Automating Instrument Science at Scale Using Globus

Vas Vasiliadis
vas@uchicago.edu

August 2, 2022

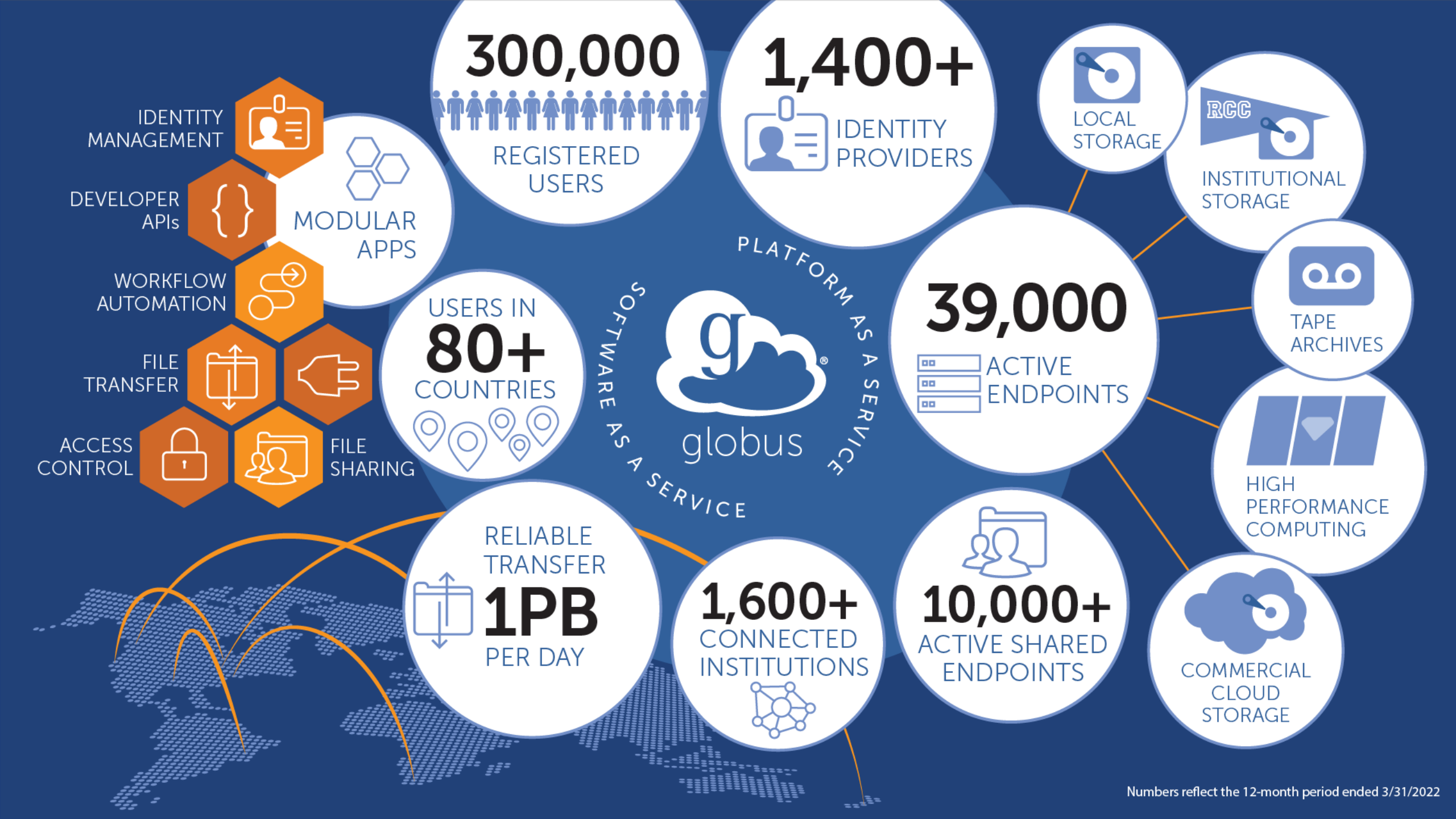
Globus is …

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


300,000
REGISTERED USERS
1,400+
IDENTITY PROVIDERS
IDENTITY MANAGEMENT
DEVELOPER APIs
WORKFLOW AUTOMATION
FILE TRANSFER
ACCESS CONTROL
FILE SHARING
MODULAR APPS
USERS IN 80+ COUNTRIES
SOFTWARE AS A SERVICE
PLATFORM AS A SERVICE
globus
LOCAL STORAGE
RCC
INSTITUTIONAL STORAGE
TAPE ARCHIVES
39,000
ACTIVE ENDPOINTS
HIGH PERFORMANCE COMPUTING
RELIABLE TRANSFER
1PB
PER DAY
1,600+
CONNECTED INSTITUTIONS
10,000+
ACTIVE SHARED ENDPOINTS
COMMERCIAL CLOUD STORAGE
Numbers reflect the 12-month period ended 3/31/2022

Development is funded by...

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

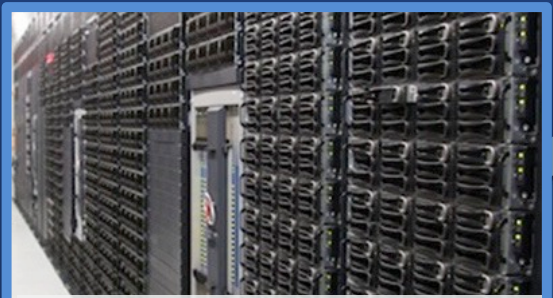
NIH

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

Argonne
NATIONAL LABORATORY

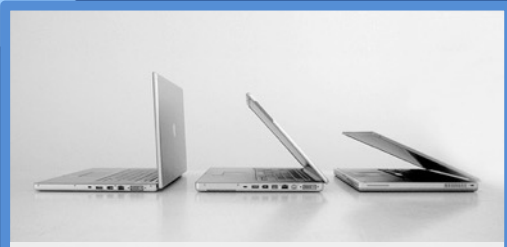
Operations are funded by subscribers

We unify data access across disparate systems…



Research Computing HPC

National Resources



Personal Systems

"I need to easily, securely and reliably move or replicate my data between systems."

