Beyond Data Management with Globus

iRODS® UGM 2023

Chapel Hill

Vas Vasiliadis

University of Chicago – Globus

Adjunct Associate Professor, Masters Program in Computer Science

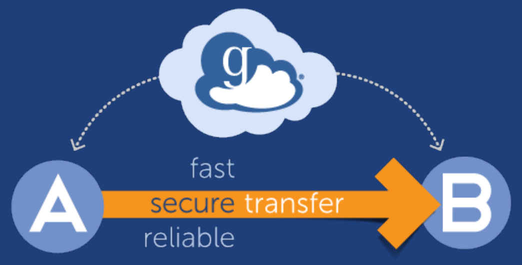
vas@uchicago.edu, vasv@anl.gov

THE UNIVERSITY OF CHICAGO

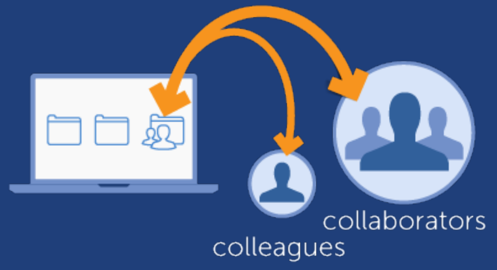
Argonne NATIONAL LABORATORY

globus®

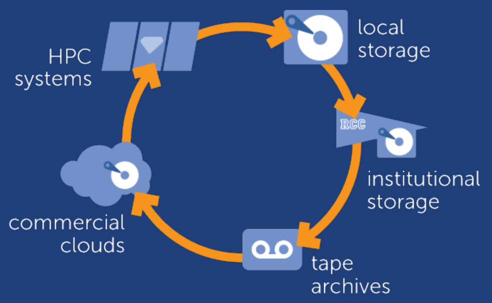
Reimagining research IT with Globus



Managed transfer & sync

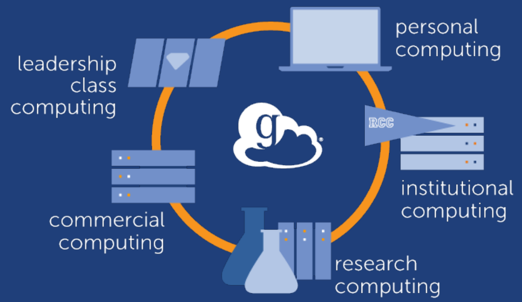


Collaborative data sharing

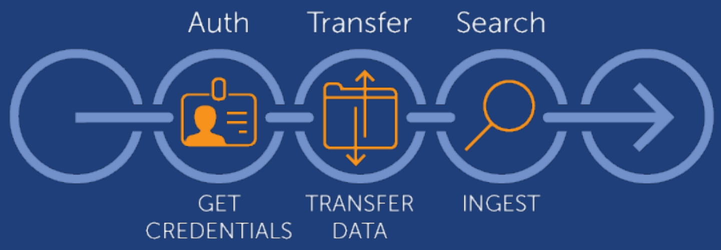


Unified data access

Publication & discovery

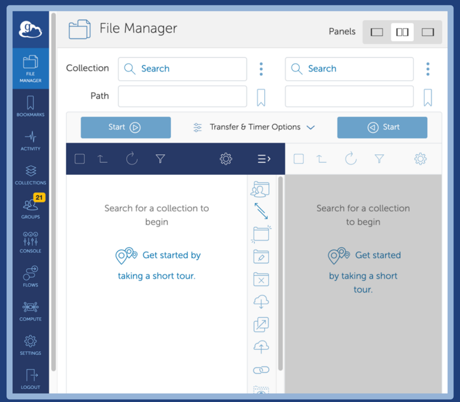


Managed remote execution



Reliable automation

Platform-as-a-Service



Search for a collection to begin

Get started by taking a short tour.

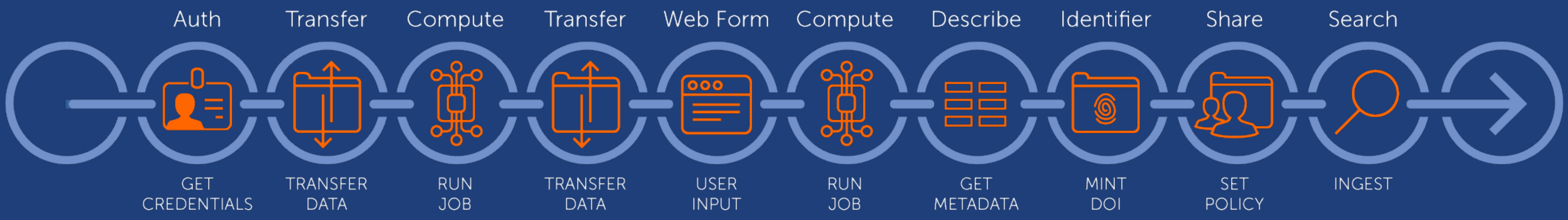
Search for a collection to begin

Get started by taking a short tour.

