing and publication that uses FAIR principles
- Operated by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI)
- Creating and sharing data and models using a variety of file formats and flexible metadata

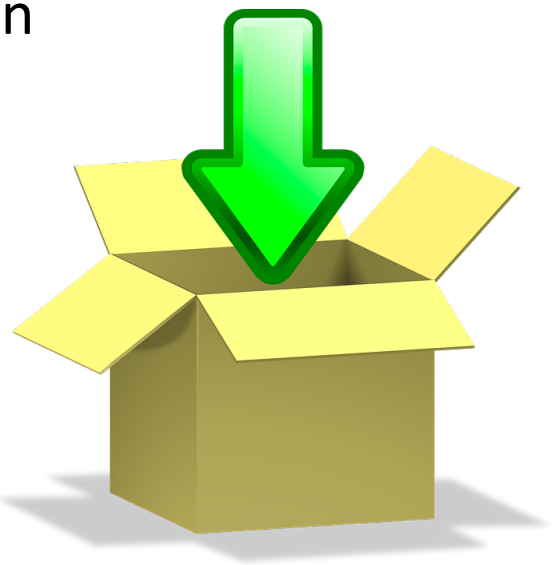
- Public-facing REST API and Python client enabling automated interactions



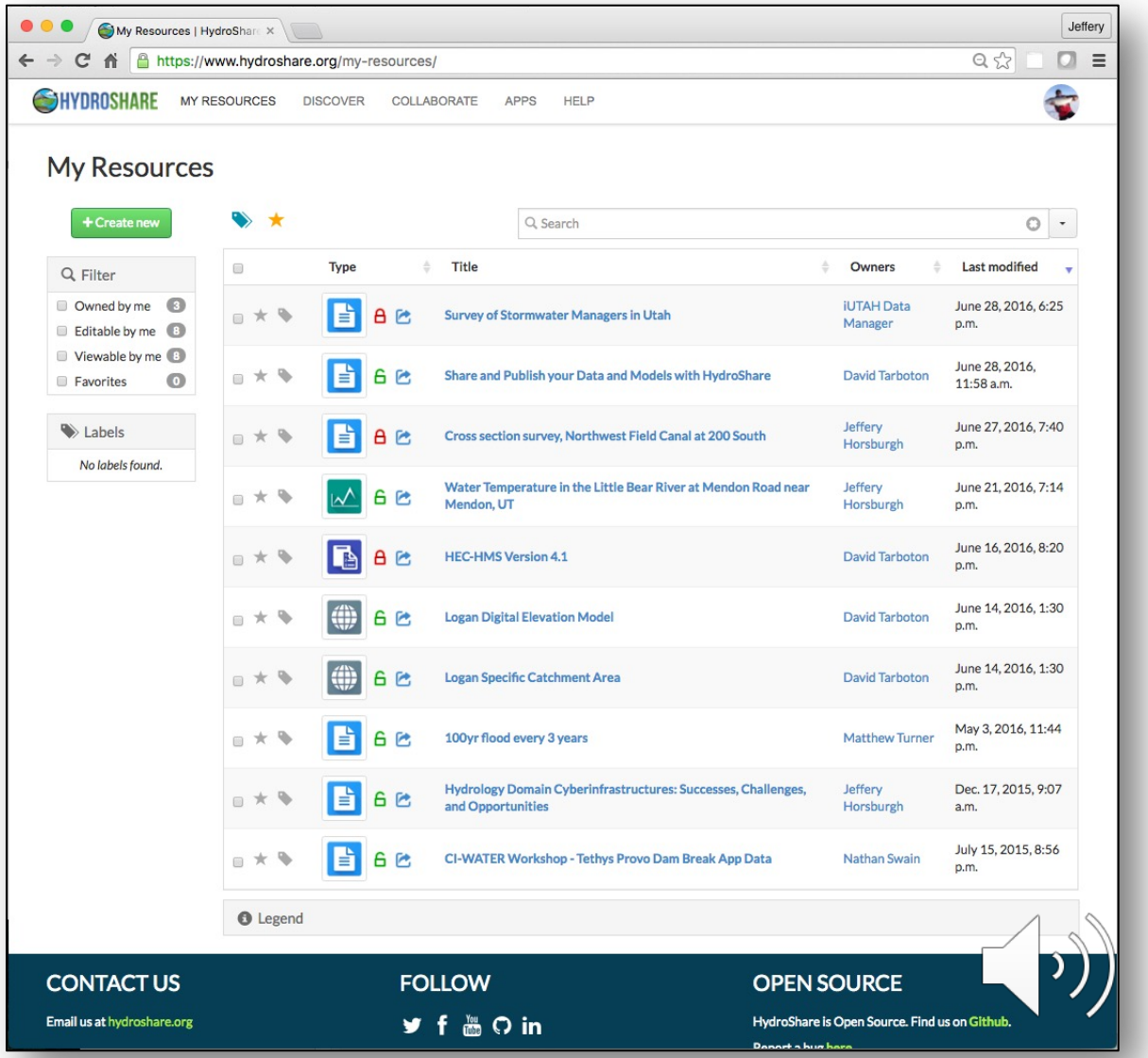
HydroShare "Resources"

- **Resource** = primary unit of digital content
  - Create, version, copy
  - Describe
  - Own, share, access
  - Discover
  - Formal Publication

