Pacific Northwest
NATIONAL LABORATORY

PNNL DataHub: Building a Central Data Capability

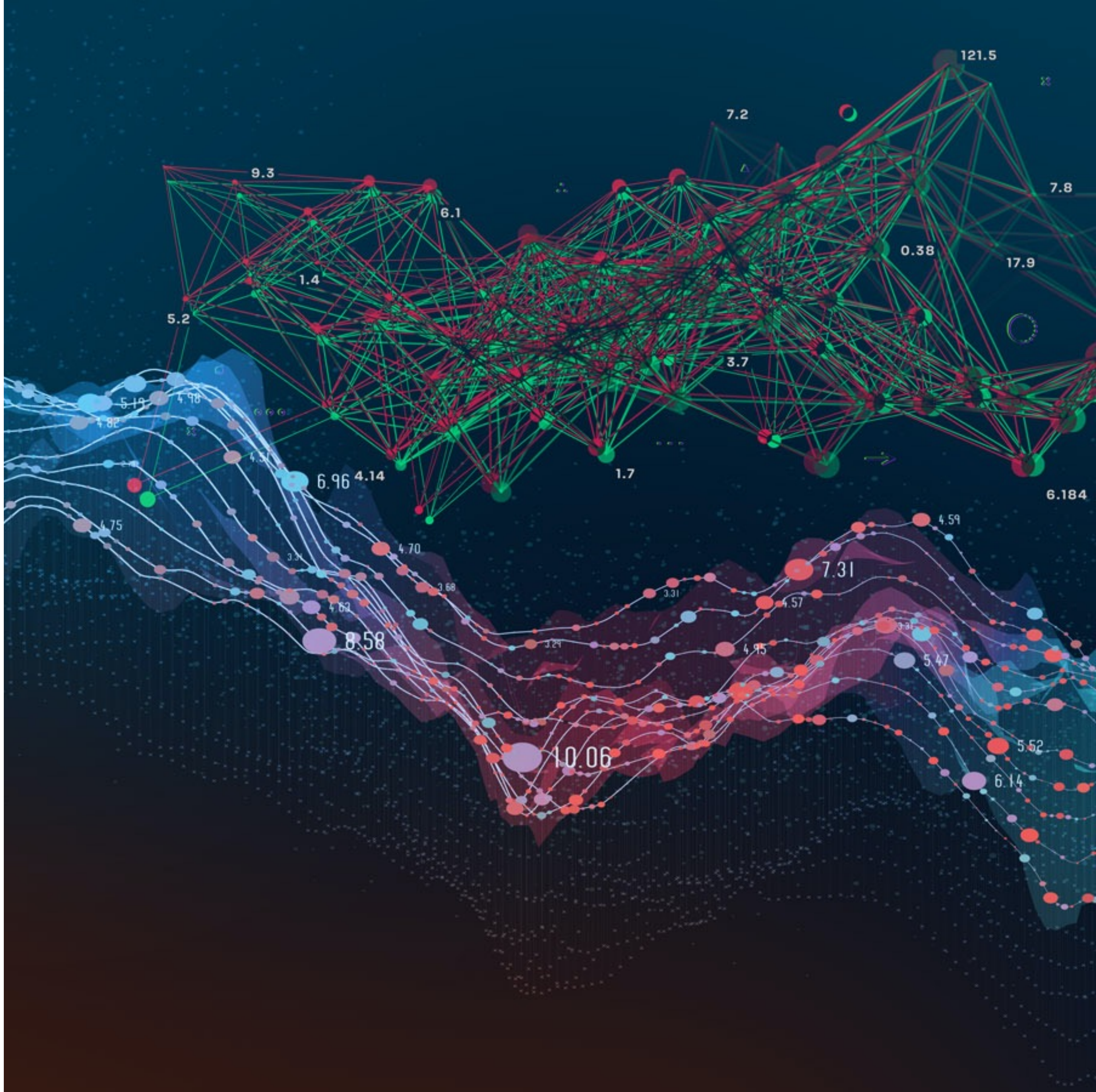
May 27, 2022

Michael Hofmockel
Ian Smith
Shannon Sheridan
Miriam Blake
Research computing, PNNL