Desktop Workstations

Archives

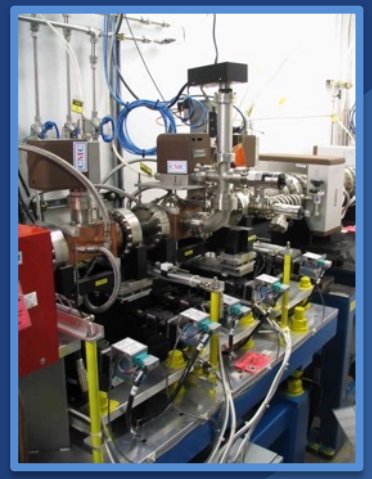
Instruments

Public Cloud Storage

…simplify secure sharing with collaborators…

…help researchers manage instrument data…

Next-Gen Sequencer



Advanced Light Source

Cryo-EM

Analysis store

High-durability, low-cost store

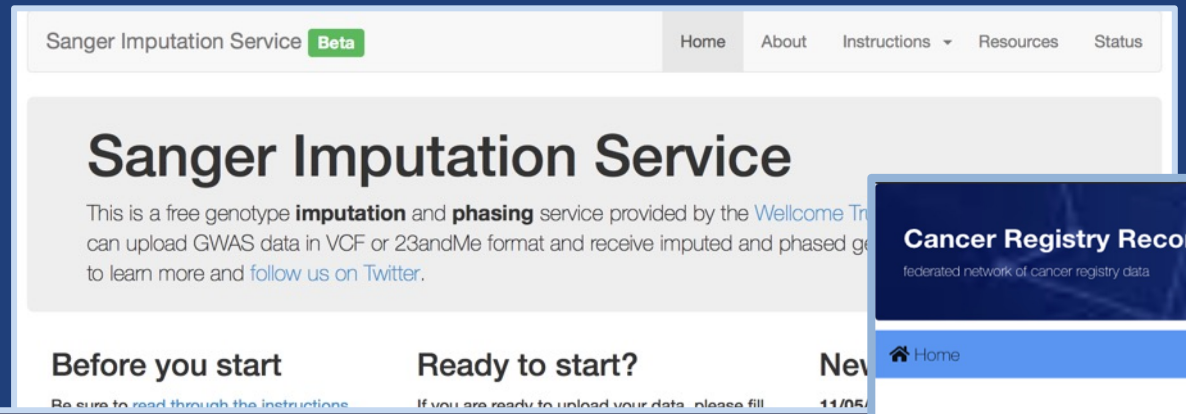
MRI

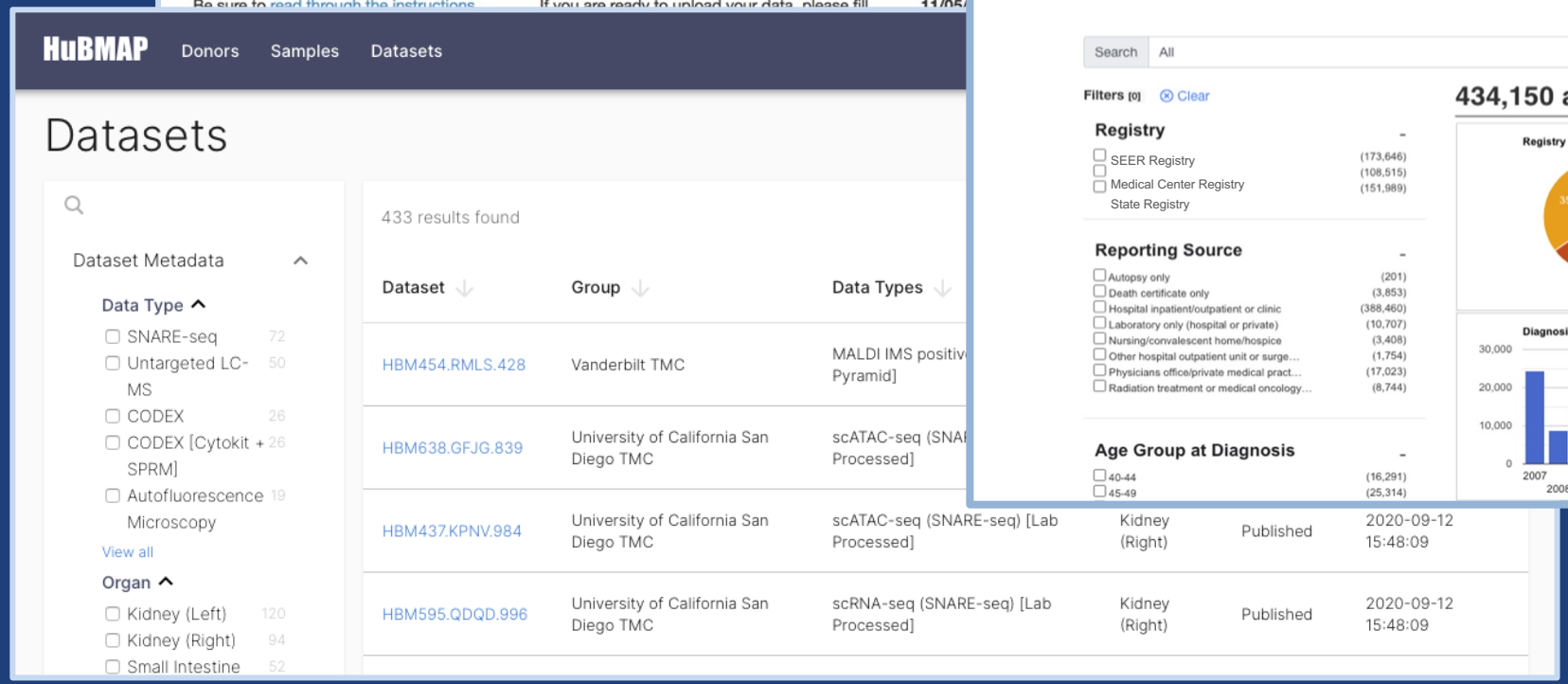
Light Sheet Microscope

Remote visualization

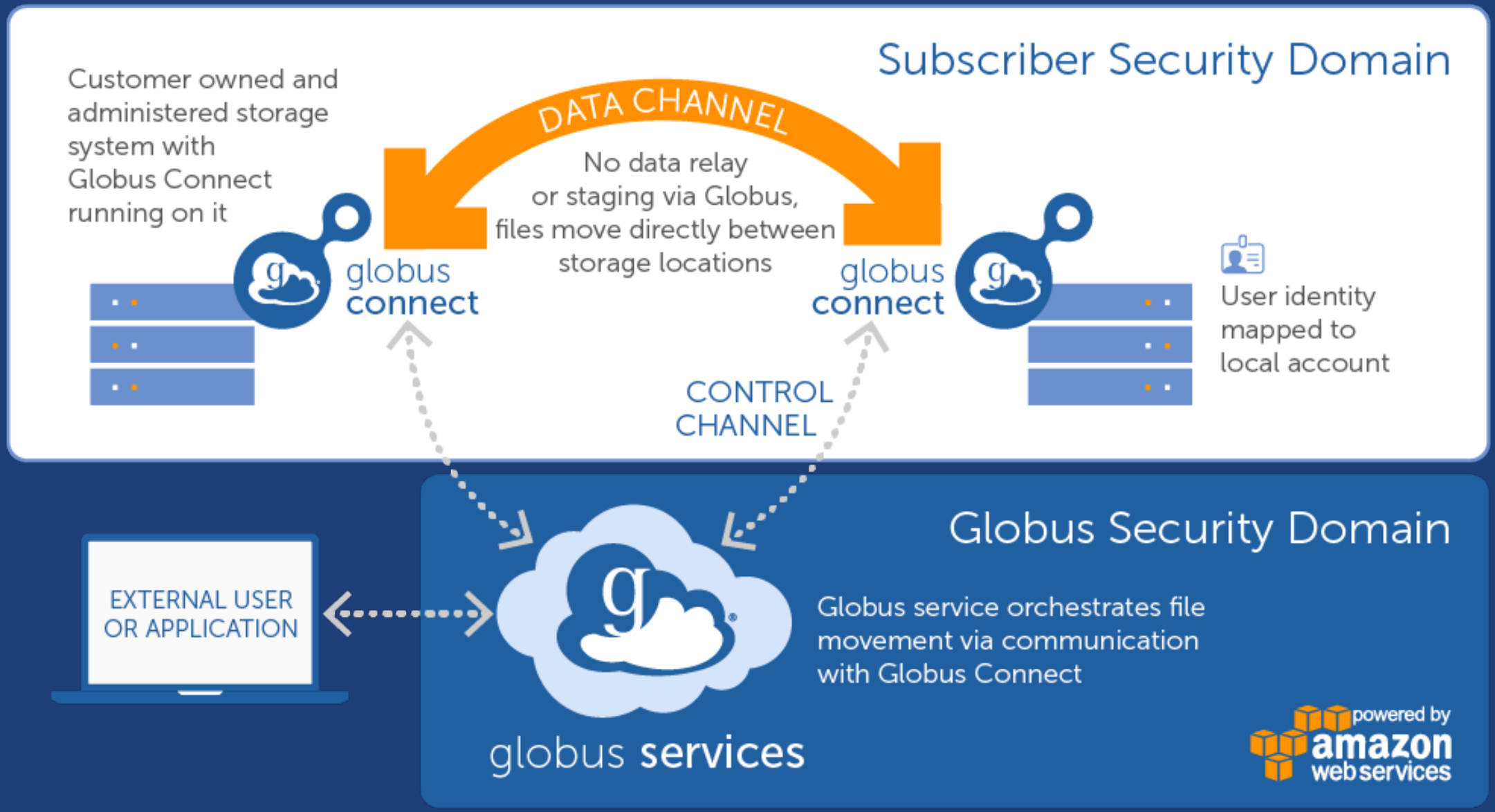
Personal system

…and build data-centric applications





Hybrid SaaS architecture



Currently supported systems

aws S3

Google Cloud

Microsoft Azure Blob Storage

DATACORE Swarm

SCALITY

IBM Cloud

Microsoft OneDrive

Google Drive

IBM Spectrum Scale

box

iRODS

Quantum

HPSS

wasabi
hot cloud storage

lustre

A quick tour…

Globus Automation Capabilities

Timer Service

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

Command Line Interface

Ad hoc scripting and integration

Globus Flows service

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

Three perspectives

- **End user: ease of use, scalability**
- **Administrator: visibility, access control**
- **Builder: tooling, rapid prototyping**

“Simple” Automation Use Cases

- **Data backup – as user, as system**
- **Stage data in or out as part of a compute job**
- **Data portal/gateway submits a transfer of compute results as the user**
- **Data portal/gateway monitors transfer, initiates processing or backup of data**

Globus Timer Service

The Globus Timer service

- **Scheduled/recurring file transfers**
- **Supports all Globus transfer and sync options**
- **Service accessible via web app and CLI**
- **Example: NIH – hpc.nih.gov/storage/globus_cron.html**

Scheduled transfers using Globus timers

Scripting with the Globus Timer service

```
$ globus-timer session {login, logout, whoami}
$ globus-timer job transfer \
--name example-job \
--label "Timer Transfer Job" \
--interval 28800 \
--start '2020-01-01T12:34:56' \
--source-endpoint ddb59aef-6d04-11e5-ba46-22000b92c6ec \
--dest-endpoint ddb59af0-6d04-11e5-ba46-22000b92c6ec \
--item ~/file1.txt ~/new_file1.txt false \
--item ~/file2.txt ~/new_file2.txt false
```

Globus Command Line Interface (CLI)

Globus Command Line Interface

```
(globus-cli) jupiter:~ vas$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

Options:
  -v, --verbose                 Control level of output
  -h, --help                    Show this message and exit.
  -F, --format [json|text]      Output format for stdout. Defaults to text
  --map-http-status TEXT        Map HTTP statuses to any of these exit codes:
                                0,1,50-99. e.g. "404=50,403=51"

