Office of Science Data Management and Sharing Plan requirements

Basic Energy Science Advisory Committee Meeting
DOE Public Access Plan and Data Management Panel Discussion

Alexander Hexemer, Advanced Light Source, Lawrence Berkeley National Laboratory

Nicholas Schwarz, Advanced Photon Source, Argonne National Laboratory

Stuart Campbell, National Synchrotron Light Source II, Brookhaven National Laboratory

Vivek Thampy, Stanford Synchrotron Radiation Lightsource, SLAC

Jana Thayer, Linac Coherent Light Source, SLAC

Argonne NATIONAL LABORATORY BERKELEY LAB Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY

Advances in BES Light Sources

Computing addresses rapid data increases

More complex experiments

- Multi-modal experiments combine data from multiple samples, techniques, and facilities
- *In situ* and *in operando* experiments require real-time feedback and autonomous control
- Spectroscopy with 1000’s of scans in just a few seconds.

Sources—orders-of-magnitude brighter

- Facility upgrades:
  • NSLS-II • LCLS-II • APS-U • ALS-U • LCLS-II HE

Detectors—orders-of-magnitude faster

- Faster readout
- Larger arrays

Analyze and reconstruct massive multimodal data volumes

Identify and classify features and patterns across modes

Merge simulation & experiment data to drive experiments and new results

Execute experiments dynamically using real-time reduction and AI/ML

Scale of the Data Volume

In 5 years, DOE light sources are projected to

– generate 1 exabyte of data/year

– 10s-of-petaflop/s to 1-exaflop/s, peak computing power

1 exabyte/y = 1.5 million movies every day

- *Analyze every frame* in near real time; guide experiments
- Hundreds of experiment types require custom solutions

1 exaflop/s = 500,000 servers

- Fast networks (multiple Tbps)
- Storage
- Analysis infrastructure

Changing Landscape for Facilities and Users

compounding the computational and data challenges

The **user community is diverse**: a wide variety of backgrounds and domains

- Varying expectations on types and scales of computing capabilities & services provided by the facilities

There is an **increasing digital divide** within the scientific user community

- Currently, few user groups have the ability to manage/process their data. This challenge will only increase.

Increased interest in **FAIR, open data,** and **data interoperability**

- The role of the facilities is unclear: facilities do not have the infrastructure in place to consistently collect, curate, archive, and disseminate data and metadata at the anticipated scale required

What does this mean for users?

- New science opportunities
- Take advantage of the wealth of facility data to augment your own science
- Develop & test new algorithms on open and shared data
- Train ML models on large shared data sets
- Use existing ML models to accelerate knowledge extraction from data
- Make data FAIR

