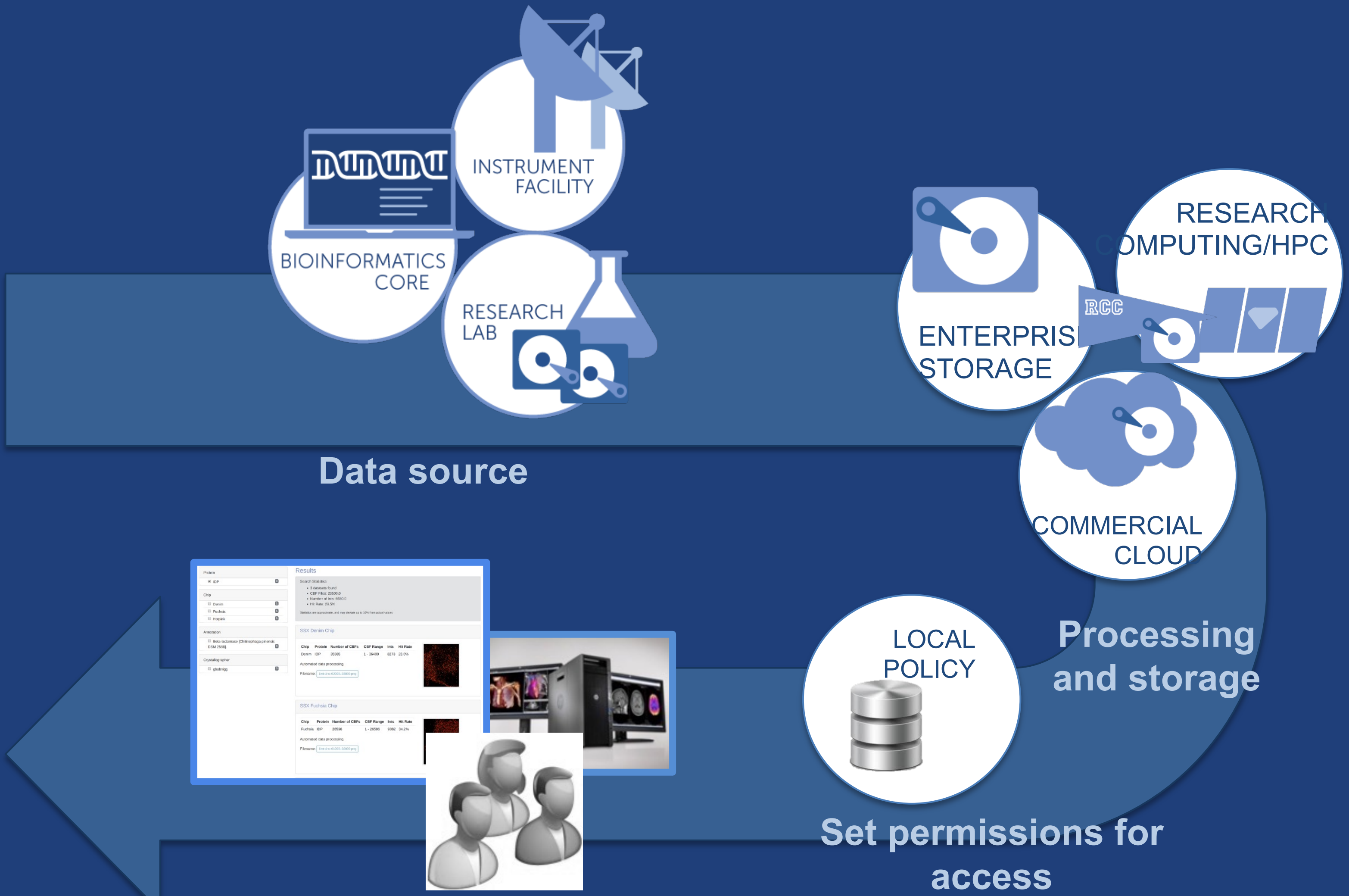
Automating Instrument Data at Scale

Rachana Ananthakrishnan
Vas Vasiliadis

Data processing pipeline pattern



Data source

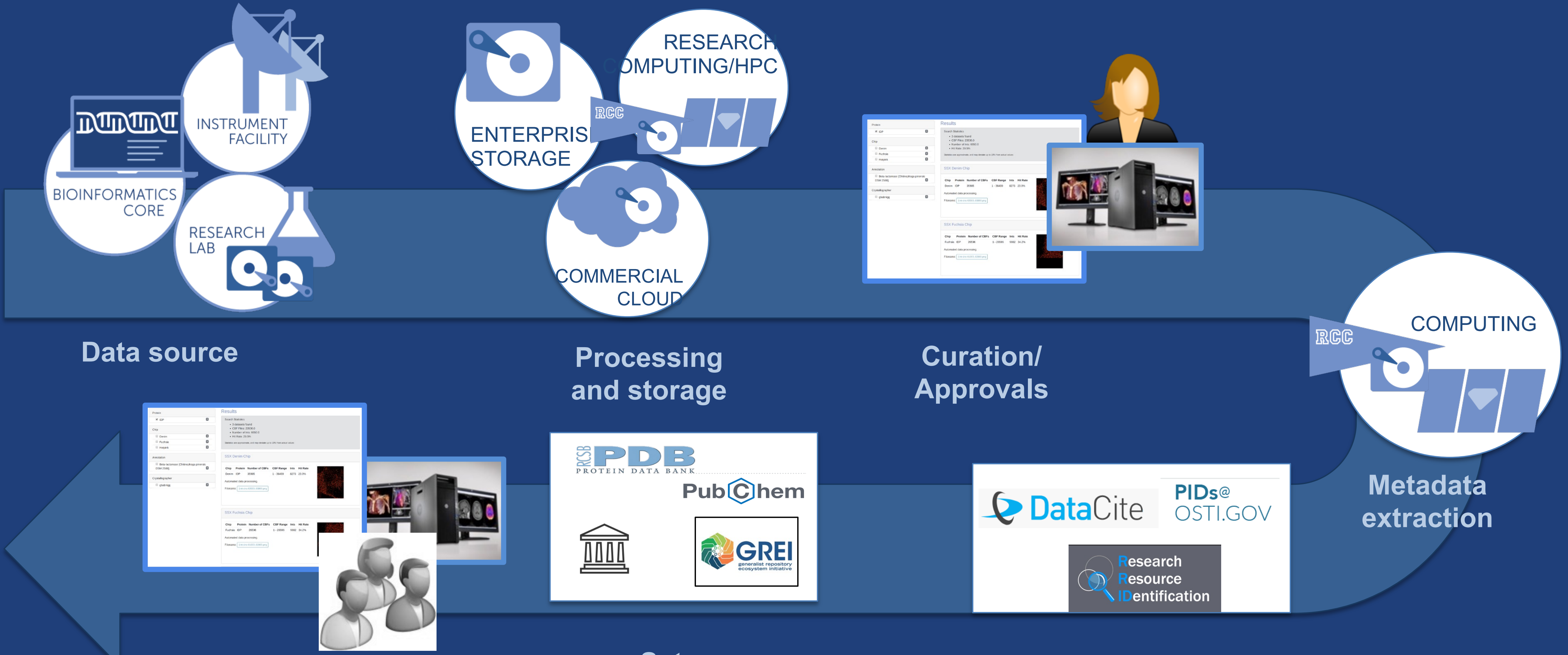
Processing and storage

Set permissions for access

Access by collaborators

At the minimum, make data available to collaborators

Data processing pipeline pattern



Data source

Processing and storage

Curation/ Approvals

Metadata extraction

Persistent Identifiers

Set access and publish for discovery

Access by collaborators

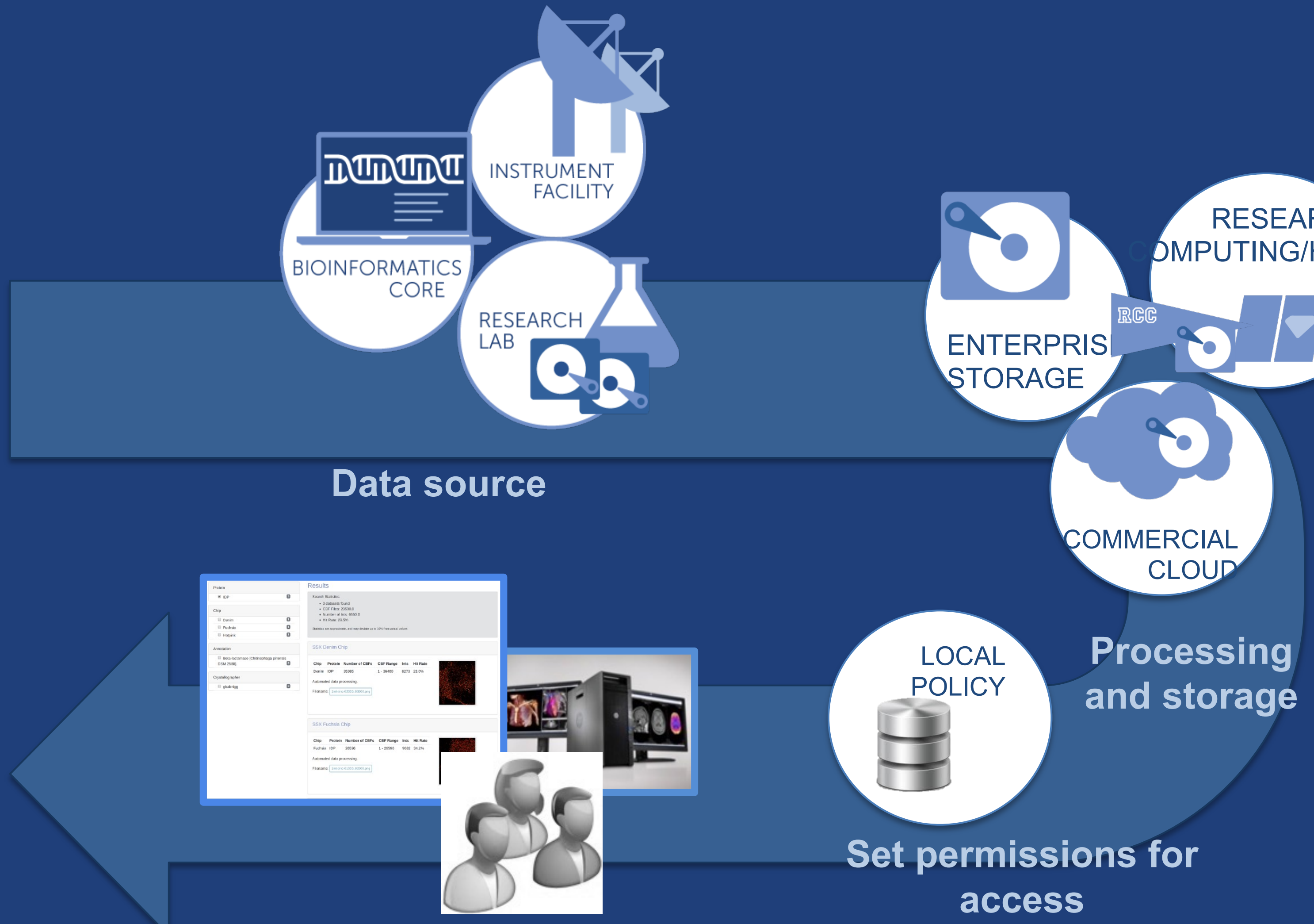
Requirements for such pipeline

- **Reliable, near-real time data access**
- **Uniformed policy for data access, based on local policy**
- **Delegation of data access management to PI**
- **Ability to compute on data across storage classes**
- **Apply best practices with data processing pipeline**
- **Support for data organization to facilitate FAIR data**
- **…**

Automation to scale

How do we use the Globus platform for full automation?