Commands:
  bookmark          Manage Endpoint Bookmarks
  config            Modify, view, and manage your Globus CLI config.
  delete            Submit a Delete Task
  endpoint          Manage Globus Endpoint definitions
  get-identities    Lookup Globus Auth Identities
  list-commands     List all CLI Commands
  login             Login to Globus to get credentials for the Globus CLI
  logout            Logout of the Globus CLI
  ls                List Endpoint directory contents
  mkdir             Make a directory on an Endpoint
  rename            Rename a file or directory on an Endpoint
  task              Manage asynchronous Tasks
  transfer          Submit a Transfer Task
  version           Show the version and exit
  whoami            Show the currently logged-in identity.
```

Automation of simple data management tasks

Integration with existing scripts (job submission …)

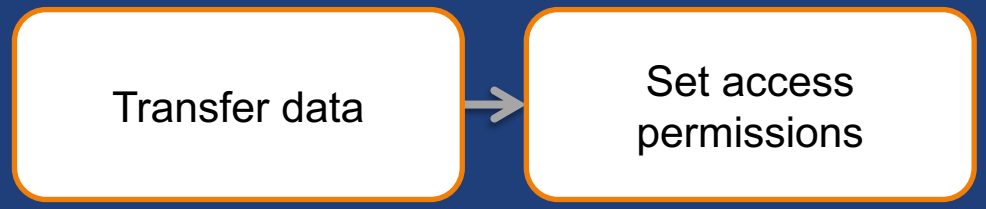
Open source, uses the Python SDK

Commands refer to resources by UUID

- **UUIDs for endpoint, task, user identity, groups…**
- **Use search/list options**
- `get-identities` **for identity username to UUID**

```
$ globus endpoint search 'Tutorial Endpoint 1'
$ globus task list
$ globus get-identities vas@globusid.org
bfc122a3-af43-43e1-8a41-d36f28a2bc0a
```

A common use case



- **Distribute results from computation**
- **Analyze raw data from an instrument**

Step 1: Transfer files

```
$ export src=<source_collection_UUID>
$ export dst=<destination_collection_UUID>
$ globus transfer --recursive $src:/carousel
$dst:/cli/images/MY_IMAGES
$ globus task show <transfer_task_UUID>
```

Step 2: Set permissions

- **Set and manage permissions on *guest collection***
- **Requires Access Manager role**

```
$ export share=<guest_collection_UUID>
$ globus endpoint permission create --permissions r --
identity demodoc@globusid.org $share:/cli/images/MY_IMAGES/
$ globus endpoint permission list $share
$ globus endpoint permission delete $share <perm_UUID>
```

Parsing CLI output

- **Default output is text; for JSON output use** `--format json`

```
$ globus endpoint search --filter-scope administered-by-me
$ globus endpoint search --filter-scope my-endpoints --
format json
```

- **Extract specific attributes using** `--jmespath <expression>`

```
$ globus endpoint search --filter-scope my-endpoints --
jmespath 'DATA[].[id, display_name]'
```

Your script is a “Native App” in Globus Auth

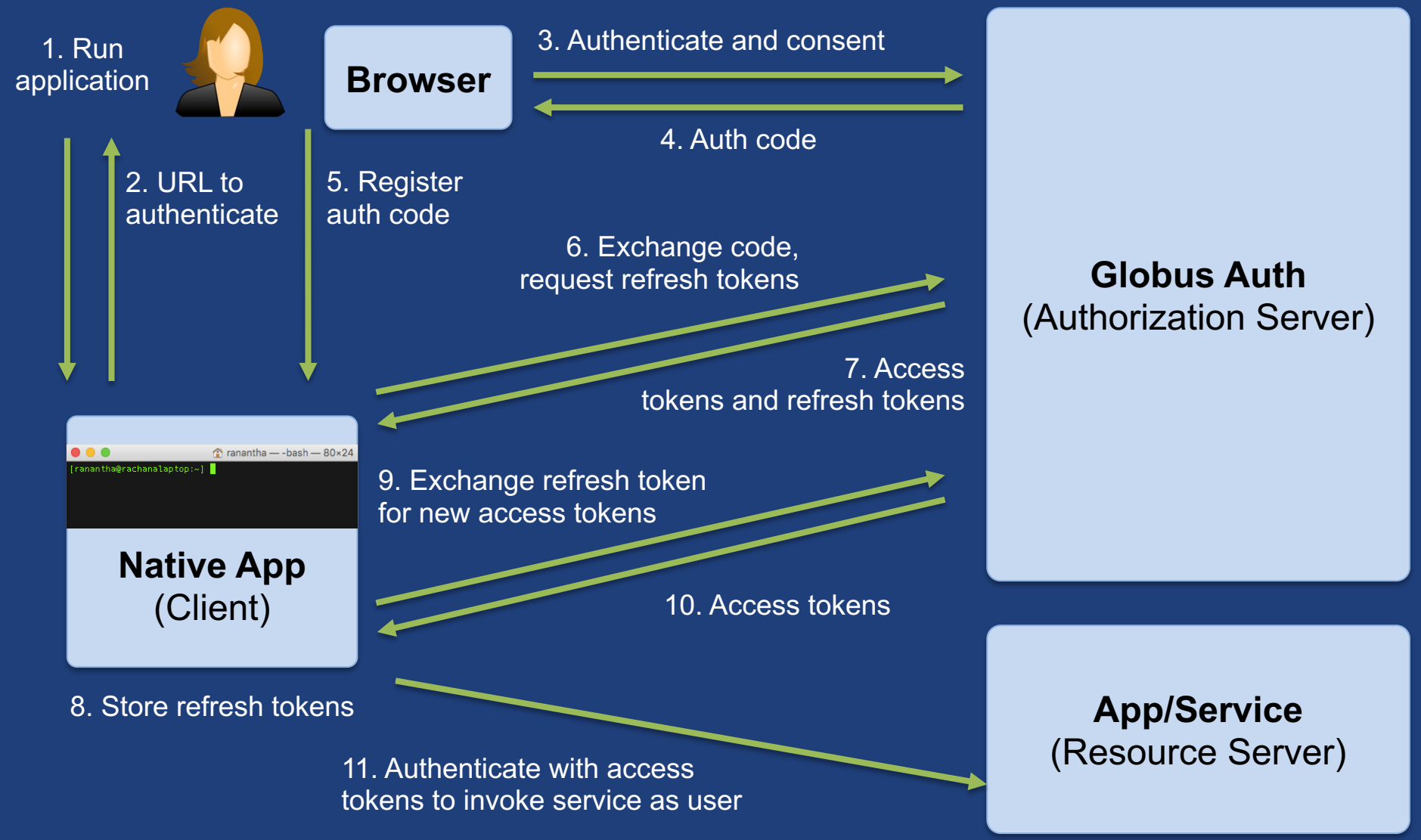
- **Native App: client that cannot keep a secret (script, mobile, Jupyter notebook, …)**
- **Register with Globus Auth → special callback URL**