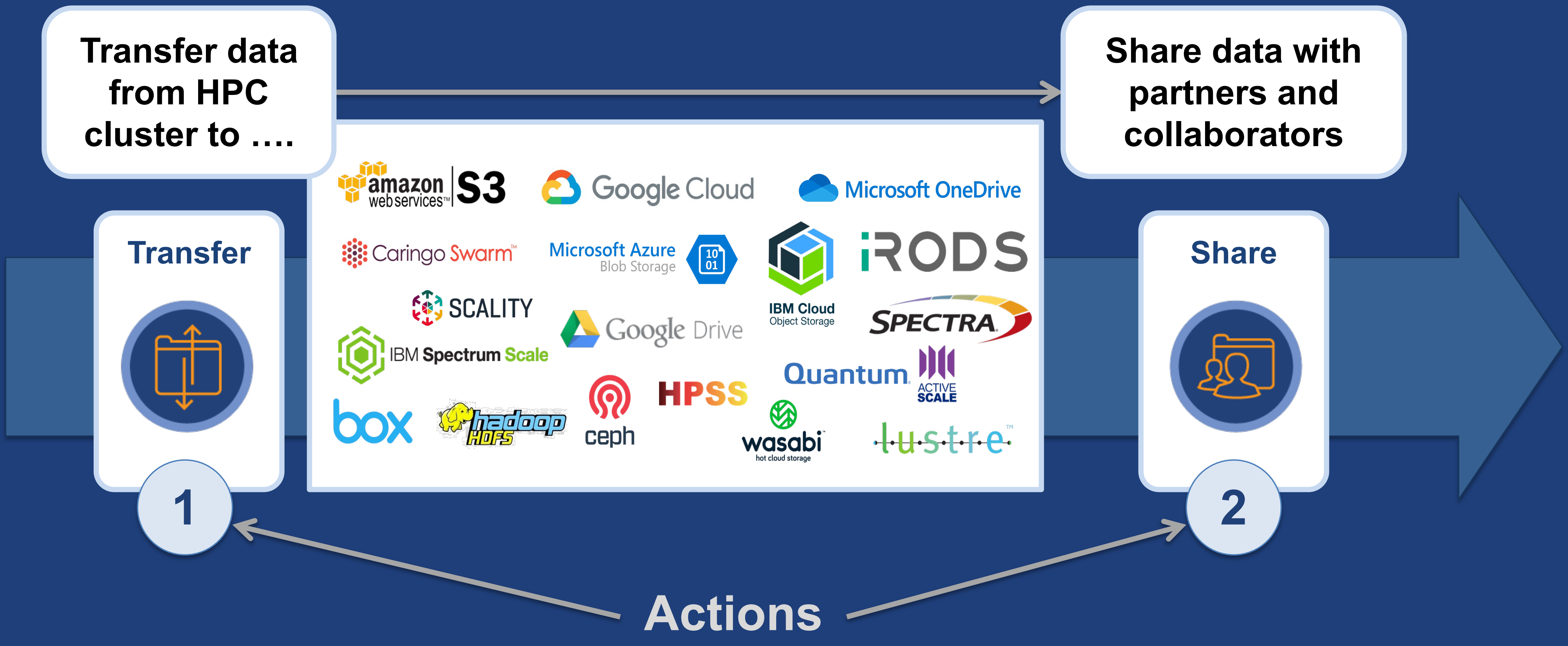
Software-as-a-Service

Automating Research with Globus Flows

- A platform for orchestrating distributed research tasks
- Flows comprise **Actions**
- **Action Providers**: Called by Flows to perform tasks
- **Triggers**: Start flows based on events
- Extensible via **Action Provider API**



A simple, and very common, use case



Transfer data from HPC cluster to ….

Share data with partners and collaborators

Transfer

Share

1

2

Actions

cryoEM automation

Globus Flows

Transfer

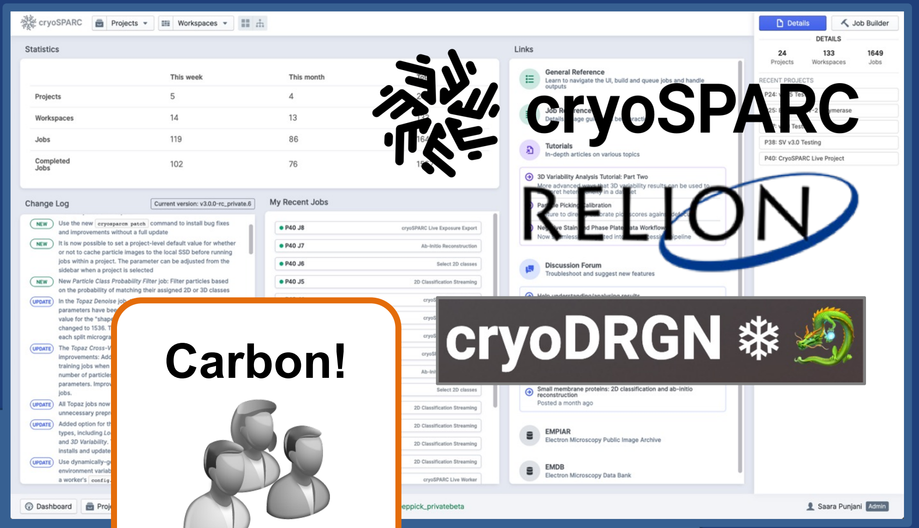
Transfer raw files

Compute

Launch analysis job

Carbon!