Data processing pipeline pattern



Data source

Processing and storage

Set permissions for access

Access by collaborators

1. Preparing acquisition machine for data access
2. Configuring data collections for automated data transfer
3. Credentials for automation
4. Create and deploy flow for automation
5. Create trigger scripts

1. Preparing acquisition machine for data access

Install Globus Connect

- **For Linux machines, install and use Globus Connect Server**
- **Acquisition machines often are Windows machine → Globus Connect Personal deployment**
  - Installed as a local user account
  - PI login vs GCP always running on a shared account
  - Outbound connections only

2. Configuring data collections for automated data transfer

Using guest collections at instrument facilities

- **Create guest collections on all storage systems for automation**
  - Acquisition machine, storage mounted on compute etc.
- **Creation of guest collection cannot generally be automated ***
- **Permission and role management on guest collections can be fully automated**

Pattern 1: Guest collection use

- **Create a guest collection at top level directory**
  - This is done by a user who has local account
- **For each experiment/project/modality**
  - Create a folder
  - Set permissions PI/collaborators to read data from the folder
- **Can automate permission management by using local policy store**

Pattern 2: Guest collection use

- **Create a guest collection for each experiment/project/modality**
  - Grant PI role to manage access to the guest collection (Access Manager role)
  - Set permissions collaborators to read data from the folder
- **Can automate role and permission management by using local policy store**

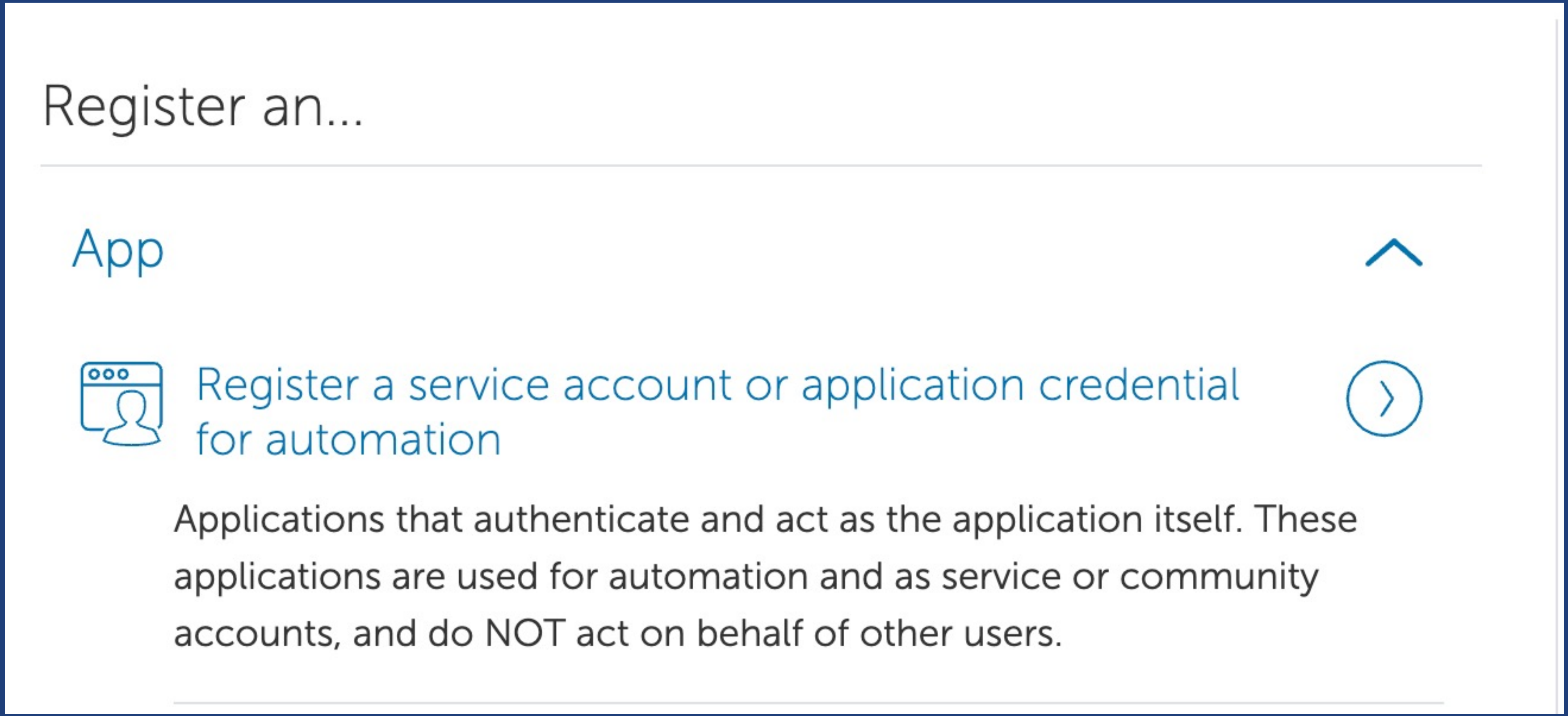
3. Credentials for automation

Managing service accounts/app credentials

- **Application Identity: appclientid@clients.auth.globus.org**
- **These are confidential apps with client id and secret**
- **Ensure application is on a secure device**
- **Set up policy for rotation of secret**
- **Assign project admins to manage the registration**

Registering a service account

- **Webapp - Settings**
  - app.globus.org/settings/developers



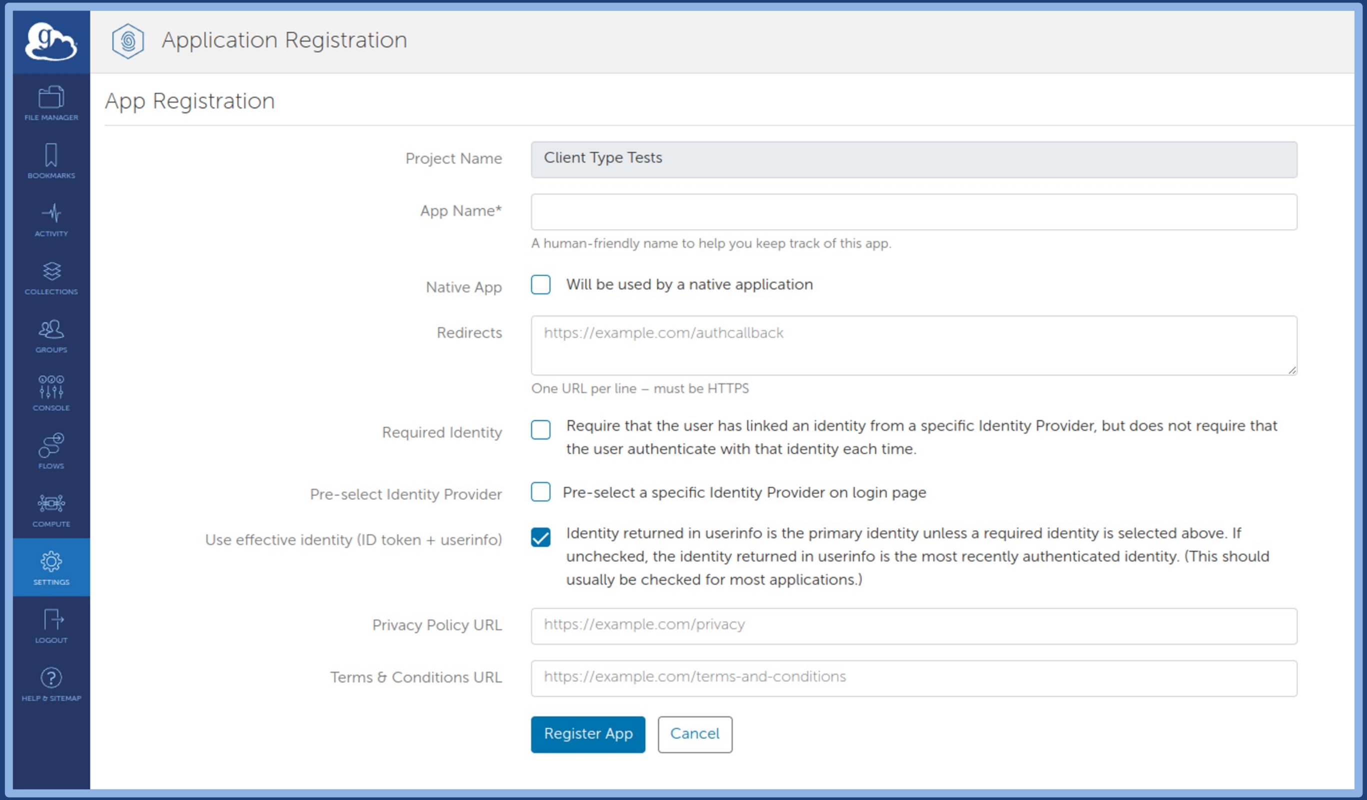
Register an...

App

Register a service account or application credential for automation

Applications that authenticate and act as the application itself. These applications are used for automation and as service or community accounts, and do NOT act on behalf of other users.

