Introduction to Research Data Management Using Globus

Rick Wagner
rick@globus.org

SDSC Summer Institute 2019
August 9, 2019

Research data management today

Index?

How do we...
...move?
...share?
...discover?
...reproduce?

Globus delivers…

Fast and reliable ~~big~~ data transfer, sharing, and platform services…

…directly from your own storage systems…

...via software-as-a-service using existing identities with the overarching goal of...

…unifying access to data across tiers


Research Computing HPC

National Resources


Personal Resources


Desktop Workstations


Mass Storage

Instruments

Public Cloud

Storage Connectors - globus.org/connectors

Current

IBM Spectrum Scale

SPECTRA

amazon web services S3

ceph

Google Drive

hadoop HDFS

box

globus connect

Lustre®

HPSS

HGST ActiveScale
a Western Digital brand

Planned

Google Cloud

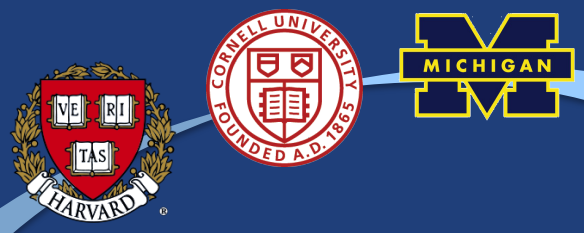
wasabi
hot cloud storage

Microsoft Azure

Share with collaborators/community

Project repositories, replication stores

Public repositories

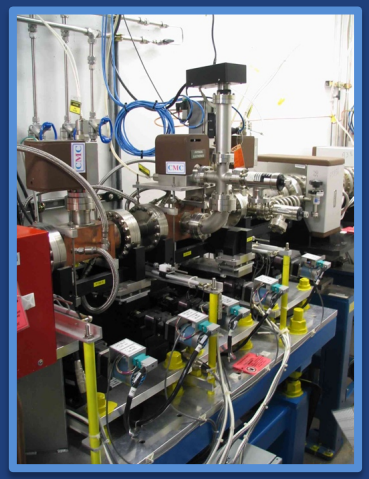


External campus storage

Public / private cloud stores

Manage data from instruments

Next-Gen Sequencer



Advanced Light Source

Cryo-EM

MRI

Light Sheet Microscope

Analysis store

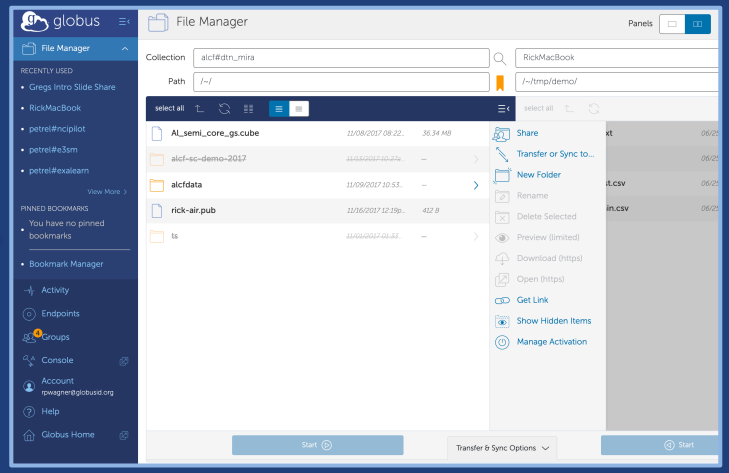
High-durability, low-cost store

Remote visualization

Personal system

Use(r)-appropriate interfaces

Globus service



Web

```
(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.
```

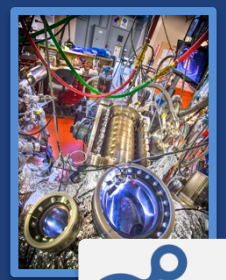
CLI

```
GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
...
```

Rest API

Globus SaaS / PaaS: Research data lifecycle

Instrument



Globus transfers files reliably, securely

2

Compute Facility

4 Globus controls access to shared files on existing storage; no need to move files to cloud storage!

6 The Globus Command Line Interface, API sets, and Python SDK provide a platform…

Build

Transfer

3

Share

1

Researcher initiates transfer request; or requested automatically by script, science gateway

Researcher selects files to share, selects user or group, and sets access permissions

jupyterhub

7 … for building science gateways, portals and publication services.

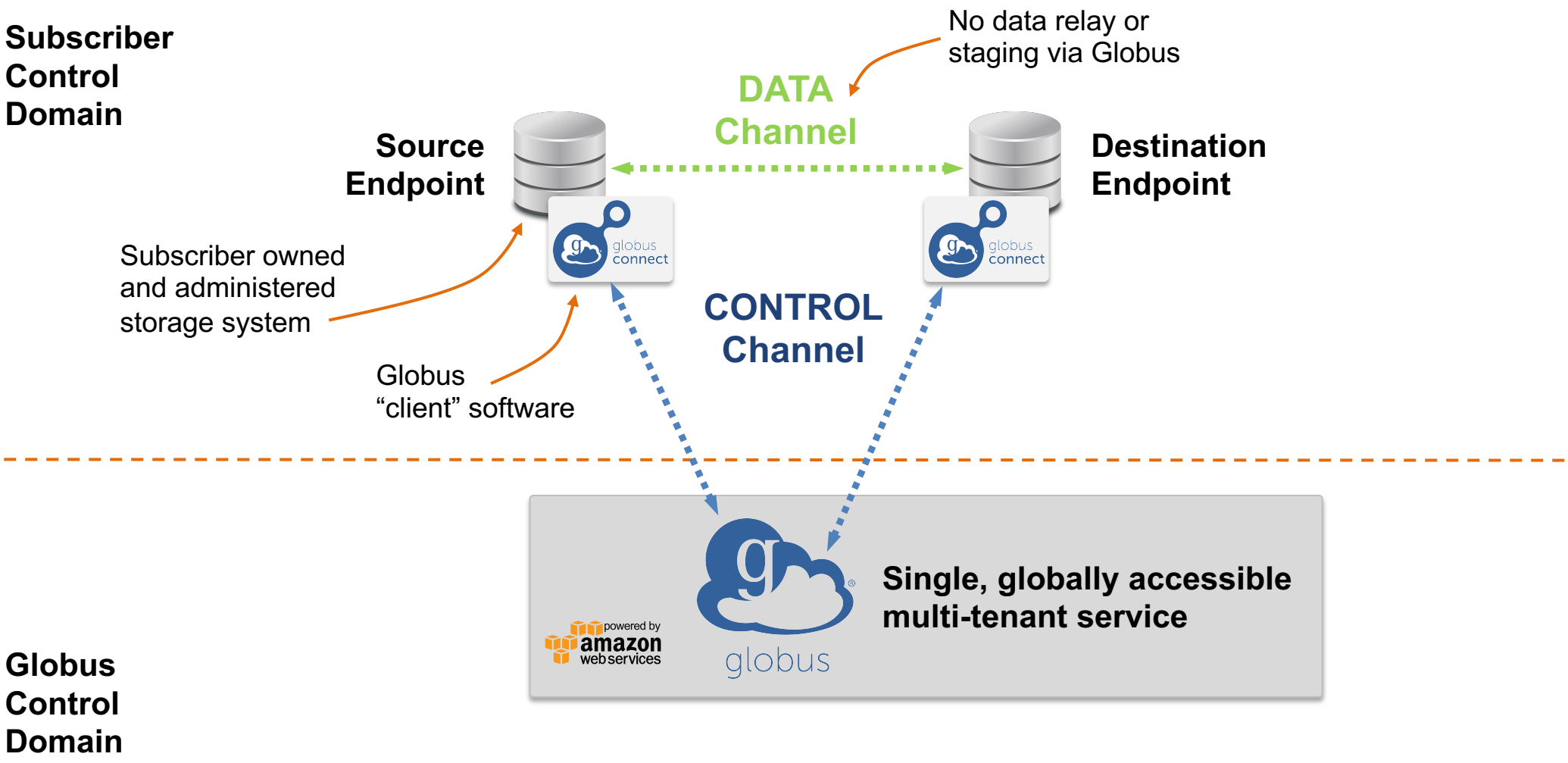
5 Collaborator logs in to Globus and accesses shared files; no local account required; download via Globus