U.S. DEPARTMENT OF ENERGY BATTELLE

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

PNNL-SA-172483



Problem

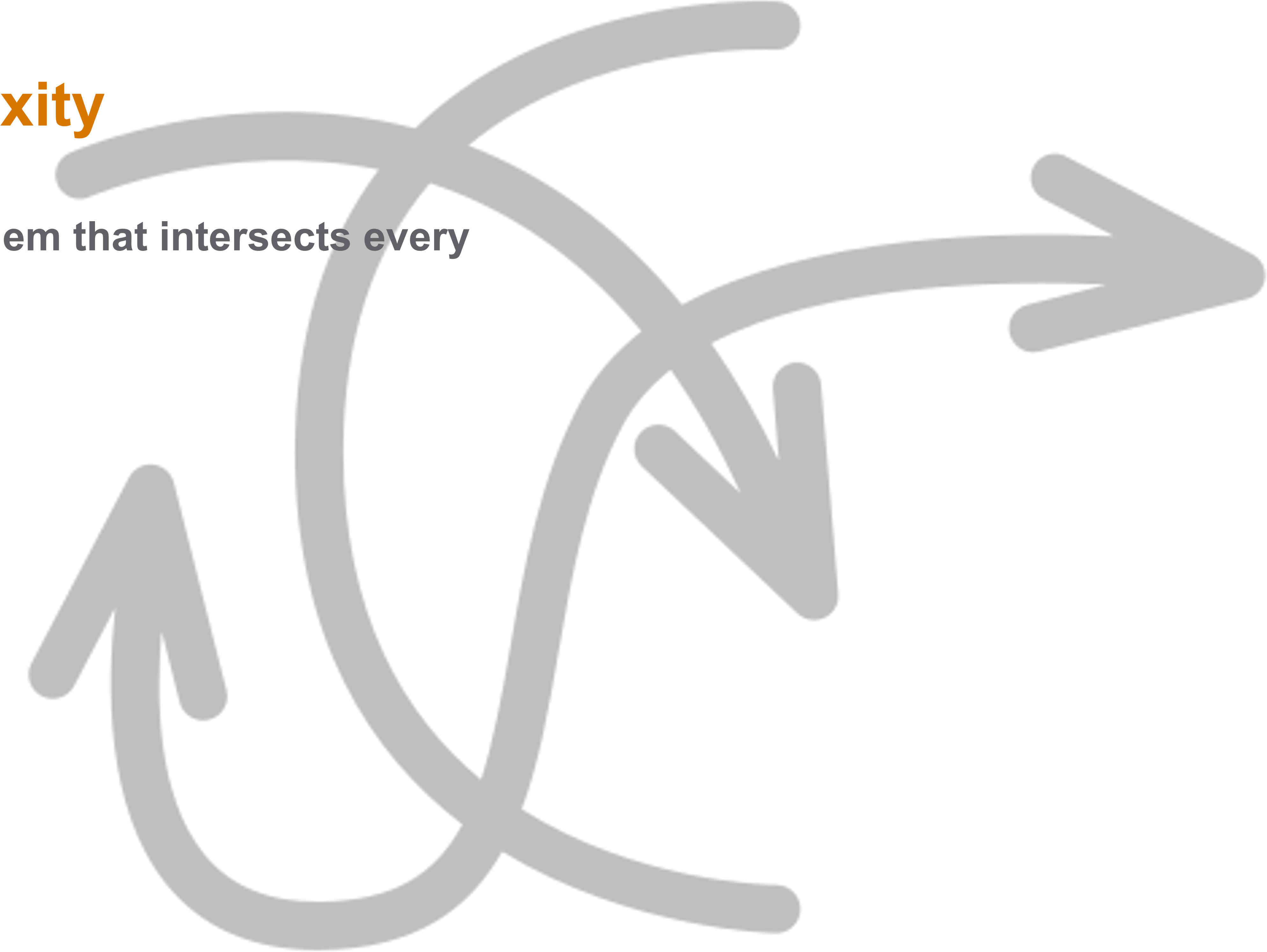
Estimate "data of record" creation at ~10PB/2,000 projects/year

- Institution cannot enumerate its research data
- Projects roll their own solutions/policies
- Preservation is limited to project lifetime and PI knowledge
- Commercial data management solutions are focused on back-office data that doesn't translate to research data.

