

# Globus: Platform for Data Driven Research

Rachana Ananthakrishnan  
Executive Director, Globus  
[ranantha@uchicago.edu](mailto:ranantha@uchicago.edu)



## IN-PERSON WORKSHOP

Empowering Secure Data-Driven Research:  
Leveraging Science DMZ, Globus, and National  
Cyberinfrastructure Resources





Globus is ...

a non-profit service  
developed and operated by



THE UNIVERSITY OF  
CHICAGO



Our mission is to...  
increase the efficiency and  
effectiveness of researchers  
engaged in data-driven  
science and scholarship  
through *sustainable* software.



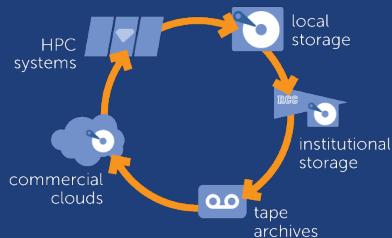
# Deliver Platform for Research IT



Managed transfer & sync



Collaborative data sharing



Unified data access



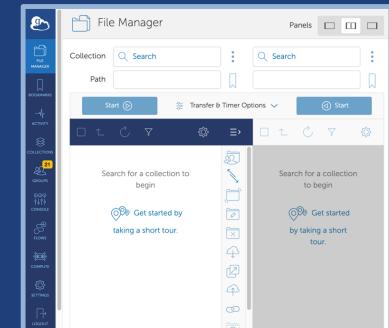
Publication & discovery



Managed remote execution



Reliable automation



Software-as-a-Service



Platform-as-a-Service



## Our freemium sustainability model

- **Basic capabilities are available free of charge to anyone engaged in non-profit research**
- **Subscriptions enable enhanced features for both researchers and system administrators**
- **Subscription pricing based on level of research supported**

[globus.org/subscriptions](https://globus.org/subscriptions)



## Which best describes your primary role at your institution?



- Research computing manager/leader
- Research computing staff/System administrator
- Compliance/security staff/leader
- Researcher/postdoc
- Student
- Infrastructure/Service provider



**In service for science and  
scholarship...**



# Digital agriculture – University of Winnipeg

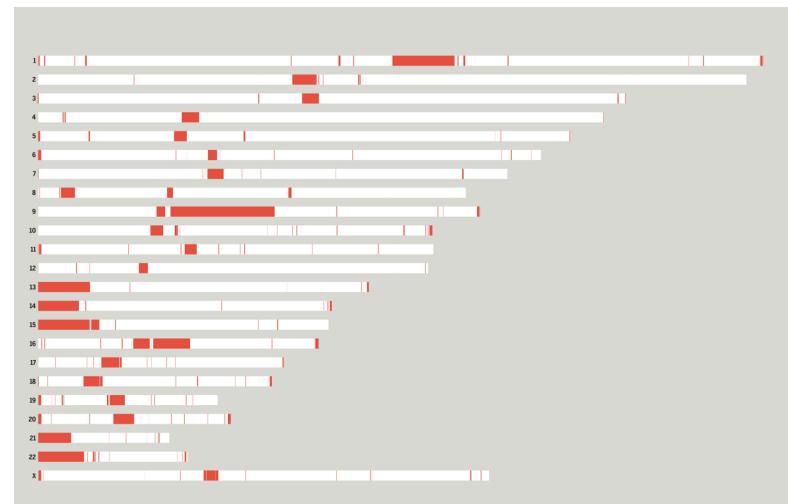
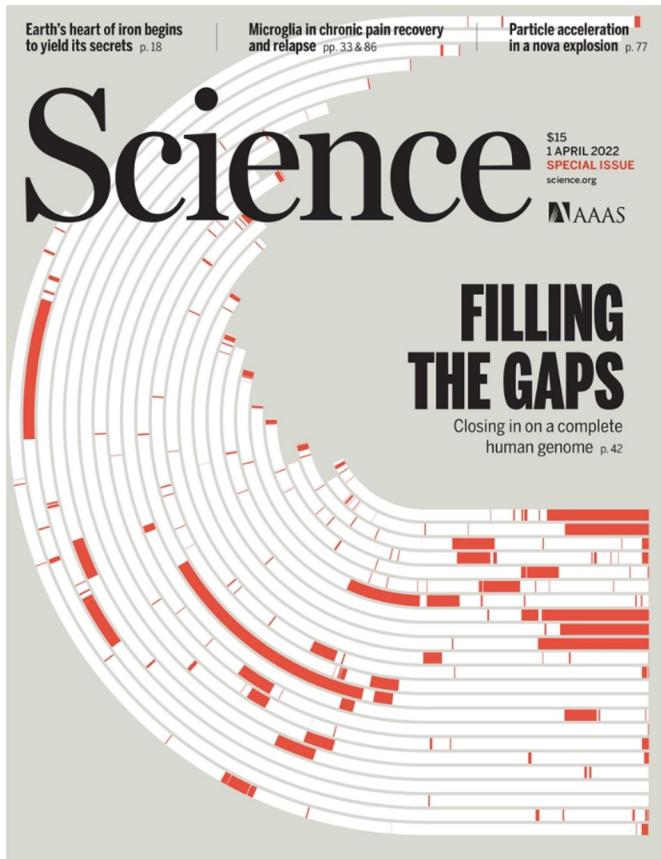
- Increasing crop yields using machine learning models
- Building training data sets
  - 40K images per day, tagged with metadata
  - Move data from diverse sources to campus storage, then onto Compute Canada HPC to run models
- Orchestrate data transfer using Globus CLI



Credit: Dilbarjot and Michael Beck,  
Physics and Applied Computer Science , University of Winnipeg



# Secure data sharing for international collaboration

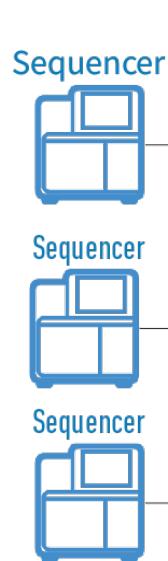


Resolved sequences by the T2T-CHM13v2.0 reference genome.  
Resource: T2T consortium

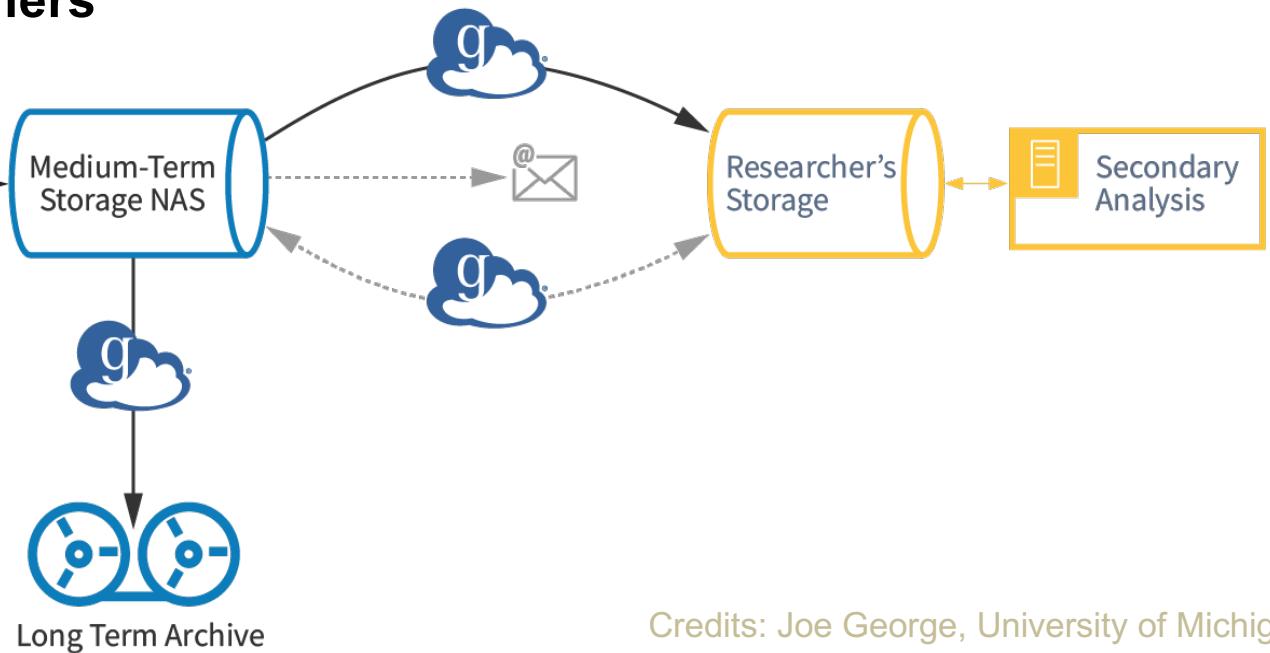
[globus.org/user-stories/globus-enables-multi-institutional-data-sharing](https://globus.org/user-stories/globus-enables-multi-institutional-data-sharing)



# Instrument data delivery at scale



**Use Globus to deliver  
100s of TB of genomic  
data to researchers**



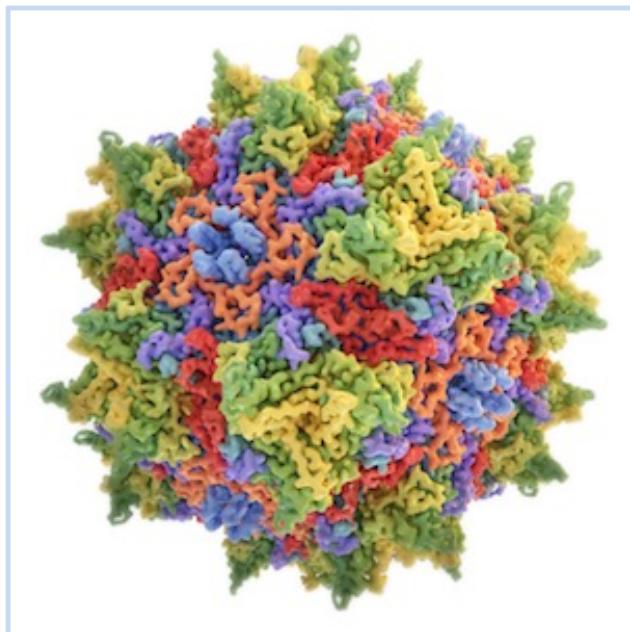
**BIOMEDICAL RESEARCH CORE FACILITIES  
ADVANCED GENOMICS CORE  
UNIVERSITY OF MICHIGAN**