Correct, classify, …



cryoSPARC

RELION

cryoDRGN

cryoSPARC | Projects | Workspaces | Details | Job Builder

Statistics

	This week	This month
Projects	5	4
Workspaces	14	13
Jobs	119	86
Completed Jobs	102	76

Change Log | Current version: v3.0.0-rc_private.6

- NEW Use the new `cryosparcm patch` command to install bug fixes and improvements without a full update
- NEW It is now possible to set a project-level default value for whether or not to cache particle images to the local SSD before running jobs within a project. The parameter can be adjusted from the sidebar when a project is selected
- NEW New *Particle Class Probability Filter* job: Filter particles based on the probability of matching their assigned 2D or 3D classes

My Recent Jobs

P40 J8	cryoSPARC Live Exposure Export
P40 J7	Ab-Initio Reconstruction
P40 J6	Select 2D classes
P40 J5	2D Classification Streaming

Links

- General Reference: Learn to navigate the UI, build and queue jobs and handle outputs
- Tutorials: In-depth articles on various topics
- Discussion Forum: Troubleshoot and suggest new features
- EMPIAR: Electron Microscopy Public Image Archive
- EMDB: Electron Microscopy Data Bank

DETAILS

24	133	1649
Projects	Workspaces	Jobs

RECENT PROJECTS

- P38: SV v3.0 Testing
- P40: CryoSPARC Live Project

Saara Punjani Admin

Compute

Extract metadata

Describe

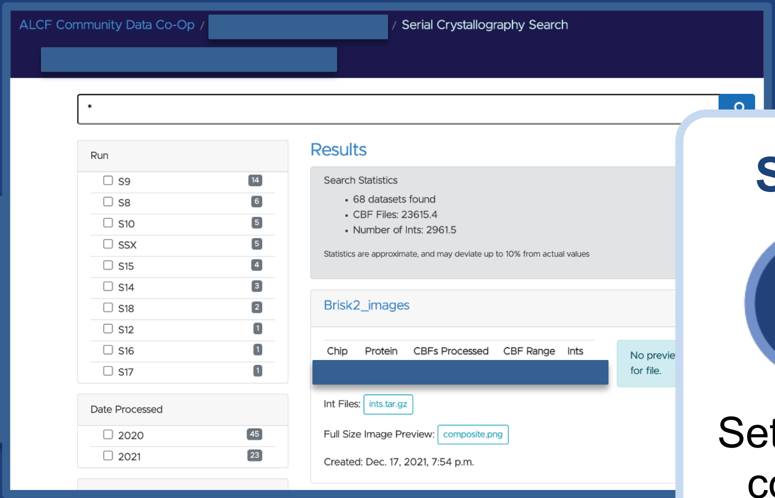
Search index ingest

Transfer

Move final files to repo

Share

Set access controls



ALCF Community Data Co-Op / [redacted] / Serial Crystallography Search

[redacted]

*

Run

□ S9	14
□ S8	6
□ S10	5
□ SSX	5
□ S15	4
□ S14	3
□ S18	2
□ S12	1
□ S16	1
□ S17	1

Date Processed

□ 2020	45
□ 2021	23

Results

Search Statistics

- 68 datasets found
- CBF Files: 23615.4
- Number of Ints: 2961.5

Statistics are approximate, and may deviate up to 10% from actual values

Brisk2_images

Chip	Protein	CBFs Processed	CBF Range	Ints
[redacted]				

No previe for file.

Int Files: ints.tar.gz

Full Size Image Preview: composite.png

Created: Dec. 17, 2021, 7:54 p.m.

iRODS

How to best enable distributed, remote compute?

Borrow page from data management playbook

- **➔“Fire-and-forget” computation**
- **➔Uniform access interface**
- **➔Federated access control**
- **➔Move closer to researchers’ environments**
  - ➔Researchers primarily work in high level languages
  - ➔Functions are a natural unit of computation

Globus Compute

Managed, federated Function-as-a-Service for reliably, scaleably and securely executing functions on remote endpoints from laptops to supercomputers

