Introduction to Globus for ALCF Users
Research Data Transfer

2021 ALCF Computational Performance Workshop
May 5, 2021

Greg Nawrocki
greg@globus.org
nawrocki@uchicago.edu

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

globus

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

Development is funded by...

U.S. DEPARTMENT OF ENERGY

NSF

THE UNIVERSITY OF CHICAGO

NIH

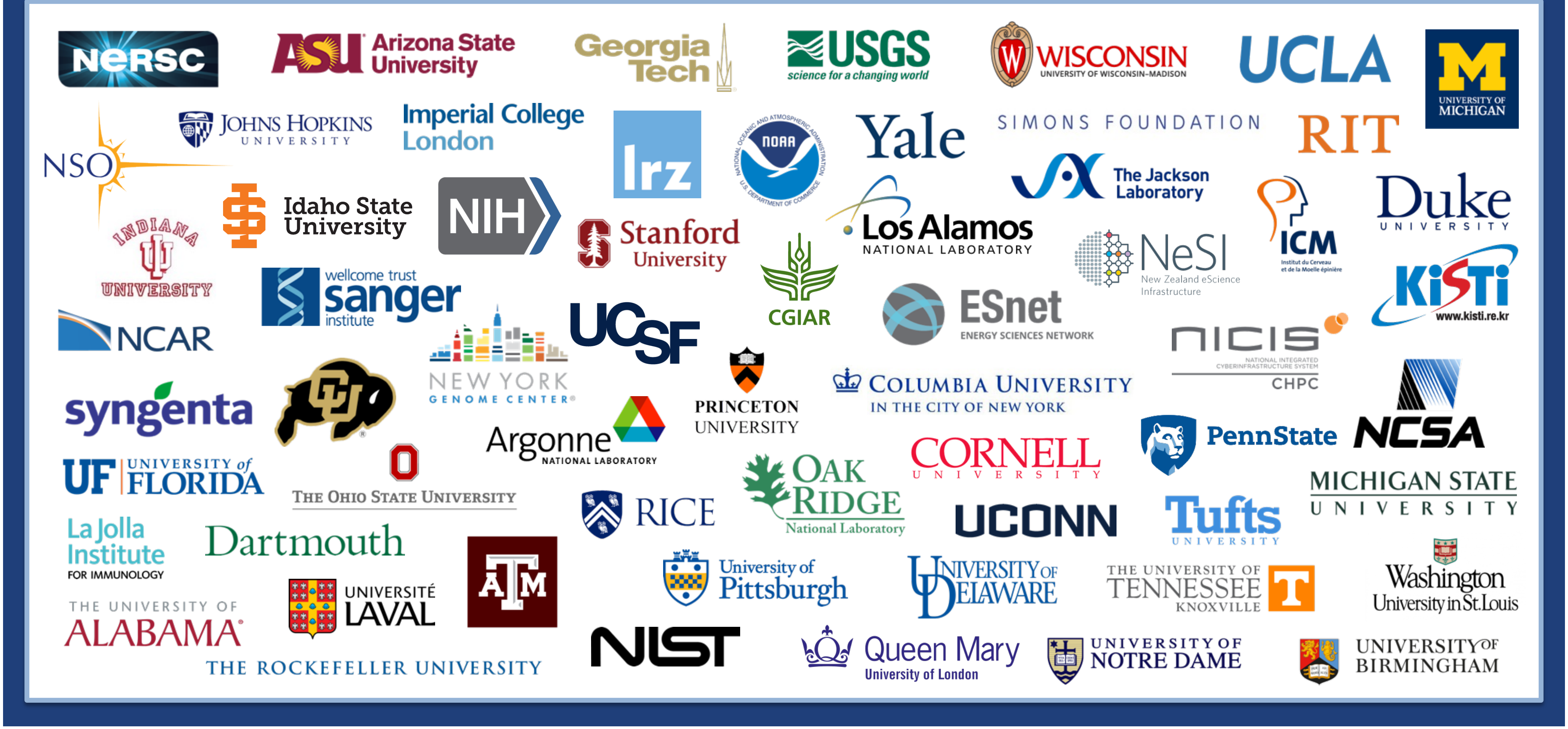
powered by amazon web services

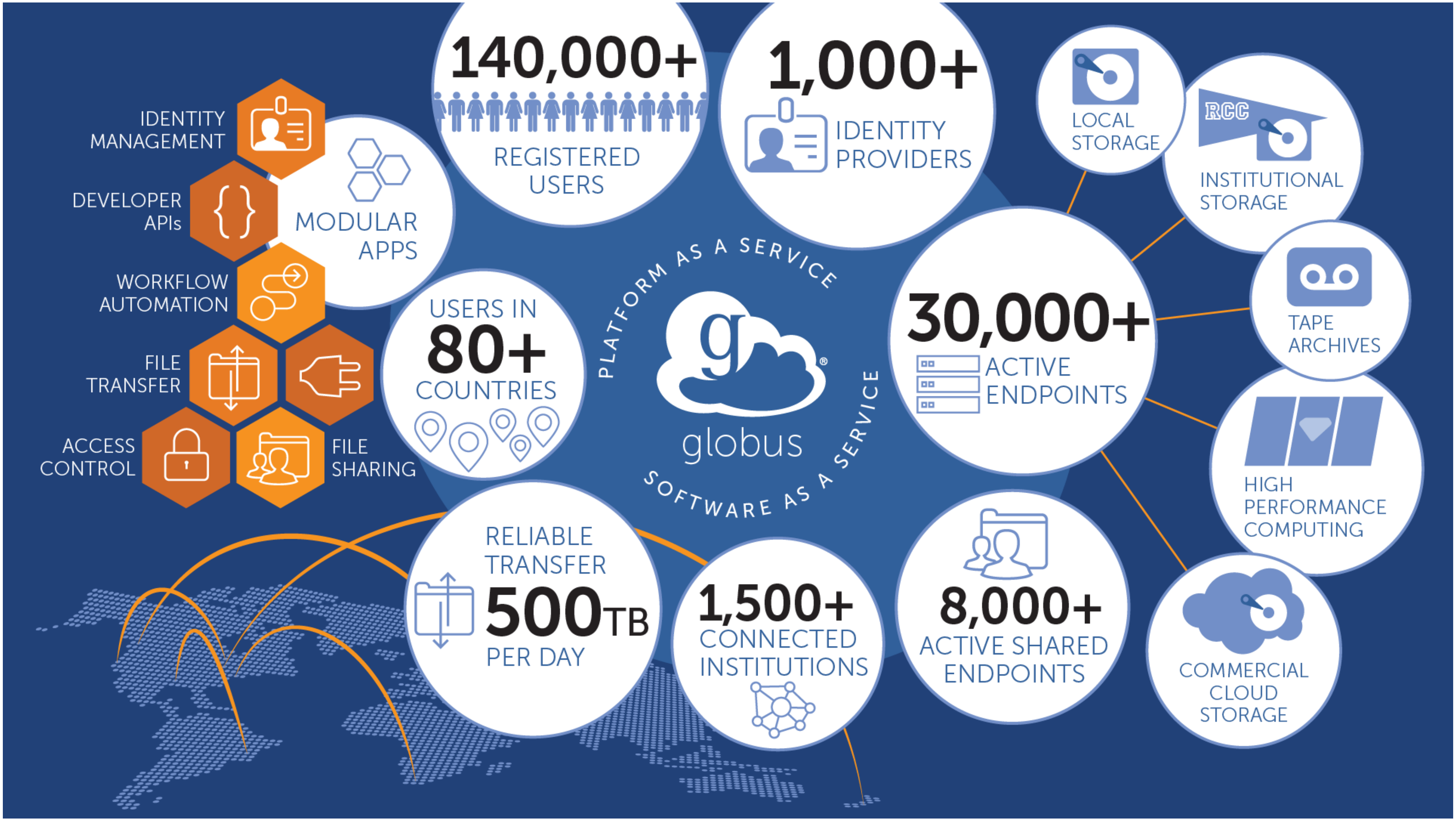
ALFRED P. SLOAN FOUNDATION 1934

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

Argonne NATIONAL LABORATORY

Operations are funded by subscribers


NERSC
ASU Arizona State University
Georgia Tech
USGS science for a changing world
W WISCONSIN UNIVERSITY OF WISCONSIN-MADISON
UCLA
M UNIVERSITY OF MICHIGAN
NSO
JOHNS HOPKINS UNIVERSITY
Imperial College London
lrz
NOAA NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION U.S. DEPARTMENT OF COMMERCE
Yale
SIMONS FOUNDATION
RIT
The Jackson Laboratory
