

Globus Flows and Compute:

Smart automation around data submission

Andy Boughton
(abought@globus.org)
APECx Demos – April 2025



THE UNIVERSITY OF
CHICAGO



Argonne
NATIONAL LABORATORY





Globus Flows: Part of an integrated platform



Learn more:

<https://www.globus.org/platform>

<https://docs.globus.org/api/flows/>

<https://globus-compute.readthedocs.io/en/latest/>



Today's notes:

<https://github.com/globus/apecx-demos/>



Try it yourself

Get an endpoint: <https://www.globus.org/globus-connect-personal>

Library of public flows: <https://app.globus.org/flows/library>



Globus Flows:

**Secure, managed automation of
complex workflows at scale**





Big projects need automation

- **Globus Flows: automate common tasks**
- **Example use cases**
 - Ensure that users obey best practices for secure data movement
 - Define complex procedures such as multi-step uploads to a secure environment
 - Hide sensitive services from being directly invoked, except in a controlled way





Imperative, distributed, and resilient

- **Distributed design allows leveraging many different services as *action providers***
- **Automation will retry and resume if a system component is unavailable**
- **Imperative flow syntax based on Amazon Step Functions**



Integrates with Globus Services

- **Action providers can invoke many Globus Features**
 - Conditional logic can take steps based on results at each step
- **Auto-generated UI for running flows and selecting files**



