I-GUIDE
Institute for Geospatial Understanding through an Integrative Discovery Environment

The I-GUIDE Cyberinfrastructure Platform: Supporting Open, Reproducible Convergence Science at Scale

Jeffery S. Horsburgh

Utah Water Research Laboratory

Utah State University

and

The I-GUIDE Core Cyberinfrastructure Capabilities and Services Team

Utah Water Research Laboratory
UtahStateUniversity

Support: 2118329

Core CI Capabilities & Services Team

Anand Padmanabhan

University of Illinois Urbana-Champaign

Carol (Xiaohui) Song

Purdue University

Objective: Integrate distributed geospatial data capabilities and advanced CI to form a composable and open I-GUIDE platform to accelerate scientific workflows and support education and workforce development as well as broader community engagement

Furqan Baig

University of Illinois Urbana-Champaign

Anthony Castronova

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

Irene Garousi-Nejad

CUAHSI

Jeff Horsburgh

Utah State University

Rajesh Kalyanam

Purdue University

Erick Li

University of Illinois Urbana-Champaign

Josh Lieberman

The Open Geospatial Consortium

Alexander Michels

University of Illinois Urbana-Champaign

Steven Manson

University of Minnesota Twin Cities

Mohan Ramamurthy

University Corporation for Atmospheric Research (UCAR)

Ingo Simonis

The Open Geospatial Consortium

Noah Samuel Oller Smith

Purdue University

Sina Taghavikish

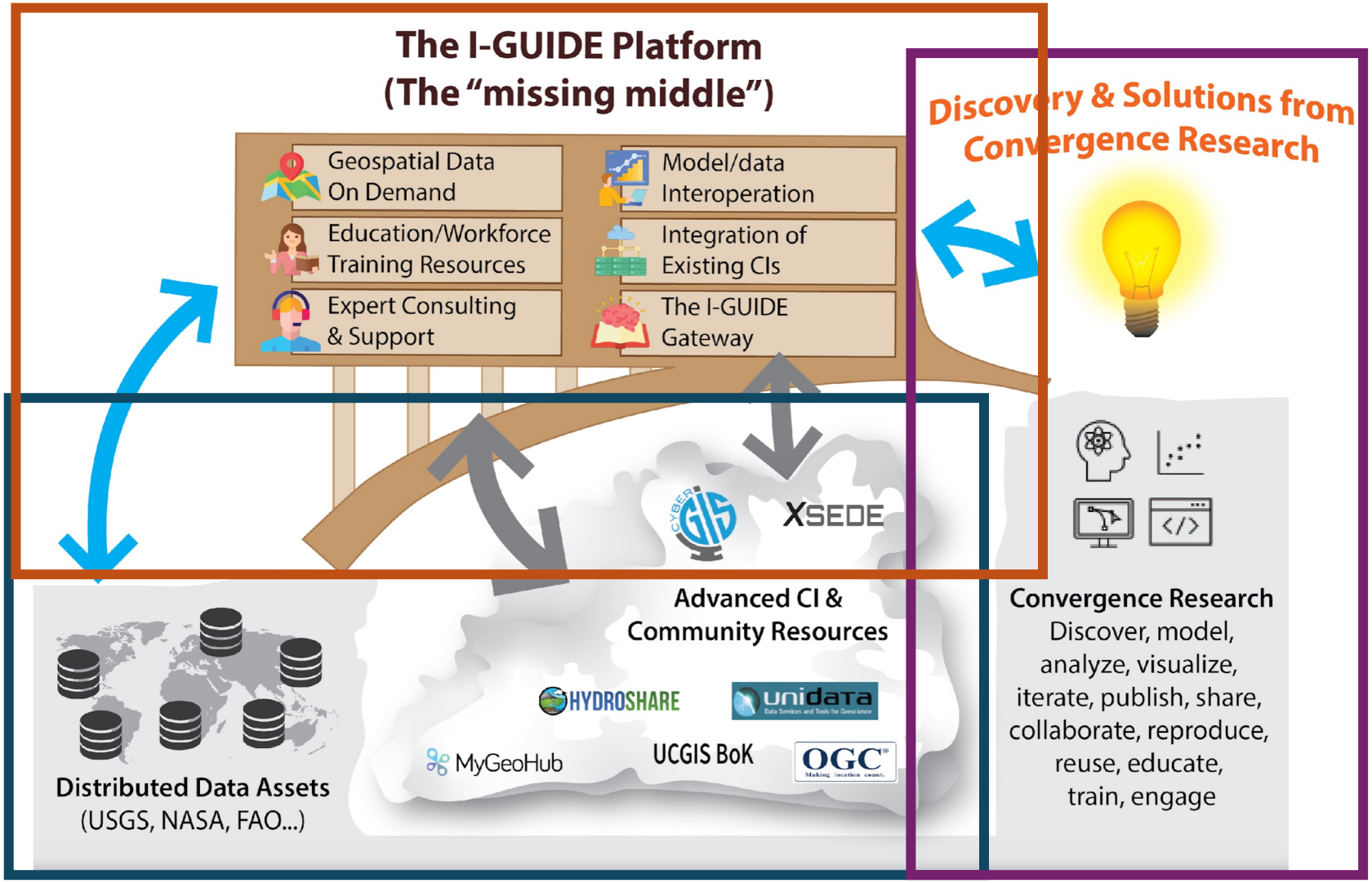
The Open Geospatial Consortium

Lan Zhao

Purdue University

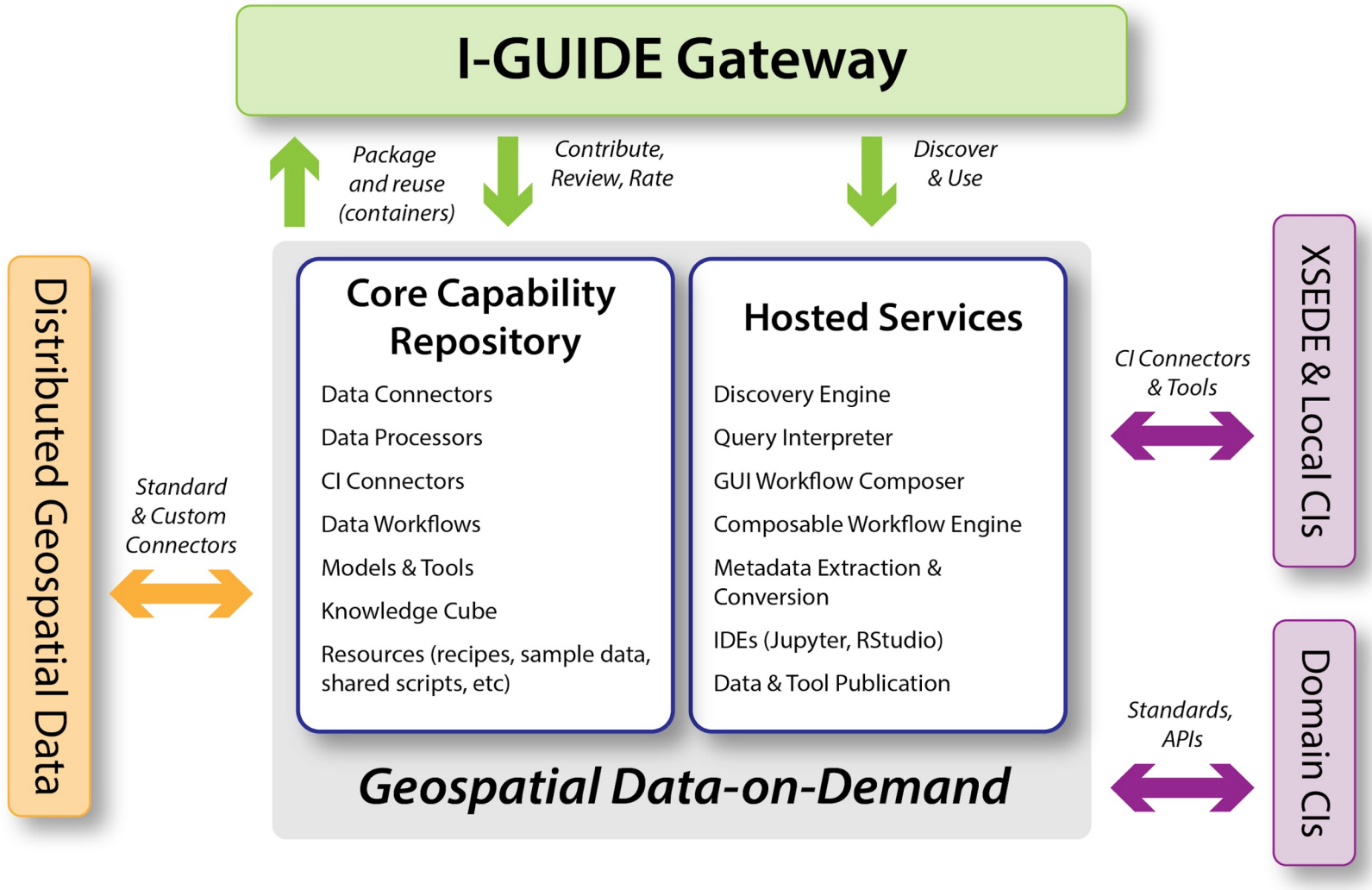
What is the I-GUIDE "Platform"?

- Existing distributed datasets and community cyberinfrastructure
- We want to get to convergence research
- The platform is the bridging infrastructure between the two



Confusing terminology…

- The “buzz words”
  - I-GUIDE Platform?
  - I-GUIDE Gateway?
  - Geospatial data-on-demand?
- Confusing – how are these things different than the “Platform”?



Some challenges in communication

- Convergence Science Catalyst (CSC) teams want to use the cyberinfrastructure, but they don't know what it is, how to use it, or how it improves their life
- The Core Cyberinfrastructure Capabilities and Services Team wants to build effective cyberinfrastructure, but we don't fully understand the science use cases

We need some common ground to move forward together!