Idaho State University
NIH
Stanford University
Los Alamos NATIONAL LABORATORY
ICM Institut du Cerveau et de la Moelle épinière
Duke UNIVERSITY
INDIANA UNIVERSITY
NeSI New Zealand eScience Infrastructure
wellcome trust sanger institute
CGIAR
ESnet ENERGY SCIENCES NETWORK
KiSTi www.kisti.re.kr
NCAR
UCSF
NEW YORK GENOME CENTER
NICIS NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM CHPC
syngenta
PRINCETON UNIVERSITY
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
Argonne NATIONAL LABORATORY
PennState
NCSA
UF UNIVERSITY of FLORIDA
O THE OHIO STATE UNIVERSITY
CORNELL UNIVERSITY
MICHIGAN STATE UNIVERSITY
RICE
OAK RIDGE National Laboratory
UCONN
Tufts UNIVERSITY
La Jolla Institute FOR IMMUNOLOGY
Dartmouth
ATM
University of Pittsburgh
UNIVERSITY OF DELAWARE
THE UNIVERSITY OF TENNESSEE KNOXVILLE
Washington University in St.Louis
THE UNIVERSITY OF ALABAMA
UNIVERSITÉ LAVAL
NIST
Queen Mary University of London
UNIVERSITY OF NOTRE DAME
UNIVERSITY OF BIRMINGHAM
THE ROCKEFELLER UNIVERSITY