Personal Computer

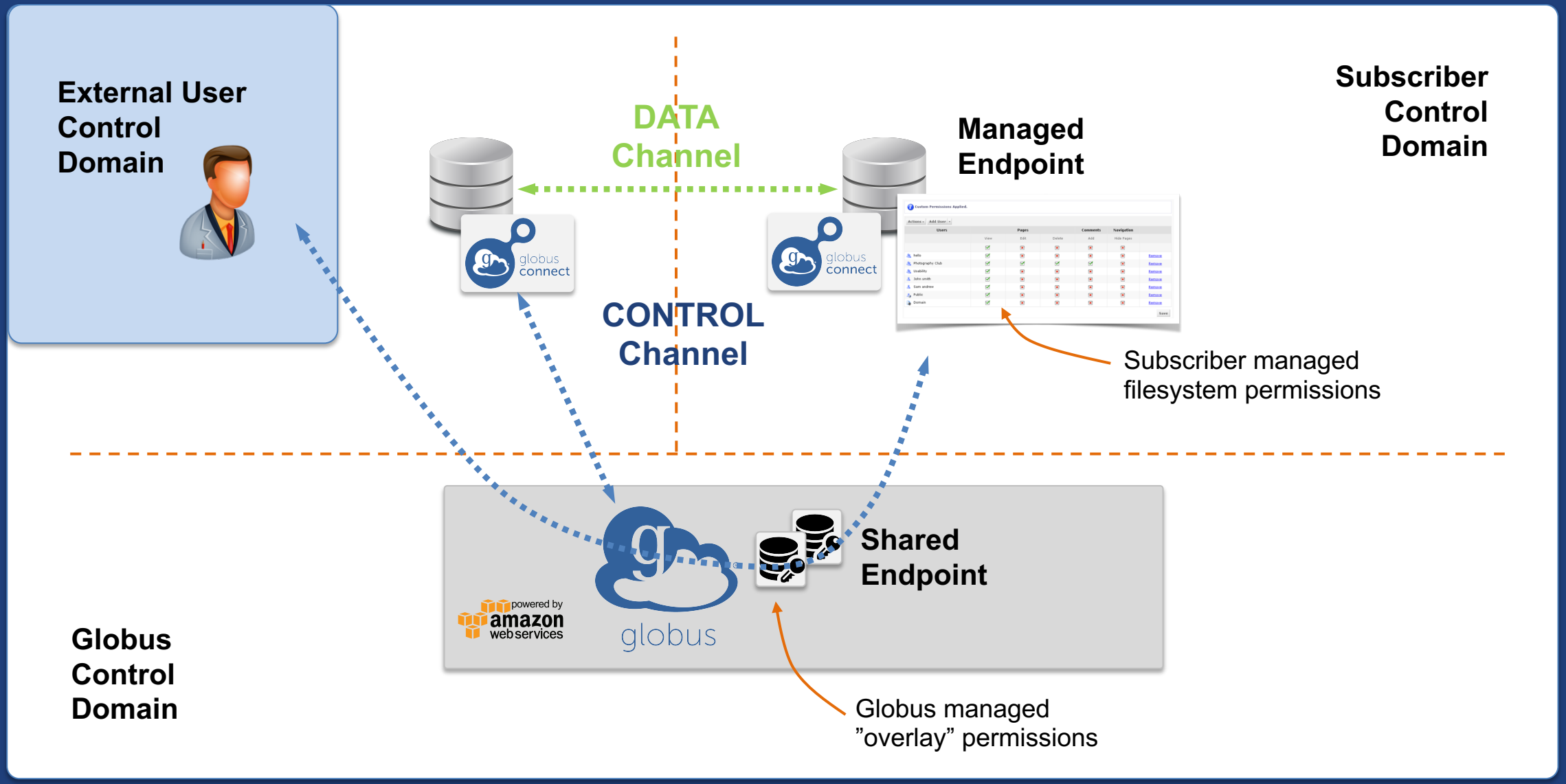
8 Automating research workflows and ensuring those that need access to the data have it.

- **Use a Web browser or platform services**
- **Access any storage**
- **Use an existing identity**

Conceptual architecture: Hybrid SaaS



Conceptual architecture: Sharing



…makes your storage system a Globus endpoint

Endpoints (Collections)

- **Storage abstraction**
  - All transfers happen between two endpoints
  - Globus Connect instantiates endpoints
- **Collection ~= Endpoint**
- **Test / Demo Endpoints**
  - Globus Tutorial Endpoint 1
  - Globus Tutorial Endpoint 2
  - ESnet Test Endpoints
    - Contain file samples of various sizes
- **Globus Connect Personal**
  - Now your laptop is an endpoint
  - https://www.globus.org/globus-connect-personal

Globus Connect Personal

- **Installers do not require admin access**
- **Zero configuration; auto updating**
- **Handles NATs**
- **Installs in seconds – easy to delete**

The Globus Web App - Accounts

- **A Globus Account is**
  - A Primary Identity
  - Possible Linked Identities
- **Linking Identities**
- **Managing Identities**
- **Consents**

Demonstration

Identities

File Transfer

File Sharing

Activity Monitoring

- **Recent / History / Filter**
- **Drilling Down**
  - File transfer statistics
  - Overview
  - Event Log
  - Cancelling an active task

Groups

- **What can they be used for?**
  - Sharing: Access permissions for more than one person
  - Roles: Endpoint management and monitoring
- **Groups**
  - Creating groups and setting the visibility
  - Members (invitations), Subgroups, Settings
  - Settings
    - Policies / Membership Fields / Terms & Conditions
  - Roles
    - Giving others authority over your groups

Endpoint Sharing and Roles

- **Sharing**
  - Select the directory and create the “share”
  - A ”share” is another type of endpoint
  - Share with: Users / Groups / All Globus Users
- **Roles**
  - Giving others (or groups of others) control or monitoring rights for your endpoints

Bookmarks

- **Just like browser bookmarks – frequently used, or maybe not used frequently enough!**
- **Creating a bookmark**
- **Using a bookmark**
- **Sorting and Filtering**
- **Editing and Deleting**

