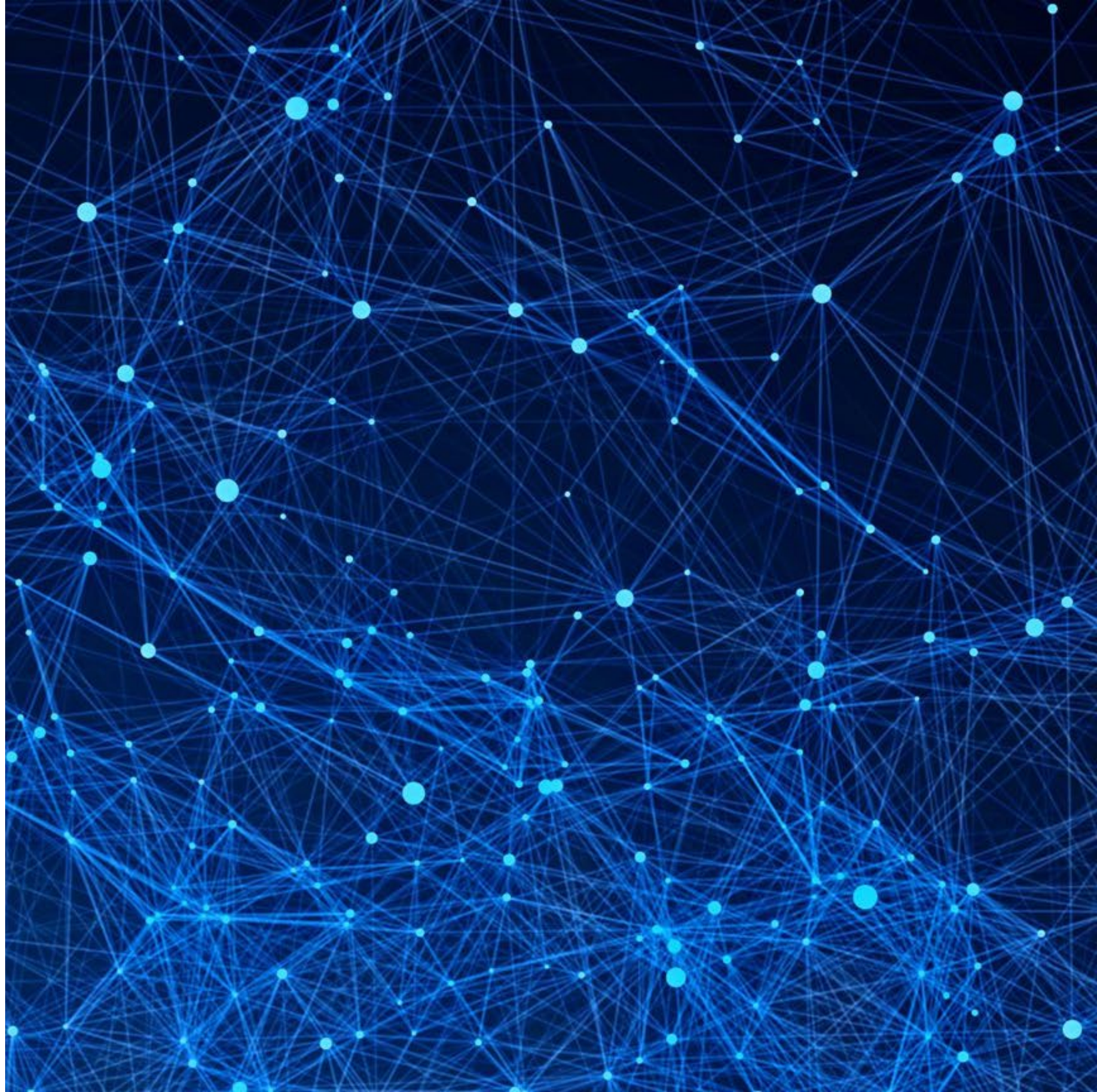
Pacific Northwest NATIONAL LABORATORY

MSD LIVE

Engaging with MSD-LIVE

Casey Burleyson, Carina Lansing, Zoë Guillen, Matt Jugovic, Matthew Macduff, Devin McAllester, and Jon Weers



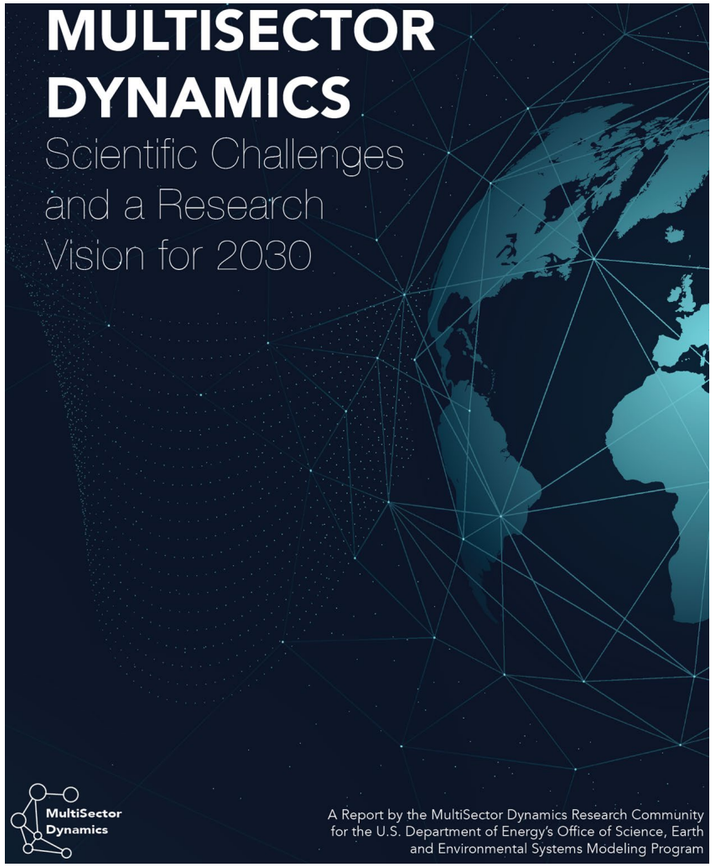
U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Open Science in the MSD CoP

