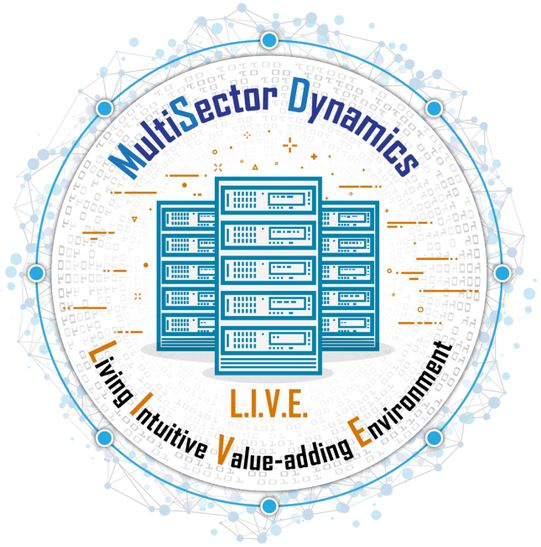
Pacific Northwest
NATIONAL LABORATORY

What is MSD-LIVE?



Casey Burleyson, Zoë Guillen, Carina Lansing, Mathew Thomas, and Jon Weers

U.S. DEPARTMENT OF ENERGY BATTELLE

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

It's Time to Shift Emphasis Away from Code Sharing

Have you ever watched a student struggle to perform a seemingly straightforward analytical procedure? It may be a routine preprocessing step, like detrending a time series or removing a seasonal cycle, but somehow the simple operations can stymie a student for weeks. It's tempting to assume that young people with their short attention spans are unwilling or unable to think through the task at hand, but a closer look suggests that students may have little choice but to blindly tinker with code until things seem to work.

before, rather than moving forward. But worse than simply wasting time, our way of doing business has ensured that scripting typos inevitably go undetected, leading to publication of incorrect findings that at best get caught by follow-on studies or at worst go unidentified and perpetually misguide science.

We Need Better Toolboxes

For efficiency, accuracy, and transparency in Earth science, we need to develop and adopt standard sets of well-tested tools for all our

VOL. 101 | NO. 8
AUGUST 2020
Eos
SCIENCE NEWS BY AGU

Europe's Biodiversity Strategy

A Virtual Hackathon Fights Locusts

MH370's Search Reveals New Science

A GEODATA FABRIC

FOR THE 21ST CENTURY

WE HAVE THE POTENTIAL TO TRANSFORM OUR UNDERSTANDING OF EARTH—IF WE CAN JUST FIGURE OUT HOW TO HARNESS EVER GROWING DATA STREAMS.

By Jeff de La Beaujardière

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

OPEN SCIENCE BY DESIGN

Realizing a Vision for 21st Century Research

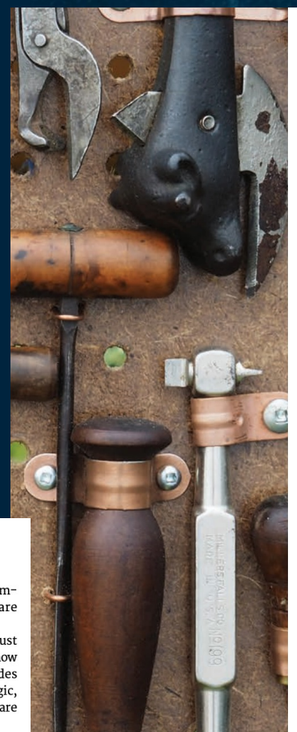
A Data Systems Perspective on *Advancing AI*

Tackling data challenges and incorporating physics into machine learning models will help unlock the potential of artificial intelligence to answer Earth science questions.

INNOVATIONS IN TECHNOLOGY GOT US INTO

THE DATA PROBLEM.

WE NEED AN EVOLUTION IN TECHNOLOGY TO GET US OUT.



CREATING DATA TOOL KITS THAT EVERYONE CAN USE

BY ZHONG LIU, VASCO MANTAS, JENNIFER WEI, MENGLIN JIN, AND DAVID MEYER

Earth scientists need to make the growing wealth of data more accessible and build data services meant for interdisciplinary use.