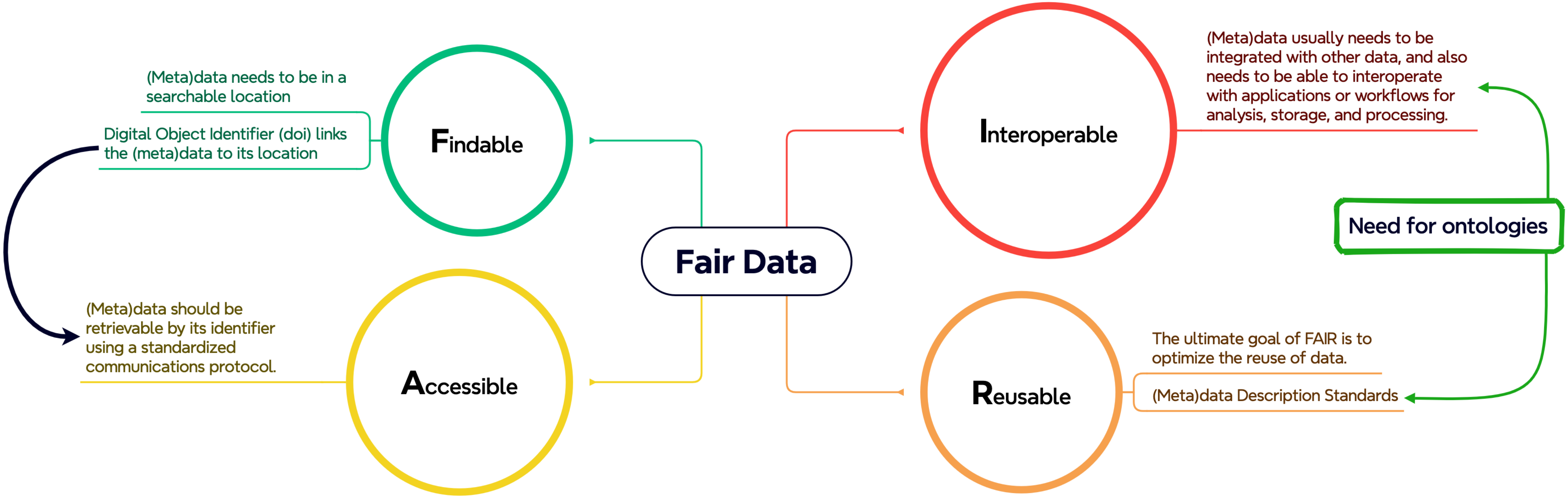
FAIR Data

Findable: Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

Accessible: (Meta)data are understandable to humans and machines. Data is deposited in a trusted repository.

Interoperable: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

Reusable: Data and collections have a clear usage licenses and provide accurate information on provenance.



Fair Data

Findable

(Meta)data needs to be in a searchable location

Digital Object Identifier (doi) links the (meta)data to its location

Accessible

(Meta)data should be retrievable by its identifier using a standardized communications protocol.

Interoperable

(Meta)data usually needs to be integrated with other data, and also needs to be able to interoperate with applications or workflows for analysis, storage, and processing.

Reusable

The ultimate goal of FAIR is to optimize the reuse of data.