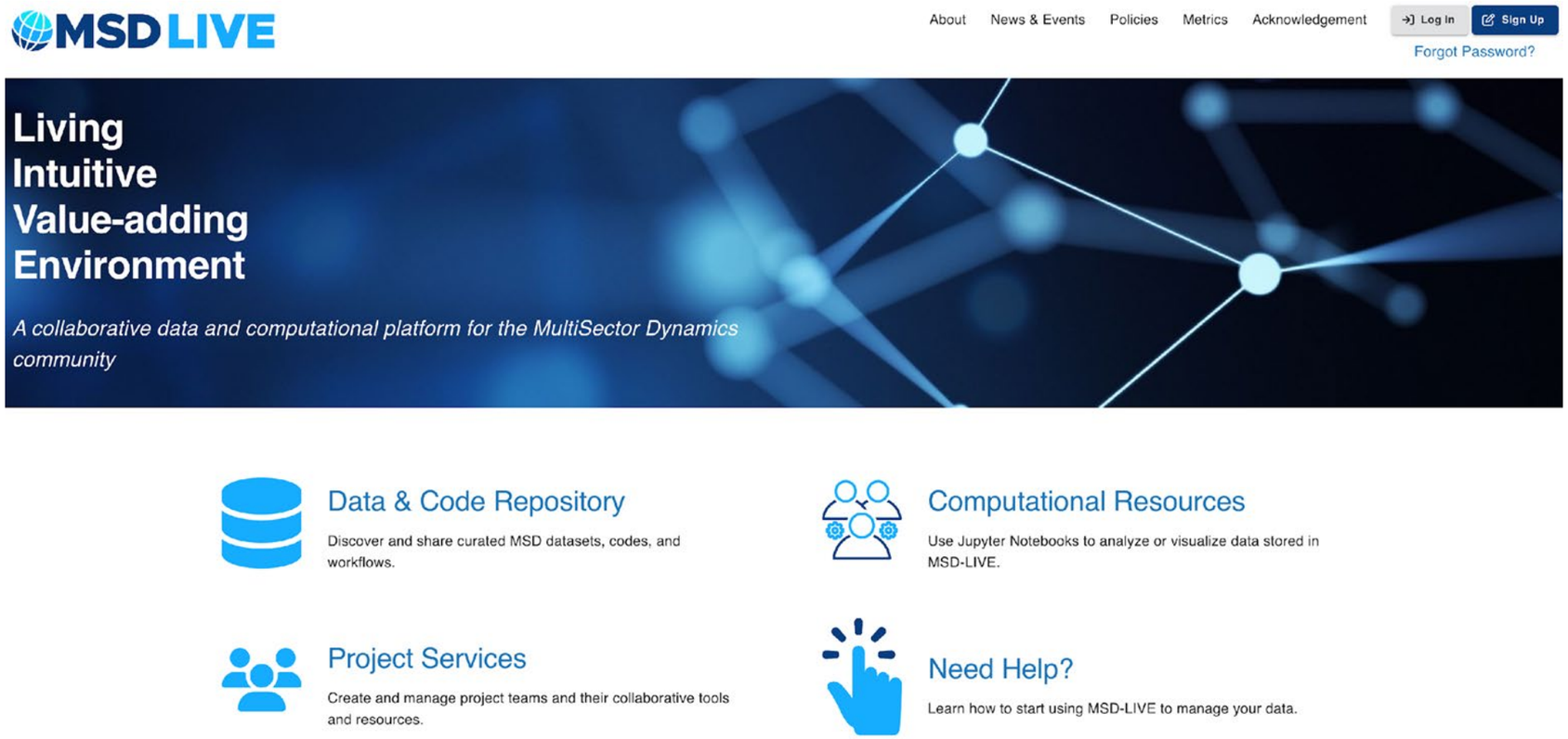
"This report summarizes current MSD research, synthesizes insights from these early investments, defines key concepts, identifies areas for further research, and ***proposes a collaborative research strategy based on open science principles and growing a diverse workforce to accelerate progress****."*

*"****Expand MSD's commitment and capacity for open science to continue developing interoperable and more easily reusable data, models, and analysis methods****; leverage emerging computational innovations (e.g., AI, natural language processing, reinforcement learning (RL), and interactive visual analytics); and grow and diversify the MSD workforce to broaden the backgrounds, technical skills, and expertise/experiences available to advance our understanding of societal risks."*



Reed, P. M., et al. (2022). MultiSector Dynamics: Scientific Challenges and a Research Vision for 2030, A Community of Practice Supported by the United States Department of Energy's Office of Science. Zenodo. https://doi.org/10.5281/zenodo.6144309

The Vision for MSD-LIVE



https://msdlive.org

- A cloud-based data management system and advanced computing platform
- Will enable researchers to document and archive their data, run their models and analysis tools, and share data, software, and multi-model workflows
- A cornerstone capability of the MSD Community of Practice