Advancing FAIR Data in Earth, Space, and Environmental Science

By Manil Maskey, Hamed Alemohammad, Kevin J. Murphy, and Rahul Ramachandran

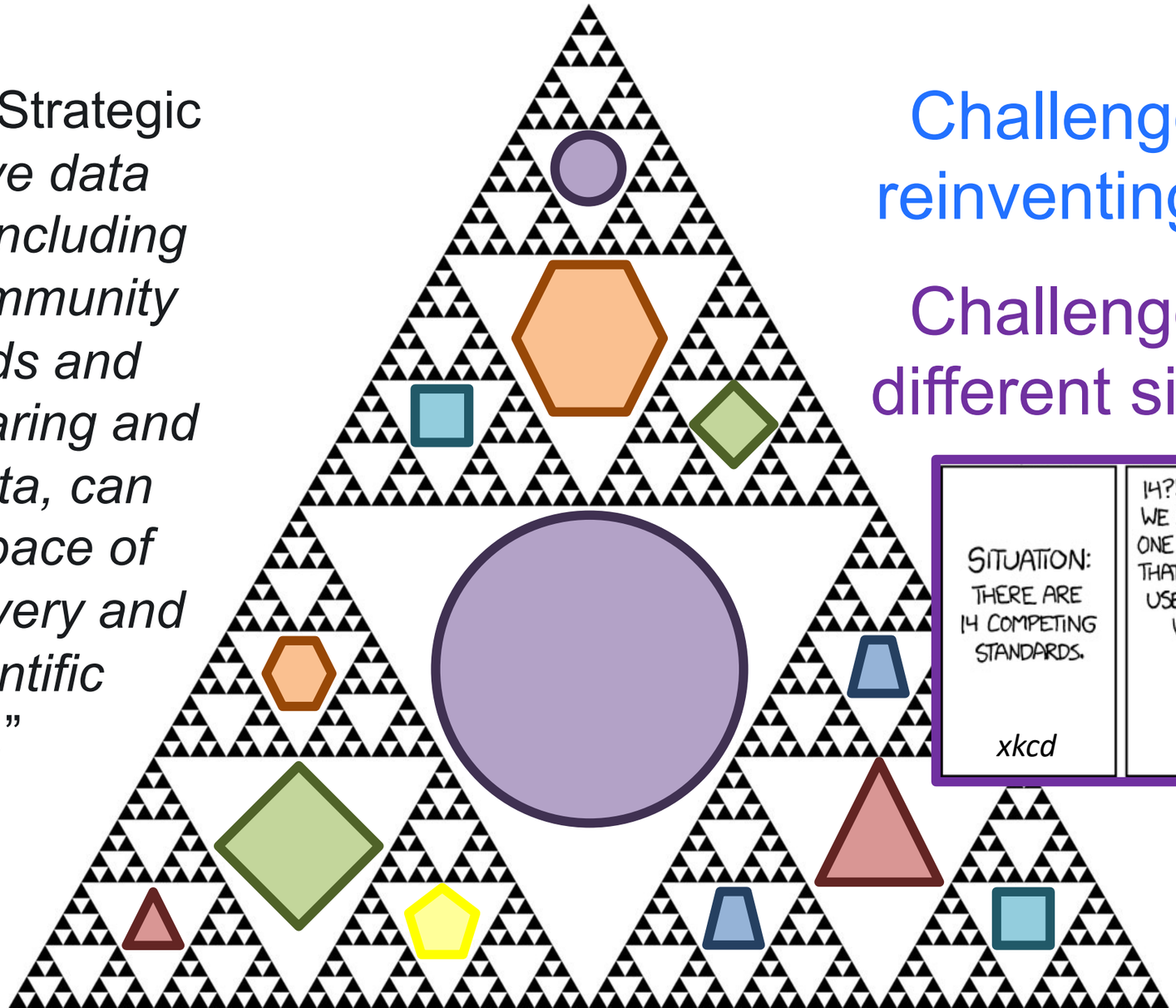
Improving Reproducibility in Earth Science Research

it is available, missing information or incomplete descriptions can make the software hard to understand.

The workflow embedded in software must be well described for others to understand how it processes data. This description includes input and output data sets, workflow logic, algorithms used, the version of the software or library used, and more.