The purpose of this VCO

- Describe the I-GUIDE Platform in a bit more detail
- Provide our vision and approach for the Platform
- Describe the principles that are guiding our development
- Describe some of the challenges we are working to address

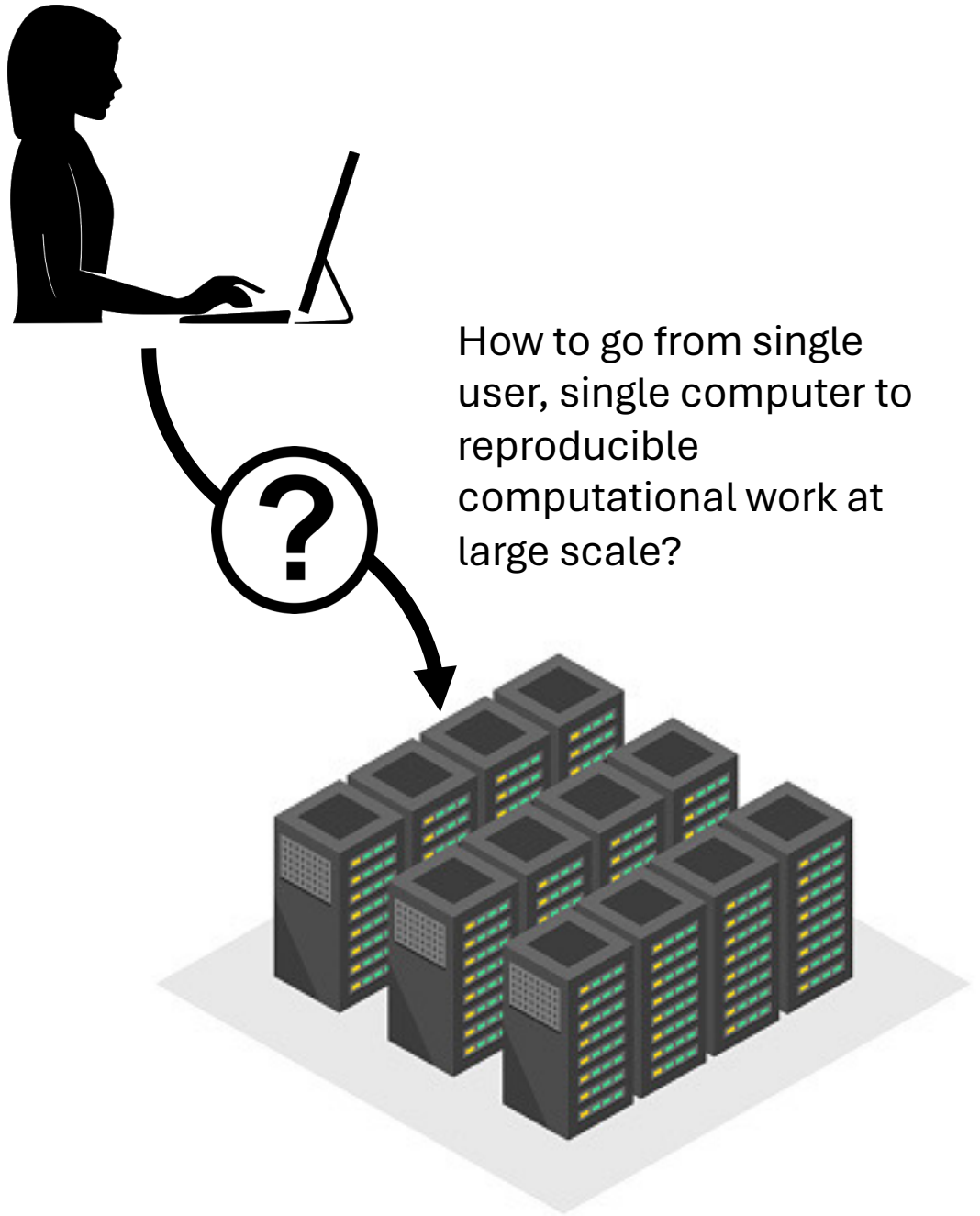
I-GUIDE Platform Vision

From the perspective of the I-GUIDE cyberinfrastructure team, we want to:

Support open, reproducible convergence science at scale using distributed geospatial datasets that may be too large to store in a single location. The I-GUIDE Platform aims to help reduce the time to science and enable more open, accessible, and reproducible computational workflows.

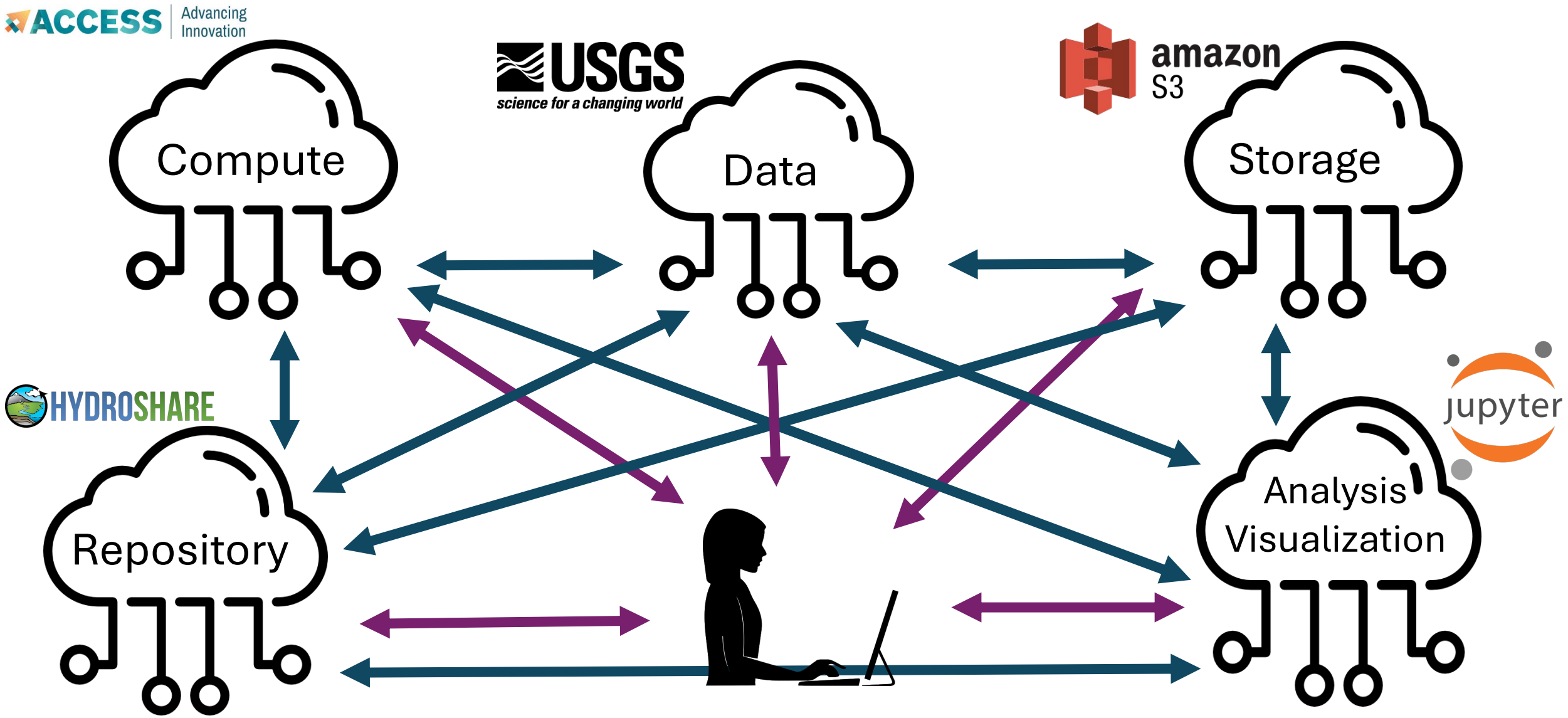
“At scale”

- Geospatial convergence science faces multiple scale-related challenges:
  - **Data**: How to efficiently store, manage, and stage data for modeling and analysis? How to share results?
  - **Computation**: How to get access to adequate computational resources?
  - **Connection**: How to efficiently connect data and computation?
  - **Reproducibility**: How to make all of these accessible for reproducing computational work?



How to go from single user, single computer to reproducible computational work at large scale?

More services are now cloud based
Composing workflows that bridge clouds is hard



Our approach to building the Platform

- Use existing cyberinfrastructure components where we can
- Find ways to facilitate “composing” workflows that use those components
- Build missing pieces where necessary
- Explore what we can do by engaging with use cases coming out of the convergence science teams

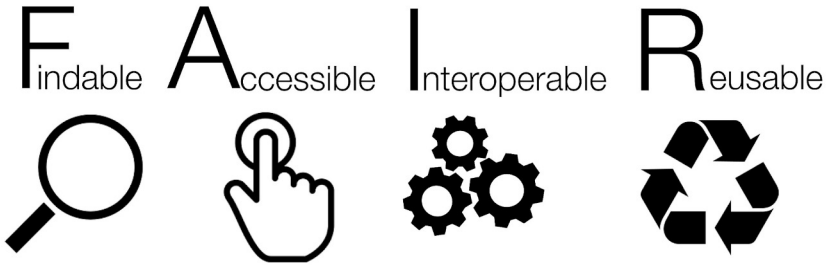
Guiding principles for working together

- We need some principles to guide our cyberinfrastructure development
- To find the common ground, these also need to be shared by I-GUIDE's convergence science teams!

1. I-GUIDE convergence science should be open - meaning shared, transparent, and accessible.
2. Inputs and outputs of analyses along with workflows used to generate them should be shared in open repositories or openly accessible storage locations.
3. Computational workflows should be defined in a way that they can be documented, shared, and reproduced/repeated using accessible computational resources.

Applying these principles will likely require some changes to the way we work.

What is driving these principles?



- Data (broadly defined) are primary research products!

- **Findable**: Data have sufficient metadata and a unique, persistent identifier making data discoverable on the Web
- **Accessible**: Metadata and data are understandable to humans and machines and are available via a trusted repository
- **Interoperable**: Metadata use formal community standards
- **Reusable**: Data have clear metadata, usage license, and information about provenance

The extent to which data are FAIR affects their value and extent of reuse!

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018, https://doi.org/10.1038/sdata.2016.18.

Funder requirements

- NSF's general Data Sharing Policy[1]:

"NSF-funded investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF awards."

- From NSF's Division of Earth Sciences Data and Sample Policy[2]:

"Possible types of "data" to be addressed in the DMP include, but are not limited to: observational, experimental, analytical, and model outputs; derived and compiled datasets; software and code; educational materials; and any other relevant digital products resulting from the project."