Top-10 Most Pressing Use Cases

Core Capabilities	Use Cases
Data Repository	Find Data
	Archive Data
	Version Control
	Training
	Control Access
Advanced Computing	Share Data
	Analyze Data
	Multi-Model Workflows
	ML/AI
	Move Code to the Data

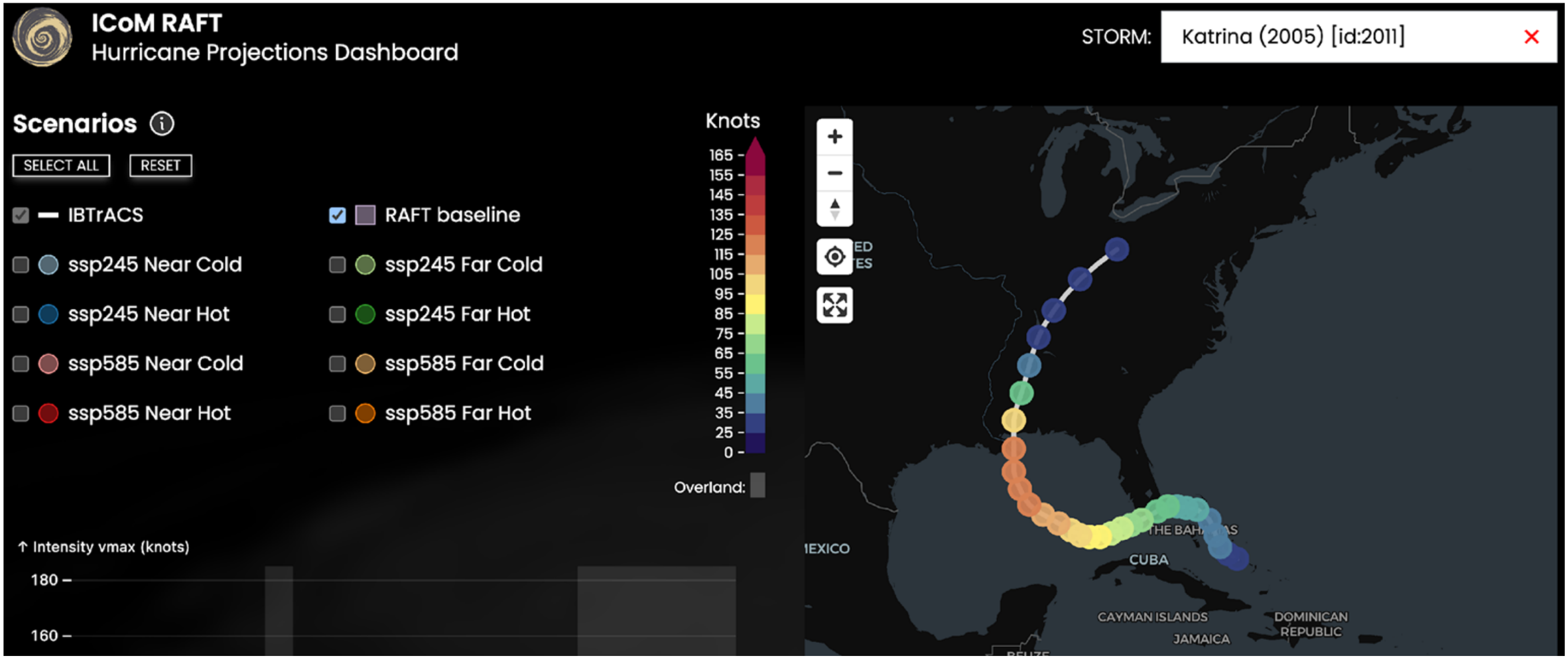
= available now = current focus

Our use cases are derived from extensive conversations with researchers from the MSD community.

Current Capabilities

- Data repository:
  - https://data.msdlive.org/
- Data dashboards:
  - PCHES - https://lafferty-sriver-2023-downscaling-uncertainty.msdlive.org/user-redirect/lab/tree/dashboard.ipynb
  - ICoM - https://raft-hurricane-projections.msdlive.org/
- Interactive model training notebooks:
  - IM3 - Uncertainty Characterization eBook: https://uc-ebook.org/
  - GCIMS - Xanthos: https://xanthos.msdlive.org
  - GCIMS - Hector: https://hector.msdlive.org
  - GCIMS - GCAM Wrapper: https://gcamwrapper.msdlive.org
  - GCIMS - Stitches: https://stitches.msdlive.org
  - IM3 - StateModify: https://statemodify.msdlive.org

Interactive Data Dashboards



https://raft-hurricane-projections.msdlive.org/

Jupyter Notebooks

The MSD community is leveraging the cloud-computing capabilities in MSD-LIVE to train new people and teams to understand and extend MSD data and models.