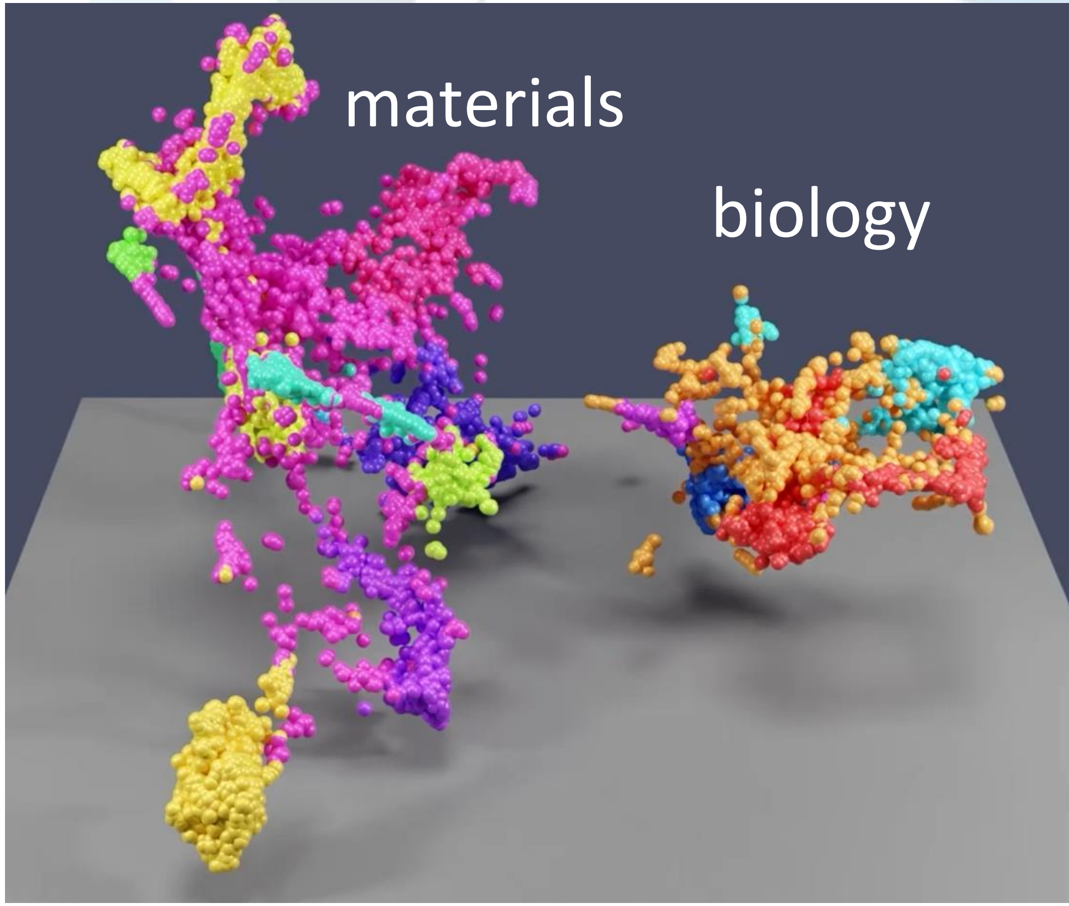
(Meta)data Description Standards

Need for ontologies

Challenges with FAIR data

(a scientific user facility perspective)

- The **ontology** in each field must be well-structured and precise to ensure clear communication and data interoperability.
- Not all the **metadata** of an experiment is available to the user facilities (e.g., material synthesis)
- Data sets can be ***very* large** & difficult to handle, stored and served
- (Meta)data of a single study can be spread across **multiple facilities**
- What are the implications of **data deletion** (raw or derived data)?
- **Authentication** and authorization challenges across facilities
- Increased **complexity** of working with data



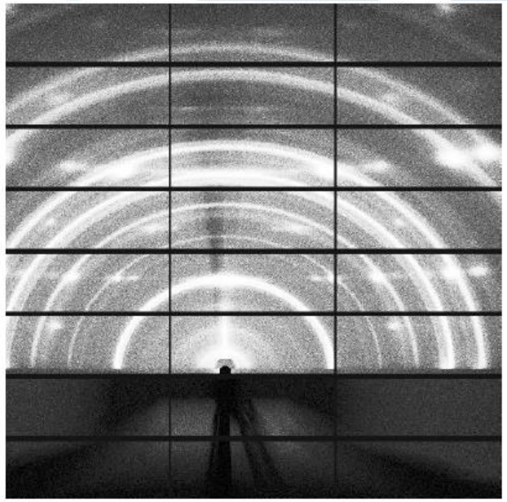
Clustering of ALS publications using a LLM and UMAP to show the wealth of information and diversity of research areas

Opportunities using FAIR data

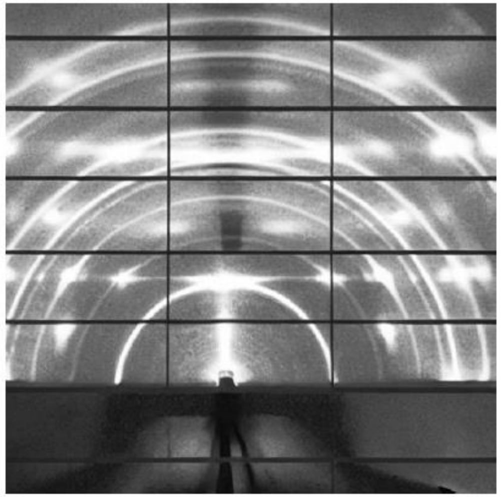
(a scientific user facility perspective)

- Data reusability
- Reproducibility of experimental results & analysis
- Development and testing of new algorithms on well-described data
- Common ontologies allows for better cross-facility collaboration
- Seamless integration of data with (HPC) compute resources
- Improved training data quality for ML models with AI ready data
- Opportunity to share data and trained ML models
- Opportunities for Unsupervised and Semi-Supervised Learning
- Using generative AI to create data sets specific to a given experiment

GIWAXS
Experimental data



GIWAXS
AI-generated data



prompt:
"GIWAXS data with rings and peaks"