Credits: Joe George, University of Michigan



# Data Management at Cryo-EM Facilities

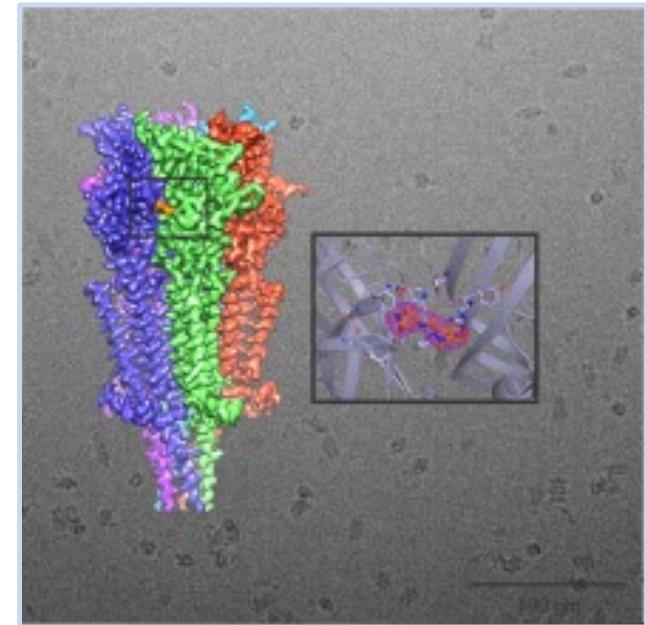
## Pacific Northwest Cryo-EM Processing Center (PNNL and Oregon Health Sciences University)



Credit: <https://pncc.labworks.org/about-us>

## Case Western Reserve – Cryo-EM Core

**Globus for**  
**automated data sync as new**  
**data is collected**  
**provisioning of data access**  
**for researchers**  
**reliable, secure data access**  
**for users**  
**monitoring and management**  
**via console**



Credit: <https://case.edu/medicine/research/som-core-facilities/cryo-electron-microscopy-core>

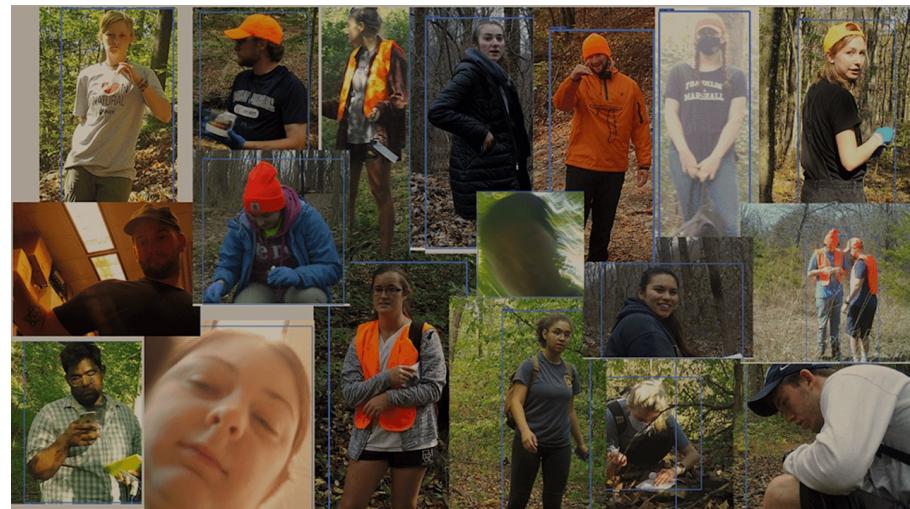


## Streamlining processing of field data

*Having a Globus Flow developed in collaboration with our Research Computing Colleagues and maintained in a library of flows allows high-speed computing to be **available to a larger number of potential users**. In my case, the Globus flow structure will allow me to incorporate collaborators and volunteers more easily into my research, **which increases community impact and engagement**.*

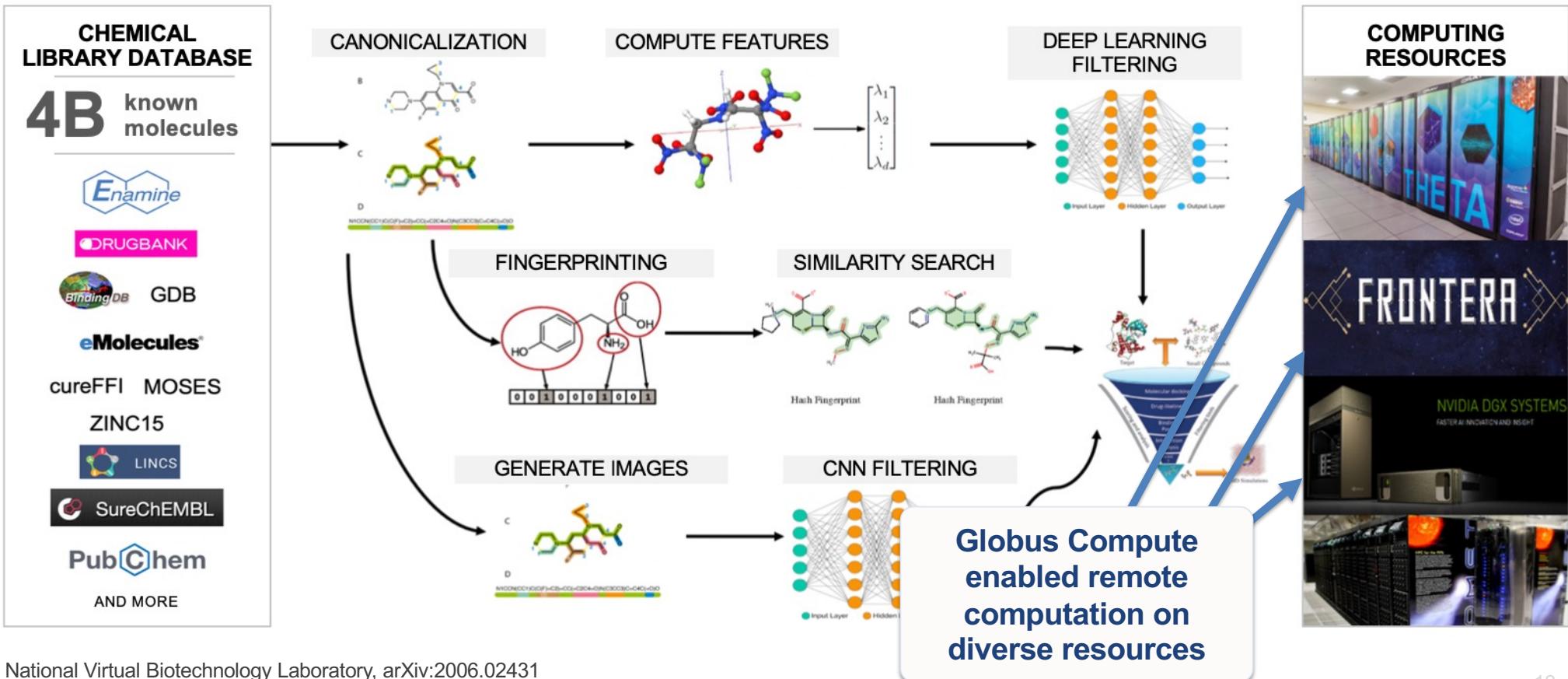
-Dan Ardia, Charles A. Dana Professor of Biology, F&M College

FRANKLIN & MARSHALL COLLEGE



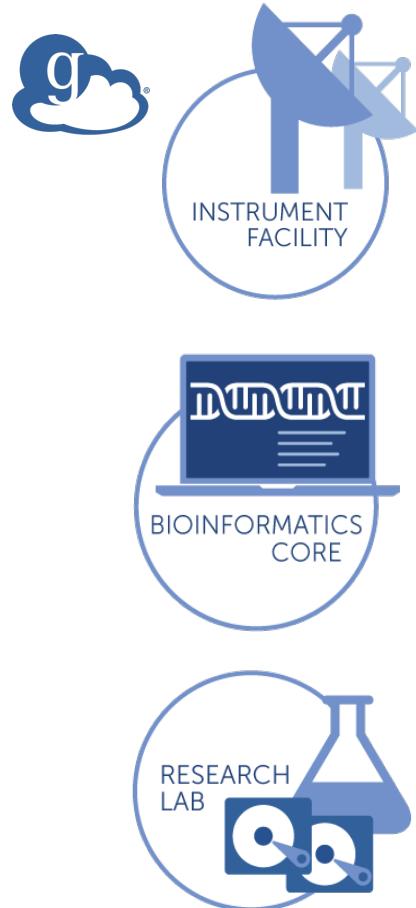


# Exemplar Use Case: ML-based drug screening



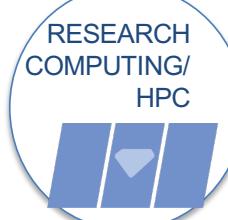


# Need to navigate...



## Experimental and core facilities

## Local/remote compute systems



## Local/remote storage systems



## Discovery and repository services

### Domain Repository



### Generalist/Institutional Repository





# Stakeholder perspectives or wishlist

## Researchers

No/low-friction path for access to data and computing systems by me and my collaborators

## Core Facility Staff

Reliable data egress maximize instrument utilization, and user support for analysis

## System Administrators

Managed, centralized visibility into system utilization, access policies, etc.