[1] https://new.nsf.gov/funding/data-management-plan#nsfs-data-sharing-policy-1c8
[2] https://www.nsf.gov/geo/geo-data-policies/ear/ear-data-policy-jul2023.pdf

Publisher requirements

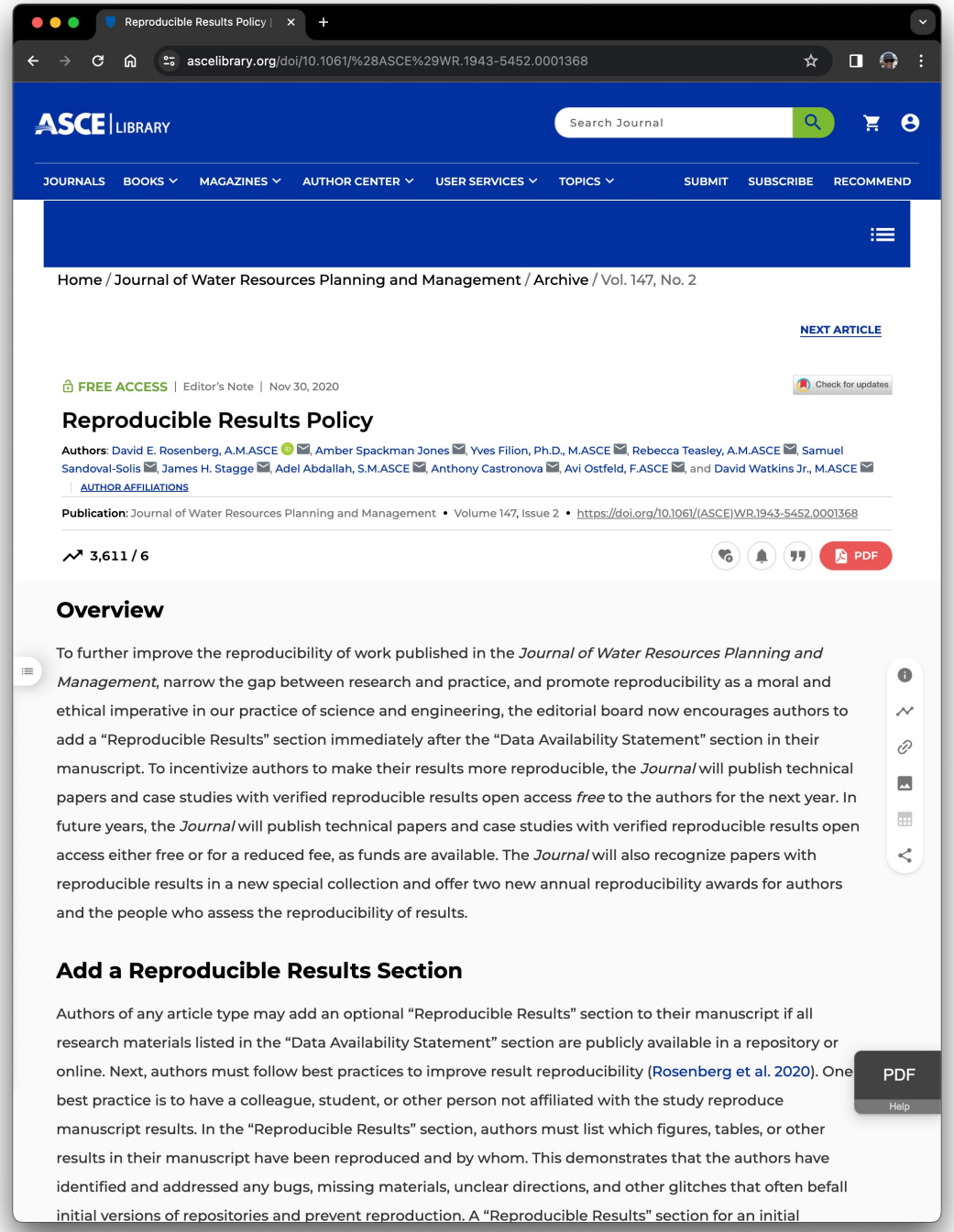
"AGU requires that the underlying data and/or software needed to understand, evaluate, and build upon the reported research be available at the time of peer review and publication. Additionally, authors should make available software that has a significant impact on the research. This entails:

1. *Depositing the data and software in a community accepted, trusted repository, as appropriate, and preferably with a DOI*
2. *Including an Availability Statement as a separate paragraph in the Open Research section explaining to the reader where and how to access the data and software*
3. *And including citation(s) to the deposited data and software, in the Reference Section*[1]*."*

[1] https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Data-and-Software-for-Authors

Publisher incentives

To incentivize authors to make their results more reproducible, the Journal will **publish technical papers** and case studies with verified reproducible results **open access free** to the authors for the next year.”



https://doi.org/10.1061/(ASCE)WR.1943-5452.0001368

From the I-GUIDE Data Management Plan

"The institute itself **will strictly adhere to NSF rules**, and share with other researchers, at no additional cost and within a reasonable time, the primary data, software and other related materials created or gathered during the project execution."

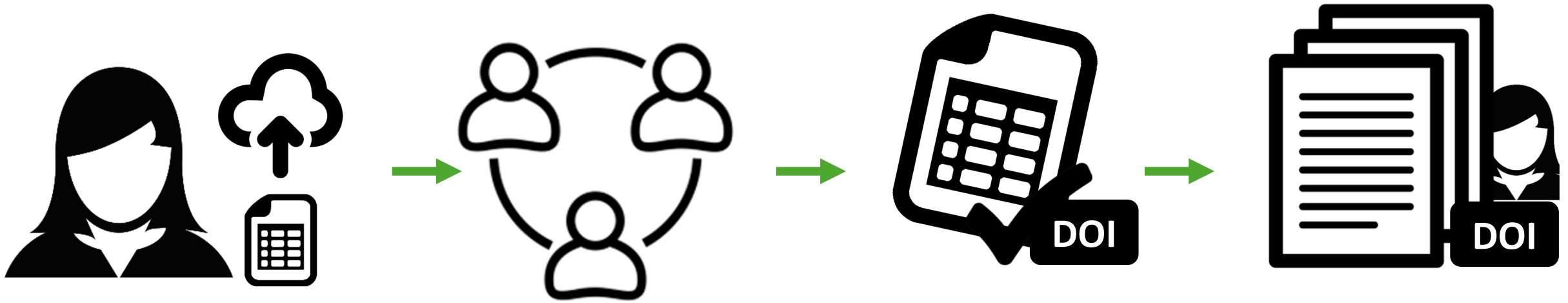
"We will archive digital data used in our experiments, as well as the associated software and programs generated by the institute and make such materials available to pertinent communities. For example, Hydroshare, will be used for hosting hydrology related datasets and associated models."

"From the point of view of this institute, such dissemination of data is crucial to stimulate new advances as quickly as possible and to allow for prompt evaluation of the results by the pertinent communities."

I-GUIDE

What do we mean by “open”?

- An example workflow for a dataset product:



User creates dataset and uploads to reputable repository

Research team iterates collaboratively

Final version is published and assigned DOI

User cites dataset in paper

Result: Research artifacts (data and analyses) are openly shared, accessible, and citable.

Open repositories

- Choose a repository that provides:
  - Data registration/publication with a persistent, globally-unique identifier such as a digital object identifier (DOI)
  - Free access to the data
  - A landing page that provides metadata
  - Support for versioning

figshare
credit for all your research

DRYAD

NSIDC
National Snow and Ice Data Center

GitHub

HYDROSHARE

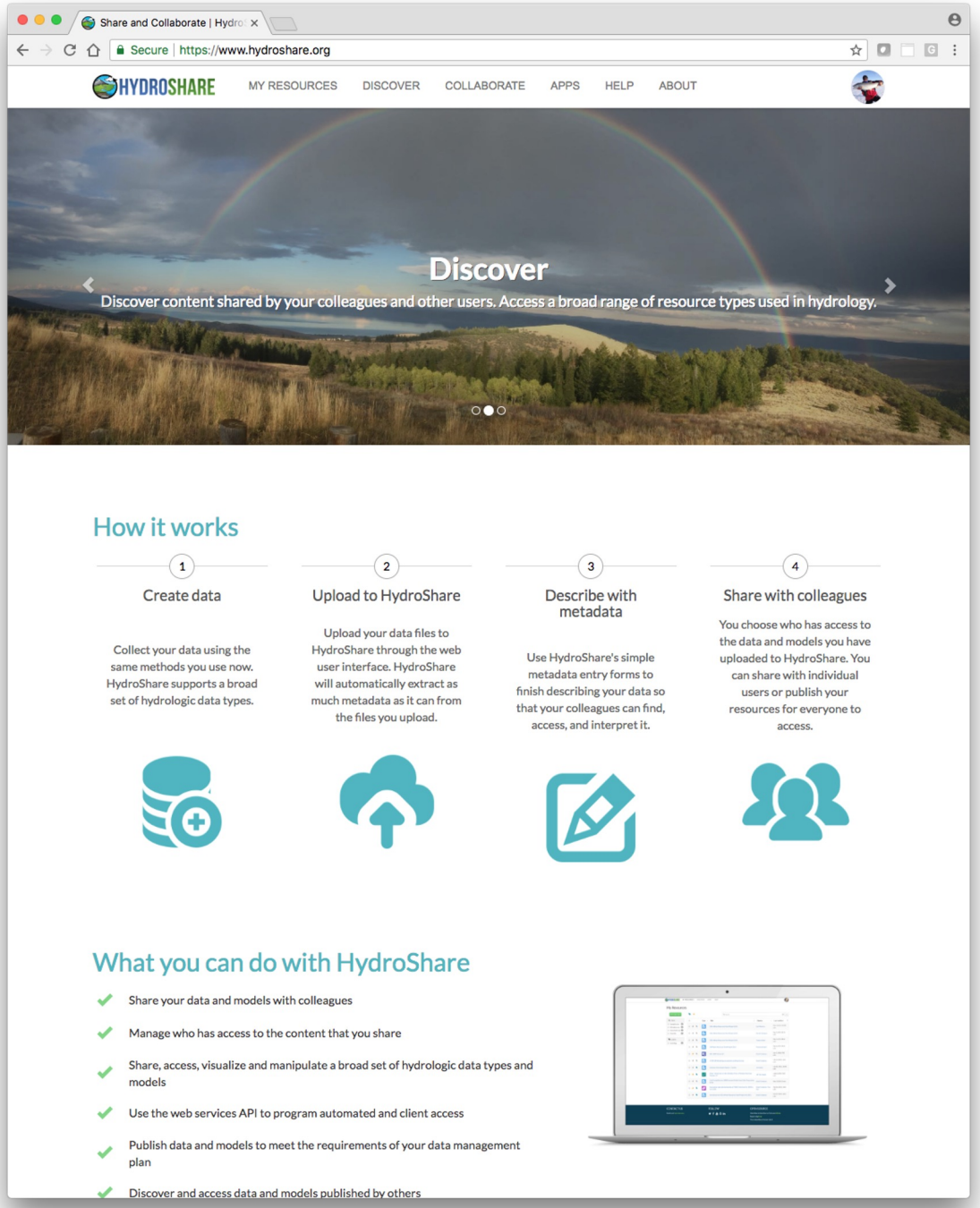
PANGAEA.
Data Publisher for Earth & Environmental Science