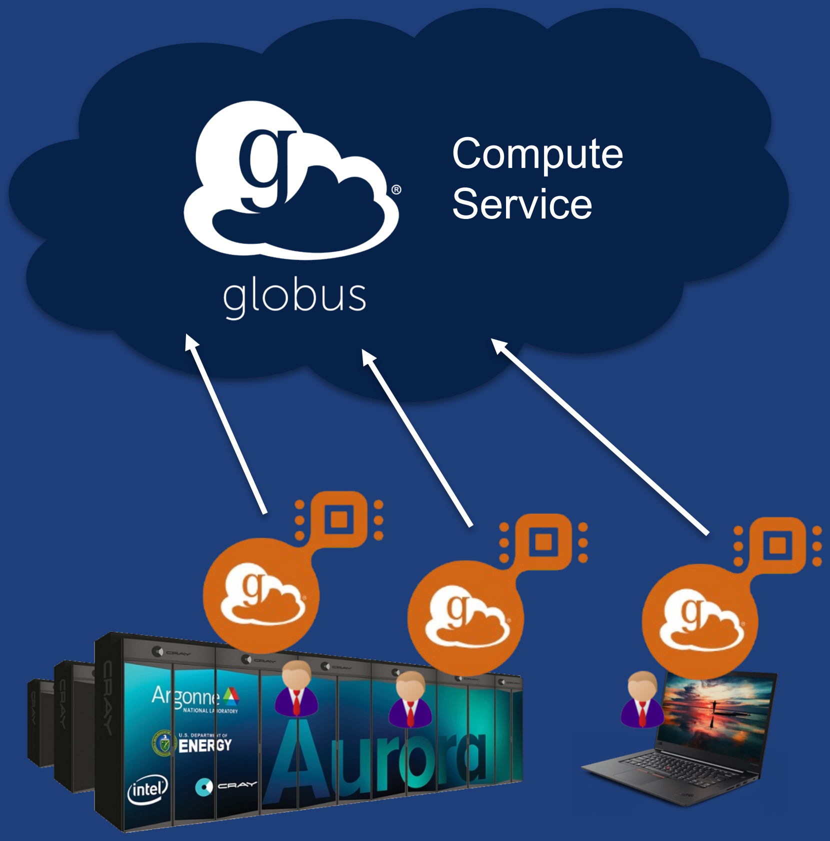
THE UNIVERSITY OF CHICAGO

ILLINOIS

Argonne NATIONAL LABORATORY

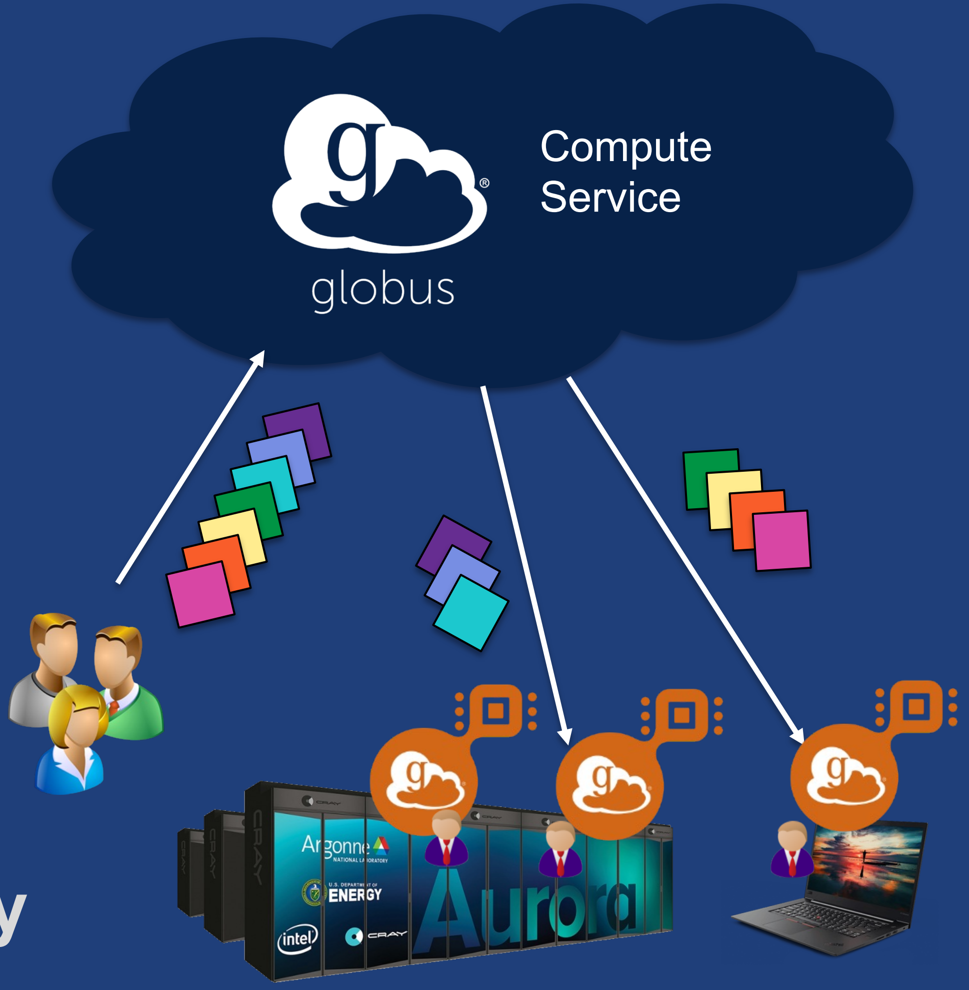
Globus Compute transforms any computing resource into a function serving endpoint

- **Python pip installable agent**
- **Elastic resource provisioning from local, cluster, or cloud system (via Parsl)**
- **Parallel execution using local fork or via common schedulers**
  - Slurm, PBS, LSF, Cobalt, K8s