## Librarians/ Repository Managers

Simplified creation and implementation of DMPs, meet FAIR data standards

## Cybersecurity and Privacy Staff

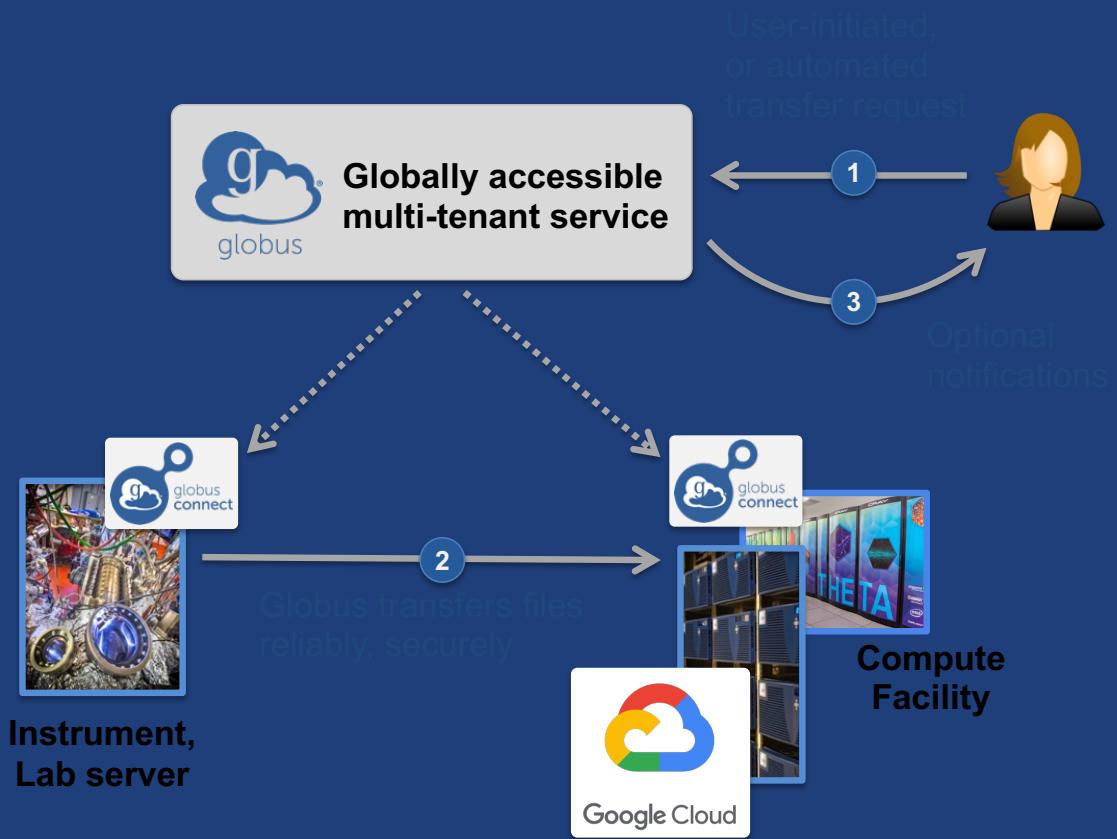
Auditable enforcement of variety of compliance requirements



# Unified research data access and sharing

# Fast, reliable file transfer ...from any to any system

- Fire-and-forget transfers/sync
- Optimized speed
- Assured reliability
- Unified view of storage
- HTTP/S access to data





# Intuitive web application interface

The screenshot shows the gCloud File Manager interface. The left sidebar contains navigation links: FILE MANAGER (selected), BOOKMARKS, ACTIVITY, COLLECTIONS (21), GROUPS, CONSOLE, FLOWS, COMPUTE, SETTINGS, LOGOUT, and HELP & SITEMAP. The main area is titled 'File Manager' and shows a 'Collection' of 'UChicago RCC Midway3'. The 'Path' is set to '/~/'. The file list includes:

NAME	LAST MODIFIED	SIZE
E099_HPPG_100_55_025C_att06...	3/17/2023, 11:2...	110.45 KB
E099_HPPG_100_55_025C_att06...	3/17/2023, 11:2...	113.91 KB
esgf_demo	3/11/2023, 12:1...	—
GW_Demo	4/18/2023, 02:...	—
TestFolder	9/30/2022, 12:...	—
TestUser1	3/20/2023, 05:...	—

On the right, a modal window titled 'ALCF#dtn\_eagle' is open, prompting for 'ALCF Username' and 'Password'. It also displays a note about the system being a Federal computer system and a warning about unauthorized use.

ALCF Username:

Password:

Cryptocard or Mobile token password

**Sign In**

This is a Federal computer system and is the property of the United States Government. It is for authorized use only. Users (authorized or unauthorized) have no explicit or implicit expectation of privacy.

Any or all uses of this system and all files on this system may be intercepted, monitored, recorded, copied, audited, inspected, and disclosed to authorized site, Department of Energy, and law enforcement personnel, as well as authorized officials of other agencies, both domestic and foreign. By using this system, the user consents to such interception, monitoring, recording, copying, auditing, inspection, and disclosure at the discretion of authorized site or Department of Energy personnel.

Unauthorized or improper use of this system may result in:



# Transfer/sync options

Start

Transfer & Timer Options

Start

Label This Transfer

Transfer Settings

NOTE: These settings will persist during this session unless changed.

