

Linking scientific instruments and computation: Patterns, technologies, and experiences

Kyle Chard

University of Chicago

Argonne National Laboratory

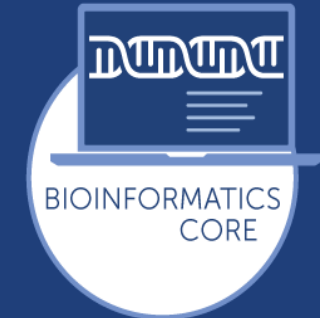
chard@uchicago.edu





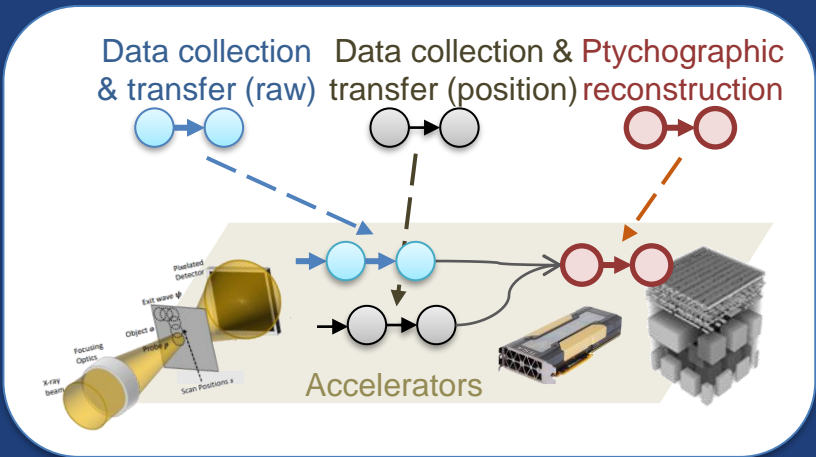
Towards self-driving instruments

- **Exponential growth in the rate that instruments perform measurements**
 - Data generated at GB+/s and 100+TB/day
 - **Analysis requires significant online computing capacity**
 - HPC resources, GPUs, AI accelerators
 - **Computation (and AI) can steer experiments**
 - Workflows involve both humans and machines
- We need new methods to automate these workflows and coordinate actions and resources across experiment and compute environments





Example instrument patterns

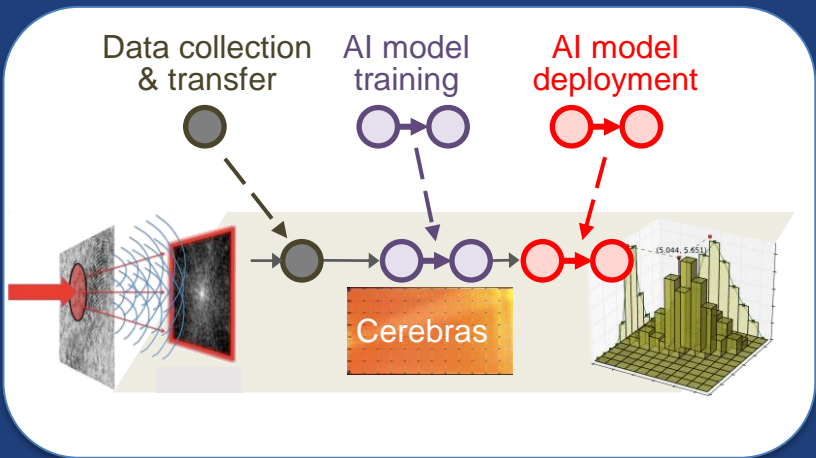
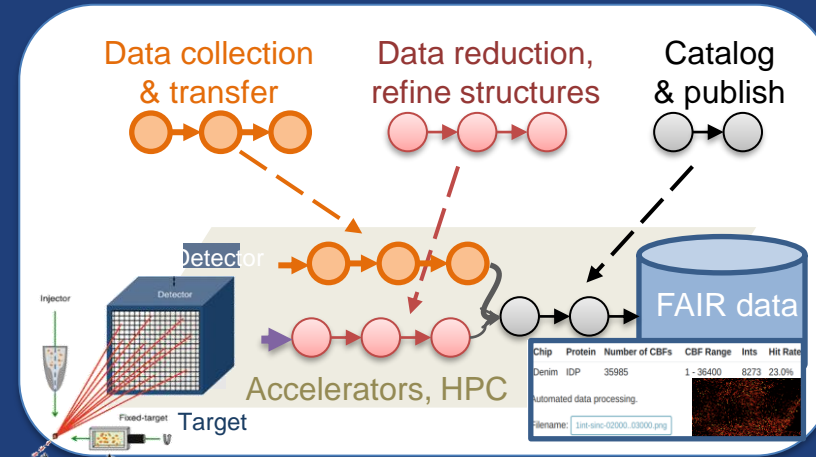


Ptychography

Reconstruction algorithms applied to data as they are acquired

Serial Crystallography

Data are reduced and made available to scientists and collaborators

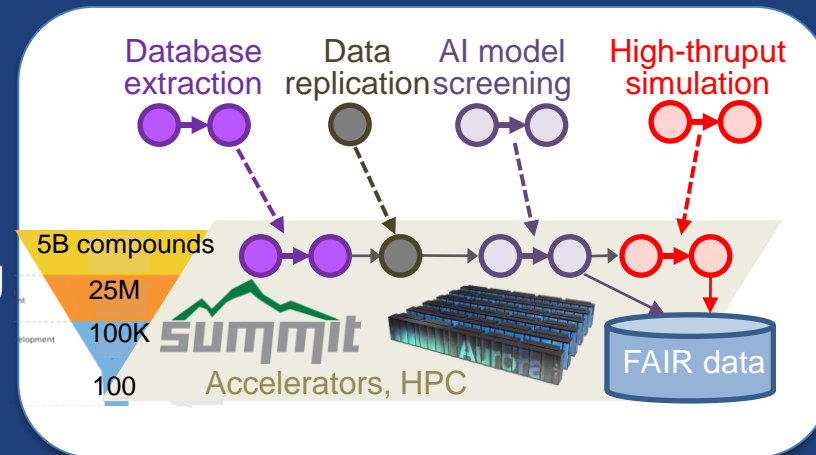


HEDM

Models trained on acquired data and deployed at the edge for fast inference

COVID

Data replicated across sites to apply AI screening methods and AI-guided simulation





These patterns highlight diverse automation needs

- **Support various *actions***
 - Transfer, compute, ingest in a search index, associate a persistent identifier, modify access permissions
- **Robust orchestration spanning several locations**
 - Enable remote control of actions in different places
- **Authentication/authorization model to provide secure management of remote operations across the computing continuum**



Globus automation capabilities



Timer Service

Scheduled and recurring transfers
(*a.k.a. Globus cron*)