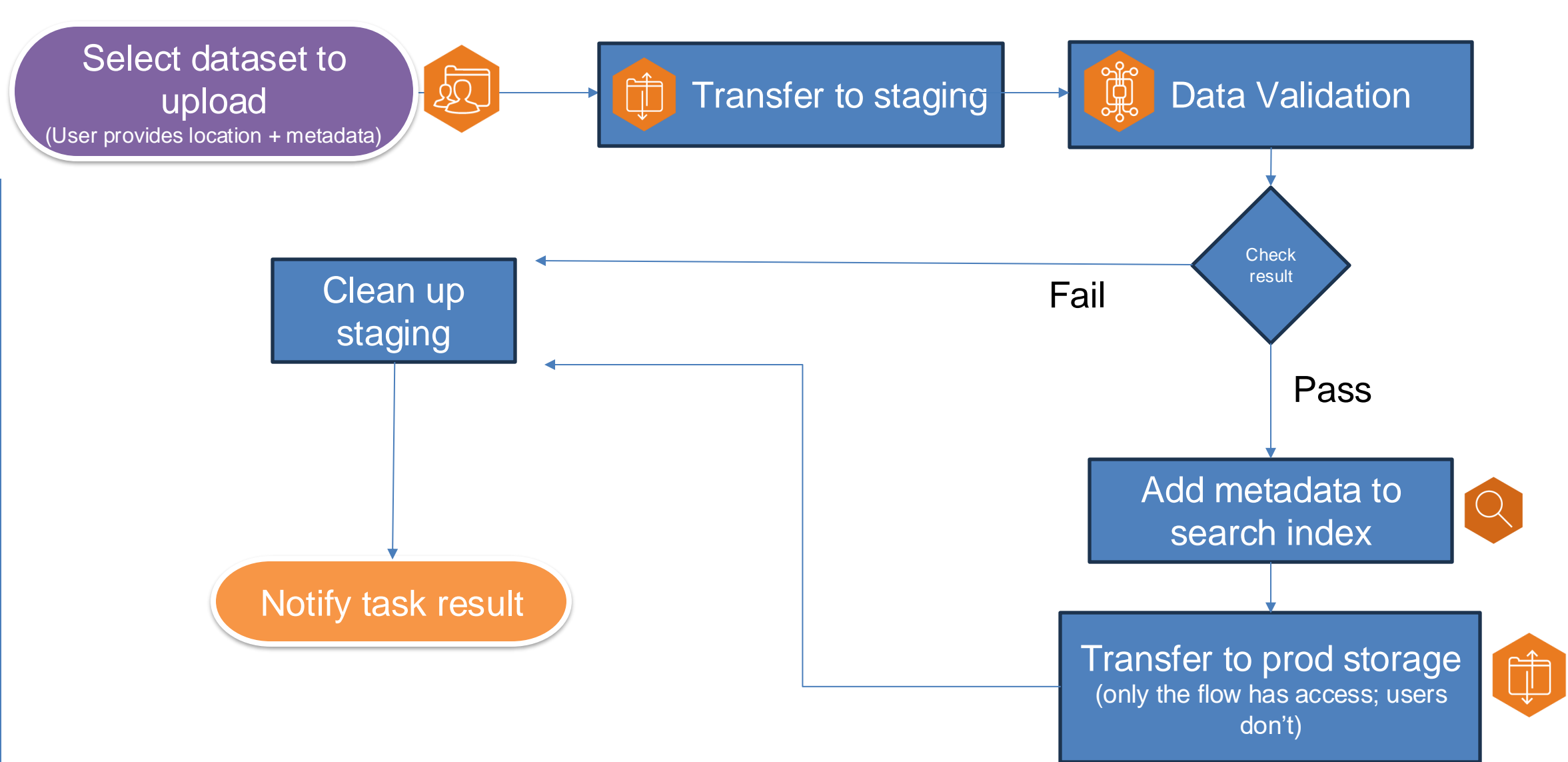
Start - Example flow

Guided Advanced

[Back to Flows Library](#)

Source [OPTIONAL]
The data's origin
Collection

Intermediate [OPTIONAL]
An intermediate location used to hold data, often used to manage network usage
Collection



Transfer + Compute + Search



Start - Example flow



Guided



Advanced



[Back to Flows Library](#)



Source [OPTIONAL]

The data's origin

Collection

Browse



Intermediate [OPTIONAL]

An intermediate location used to hold data, often used to manage network usage

Collection

Browse

Above: UI to start the flow

Right: UI to view flow results, step-by-step

Started: 4/2/2025, 04:58 PM

Duration: 25 seconds

Sort



Download 20 Log Entries



FlowSucceeded

[View details](#)



SuccessReportValidation – PassCompleted

[View details](#)



SuccessReportValidation – PassStarted

[View details](#)



SearchIngest – ActionCompleted (7 seconds)

[View details](#)



SearchIngest – ActionStarted

[View details](#)



EvaluateValidationResult – ChoiceCompleted

[View details](#)



EvaluateValidationResult – ChoiceStarted

[View details](#)



RunValidation – ActionCompleted (6 seconds)

[View details](#)



RunValidation – ActionStarted

[View details](#)



CopySourceToIntermediate – ActionCompleted (6 seconds)

[View details](#)



CopySourceToIntermediate – ActionStarted

[View details](#)



MakeIntermediate – ActionCompleted (1 second)

[View details](#)



MakeIntermediate – ActionStarted

[View details](#)



ComputeIntermediatePath – ActionCompleted (723 milliseconds)

[View details](#)