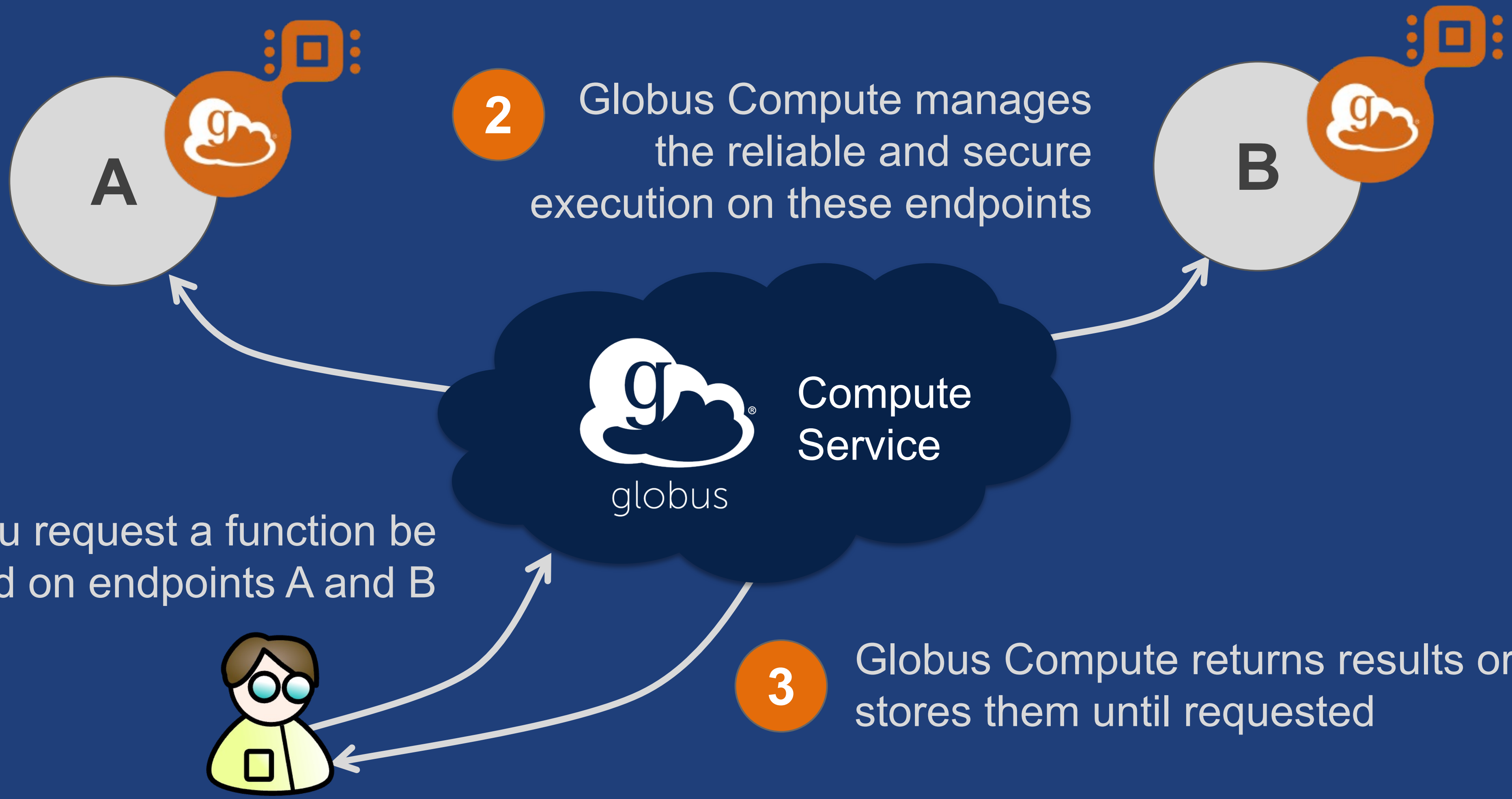


Executing functions with Globus Compute

- **Users invoke functions as tasks**
  - Register Python function
  - Pass input arguments
  - Select endpoint(s)
- **Service stores tasks in the cloud**
- **Endpoints fetch waiting tasks (when online), run tasks, and return results**
- **Results stored in the cloud and on Globus storage endpoints**
- **Users retrieve results asynchronously**



User interaction with Globus Compute



1 You request a function be executed on endpoints A and B

2 Globus Compute manages the reliable and secure execution on these endpoints

3 Globus Compute returns results or stores them until requested

Use Case

Executing a bag of tasks, e.g., running simulations with different parameters, executing ML inferences, on multiple remote computers directly from your environment, e.g., Jupyter notebook

Use Case

Constructing and running automated analysis pipelines with data processing steps that need to be executed in different locations