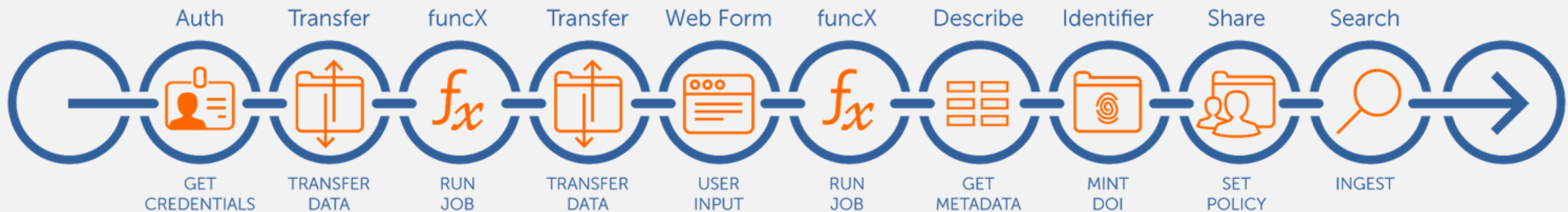
Globus Flows service

Comprehensive task (data and compute) orchestration with human in the loop interactions



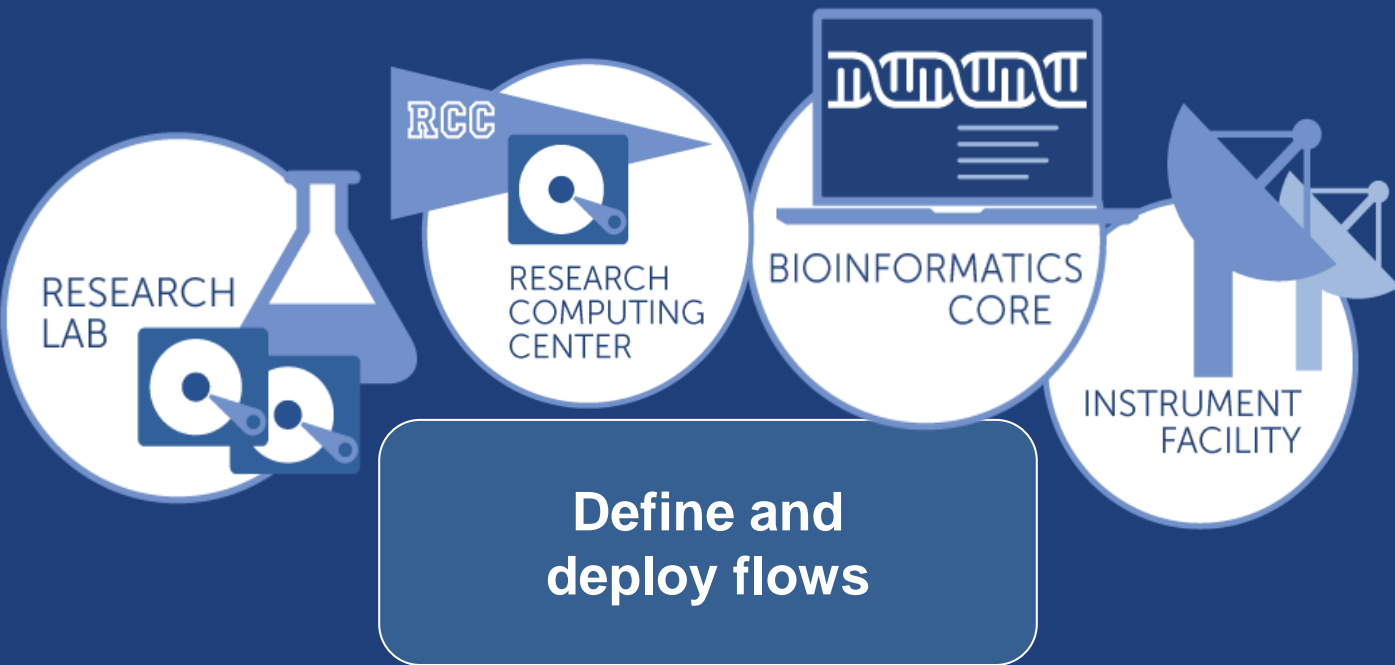
The Globus Flows service

- A platform for defining, executing, and sharing distributed research automation flows
- Flows comprise **Actions**
- **Action Providers:** Called by Flows to perform tasks





Create and deploy flows



- Use declarative language (JSON or YAML)
- Set input schema
- Set policy for use of the flow

```
{
  "States": {
    "SetPermission": {
      "End": true,
      "Type": "Action",
      "Comment": "Grant read permission on the data to a Globus user or group",
      "ActionUrl": "https://actions.automate.globus.org/transfer/set_permission",
      "Parameters": {
        "path.$": "$.input.destination.path",
        "operation": "CREATE",
        "permissions": "r",
        "principal.$": "$.input.principal_identifier",
        "endpoint_id.$": "$.input.destination.id",
        "principal_type.$": "$.input.principal_type"
      },
      "ResultPath": "$.SetPermission"
    },
    "TransferFiles": {
      "Next": "SetPermission",
      "Type": "Action",
      "Comment": "Transfer to a guest collection",
      "WaitTime": 60,
      "ActionUrl": "https://actions.automate.globus.org/transfer/transfer",
      "Parameters": {
        "transfer_items": [
          {
            "recursive.$": "$.input.recursive_tx",
            "source_path.$": "$.input.source.path",
            "destination_path.$": "$.input.destination.path"
          }
        ],
        "source_endpoint_id.$": "$.input.source.id",
        "destination_endpoint_id.$": "$.input.destination.id"
      },
      "ResultPath": "$.TransferFiles"
    }
  },
  "Comment": "Transfer files to a guest collection and set access permissions",
  "StartAt": "TransferFiles"
}
```



Start flows: Guided input

Start - Two Stage Globus Transfer

Guided Advanced

Source
Globus-provided flows require that at least one collection is managed under a subscription.

Collection

mid

- UChicago RCC Midway**
Owner: ucrcc@globusid.org
University of Chicago Research
Computing Center Midway cluster server
- UChicago RCC Midway3**
Owner: ucrcc@globusid.org
University of Chicago Research
Computing Center Midway3 Cluster

Path

/~/my-data-for-sharing Browse

Destination
Globus-provided flows require that at least one collection is managed under a subscription.

Collection

Guided Advanced

of layout and function

Timeout
This is an example property description for sleep (number)

Label

Notify user

true
 false

Choose input type:

null boolean string array number

Dynamic forms generated from input schema



Managing runs at scale

Runs started on

01/09/2021

12:00 AM

mm/dd/yyyy

--:-- --

Run Status

ACTIVE (365)

WAITING (0)

FAILED (8113)

COMPLETED (49872)

Started By

Me (includes all your linked identities)

Nickolaus Saint (nickolaussaint@globusid.org)

Rafael Vescovi (ravescovi@globusid.org)

APS_8IDI (5121b12c-9ef4-4c76-91aa-0dc6a8c9bcad@clients.auth.globus.org)

Hannah Parraga (hparraga@anl.gov)

Rachana Ananthakrishnan (ranantha@uchicago.edu)

Limit to Flow

search Flows or lookup by UUID

Run Tags

aps (44936)

xpcs (40547)

ssx (4409)

PEARC_Test (57)

Trigger_Tutorial (26)

RA_Demo (20)

demo (19)

glacier (19)

tutorial (17)

AMNH_Demo (4)

globusworld2022 (3)

my-first-flow (3)

Demo (2)

NCAR Tutorial (2)

Automate_Demo (1)

DemoTest (1)

Demo_RA (1)

Feb2023 (1)

MTB (1)

PEARC (1)

PEARC22 (1)

Sample 67B (1)

Tag1 (1)

TestTag (1)