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.
```

Open source, uses Python SDK

docs.globus.org/cli

github.com/globus/globus-cli

The Globus CLI

- **Installation**
  - docs.globus.org/cli/installation
  - Prerequisites
- **Logging On (remember the consents?)**
  - globus login / logout
- **Getting help / list of commands**
  - globus –help
  - globus list-commands
- **Doing something**
  - It all about the UUIDs
  - Don’t forget the file paths!

The Globus CLI – Let's do a few things…

- **Find endpoints**
  - globus endpoint search Midway
  - globus endpoint search ESNet
  - globus endpoint search --filter-scope=recently-used
- **Find endpoint contents**
  - globus ls af7bda53-6d04-11e5-ba46-22000b92c6ec
  - globus ls af7bda53-6d04-11e5-ba46-22000b92c6ec:RMACC2018
- **Transfer a file**
  - From ESnet Read-Only Test DTN at CERN to Midway
  - Note the specific paths
  - globus transfer d8eb36b6-6d04-11e5-ba46-22000b92c6ec:/~/data1/1M.dat af7bda53-6d04-11e5-ba46-22000b92c6ec:/~/1M.dat
- **Transfer a directory**
  - From Globus Tutorial Endpoint 2 to Midway (create directory and contents)
  - globus transfer --recursive ddb59af0-6d04-11e5-ba46-22000b92c6ec:/~/sync-demo af7bda53-6d04-11e5-ba46-22000b92c6ec:/~/syncDemo
- **https://docs.globus.org/cli/examples/**

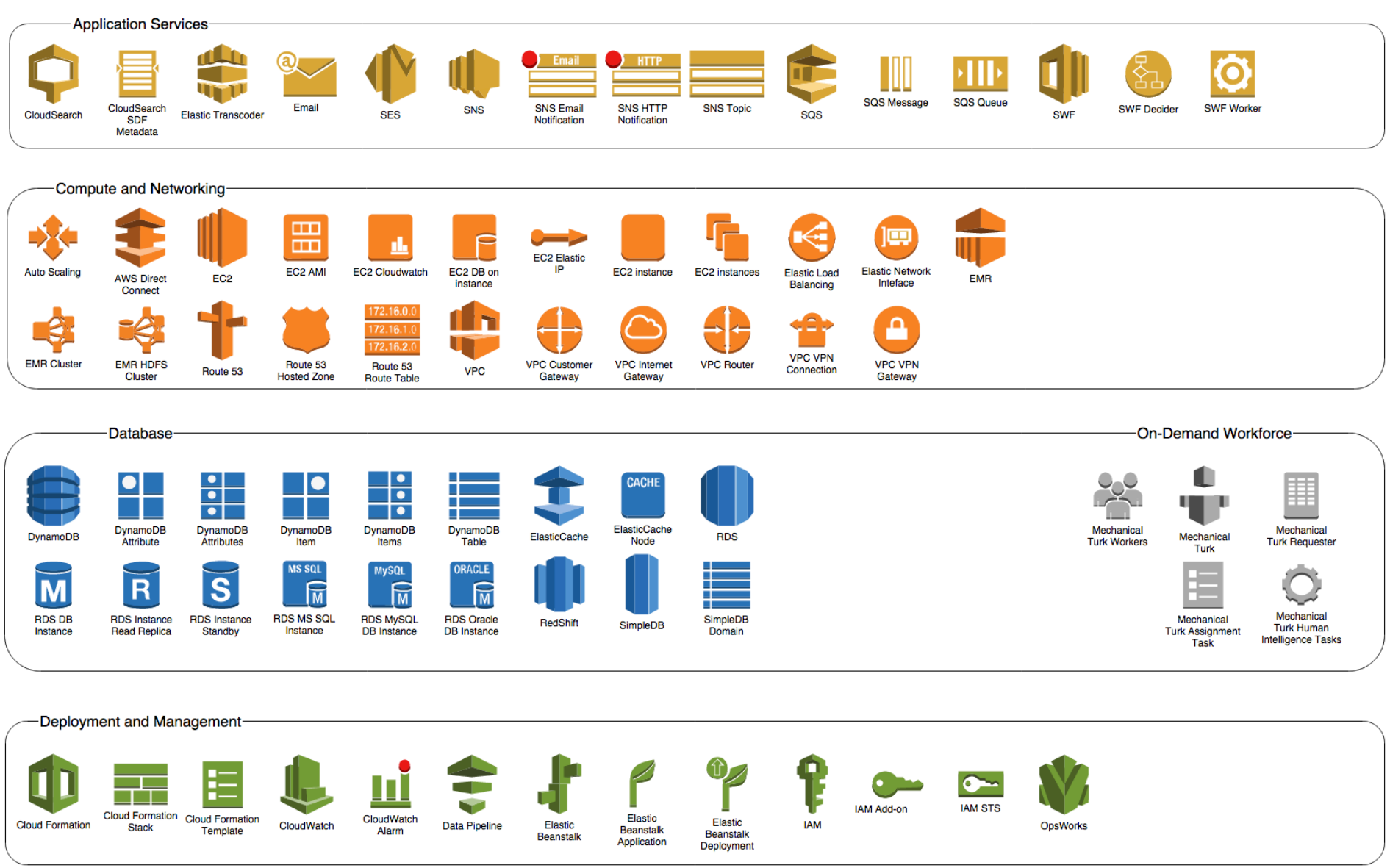
Globus CLI

- **Easy install and updates**
- **It's a native application distributed by Globus**
  - https://docs.globus.org/cli/
  - https://github.com/globus/globus-cli