Use Case

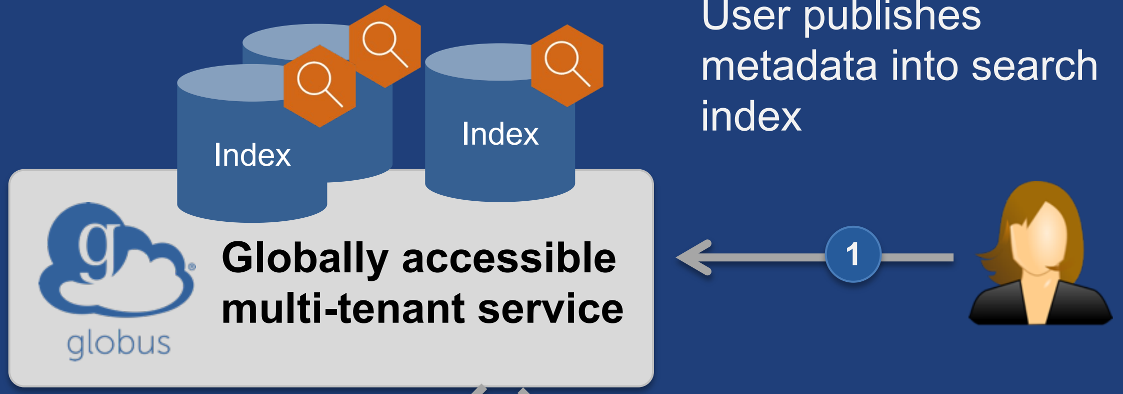
Building new applications and services that seamlessly execute application components or user workloads on remote resources

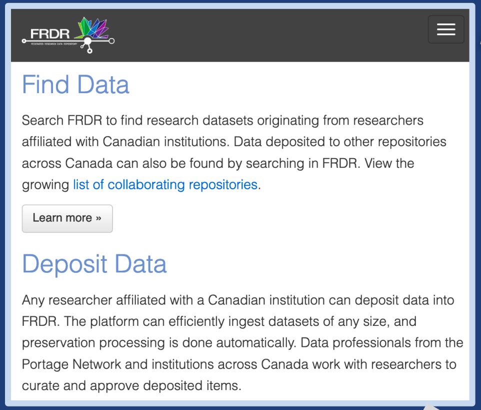
Scalable data discovery with Globus Search

- **Scalable metadata store**
- **Fine-grained visibility controls**
- **Schema agnostic**
  → dynamic schemas
- **Federated auth integration**
- **Robust query API**
  - **GET with URL parameters**
  - **POST with facets**

docs.globus.org/api/search

User publishes metadata into search index





FRDR

Find Data

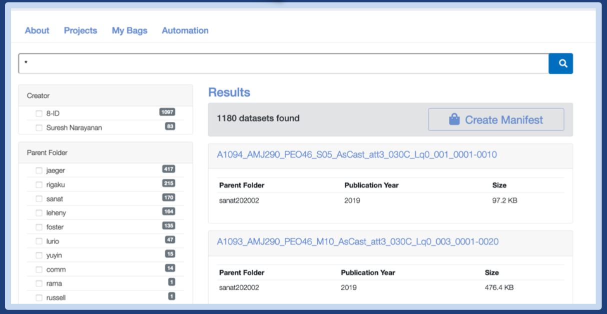
Search FRDR to find research datasets originating from researchers affiliated with Canadian institutions. Data deposited to other repositories across Canada can also be found by searching in FRDR. View the growing list of collaborating repositories.

Learn more »

Deposit Data

Any researcher affiliated with a Canadian institution can deposit data into FRDR. The platform can efficiently ingest datasets of any size, and preservation processing is done automatically. Data professionals from the Portage Network and institutions across Canada work with researchers to curate and approve deposited items.

Globus manages metadata access



About   Projects   My Bags   Automation

*

Creator

□ 8-ID	1097
□ Suresh Narayanan	83

Parent Folder

□ jaeger	417
□ rigaku	215
□ sanat	170
□ leheny	164
□ foster	135
□ lurio	47
□ yuyin	15
□ comm	14
□ rama	1
□ russell	1