A "Resource" is a container into which users can put digital content

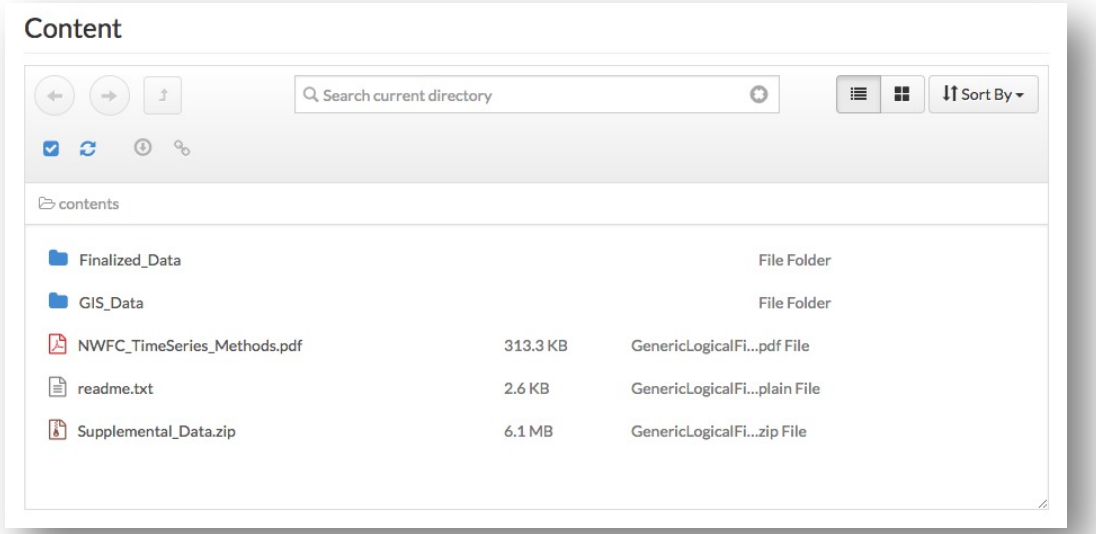


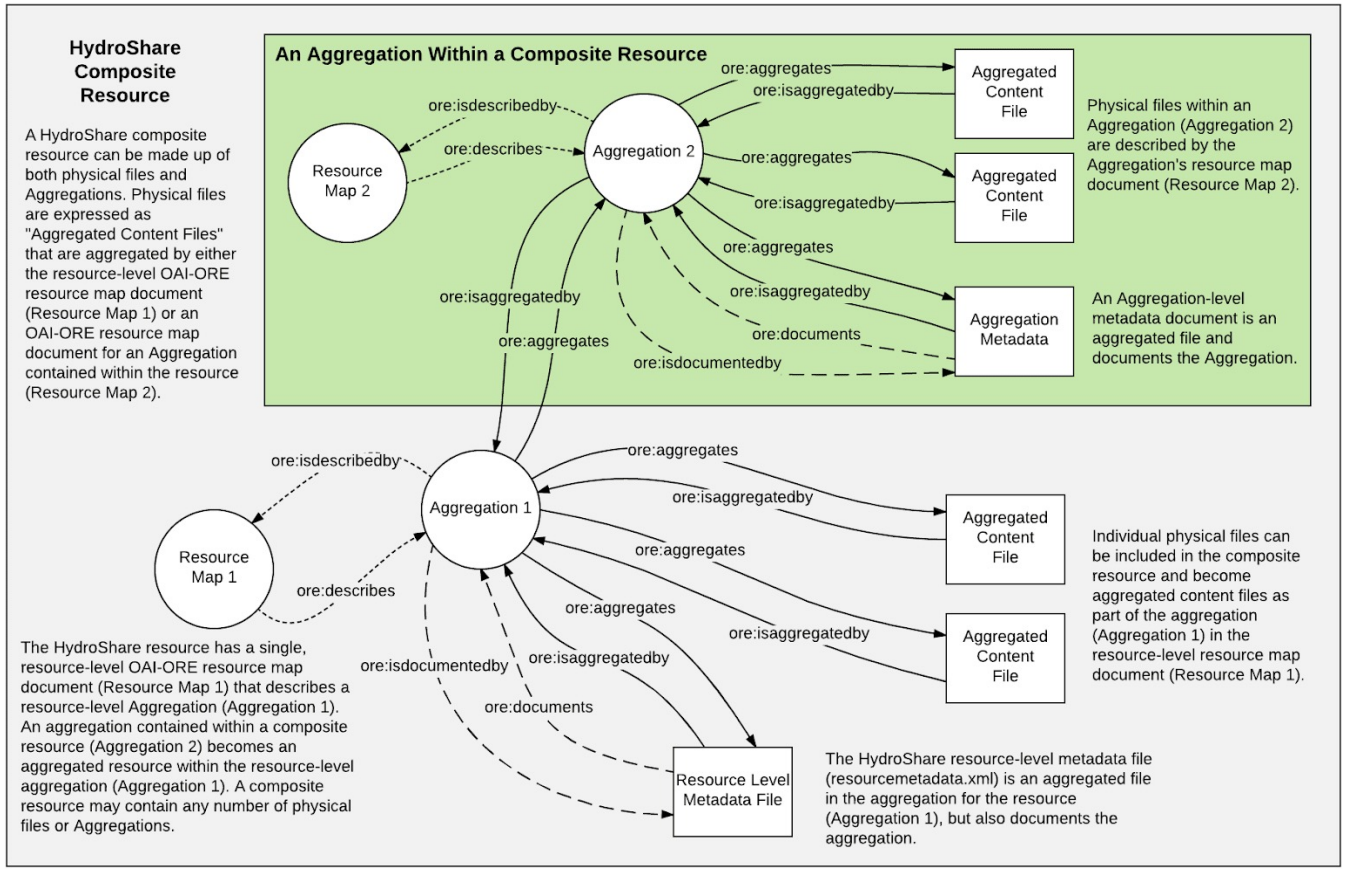
Resources can be datasets, models, or other research products



HydroShare "Resources"

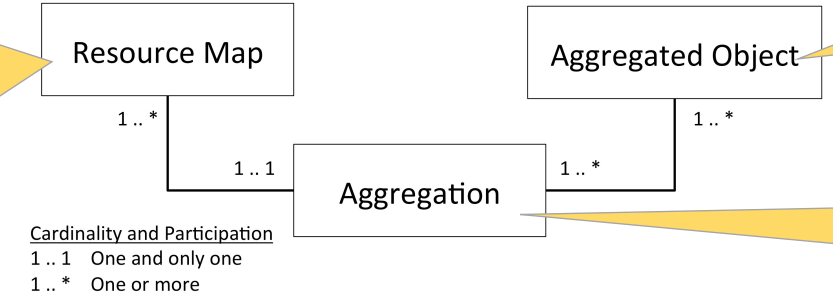
A file/content – based definition





A semantic definition based on OAI-ORE

An XML document that encodes the description of a Resource and the Aggregation



A file that is part of a Resource

A list of all of the objects/files aggregated within the resource

A profile of the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) standard

Horsburgh, J. S., Morsy, M. M., Castronova, A., Goodall, J. L., Gan, T., Yi, H., Stealey, M. J., and D.G. Tarboton (2015). HydroShare: Sharing diverse hydrologic data types and models as social objects within a Hydrologic Information System, Journal Of the American Water Resources Association (JAWRA), http://dx.doi.org/10.1111/1752-1688.12363.