sync - only transfer new or changed files

where the **✓ modification time is newer**

File size is different

File does not exist on destination

Checksum is different

delete files on destination that do not exist on source

preserve source file modification times

do NOT verify file integrity after transfer

encrypt transfer

Skip files on source with errors

Fail on quota errors

Notification Settings

Disable success notification

Disable failure notification

Disable inactive notification



# Globus Connectors support diverse systems



Google Cloud

Microsoft Azure  
Blob Storage



IBM Cloud



Google Drive



IBM Spectrum Scale



iRODS®

Quantum. ACTIVE SCALE.™



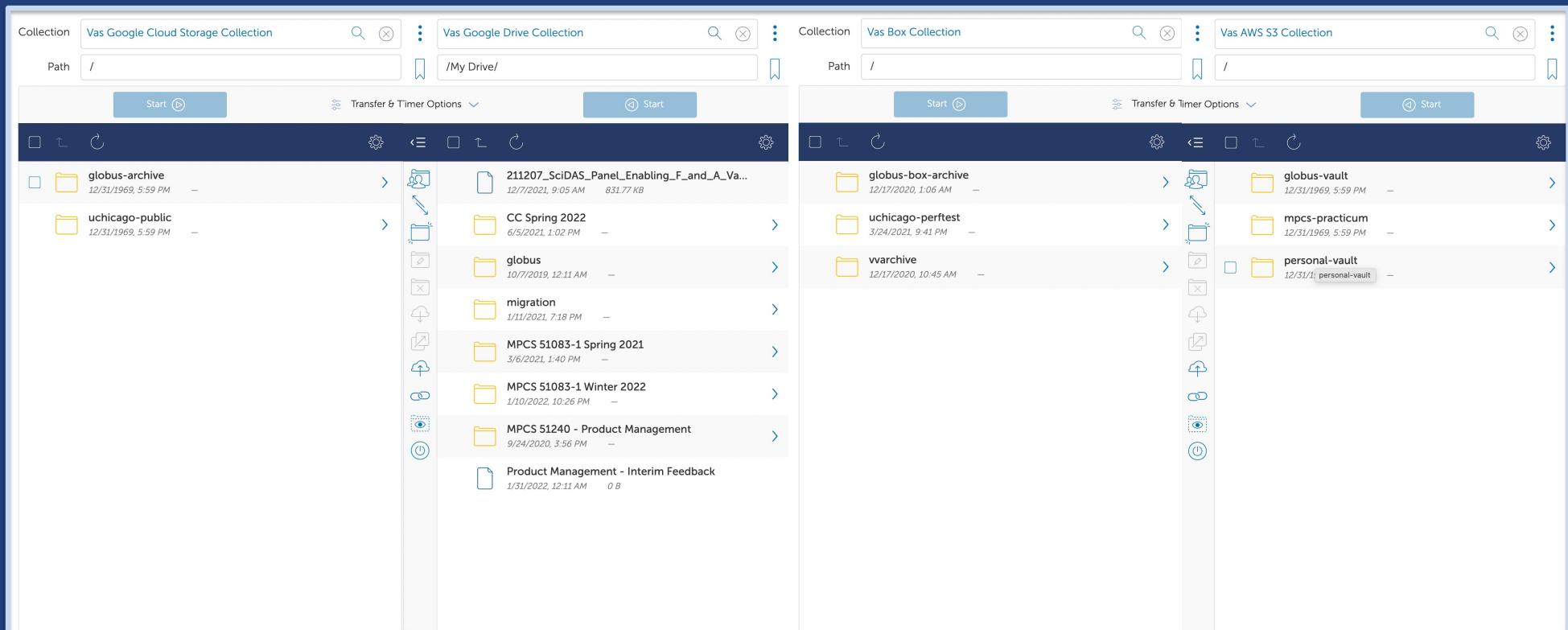
HPSS

wasabi  
hot cloud storage

lustre™



# Uniform interface, consistent user experience





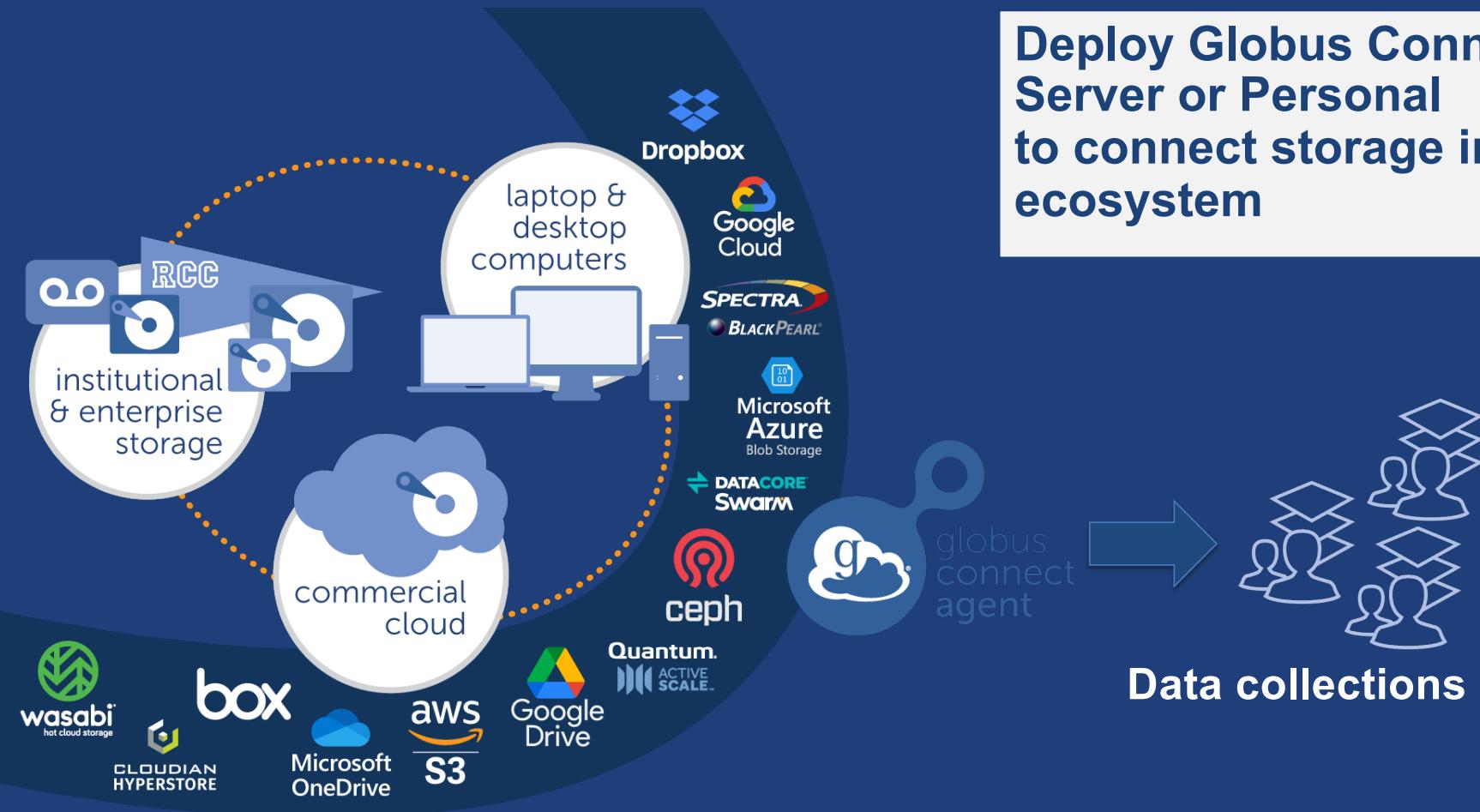
## What types of storage systems do you use/offer?



- Object store – on prem
- Tape archives
- AWS S3
- Google – Cloud/Drive
- MS – Blob/OneDrive
- Dropbox
- Box



# Enabling data ecosystem



**Deploy Globus Connect Server or Personal to connect storage into the ecosystem**

**Data collections**



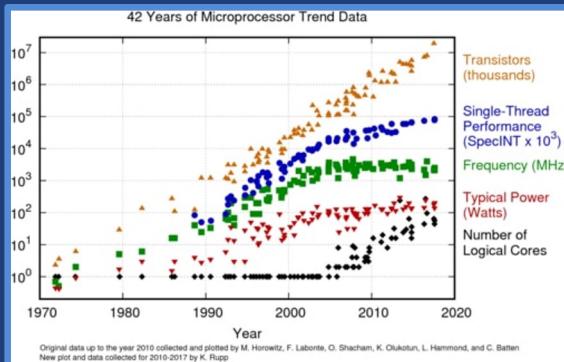
# Unified access to compute resources



# Why do we need to rethink research computing?

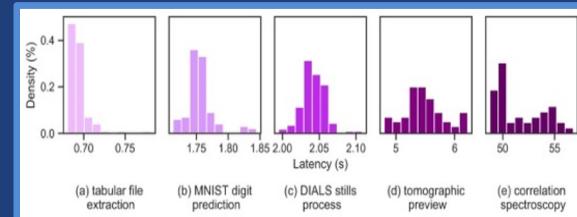
## Resources

- Hardware specialization
- Specialization leads to distribution



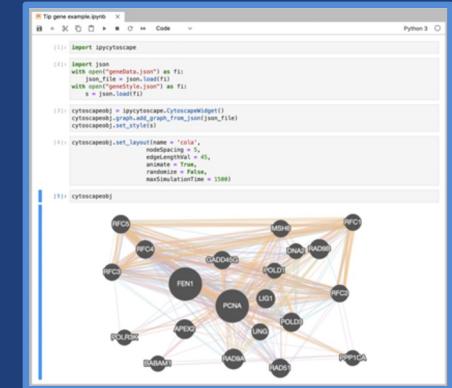
## Workloads

- Interactive, real-time workloads
- Machine learning training and inference
- Components may best be executed in different places



## Users

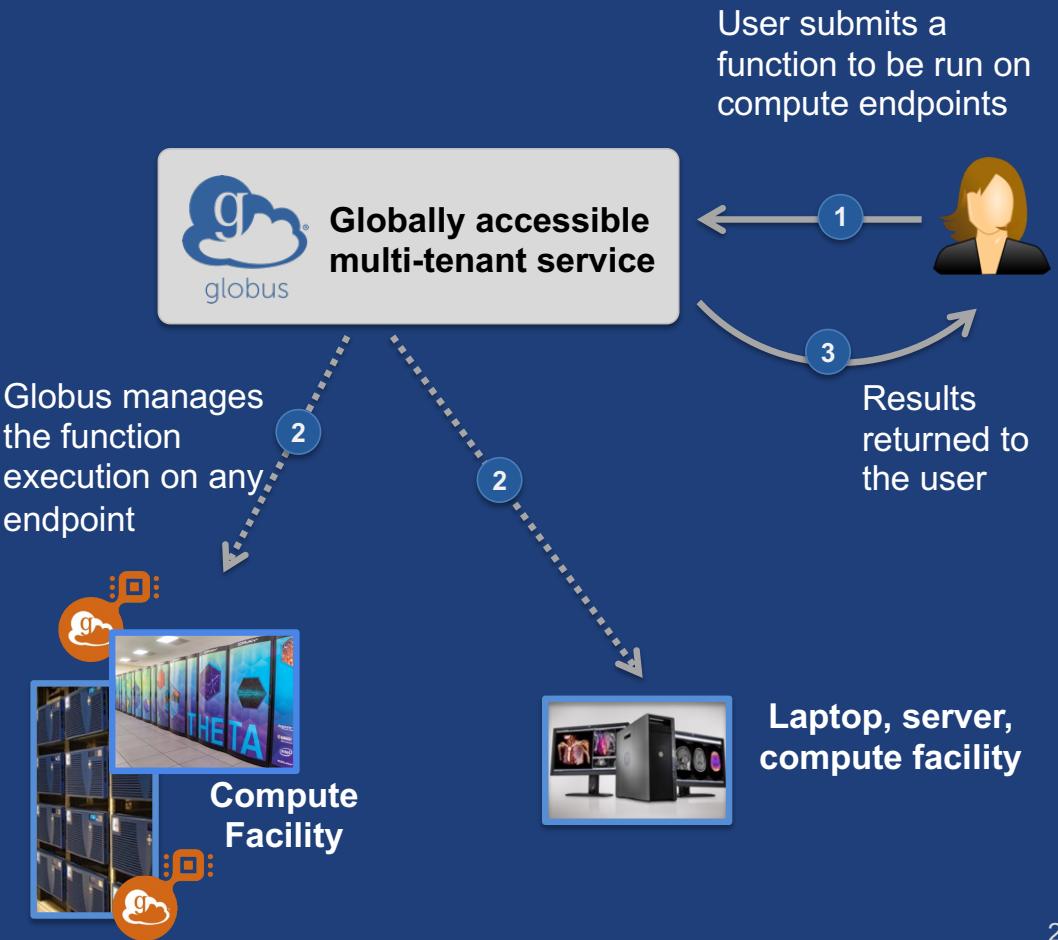
- Diverse backgrounds and expertise
- Different user interfaces (e.g., notebooks)