Complexity

Multi-dimensional problem that intersects every part of the institution

- Commitment
- Understanding
- Standards
- Implementation
- Over generalization
- Education
- Training
- Sustainability



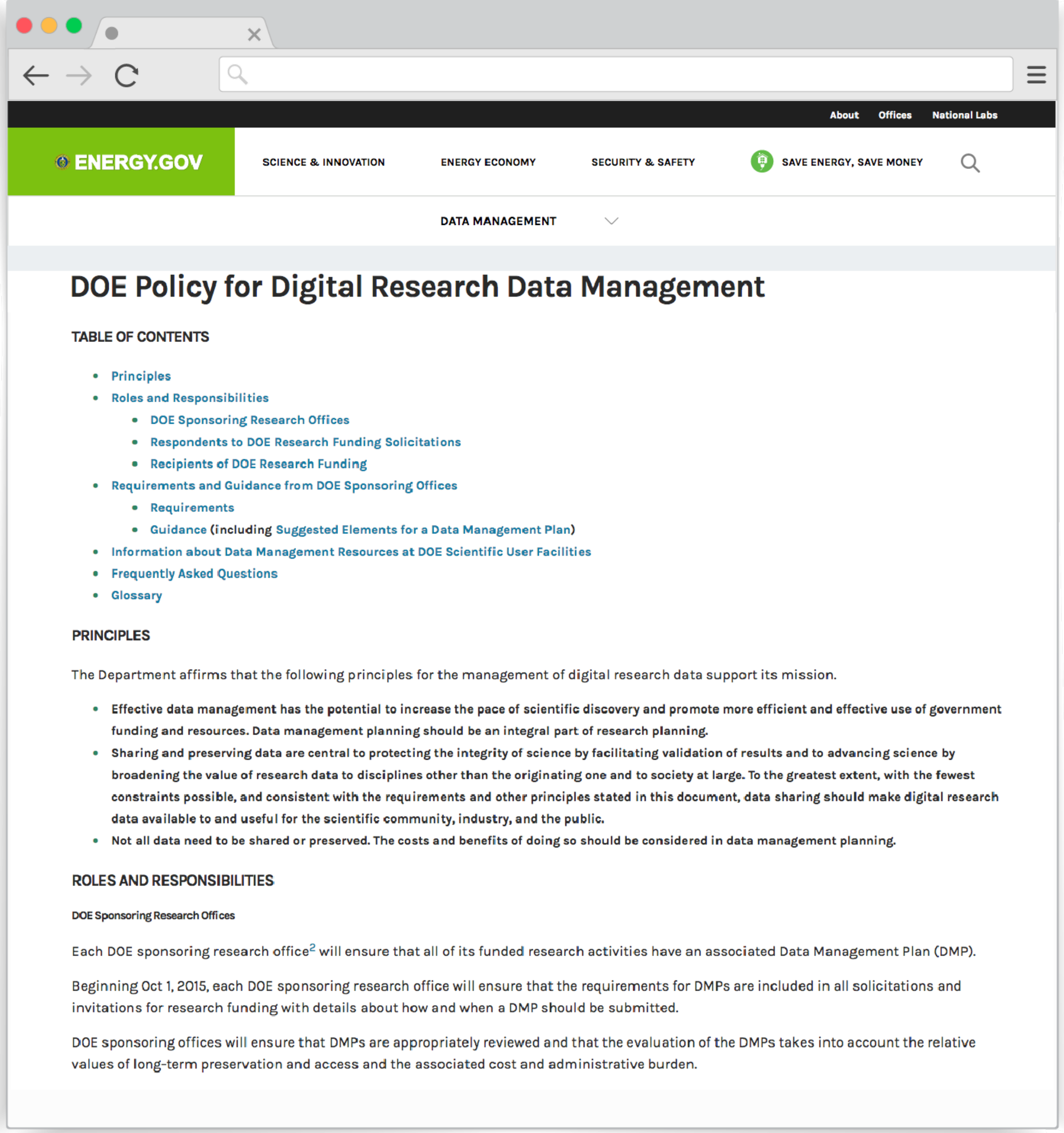
Pacific Northwest
NATIONAL LABORATORY

Our Sponsor’s Policy Should Be Our Policy

> ***“Sharing and preserving data are central to protecting the integrity of science** by facilitating validation of results and to advancing science by broadening the value of research data to disciplines other than the originating one and to society at large.”*

- Department of Energy

DOE Policy for Digital Research Data Management



DOE Policy for Digital Research Data Management

TABLE OF CONTENTS

PRINCIPLES

The Department affirms that the following principles for the management of digital research data support its mission.

- Effective data management has the potential to increase the pace of scientific discovery and promote more efficient and effective use of government funding and resources. Data management planning should be an integral part of research planning.
- Sharing and preserving data are central to protecting the integrity of science by facilitating validation of results and to advancing science by broadening the value of research data to disciplines other than the originating one and to society at large. To the greatest extent, with the fewest constraints possible, and consistent with the requirements and other principles stated in this document, data sharing should make digital research data available to and useful for the scientific community, industry, and the public.
- Not all data need to be shared or preserved. The costs and benefits of doing so should be considered in data management planning.

ROLES AND RESPONSIBILITIES

DOE Sponsoring Research Offices

Each DOE sponsoring research office[2] will ensure that all of its funded research activities have an associated Data Management Plan (DMP).

Beginning Oct 1, 2015, each DOE sponsoring research office will ensure that the requirements for DMPs are included in all solicitations and invitations for research funding with details about how and when a DMP should be submitted.

DOE sponsoring offices will ensure that DMPs are appropriately reviewed and that the evaluation of the DMPs takes into account the relative values of long-term preservation and access and the associated cost and administrative burden.

Research Data Management Vision

Increase the impact and scientific value of our research by elevating research data to be a principal PNNL product.

This will require transformation of our institutional culture, policies, and infrastructure to fully support the Research Data Life Cycle.

DataHub Advance Team - DATa

Data stewardship is the practice that ensures data is accessible, usable, safe, and trusted. It includes overseeing every aspect of the data lifecycle, promoting data quality and integrity.