DataONE

HYDROSHARE

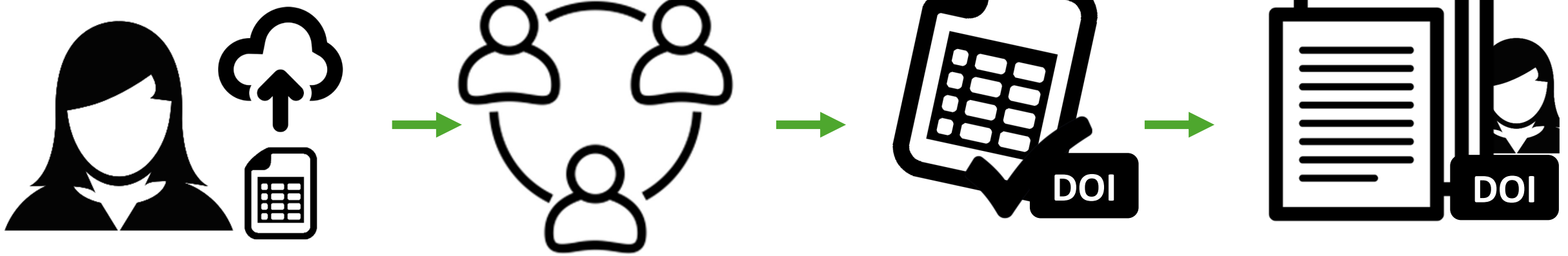
http://www.hydroshare.org

- Web-based repository and Hydrologic Information System
- 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
- Formal publication of data and models with linkages to literature



From our prior VCO about Hydroshare:

HYDROSHARE enables transparency



User creates dataset and uploads to Hydroshare

Research team iterates collaboratively

Final version is published and assigned DOI

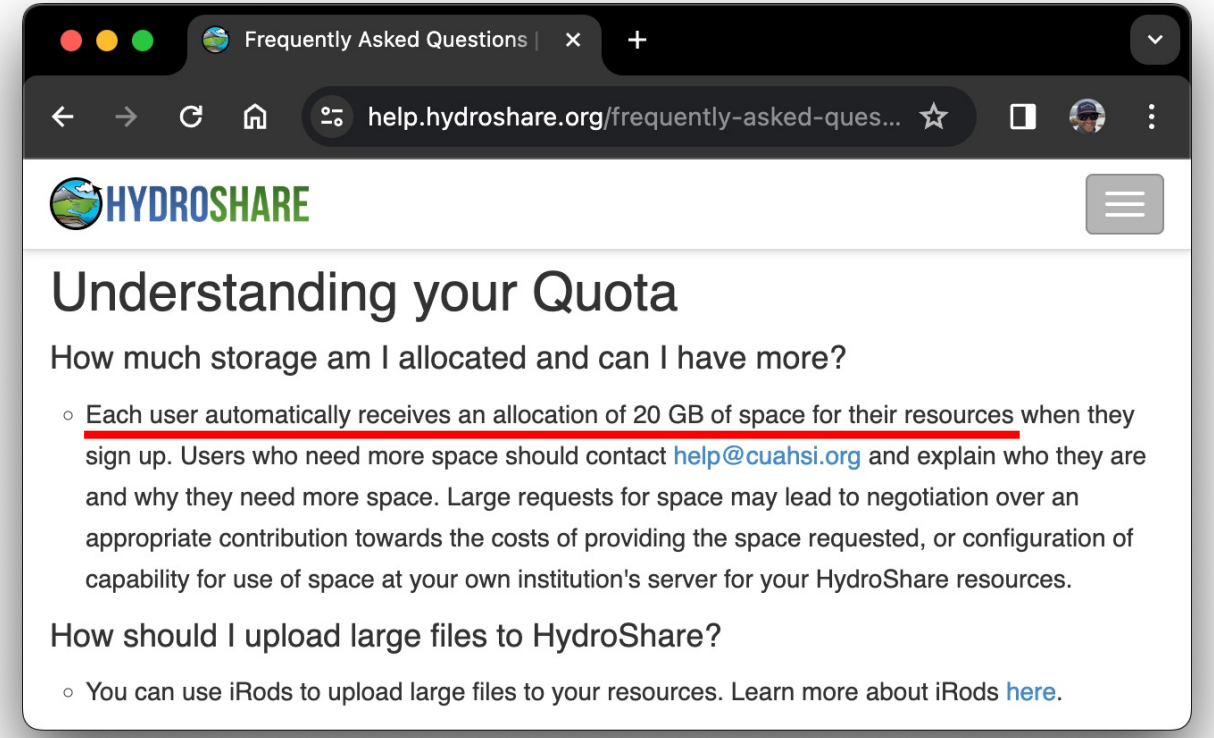
User cites dataset in paper

What happens when this breaks down?

Hey CUAHSI. My dataset is 1.5 TB in size. Can I share it in HydroShare?

What do I do now?

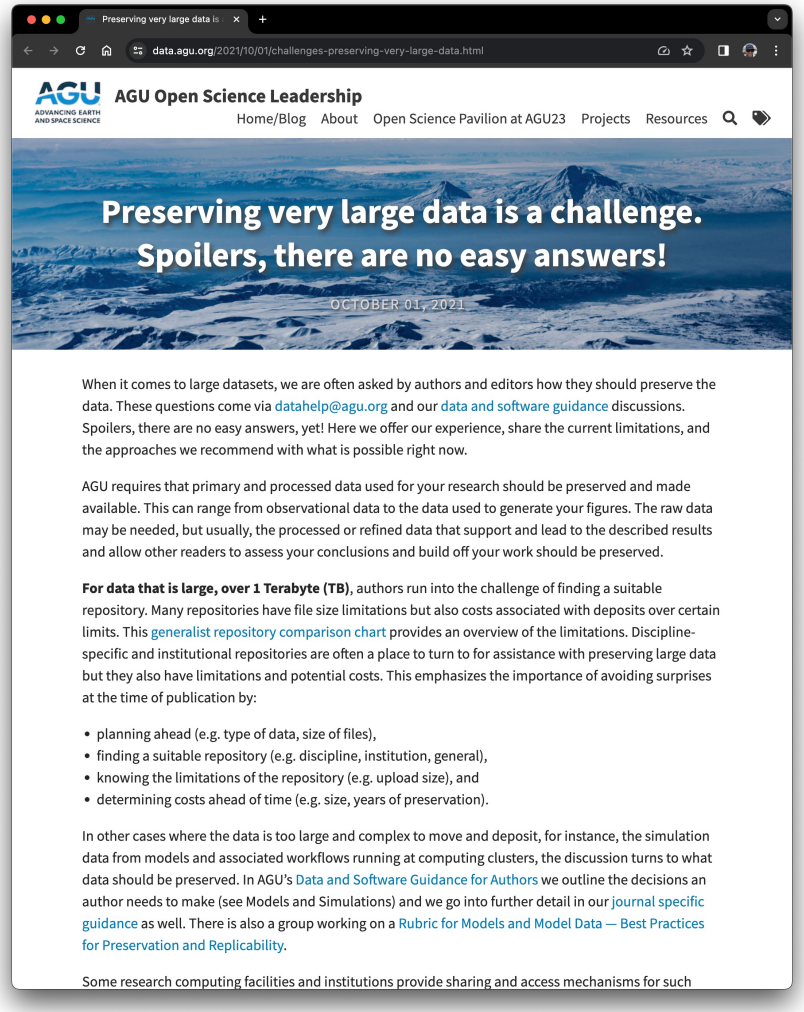
User creates dataset and wants to upload



Data constraints

- Limitations on data size represent a constraint imposed by current repository technology
- Some important datasets are too large to store in existing repositories like HydroShare
- Some models generate results that overwhelm repository storage
- As we scale up, we need other options!

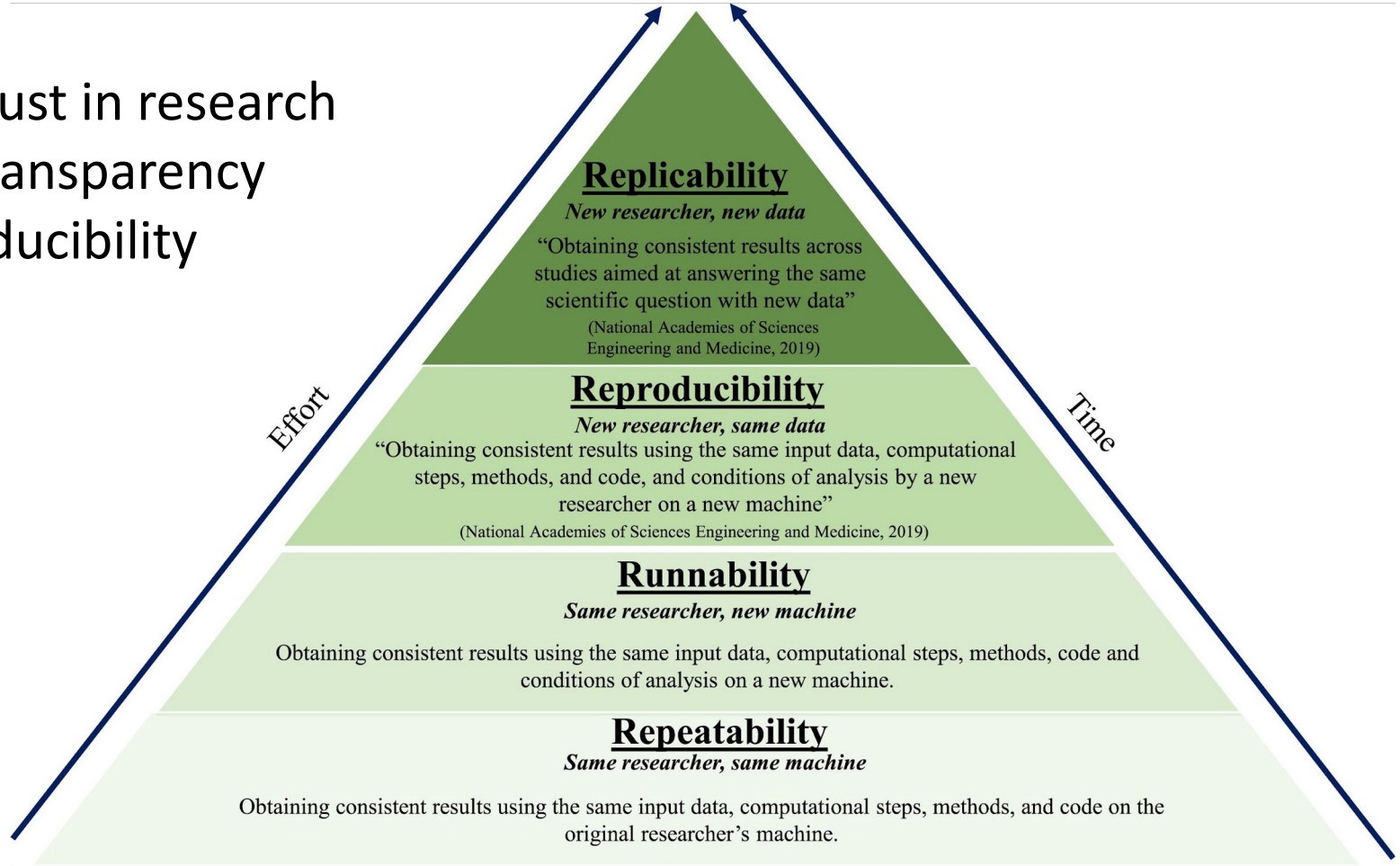
Despite policies, journals don't have all the answers either!



https://data.agu.org/2021/10/01/challenges-preserving-very-large-data.html

What does it mean to be reproducible?