Clear All Filters



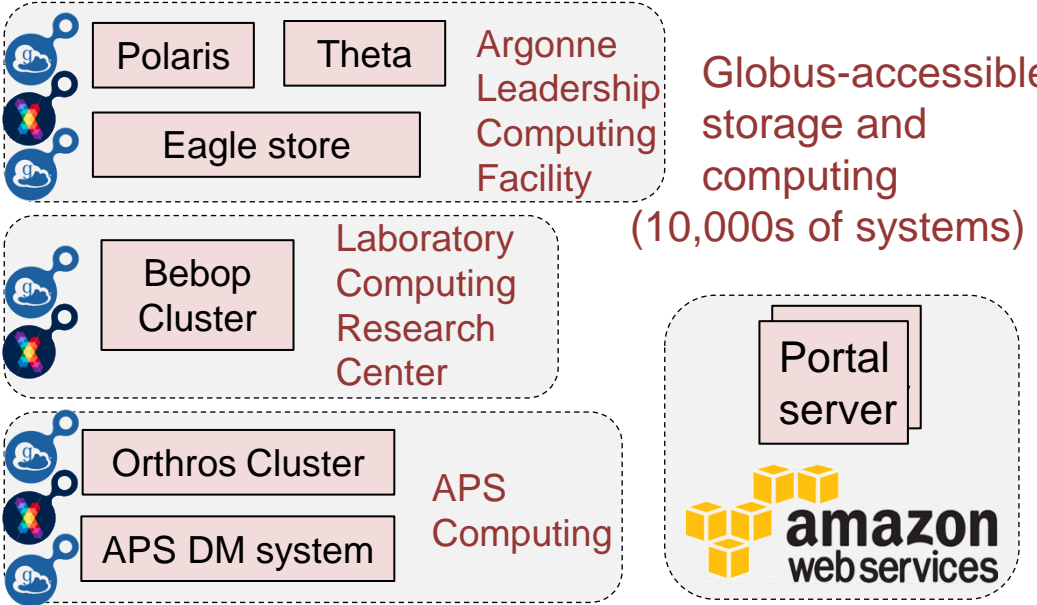
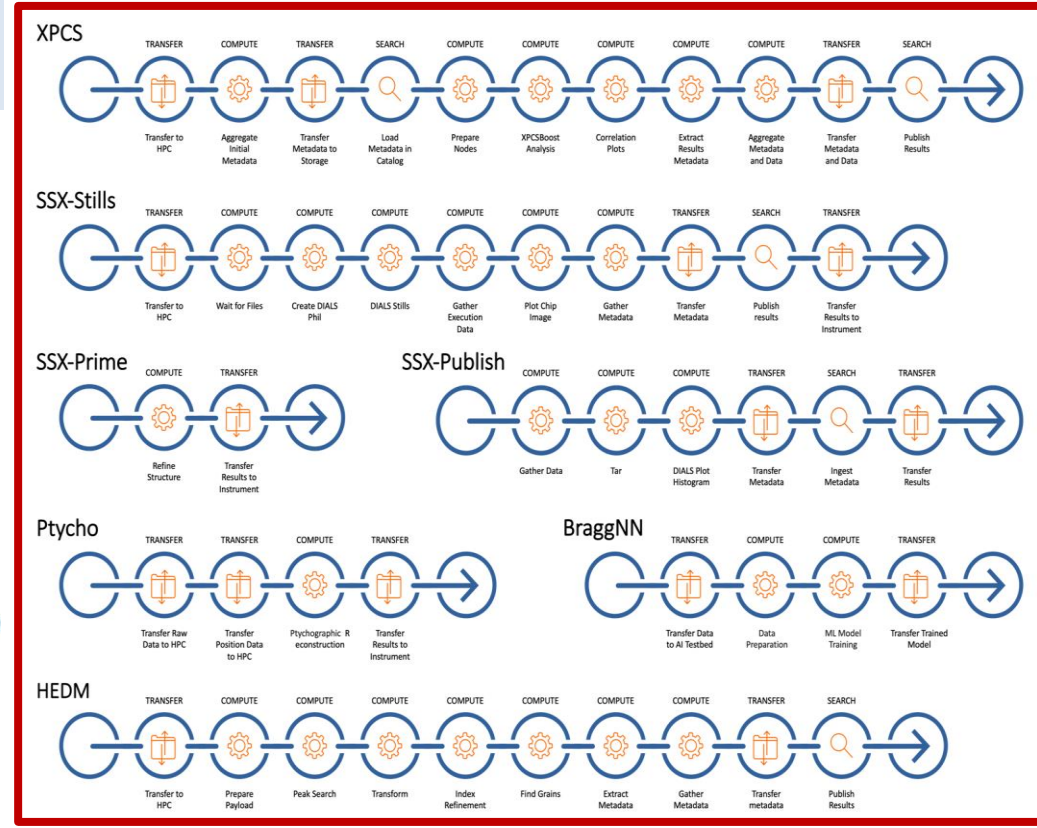
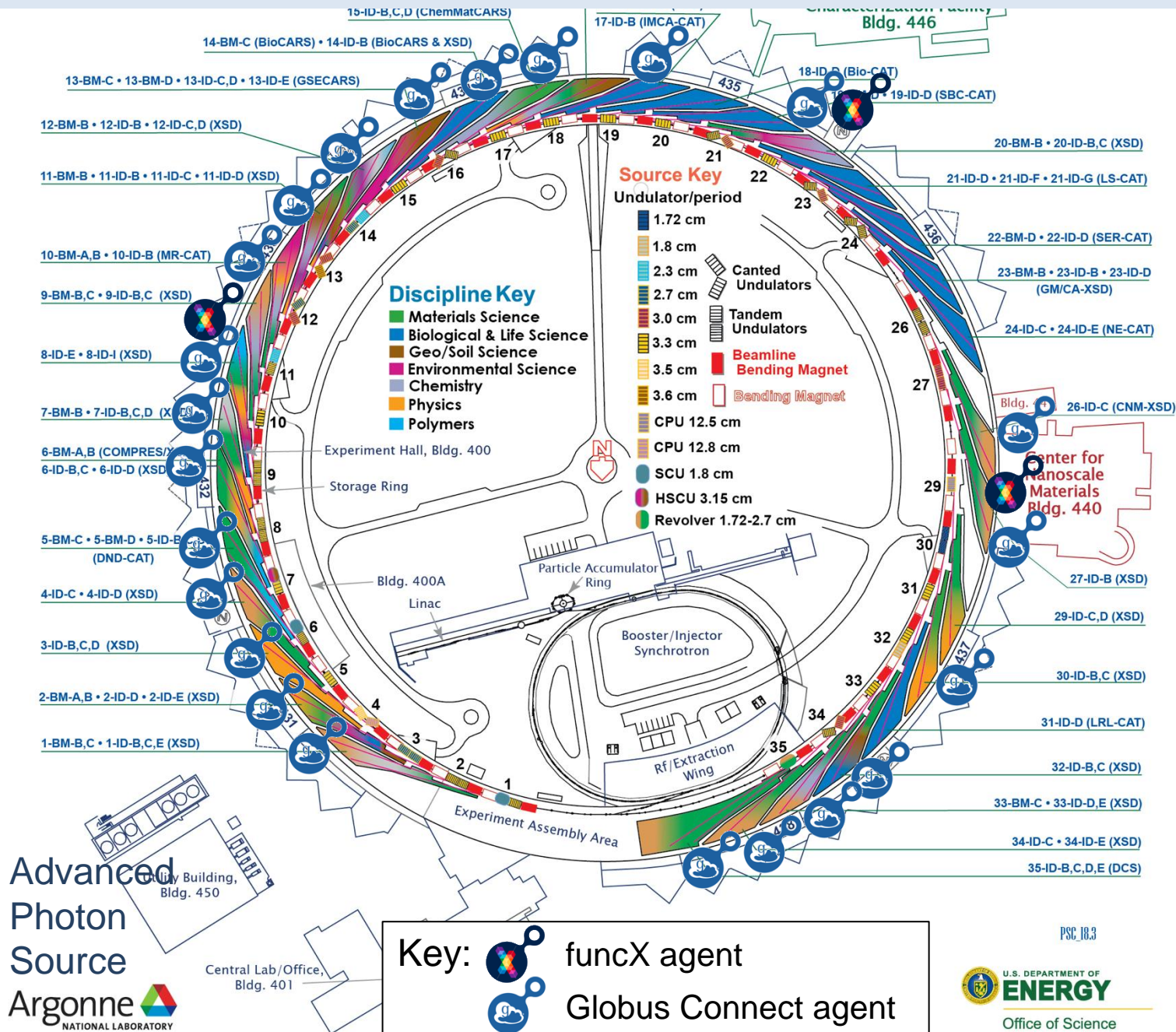
Running flows across the computing continuum requires a universal data and compute fabric