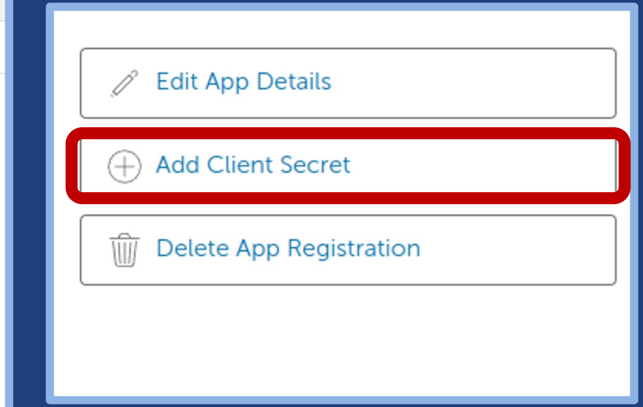
Get app credentials at **app.globus.org/settings/developers**

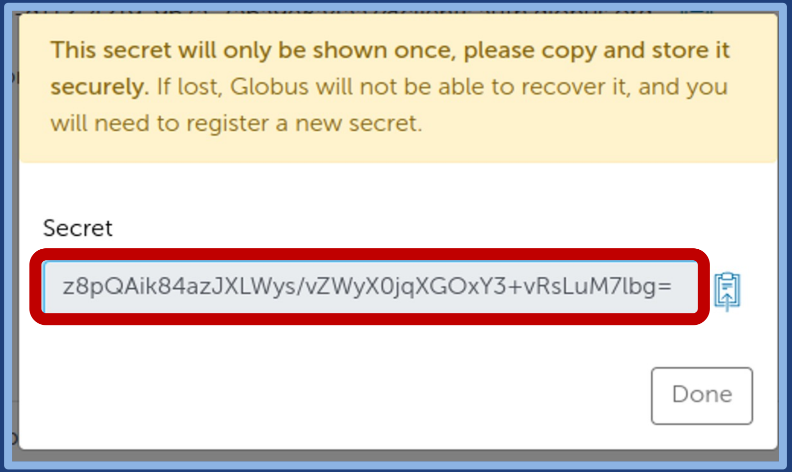


Application Registration

App Registration

Project Name	Client Type Tests
App Name*	 A human-friendly name to help you keep track of this app.
Native App	□ Will be used by a native application
Redirects	https://example.com/authcallback One URL per line – must be HTTPS
Required Identity	□ Require that the user has linked an identity from a specific Identity Provider, but does not require that the user authenticate with that identity each time.
Pre-select Identity Provider	□ Pre-select a specific Identity Provider on login page
Use effective identity (ID token + userinfo)	☑ Identity returned in userinfo is the primary identity unless a required identity is selected above. If unchecked, the identity returned in userinfo is the most recently authenticated identity. (This should usually be checked for most applications.)
Privacy Policy URL	https://example.com/privacy
Terms & Conditions URL	https://example.com/terms-and-conditions



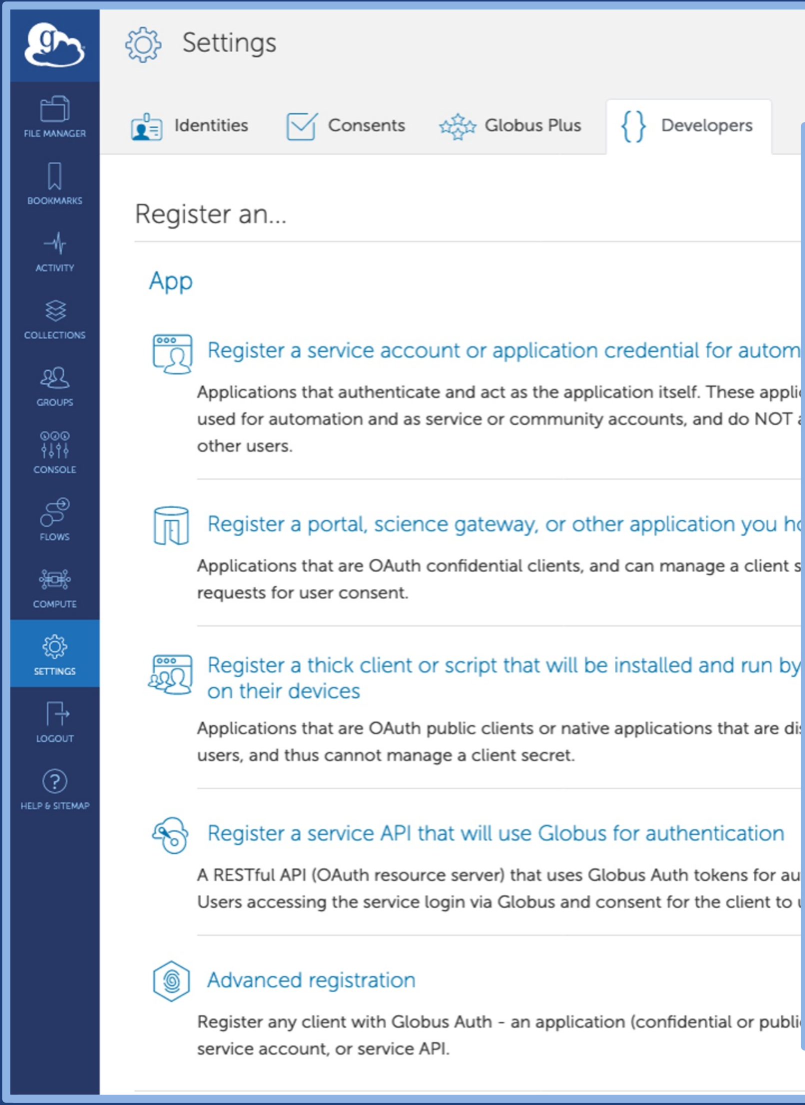


This secret will only be shown once, please copy and store it securely. If lost, Globus will not be able to recover it, and you will need to register a new secret.

Secret

z8pQAik84azJXLWys/vZWyX0jqXGOxY3+vRsLuM7lbg=

Get app credentials at **app.globus.org/settings/developers**



Settings

Register an...

App

Register a service account or application credential for autom

Applications that authenticate and act as the application itself. These appli used for automation and as service or community accounts, and do NOT other users.

Register a portal, science gateway, or other application you h

Applications that are OAuth confidential clients, and can manage a client s requests for user consent.

Register a thick client or script that will be installed and run by on their devices

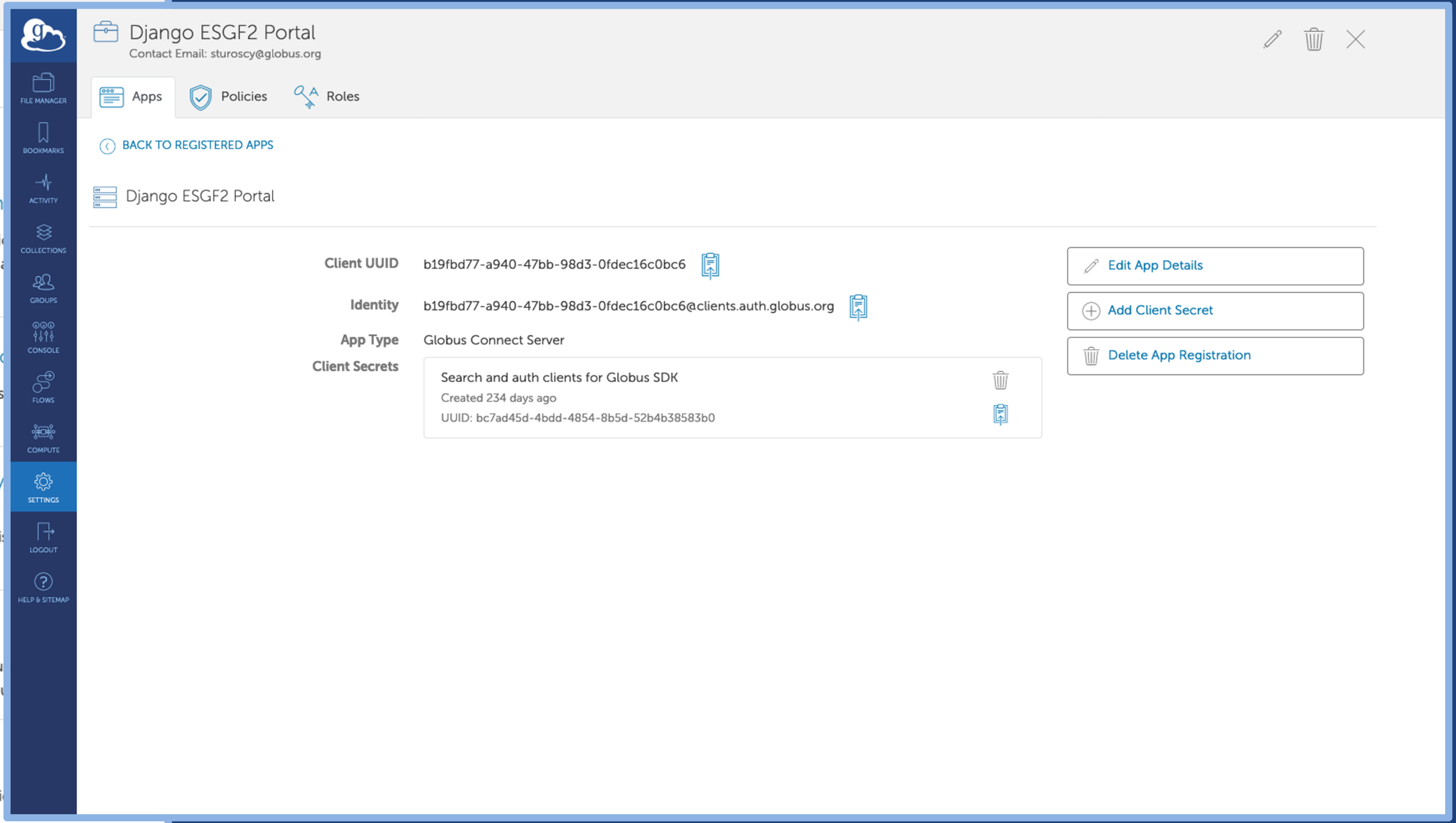
Applications that are OAuth public clients or native applications that are di users, and thus cannot manage a client secret.

Register a service API that will use Globus for authentication

A RESTful API (OAuth resource server) that uses Globus Auth tokens for au Users accessing the service login via Globus and consent for the client to

Advanced registration

Register any client with Globus Auth - an application (confidential or publi service account, or service API.