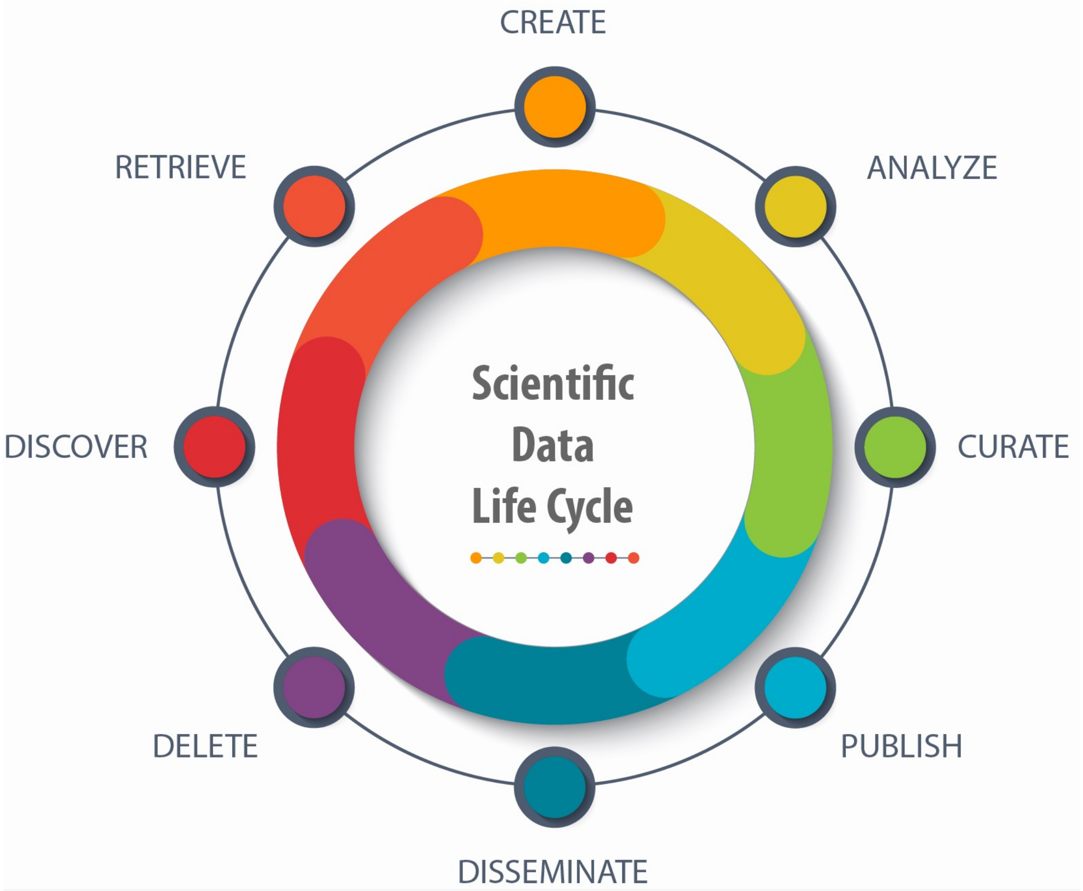
The **DataHub Advance Team** is the source for data stewardship support within Research Computing at PNNL. Utilizing DataHub, the Advance Team helps researchers address their institutional projects' data stewardship. Bringing together a wide variety of subject matter experts, the team can provide curated assistance.

Research groups around the lab have utilized the Advance Team to:

- Empower project team data stewards to create data policies for entire projects
- Develop data catalogs to track all data outputs from multiple collaborators internally
- Create project-wide naming conventions and directory structures
- Deploy a public web presence that semantically links data, publications, research staff, code, and scientific instruments