Globus Auth:

standards compliant identity and access management platform

Endpoints at experimental facilities



Key:

- funcX agent
- Globus Connect agent

Advanced Photon Source
Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY
Office of Science



Integrating APS and ALCF

One-time configuration at ALCF

APS experiments

Using a service account per beam line

No human involved in the workflows

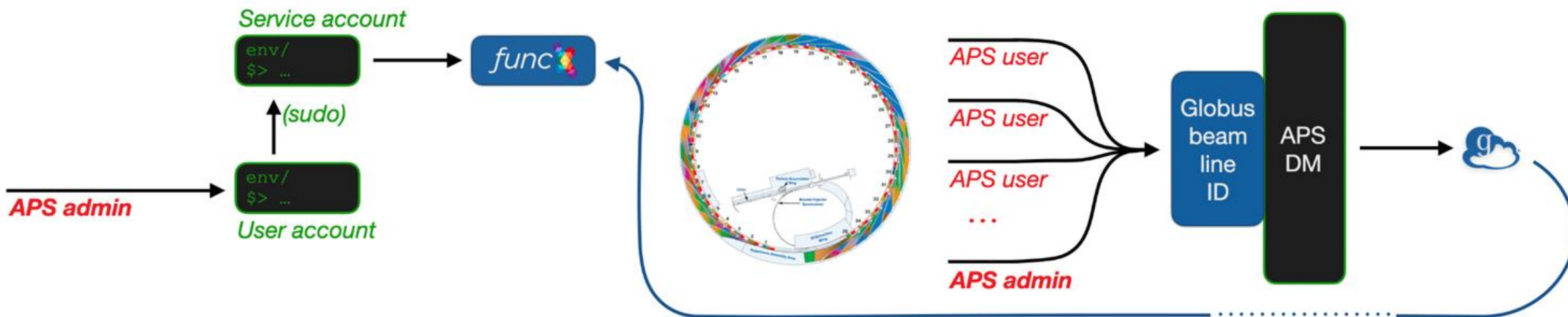
APS requests service account at ALCF

ALCF creates service account

APS admin user setups a **shared environment** and funcX endpoint

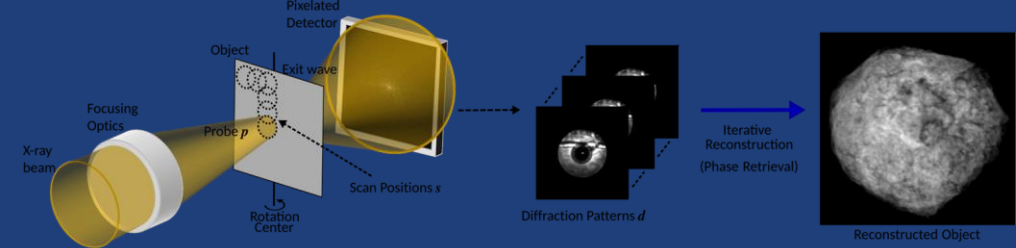
APS users log into DM system using **shared Globus identity**