HydroShare “Resource” as a Data Science Enabler

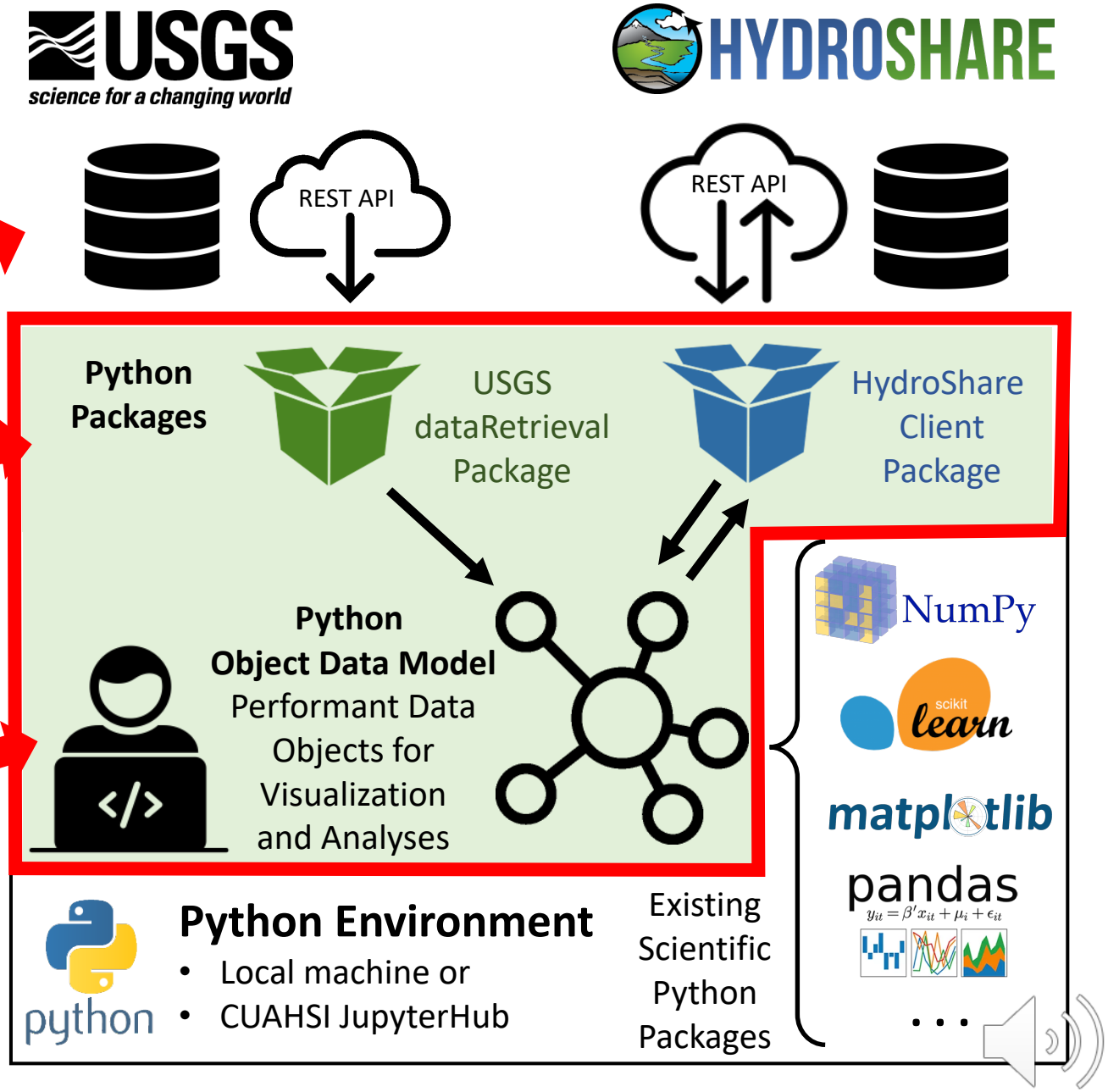
- When creating reproducible data science workflows – how to organize?
  - Eventual goal is to share the analysis
  - Need to be able to get data/code into a repository
  - Need straightforward ways to organize the content used
    - Potentially inputs, outputs, code, etc.
- The HydroShare Resource is a great organizing container
  - Think of it as a “Project Directory”
  - Existing Resource Data Model
  - Machine readable semantic representation of structure
  - Flexible
  - Existing “aggregation” types identify commonly used data
  - Can map the whole thing to Python for easy manipulation
  - Already handled by a repository!

The tools needed to make this work

Data repositories

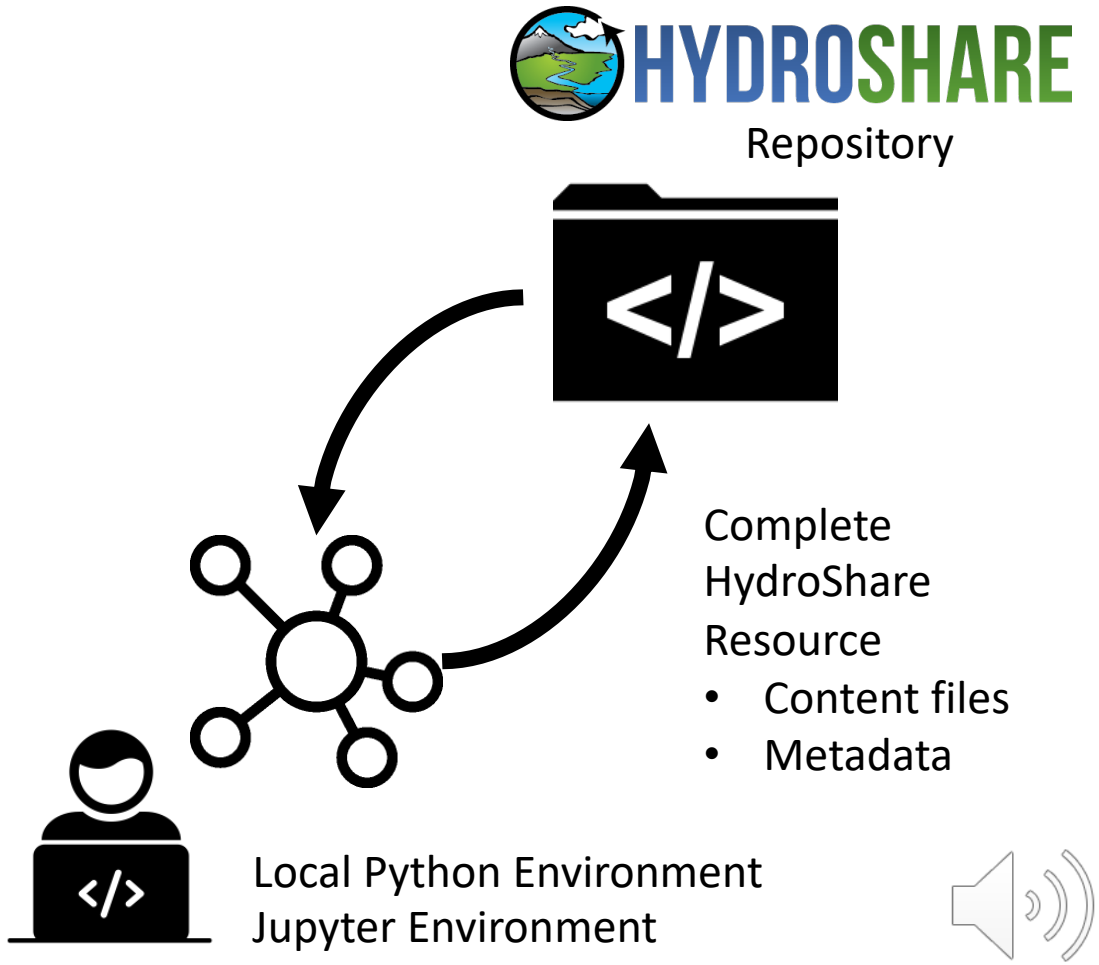
Tools for accessing and interacting with those repositories

A Python representation of the data retrieved that can be operated on using existing data science tools



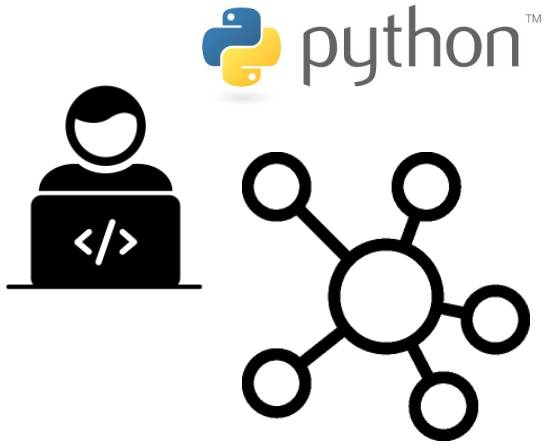
HydroShare Python Client (hsclient)