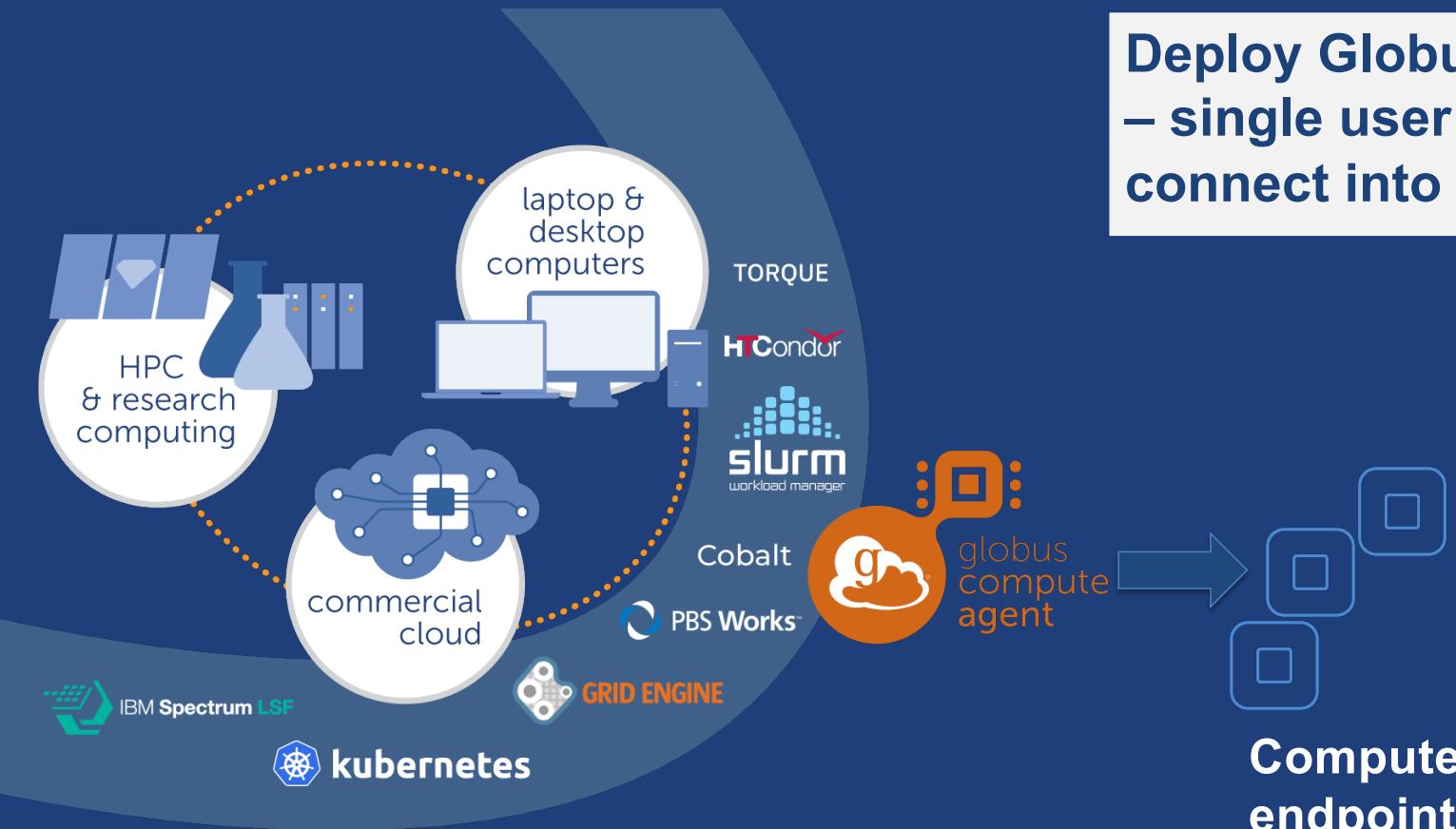
# Managed compute ...on any system

- **Programmatic access compute resources**
- **Consistent user interface**
- **“Fire and forget” function execution**
- **Federated authentication, and local access control**
- **Support use of Python for functions**





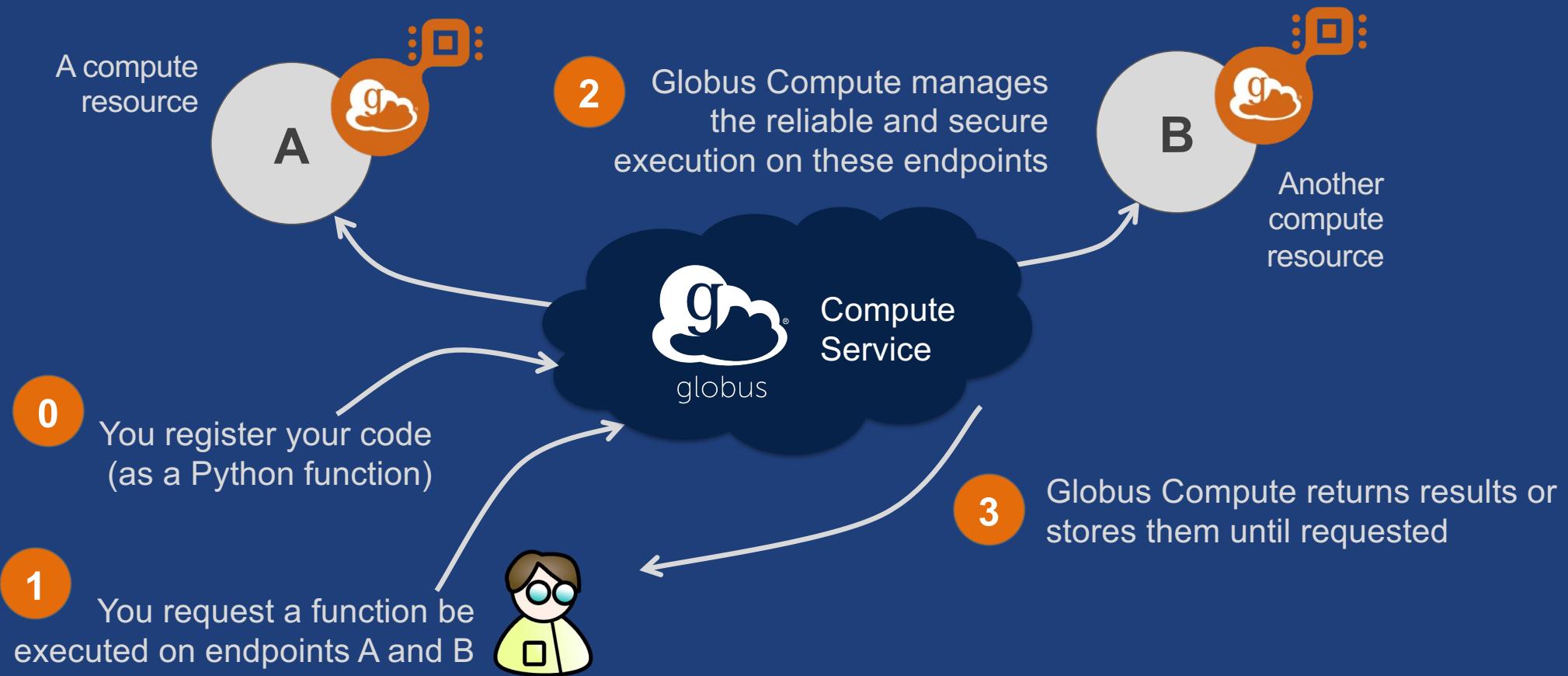
# Compute ecosystem



**Deploy Globus Compute agent**  
– single user or multi user to connect into the ecosystem



# How does it look from the researcher's PoV?



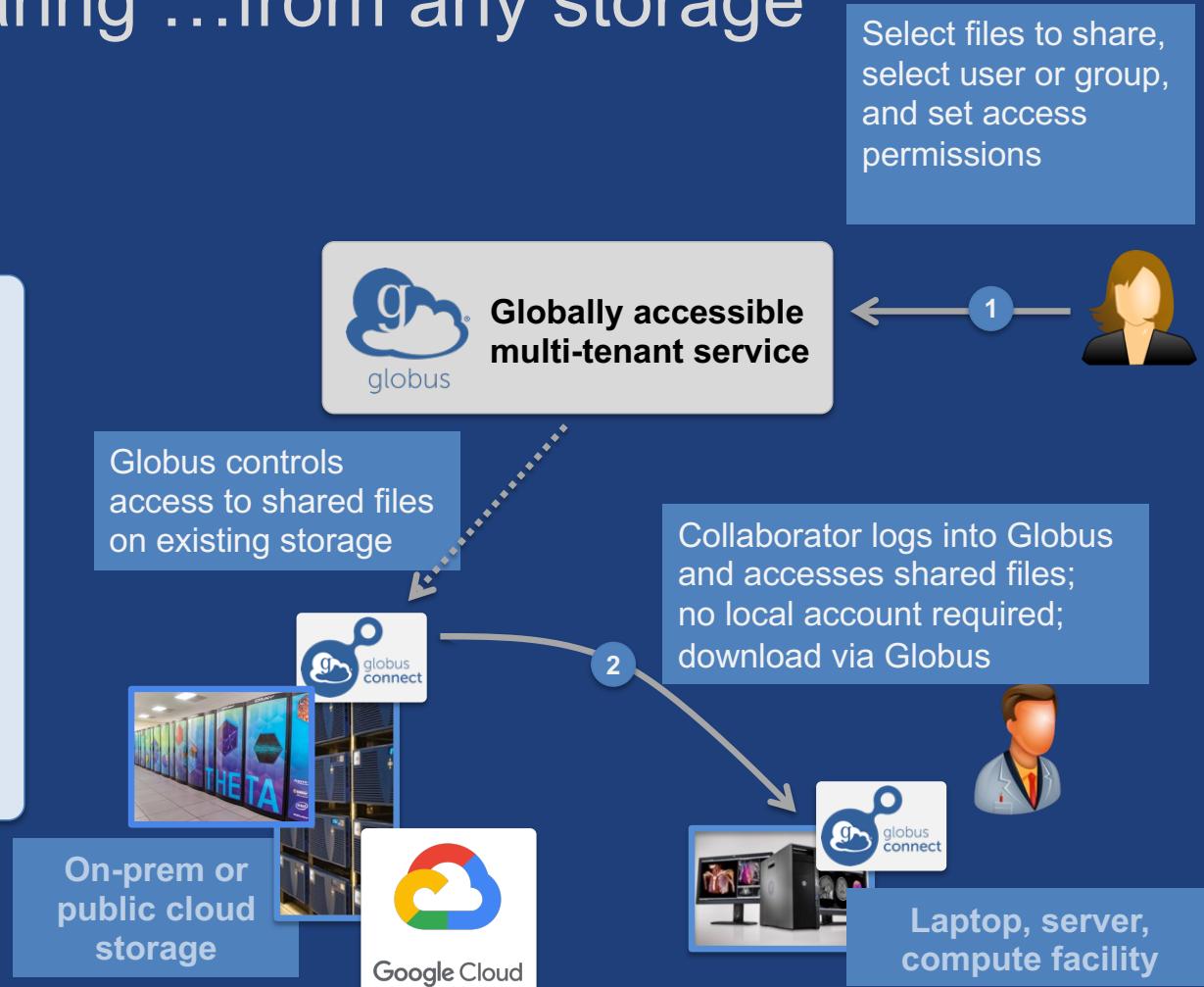


# Compliant and secure collaboration



# Secure data sharing ...from any storage

- Fine-grained access control “overlay” on storage system
- Share with any identity, email, group
- No need to stage data just for sharing





# Data sharing – permissions & roles

Presentation Materials - RA

Overview Permissions Roles

USER OR GROUP	CREATED	EXPIRATION	READ	WRITE
Brigitte Raumann (braumann@uchicago.edu)	3/8/2024, 04:33 PM	never expires	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Identity 6afa2dc5-d219-4078-9a06-ba37aa32c739 (6afa2dc5-d219-4078-9a06-ba37aa32c739@clients.auth.globus.org)	12/11/2023, 06:29 PM	never expires	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Path: /