Django ESGF2 Portal

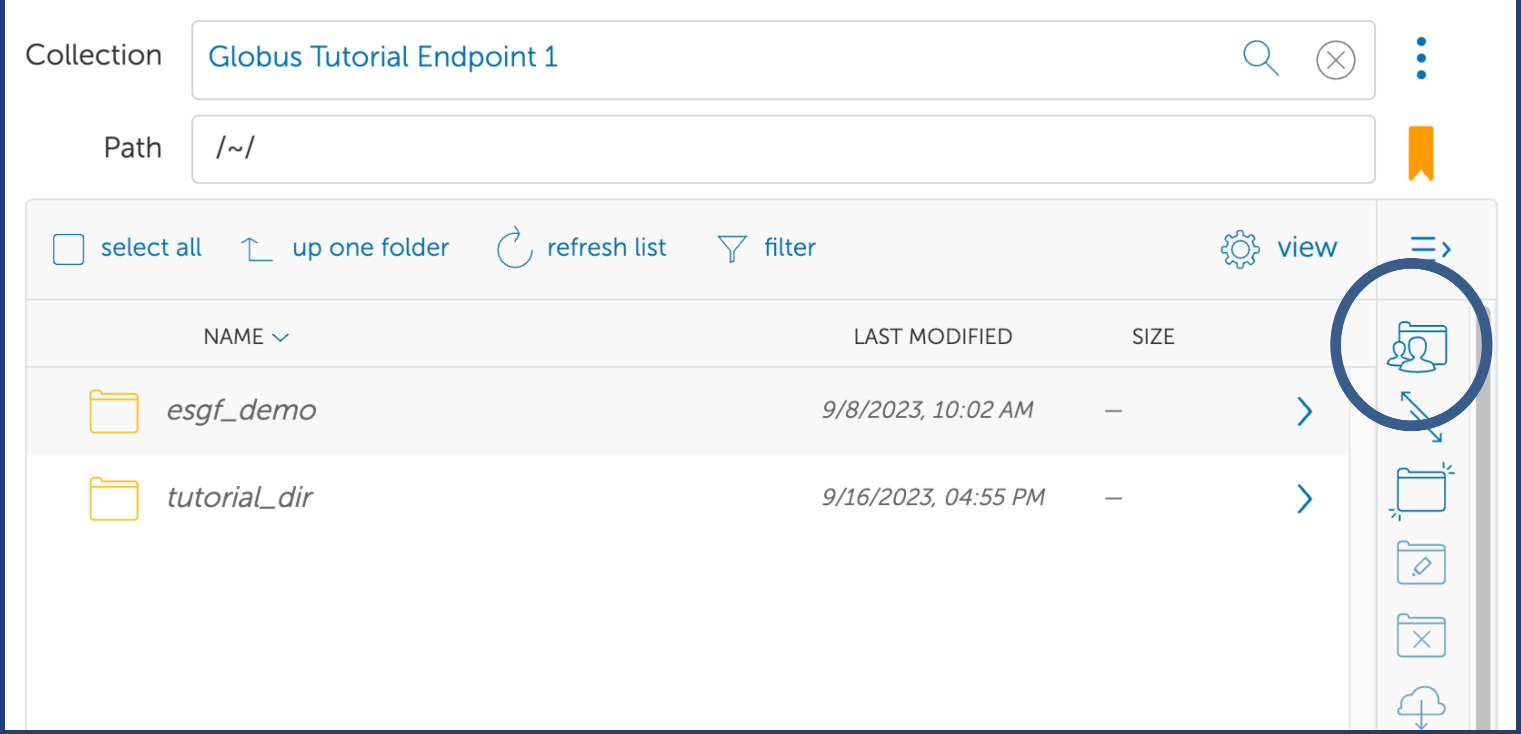
Contact Email: sturoscy@globus.org

Django ESGF2 Portal

Client UUID	b19fbd77-a940-47bb-98d3-0fdec16c0bc6
Identity	b19fbd77-a940-47bb-98d3-0fdec16c0bc6@clients.auth.globus.org
App Type	Globus Connect Server
Client Secrets	Search and auth clients for Globus SDK Created 234 days ago UUID: bc7ad45d-4bdd-4854-8b5d-52b4b38583b0

Set permission for the service account

- **Use the web app to create guest collection on guest collection you created**

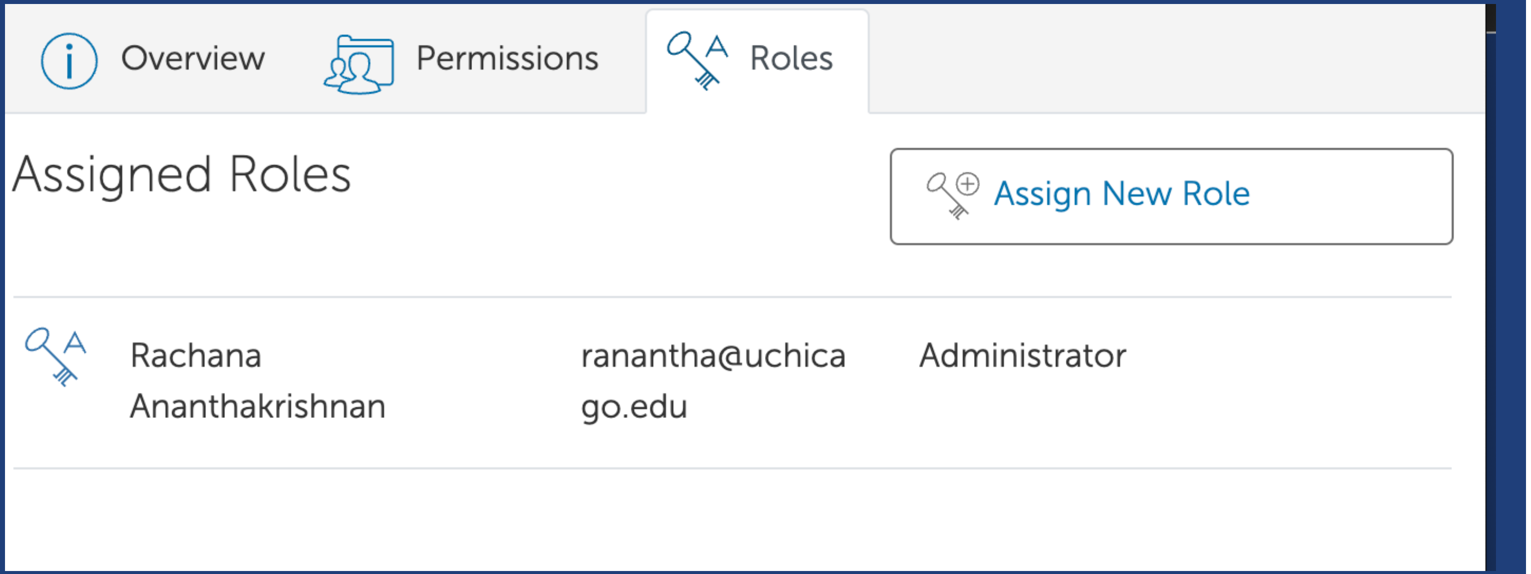


Collection: Globus Tutorial Endpoint 1

Path: /~/

NAME	LAST MODIFIED	SIZE
esgf_demo	9/8/2023, 10:02 AM	–
tutorial_dir	9/16/2023, 04:55 PM	–

Set permission for the service account

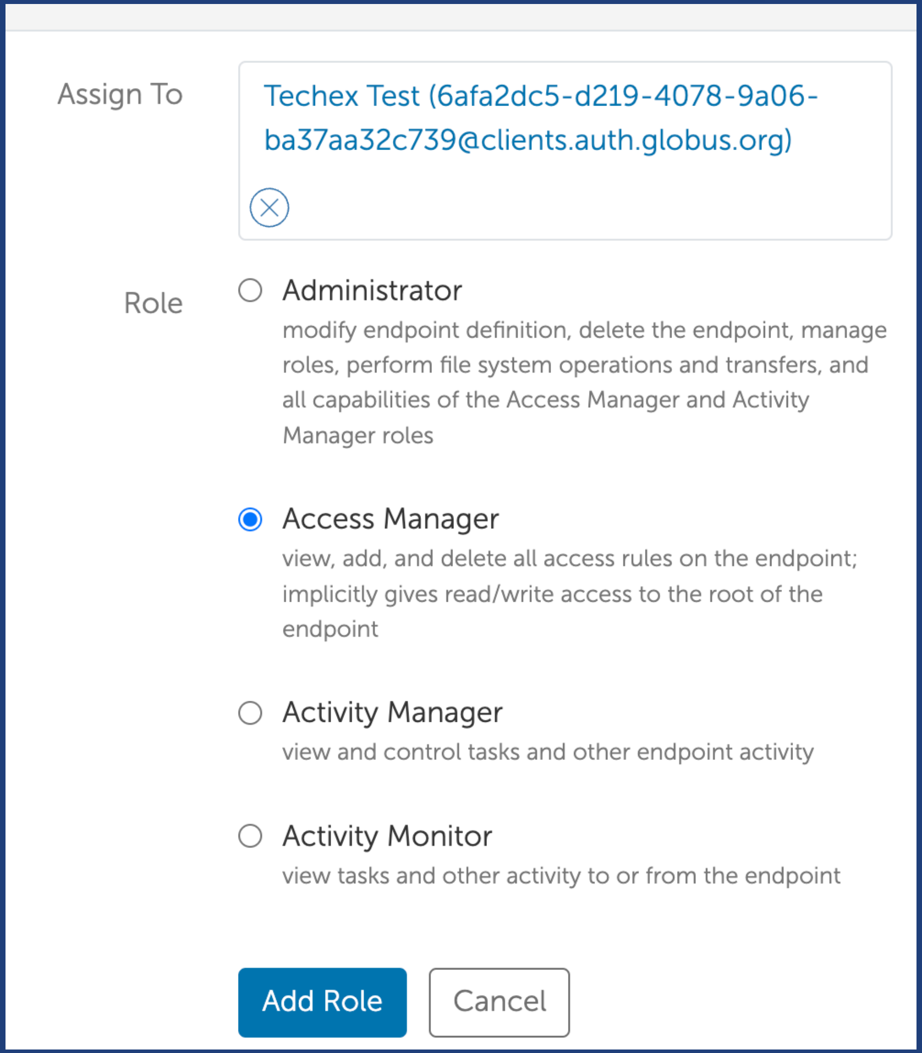


Overview | Permissions | Roles

Assigned Roles

Assign New Role

	Rachana Ananthakrishnan	ranantha@uchicago.edu	Administrator



Assign To: Techex Test (6afa2dc5-d219-4078-9a06-ba37aa32c739@clients.auth.globus.org)

Role:

○ Administrator
modify endpoint definition, delete the endpoint, manage roles, perform file system operations and transfers, and all capabilities of the Access Manager and Activity Manager roles

● Access Manager
view, add, and delete all access rules on the endpoint; implicitly gives read/write access to the root of the endpoint