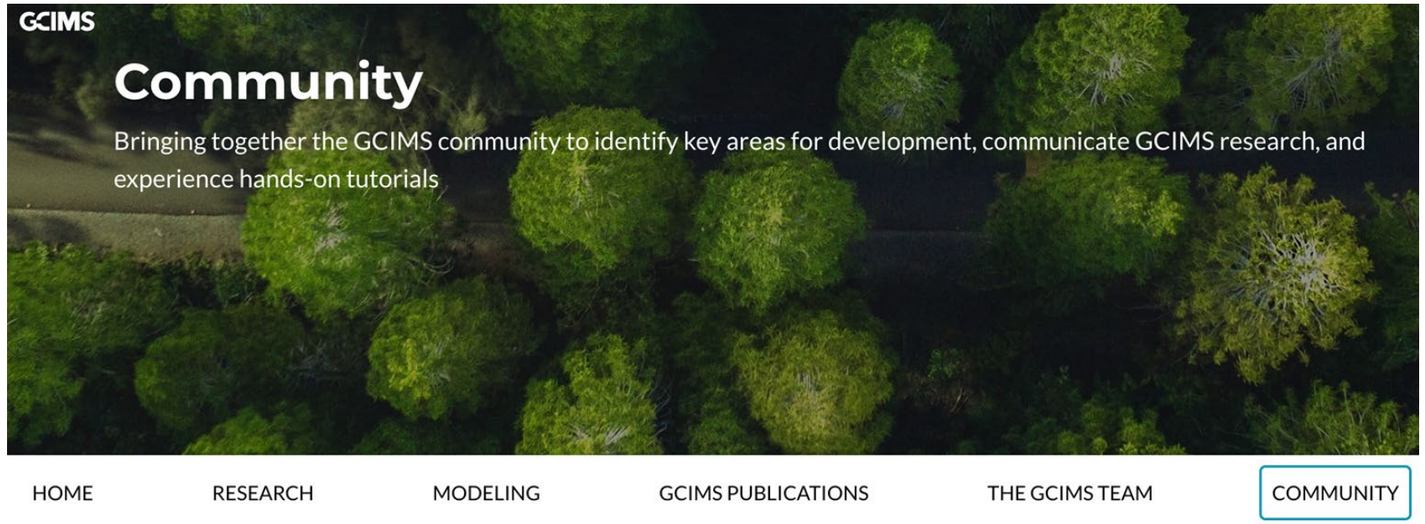
Sandbox

Classroom

Library

Training New Users

- At the 2024 GCAM Annual Meeting we used MSD-LIVE to train attendees on seven GCAM ecosystem models (demeter, gcamwrapper, hector, matilda, rgcam xanthos, and stitches).
- Attendees were able to run the models and analyze the output in real-time on the AWS cloud.
- 250+ people launched live computing instances over the course of the two-day training.



Global Change Analysis Model (GCAM) Annual Meetings

The Joint Global Change Research Institute's next annual GCAM Community Modeling Meeting will be held **June 5-6, 2024** virtually. Registration is now open.

The GCAM Community Modeling Meeting provides a forum for the GCAM community to hear about the latest developments and current uses of GCAM. For those new to GCAM, this workshop continued to serve as an excellent in-depth opportunity to learn and gain a better understanding of the model.

Check out what was presented here:

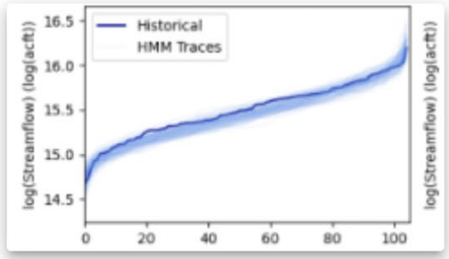
-2023 GCAM Annual Meeting Presentations and Posters
-2022 GCAM Annual Meeting Presentations
-2022 GCAM Annual Meeting Posters
-2019 (and earlier) GCAM Annual Meeting Presentations.

Building a Library of MSD Models

Computational Resources / Model Training Notebooks

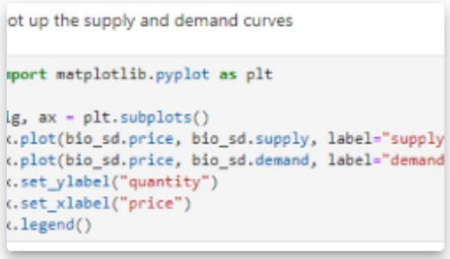
Model Training Notebooks

MSD projects are using the cloud-computing capabilities in MSD-LIVE to create interactive Jupyter notebooks to train new users to configure, run, and analyze MSD models.



statemodify

statemodify is an open-source Python package that provides users with a way to easily modify StateMod's input and output files to enable exploratory modeling.



```
ot up the supply and demand curves

mport matplotlib.pyplot as plt

lg, ax = plt.subplots()
x.plot(bio_sd.price, bio_sd.supply, label="supply
x.plot(bio_sd.price, bio_sd.demand, label="demand
x.set_ylabel("quantity")
x.set_xlabel("price")
x.legend()
```

gcamwrapper

gcamwrapper contains all C++ and R/python source code to wrap the GCAM model such that simulations can be run interactively.

```
:param coord:          data frame for coordinates [lat, lon]
:param legend_label:   string for legend label
:param cmap:           string for color palette

---

cols_str, cols_dict = name_columns(tas, start_yr=1971)

df = pd.DataFrame(data=arr, columns=cols_str).mean(axis=1).rename('value')

df = pd.concat([df, coord], axis=1)

fig = plt.figure(figsize=(15,6), constrained_layout=True)
spec = fig.add_gridspec(1, 1)
ax = fig.add_subplot(spec[0, 0])
p = ax.scatter(df.lon, df.lat, s=3, c=df.value, cmap=cmap)
legend_kwds = {'label':legend_label,
               'shrink':1}
```

xanthos

xanthos is an open-source hydrologic model written in Python that simulates historical and future global water availability on a monthly time step.

https://msdlive.org/sb/computational-resources

- MSD-LIVE is becoming a living library for MSD models.
- Our innovative approach using Jupyter notebooks is making it easier to train new users to configure, run, and analyze MSD models.
- Can be used to entrain new users, including those from small and resource-limited colleges, into the MSD research community.