Show link for sharing

Path: /RPI/

Show link for sharing

USER OR GROUP	CREATED	EXPIRATION	READ	WRITE
Public	2/27/2024, 03 PM			
Identity a7a2bd06-919b-478e-9477-447f14198a63 (foster@anl.gov)	5/3/2024, 04:22 PM			

Path: /Stanford/

Show link for sharing

USER OR GROUP	CREATED	EXPIRATION	READ	WRITE
All Users	12/11/2023, 06 PM			

Assigned Roles

Assign New Role

Role	Owner	Administrator	Access Manager
Rachana Ananthakrishnan	ranantha@uchicago.edu		
Rachana Ananthakrishnan	ranantha@uchicago.edu		
7d69a95e-48a8-491e-8e72-ed1631e81954	7d69a95e-48a8-491e-8e72-ed1631e81954@clients.auth.globus.org		



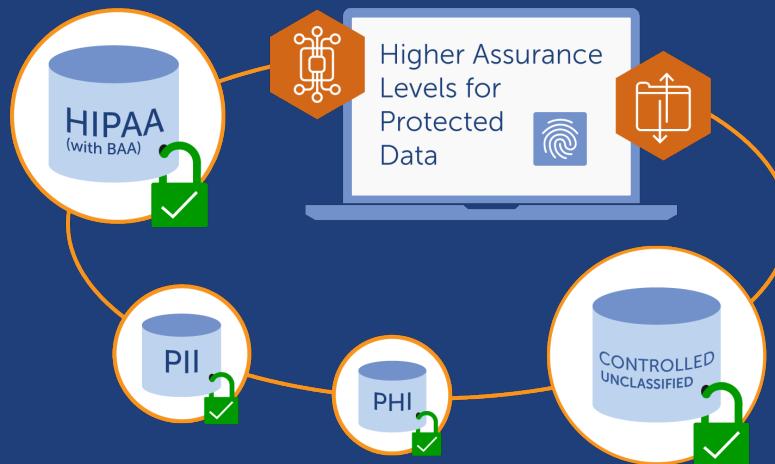
## Common sharing scenarios enabled by Globus

- **Disable sharing; not very FAIR, but may be necessary**
- **Allow sharing only by specific users/groups**
- **Allow sharing only from specific systems/directories**
- **Limit sharing to specific institution(s)**
- **Limit access based on time**



# Globus High Assurance for managing protected data

## Security controls



**Restricted data handling**  
→ PHI, PII, CUI  
→ Compliant data sharing

**BAA w/UChicago**



## What are some of your compliance requirements?



- NIST 800-53
- NIST 800-171
- SOC-2
- HIPAA
- CMMC
- GDPR/State Privacy laws
- Others



# Core Facility data automation

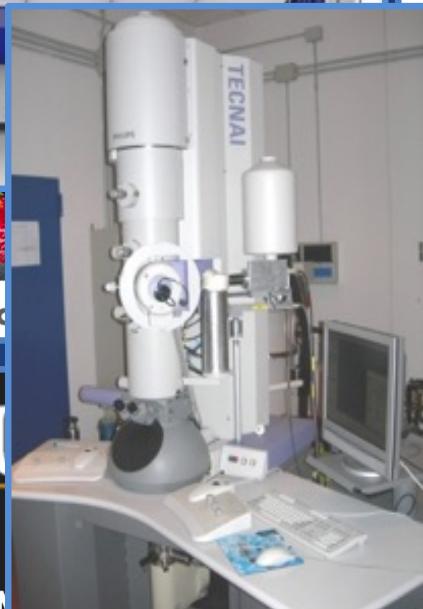


High Resolution  
Cryo-EM

National Cryo-Electron Microscopy

**PURD**  
SCIEN

Purdue Cryo-EM Facility



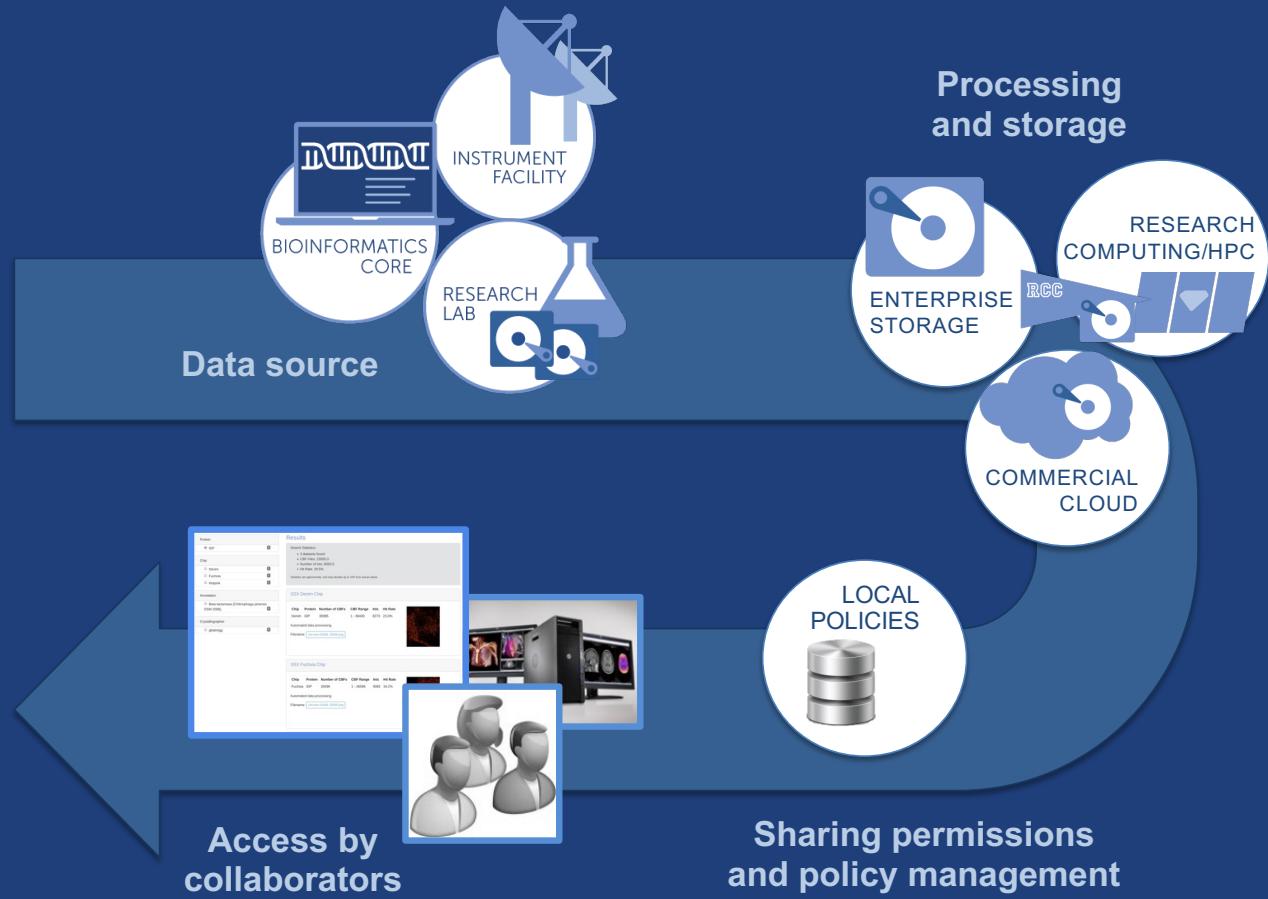
**ADVANCED LIGHT SOURCE**

**NISC**  
NIH Intramural Sequencing Center





# Instrument data processing pattern



At minimum, make data available to others in your lab, campus colleagues and/or external collaborators



## How do you interact with core/instrument facilities?

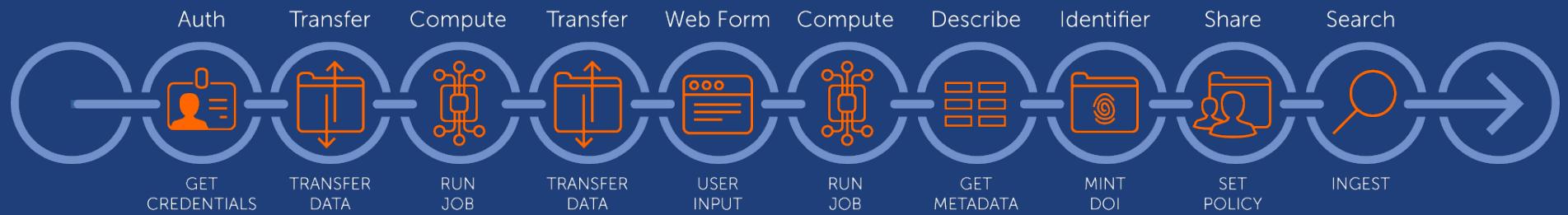


- User of a local facility
- User of remote facility
- Administrator/manager for facility
- Provide services to facility (storage/compute/etc)
- Other interactions



# Globus Flows for managed automation

- **Managed, secure, reliable task orchestration—at scale**
- **Define, run, and share distributed research pipelines**
- **Event driven execution model**





# Automating cryoEM

Globus  
Flows



Transfer



Transfer  
raw files

Compute



Launch  
analysis job

Carbon!

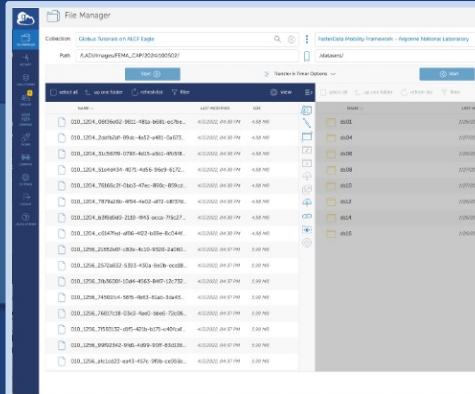


Correct,  
classify, ...