Registering your script as a native app

Key Globus requirements for automation

- **Guest collections → no human in the loop for auth**
- **Delegate permissions management → app as Access Manager (apps are people too!)**
- **Evergreen auth → native app grant w/refresh tokens**
  - Client gets access + refresh tokens, in particular scope
  - Client uses refresh token to get access token
  - Refresh token good for 6 months after last use
  - Consent rescindment revokes resource token

Refresh tokens



Native App/Refresh Tokens Sample Code

github.com/globus/native-app-examples

- `./example_copy_paste.py`
  - User copies and pastes code to the app
- `./example_copy_paste_refresh_token.py`
  - Stores refresh token locally, uses it to get new access tokens

Globus Flows Service

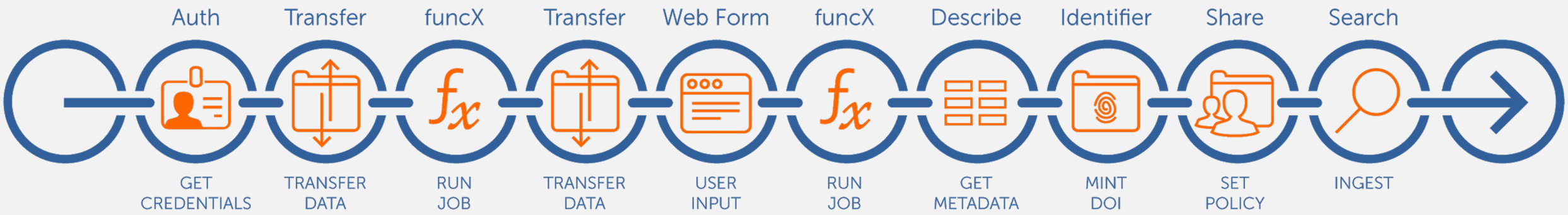
Automation using the Globus platform

Managed, secure, reliable task orchestration across heterogenous resources, using a declarative language for composition and an **event driven execution model**, **extensible** via custom actions, for automation at **scale**

The Globus Flows service

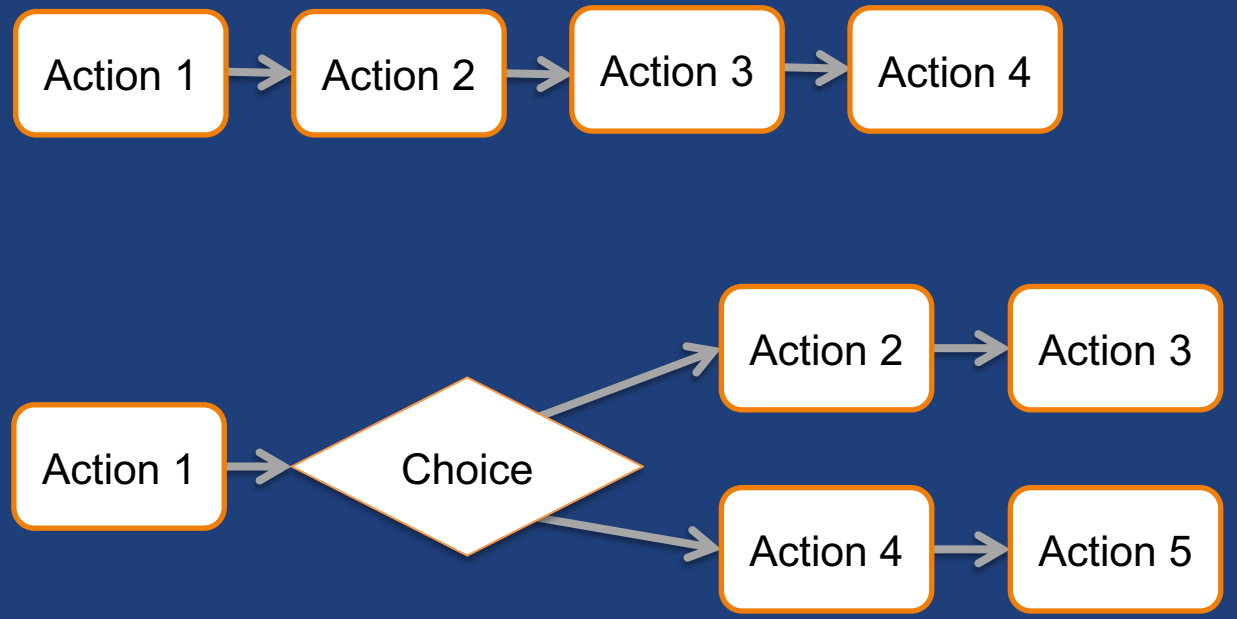
- **Flows**: A platform service for defining, applying, and sharing distributed research automation flows
- Flows comprise **Actions**
- **Action Providers**: Called by Flows to perform tasks
- **Triggers***: Start flows based on events