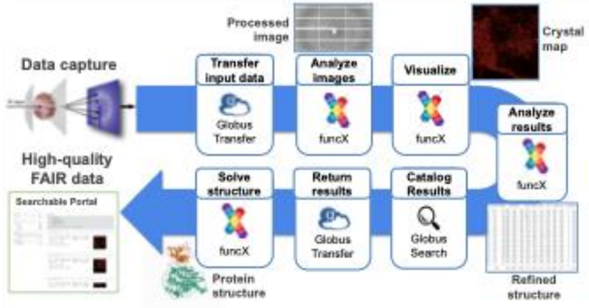
Data Portals and Access across Light Sources

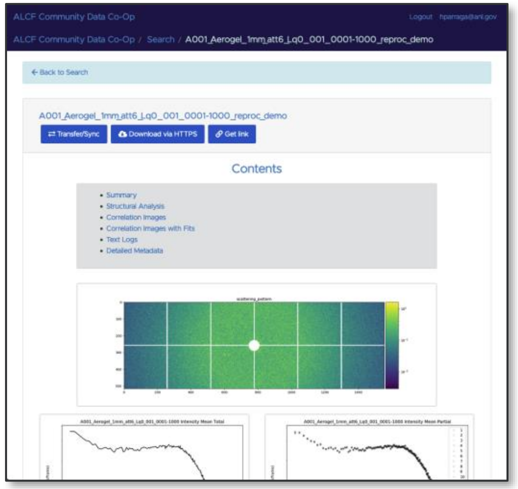
APS

Argonne NATIONAL LABORATORY

globus

bluesky





ALCF Community Data Co-Op   Logout hparraga@anl.gov

ALCF Community Data Co-Op / Search / A001_Aerogel_1mm_att6_Lq0_001_0001-1000_reproc_demo

← Back to Search

A001_Aerogel_1mm_att6_Lq0_001_0001-1000_reproc_demo

Contents

- Summary
- Structural Analysis
- Correlation Images
- Correlation Images with Fits
- Text Logs
- Detailed Metadata

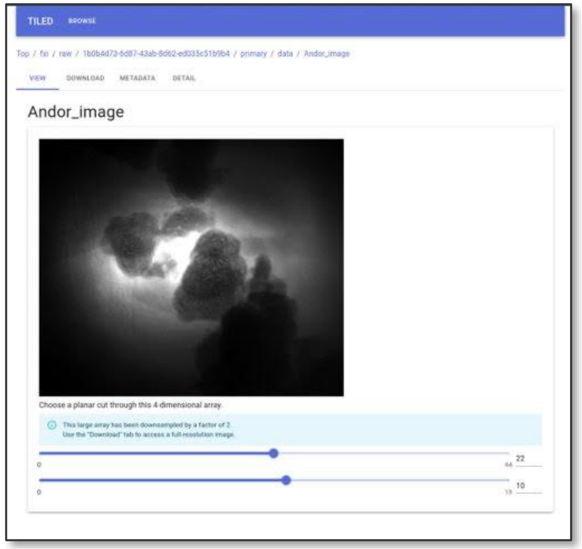
NSLS II

Brookhaven National Laboratory

bluesky

globus

tiled



Top / fxi / raw / 1b0b4d73-6d87-43ab-8d62-ed033c51b9b4 / primary / data / Andor_image

Andor_image

Choose a planar cut through this 4 dimensional array.

This large array has been downsampled by a factor of 2.
Use the "Download" tab to access a full-resolution image.