○ Activity Manager
view and control tasks and other endpoint activity

○ Activity Monitor
view tasks and other activity to or from the endpoint

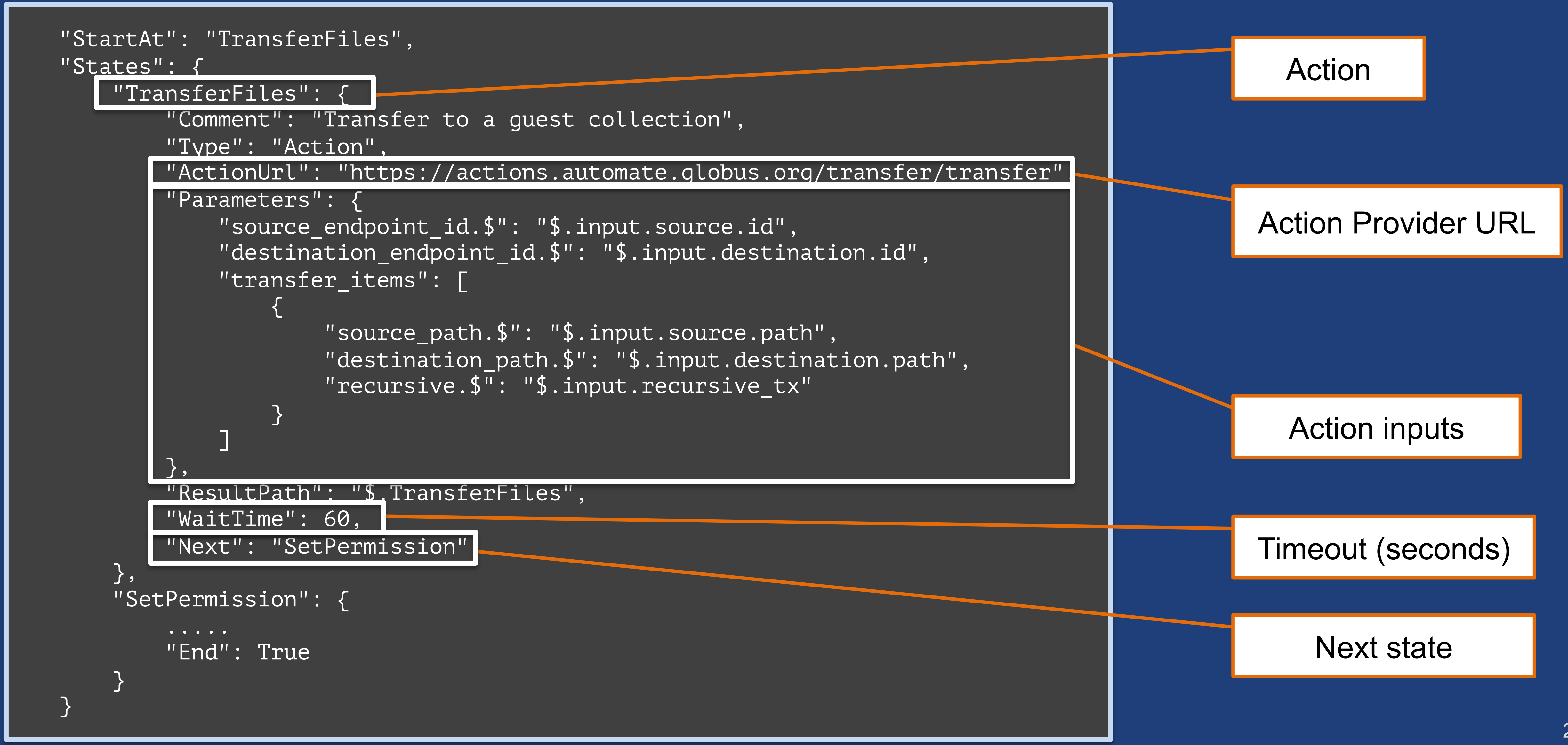
Add Role | Cancel

4. Create and deploy flow for automation

Create a flow for your use case

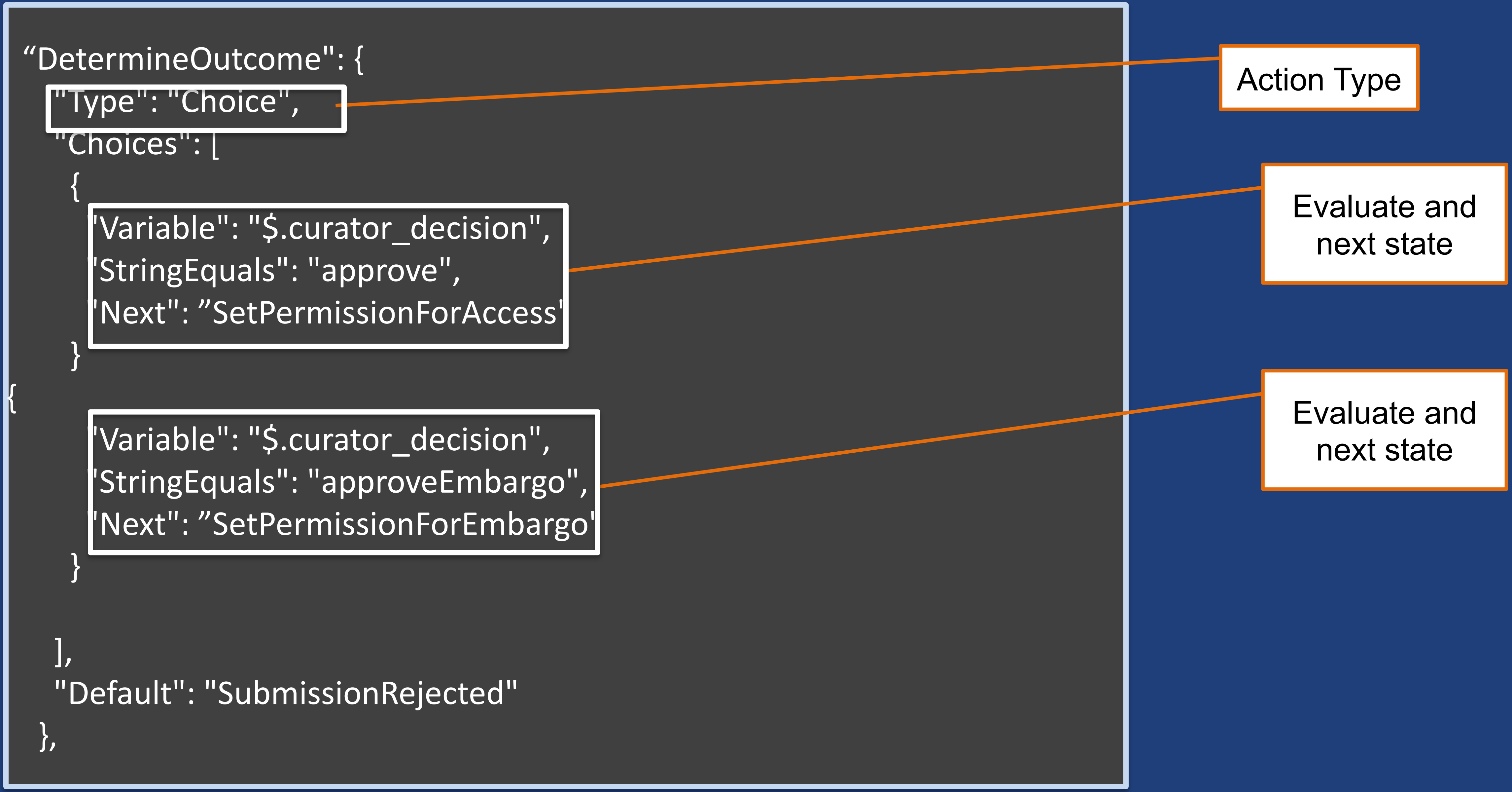
- **Start with flow definitions that are published:**
  - github.com/globus/globus-flows-trigger-examples
  - docs.globus.org/api/flows/authoring-flows/examples/
- **Manage flow definitions in a version controlled system**
- **Validation tools**
  - Flows IDE: https://globus.github.io/flows-ide/
  - Globus CLI: *globus flows validate*

Flow definition



```
"StartAt": "TransferFiles",
"States": {
    "TransferFiles": {
        "Comment": "Transfer to a guest collection",
        "Type": "Action",
        "ActionUrl": "https://actions.automate.globus.org/transfer/transfer"
        "Parameters": {
            "source_endpoint_id.$": "$.input.source.id",
            "destination_endpoint_id.$": "$.input.destination.id",
            "transfer_items": [
                {
                    "source_path.$": "$.input.source.path",
                    "destination_path.$": "$.input.destination.path",
                    "recursive.$": "$.input.recursive_tx"
                }
            ]
        },
        "ResultPath": "$.TransferFiles",
        "WaitTime": 60,
        "Next": "SetPermission"
    },
    "SetPermission": {
        .....
        "End": True
    }
}
```

Choice in flows



```
“DetermineOutcome": {
   "Type": "Choice",
   "Choices": [
     {
       'Variable": "$.curator_decision",
       'StringEquals": "approve",
       'Next": ”SetPermissionForAccess'
     }
{
       'Variable": "$.curator_decision",
       'StringEquals": "approveEmbargo",
       'Next": ”SetPermissionForEmbargo'
     }

   ],
   "Default": "SubmissionRejected"
  },
```

Action Type

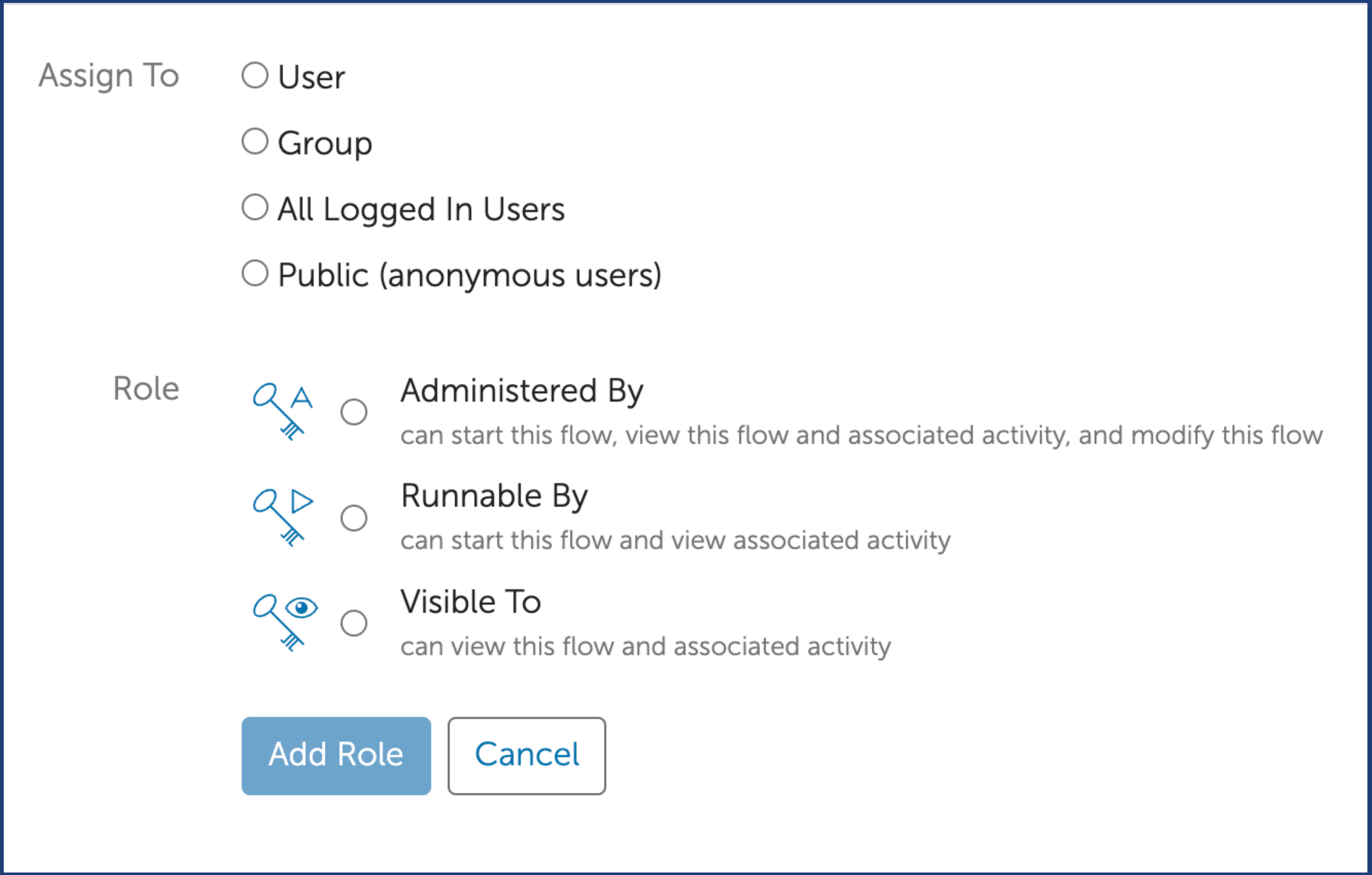
Evaluate and next state

Evaluate and next state

Run context

- **Flow and user properties as read-only values**
- **Available in $._context**
  - – Flow id
  - – Run id
  - – All identities of the user invoking the flow
  - – Email id of the user invoking the flow
  - – Token information (issued time, and expiration information)