- A set of Python functions for interacting with the HydroShare repository
- (Object) Structure of HydroShare resources is specified in the OAI-ORE RDF/XML resource map documents
- hsclient translates this structure to a Python object representation
  - Read the structure and metadata of a resource into Python objects
  - Manipulate it in your Python environment (local or Jupyter)
  - Save that structure back to HydroShare
  - Modify RDF/XML outside of HydroShare and send those files back to be ingested



A flexible water-data science object data model (hsmodels)

- Extending the HydroShare Resource Data Model to Python analysis environments
- Maps HydroShare's resource metadata to a set of Python objects (classes) defined using pydantic models
- Maps common water-related data types (HydroShare content types) to performant data structures within Python
- Load and stage data for visualization/analysis using common Python tools (pandas, matplotlib, etc.)

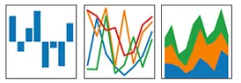


- Object Data Model
- Performant objects for visualization and analysis

Existing Scientific Python Packages

pandas

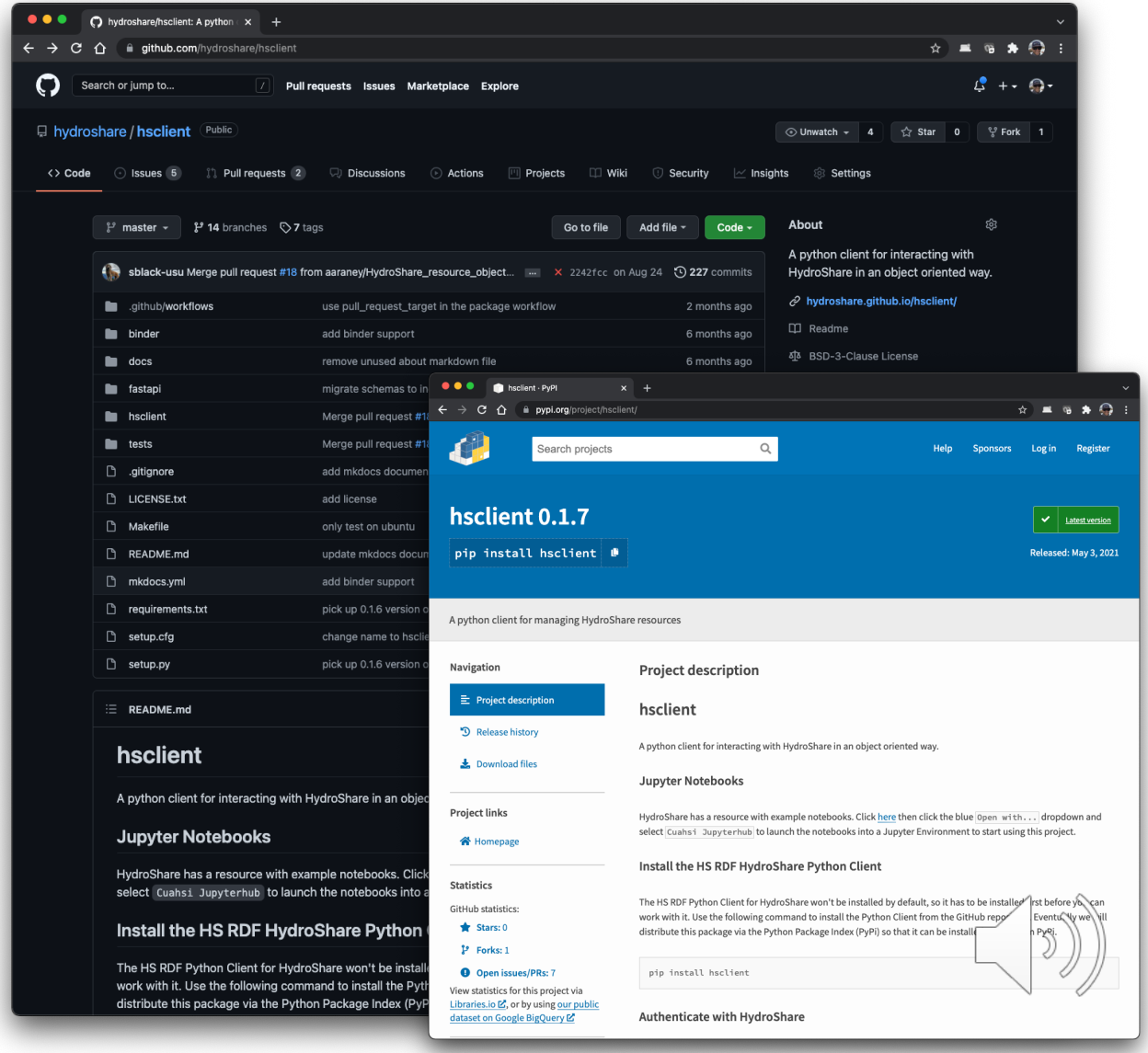
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



. . .