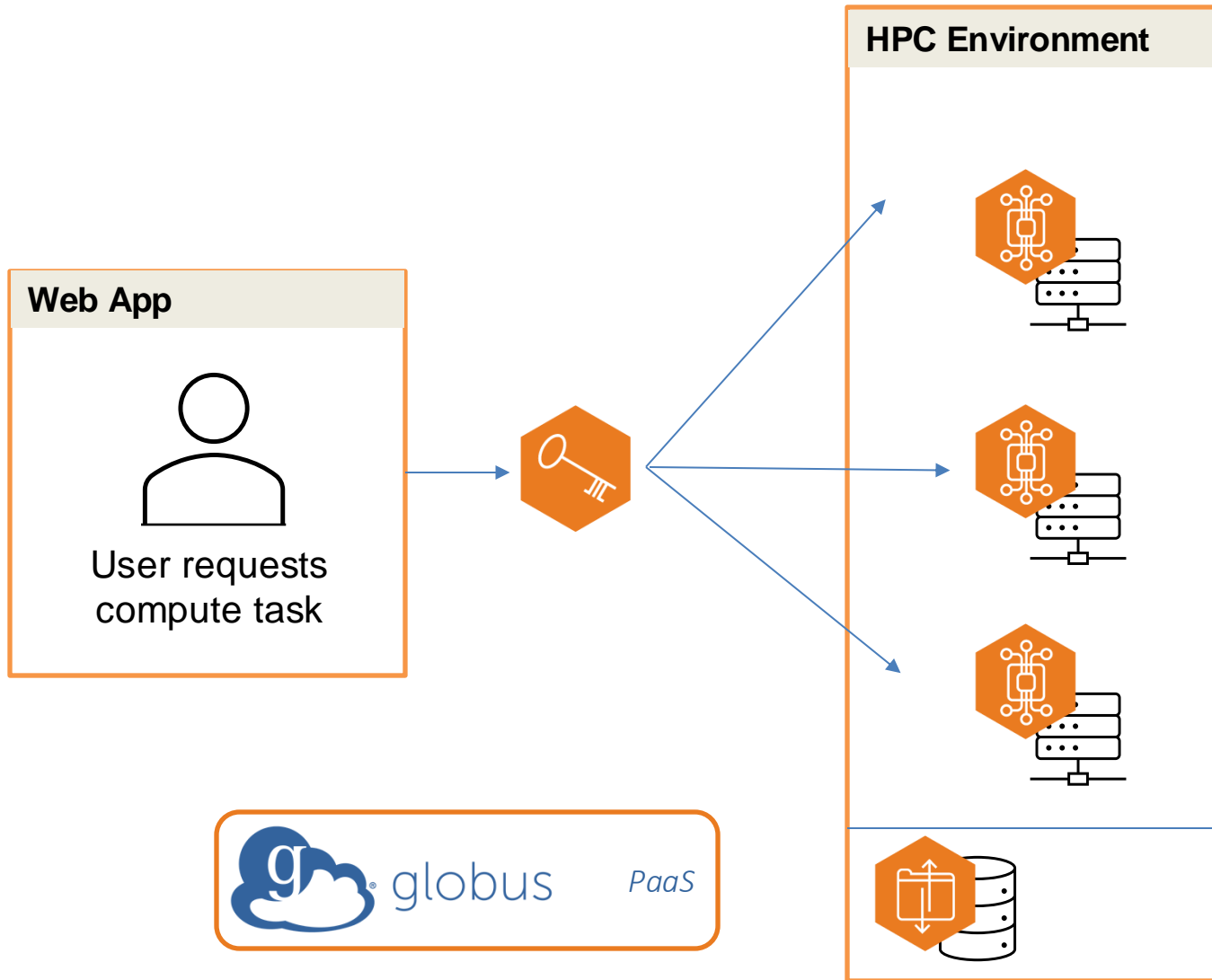


ComputeIntermediatePath – ExpressionEvalStarted

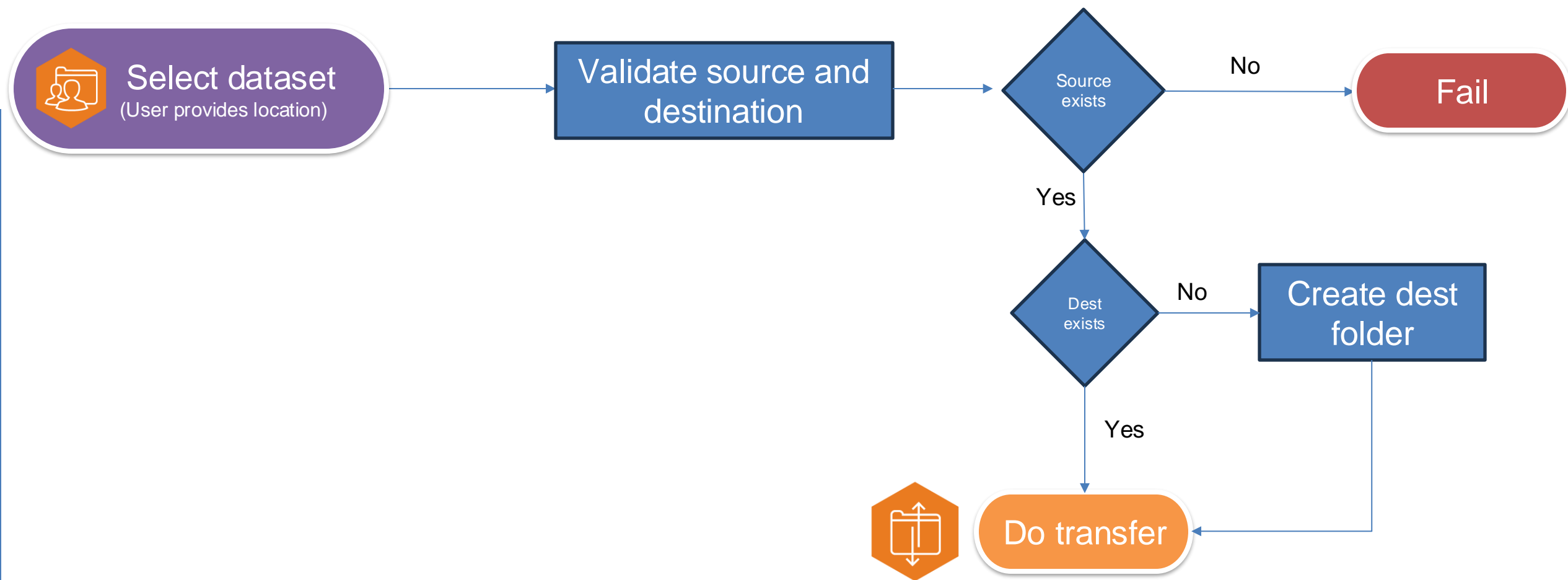
[View details](#)