DM system starts Globus flows involving **shared funcX endpoint**

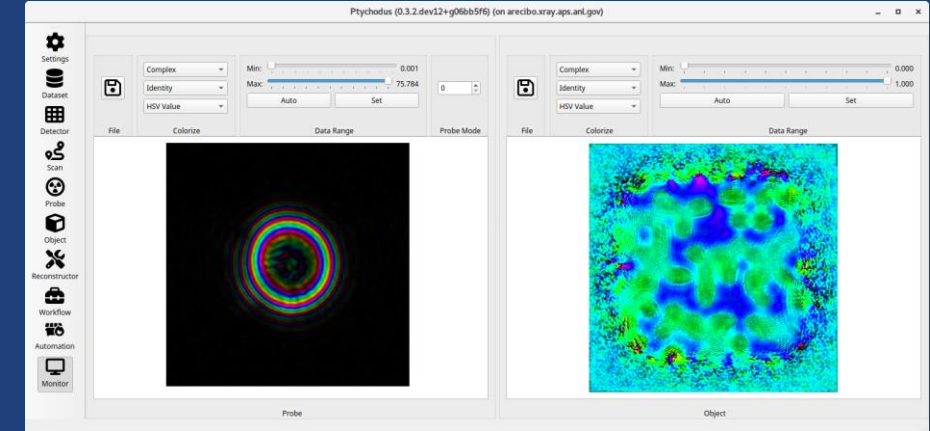




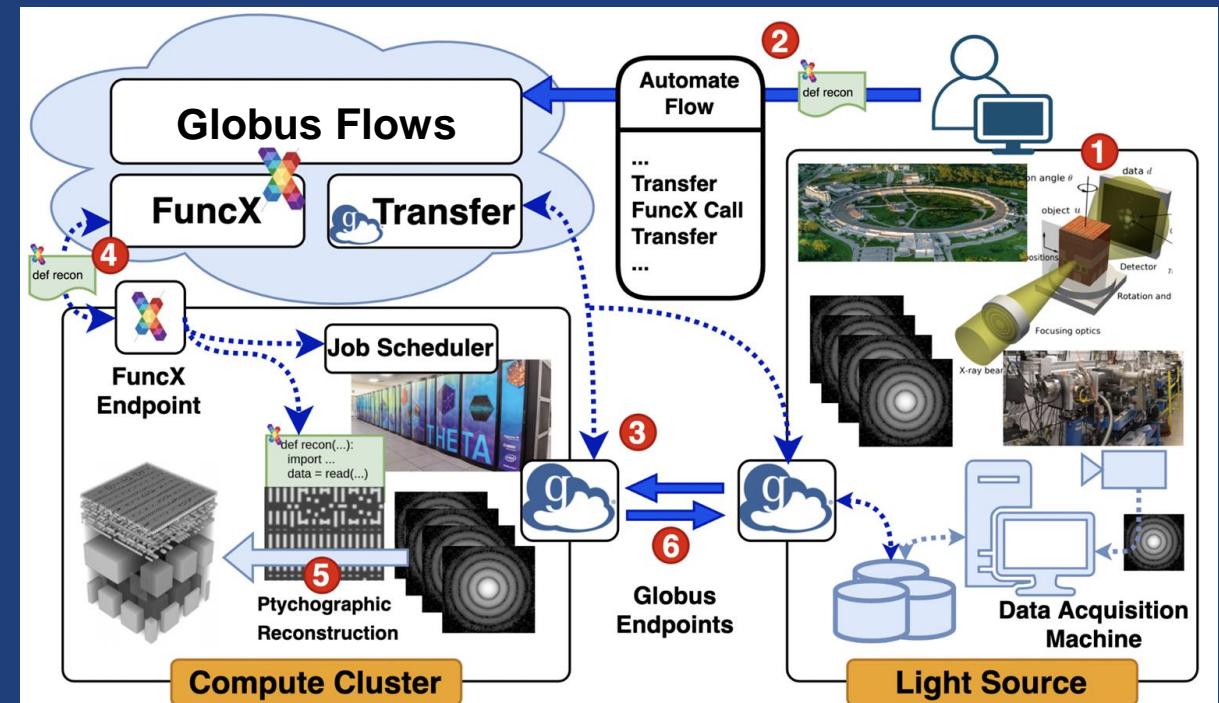
Ptychography at 26ID



- Ptychography is a computational microscopy technique for reconstructing the complex-valued transmission function of an object
- Diffraction patterns are recorded at many overlapping scan positions
- **Flow:**
 - Ptychodus monitors the local filesystem to trigger flows as data are collected
 - Scans are reconstructed at ALCF using on-demand queue
 - Results returned to APS where Ptychodus loads them for visualization
 - Users can customize where the compute needs to run via flow configuration



Full automation using service accounts; on-demand queue for timely runs

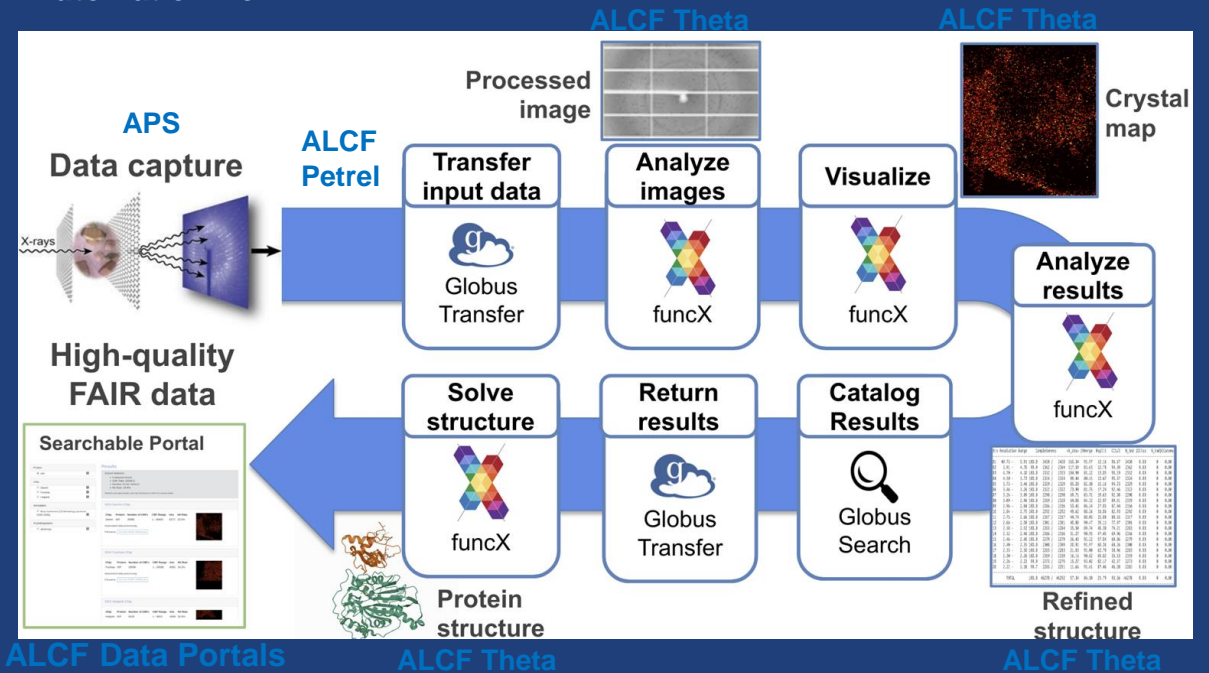




Solving Protein Structures an Order of Magnitude Faster

- Flow: collect data, analyze and visualize the data, solve protein structure and load results into a searchable portal for real-time feedback
- Achieved >order of magnitude speed up in time to solution of protein structures at APS
- Leveraged unique DOE facilities at Advanced Photon Source (SBC Sector 19) and ALCF (Theta/ ThetaGPU, Petrel, and Data Portals)

Automation flow



Deposited results in open repositories

Argonne researchers use Theta for real-time analysis of COVID-19 proteins

Author: Nils Heinonen
Published: 07/28/2020
Domain: Biological Sciences Systems Theta

Argonne's User Facilities Continue to Enable Critical Work Combating and Addressing the Impacts of the COVID-19 Epidemic
June 12, 2020

7JIB
Room Temperature Crystal Structure of Nsp10/Nsp16 from SARS-CoV-2 with Substrates and Products of 2'-O-methylation of the Cap-1
Released: 2020-08-26
Method: X-RAY DIFFRACTION 2.65 Å
Organisms: Severe acute respiratory syndrome coronavirus 2
Macromolecule: 2'-O-methyltransferase (protein)
Unique Ligands: Cl-, GTA, MGP, SAH, SAM, V9G, ZN

7JPE
Room Temperature Structure of SARS-CoV-2 Nsp10/Nsp16 Methyltransferase in a Complex with m7GpppA Cap-0 and SAM Determined by Fixed-Target Serial Crystallography
Released: 2020-08-26
Method: X-RAY DIFFRACTION 2.18 Å
Organisms: Severe acute respiratory syndrome coronavirus 2
Macromolecule: 2'-O-methyltransferase (protein)
Unique Ligands: 8NK, GTA, SAM, ZN

"These data services have taken the time to solve a structure from weeks to days and now to hours"