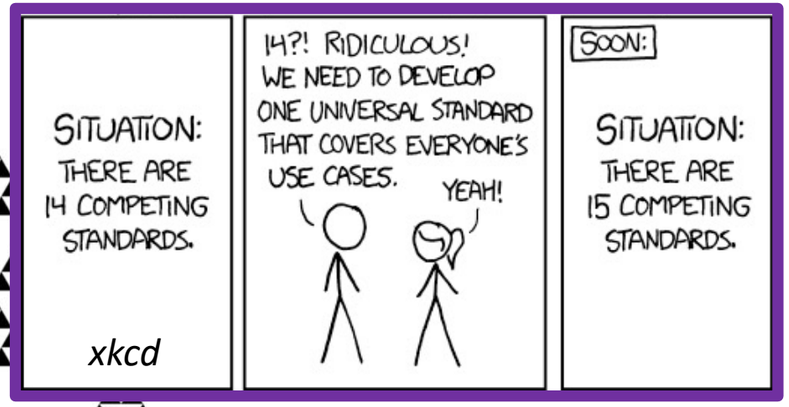
Project-Specific Data Management

From EESSD's Strategic Plan: "*Effective data management, including developing community data standards and formats and sharing and preserving data, can increase the pace of scientific discovery and ensure scientific integrity.*"



Challenge #1: Lots of reinventing the wheel…

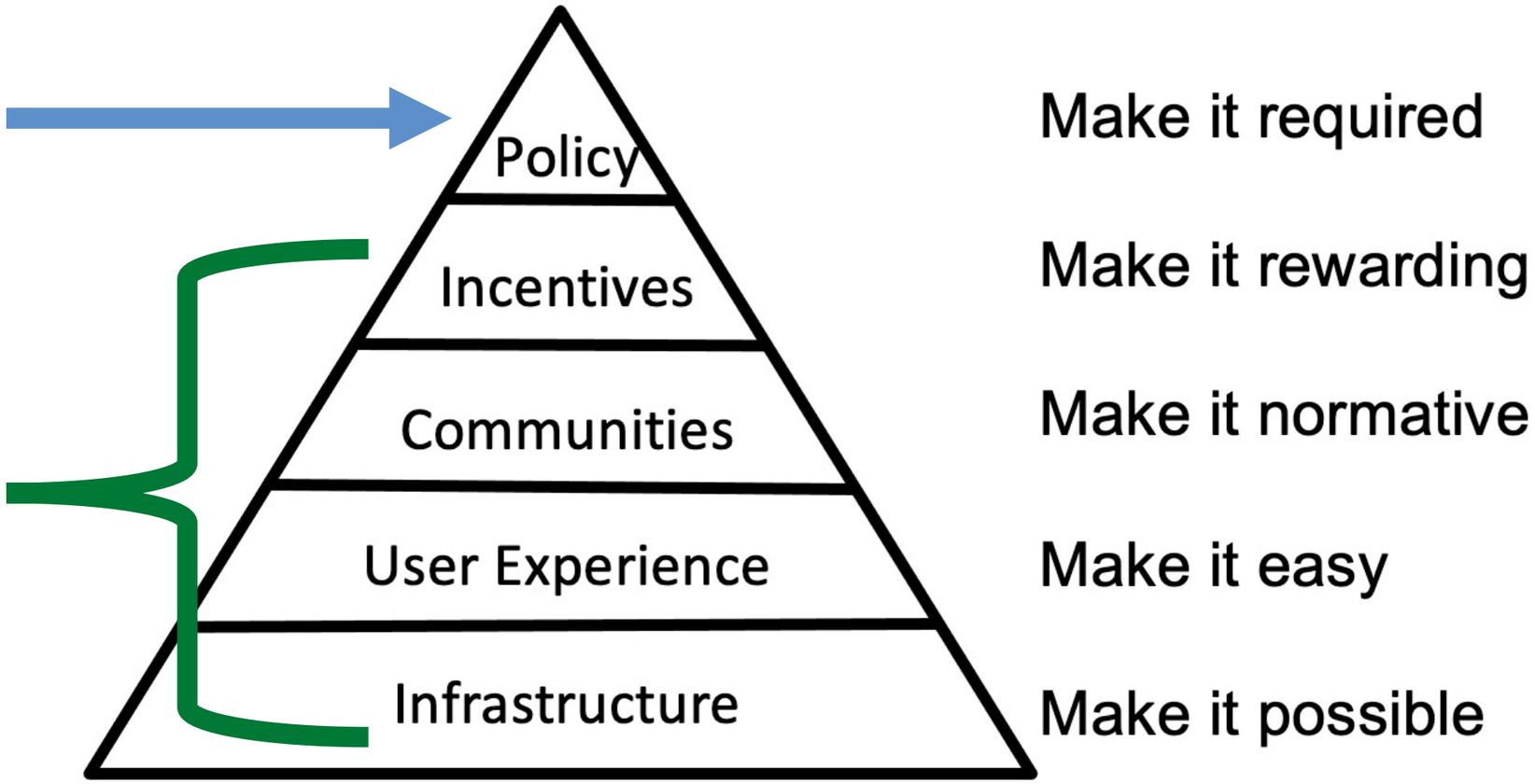
Challenge #2: Lots of different sized wheels…