* Coming soon



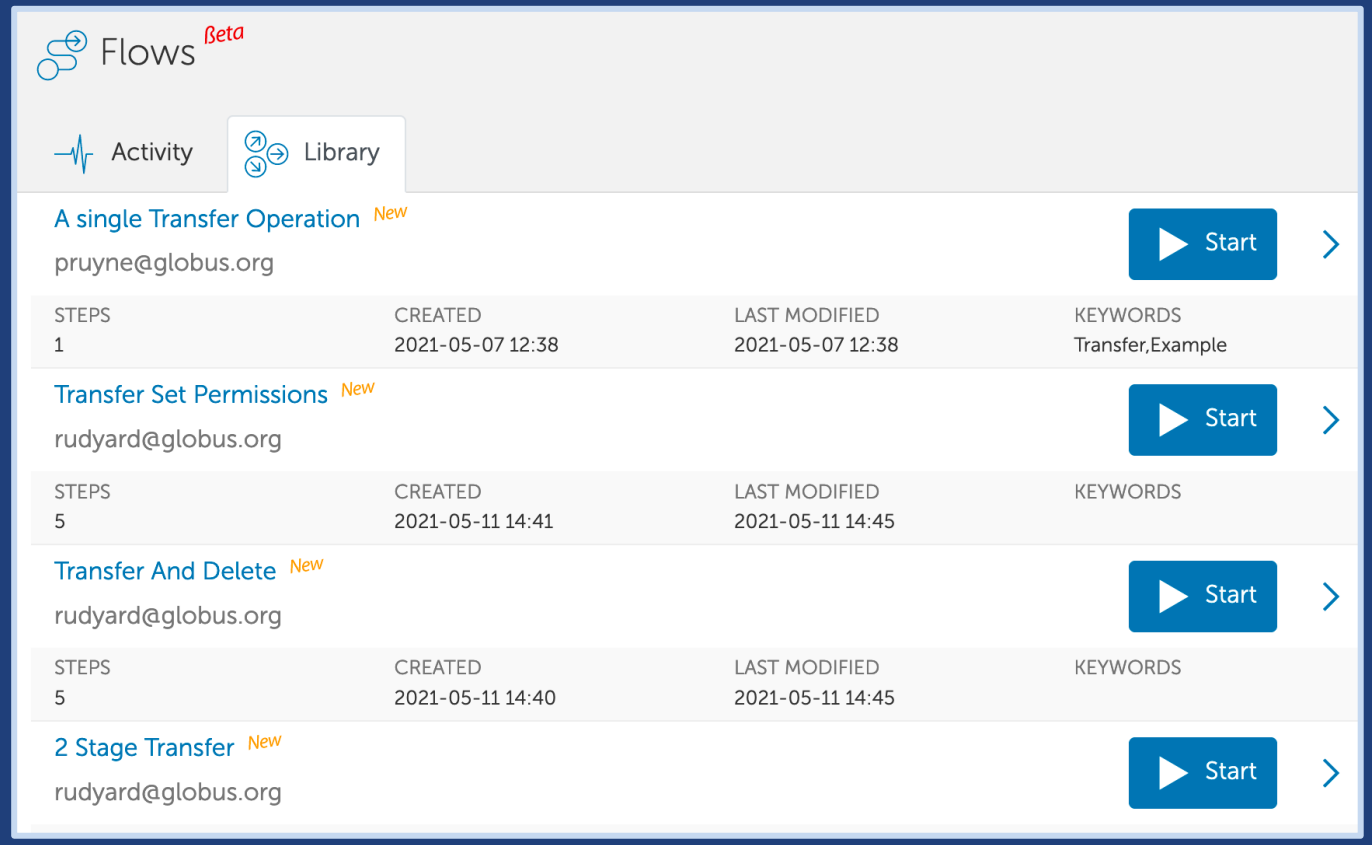
Creating and deploying flows

- **Define flows using a declarative language (JSON or YAML)**
- **Deploy flows to the Globus Flows service**
- **Set access policy for visibility and execution**



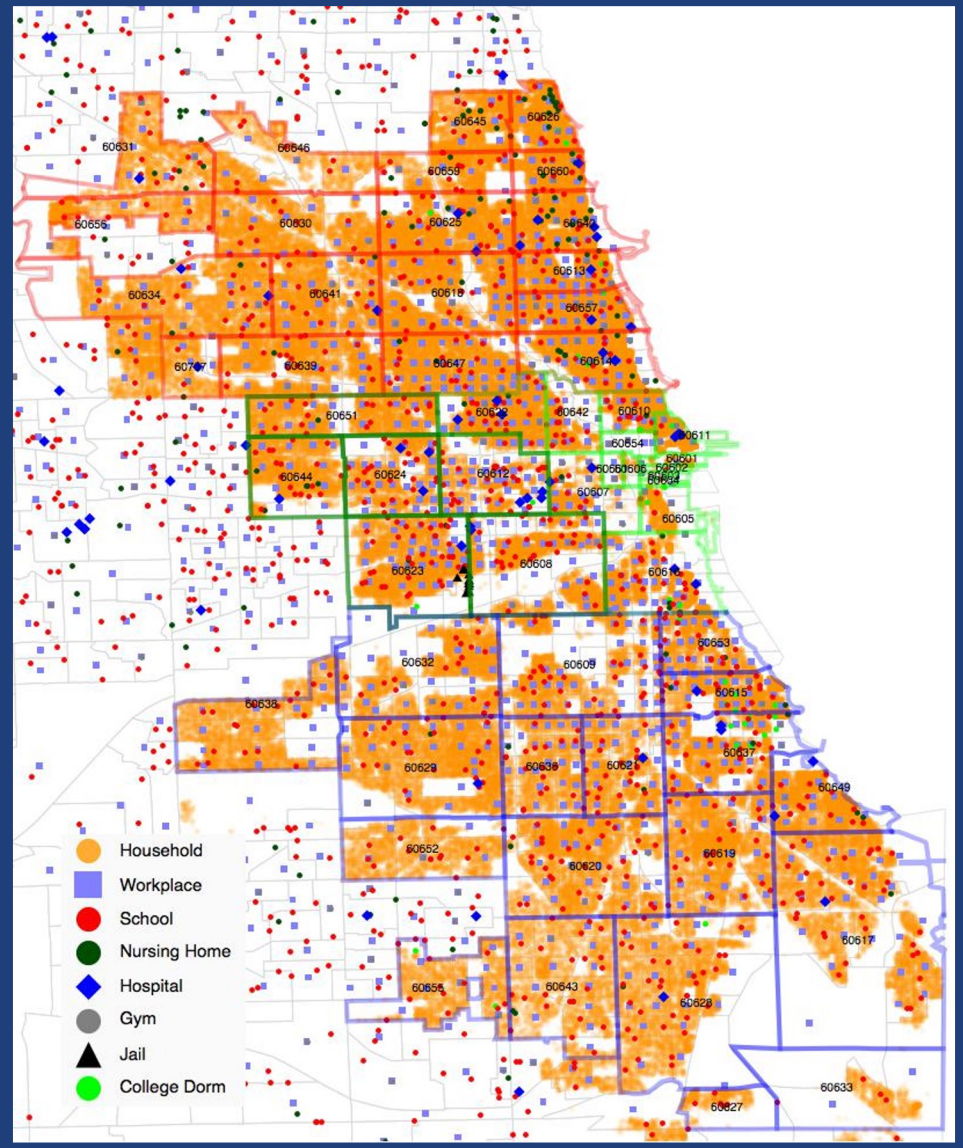
Managing flow execution

- **Provide inputs and run an instance of the flow**
  - Dynamic forms for manually running a flow
  - Triggers for starting unattended flow run
- **Check run progress/ status, cancel run**
- **Set access policy for monitoring, managing**



CityCOVID

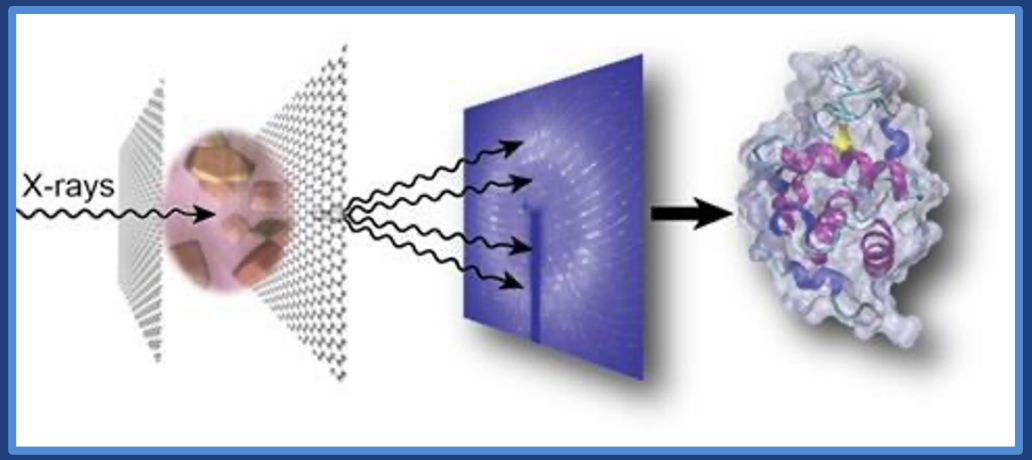
- Integrated COVID-19 pandemic monitoring, modeling, and analysis capability.
- CityCOVID is a city-scale agent-based model
- Automate flow
  - Scrape daily Chicago reports
  - Perform simulations at ALCF
  - Postprocess data at LCRC