HydroShare Python Client ‘hsclient’ package

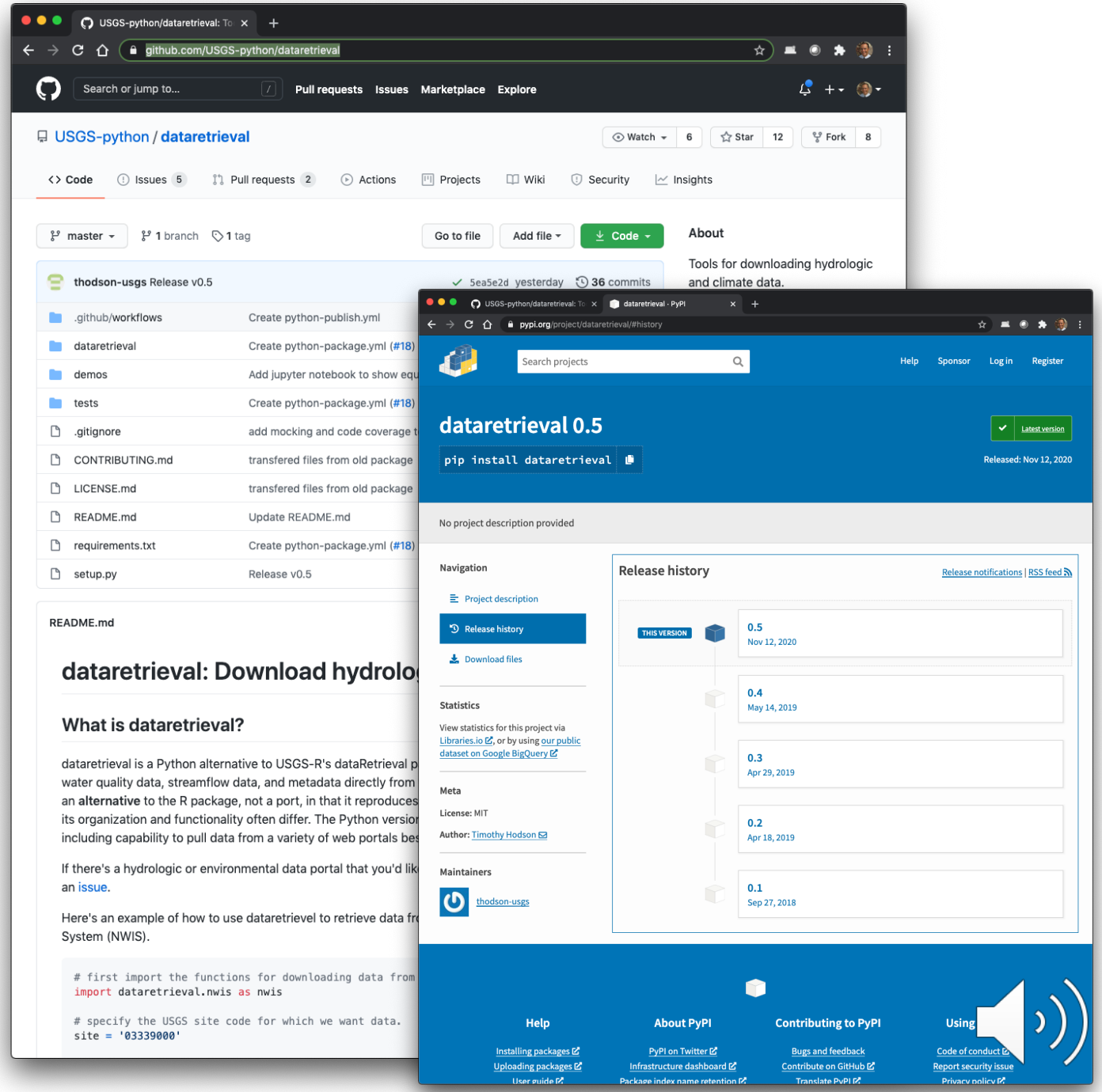
- A set of Python functions for interacting with HydroShare
  - Resource creation/editing
  - Interact with resources in an interactive, object-oriented way
  - Integrate HydroShare resources into data science workflows
  - Reduce the time required to get data for analysis and then save results
- Example Jupyter Notebooks: https://www.hydroshare.org/resource/7561aa12fd824ebb8edbee05af19b910/
- GitHub Repository: https://github.com/hydroshare/hsclient



USGS dataretrieval Python package

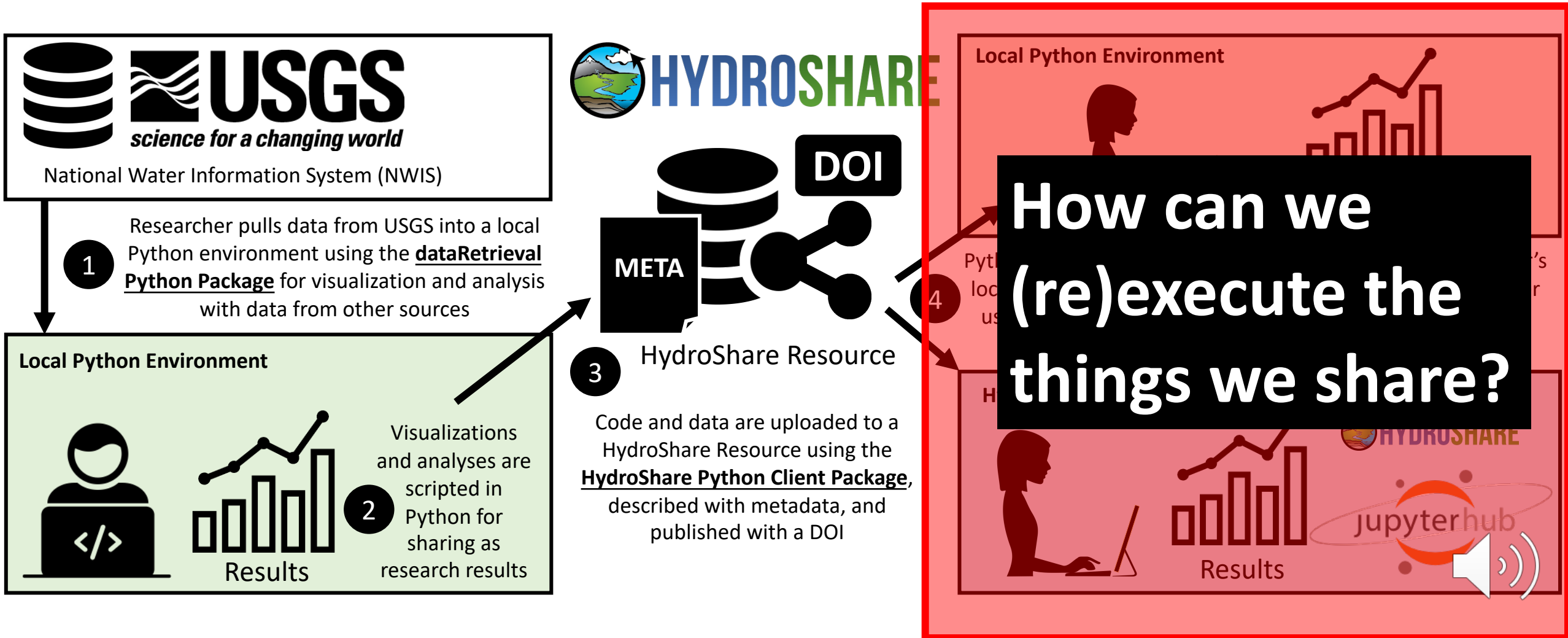
- Python mirror of the R dataRetrieval tool
- Currently has most of the same functions
- Very similar results
- Collaborating with Timothy Hodson at USGS
- Example Jupyter Notebooks: https://www.hydroshare.org/resource/c97c32ecf59b4dff90ef013030c54264/