Facilitating Open Data and Code

Journals and funders largely skipped to this end of the pyramid…

MSD-LIVE is about tackling these foundational elements of the pyramid…

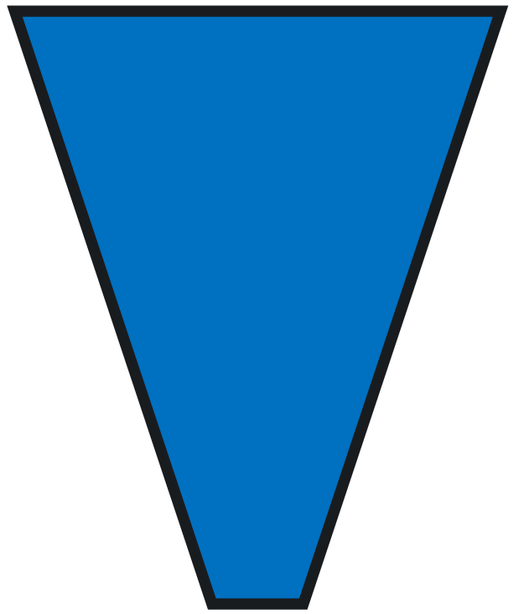


Really nice conceptual diagram shamelessly ~~stolen~~ borrowed from Brian Nosek's talk yesterday

MSD-LIVE Evolved in Multiple Phases

- **Phase 1** (Oct 2018 – May 2019):
  - Initial identification of user needs
  - Initial technical scoping and feasibility assessment
  - Review of other archives (e.g., ESS-DIVE, ARM, EMSL, A2E, NCAR, and NASA platforms)
  - White paper to BER
- **Phase 2** (May 2019 – Feb 2021):
  - Identify and evaluate potential off-the-shelf tools
  - Stakeholder meeting to validate assumptions, refine technical requirements, and prioritize development activities
  - Revised and expanded white paper to BER
- **Phase 3** (Feb 2021+)
  - Funding and implementation!

What is possible?



What will revolutionize MSD?

Top-10 Most Pressing Use Cases

Name	Description
Find Data	Find datasets produced by other users and projects
Archive Data	Archive data and generate DOIs in order to meet journal requirements
Version Control	Manage multiple versions of a dataset
Training	Train new team members to effectively manage their data and code
Control Access	Create teams that cross institutions in order to manage access to data and code
Share Data	Share datasets across multiple institutions collaborating on a project in real-time
Analyze Data	Create, run, and share code to analyze or visualize data
Multi-Model Workflows	Create, execute, document, and publish multi-model workflows
ML/AI	Assemble data lakes and execute ML/AI algorithms on large pools of data
Move Code to Data	Deploy models to run on new computational resources to avoid large transfers

Our Vision for MSD-LIVE

- A data and code management system combined with a distributed computational platform.
- Enable MSD researchers to document and archive data, run models and analysis tools, and share data, software, and multi-model workflows.
- A cornerstone capability of the MSD Community of Practice.