Results

1180 datasets found   Create Manifest

A1094_AMJ290_PEO46_S05_AsCast_att3_030C_Lq0_001_0001-0010

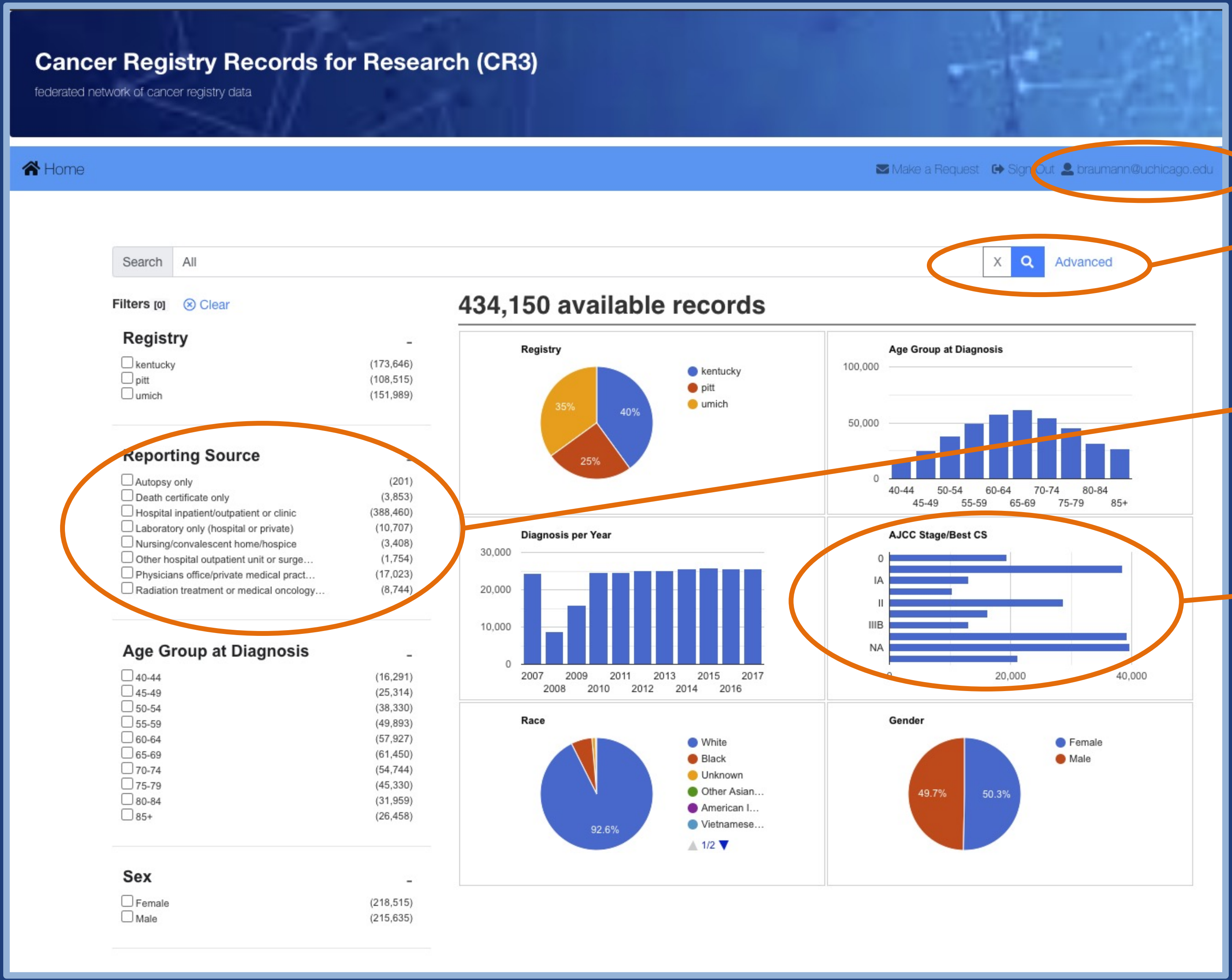
Parent Folder	Publication Year	Size
sanat202002	2019	97.2 KB

A1093_AMJ290_PEO46_M10_AsCast_att3_030C_Lq0_003_0001-0020

Parent Folder	Publication Year	Size
sanat202002	2019	476.4 KB

Users query and discover data of interest

CR3 Portal (simulated data)



Cancer Registry Records for Research (CR3)

federated network of cancer registry data

Search: All

Filters [0] ⊗ Clear

434,150 available records

Registry -

□ kentucky	(173,646)
□ pitt	(108,515)
□ umich	(151,989)

Reporting Source -

□ Autopsy only	(201)
□ Death certificate only	(3,853)
□ Hospital inpatient/outpatient or clinic	(388,460)
□ Laboratory only (hospital or private)	(10,707)
□ Nursing/convalescent home/hospice	(3,408)
□ Other hospital outpatient unit or surge…	(1,754)
□ Physicians office/private medical pract…	(17,023)
□ Radiation treatment or medical oncology…	(8,744)

Age Group at Diagnosis -

□ 40-44	(16,291)
□ 45-49	(25,314)
□ 50-54	(38,330)
□ 55-59	(49,893)
□ 60-64	(57,927)
□ 65-69	(61,450)
□ 70-74	(54,744)
□ 75-79	(45,330)
□ 80-84	(31,959)
□ 85+	(26,458)

Sex -

□ Female	(218,515)
□ Male	(215,635)