ALS

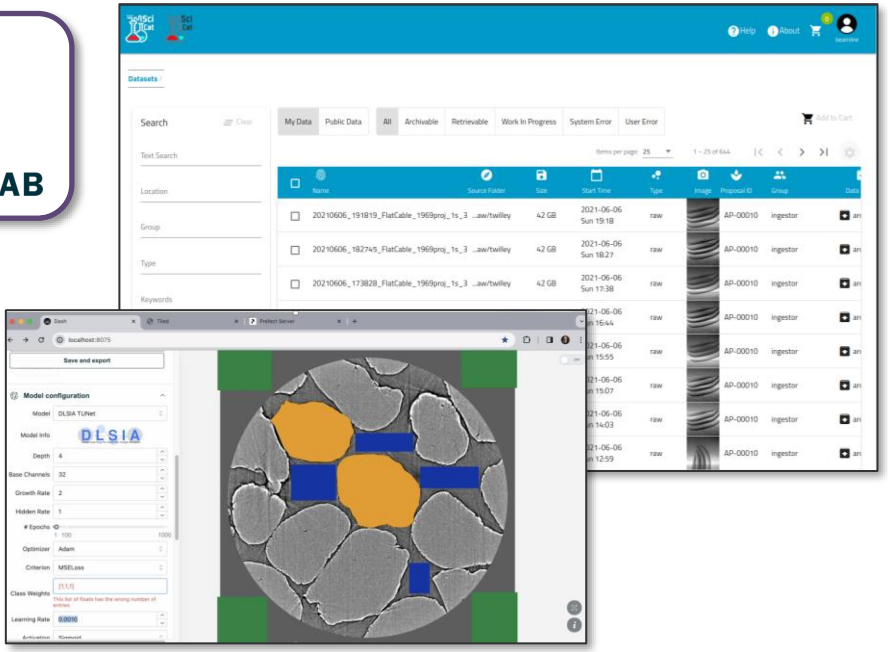
BERKELEY LAB

tiled

globus

SciCat

bluesky



Items per page: 25   1 – 25 of 644

Name	Source Folder	Size	Start Time	Type	Image	Proposal ID	Group
20210606_191819_FlatCable_1969proj_1s_3	...aw/twilley	42 GB	2021-06-06 Sun 19:18	raw		AP-00010	ingestor
20210606_182745_FlatCable_1969proj_1s_3	...aw/twilley	42 GB	2021-06-06 Sun 18:27	raw		AP-00010	ingestor
20210606_173828_FlatCable_1969proj_1s_3	...aw/twilley	42 GB	2021-06-06 Sun 17:38	raw		AP-00010	ingestor
			...21-06-06 ...n 16:44	raw		AP-00010	ingestor
			...21-06-06 ...n 15:55	raw		AP-00010	ingestor
			...21-06-06 ...n 15:07	raw		AP-00010	ingestor
			...21-06-06 ...n 14:03	raw		AP-00010	ingestor
			...21-06-06 ...n 12:59	raw		AP-00010	ingestor

localhost:8075

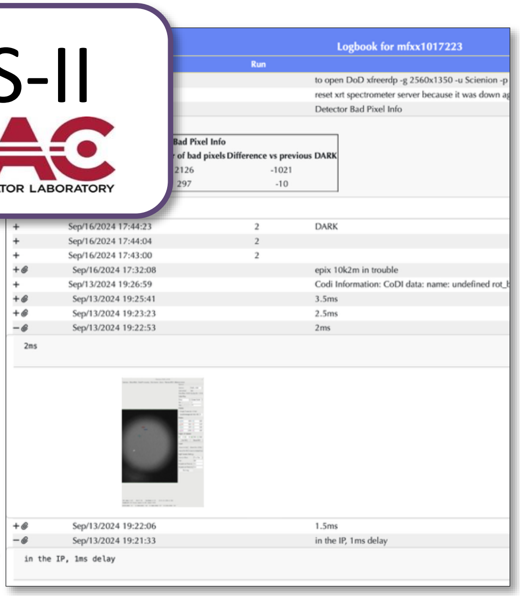
Model configuration

Model	DLSIA TUNet
Model Info	DLSIA
Depth	4
Base Channels	32
Growth Rate	2
Hidden Rate	1
# Epochs	1 100 1000
Optimizer	Adam
Criterion	MSELoss
Class Weights	[1,1,1]
Learning Rate	0.0010
Activation	Sigmoid