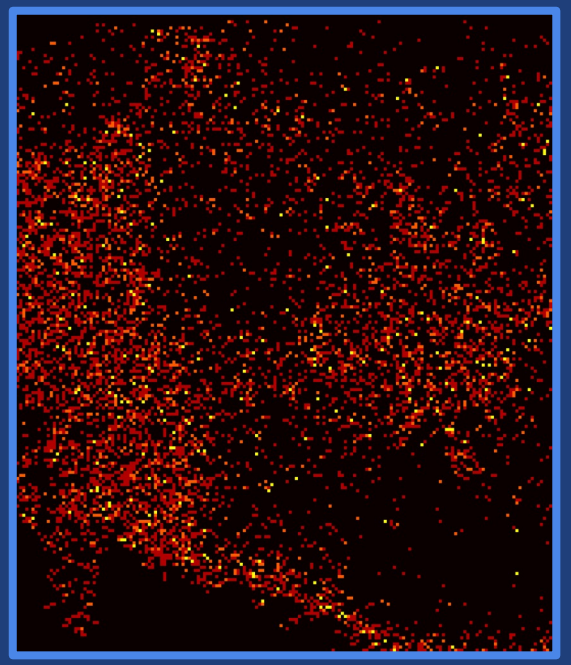


Jonathan Ozik, Nick Collier, and Charles Macal

Enabling serial crystallography at scale

- Serially image chips with thousands of embedded crystals
- Quality control first 1,000 to report failures
- Analyze batches of images as they are collected
- Report statistics and images during experiment
- Return crystal structure to scientist





Darren Sherrell, Gyorgy Babnigg, Andrzej Joachimiak

Automating cryoEM flows

Globus Flows

Auth

Get credentials

Transfer

Transfer raw files

funcX

Launch analysis job

Carbon!

Correct, classify, …

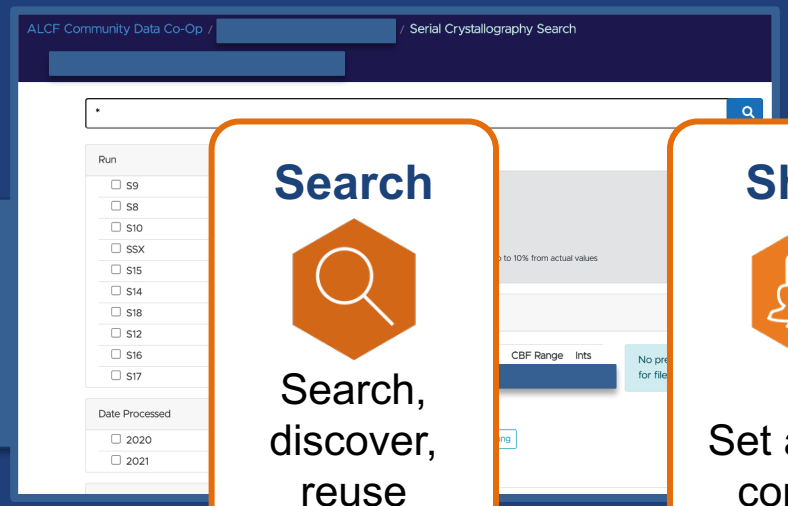
cryoSPARC

RELION

cryoDRGN

funcX

Extract metadata



Search

Search, discover, reuse

Share

Set access controls

Transfer

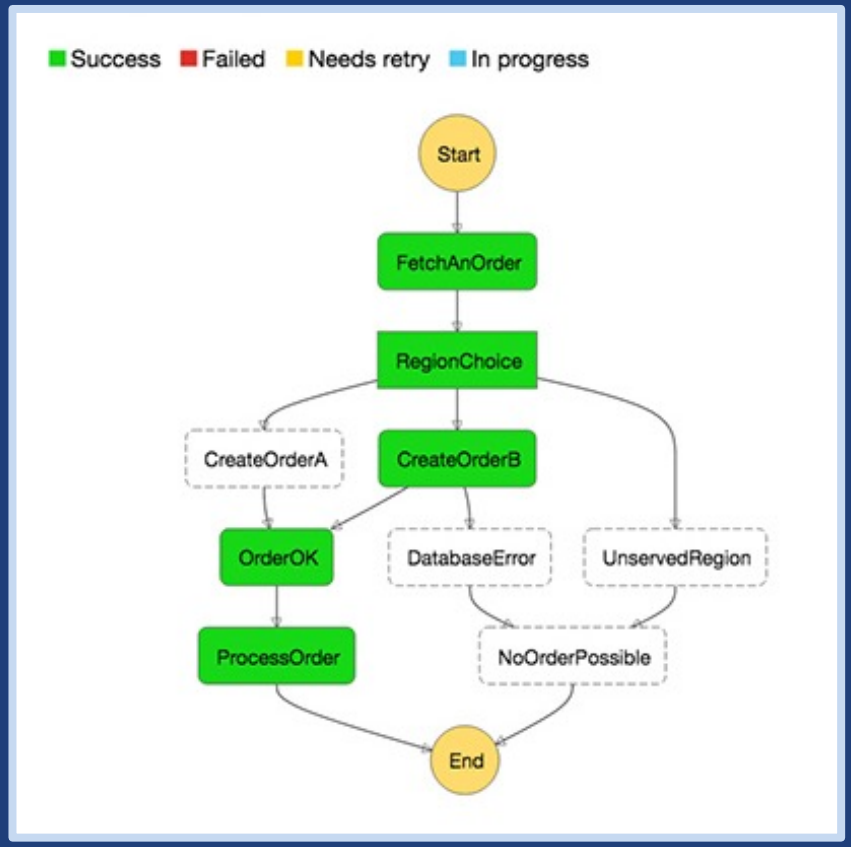
Move final files to repo

Search

Index ingest

Globus Flows service implementation

- **Built on AWS Step Functions**
  - Simple state machine language
  - Conditions, loops, fault tolerance, etc.
  - Propagates state through the flow
- **Standardized API for integrating custom event and action services**
  - Actions: synchronous or asynchronous
  - Custom Web forms prompt for user input
- **Actions secured with Globus Auth**



Developing Globus Flows

jupyter.demo.globus.org

Globus-provided flows

Two Stage Globus Transfer

kurt@globus.org

This flow requires at least one collection to be managed under a Globus subscription. The flow will perform a data transfer between source and destination collections in two stages. The first stage transfers from the source collection to an intermediate collection, and the second stage transfers from the intermediate collection to the destination collection. Data used in this flow are deleted from the intermediate collection after the final transfer is complete. Transferring data through an intermediate location can enable or improve performance in some firewalled or other network configurations.

Start

STEPS	CREATED	LAST MODIFIED	KEYWORDS
25	2022-03-30 11:24	2022-03-30 11:24	Two Stage,Two Hop,Intermediate,Globus Transfer,Transfer,Globus Production,Production

Move (copy and delete) files using Globus

This flow requires at least one collection to be managed under a Globus subscription. Following the transfer operation, data in the source collection will be deleted if the transfer to the destination collection is successful.