Building trust in research requires transparency and reproducibility



Essawy, B., J. Goodall, D. Voce, M. Morsy, J. Sadler, Y. D. Choi, D. Tarboton and T. Malik (2020), A taxonomy for reproducible and replicable research in environmental modeling, *Environmental Modelling & Software*, 134:104753, https://doi.org/10.1016/j.envsoft.2020.104753

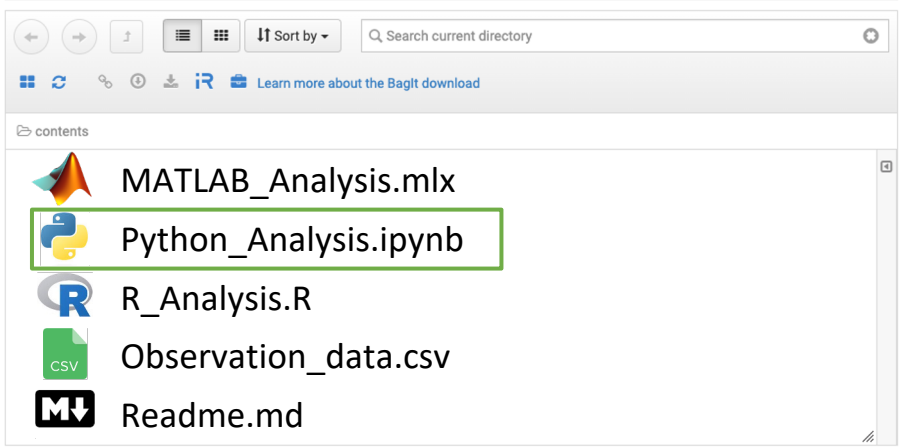
From our prior VCO about Hydroshare:

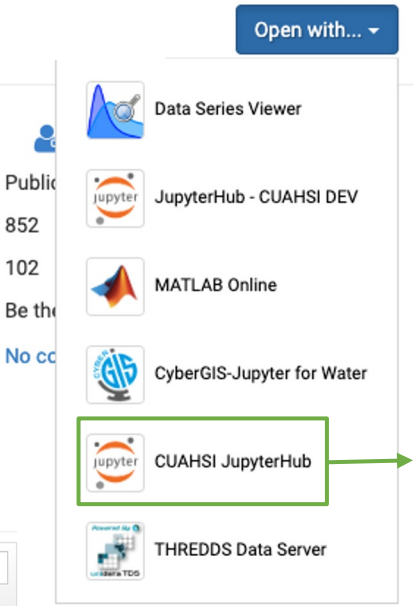
Cloud Computing Makes Resources Actionable

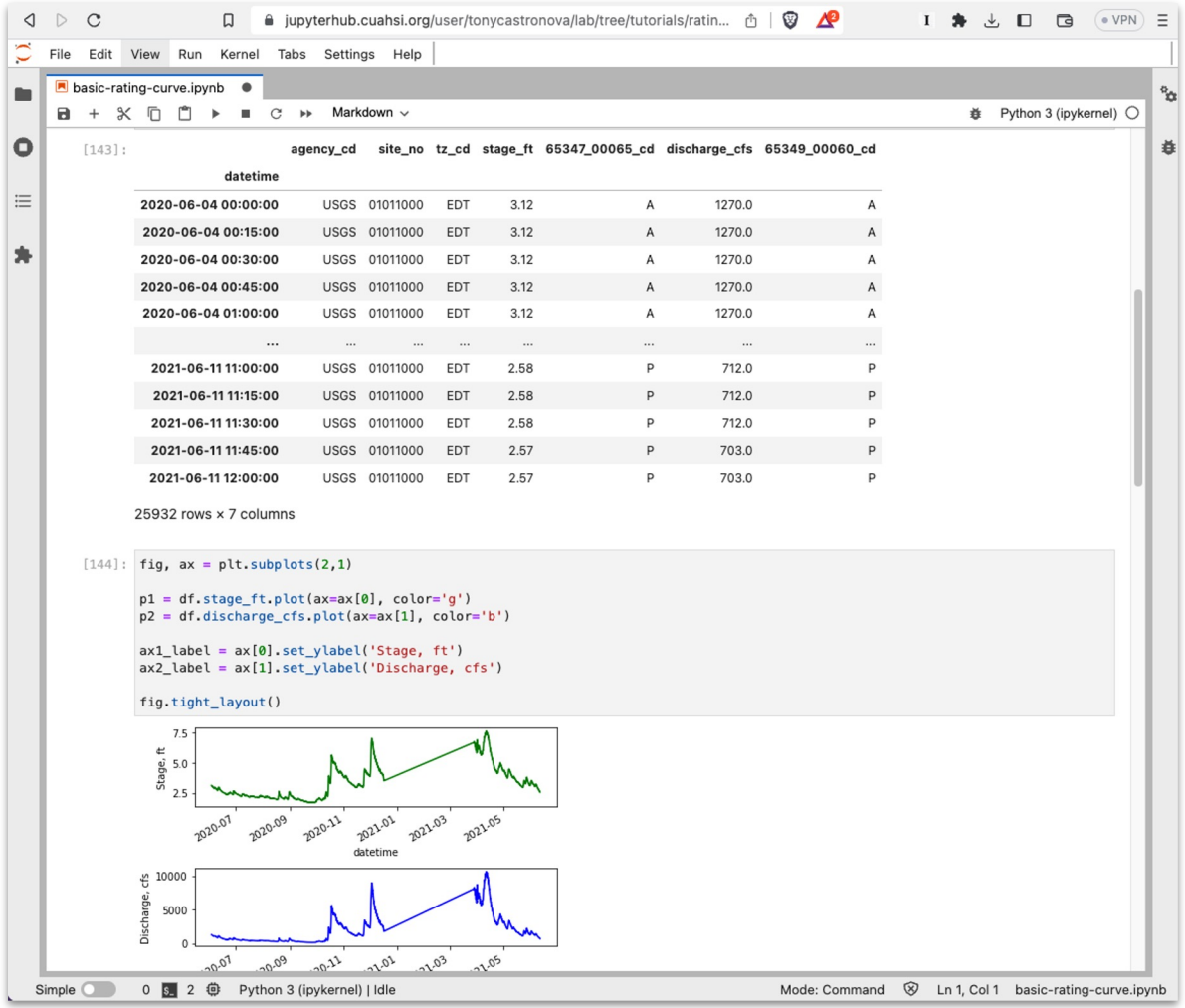
Sample Computational Resource

Authors:	Anthony Michael Castronova	**Sharing Status:**	Public
Owners:	Anthony M. Castronova	**Views:**	852
Type:	Resource	**Downloads:**	102
Storage:	The size of this resource is 2.4 MB	**+1 Votes:**	Be the
Created:	Jul 24, 2020 at 4:29 p.m.	**Comments:**	No cc
Last updated:	Feb 22, 2021 at 5:29 p.m. Anthony M. Castronova		
Citation:	See how to cite this resource		

Content



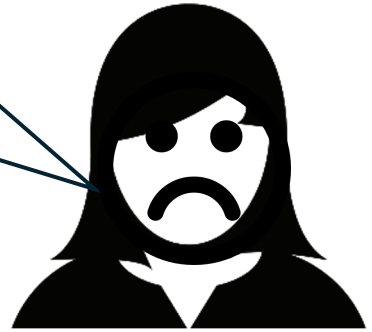




https://jupyterhub.cuahsi.org

What happens when this breaks down?

I want to create a Jupyter Notebook that enables people to run subsets of the National Water Model!



This workflow requires:

- Staging forcing data (~30 TB) and static data (~20 GB) somewhere
- Executing a domain subsetter program to get static and forcing data for a model domain
- Storing the subset of data for the model run (~5 GB)
- Executing the model on the resulting subset using a high-performance computational resource
- Storing the resulting model output data (~450 GB)
- Providing visualization and analysis of model outputs

- Rinse and repeat…

Computational constraints

- Availability and accessibility of computational resources represent constraints on the reproducibility of computational work
  - Scientists may not know which environment to use or how to use it
  - They may not have access to adequate computational environments
  - They may not be able to give others access to the environment they used
  - The computational environment they can access may not have convenient access to the data they need

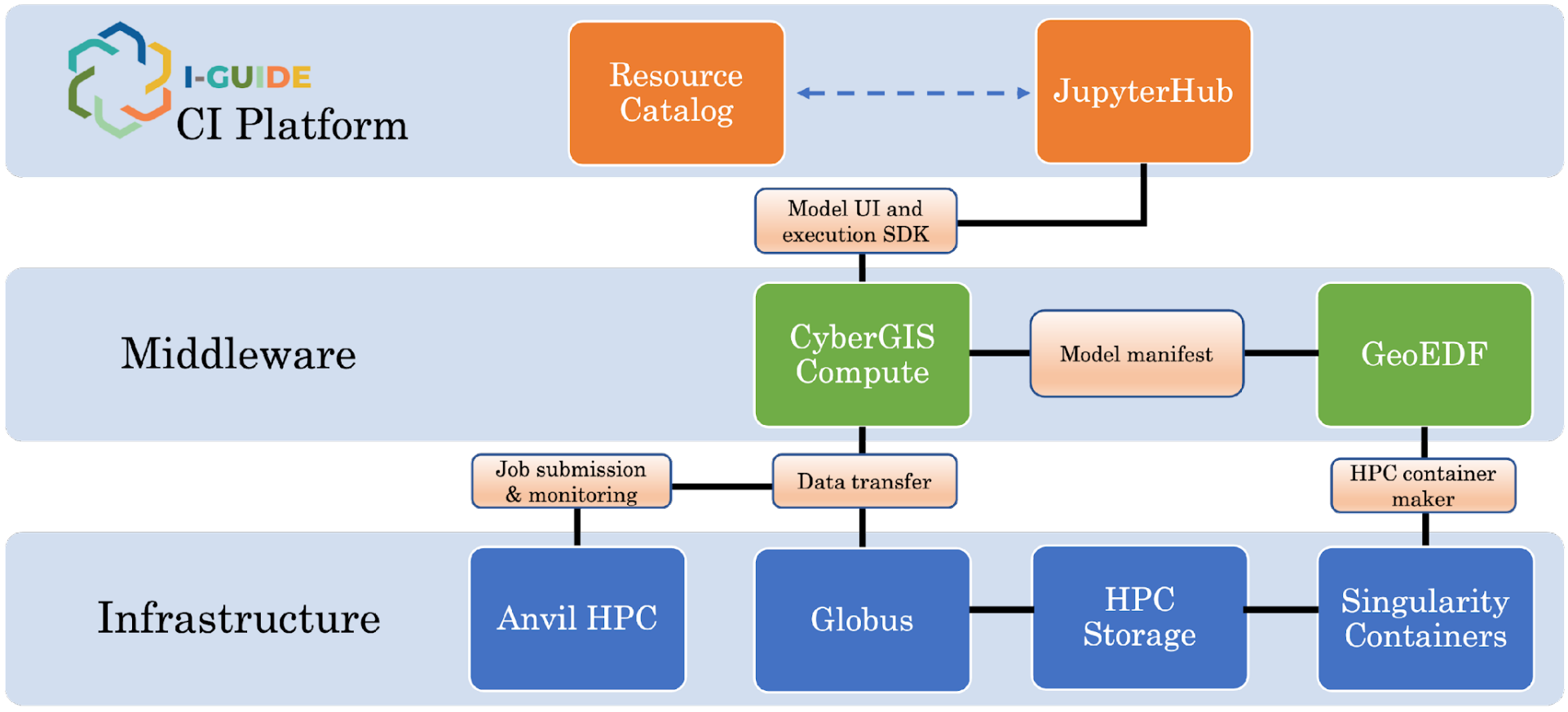
How is the I-GUIDE Platform addressing these constraints?