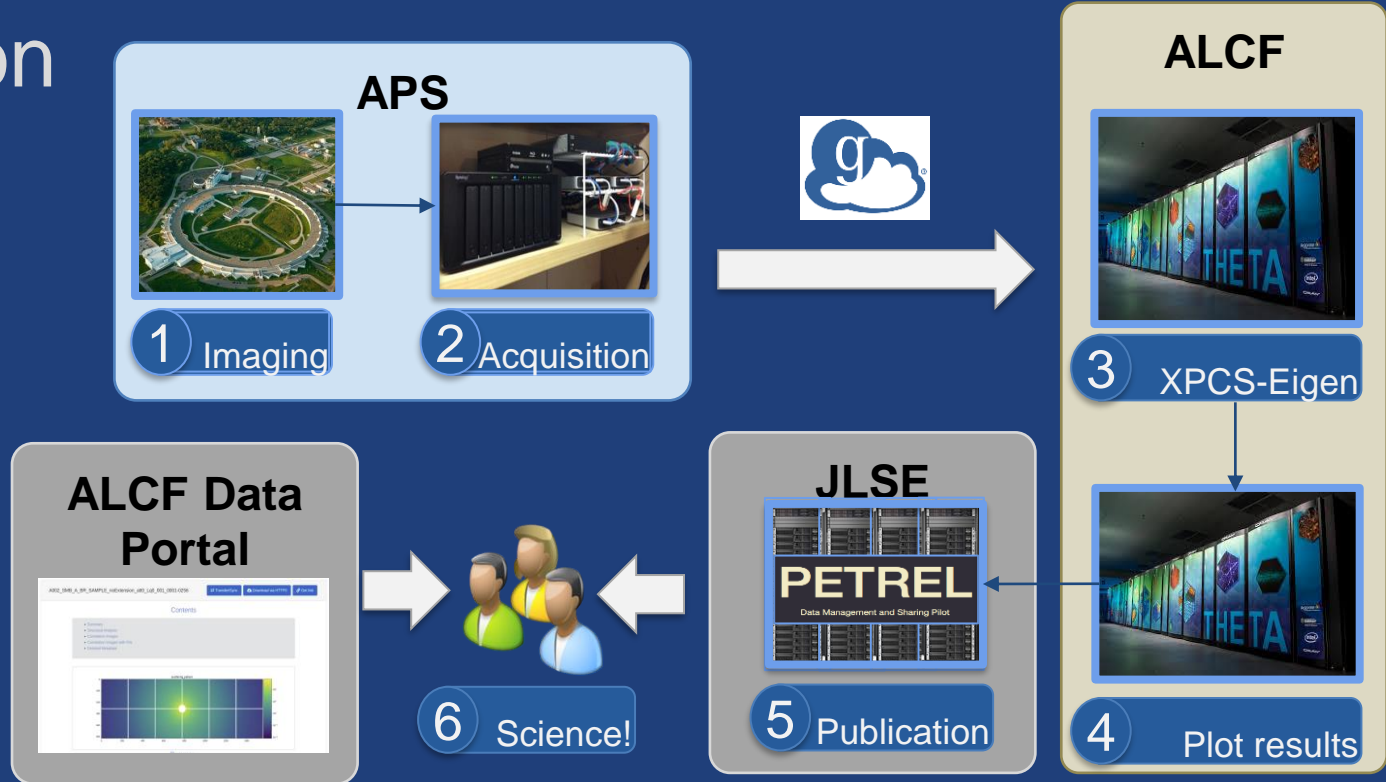
Darren Sherrell, SBC beamline scientist APS Sector 19

(R. Chard, Vescovi, Foster, Blaiszik, Sherrell, Joachimiak, et al.)



X-ray Photon Correlation Spectroscopy (XPCS)

- XPCS studies dynamical properties of materials by recording speckle patterns over time, constructing a time correlation function, and measuring processes of interest (e.g., diffusion)
- **Flows:**
 - Integrated with APS's Data Management system to automatically invoke flows
 - Data analyzed and published to a searchable ALCF portal
 - HTTPS-enabled portals to visualize results
 - Reprocessing capabilities in portal to invoke flows



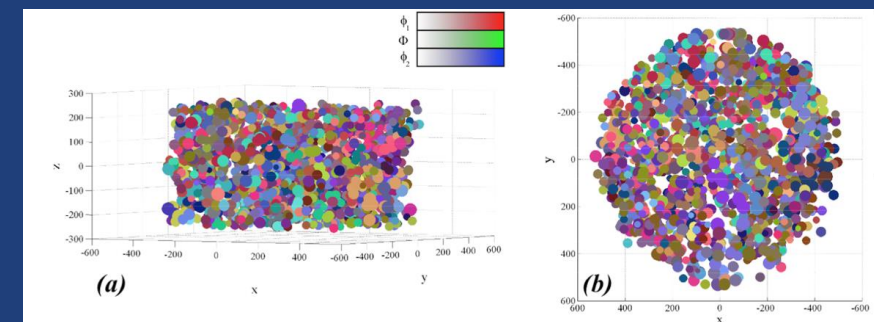
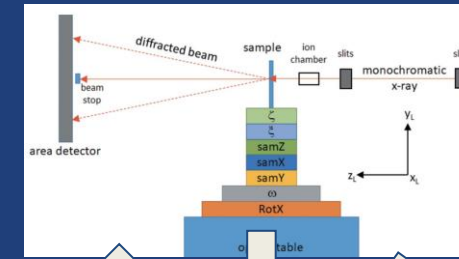
The image shows two screenshots of the ALCF Data Portal and PETREL interface. The left screenshot shows a search results page with a table of datasets:

Parent Folder	Publication Year	Size
sanat202002	2019	97.2 KB
sanat202002	2019	476.4 KB

The right screenshot shows a 'Contents' page with a 'scattering pattern' plot. The plot displays a central bright spot surrounded by a diffuse scattering pattern, with axes ranging from 0 to 1600. A color scale on the right indicates intensity from 10^0 to 10^1 .

High-Energy X-ray Diffraction Microscopy (HEDM)