Fan out across many workers



- **Allow users to request big compute from a website**
- **Scale compute separately from webapp**
- **Enable remote access to compute resources, without SSH**



Process Automation: Ensure that data upload complies with internal policies

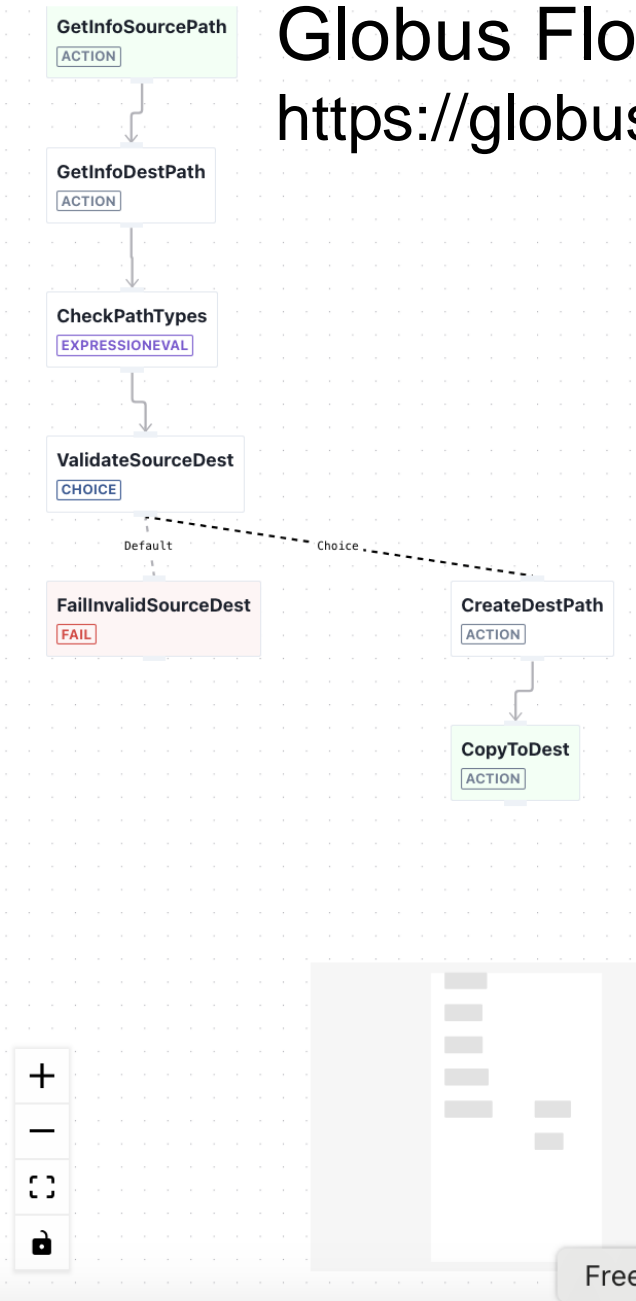
Demo repo: [flows-compute/flow/data/transfer_with_filters](https://github.com/flows-compute/flow/data/transfer_with_filters)

```
{
  "Comment": "Transfer files from a sou",
  "StartAt": "GetInfoSourcePath",
  "States": {
    "GetInfoSourcePath": {
      "Type": "Action",
      "ActionUrl": "https://transfer.ac",
      "Comment": "Get the source path i",
      "Next": "GetInfoDestPath",
      "ResultPath": "$._run.GetInfoSour",
      "Parameters": {
        "endpoint_id.$": "$.source.id",
        "path.$": "$.source.path"
      }
    },
    "GetInfoDestPath": {
      "Type": "Action",
      "ActionUrl": "https://transfer.ac",
      "Comment": "Get the dest path inf",
      "Next": "CheckPathTypes",
      "ResultPath": "$._run.GetInfoDest",
      "Parameters": {
        "endpoint_id.$": "$.dest.id",
        "path.$": "$.dest.path"
      }
    },
    "CheckPathTypes": {
      "Type": "ExpressionEval",
      "Comment": "Check the paths to se",
      "Next": "ValidateSourceDest",
      "ResultPath": "$._run.CheckPathTy",
      "Parameters": {
        "dest_exists.=": "'code' not in",
        "source_is_dir.=": "'type' in _"
      }
    },
    "ValidateSourceDest": {
      "Type": "Choice",
      "Comment": "Ensure that path crit
```