- General I-GUIDE Platform Functionality
  - **Data staging and preparation**: Where can we put large datasets to make them available for processing and modeling?
  - **Data exploration**: How can we interact with large datasets given that we can't download (or don't want to move) the whole thing?
  - **Data integration/wrangling**: How can we subset large datasets? Aggregate in space or time? Combine with other datasets?
  - **Model/workflow execution**: Where and how can we run models over large domains or workflows that are computationally intensive?
  - **Data/workflow/results sharing**: What do we do with the results when we are done? How can we give others access?

I-GUIDE Platform Design

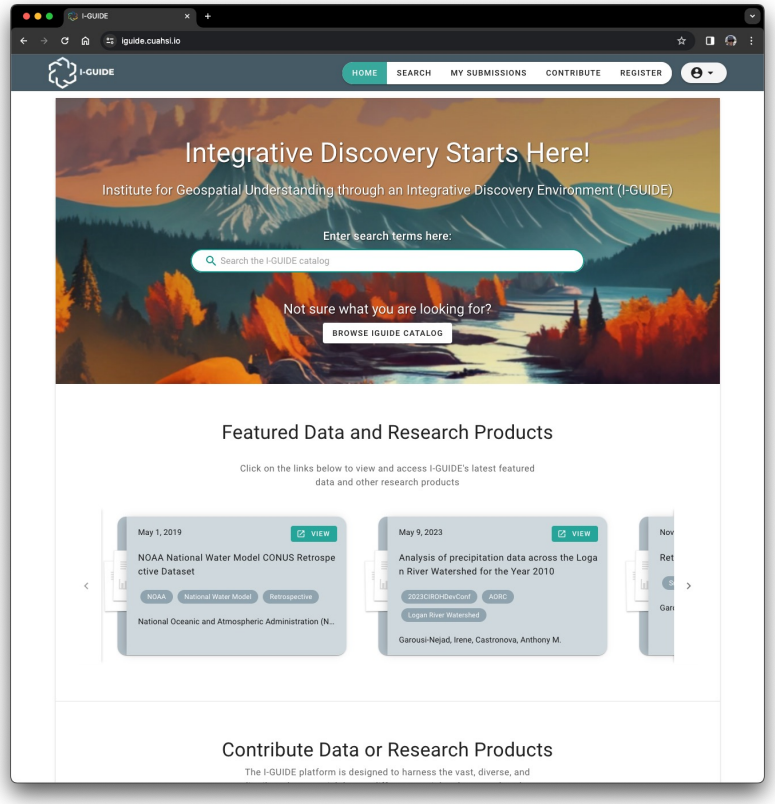
The entry points for people

The stuff you use, but don't really see



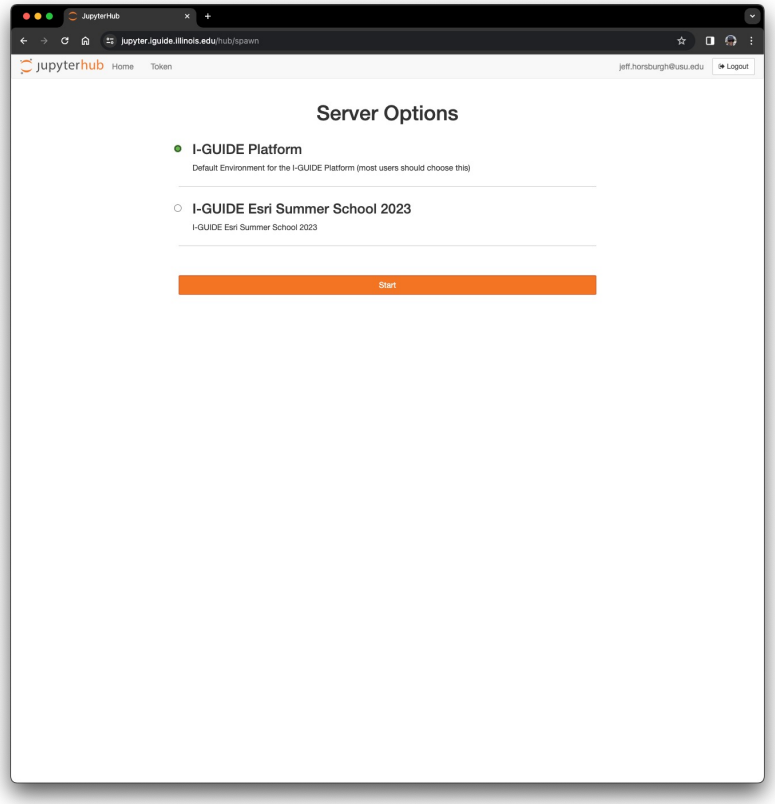
Two main user interfaces

I-GUIDE Catalog: Accessing shared data, workflows, etc.



Current Prototype: https://iguide.cuahsi.io

I-GUIDE JupyterHub(s): Multiple options for data exploration, integration, wrangling, model/workflow execution

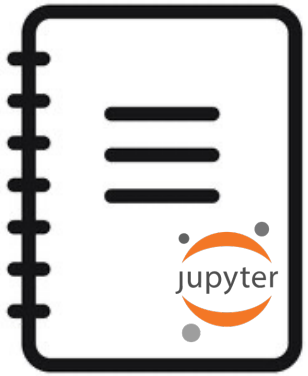


Example: https://jupyter.iguide.illinois.edu/

How do we envision people using the Platform?

1

Develop your workflow as a computational notebook

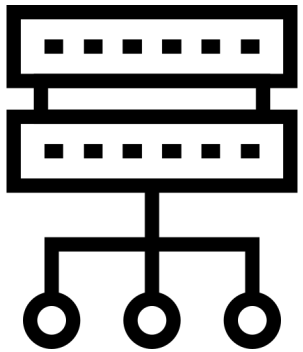


2

Stage your data in an appropriate storage location or repository

3

Choose an appropriate computational resource

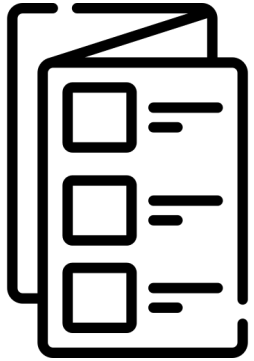


4

Deposit data, notebooks, results in a repository for publication

5

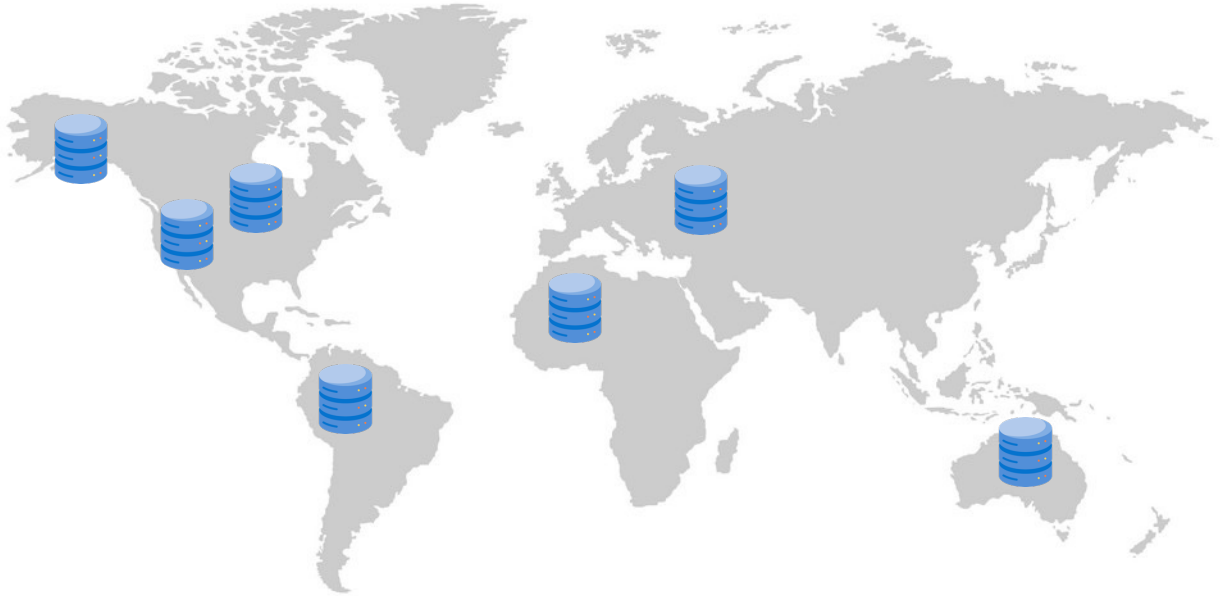
Register your products with the I-GUIDE catalog