Permissions to run the flow



Assign To

○ User

○ Group

○ All Logged In Users

○ Public (anonymous users)

Role

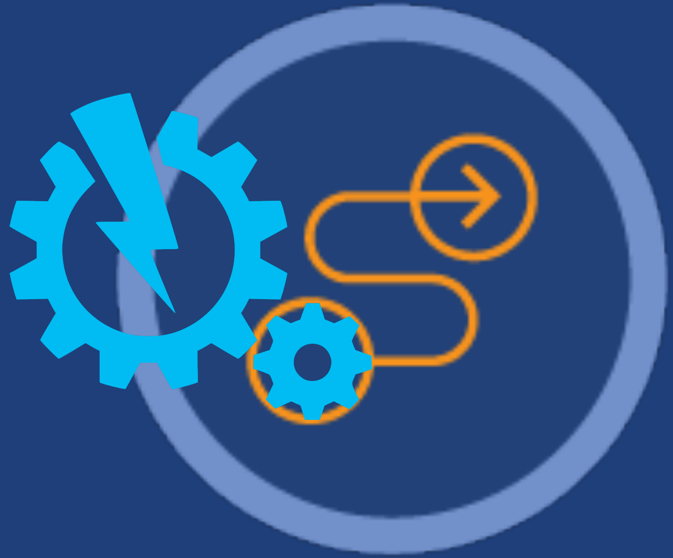
○ Administered By
can start this flow, view this flow and associated activity, and modify this flow

○ Runnable By
can start this flow and view associated activity

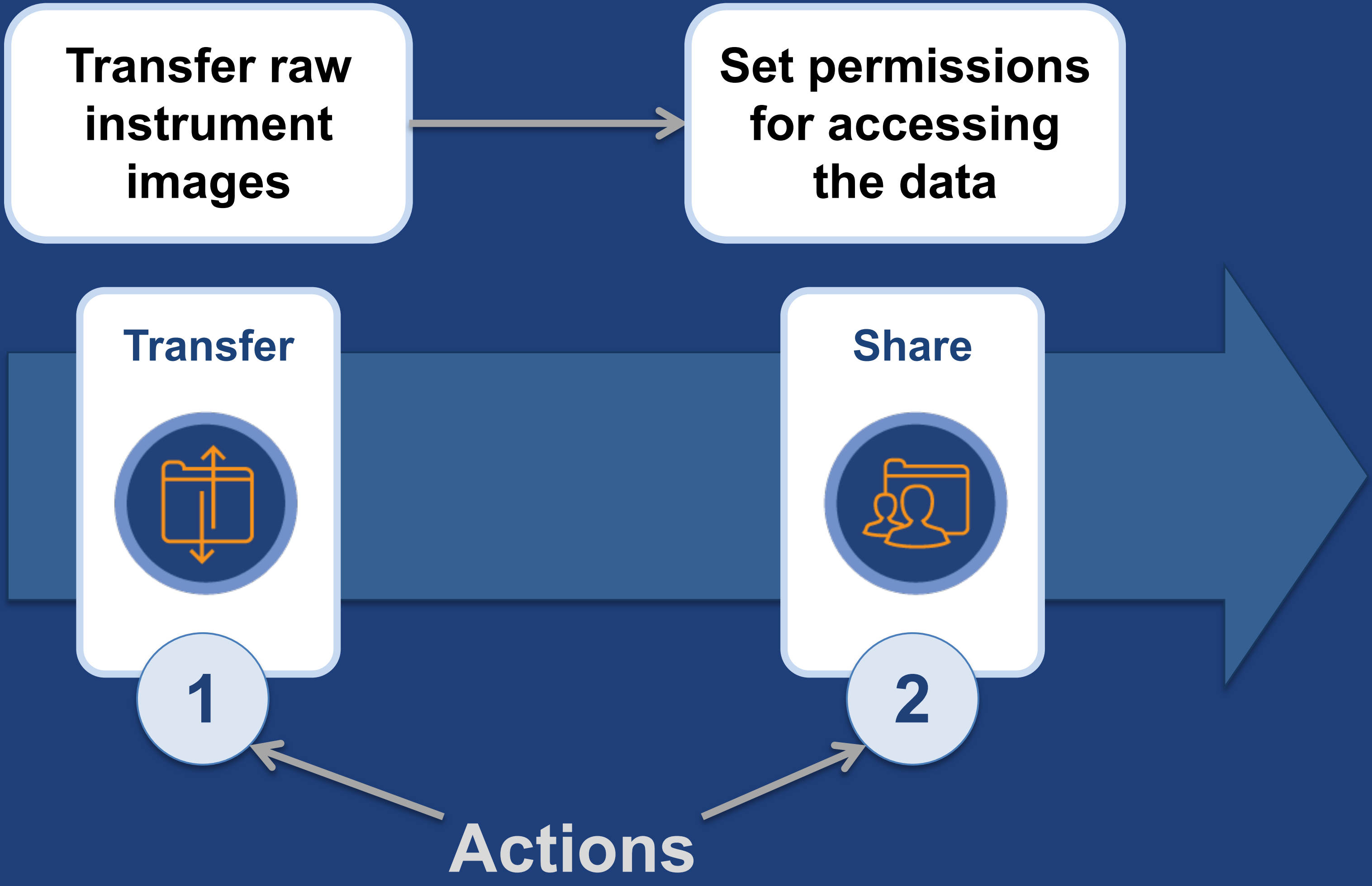
○ Visible To
can view this flow and associated activity

Set permission for the service account to run the flow

Triggering flows on instruments …and other resources



A simple, and very common, use case



Transfer raw instrument images

Set permissions for accessing the data

Actions

Creating and triggering runs of this flow

1. Log into the Globus CLI
2. Create (deploy) the transfer-and-share flow
3. Edit the monitor (trigger) script
4. Ensure GCP is running on the instrument
5. Run the monitor script
6. Trigger the flow

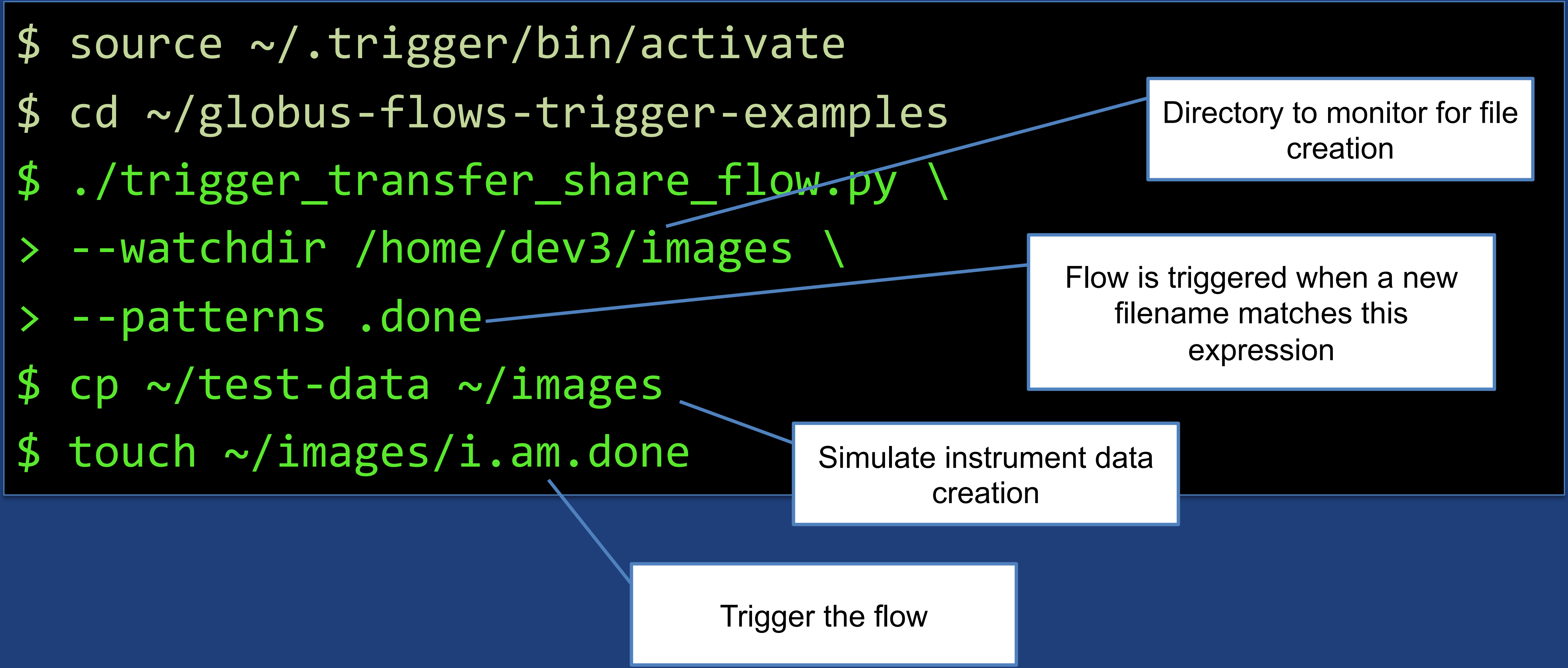
github.com/globus/**globus-flows-trigger-examples**

Create the flow

```
$ globus login
$ cd ~/globus-flows-trigger-examples/transfer_share
$ globus flows create FLOW_NAME \
> definition.json --input-schema schema.json
$ ~/globusconnectpersonal-3.2.5/globusconnectpersonal –start &
```

- **Success returns the flow ID**
- **Inspect the flow using the web app**

Run the monitor script and trigger the flow



```
$ source ~/.trigger/bin/activate
$ cd ~/globus-flows-trigger-examples
$ ./trigger_transfer_share_flow.py \
> --watchdir /home/dev3/images \
> --patterns .done
$ cp ~/test-data ~/images
$ touch ~/images/i.am.done
```