IDENTITY MANAGEMENT

DEVELOPER APIs

WORKFLOW AUTOMATION

FILE TRANSFER

ACCESS CONTROL

FILE SHARING

MODULAR APPS

140,000+ REGISTERED USERS

1,000+ IDENTITY PROVIDERS

USERS IN 80+ COUNTRIES

PLATFORM AS A SERVICE

globus

SOFTWARE AS A SERVICE

RELIABLE TRANSFER 500TB PER DAY

1,500+ CONNECTED INSTITUTIONS

30,000+ ACTIVE ENDPOINTS

8,000+ ACTIVE SHARED ENDPOINTS

LOCAL STORAGE

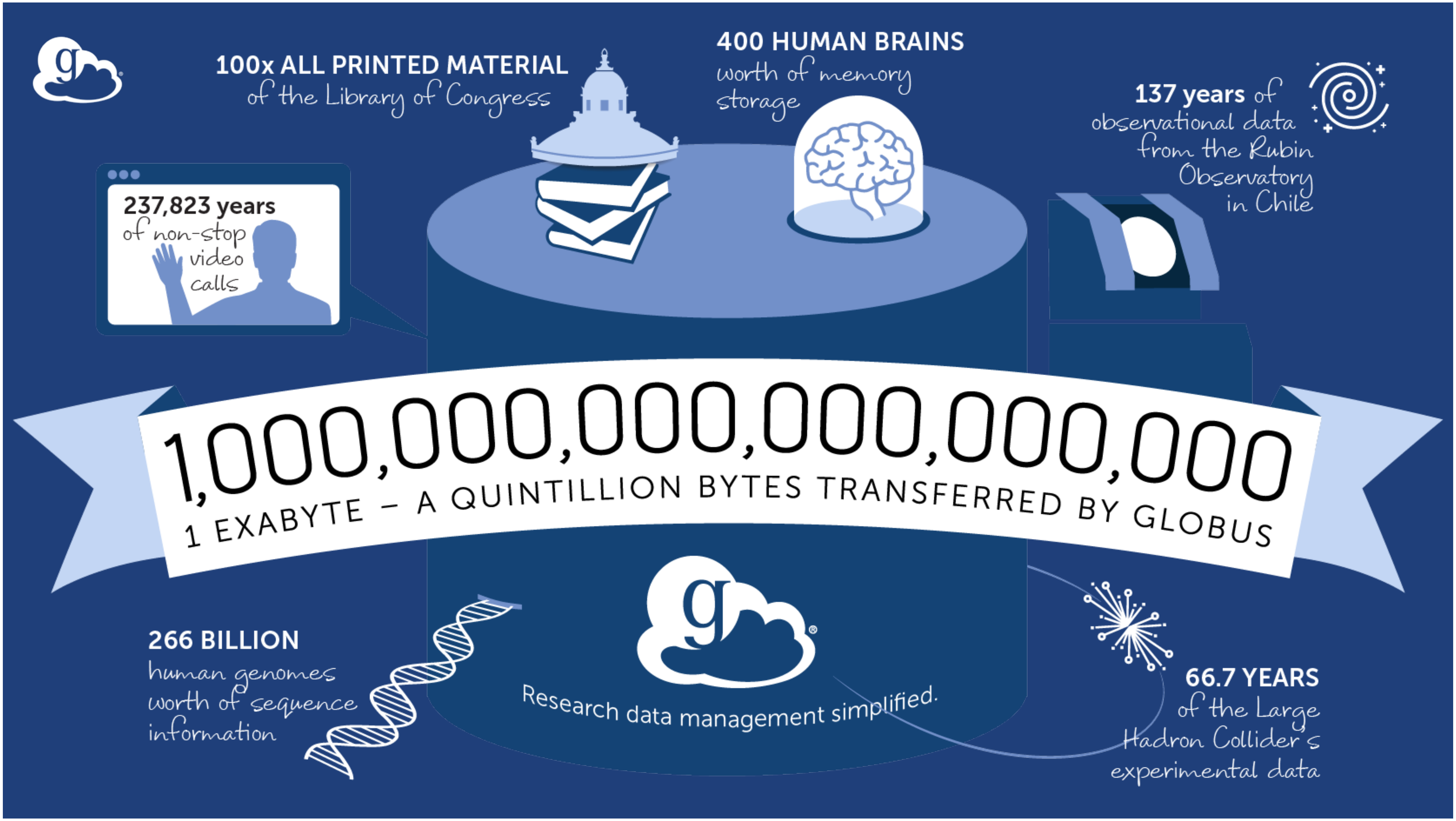
RCC

INSTITUTIONAL STORAGE

TAPE ARCHIVES

HIGH PERFORMANCE COMPUTING

COMMERCIAL CLOUD STORAGE



100x ALL PRINTED MATERIAL
of the Library of Congress

400 HUMAN BRAINS
worth of memory storage

137 years of observational data from the Rubin Observatory in Chile

237,823 years
of non-stop video calls

1,000,000,000,000,000,000

1 EXABYTE – A QUINTILLION BYTES TRANSFERRED BY GLOBUS

266 BILLION
human genomes worth of sequence information

Research data management simplified.