Compute



Extract  
metadata



Share



Set access  
controls

Transfer

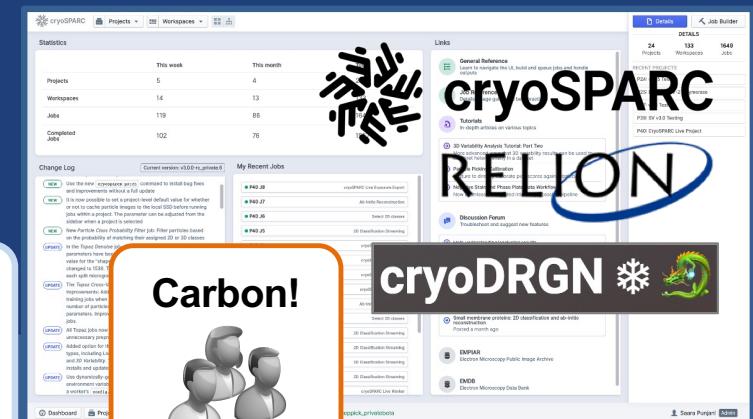


Move final  
files

Search



Ingest to  
index



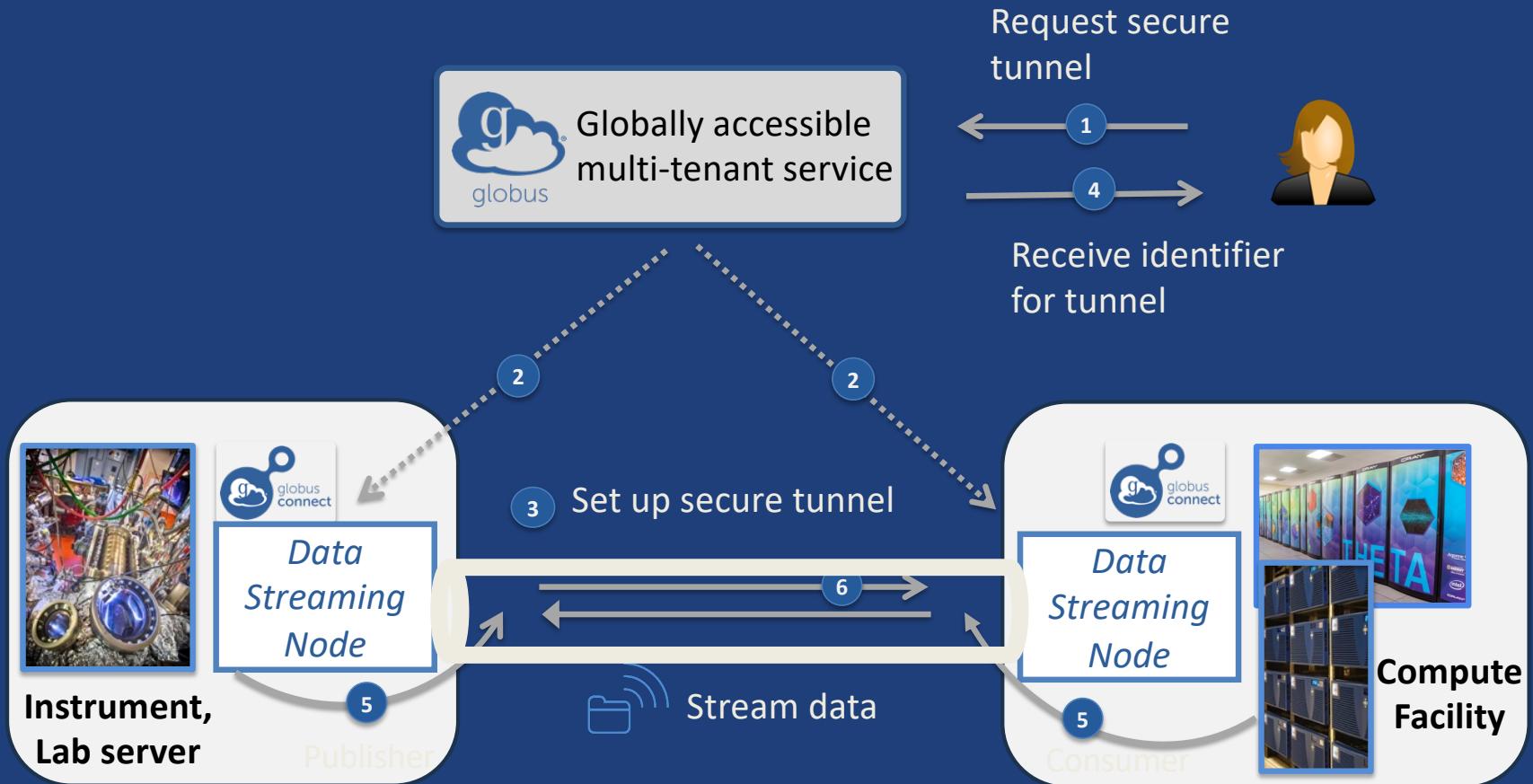
cryoDRGN



RELION

RELION

# Streaming data using Globus

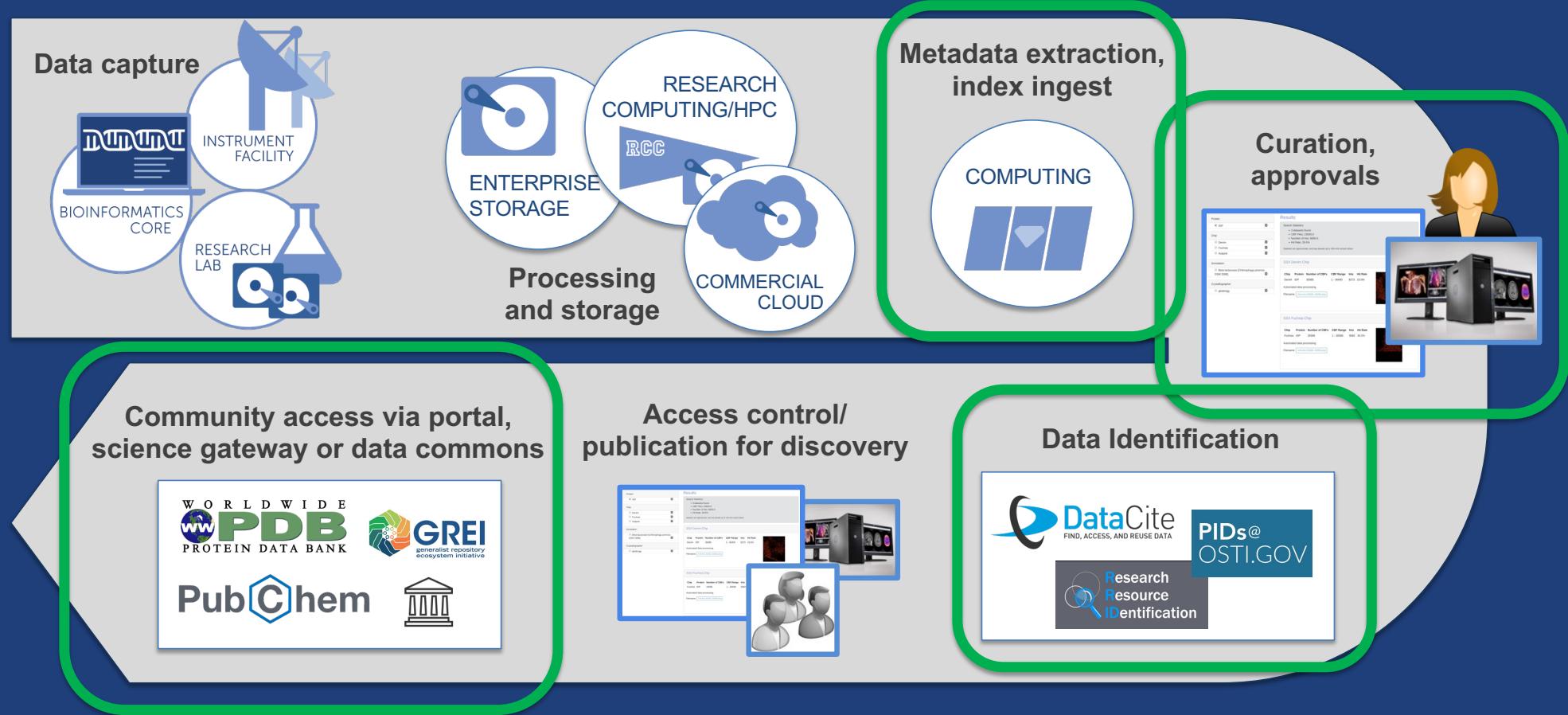




# Beyond automation



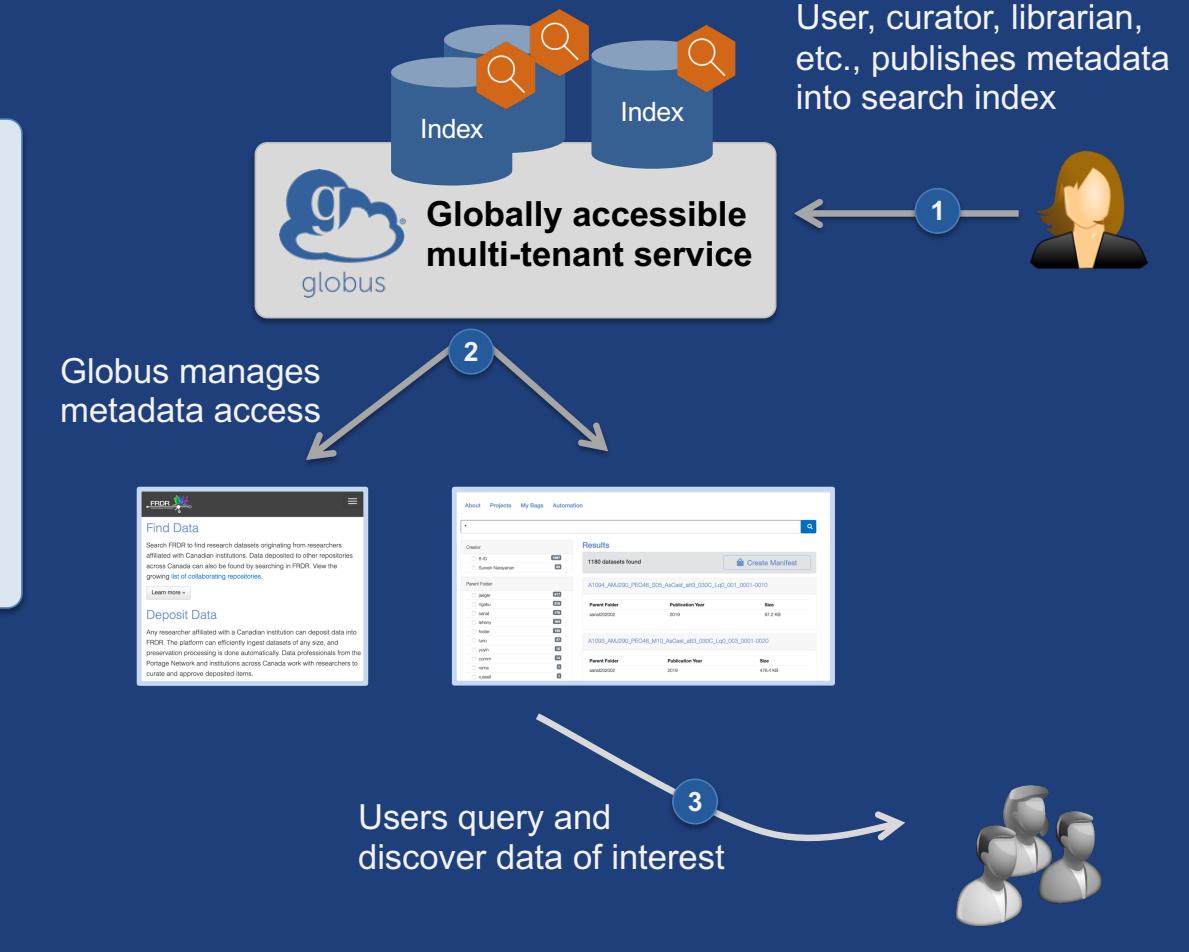
# Flows to ensure “FAIRness”





# Globus Search: Data discovery for any domain

- Scalable metadata store
- Fine grained visibility controls
- Schema agnostic
- Federated auth integration
- Queried via API with facets





# An example of how we can get to FAIR data

**Cancer Registry Records for Research (CR3)**  
federated network of cancer registry data

Home

Search All

Filters [0]

**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 (36,330)
- 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)

**434,150 available records**

Registry

Age Group at Diagnosis

Diagnosis per Year

AJCC Stage/Best CS

Race

Gender

Make a Request  braumann@uchicago.edu

X

Federated login → instant access using your institutional credentials

Google-like text search with facets for filtering