These steps aren't trivial – some iteration may be required

Platform cyberinfrastructure components

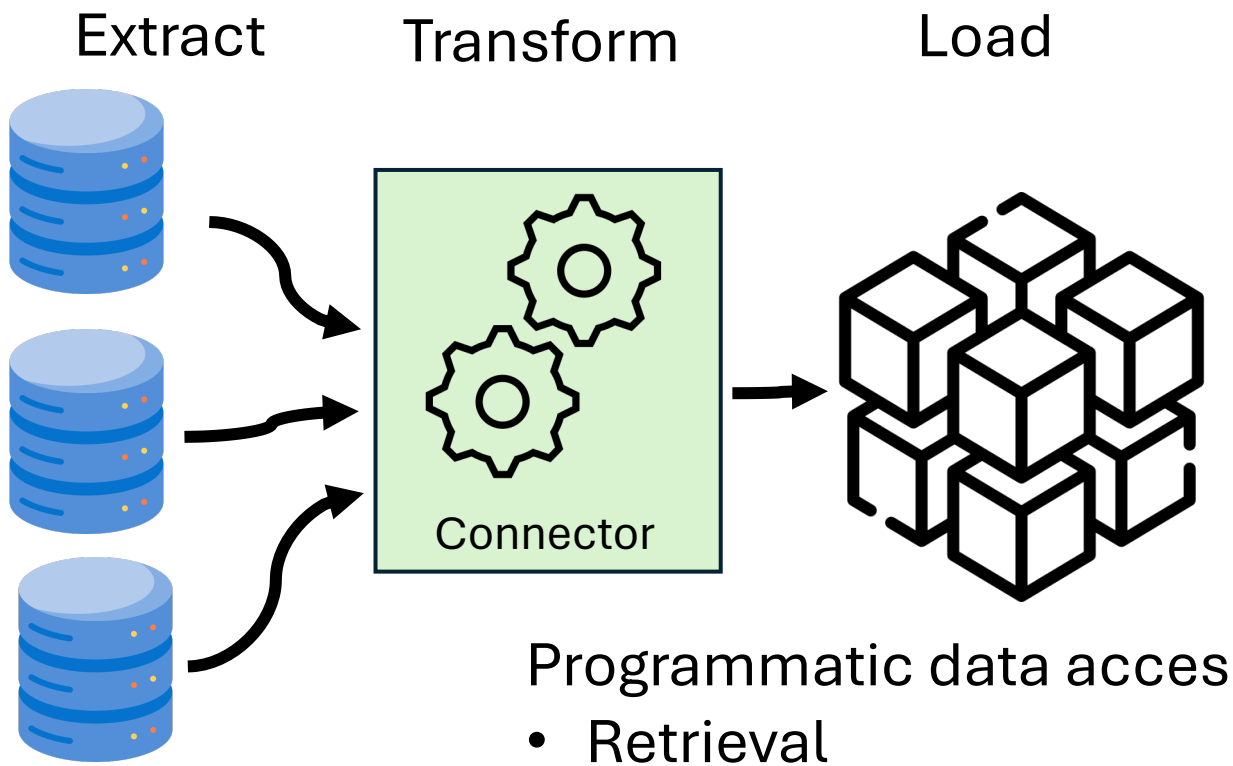
- Distributed geospatial data assets
- Data connectors and data processors
- Workflow orchestration
- Open data portals/repositories
- HPC and Cloud storage
- Computational environments including HPC
- Actionable catalog of I-GUIDE products



- Core to I-GUIDE convergence science
- Data located across websites/repositories
- May be too big to copy/move

Platform cyberinfrastructure components

- Distributed geospatial data assets
- Data connectors and data processors
- Workflow orchestration
- Open data portals/repositories
- HPC and Cloud storage
- Computational environments including HPC
- Actionable catalog of I-GUIDE products



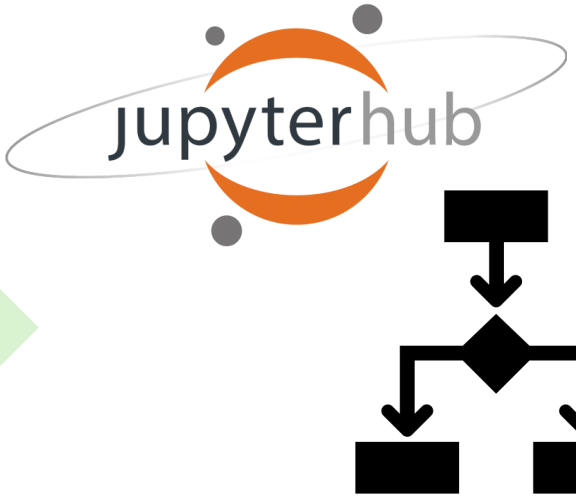
GeoEDF

Programmatic data access

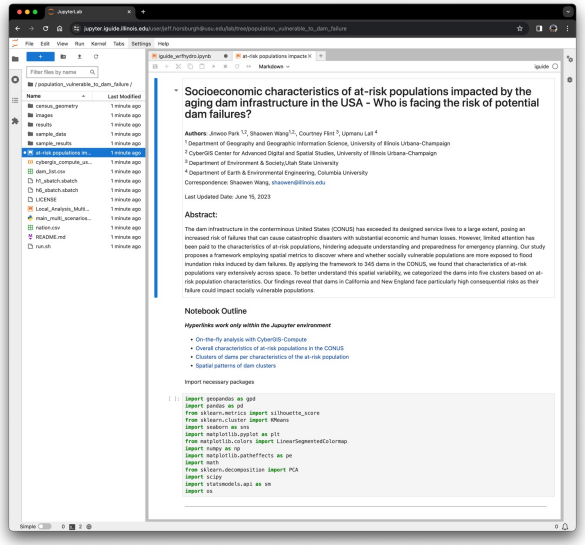
- Retrieval
- Subsetting
- Regridding
- Transformation

Platform cyberinfrastructure components

- Distributed geospatial data assets
- Data connectors and data processors
- Workflow orchestration
- Open data portals/repositories
- HPC and Cloud storage
- Computational environments including HPC
- Actionable catalog of I-GUIDE products



Computational Notebook

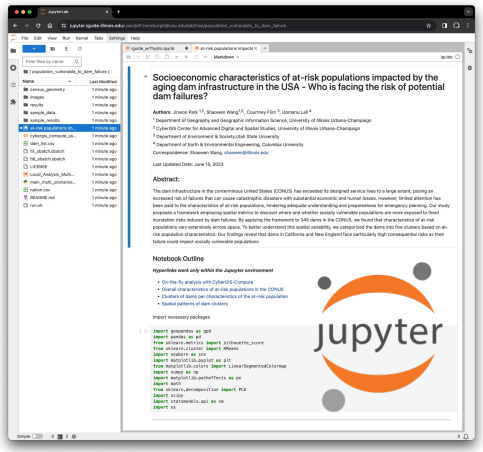


- Connecting multiple, complex processing/analysis/modeling tasks
- Repeatable set of steps that can be easily executed and documented
- Computational notebooks

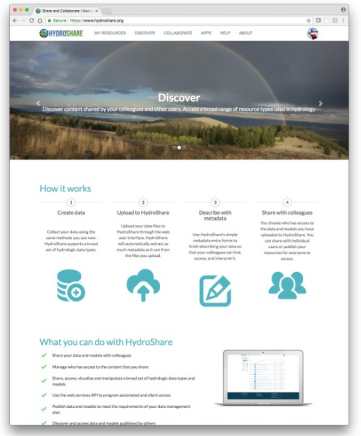
Platform cyberinfrastructure components

- Distributed geospatial data assets
- Data connectors and data processors
- Workflow orchestration
- Open data portals/repositories
- HPC and Cloud storage
- Computational environments including HPC
- Actionable catalog of I-GUIDE products

Computational Notebook



HYDROSHARE

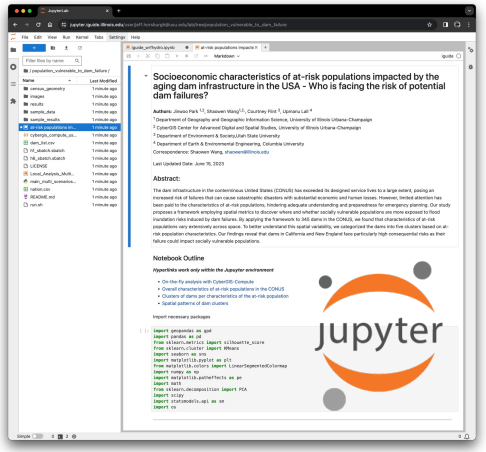


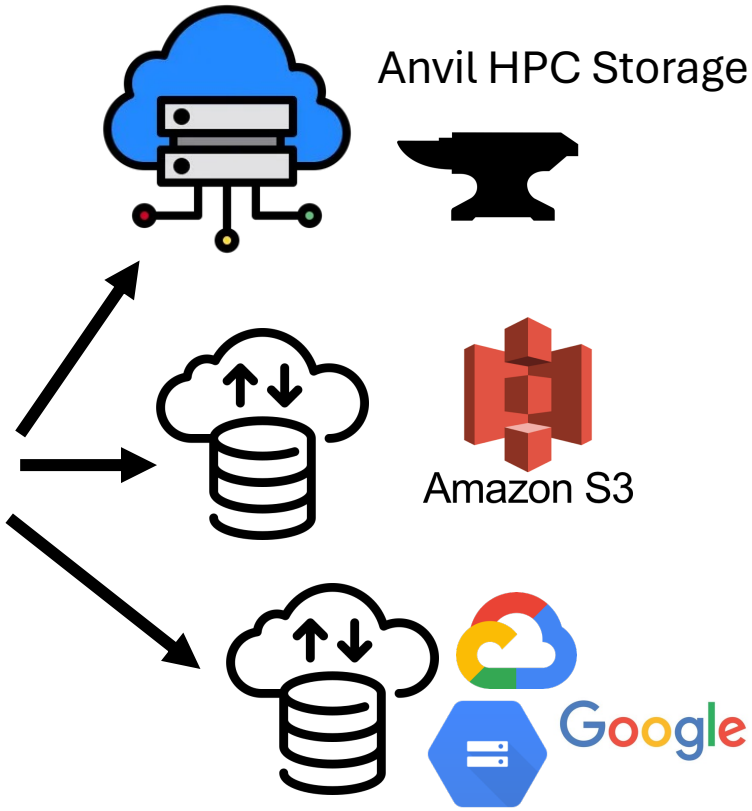
- Store and retrieve content
- Staging data for analysis
- Sharing data/notebooks/content
- Launching computational notebooks
- Permanent publication

Platform cyberinfrastructure components

- Distributed geospatial data assets
- Data connectors and data processors
- Workflow orchestration
- Open data portals/repositories
- HPC and Cloud storage
- Computational environments including HPC
- Actionable catalog of I-GUIDE products