66.7 YEARS
of the Large Hadron Collider's experimental data

Key capabilities

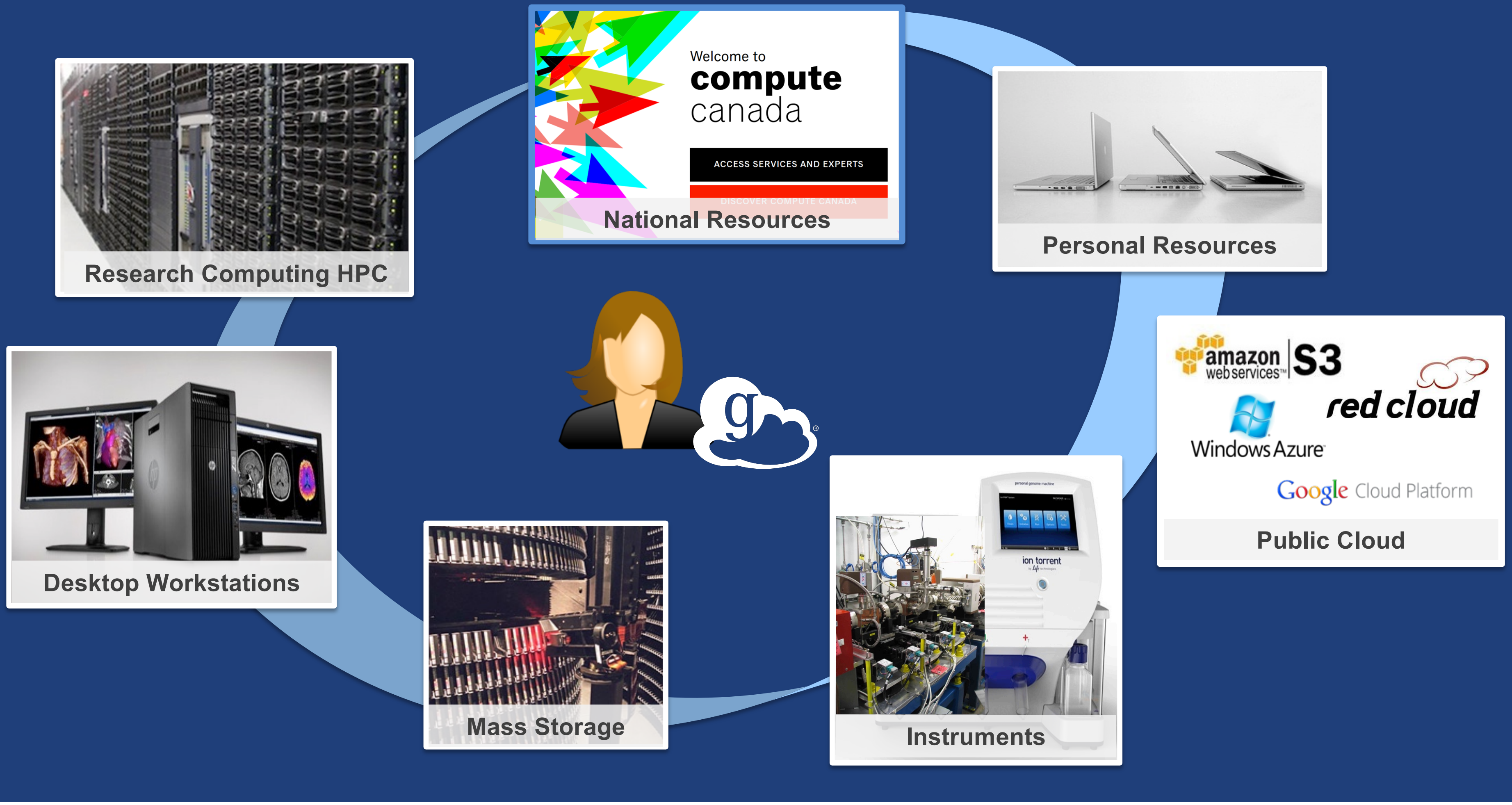
Globus delivers…

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



National Resources

Research Computing HPC

Personal Resources

Public Cloud

Desktop Workstations

Mass Storage

Instruments

Globus Connectors

amazon web services™ S3

Google Cloud

SPECTRA

IBM Cloud Object Storage

box

Caringo Swarm™

Quantum

ActiveScale Object Storage

iRODS®

ceph

Google Drive

SCALITY

IBM Spectrum Scale

HPSS

wasabi™ hot cloud storage

lustre™

OneDrive

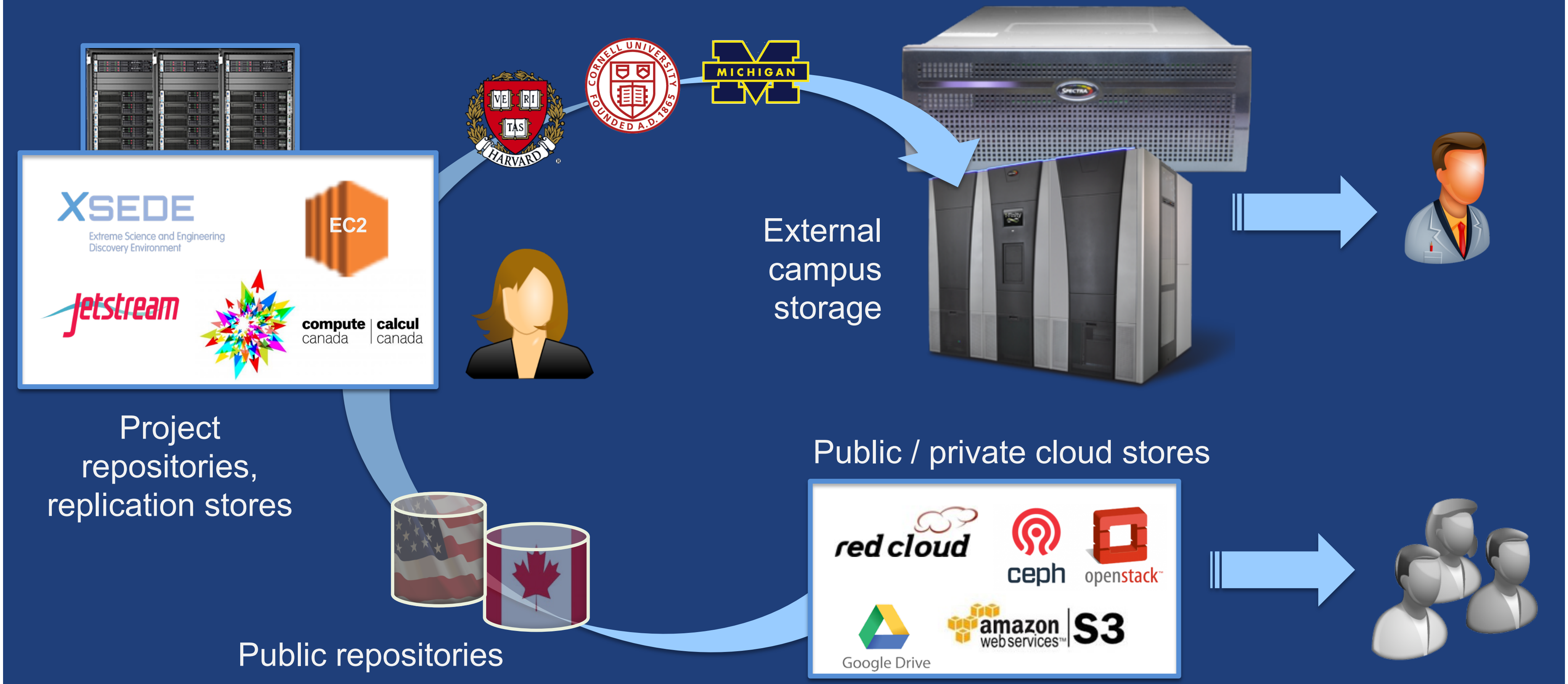
hadoop HDFS

Planned

Microsoft Azure Blob Storage

Dropbox

Share with collaborators/community



External campus storage