- **Command *globus login* gets access tokens and refresh tokens**
  - Stores the token locally (~/.globus.cfg)
  - The CLI "acts as" the logged in user
- **All interactions with the service use the tokens**
  - Tokens for Globus Auth and Transfer services
- **Command *globus logout* deletes those**
- **https://docs.globus.org/cli/examples/**
- **https://github.com/globus/automation-examples**

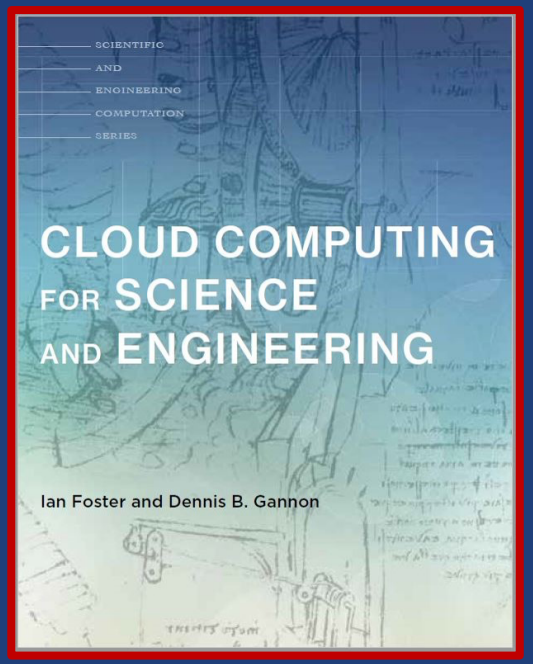
Demonstration

Globus CLI

Industry software builds on platform services



cloud4scieng.org



Globus delivers… with applications and as a platform…

Fast and reliable data transfer, sharing, and file management…

…directly from your own storage systems…

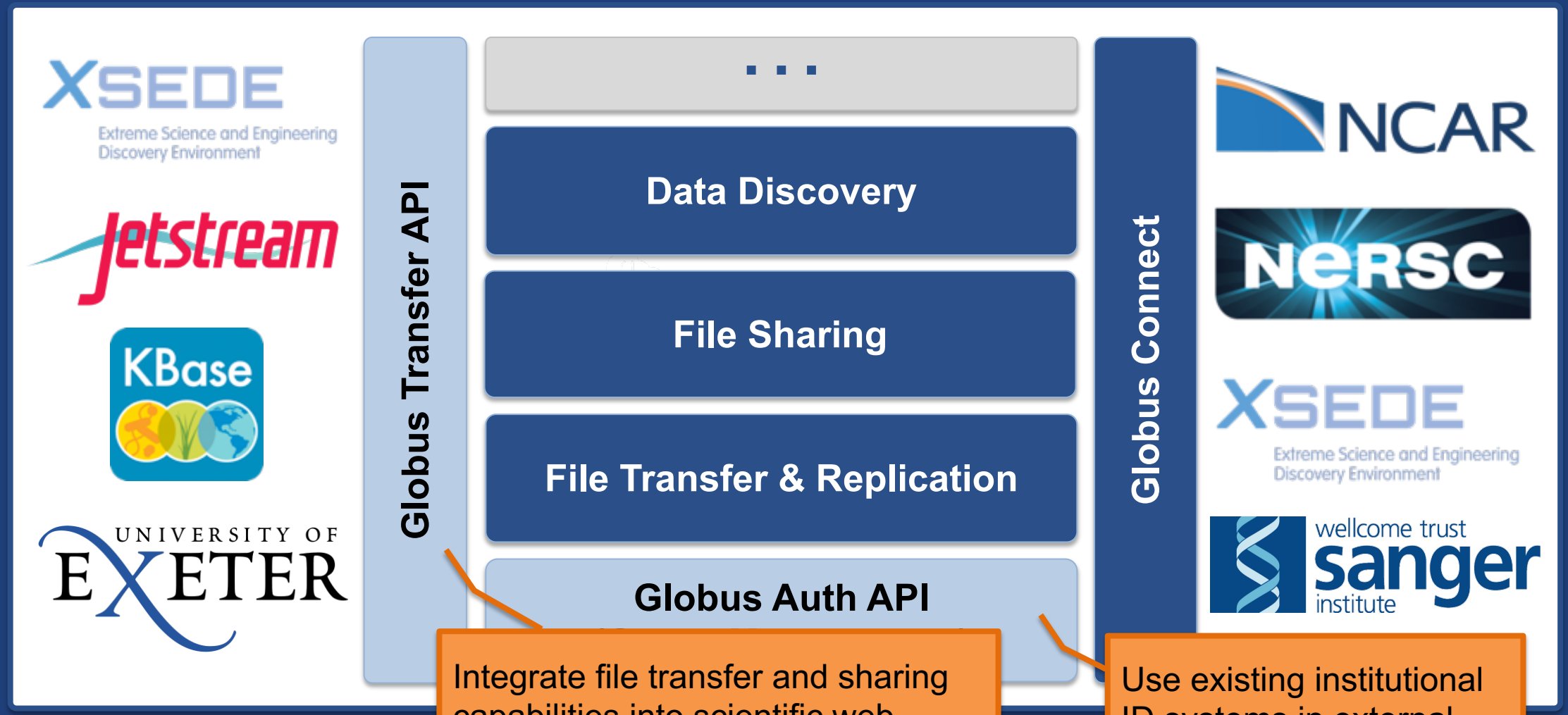
…via software-as-a-service using existing identities.

How can I integrate Globus into my research workflows?

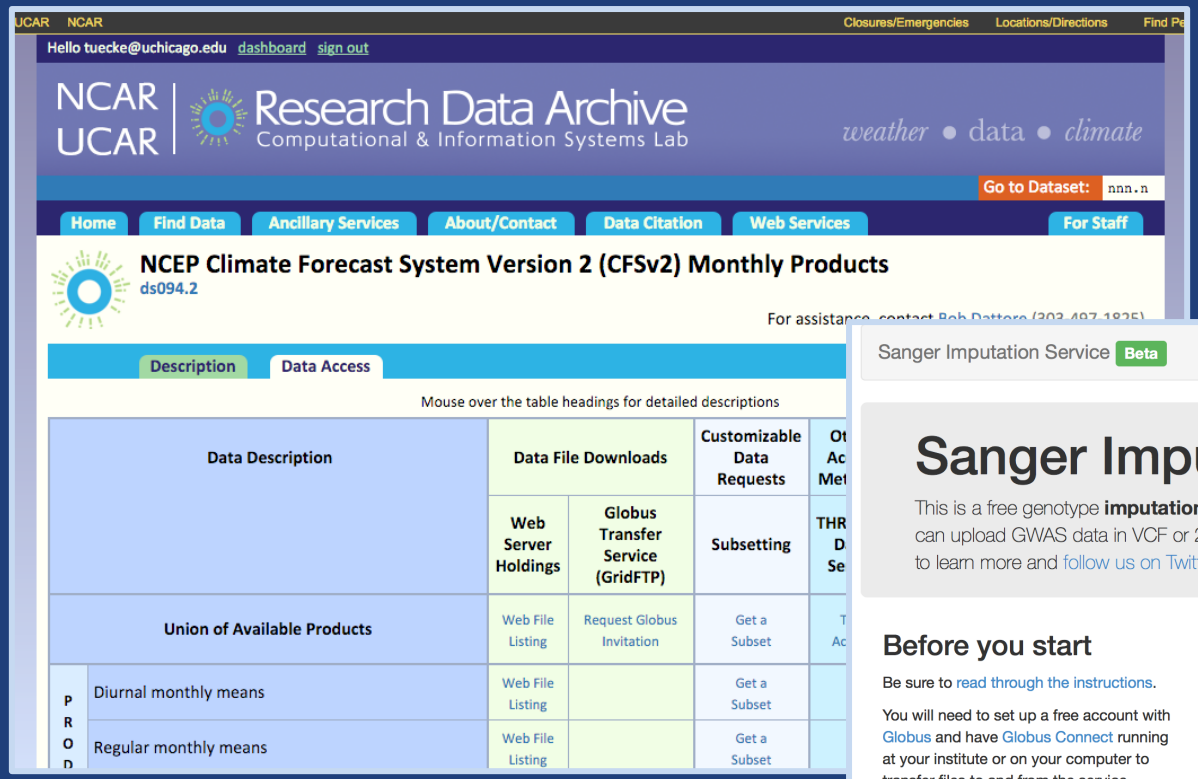
Globus serves as...

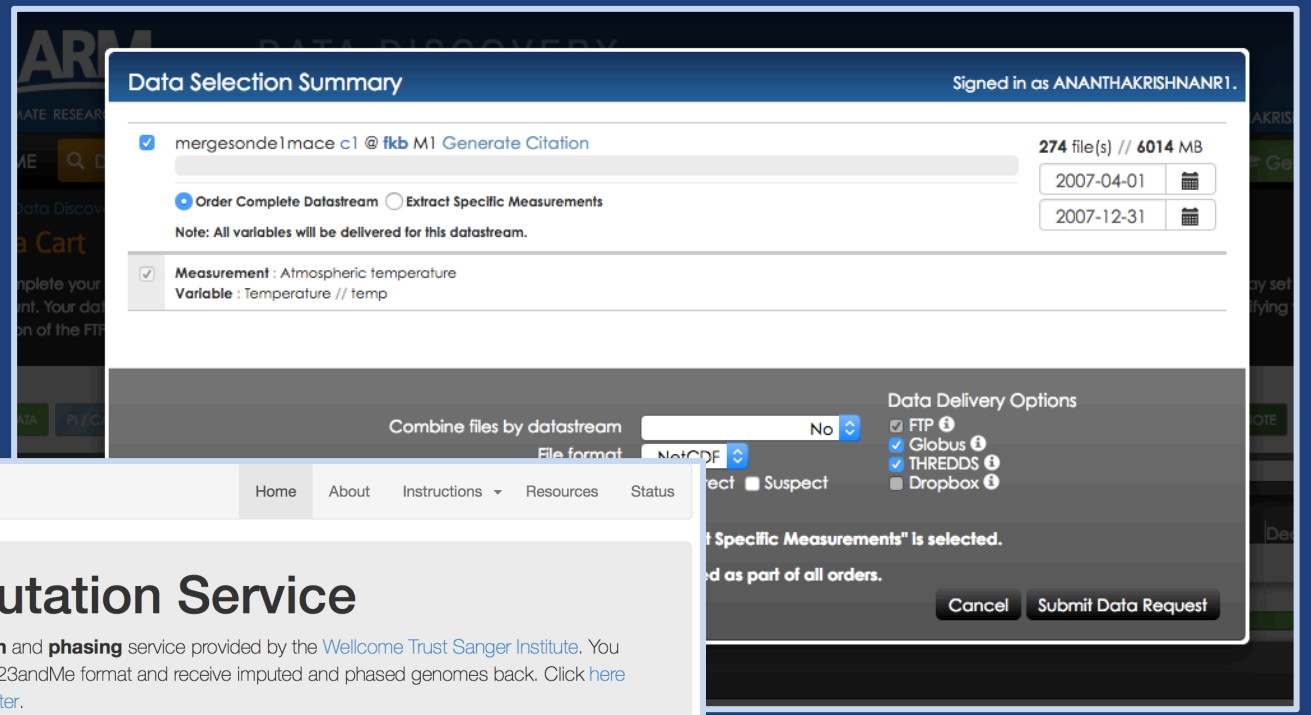
A platform for building science gateways, portals and other web applications in support of research and education.

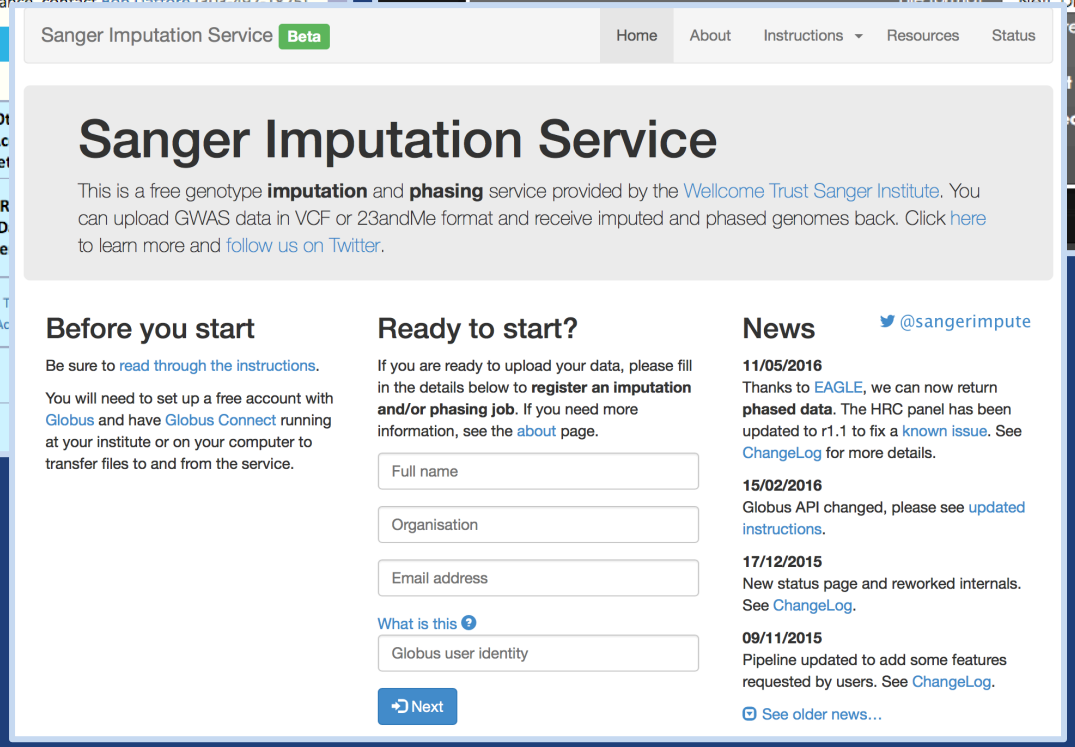
Globus Platform-as-a-Service



Example web apps that leverage Globus







Globus Transfer API Set

- **Doc**
  - https://docs.globus.org/api/transfer/
- **Sample data portal**
  - https://docs.globus.org/modern-research-data-portal/
- **Jupyter notebook**
  - https://github.com/globus/globus-jupyter-notebooks

Globus Auth API Set

- **Doc**