Validate

Diagram

Documentation



Globus Flows IDE (Beta)

<https://globus.github.io/flows-ide>

Flows

Runs Library Deploy a Flow Beta

List of flows you may view or use.

search flow library

60 flows available to you in the library

☒ GLOBUS-PROVIDED FLOW ☐ ADMINISTERED BY ME ☐ RUNNABLE BY ME

Two-Stage Transfer

Globus Team

Transfer from a source collection to a destination collection using an intermediate endpoint

STEPS	CREATED	LAST MODIFIED	KEYWORDS
27	7/27/2023, 01:10 PM	10/21/2024, 10:58 AM	Globus-Provided,Two Stage,Two Hop,Intermediate,Globus Transfer,Transfer,Production

Move (Copy and Delete) Files

Globus Team

Transfers data from a source collection to a destination collection, then deletes the data from the source collection

STEPS	CREATED	LAST MODIFIED	KEYWORDS
20	7/27/2023, 01:09 PM	10/21/2024, 10:56 AM	Globus-Provided,Move,Data Move,Globus Transfer,Transfer,Production

Transfer and Set Permissions

Globus Team

Transfer to a guest collection that you own or manage and grant an identity or group access to the data.

STEPS	CREATED	LAST MODIFIED	KEYWORDS
31	9/30/2024, 05:49 PM	10/2/2024, 05:58 PM	Globus-Provided,Transfer,Globus Transfer,Guest Collection,Sharing,Permission,Access,Read,Write

Globus Flows Library

<https://app.globus.org/flows/library>



Documentation and Tooling

- **Docs:**
 - <https://docs.globus.org/api/flows/>
- **Tools for writing flows:**
 - <https://globus.github.io/flows-ide>
 - <https://glacier.readthedocs.io/>

```
{
  "Comment": "Transfer files from a source folder (that must exist) to",
  "StartAt": "GetInfoSourcePath",
  "States": {
    "GetInfoSourcePath": {
      "Type": "Action",
      "ActionUrl": "https://transfer.actions.globus.org/stat",
      "Comment": "Get the source path info",
      "Next": "GetInfoDestPath",
      "ResultPath": "$._run.GetInfoSourcePath",
      "Parameters": {
        "endpoint_id.$": "$.source.id",
        "path.$": "$.source.path"
      }
    },
    "GetInfoDestPath": {
      "Type": "Action",
      "ActionUrl": "https://transfer.actions.globus.org/stat",
      "Comment": "Get the dest path info",
      "Next": "CheckPathTypes",
      "ResultPath": "$._run.GetInfoDestPath",
      "Parameters": {
        "endpoint_id.$": "$.dest.id",
        "path.$": "$.dest.path"
      }
    }
  }
},
```



When to use Globus Flows

Common use cases

- **Manage complex permissions models (such as submission to data repository)**
- **Enforce best practices (such as filtering out sensitive data)**
- **Perform actions after transfer**
 - Add to search index
 - Data validation (quick compute)
 - Notify curators

Not aimed at these scenarios

- **Tasks focused on compute, rather than data management**
- **Very big workflows with many complex steps**
- **Makefile-like “only rebuild what changed” behavior**