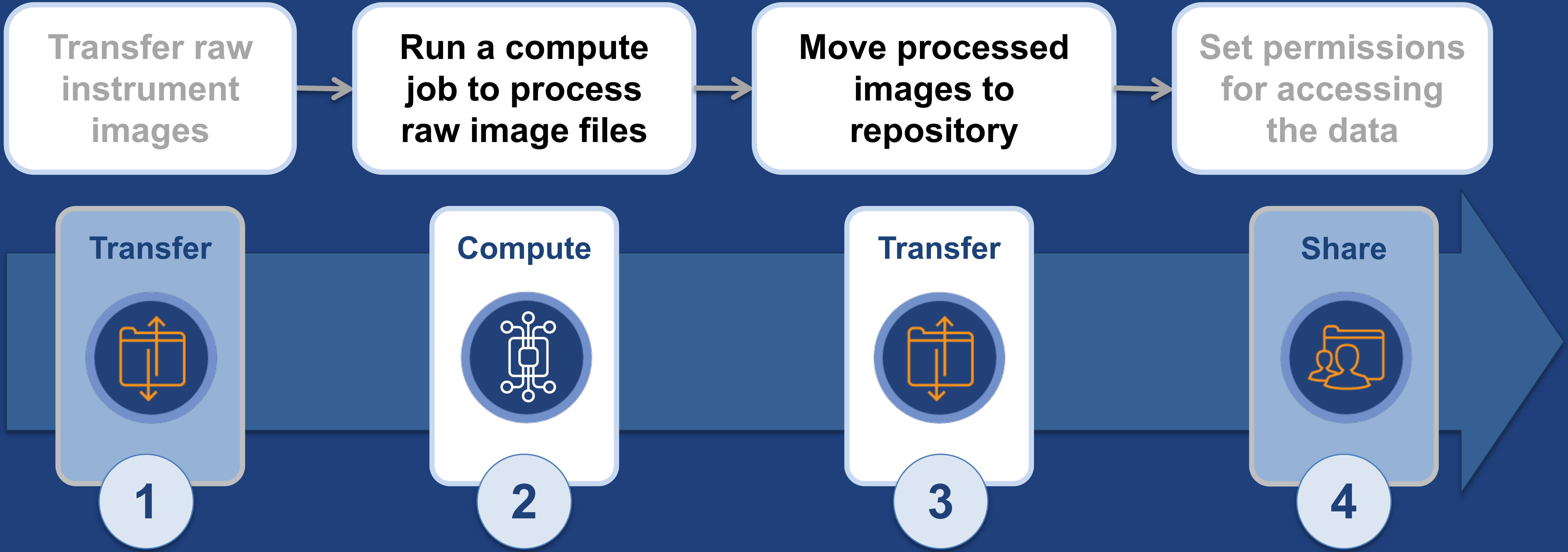
Directory to monitor for file creation

Flow is triggered when a new filename matches this expression

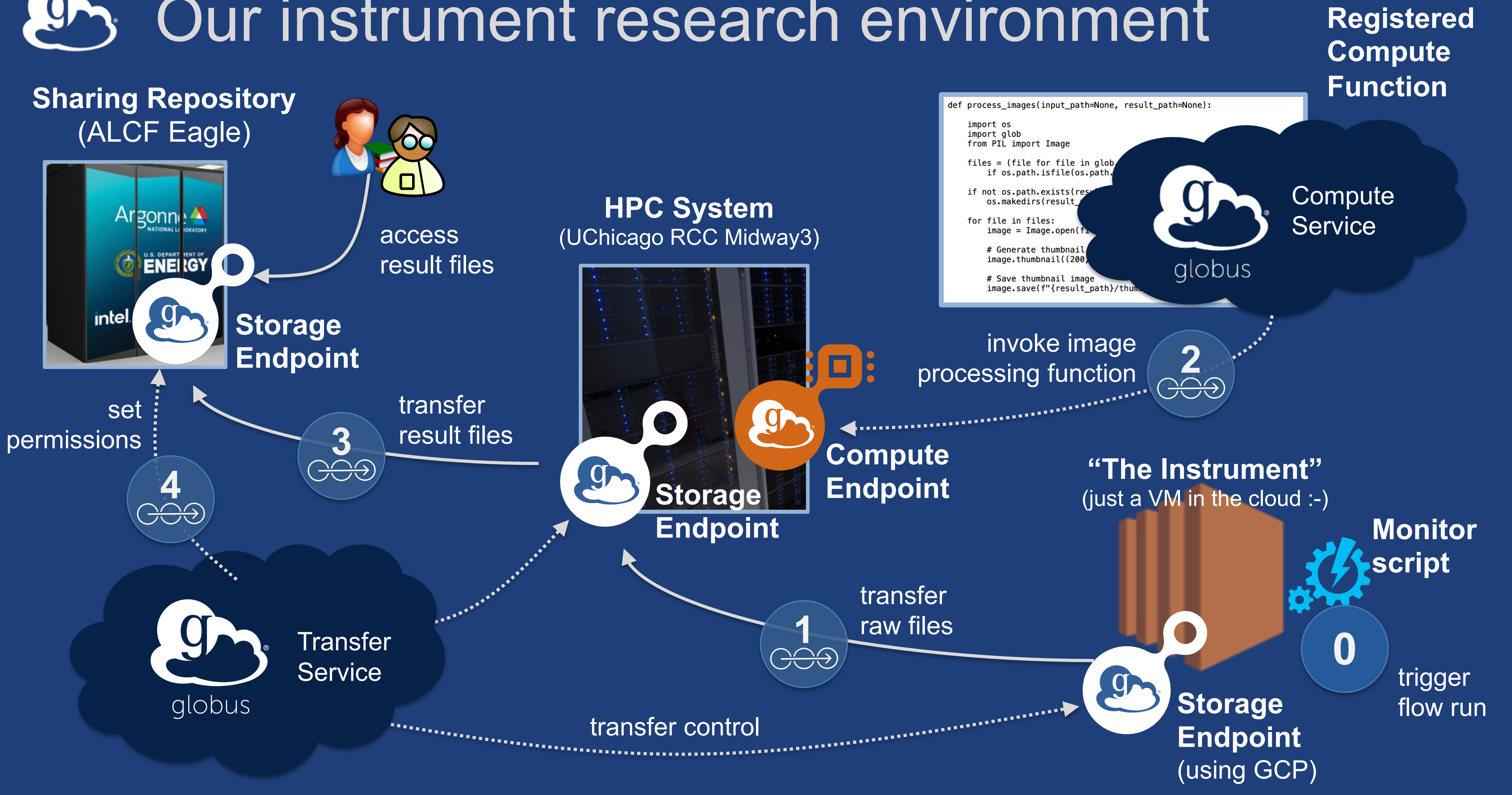
Simulate instrument data creation

Trigger the flow

Adding computation to our instrument flow


Transfer raw instrument images
Run a compute job to process raw image files
Move processed images to repository
Set permissions for accessing the data
Transfer
1
Compute
2
Transfer
3
Share
4

Our instrument research environment



```
def process_images(input_path=None, result_path=None):

    import os
    import glob
    from PIL import Image

    files = (file for file in glob
        if os.path.isfile(os.path.

    if not os.path.exists(res
        os.makedirs(result_

    for file in files:
        image = Image.open(fi

        # Generate thumbnail
        image.thumbnail((200,

        # Save thumbnail image
        image.save(f"{result_path}/thum
```

Key consideration: Be identity aware

- **What identity is the flow running as?**
- **Does identity have access to target resources?**
  - Collections (ideally, guest collections)
  - Compute endpoint
  - Compute function
- **Does identity have the required role?**
  - Access Manager, if granting/revoking permissions

Making Data Findable with Globus Search

Data description and discovery

- Metadata store with fine-grained visibility controls
- Schema agnostic dynamic schemas
- Simple search using URL query parameters
- Complex search using search request document

Search Index

globus

docs.globus.org/api/search

Data ingest with Globus Search

POST /index/{index_id}/ingest'

```
{
  "ingest_type": "GMetaList",
  "ingest_data": {
  "gmeta": [
    {
      "id": "filetype",
      "subject”: "https://search.api.globus.org/abc.txt",
      "visible_to": ["public"],
      "content": {
        "metadata-schema/file#type": "file”
      }
    },
    ...
  ]
}
```

Search Index
globus

- Bulk create and update
- Task model for ingest at scale

Data ingest with Globus Search

POST /index/{index_id}/ingest'

```
{
  "ingest_type": "GMetaList",
  "ingest_data": {
  "gmeta": [
    {
      "id": ”weight",
      "subject": "https://search.api.globus.org/abc.txt",
      "visible_to": ["urn:globus:auth:identity:46bd0f56-
                      e24f-11e5-a510-131bef46955c"],
      "content": {
        "metadata-schema/file#size": ”37.6",
        "metadata-schema/file#size_human": ”<50lb”
      }
    },
    ...
  ]
}
```

Search Index
globus

Visibility limited to Globus Auth identity
- Single user
- Globus Group
- Registered client application

Data discovery with Globus Search

GET /index/{index_id}/search?q=type%3Ahdf5

```
{
  "@datatype": "GSearchResult",
  "@version": "2017-09-01",
  "count": 1,
  "gmeta": [
    {
      "@datatype": "GMetaResult",
      "@version": "2019-08-27",
      "entries": [
        { ... }
      ],
      "subject": "https://..."
    }
  ],
  "offset": 0,
  "total": 1
}
```