  - https://docs.globus.org/api/auth/
- **Sample data portal**
  - https://docs.globus.org/modern-research-data-portal/
- **Native app examples**
  - https://github.com/globus/native-app-examples

JupyterHub

jupyter.demo.globus.org

Enabling large-scale data intensive science with Jupyter

MDF: Advanced materials research

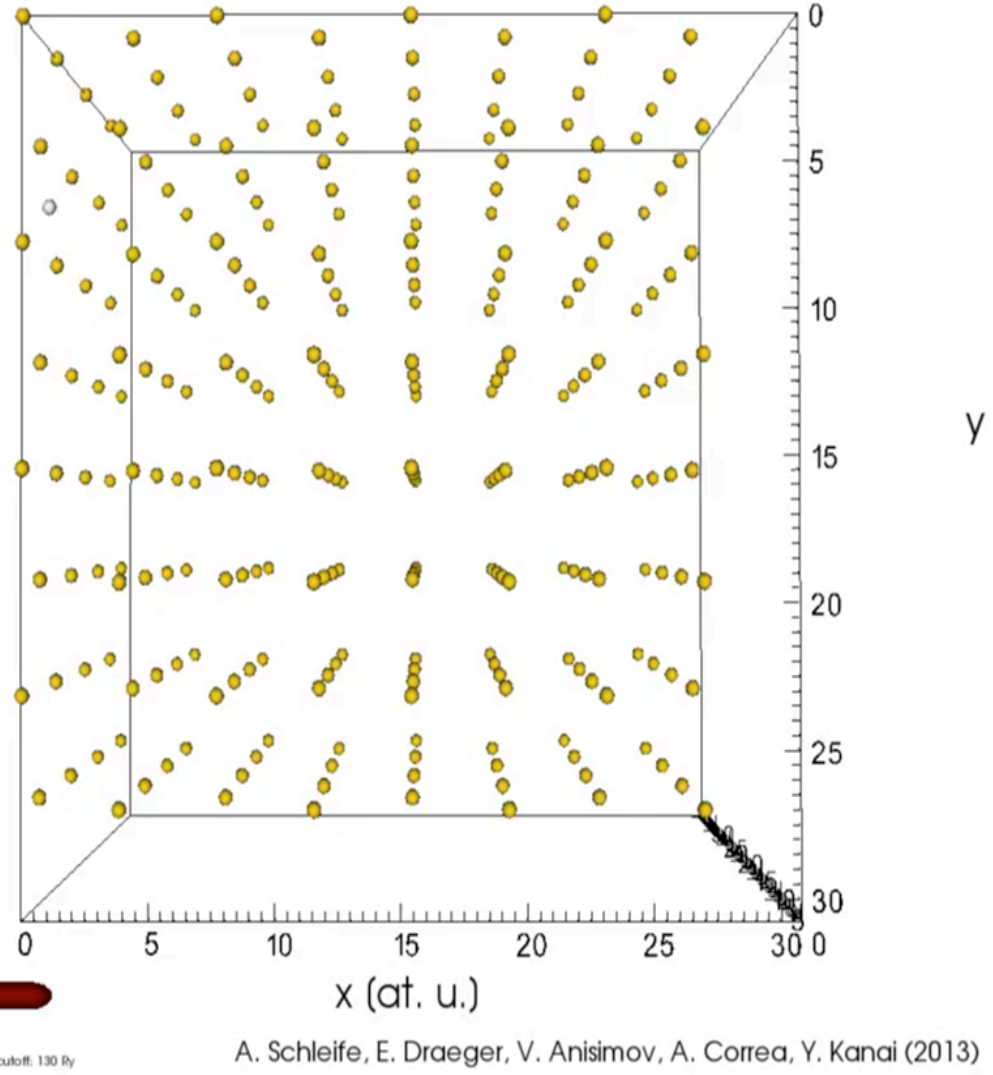
Modeling stopping power with time-dependent density functional theory

786,432 CPUs, 10 PFLOPS supercomputer
Argonne Leadership Computing Facility

Andre Schleife, UIUC

Hydrogen in Gold
(v=2.0 at. u.)

16,000 CPU-hours per simulation



A. Schleife, E. Draeger, V. Anisimov, A. Correa, Y. Kanai (2013)

Jupyter notebooks enable rapid iteration/results

In [35]:
```
@python_app
def get_stopping_power(lattice_vector, traj_computer):
    return traj_computer.compute_stopping_power([0,0.8,0.85], lattice_vector, 1.0, abserr=0.001,
                                                hit_threshold=2.5, full_output=1)
```

In [37]:
```
stopping_power_results = []
for d in tqdm(dirs, desc='Submitting'):
    stopping_power_results.append(get_stopping_power(d, traj_computer))
```

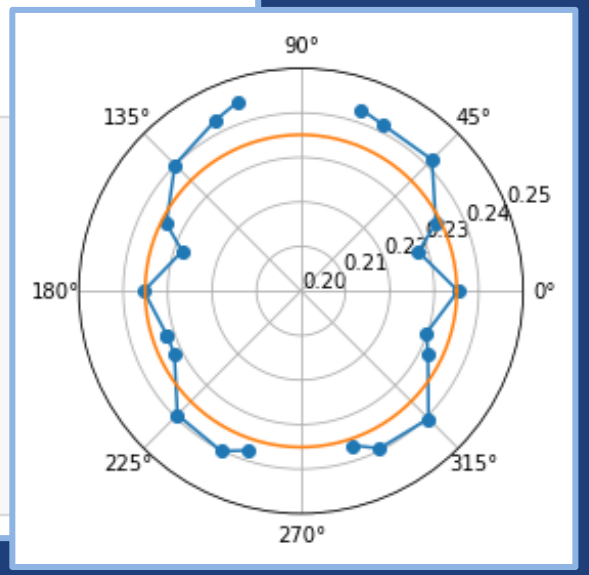
Submitting 100% 24/24 [00:00<00:00, 166.06it/s]

In [38]:
```
stopping_power_results = [s.result() for s in tqdm(stopping_power_results, desc='Waiting')]
```

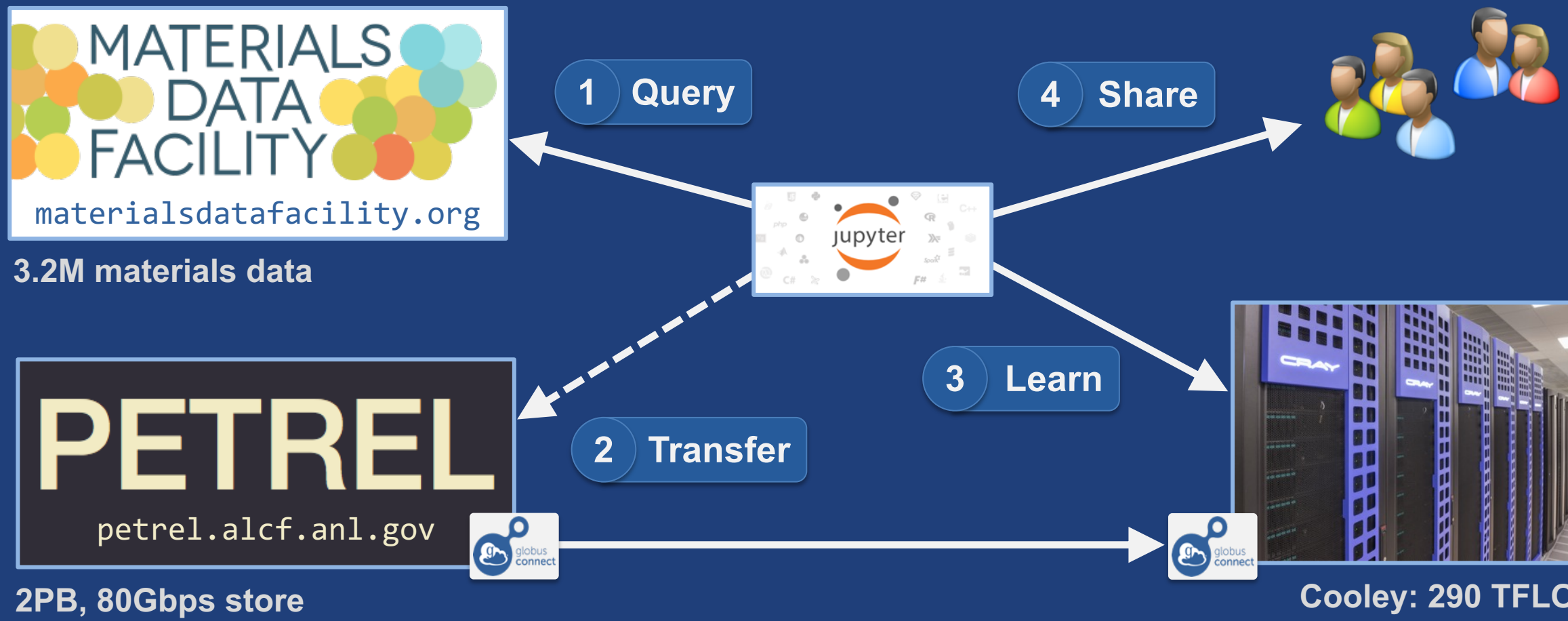
Waiting 100% 24/24 [18:47:19<00:00, 2818.33s/it]

In [62]:
```
ax = plt.subplot(111, projection='polar')
fig = plt.gcf()