Globus Compute:

**Add custom FaaS steps into your
workflow**





Function as a Service (FaaS)

- User-defined functions that can be run on a **remote** environment
- Enable **access** to compute resources without using SSH
- Retrieve results in workflows, notebooks, etc

```
def process_images(input_path=None, result_path=None):  
    import os  
    import glob  
    from PIL import Image  
  
    files = (file for file in glob.glob(os.path.join(input_path, '*')) \  
             if os.path.isfile(os.path.join(input_path, file)))  
  
    if not os.path.exists(result_path):  
        os.makedirs(result_path)  
  
    for file in files:  
        image = Image.open(file)  
  
        # Generate thumbnail  
        image.thumbnail((200, 200))  
  
        # Save thumbnail image  
        image.save(f"{result_path}/thumb_{os.path.basename(file)}")
```



FaaS as an interface to the advanced computing ecosystem



Still need...

- **Single interface**
- **Homogenous execution environment**
- **Transparent and elastic execution**
- **Integrated with data management**



Example: Automating cryoEM flows

Globus
Flows



Transfer



Transfer
raw files

Compute



Launch
analysis job

Carbon!



Correct,
classify, ...

Compute



Extract
metadata

Share



Set access
controls

Transfer

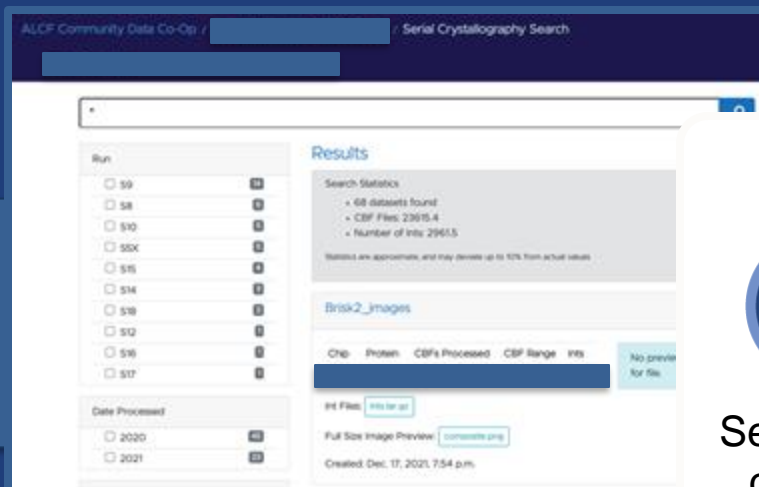


Move final
files to repo

Search



Ingest to
index



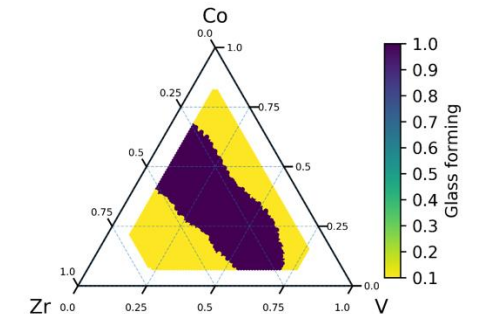
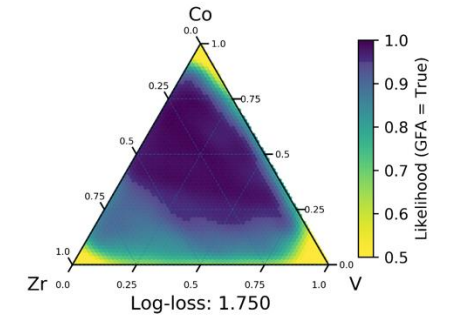
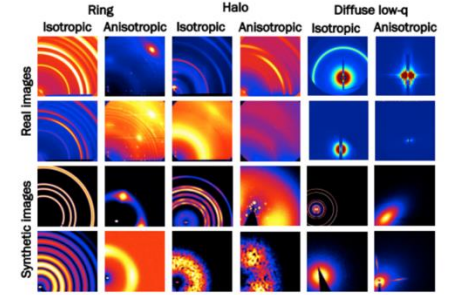
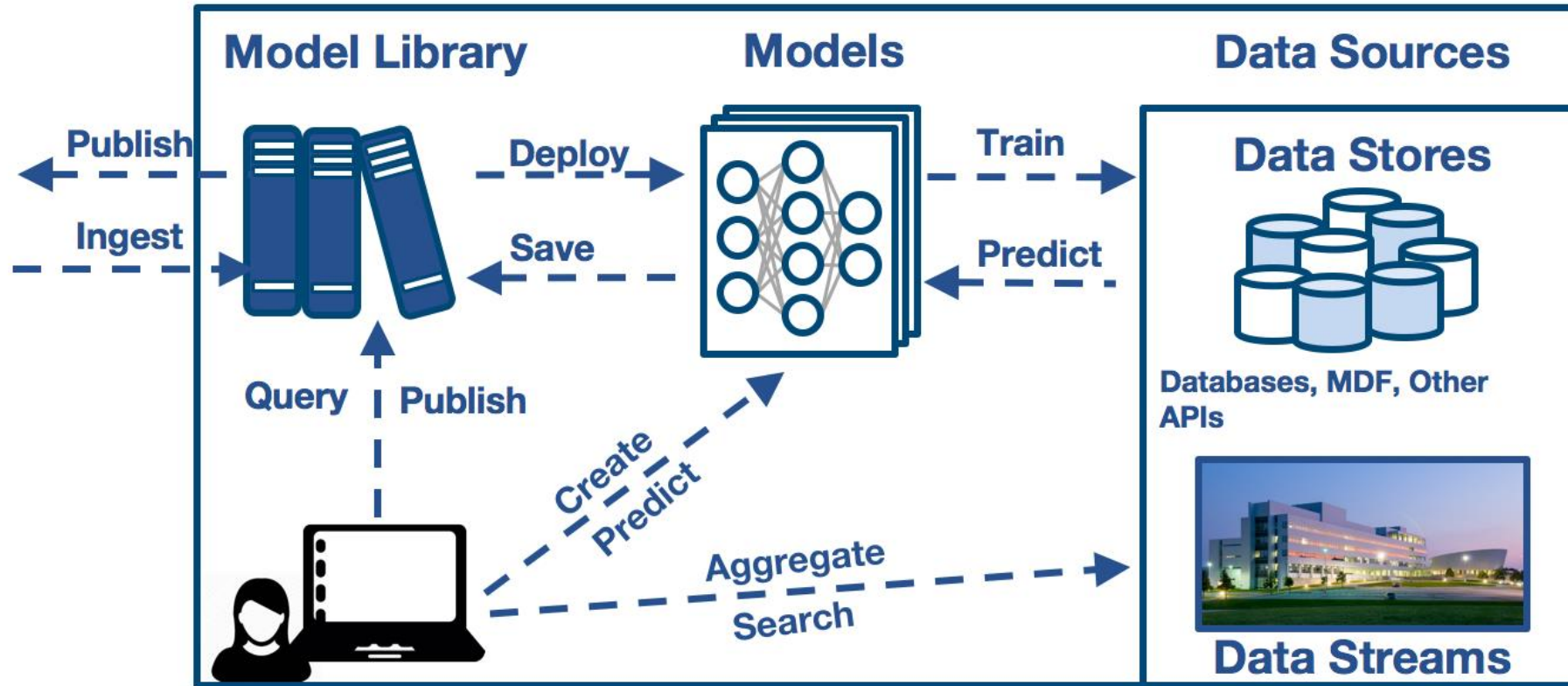


Integrates with Globus **Flows**

- **Can add custom data processing or validation into an upload step**
 - Verify that data is AI/ML ready before adding to catalog
- **Can use workflows to initiate compute functions in a controlled way**
- **Flows can directly use the output of this function in following steps**



An ecosystem of AI/ML models: DLHub, Garden.ai, etc



<https://thegardens.ai/>



Limits

- **Not tightly integrated with Globus Storage/Transfer**
- **More complex functions may take sysadmin help to deploy the first time**
 - Not endpoint-agnostic
 - Pick a few well-defined AI/ML models
- **Compute continues to *evolve* to support new use-cases**