Project repositories, replication stores

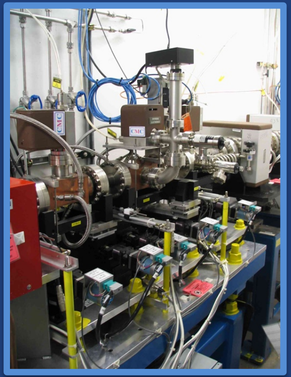
Public repositories

Public / private cloud stores

VE RI TAS
HARVARD
CORNELL UNIVERSITY FOUNDED A.D. 1865
MICHIGAN
XSEDE
Extreme Science and Engineering Discovery Environment
EC2
Jetstream
compute canada | calcul canada
red cloud
ceph
openstack
Google Drive
amazon web services
S3

Manage data from instruments

Next-Gen Sequencer



Advanced Light Source

Cryo-EM

MRI

Light Sheet Microscope

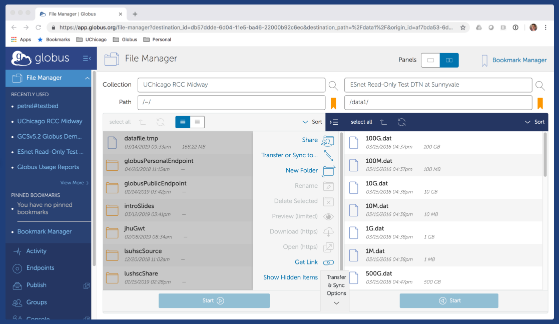
Analysis store

High-durability, low-cost store

Remote visualization

Personal system

Use(r)-appropriate interfaces



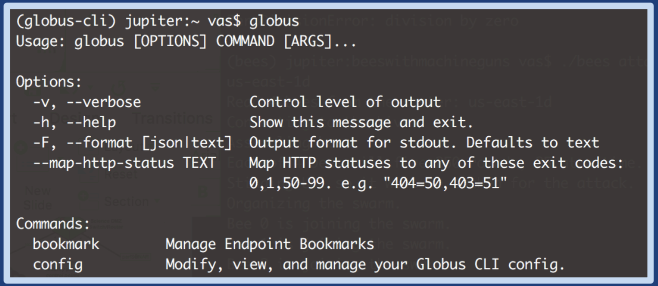
Collection: UChicago RCC Midway — ESnet Read-Only Test DTN at Sunnyvale