Start

STEPS	CREATED	LAST MODIFIED	KEYWORDS
23	2021-10-21 13:53	2022-03-30 11:20	Move,Data Move,Globus Transfer,Transfer,Globus Production,Production

Flow 1: Transfer and share data

- **Go to jupyter.demo.globus.org**
- **Open “Automation Using Globus Flows”**
- **Run sections A and B to define and deploy the flow**
- **Click on the link in the notebook to view the flow in the Globus web app**

Running Globus Flows

Running Globus flows

- **Run/manage flow via the Globus web app**
- **API to start and manage runs**
- **Globus Automate CLI and SDK: globus-automate-client.readthedocs.io**
- **Event driven execution of flows: Triggers**
  - e.g., when a file of specific type is created
  - e.g., every 2 hours

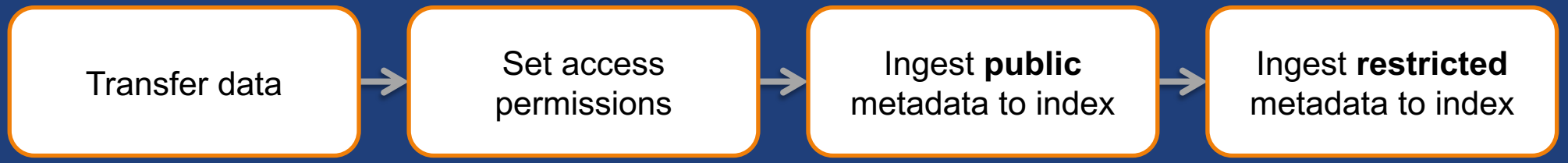
Trigger: Run flow when file is created

- **SSH into your tutorial instance**
- **Set up Globus Connect Personal (if not done)**
- **Edit `trigger_transfer_share_flow.py`**
  - Set it to run flow created using notebook
- **Run the trigger script**
- **Create files; monitor runs in the web app**

bit.ly/**gw-tut**

End-to-end instrument data management

- **Trigger**
  - Watch for file of specific type
  - Start a flow with folder and metadata about folder
- **Flow**



Making Data Findable with Globus Search

Globys Search: Data description and discovery

- **Metadata store with fine-grained visibility controls**
  - Distinct access control for metatdata
- **Schema agnostic → dynamic schemas**
- **Simple search using URL query parameters**
- **Complex search using search request document**

docs.globus.org/api/search

Flow 2: Transfer and publish data

- **Notebook at jupyter.demo.globus.org**
- **Open “Automation Using Flows with Search”**
- **Define and deploy flow using notebook (Section A and B)**

Trigger: Run flow when file is created

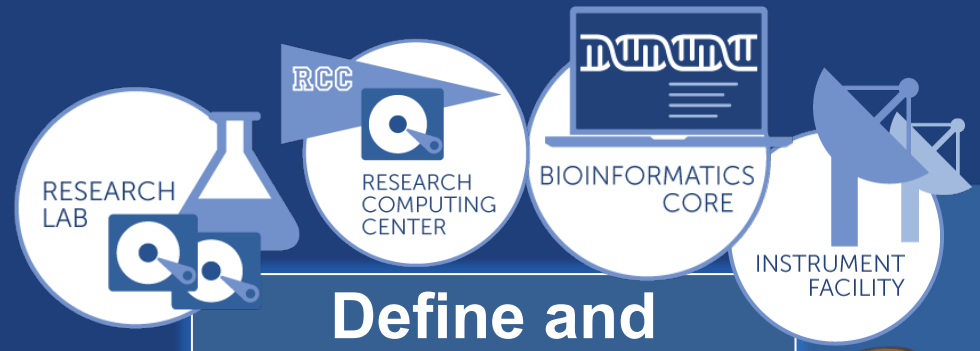
- **SSH into your tutorial instance**
- **Edit `trigger_transfer_publish_flow.py`**
  - Set it to run the flow you just deployed
- **Run the trigger script**
- **Create files; monitor runs in the web app**

bit.ly/**gw-tut**

Optional: Search your index

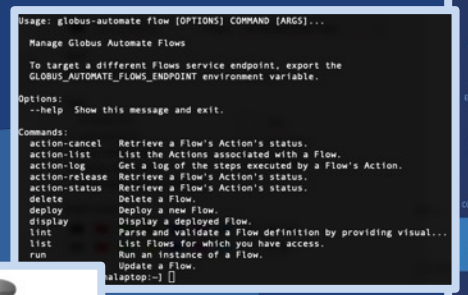
- **Notebook at jupyter.demo.globus.org**
- **Open “Metadata Search and Discovery”**
- **Run Section 1 and the first block in Section 5 (to set your search index)**
- **Run Section 6 to search your metadata**

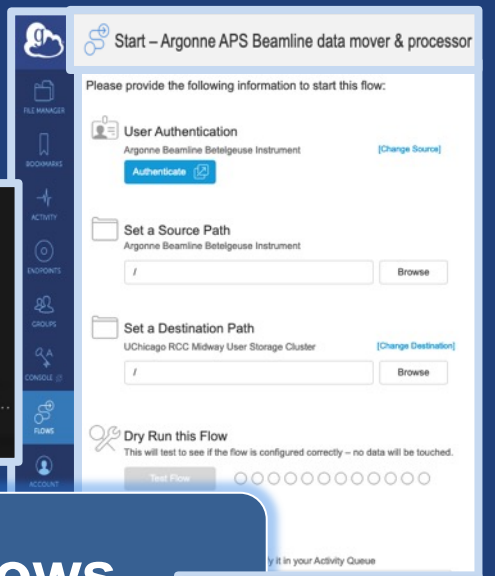
Automation services ecosystem



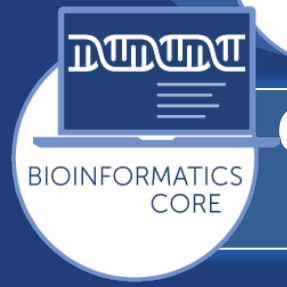
Define and deploy flows

```
{ "StartAt": "ToProject",
  "States" : {
    "ToProject" : { … },
    "SetPermission" : { …},
    "ProcessData" : { … } … }}
```