Simple query

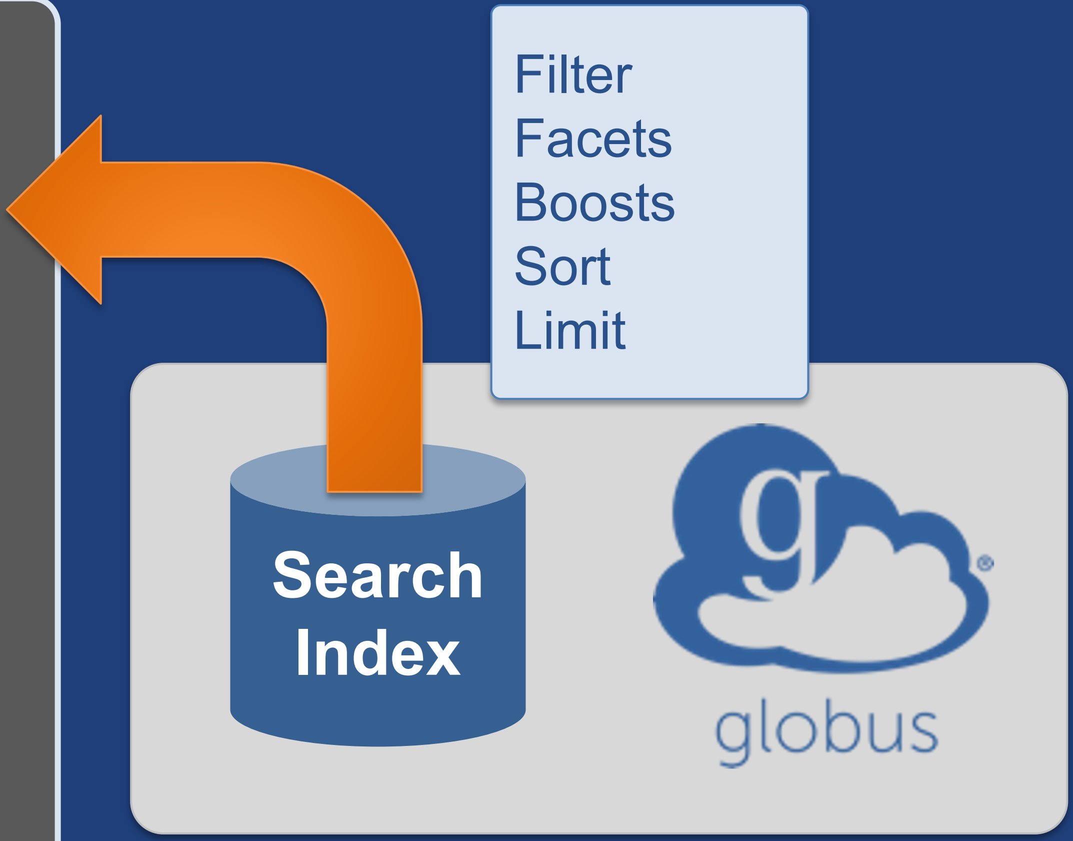
Search Index
globus

Data discovery with Globus Search

POST /index/{index_id}/search

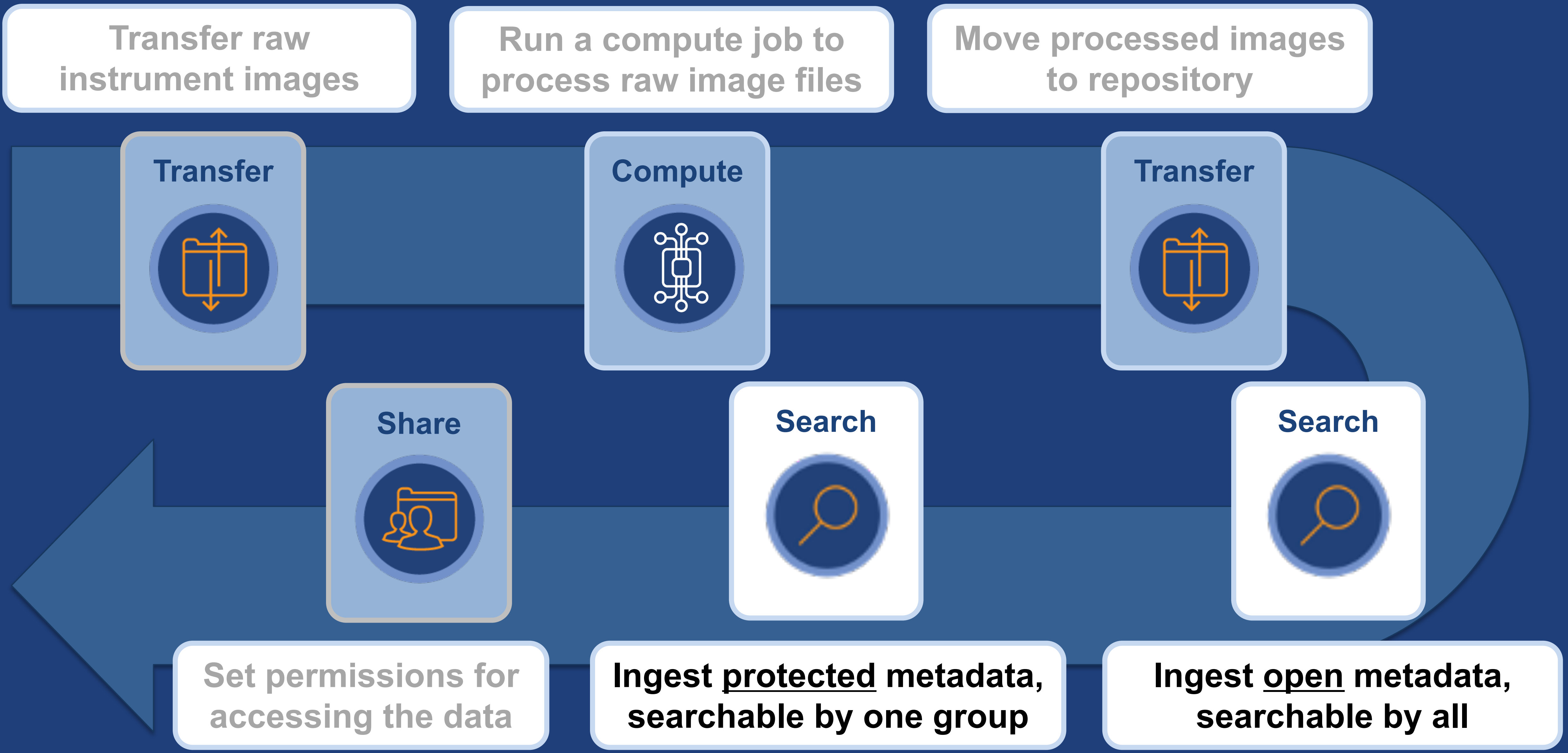
```
{
  "filters": [
    {
      "type": "range",
      "field_name": ”pubdate",
      "values": [
        {
          "from": "*",
          "to": "2020-12-31"
        }
      ]
    }
  ],
  "facets": [
    {
      "name": "Publication Date",
      "field_name": "pubdate",
      ...
    }
  ]
}
```

Complex query



Filter
Facets
Boosts
Sort
Limit

Making our instrument data FAIR



Transfer raw instrument images

Transfer

Run a compute job to process raw image files

Compute

Move processed images to repository

Transfer

Search

Ingest open metadata, searchable by all

Search

Ingest protected metadata, searchable by one group

Share

Set permissions for accessing the data

End-to-end automation in practice: XPCS

Globus Flows

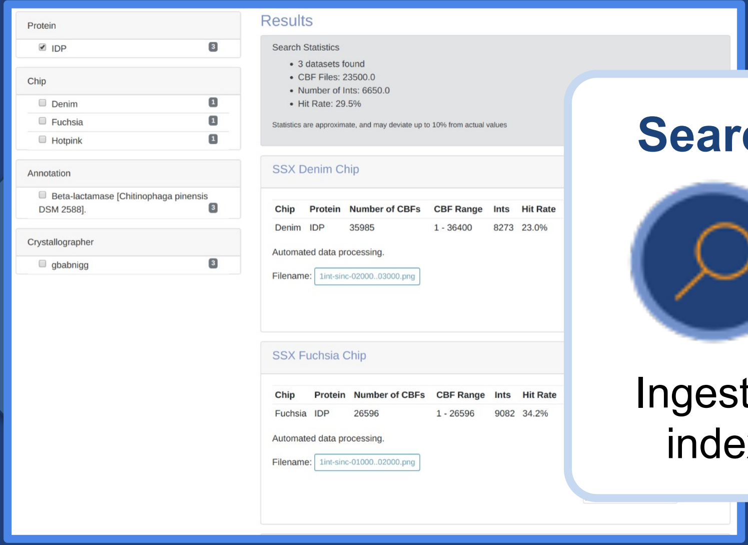
Data capture

Transfer
Transfer HDF5 files

Transfer
Transfer IMM

Compute
Run Corr

FAIR data, ready for discovery!



Protein
☑ IDP 3

Chip
□ Denim 1
□ Fuchsia 1
□ Hotpink 1

Annotation
□ Beta-lactamase [Chitinophaga pinensis DSM 2588]. 3

Crystallographer
□ gbabnigg 3

Results

Search Statistics

- 3 datasets found
- CBF Files: 23500.0
- Number of Ints: 6650.0
- Hit Rate: 29.5%

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

SSX Denim Chip

Chip	Protein	Number of CBFs	CBF Range	Ints	Hit Rate
Denim	IDP	35985	1 - 36400	8273	23.0%

Automated data processing.

Filename: 1int-sinc-02000..03000.png

SSX Fuchsia Chip

Chip	Protein	Number of CBFs	CBF Range	Ints	Hit Rate
Fuchsia	IDP	26596	1 - 26596	9082	34.2%

Automated data processing.

Filename: 1int-sinc-01000..02000.png

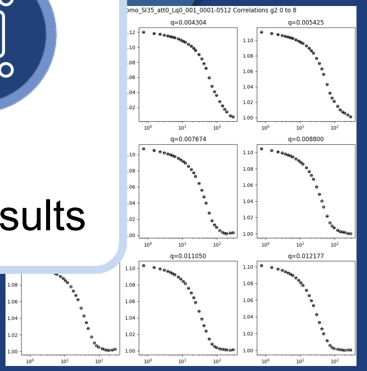
Search
Ingest to index

Share
Set access controls

Transfer
Move results to repo

Compute
Gather metadata

Compute
Plot results

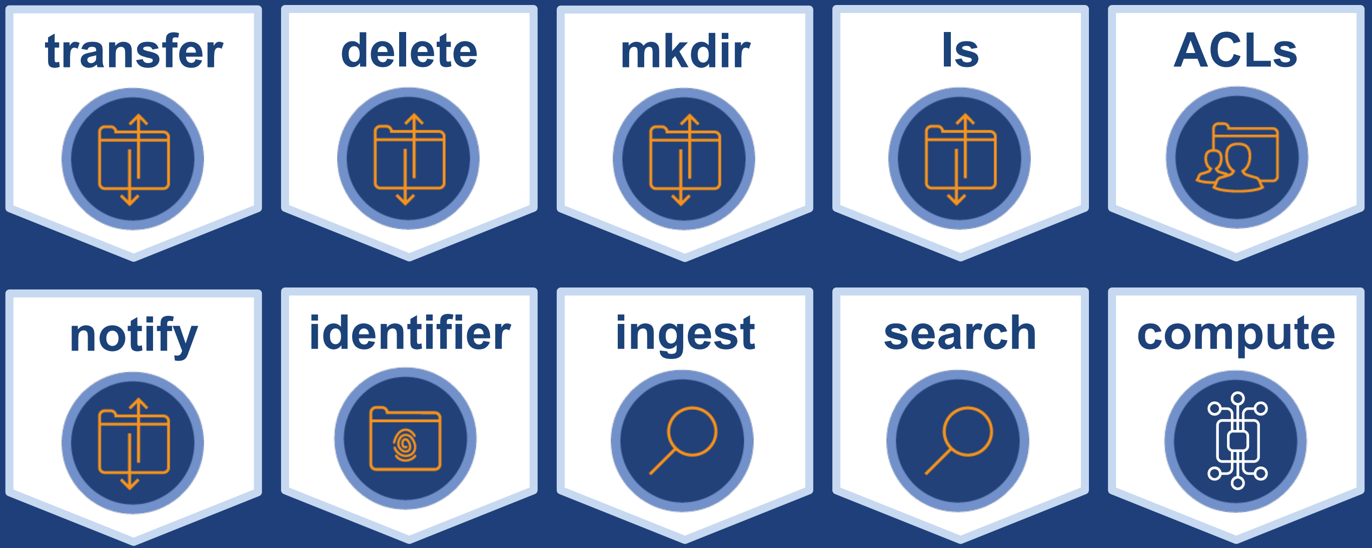


Take a look…

acdc.alcf.anl.gov

Extending the ecosystem: Action Providers

- **Action Provider is a service endpoint**
  - Run
  - Status
  - Cancel
  - Release
  - Resume



docs.globus.org/api/flows/hosted-action-providers

- **Action Provider Toolkit**

  action-provider-tools.readthedocs.io

web form
Xtract
describe

Custom developed