Variable facets based on source registry index

Dynamically updating charts as facets change

Restricted visibility to sensitive data



# Flows and Transfer for data lifecycle management

Tar and transfer

Secure data  
egress

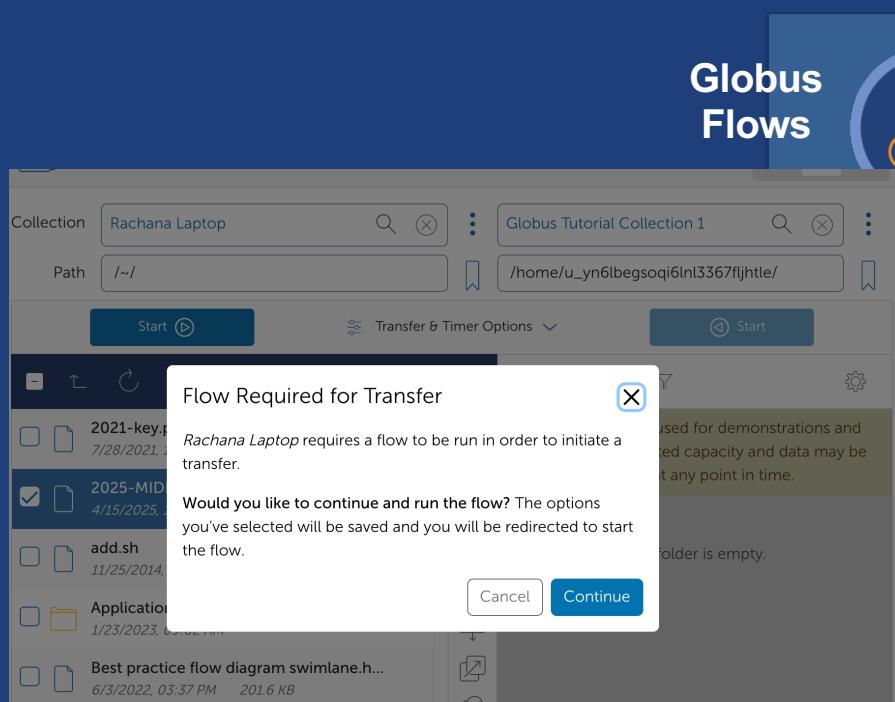
Extract metadata  
and push to an  
index

Automated  
processing on data  
deposit

***Configure a flow to run when user transfers data to or  
from a collection***



# Secure egress



## Globus Flows

### Transfer



Transfer files staging area

### Compute



Launch screening job

### Notify



Notify admin

### Transfer



Move final files

### Choice



Admin approval

User initiates a transfer, and policy triggers a flow to process request



# Using Globus Platform



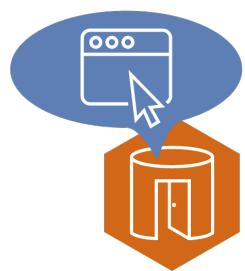
Globus Python  
SDK



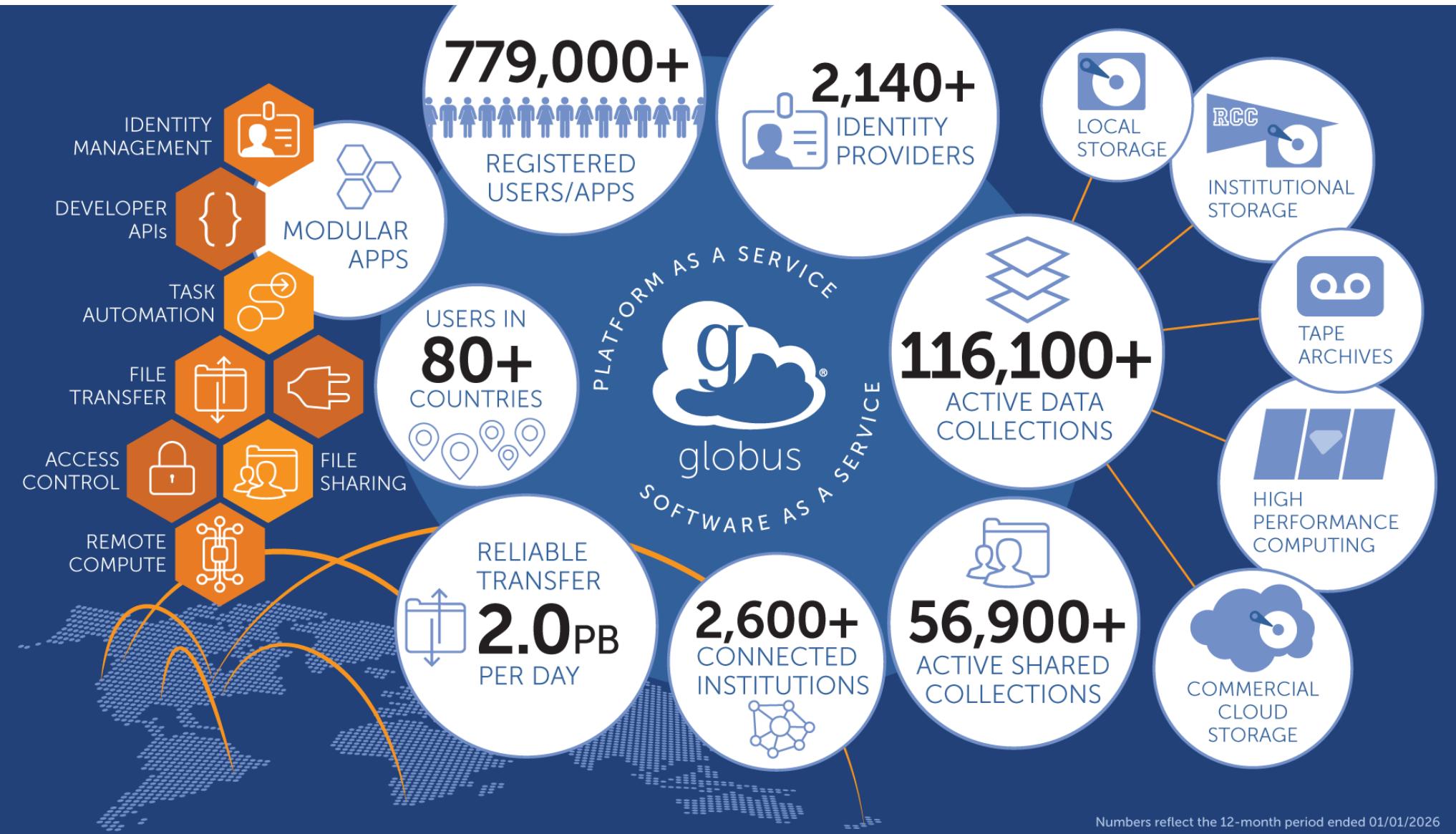
Globus Django  
Portal Framework



Globus JavaScript  
SDK



Globus Serverless  
Portal





# Thank you!

- **Engage: [ranantha@uchicago.edu](mailto:ranantha@uchicago.edu)**
- **Website: [globus.org](https://globus.org)**
- **Documentation: [docs.globus.org](https://docs.globus.org)**
- **Support: [support@globus.org](mailto:support@globus.org)**