Ways You Can Use MSD-LIVE Today

1. Archive your data.
2. Get a data DOI.
3. Find data produced by other MSD projects.
4. Share working data with collaborators.
5. Create a data dashboard so that others can visualize or interact with your data without having to download it.
6. Create a model training notebook to teach others to install, configure, run, and analyze your model.
7. Access training and documentation on open science best practices (e.g., writing readmes, using DOIs, licenses).
8. Find and watch archived MSD CoP webinars.

Core Capabilities	Use Cases
Data Repository	Find Data
	Archive Data
	Version Control
	Training
	Control Access
Advanced Computing	Share Data
	Analyze Data
	Multi-Model Workflows
	ML/AI
	Move Code to the Data

Lessons Learned

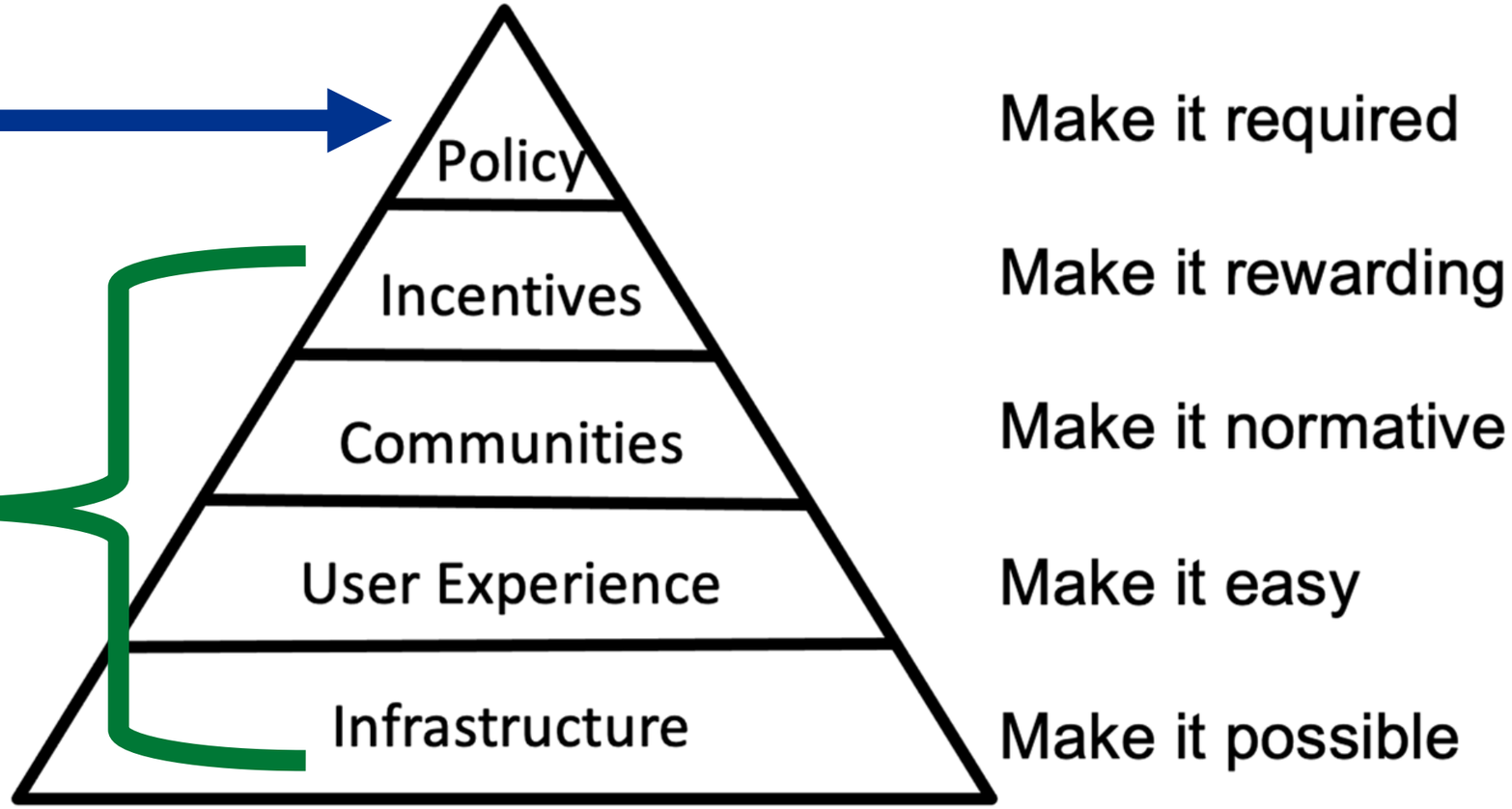
- Facilitate, not complicate
- Demonstrate value to avoid box-checking
- Listen to your users: Often, actively, and with an open mind
- Don’t underestimate how much training you will need to do
- Encourage small steps while you wait for “the big thing”

Backup Slides

How Do We Make a Transition to Open Science?