Run flows

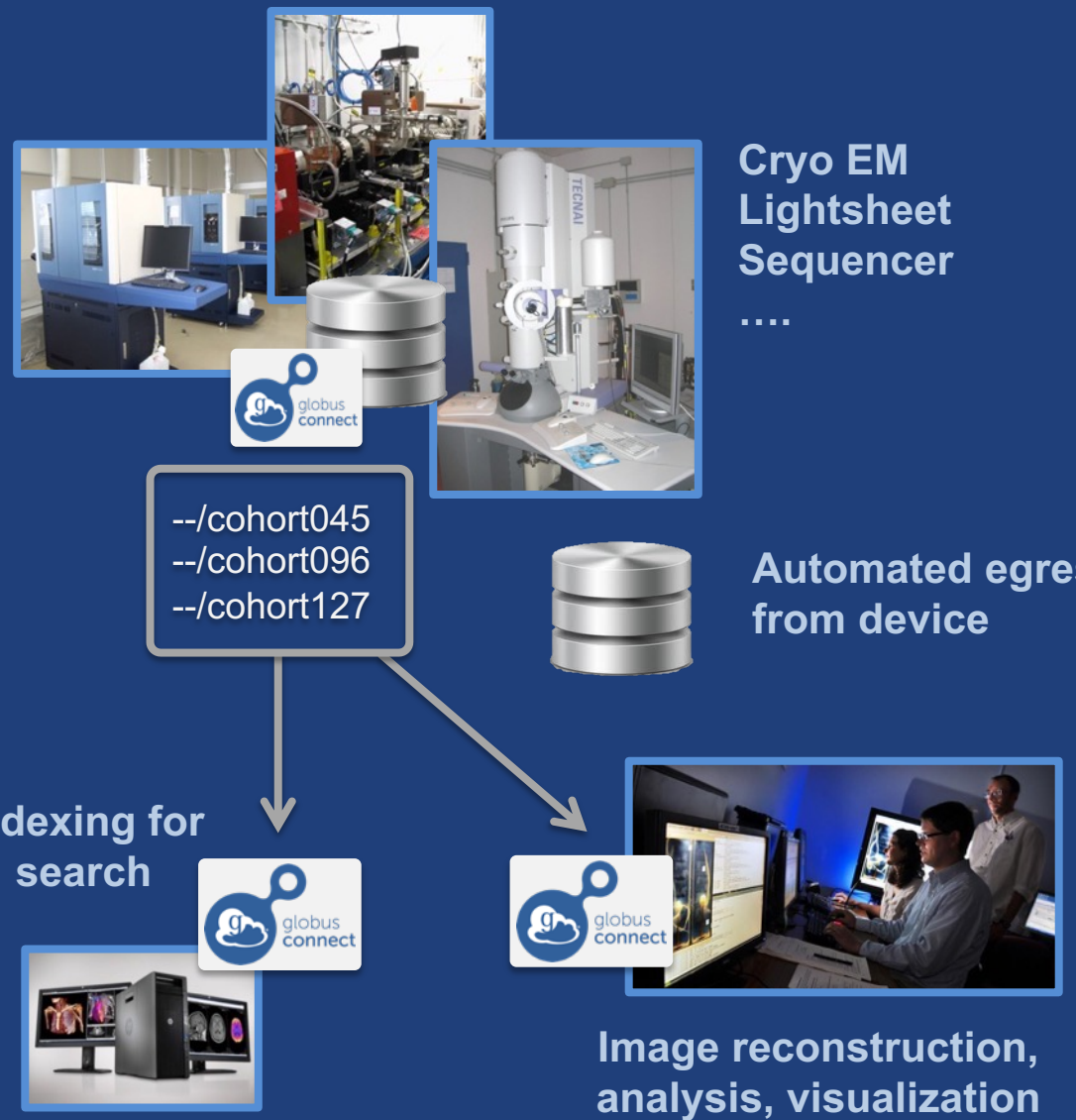


Create Action Providers

```
GET /provider_url/
POST /provider_url/run
GET /provider_url/action_id/status
GET /provider_url/action_id/cancel
GET /provider_url/action_id/status
```

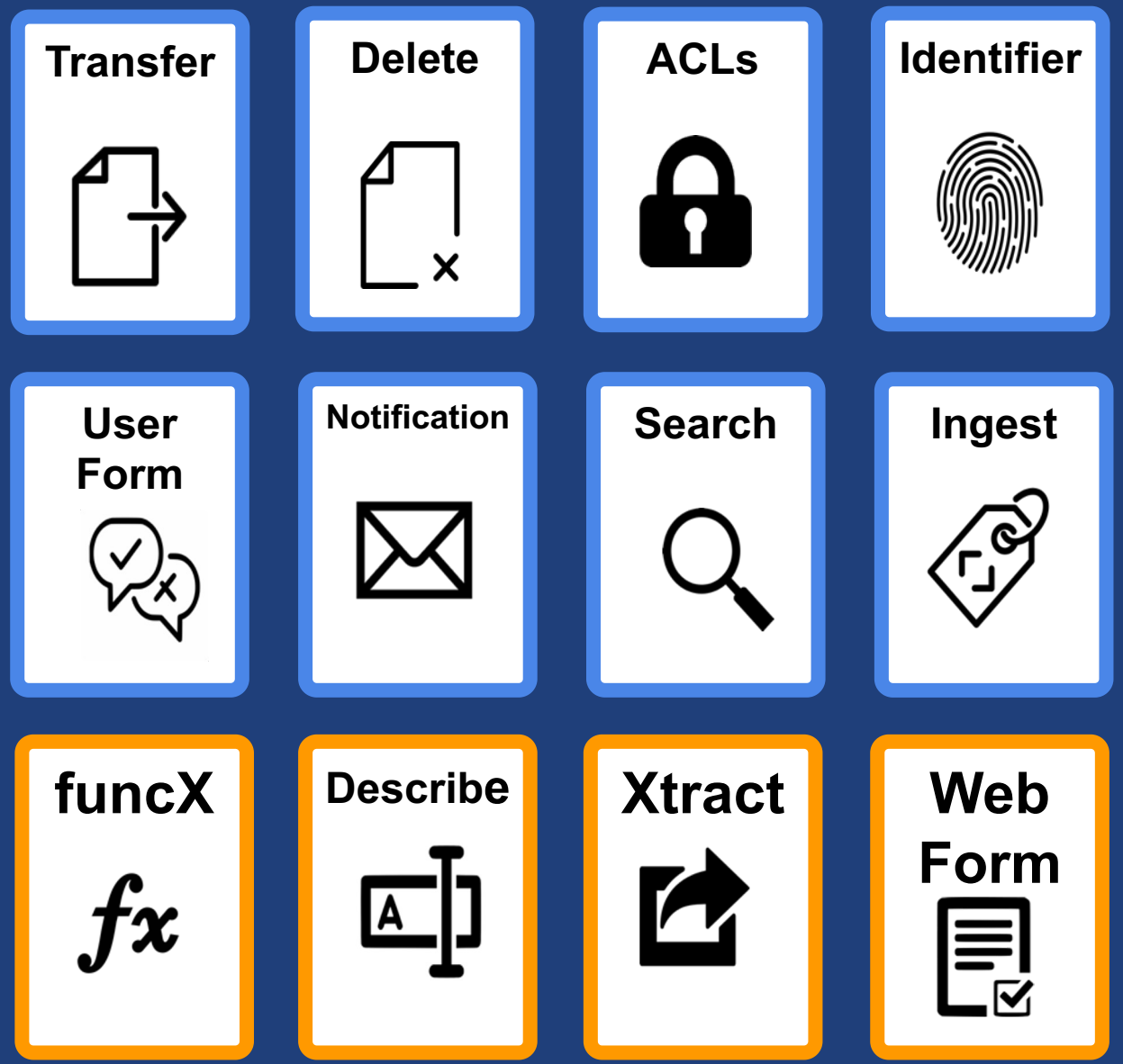
SaaS: instrument data management

- **Configurable…**
  - Set transfer triggers
  - Select destination(s)
  - Define mestadata
- **Extensible…**
  - Add/remove actions
  - Change action providers
- **No development required**



Extending the ecosystem: Action providers

- **Action Provider is a service endpoint**
  - Run
  - Status
  - Cancel
  - Release
  - Resume
- **Action Provider Toolkit**
  action-provider-tools.readthedocs.io/en/latest



Globus Provided | **Custom built**

Automating computation with ***funcX***

- **Managed, federated Functions-as-a-Service**
- **Cloud-hosted service for managing compute**
- **Register and share *compute* endpoints**
- **Register and share Python functions**
- **Reliably, scalable and securely execute functions on remote endpoints**
- **Integrated with Globus Auth and data ecosystem**

THE UNIVERSITY OF CHICAGO ILLINOIS Argonne NATIONAL LABORATORY

Support resources

- Globus documentation: docs.globus.org
- YouTube channel: youtube.com/GlobusOnline
- Helpdesk: support@globus.org
- Mailing Lists: globus.org/mailing-lists
- Customer engagement team (office hours)
- Professional services team (advisory, custom work)