Pacific Northwest NATIONAL LABORATORY

PNNL DataHub Mission

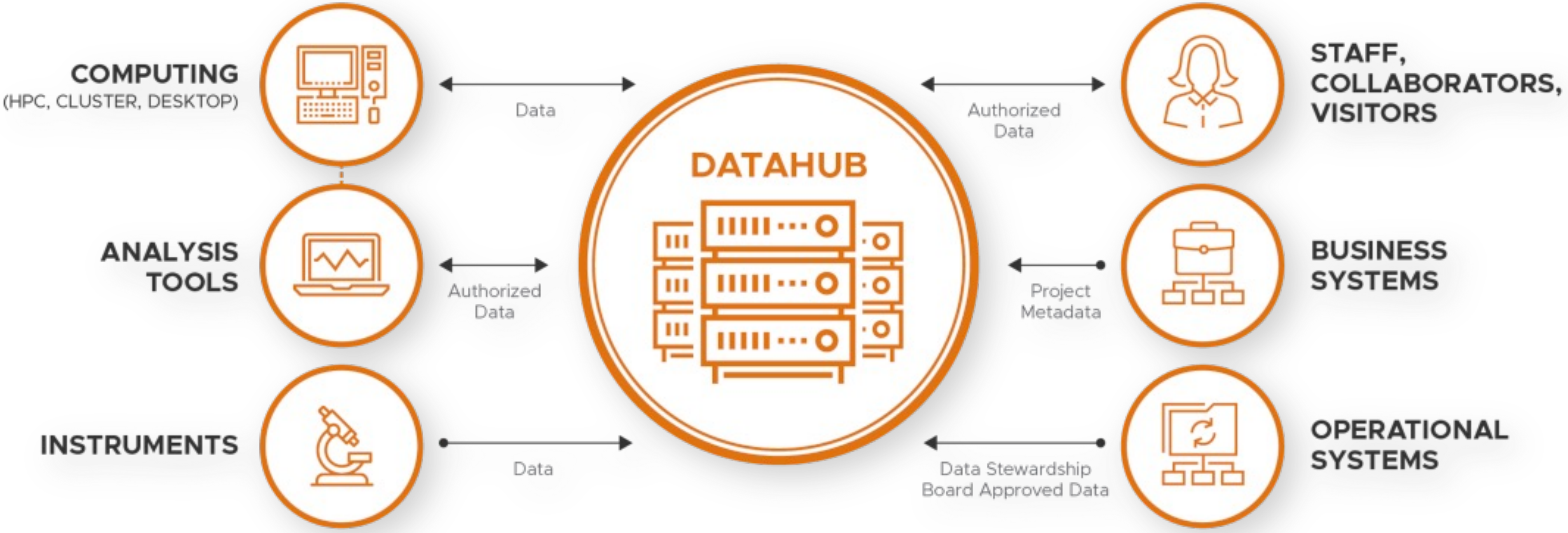


Findable Accessible Interoperable Reusable