Path: /~/ — /data1/

datafile.tmp 03/14/2019 09:33am 168.22 MB
globusPersonalEndpoint 04/26/2018 11:15am —
globusPublicEndpoint 01/14/2019 03:42pm —
introSlides 03/12/2019 03:41pm —
jhuGwt 02/08/2019 08:34am —
lsuhscSource 12/20/2018 11:02am —
lushscShare 01/15/2019 02:28pm —

100G.dat 03/15/2016 04:37pm 100 GB
100M.dat 03/15/2016 04:37pm 100 MB
10G.dat 03/15/2016 04:38pm 10 GB
10M.dat 03/15/2016 04:38pm 10 MB
1G.dat 03/15/2016 04:38pm 1 GB
1M.dat 03/15/2016 04:38pm 1 MB
500G.dat 03/15/2016 04:47pm 500 GB

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.
```

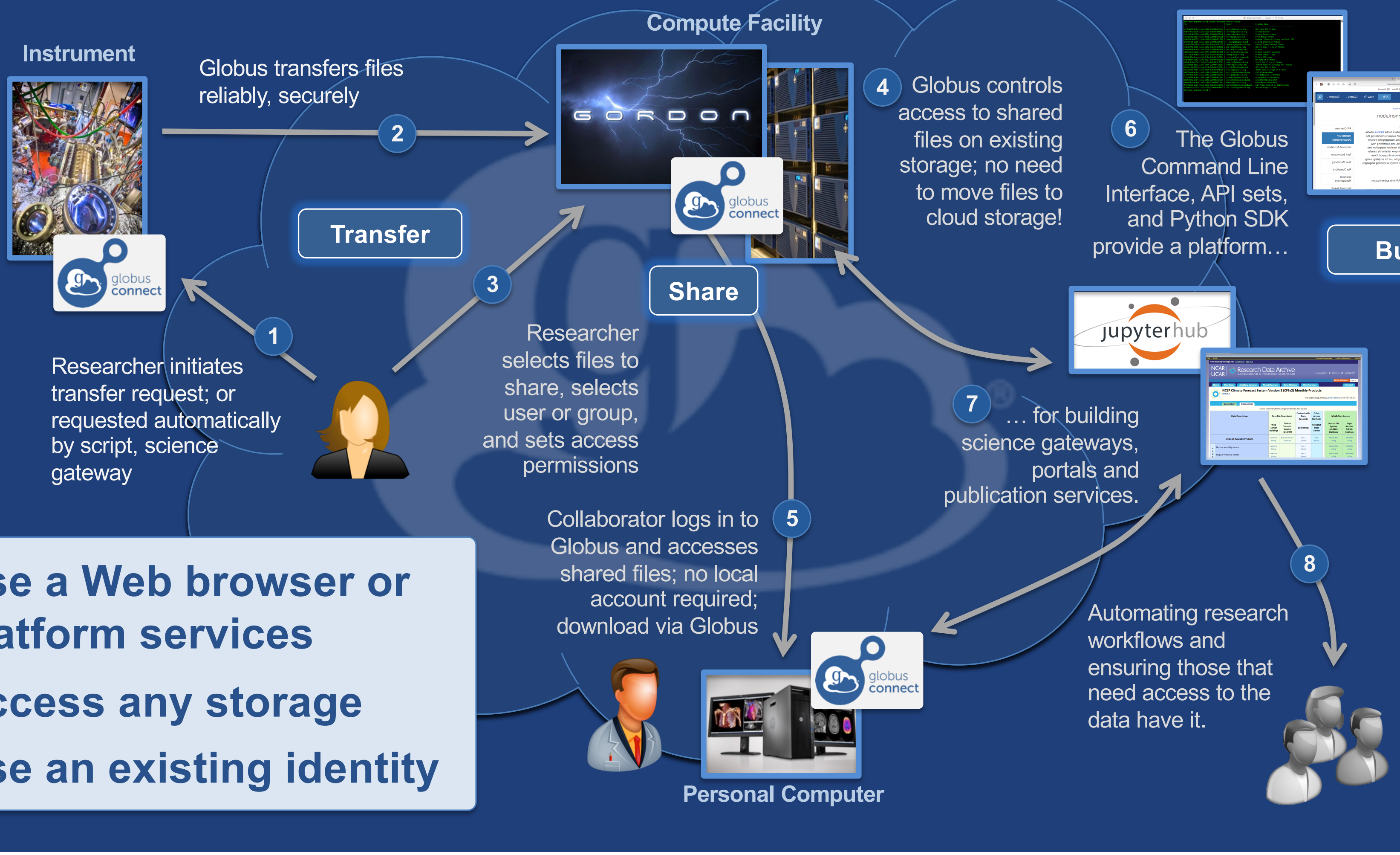
CLI

Globus service