ax.plot(angles + angles[:1], stopping_power + stopping_power[:1], marker='o')

# Plot the 'channel value'
ax.plot(np.linspace(0, 2*np.pi, 100), [ml_stopping_new,]*100)
ax.set_rmax(0.25)
ax.set_rmin(0.2)#min(stopping_power) * 0.99)

fig.set_size_inches(4, 4)
```

Logan Ward



But the data are big, distributed…
…and the science is collaborative



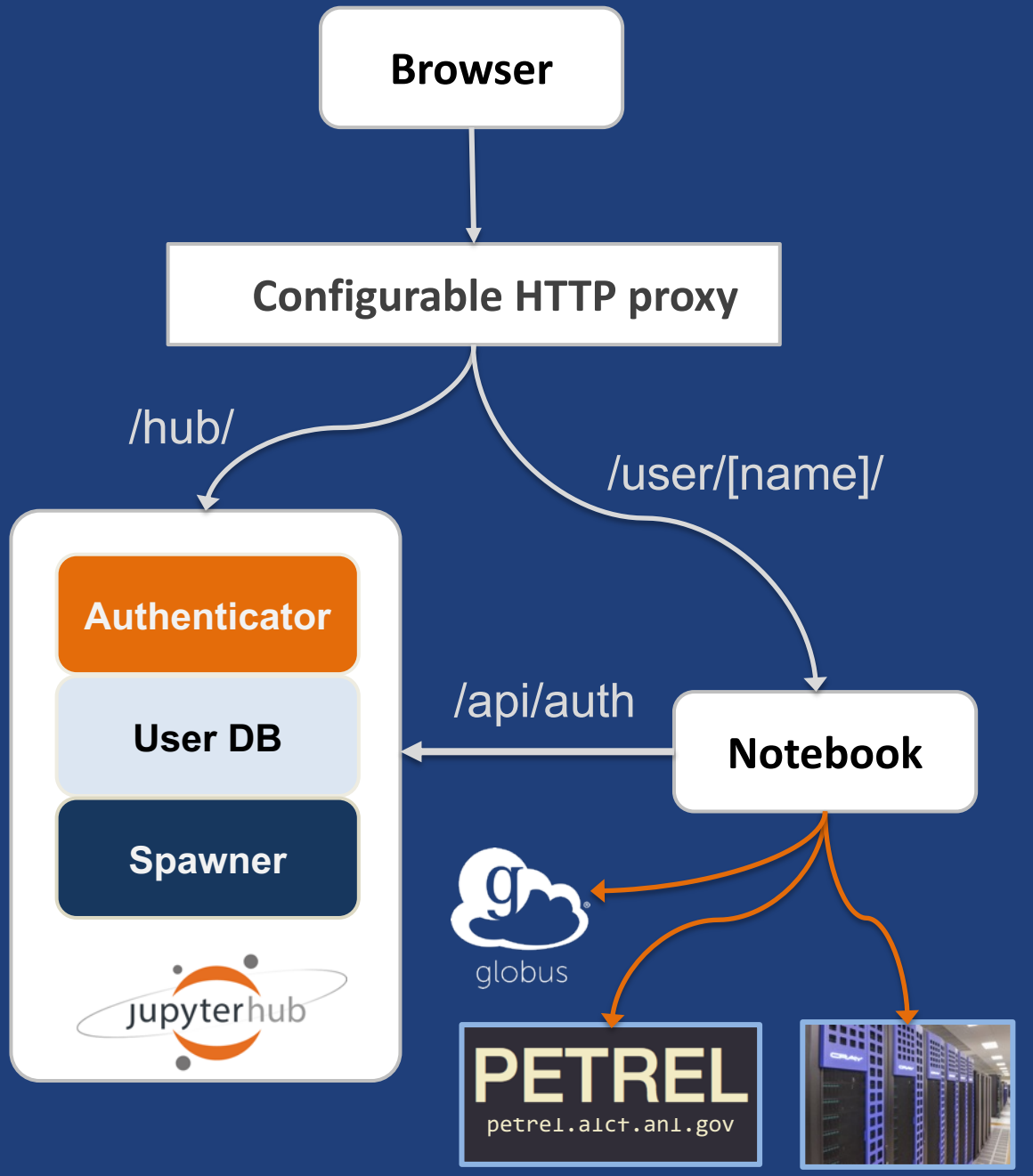
Need multi-credential, multi-service authentication and data management

JupyterHub

- **Multi-user hub**
- **Manages multiple instances of Jupyter notebook server**
- **Configurable HTTP proxy**

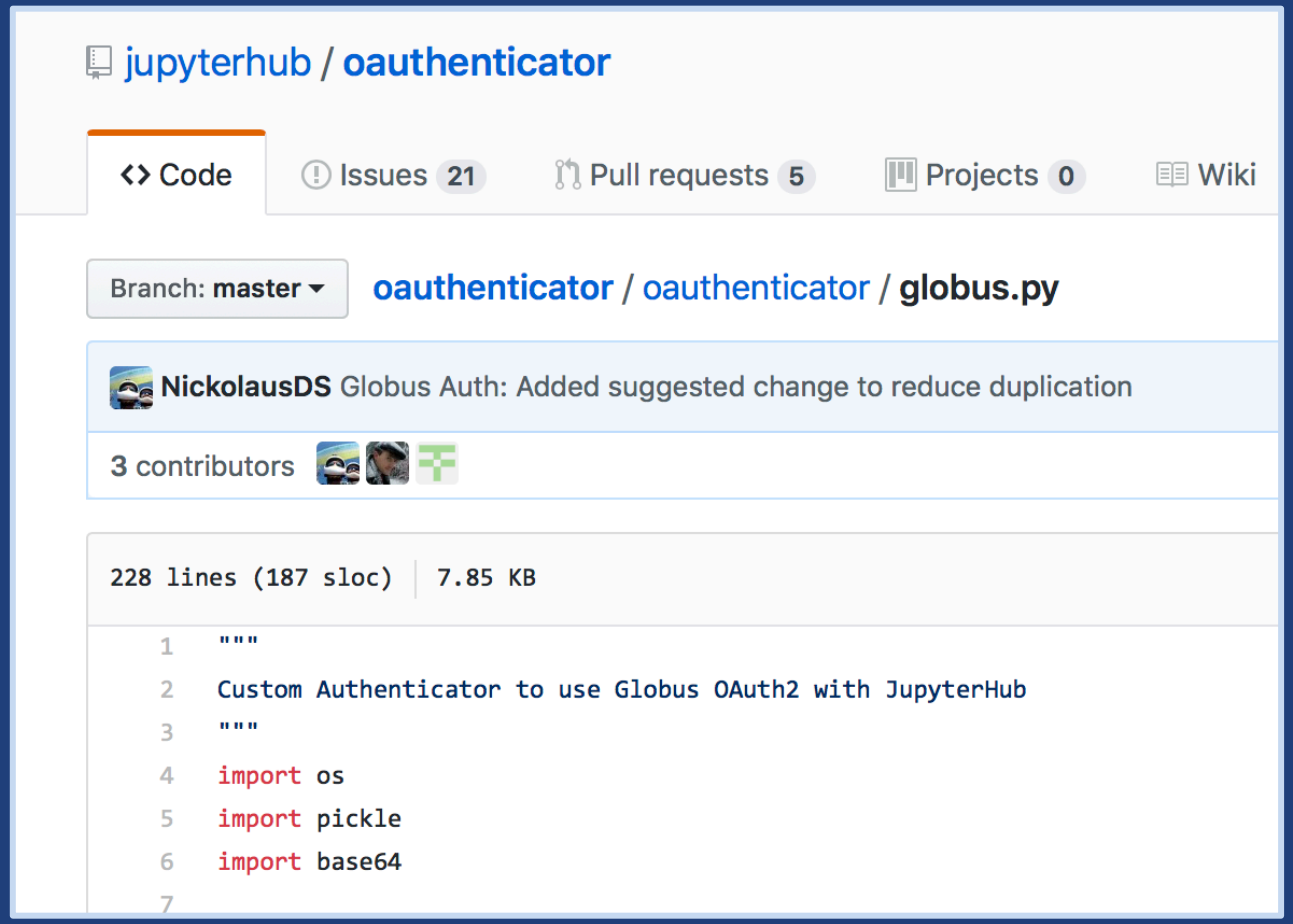
Goal: Liberate the notebook!

- **Tokens for remote services**
- **APIs for remote actions, e.g. data management via Globus service**



Securing JupyterHub with Globus Auth plugin

- **Existing OAuth framework**
- **Can restrict IdP**
- **Custom scopes**
- **Tokens passed into notebook environment**



github.com/jupyterhub/oauthenticator

Securing JupyterHub with Globus Auth

Visit https://developers.globus.org/ to set up your app. Ensure *Native App* is unchecked and make sure the callback URL looks like:

```
https://[your-host]/hub/oauth_callback
```

Set scopes for authorization and transfer. The defaults include:

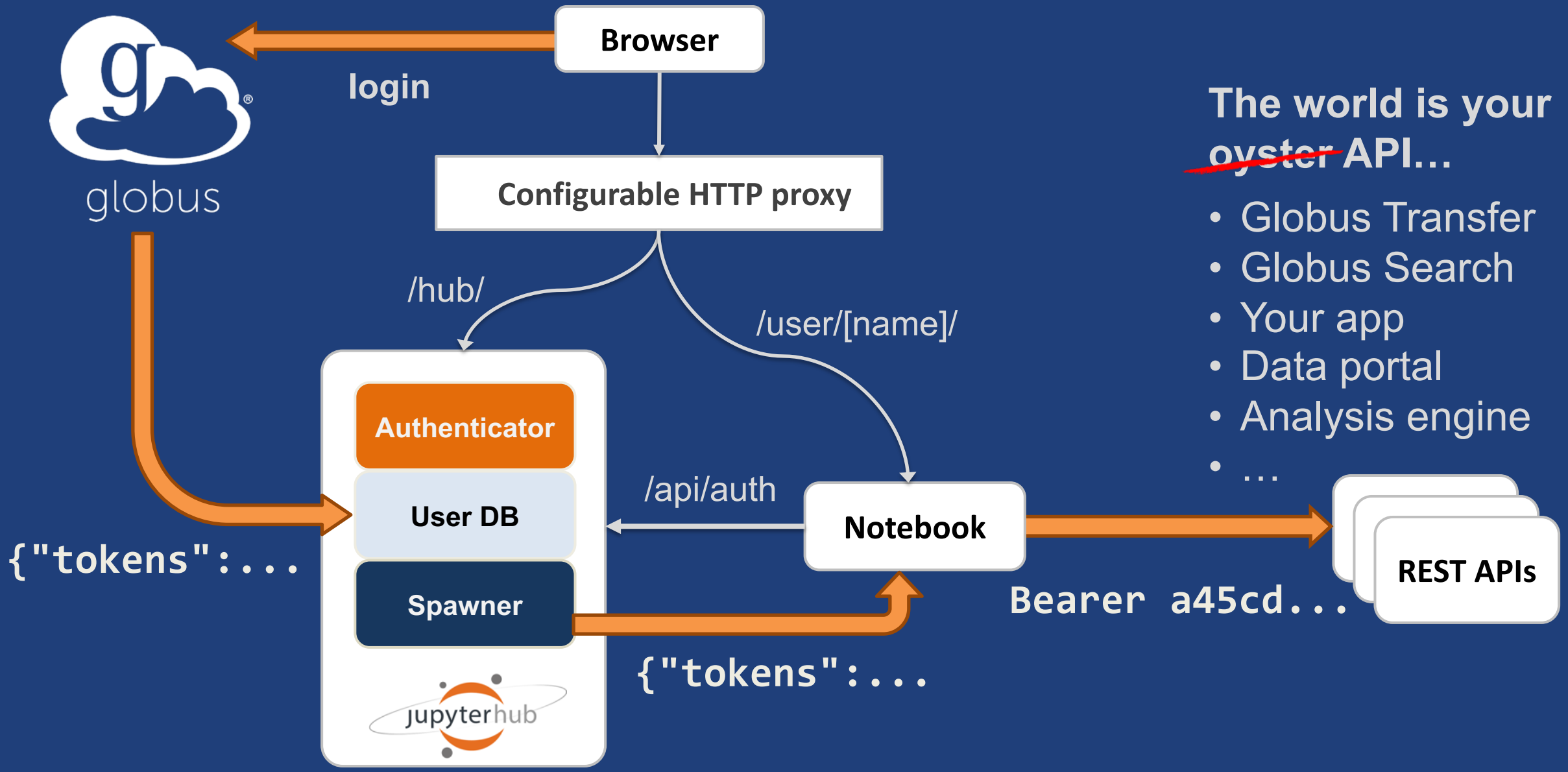
```
openid profile urn:globus:auth:scope:transfer.api.globus.org:all
```

Set the above settings in your `jupyterhub_config`:

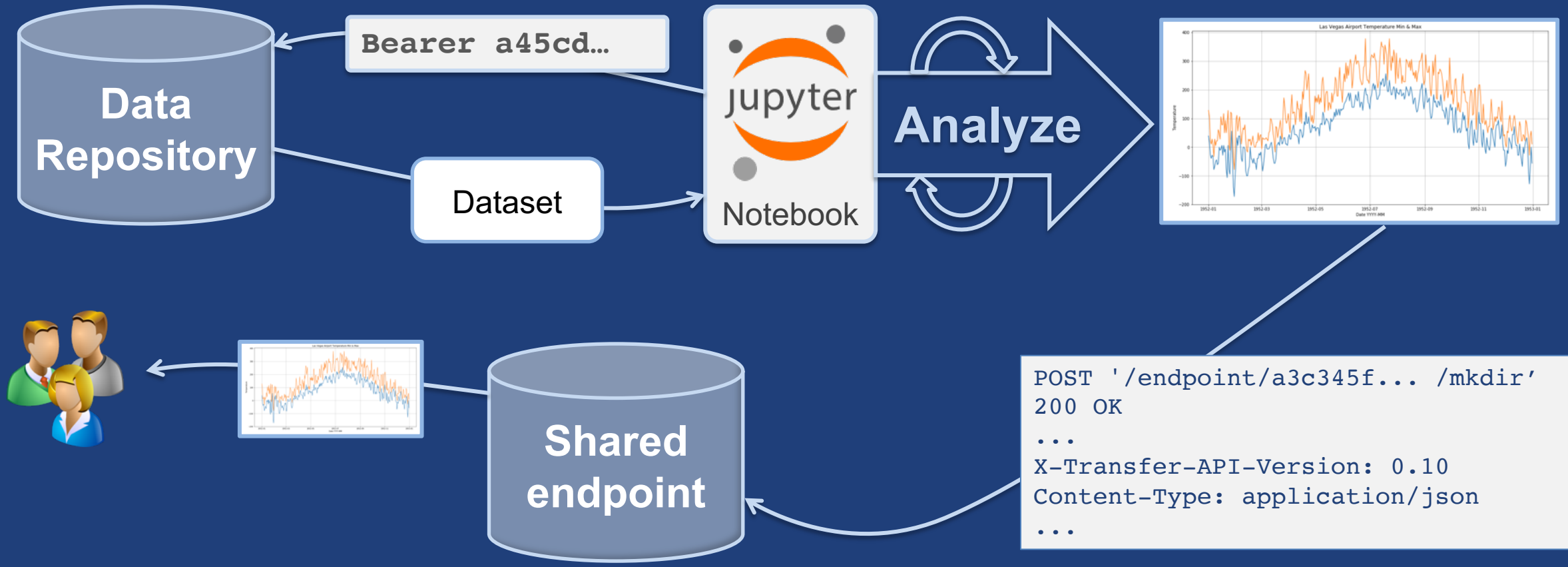
```
# Tell JupyterHub to create system accounts
from oauthenticator.globus import LocalGlobusOAuthenticator
c.JupyterHub.authenticator_class = LocalGlobusOAuthenticator
c.LocalGlobusOAuthenticator.enable_auth_state = True
c.LocalGlobusOAuthenticator.oauth_callback_url = 'https://[your-host]/hub/oauth_callback'
c.LocalGlobusOAuthenticator.client_id = '[your app client id]'
c.LocalGlobusOAuthenticator.client_secret = '[your app client secret]'
```

github.com/jupyterhub/oauthenticator#globus-setup

Tokens in Jupyter notebooks



Ad hoc data analysis/results distribution



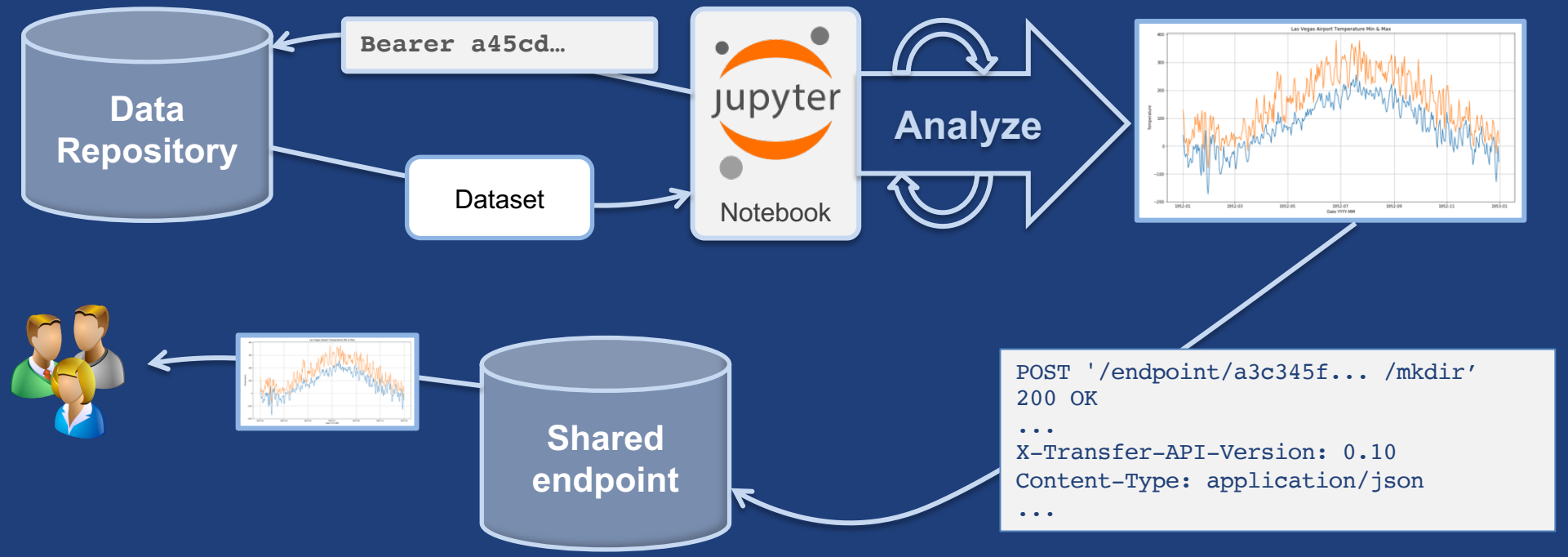
Experiment with the demo notebook

- **Login into our JupyterHub*: jupyter.demo.globus.org**
- **Launch (spawn) a notebook server; get tokens**
- **Using the `JupyterHub_Integration.ipynb` notebook:**
  - Access Globus APIs; download some data
  - “Analyze” data (generate plot)
  - PUT results (graph) on an HTTPS endpoint
  - Share the URL with others so they can access the results

*zero-to-jupyterhub.readthedocs.io

Leveraging the next generation of services

Our (simplistic) data flow thus far…



- Adequate for *ad hoc* sharing (implicit knowledge)
- Broader access, reuse requires “formalization”
- Leverage additional Globus platform services

Globus Search

- **Scalable service → billions of entries**
- **Schema agnostic: use standard (e.g. DataCite) or custom metadata**
- **Fine grained access control: only returns results that are visible to user**
- **Plain text search:** ranked results
- **Faceted search:** facilitates data discovery
- **Rich query language:** ranges, expressions, regex, etc.

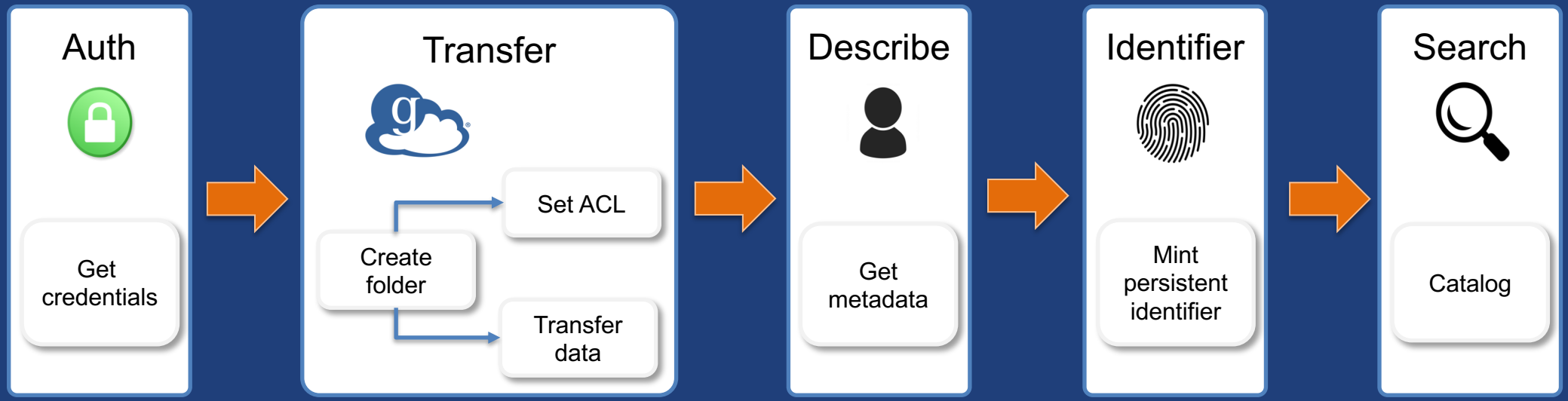
docs.globus.org/api/search

Persistent identifiers

- **Developing service for issuing persistent identifiers**
  - DOI, ARK, Handle, Globus
  - e.g. https://identifiers.globus.org/doi:10.1145/2076450.2076468
- **Within a namespace, e.g. your DataCite namespace**
  - Control which identities/groups can create identifiers