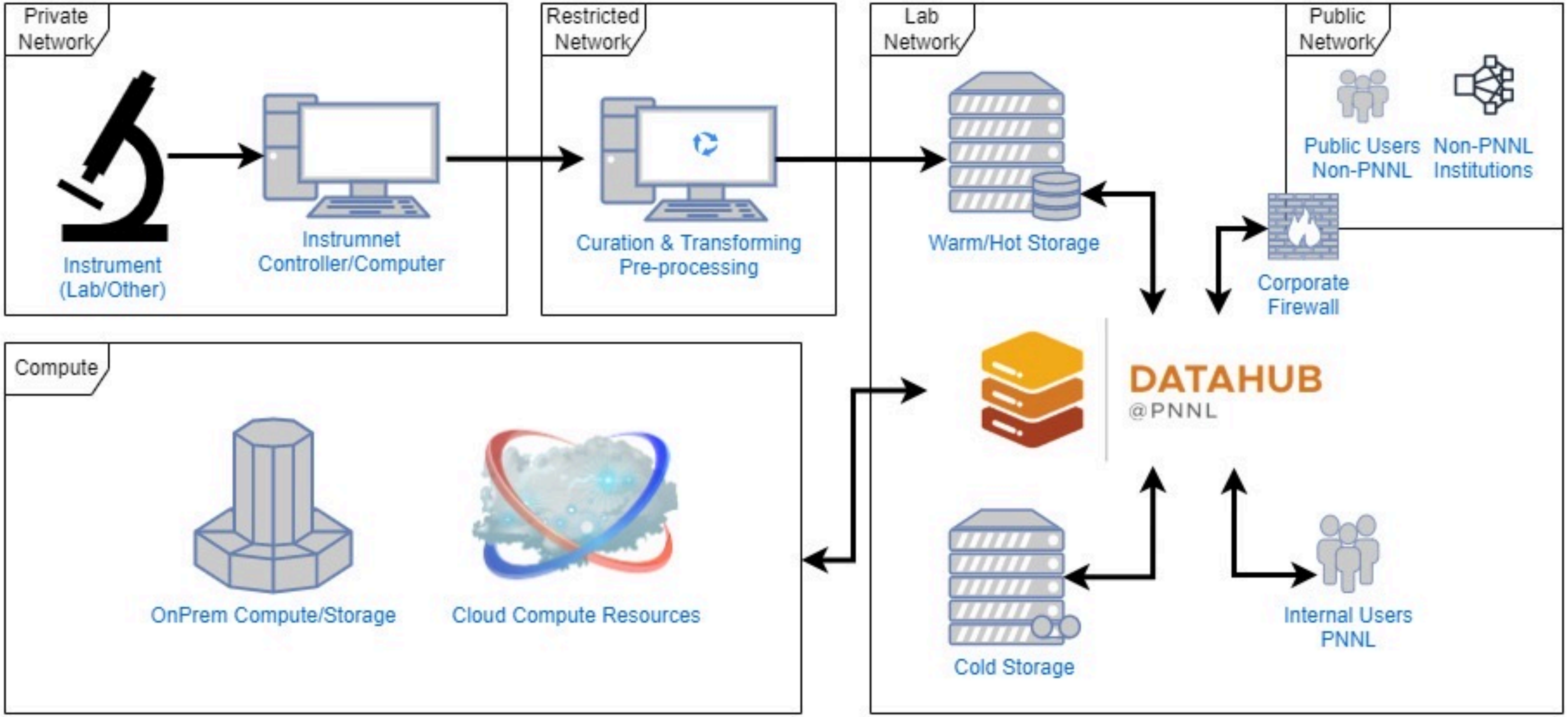
- Singular Registry for ALL data
- Easy and secure access
- Interoperable and standards based
- Research is reusable AND **reproducible**