This list of floats has the wrong number of entries

LCLS-II

SLAC NATIONAL ACCELERATOR LABORATORY



Logbook for mfxx1017223

Run

to open DoD xfreerdp -g 2560x1350 -u Scienion -p

reset xrt spectrometer server because it was down ag

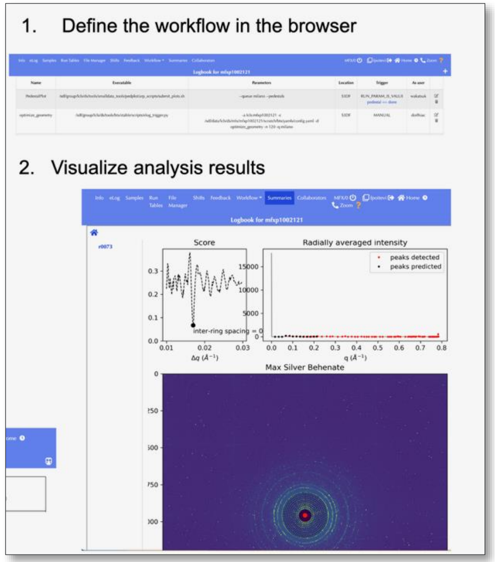
Detector Bad Pixel Info

Bad Pixel Info	
of bad pixels	Difference vs previous DARK
2126	-1021
297	-10

```
+   Sep/16/2024 17:44:23   2   DARK
+   Sep/16/2024 17:44:04   2
+   Sep/16/2024 17:43:00   2
+   Sep/16/2024 17:32:08       epix 10k2m in trouble
+   Sep/13/2024 19:26:59       Codi Information: CoDI data: name: undefined rot_b
+   Sep/13/2024 19:25:41       3.5ms
+   Sep/13/2024 19:23:23       2.5ms
−   Sep/13/2024 19:22:53       2ms
2ms
+   Sep/13/2024 19:22:06       1.5ms
−   Sep/13/2024 19:21:33       in the IP, 1ms delay
in the IP, 1ms delay
```

LCLS

eLog

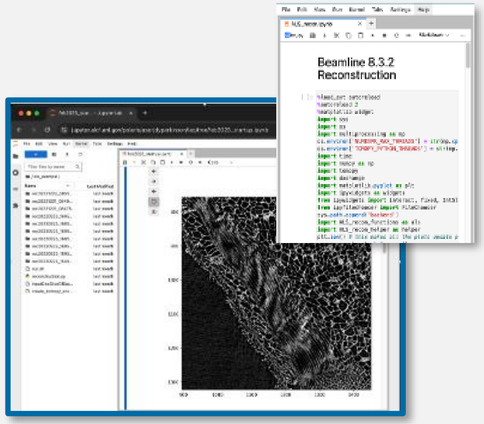


1. Define the workflow in the browser

2. Visualize analysis results

Working Together Across BES & ASCR Facilities

ALS Reconstructions: **ALCF & NERSC**

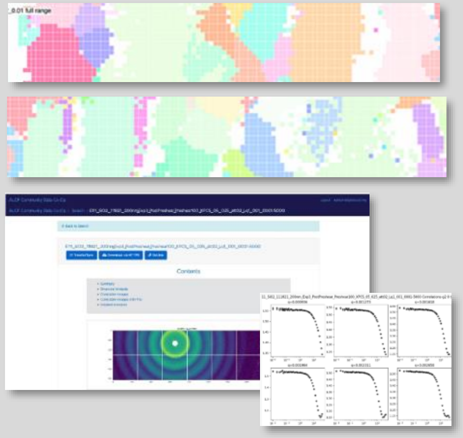


Beamline 8.3.2 Reconstruction

```
%load_ext autoreload
%autoreload 2
%matplotlib widget
import sys
import os
import multiprocessing as mp
os.environ['NUMEXPR_MAX_THREADS'] = str(mp.cp
os.environ['TOMOPY_PYTHON_THREADS'] = str(mp.
import time
import numpy as np
import tomopy
import dxchange
import matplotlib.pyplot as plt
import ipywidgets as widgets
from ipywidgets import interact, fixed, IntSl
from ipyfilechooser import FileChooser
sys.path.append('backend')
import ALS_recon_functions as als
import ALS_recon_helper as helper
```