What Our Users Want

- Secured Infrastructure
- ML/AI Entry Points
- Job/Workflow Management
- Configurable Storage
- Identity and Access Control
- Virtual File Catalogue
- DOI Minting
- User Training and Support
- Jupyter Notebooks
- Hook to Code Repositories
- Data Repository
- Configurable Metadata
- Command Line Interface (CLI)
- High-Speed Data Transfer
- Computational Resources
- Visualization Tools
- Team Management
- Search
- Code Container Registry
- Graphical User Interface (GUI)

Pacific Northwest NATIONAL LABORATORY

What Our Users Want

Enabling Infrastructure

- High-Speed Data Transfer
- Secured Infrastructure
- User Training and Support

Core Data Repository

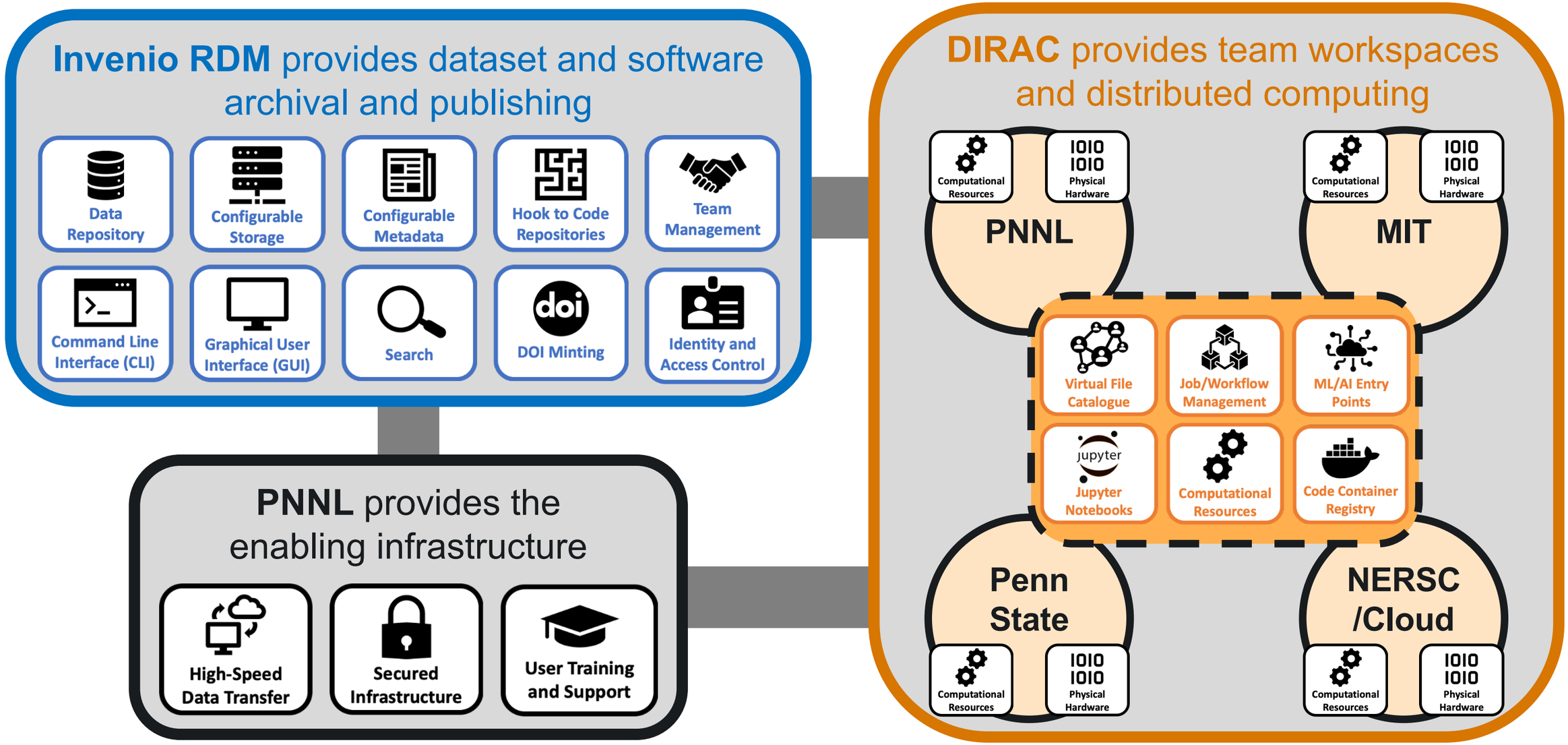
- Data Repository
- Configurable Storage
- Configurable Metadata
- Hook to Code Repositories
- Team Management
- Command Line Interface (CLI)
- Graphical User Interface (GUI)
- Search
- DOI Minting
- Identity and Access Control