Research Data Management and data literacy is embedded in institutional process and culture

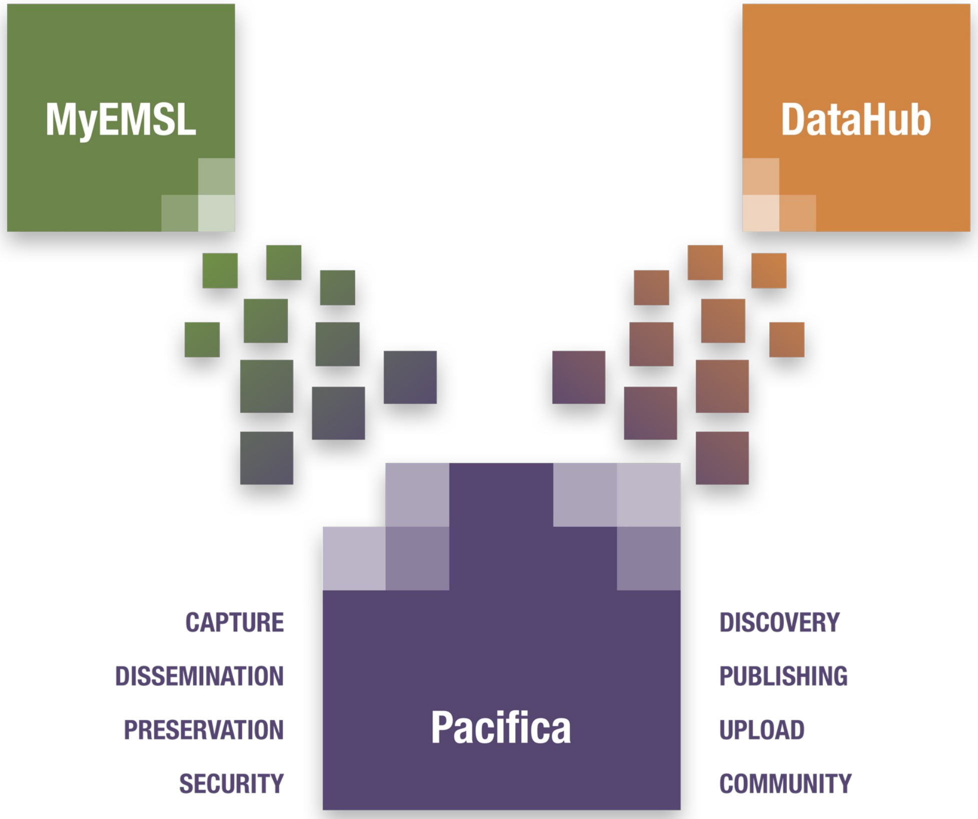
DataHub System at a glance



Example Data Pipeline



Lessons Learned



- Change is difficult
- Everyone wants access NOW!
- Starting simple is ok
- Document decisions and communicate
- Agree that a problem exists
- Commit to change
- Start a community around data

Pacific Northwest
NATIONAL LABORATORY

What are we working on now…

- Automation
- Standards
- Workforce
- Education
- Community
- Influencing Behaviors

Thank you

This research was funded by the National Nuclear Security Administration, Defense Nuclear Nonproliferation Research and Development (NNSA DNN R&D). The authors acknowledge important interdisciplinary collaboration with scientists and engineers from LANL, LLNL, MSTS, PNNL, and SNL.