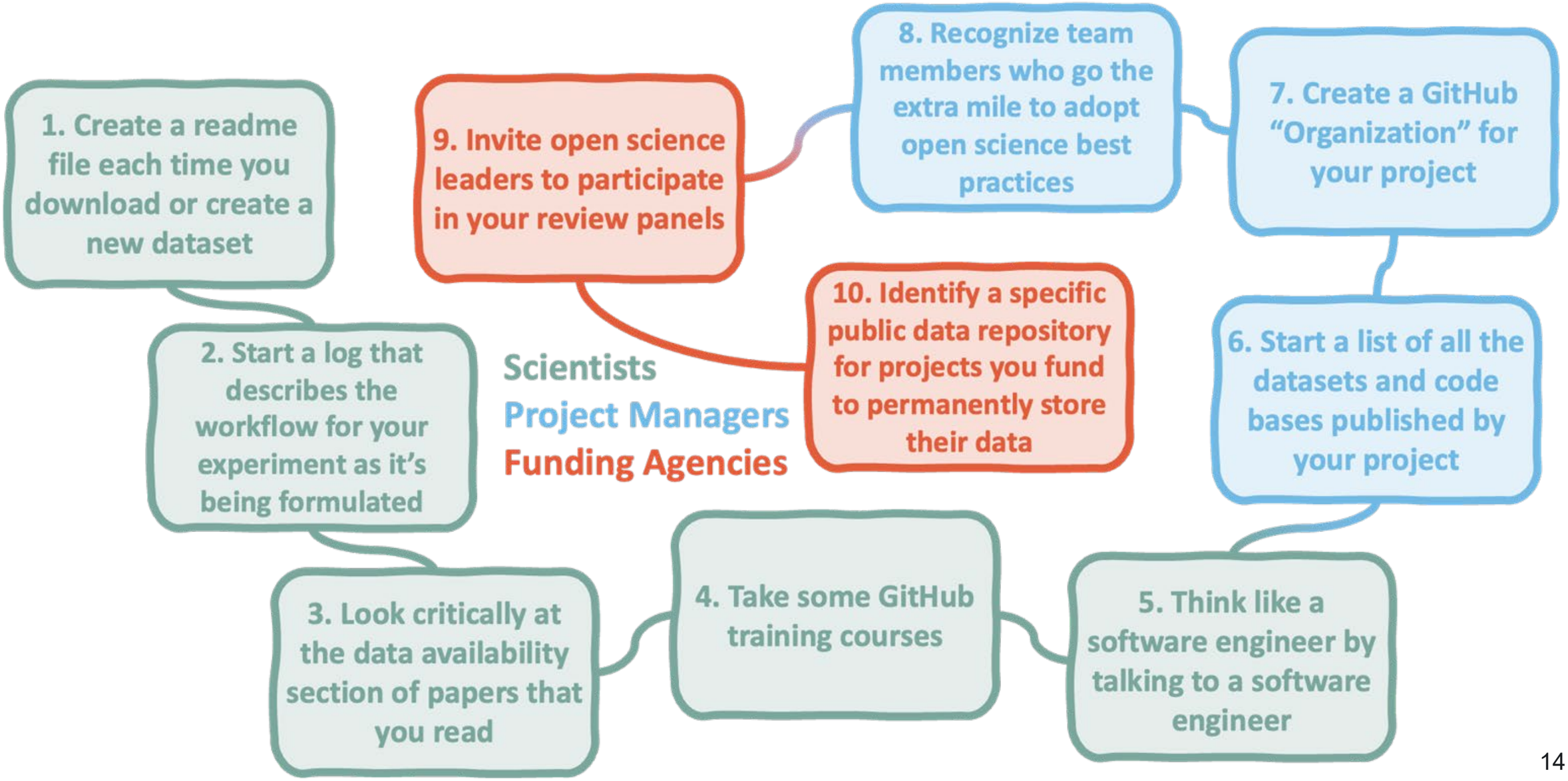
Journals largely skipped to this end of the pyramid…

MSD-LIVE and the MSD CoP are tackling these foundational elements of the pyramid…