Federated logon using Globus Auth with 1,800+ identity providers

Google-like text search with facets for filtering

Variable facets based on source registry index

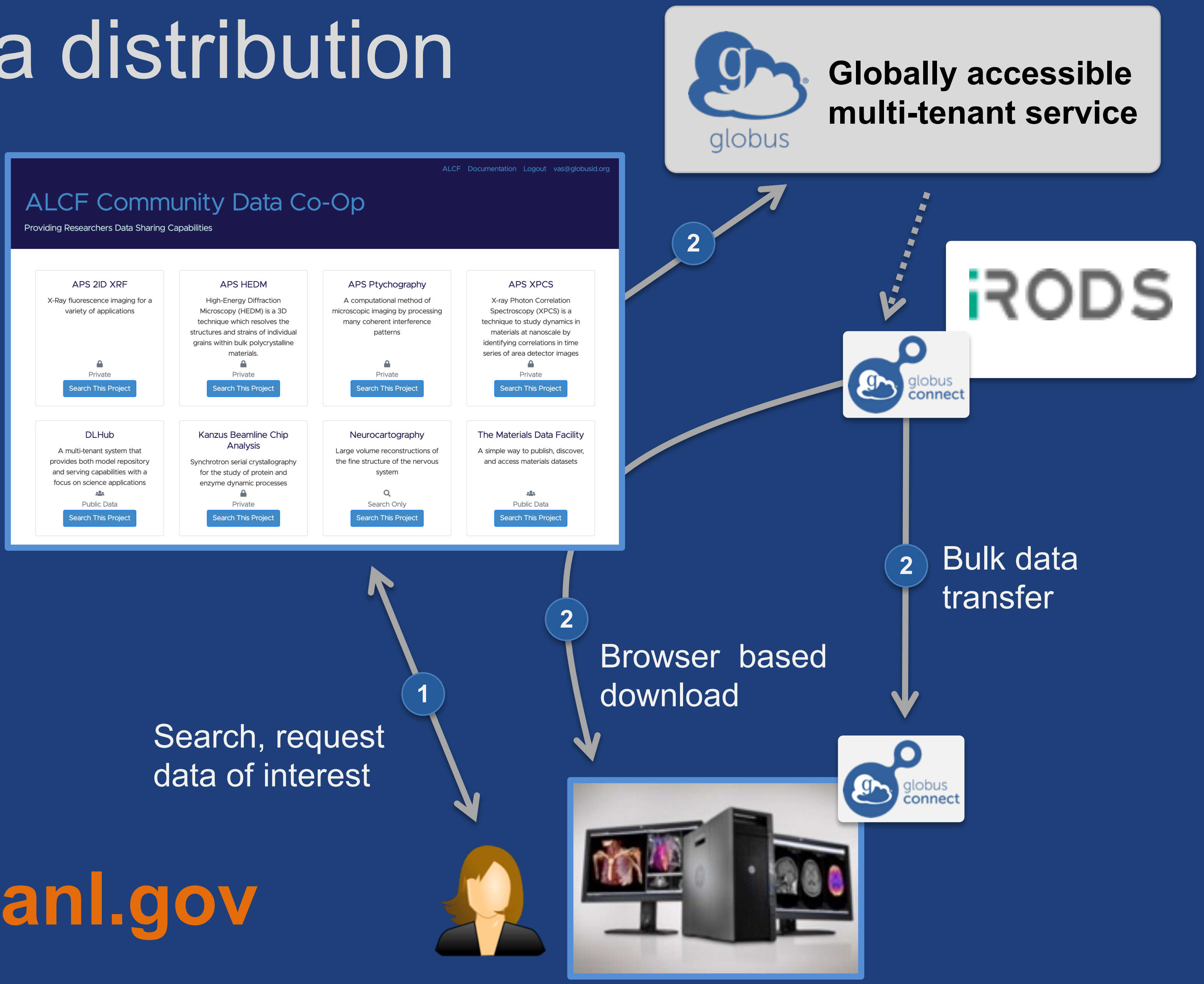
Dynamically updating charts as facets change

Developed using a framework based on the Globus Modern Research Data Portal* design pattern (**docs.globus.org/mrdp**)

* PeerJ Articles:cs-144 https://peerj.com/articles/cs-144/

Repository data distribution

- **Faceted search via data portal**
- **Enforces fine-grained authZ**
- **HTTPS download for “small” data**
- **Managed file transfer for larger data sets**



ALCF Documentation Logout vas@globusid.org

ALCF Community Data Co-Op

Providing Researchers Data Sharing Capabilities

APS 2ID XRF
X-Ray fluorescence imaging for a variety of applications
Private
Search This Project

APS HEDM
High-Energy Diffraction Microscopy (HEDM) is a 3D technique which resolves the structures and strains of individual grains within bulk polycrystalline materials.
Private
Search This Project

APS Ptychography
A computational method of microscopic imaging by processing many coherent interference patterns
Private
Search This Project

APS XPCS
X-ray Photon Correlation Spectroscopy (XPCS) is a technique to study dynamics in materials at nanoscale by identifying correlations in time series of area detector images
Private
Search This Project

DLHub
A multi-tenant system that provides both model repository and serving capabilities with a focus on science applications
Public Data
Search This Project

Kanzus Beamline Chip Analysis
Synchrotron serial crystallography for the study of protein and enzyme dynamic processes
Private
Search This Project

Neurocartography
Large volume reconstructions of the fine structure of the nervous system
Search Only
Search This Project

The Materials Data Facility
A simple way to publish, discover, and access materials datasets
Public Data
Search This Project

Globally accessible multi-tenant service

Bulk data transfer

Browser based download

Search, request data of interest

Example: **acdc.alcf.anl.gov**

Resources

- Web app access: app.globus.org
- Documentation: docs.globus.org
- Helpdesk: support@globus.org
- Mailing Lists: globus.org/mailing-lists