https://github.com/USGS-python/dataretrieval



Connecting Visualization and Analysis with an Online Repository

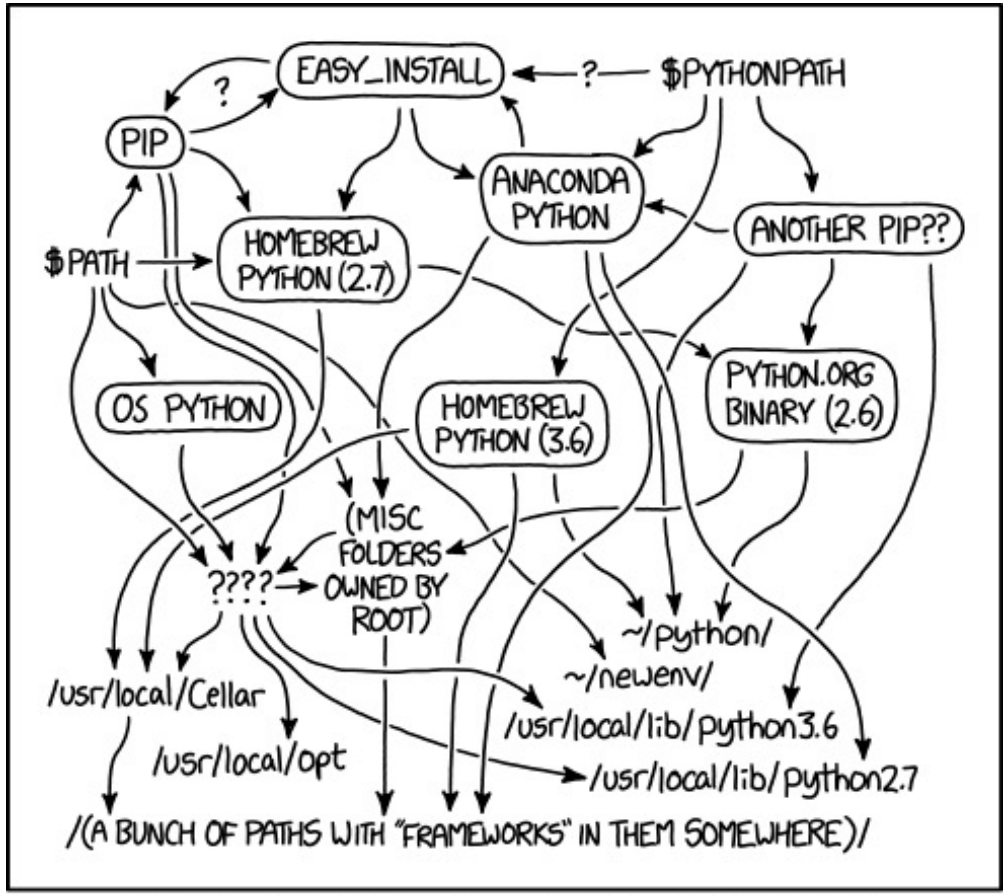
- Better enabling collaborative data science workflows and reproducibility



One Option: Local Python Environment

- Set up a local environment
- Get the Python version right
- Install the right versions of all of the packages
- Cross your fingers and hope it will run . . .

- Virtual environments can help, but this can still be challenging



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

https://xkcd.com/1987/

Collaborative and interactive computing for water-data scientists

CUAHSI JupyterHub – Google Cloud

- Supports “unlimited” users ($)
- Capable of creating classroom/workshop specific instances
- Completely customizable and uses the latest JH software