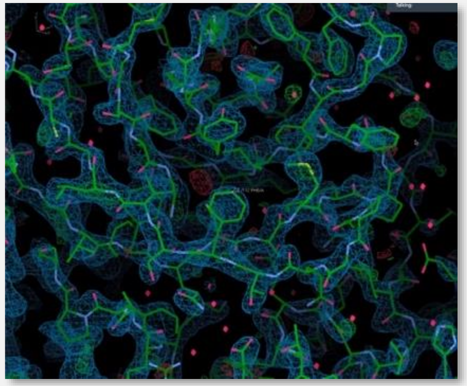
- Globus Transfer between ALS, NERSC and ALCF
- Tomography reconstruction on NERSC and using Globus Compute on ALCF
- Results are transferred back to ALS or NERSC

APS On-Demand Workflows: **ALCF**



- ALCF Polaris system: continuous on-demand data processing
- Operational workflows for over 10 techniques
- Globus tools provide workflow and web portal services

LCLS Data Analytics: **NERSC, ALCF, OLCF**

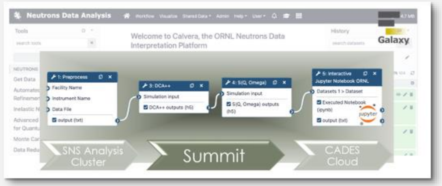


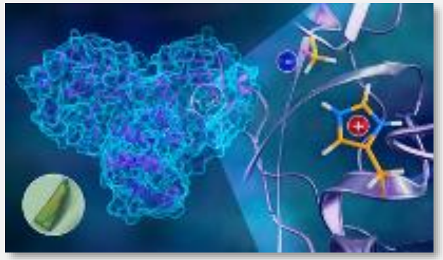
- Automated SFX pipeline at NERSC; demonstrated during COVID-19 LCLS experiments
- AI training on OLCF Summit+
- Prototype ptychography workflow demonstrated on ALCF Polaris

NSLS-II Prototype Pipeline: **ALCF**

- Export to file, transfer using Globus
- Processing XPCS using Jupyter Notebooks on ALCF using Globus Flows
- Next, integrate with bluesky/tiled for data access and output