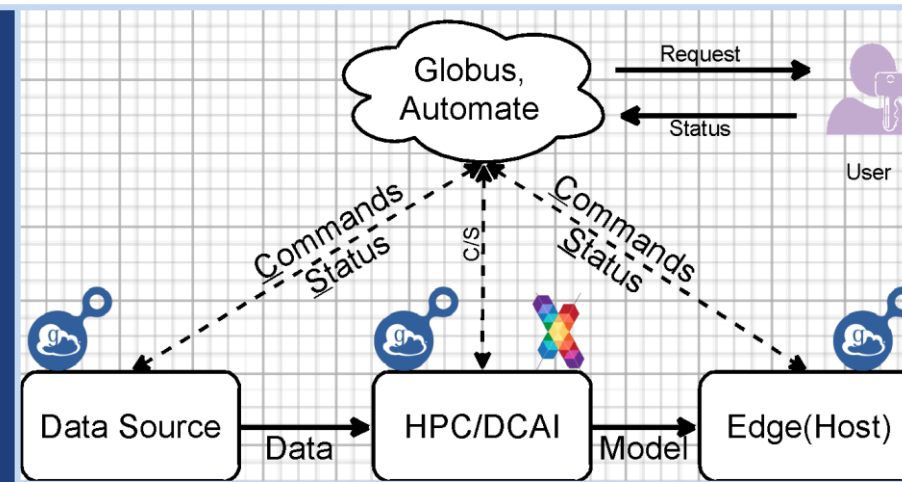
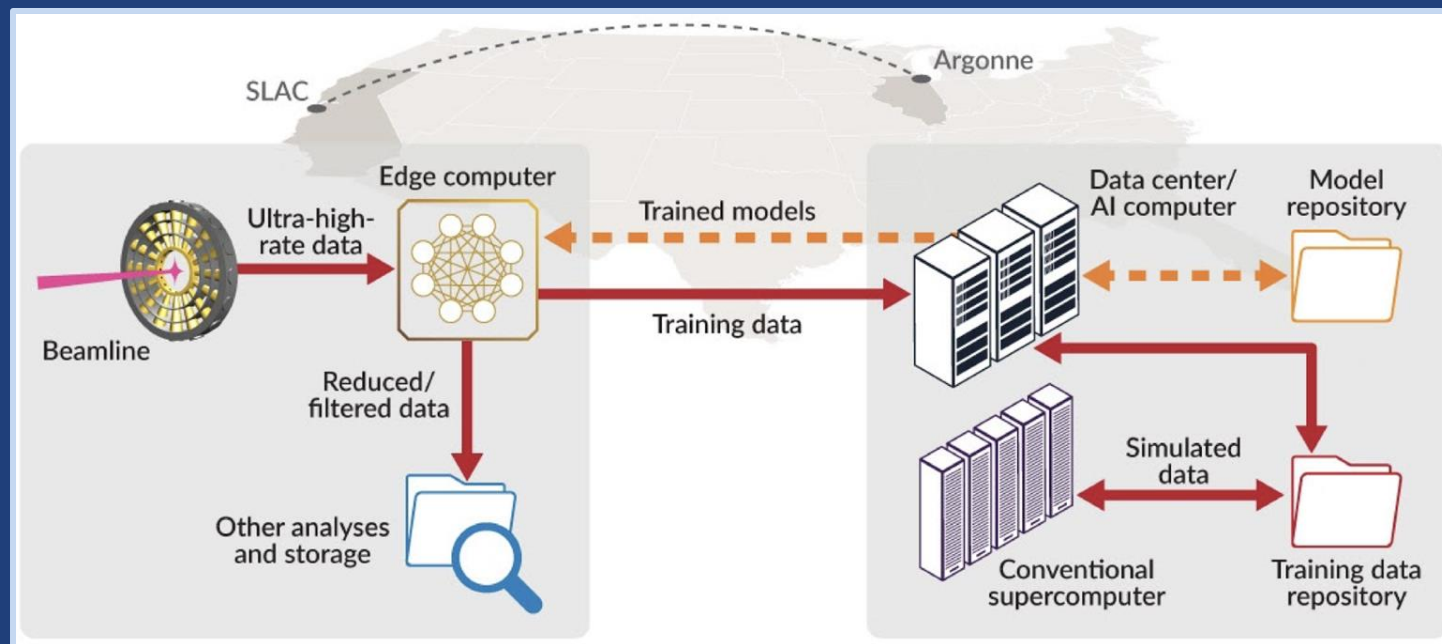
- HEDM combines imaging and crystallography to characterize polycrystalline microstructure in 3D under various in situ thermo-mechanical conditions
- **Flow**
 - Select analysis to run at APS Orthros, ALCF ThetaGPU/Cooley
 - Globus Transfer data
 - Deploy containers with MIDAS software to perform tasks
 - Results assembled and returned to APS user
 - Mechanism for users to run analysis at home institute





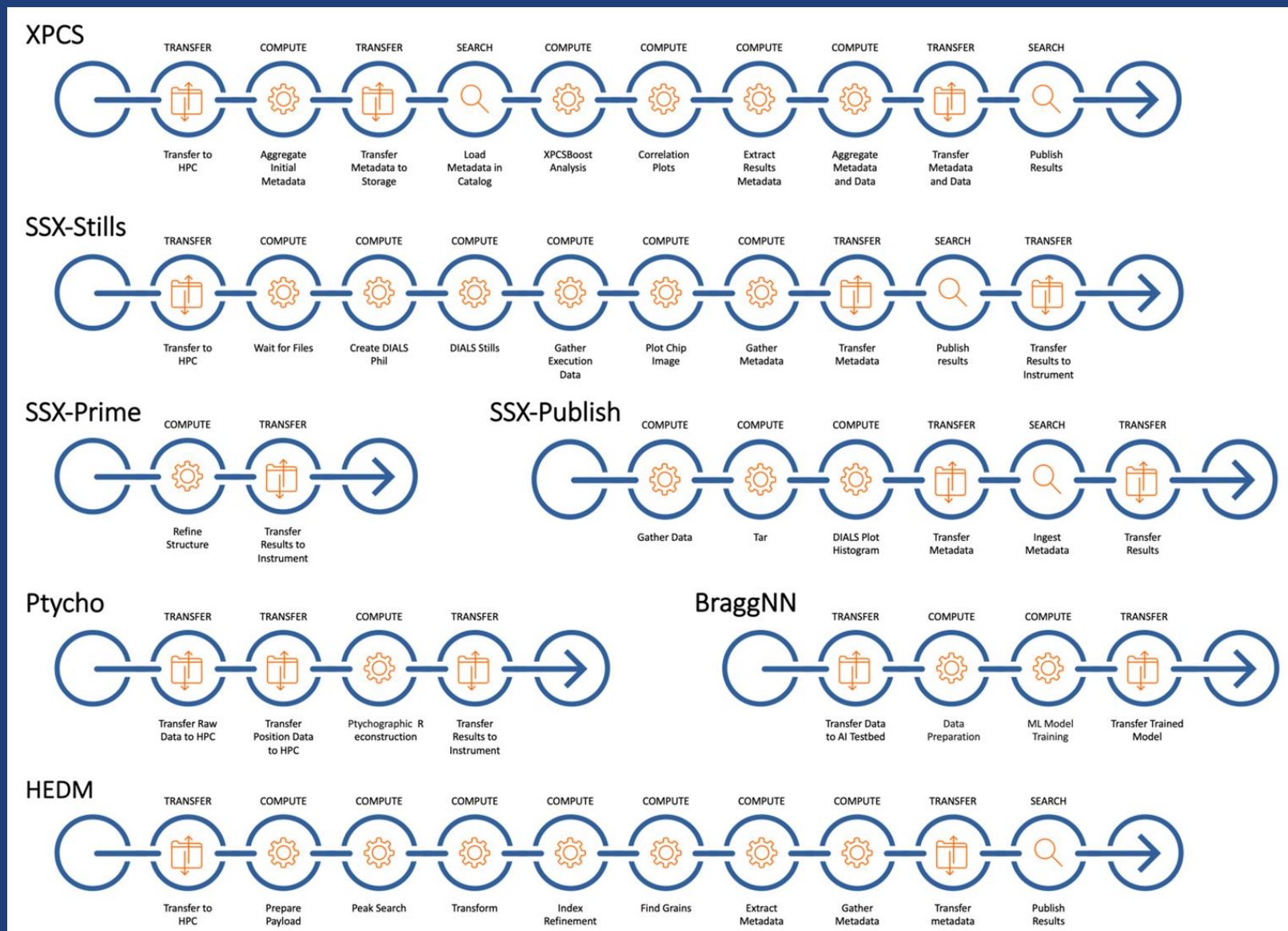
Rapid Training of Deep Neural Networks using Remote Resources

- HEDM workflow that deploys BragNN at the edge for real-time diffraction peak analysis (e.g., for experiment steering and anomaly detection)
- Tight coupling with simulation and training with real-time data
- Flow:
 - Globus to rapidly move data for training
 - funcX for simulation and model training
 - Globus to move models to the edge