Computational Notebook



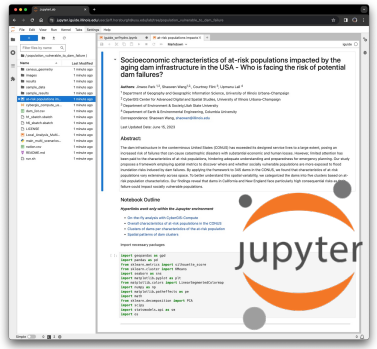


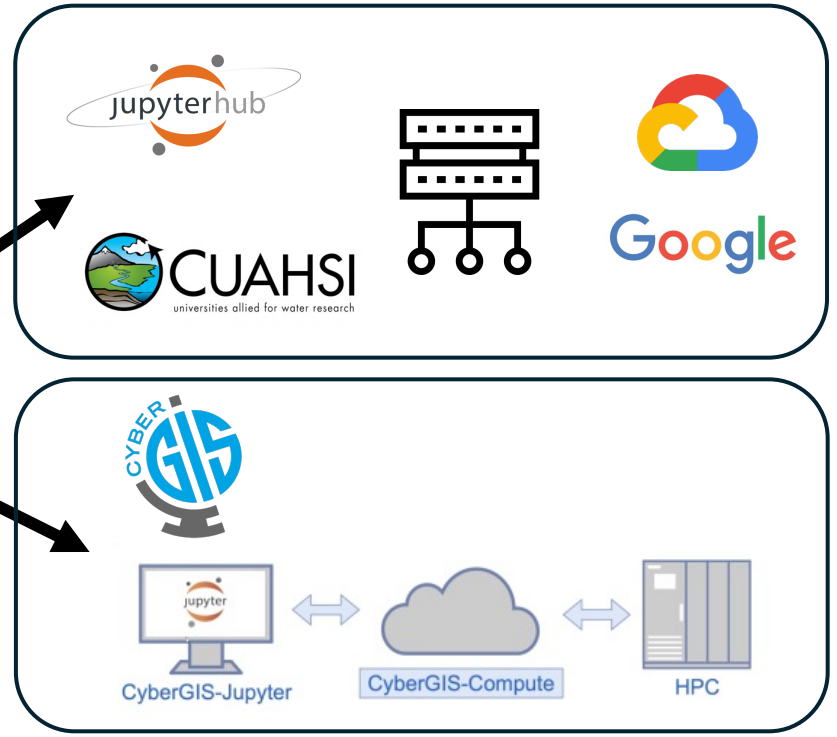
- Moving data
- Staging data for modeling
- Matching storage with computation
- Large data output

Platform cyberinfrastructure components

- Distributed geospatial data assets
- Data connectors and data processors
- Workflow orchestration
- Open data portals/repositories
- HPC and Cloud storage
- Computational environments including HPC
- Actionable catalog of I-GUIDE products

Computational Notebook



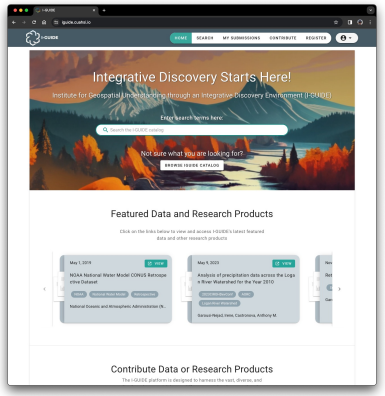


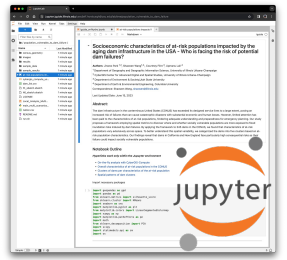
- Access an appropriate computational resource
- Link to HPC where needed

Platform cyberinfrastructure components

- Distributed geospatial data assets
- Data connectors and data processors
- Workflow orchestration
- Open data portals/repositories
- HPC and Cloud storage
- Computational environments including HPC
- Actionable catalog of I-GUIDE products

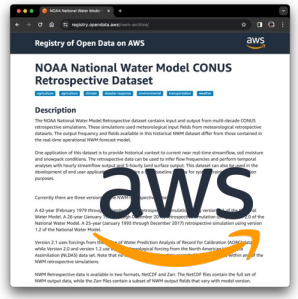
I-GUIDE Catalog





Discover and Launch Notebooks

Discover and access repository data

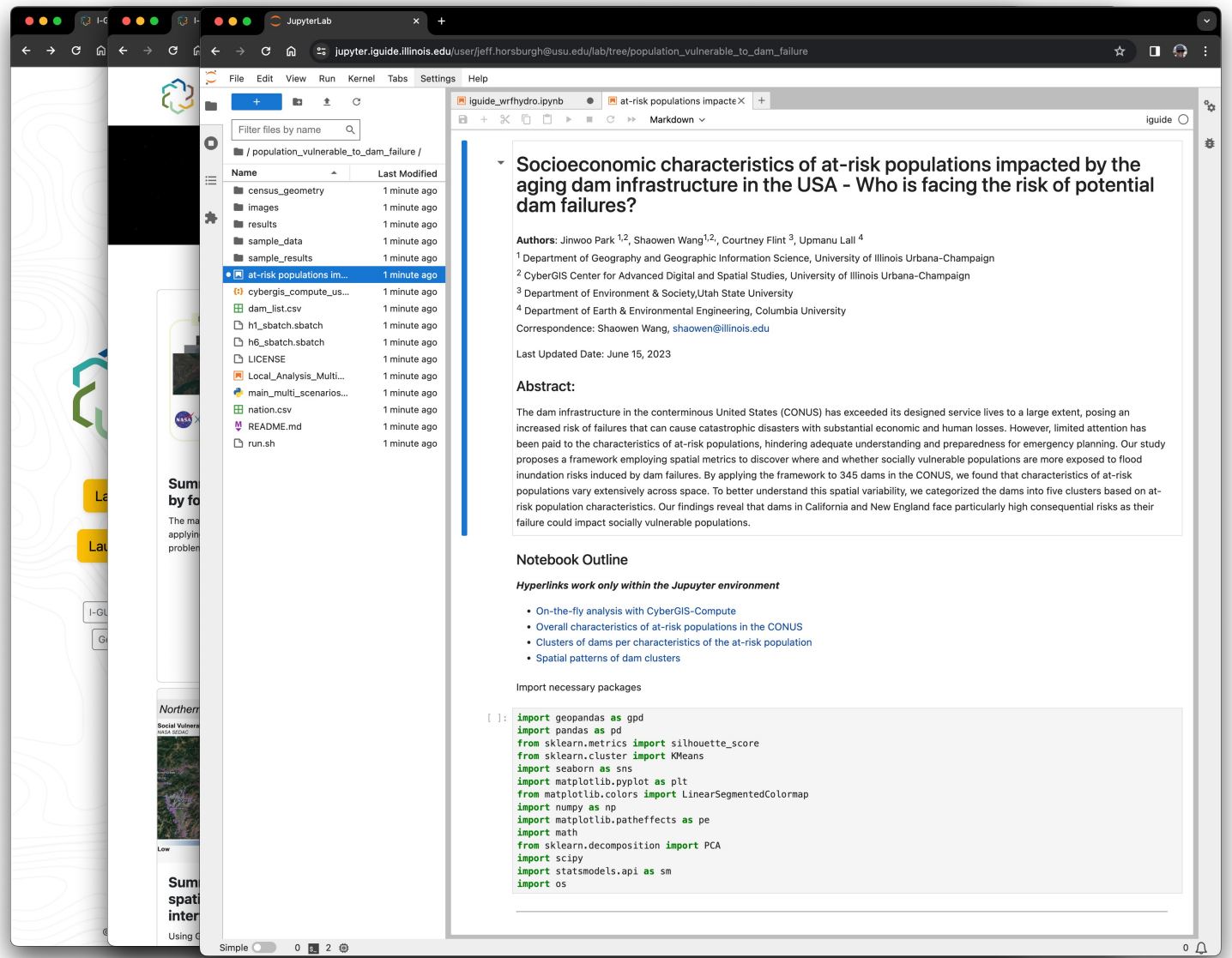


Discover and access cloud data

- Discover notebooks, data, other products
- Access data and metadata
- Launch “actionable products” (e.g., Notebooks)

Example Platform applications

- Jupyter Notebooks that demonstrate capabilities of the platform

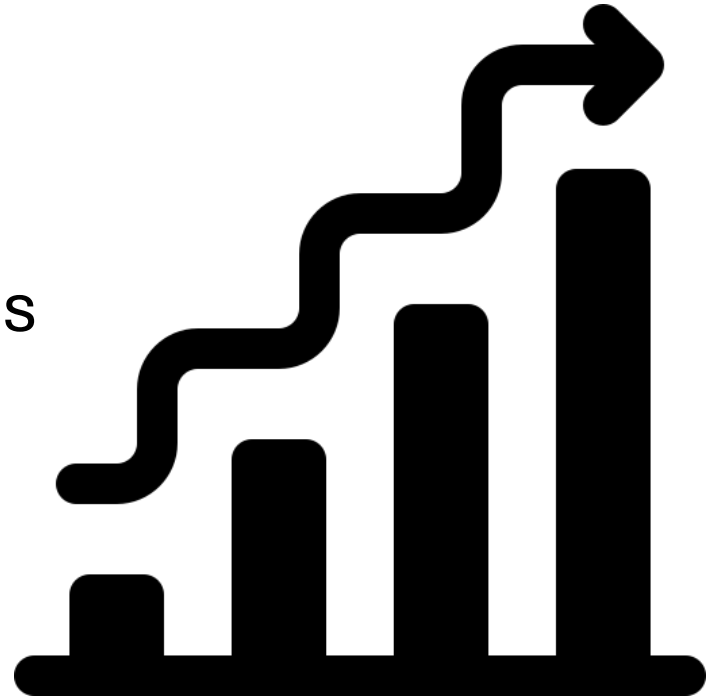


https://iguide.illinois.edu/platform/

Where do we need to improve the Platform?

- The entry point for using the Platform isn't clear – where do I start?
- There are a lot of options
  - Which storage should I use?
  - Which JupyterHub should I use?
  - Which repository should I use?

- We developed a generalized system, but we need to further test and exercise with I-GUIDE use cases
- In other words – do the components we have assembled meet the needs of the CSC teams?



Why invest in using the I-GUIDE Platform together?

- Greater transparency for our scientific work and increased trust in results
- Greater opportunity for enabling reproducibility of results
  - *Reproducibility by design* rather than requirement
- Published research products (with citable DOIs)
- Tools for sharing and collaboration
- Facilitated access to repositories and computational resources
- Tools for teaching and active learning
- Compliance with funding agency and journal publisher requirements
- Because we said we would!

I-GUIDE
Institute for Geospatial Understanding through an Integrative Discovery Environment

Questions?

Jeffery S. Horsburgh
jeff.horsburgh@usu.edu

Utah Water Research Laboratory
UtahStateUniversity

Support: 2118329