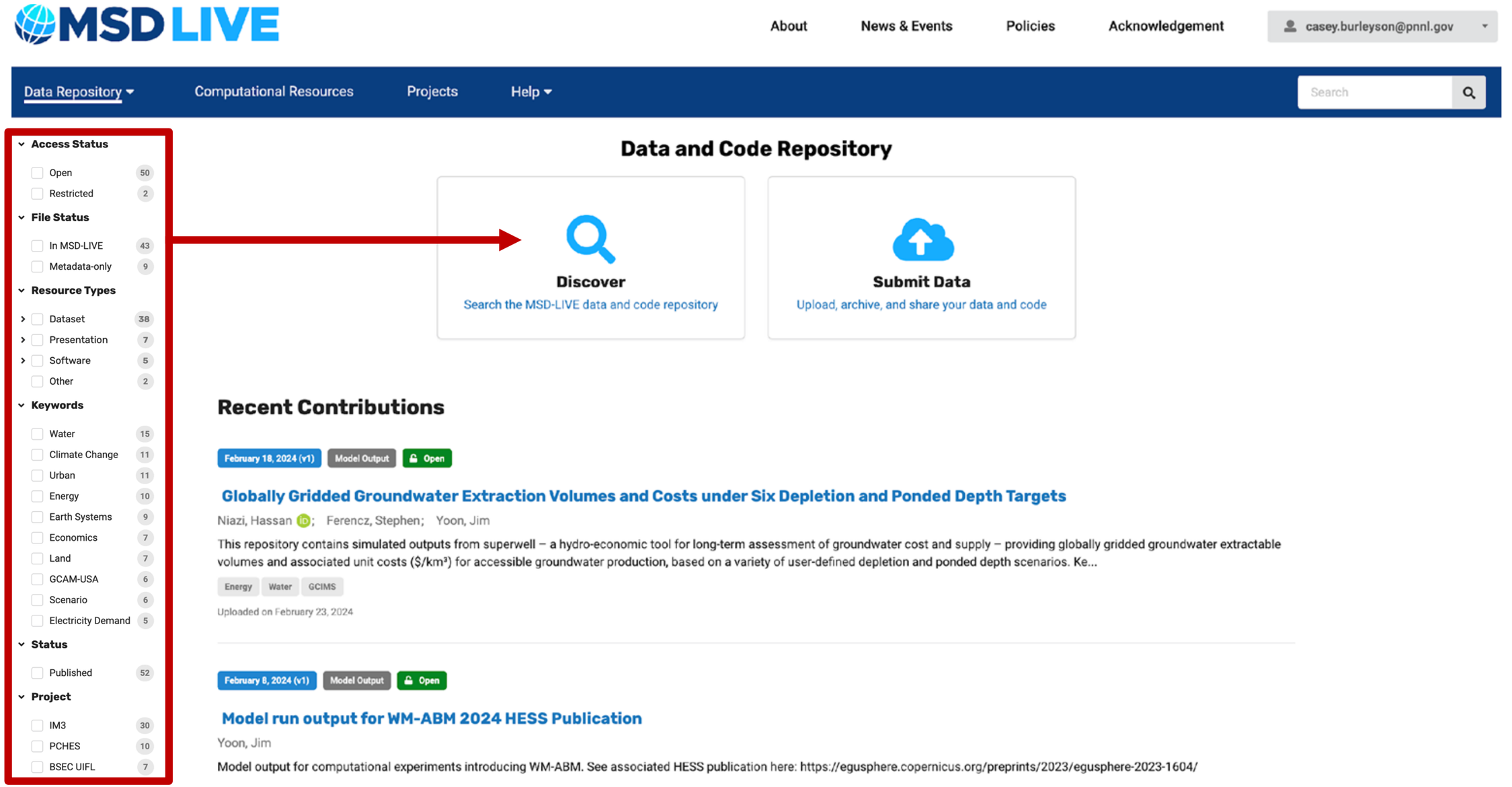


Conceptual diagram from Brian Nosek of the University of Virginia and the Center for Open Science