```
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



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

2 Globus transfers files reliably, securely

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

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

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

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

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

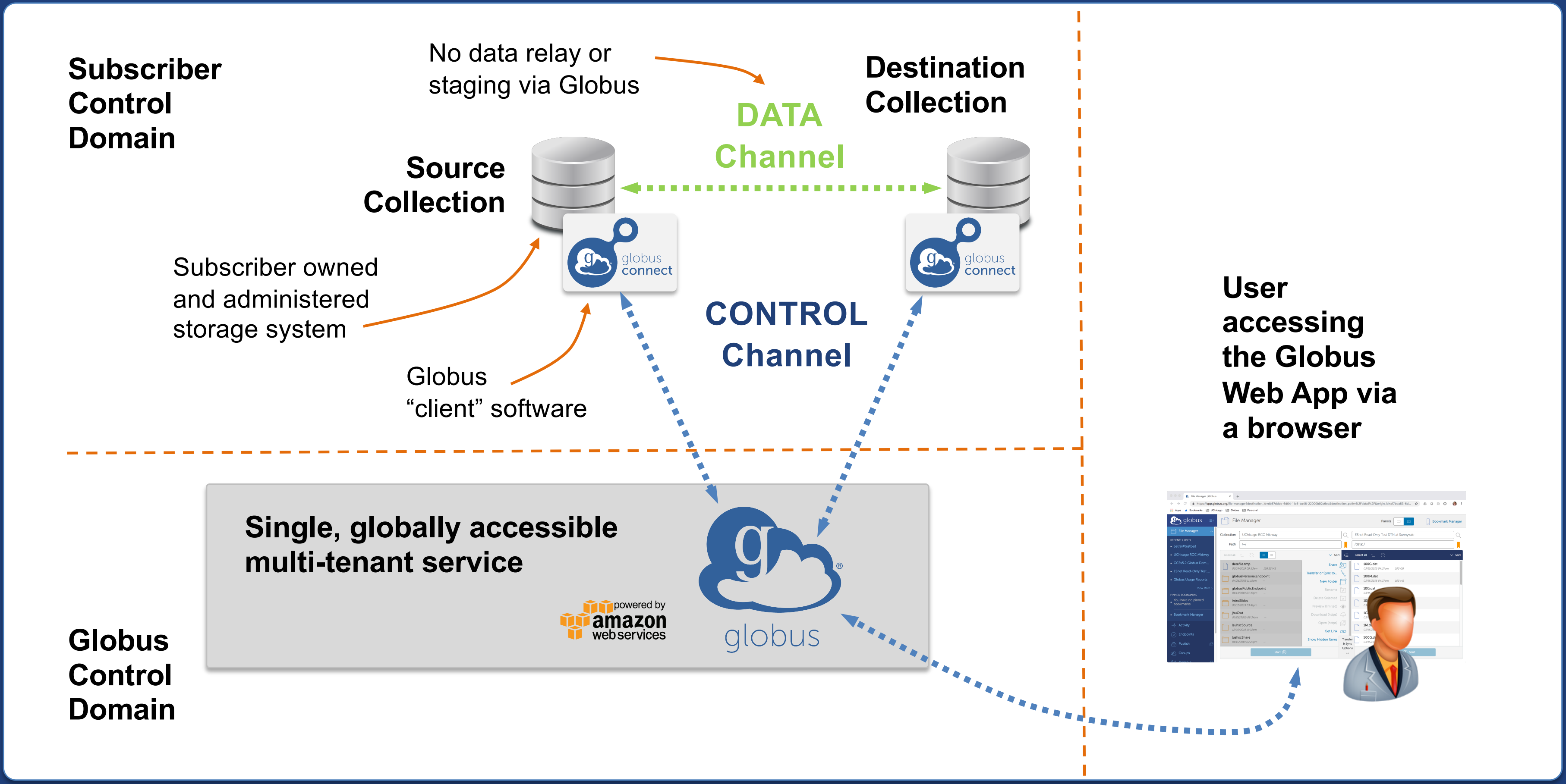
8 Automating research workflows and ensuring those that need access to the data have it.