CyberGIS-Jupyter for Water

- More available compute resources

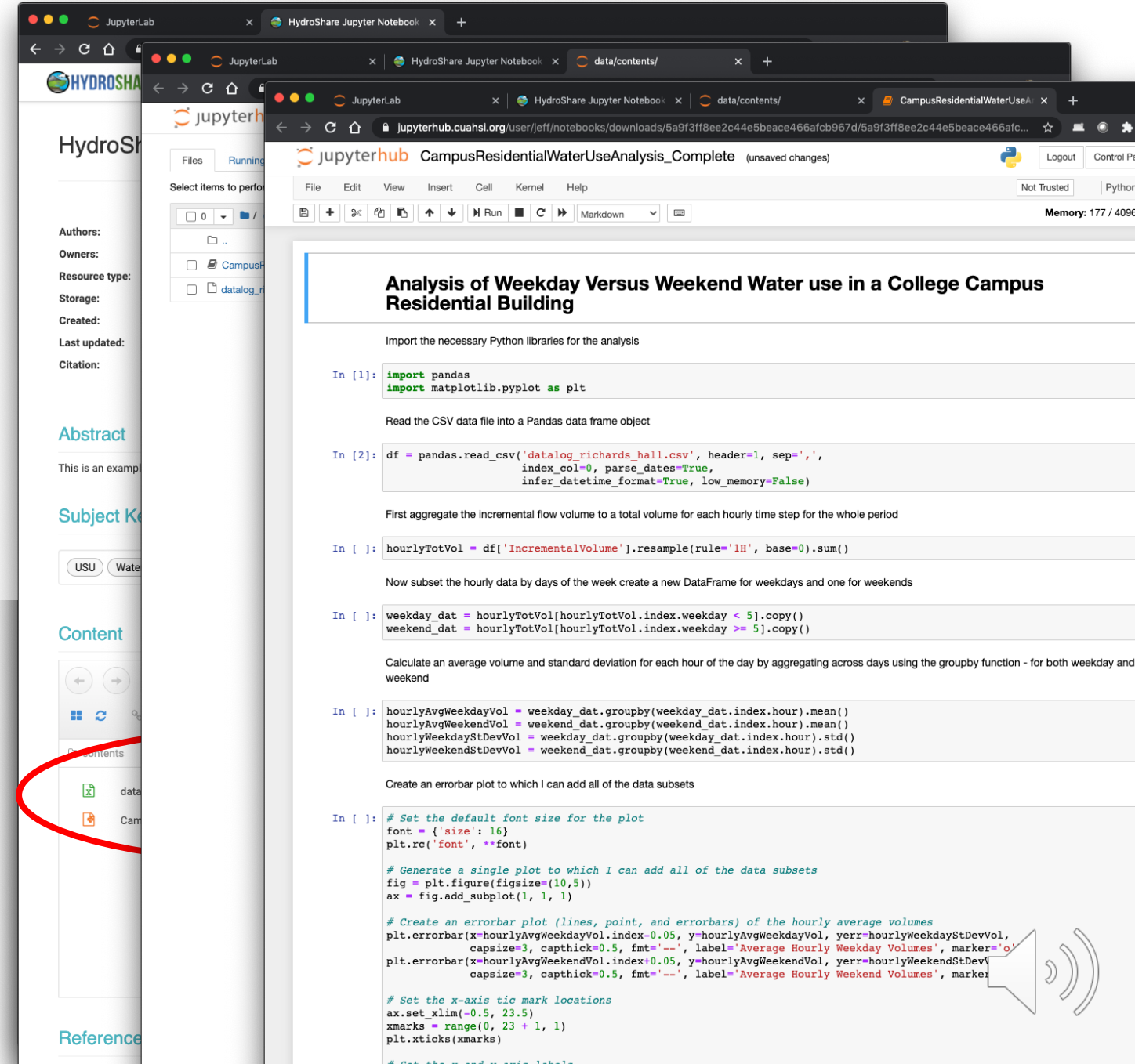
MATLAB Online

- 50 concurrent users
- Livescript support and m-file
- 20+ toolboxes

Slide from Tony Castronova at CUAHSI

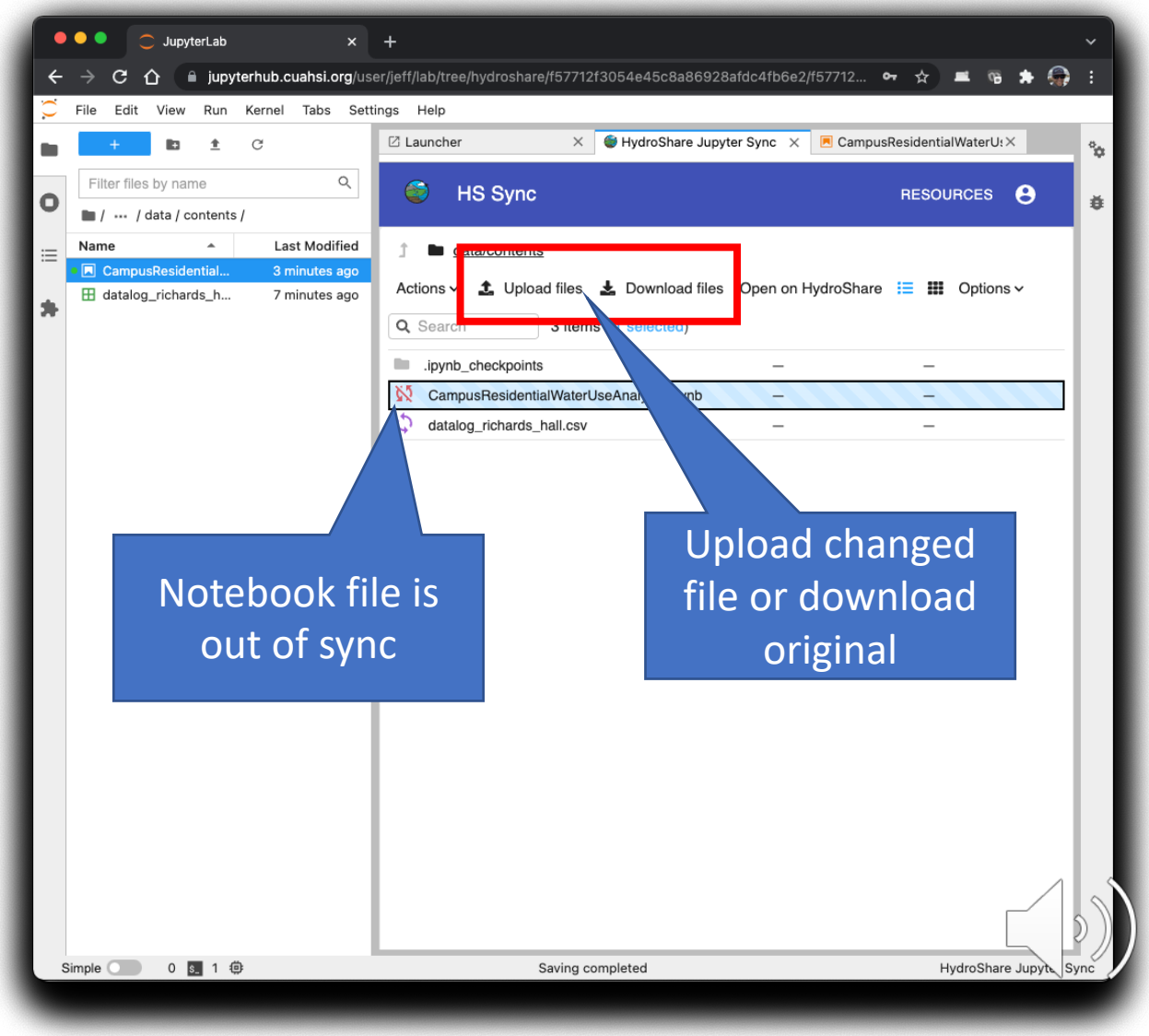
Creating and Sharing Reproducible Analyses

- Reproducible analyses: Sharing data and code together in a repository
- Linking repositories with computational environments
- Repositories as a gateway to high performance computing and cloud services



CUAHSI JupyterSync App using hsclient

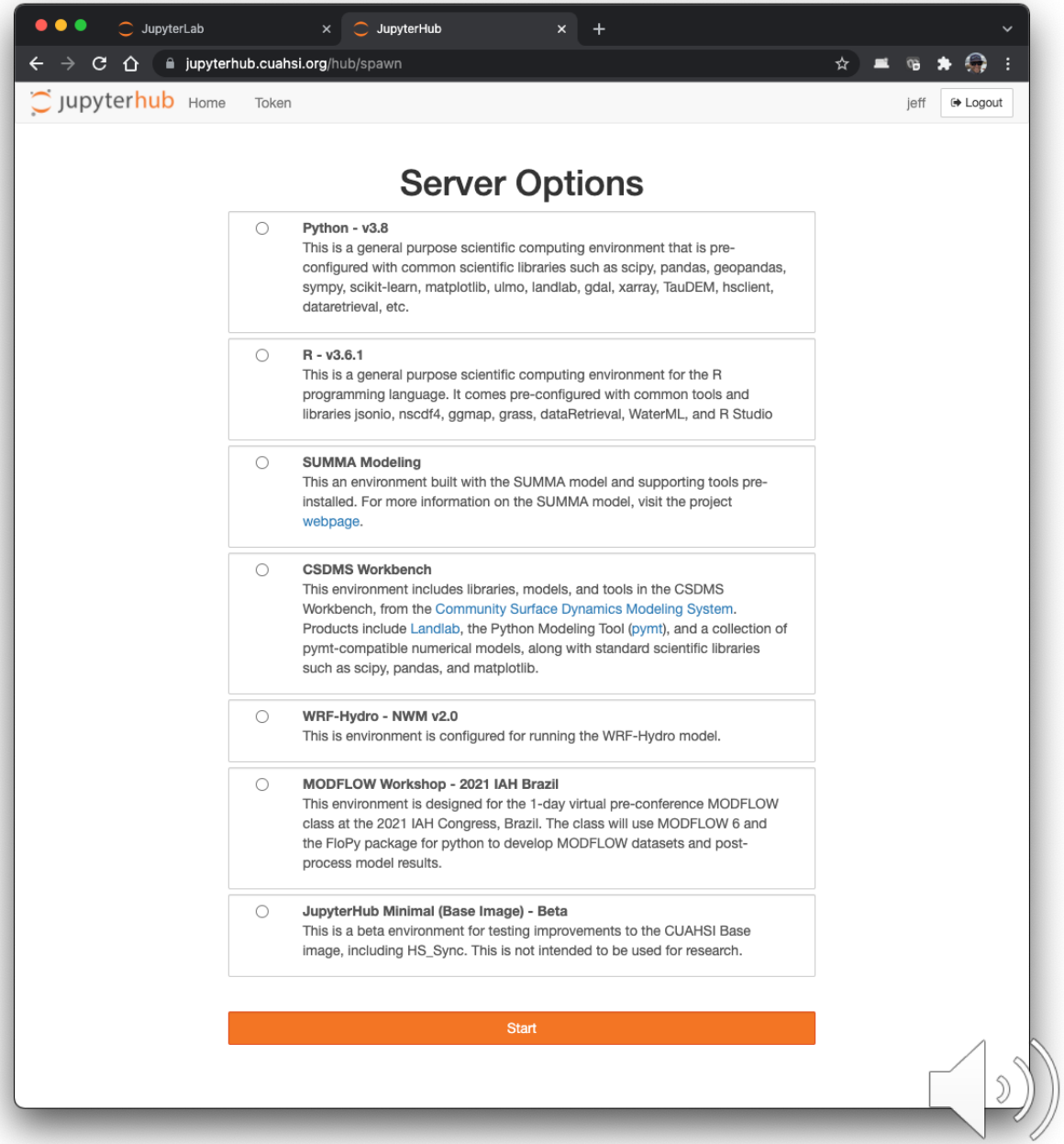
1. Launch CUAHSI Jupyterhub (jupyterhub.cuahsi.org)
2. Launch the Jupyter Sync App
3. Choose a HydroShare resource to work with
4. Select files to download to Jupyter environment
5. Open a file to edit or execute
6. Make changes to the file in the Jupyter Environment
7. Upload changed file to HydroShare or download original file to replace