Distributed Computing

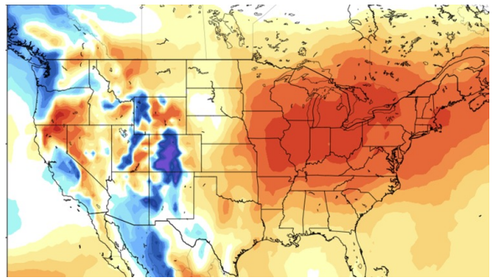
- Virtual File Catalogue
- Job/Workflow Management
- ML/AI Entry Points
- Jupyter Notebooks
- Computational Resources
- Code Container Registry

Pacific Northwest NATIONAL LABORATORY

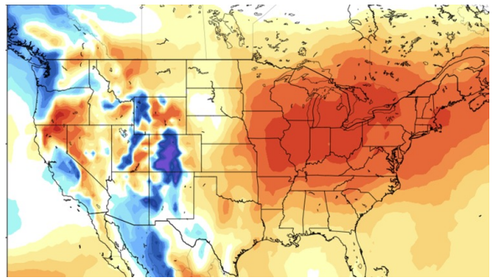
Planned Architecture of MSD-LIVE



Example Usage of MSD-LIVE



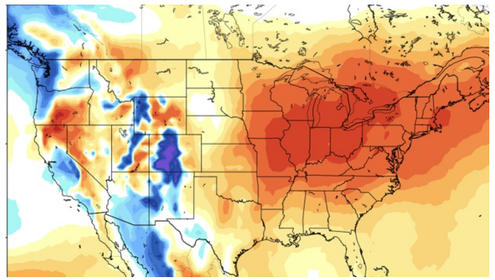
Example Usage of MSD-LIVE



MOSART-WM
+
GCAM-USA
+
TELL
+
GOWEST
=

nature

Example Usage of MSD-LIVE



MOSART-WM

+

GCAM-USA

+

TELL

+

GOWEST

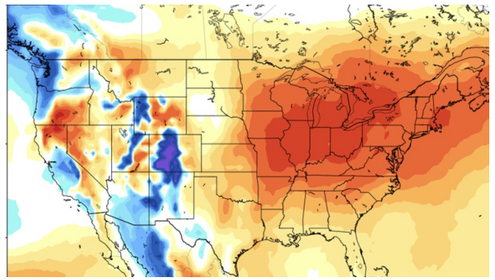
=

nature

Example Usage of MSD-LIVE

Pacific Northwest NATIONAL LABORATORY

IM3 — IM3 — PCHES Program on Coupled Human and Earth Systems



MOSART-WM

+

GCAM-USA

+

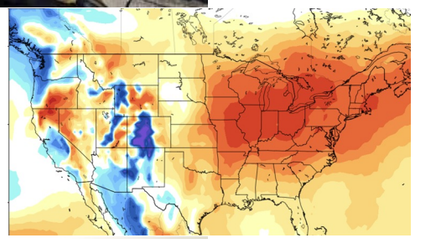
TELL

+

GOWEST

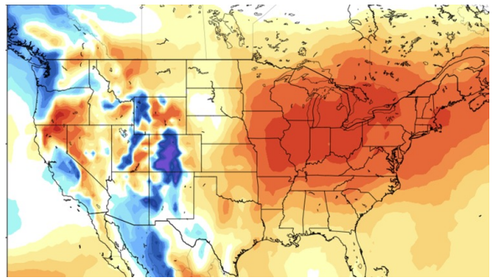
=

nature



Example Usage of MSD-LIVE

IM3 —— IM3 —— MultiSector Dynamics L.I.V.E. (Living Intuitive Value-adding Environment) —— PCHES Program on Coupled Human and Earth Systems



MOSART-WM
+
GCAM-USA
+
TELL
+
GOWEST
=
nature

Retrieve

- Input data + metadata
- Versioned model source code
- Initial and boundary conditions
- Data transformation scripts
- Model outputs