Recap



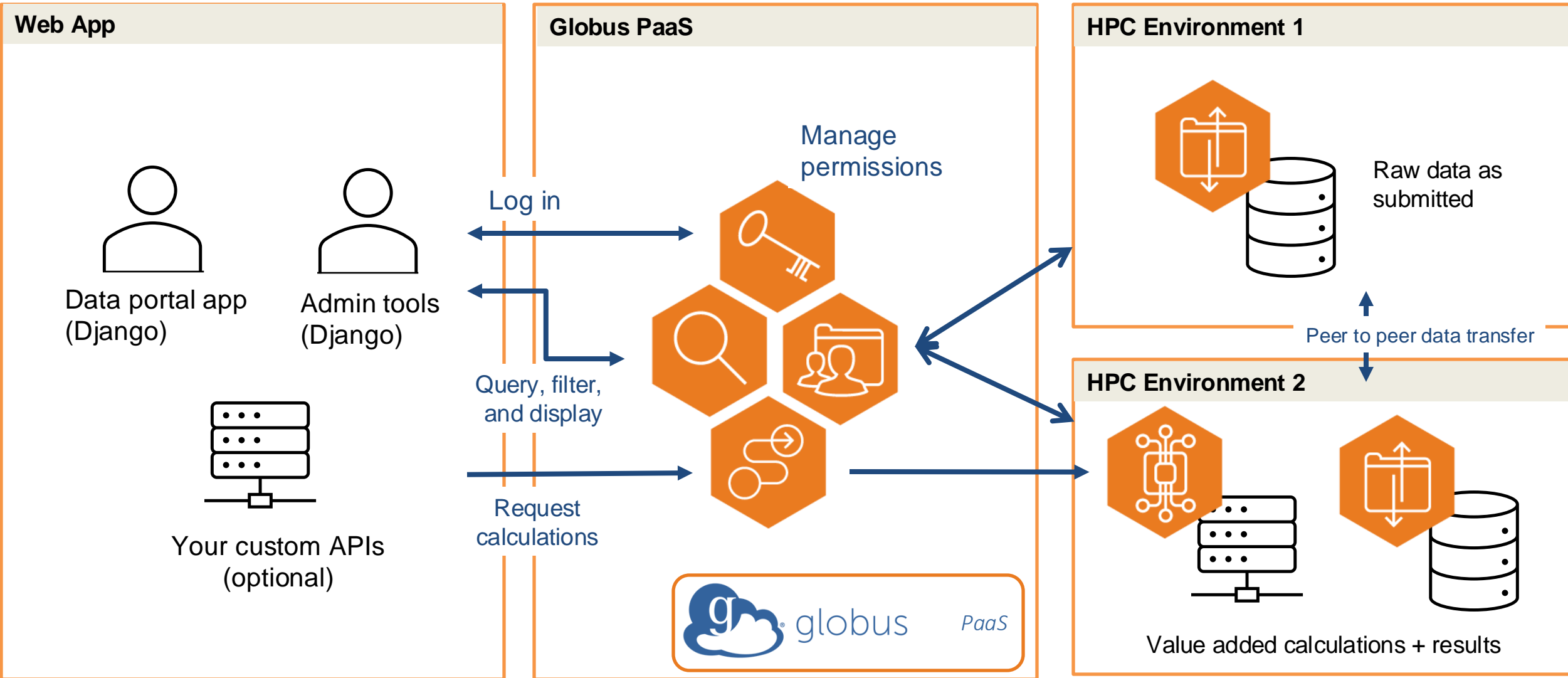


User stories recap

- **Transfer to get files in**
- **Flows to provide automation + data validation**
- **Compute to provide access to advanced resources**
- **Search index + DGPF UI to make data discoverable**
- ***...and a whole lot of science from you***



Portal Platform: Recap



Discussion questions

**(...To be continued at ANL
April Gathering!)**





What do we want to automate?

- **Data deposition: teams with sensitive data management concerns**
- **Data validation: basic checks of common data types**
 - What teams are developing these pieces?
 - What is the path to integrating with a data repository for AI/ML ready models?
- **Data curation / required approval**
 - How can curation flows be integrated with data ingest?



What **capacity** is needed?

- **Compute can provide access to big compute resources, but requires setup**
 - Are there methods / tools we want to expand access to?
- **Are people interested in collocating compute + data?**
 - This may require advance planning, but could leverage a strength of the data repository





Who needs access?

- Which team(s) would need to use compute? What environments need access?
- Touchpoints: Who is developing the data harmonization and curation? How do we integrate them with the catalog?