SNS/HFIR Data Processing: **OLCF**



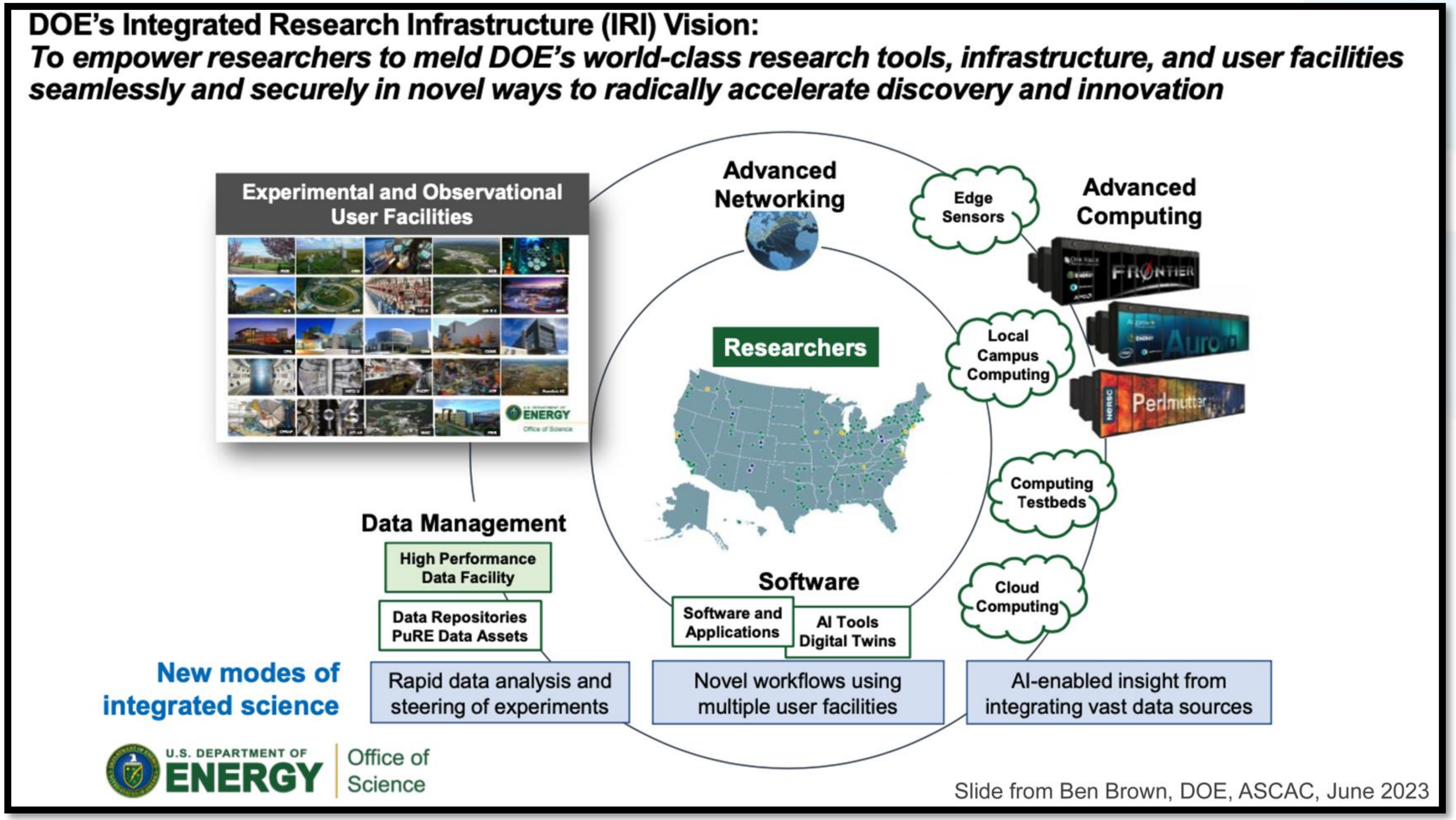


- Web-based data platform, integration with OLCF
- Neutron users perform calculations as part of an analysis workflow
- Intra-experiment training to predict protonation states of active protein sites for neutron MX

Integrated Research Infrastructure (IRI)

The BES Light Sources are one of the initial *Pathfinder* projects

- The ALS, APS and NSLS-II are a single combined pathfinder
- Target beamlines and techniques are being finalizes



DOE's Integrated Research Infrastructure (IRI) Vision:
To empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation

Experimental and Observational User Facilities

Advanced Networking

Edge Sensors

Advanced Computing

Researchers

Local Campus Computing

Computing Testbeds

Cloud Computing

Data Management

High Performance Data Facility

Data Repositories PuRE Data Assets

Software

Software and Applications

AI Tools Digital Twins

New modes of integrated science

Rapid data analysis and steering of experiments

Novel workflows using multiple user facilities

AI-enabled insight from integrating vast data sources

U.S. DEPARTMENT OF ENERGY | Office of Science

Slide from Ben Brown, DOE, ASCAC, June 2023

U.S. DEPARTMENT OF ENERGY | Office of Science

THE DOE OFFICE OF SCIENCE

Integrated Research Infrastructure Architecture Blueprint Activity

FINAL REPORT
2023

Summary of High-Priority Computing Developments across Light Sources

1. **Data management and workflow tools** that integrate beamline instruments with computing & storage, for use during experiment, as well as facile user access for post-experiment analysis.

2. **Real-time data analysis capabilities** to significantly reduce data volumes and provide feedback during experiments, improving data quality and driving autonomous experiments.

3. **On-demand utilization of computing environments** to enable quasi-real-time data processing

4. **Data storage and archival resources** to house the increasing amounts of valuable scientific data produced by the BES Light Sources in a common, smart data portal.

5. **Easy-to-use** solution to provide an inclusive environment for researchers at SUFs