Steppingstones to Open Science

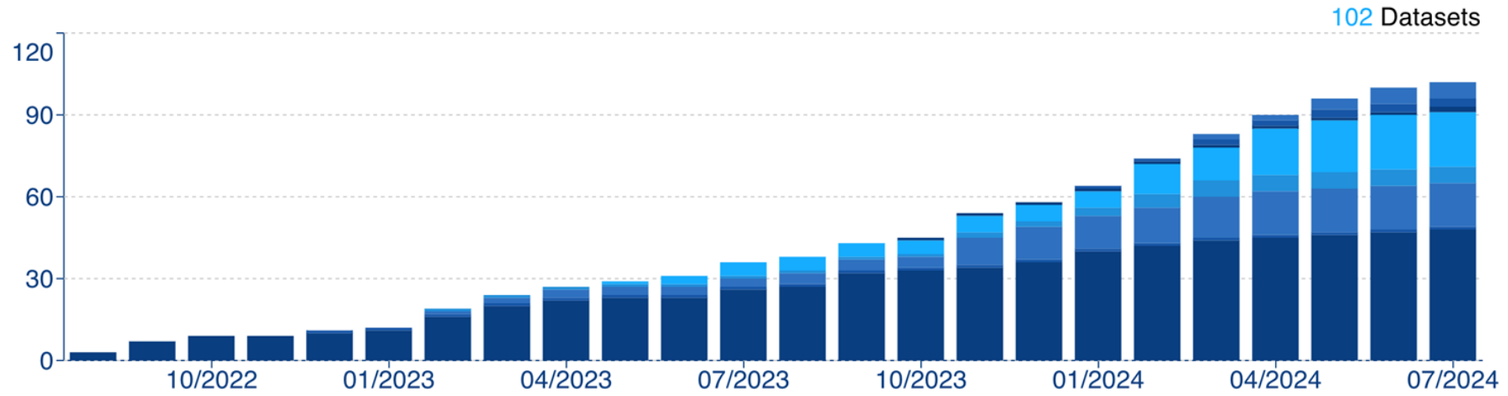


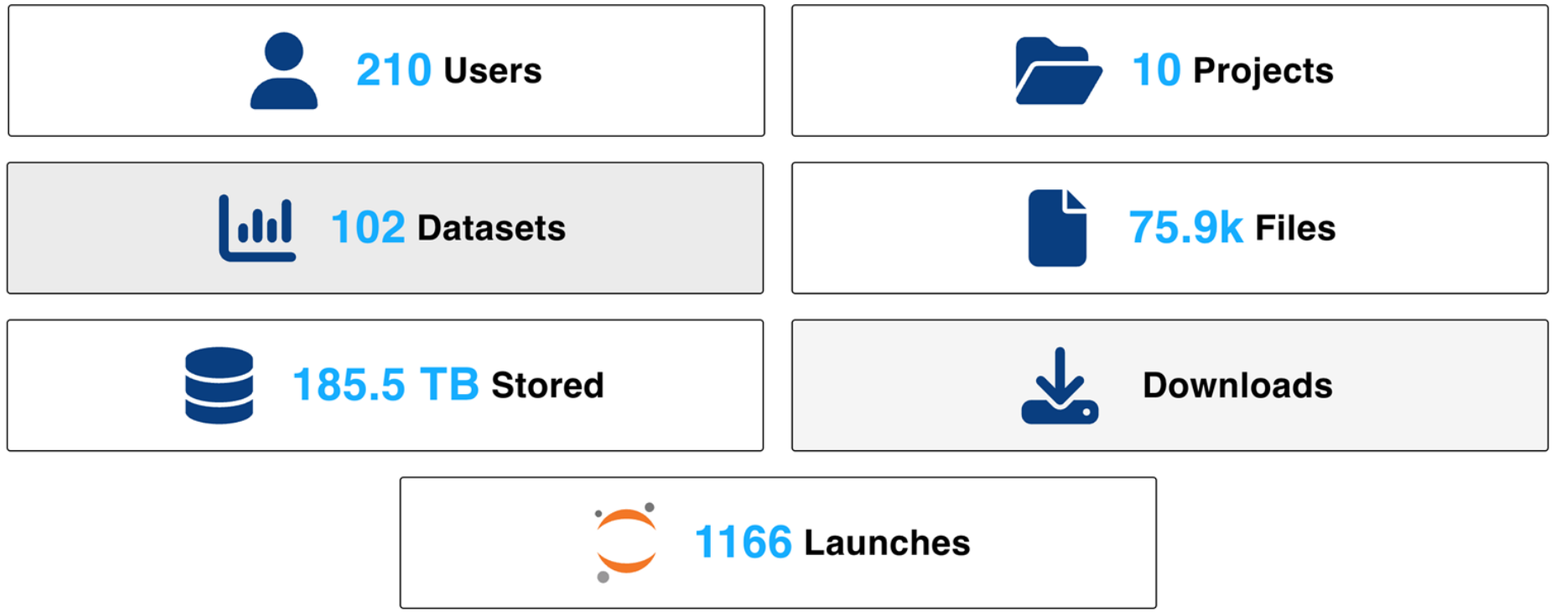
Data Repository



Usage Statistics

- v1 released 22-Aug 2022
- v2 released 28-Jun 2023
- 10 projects
- 210 registered users
- 102 published datasets
- 185+ TB of total data
- ~1150 usages of the Jupyter notebooks in the past three months

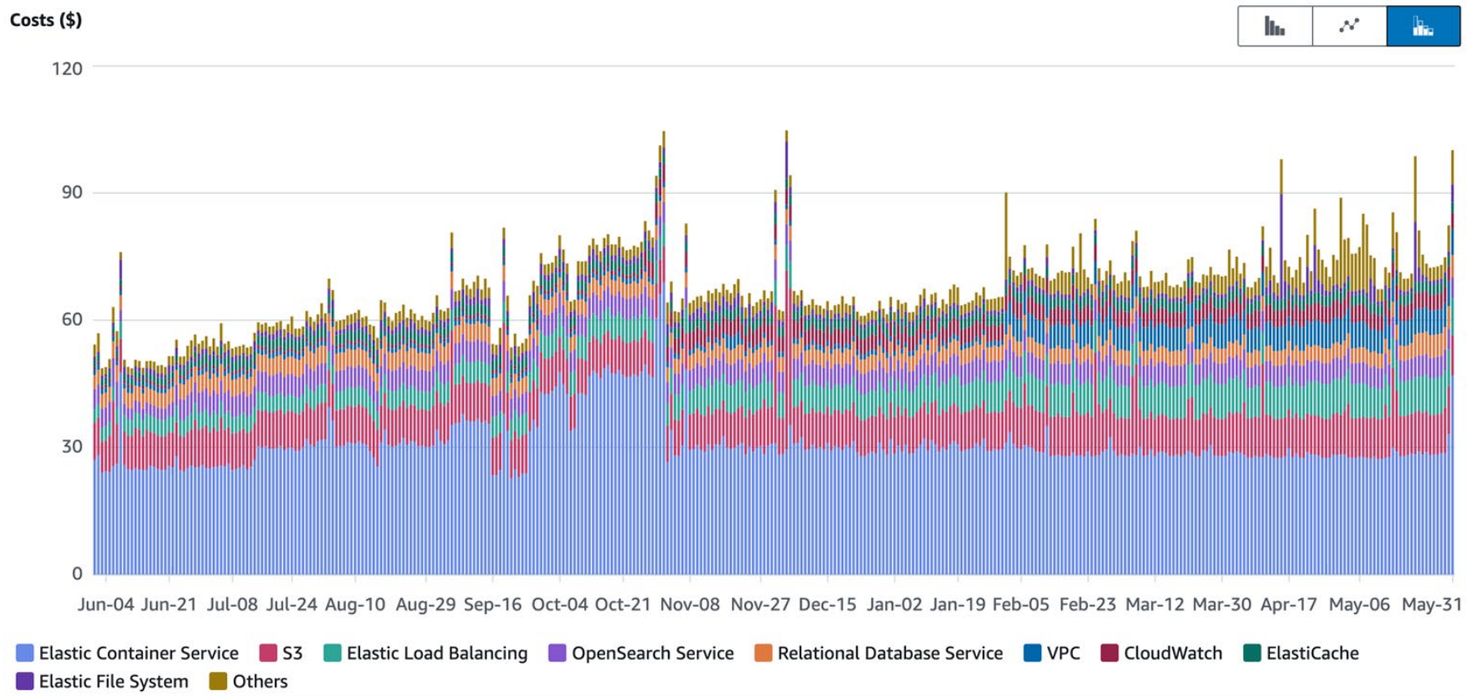




https://msdlive.org/metrics

Cloud Cost Statistics

Total cost	Average daily cost	Service count
$24,697.14	$67.48	14



AWS Service	Daily Cost
ECS (runs basically everything on the site)	~$30
S3 (data storage)	~$8
Jupyter notebooks	~$9
Everything else	~$20