Production flows linking instruments and computation



R. Vescovi et al.,

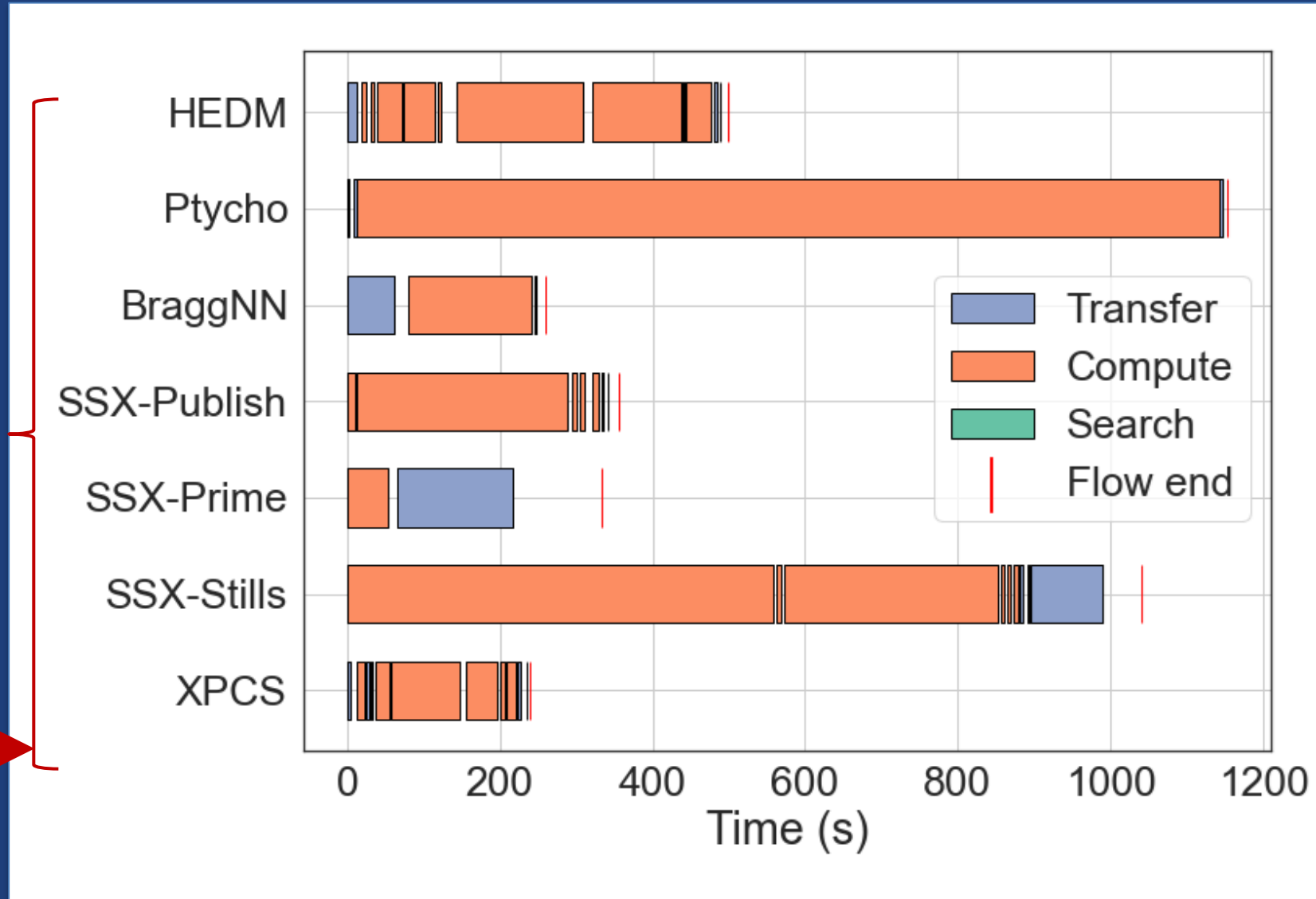
<https://doi.org/10.1016/j.patter.2022.100606>

Flows span spatial and temporal ranges

Reliable flow
orchestration across
resources

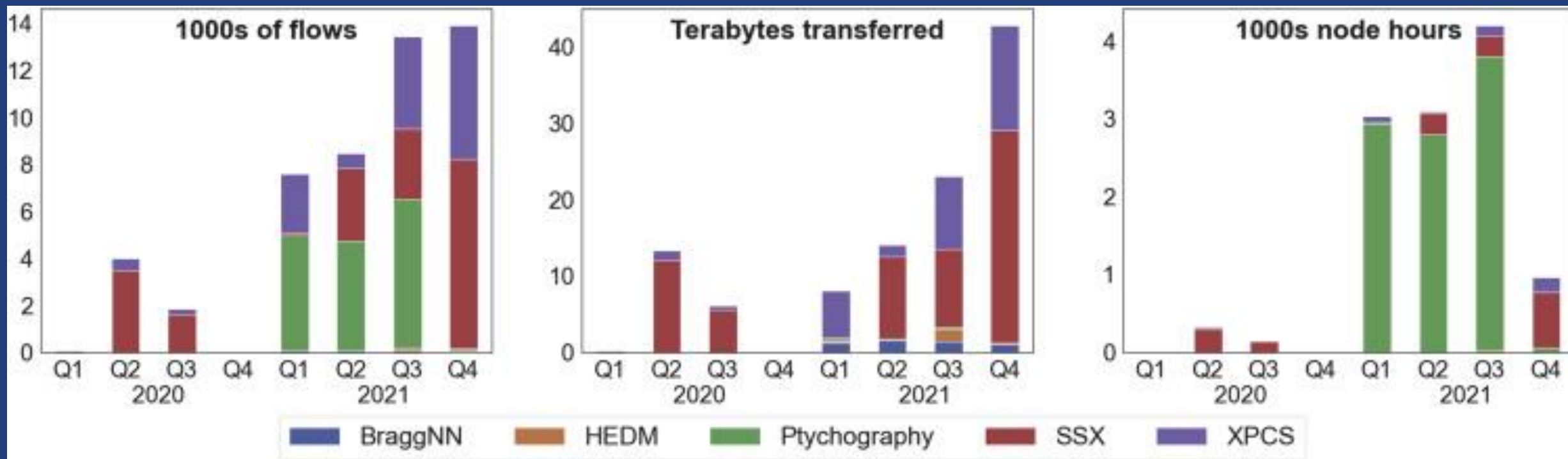
Functions executed in
various locations: at a
beamline, local server,
cluster, cloud

Execution times at the
Argonne Leadership
Computing Facility





Flows are increasingly critical to APS science





Globus documentation: docs.globus.org

YouTube: youtube.com/GlobusOnline

Helpdesk: support@globus.org

Patterns

 **CellPress**
OPEN ACCESS

<https://doi.org/10.1016/j.patter.2022.100606>

Article

**Linking scientific instruments and computation:
Patterns, technologies, and experiences**

Rafael Vescovi,¹ Ryan Chard,¹ Nickolaus D. Saint,⁶ Ben Blaiszik,^{1,6} Jim Pruyne,^{1,6} Tekin Bicer,^{1,3} Alex Lavens,⁴ Zhengchun Liu,¹ Michael E. Papka,^{2,7} Suresh Narayanan,³ Nicholas Schwarz,³ Kyle Chard,^{1,5} and Ian T. Foster^{1,5,*}



Globus
subscribers