Work by Tony Castronova and Austin Raney and students from Olin College of Engineering

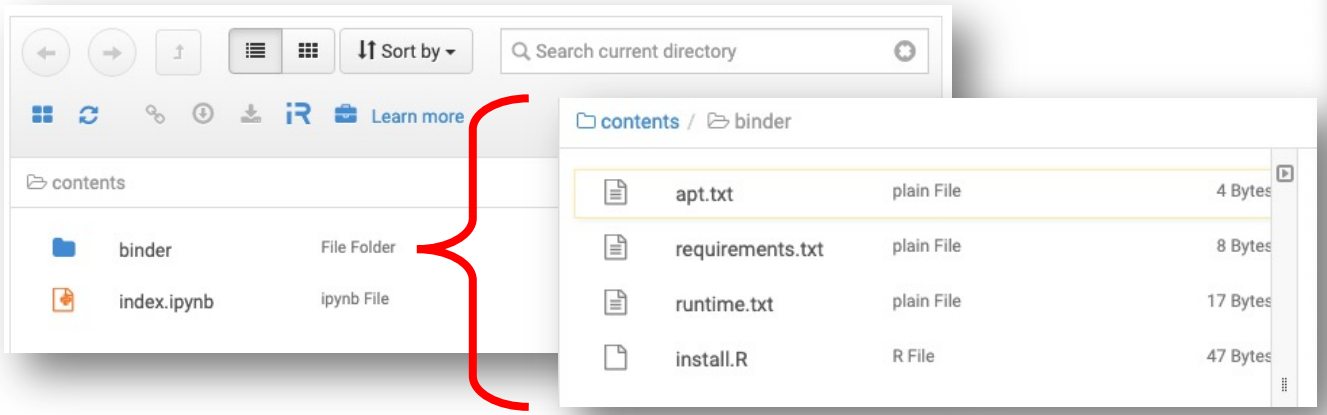
HydroShare's Linked JupyterHub Environments

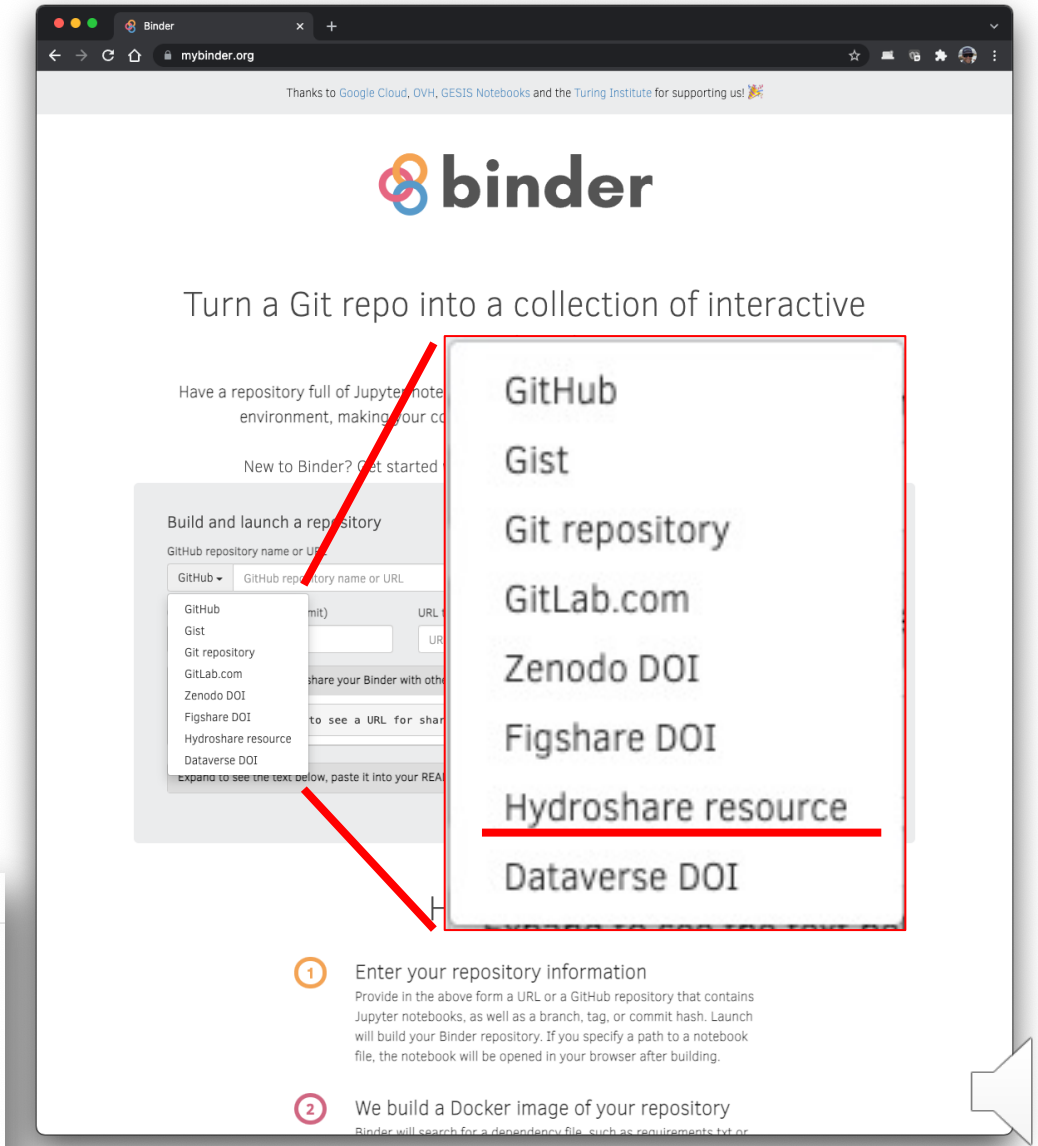
- Better because I don't have to set up the environment
- Some nice tools for interacting with HydroShare
- But, I still need an environment
- Some potential limitations:
  - Software dependencies
  - Legacy code
  - Long run times
  - Complicated and large input/output files



Improving Reproducibility with Binder

- Custom computing environments
- Free, but limited resources
- Can lower the barrier of entry for water scientists
- Integrated with HydroShare
- Users can start with a HydroShare base image





Slide from Tony Castronova at CUAHSI

1664061
1931297
1931278

Questions?

Jeffery S. Horsburgh

jeff.horsburgh@usu.edu

Utah Water Research Laboratory
UtahStateUniversity

CUAHSI
universities allied for water research

HYDROSHARE