- **Identifier attributes:**
  - **Link to data**: one or more https URLs, to file, folder or manifest
  - **Landing page**: provided by service, or by user
  - **Visibility**: identities, groups that can see identifier
  - **Checksum**: of the file or manifest
  - **Metadata**: as required by identifier (e.g., DataCite), extensible
  - **Replaces/replaced-by**: for versioning

Extending the automation flow

- **How can we enable more structured/robust data discovery using Globus platform services?**



Other Globus integrations

globus-integration-examples.readthedocs.io

- Web app development frameworks (Flask, Django)
- Content management systems (WordPress, Drupal)
- Development tools (Confluence, Jira)
- Scalable cyberinfrastructure (Kubernetes)

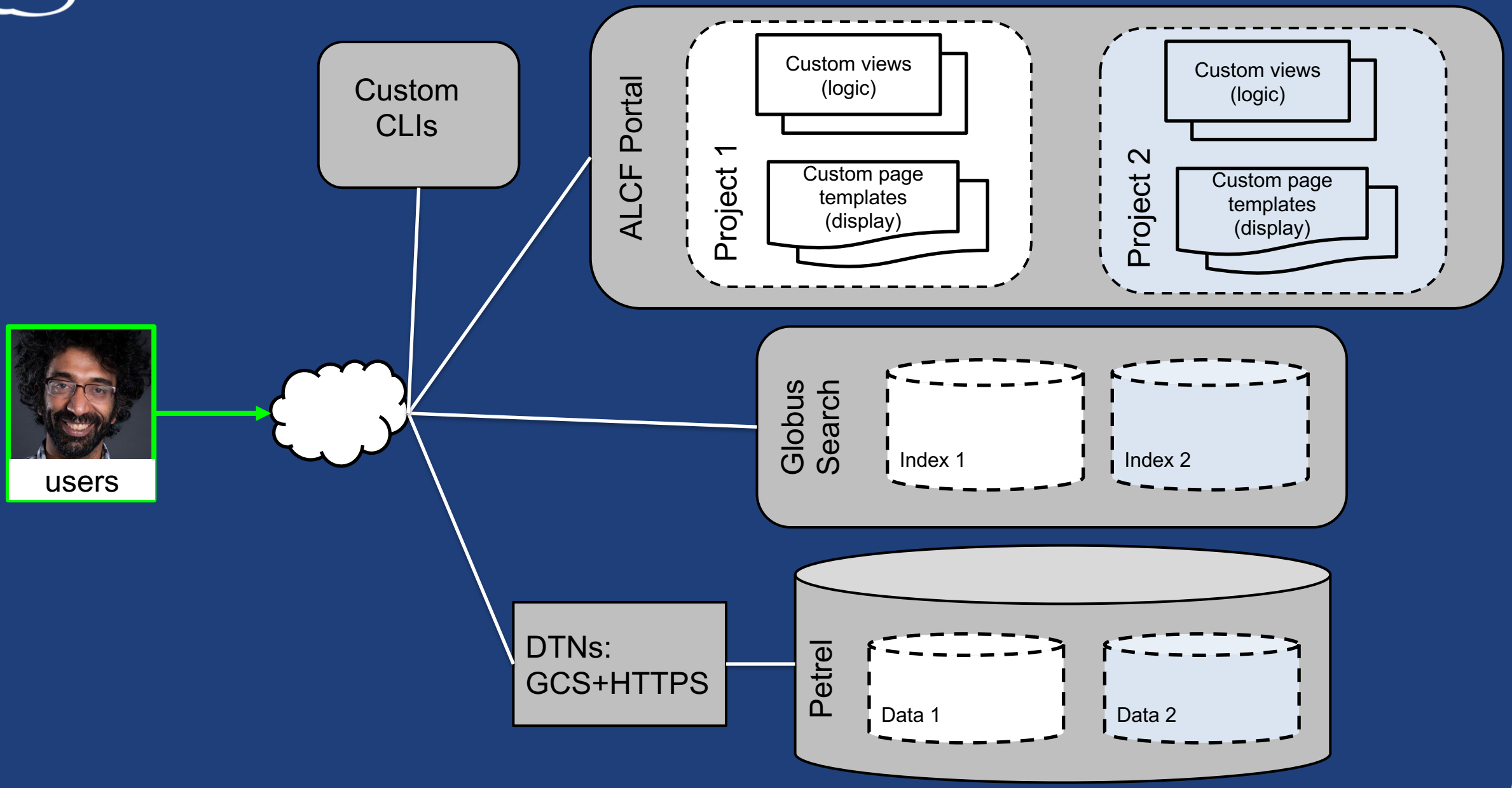
- Genomics analysis (Galaxy)
  - galaxyproject.org/authnz/use/oidc/idps/globus

Example

ALCF Data Discovery Portal

https://petreldata.net

ALCF Project Portals



A bit of Globus history

U.S. DEPARTMENT OF
ENERGY

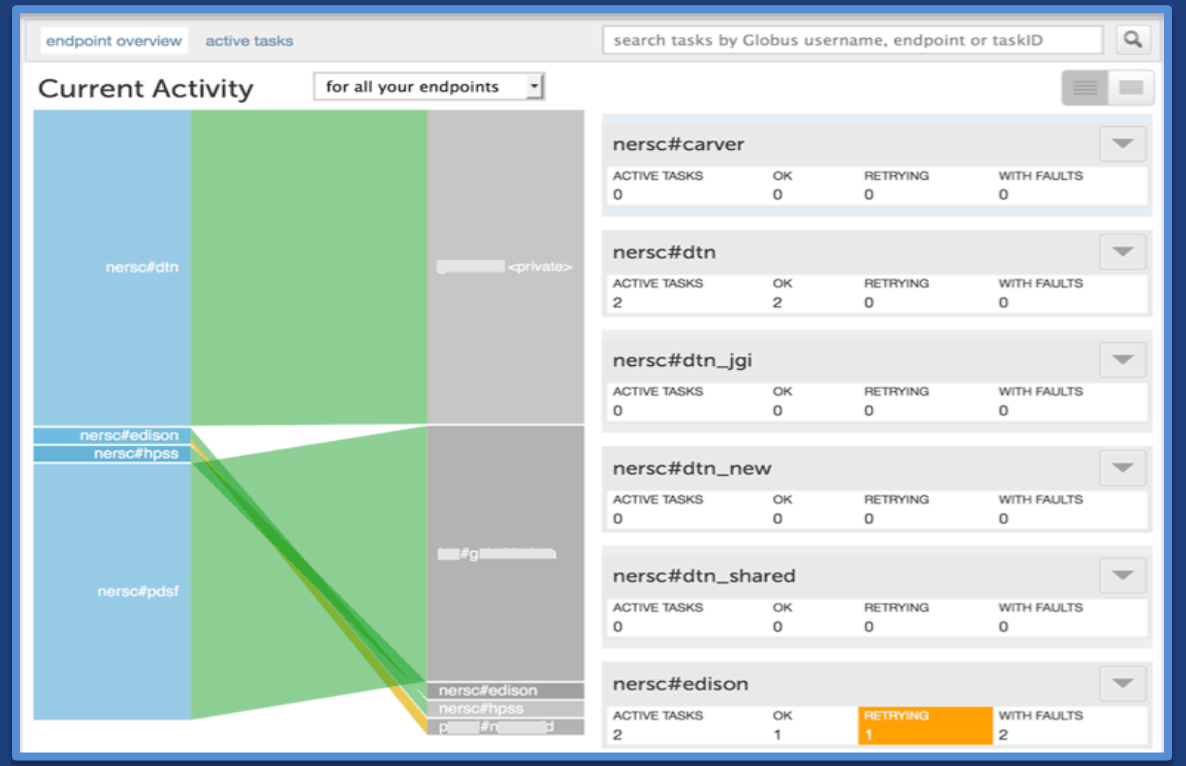
THE UNIVERSITY OF
CHICAGO

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

Argonne
NATIONAL LABORATORY

Globus sustainability model

- **Standard Subscription**
  - Shared endpoints
  - Management console
  - Usage reporting
  - Priority support
  - Application integration
  - HTTPS support (coming soon)
- **Branded Web Site**
- **Premium Storage Connectors**
- **Alternate Identity Provider (InCommon is standard)**



The path to sustainability

Globus by the numbers

1,042
most shared endpoints at a single institution

635+ PB
transferred

73 billion
files processed

1,700
active GCS endpoints

100+
subscribers

100,000+
users

3 months
longest running transfer

18,000
active GCP endpoints

500+
identity providers

1 PB
largest single transfer to date

8,000
active shared endpoints

99.9%
availability

Globus support resources

- **Globus documentation: docs.globus.org**
- **Helpdesk and issue escalation: support@globus.org**
- **Mailing Lists**
  - https://www.globus.org/mailing-lists
- **Customer engagement team**
- **Globus professional services team**
  - Assist with portal/gateway/app architecture and design
  - Develop custom applications that leverage the Globus platform
  - Advise on customized deployment and integration scenarios