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

Globus core security features

- **Access Control**
  - Identities provided and managed by institution
  - Institution controls all access policies
  - Globus is identity broker; no access to/storage of user credentials
  - Fine grained access control on the collections
- **Data remain at institutions, not stored by Globus**
- **Data does not flow through the Globus Service but directly between Endpoints and their Collections**
- **Integrity checks of transferred data**
- **High availability and redundancy**
- **Encryption of user files and Globus control data**

Conceptual architecture: Hybrid SaaS



Subscriber Control Domain

No data relay or staging via Globus

Destination Collection

DATA Channel

Source Collection

Subscriber owned and administered storage system

Globus "client" software

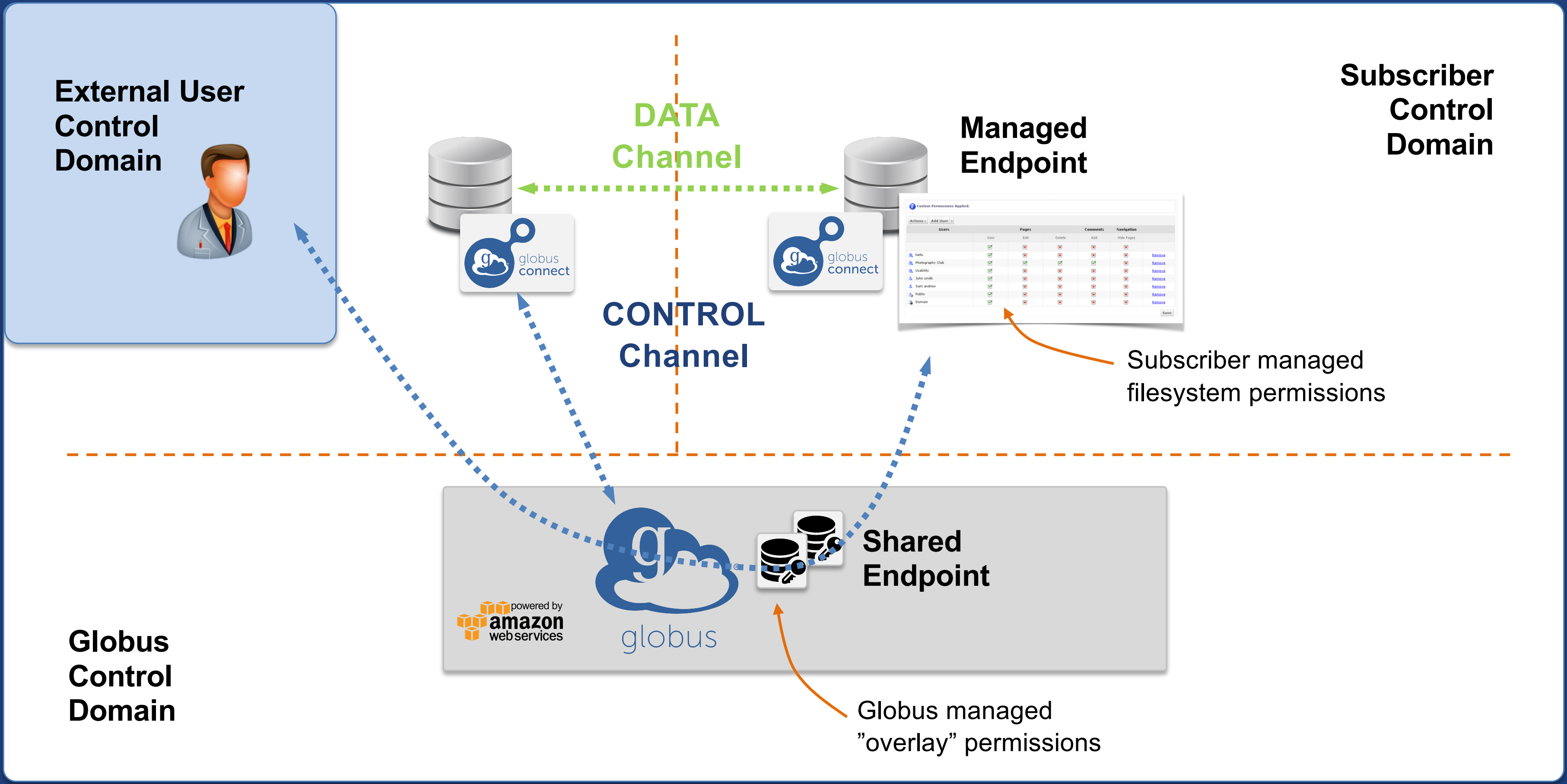
CONTROL Channel

User accessing the Globus Web App via a browser

Single, globally accessible multi-tenant service

Globus Control Domain

Conceptual architecture: Sharing



Custom Permissions Applied.

Actions ▾ Add User ▾

Users	Pages			Comments	Navigation	
	View	Edit	Delete	Add	Hide Pages	
	☑	☒	☒	☒	☒	
hello	☑	☒	☒	☒	☒	Remove
Photography Club	☑	☑	☑	☑	☒	Remove
Usability	☑	☒	☒	☒	☒	Remove
John smith	☑	☒	☒	☒	☒	Remove
Sam andrew	☑	☒	☒	☒	☒	Remove
Public	☑	☒	☒	☒	☒	Remove
Domain	☑	☒	☒	☒	☒	Remove

Save

Subscriber managed filesystem permissions

Globus managed ”overlay” permissions

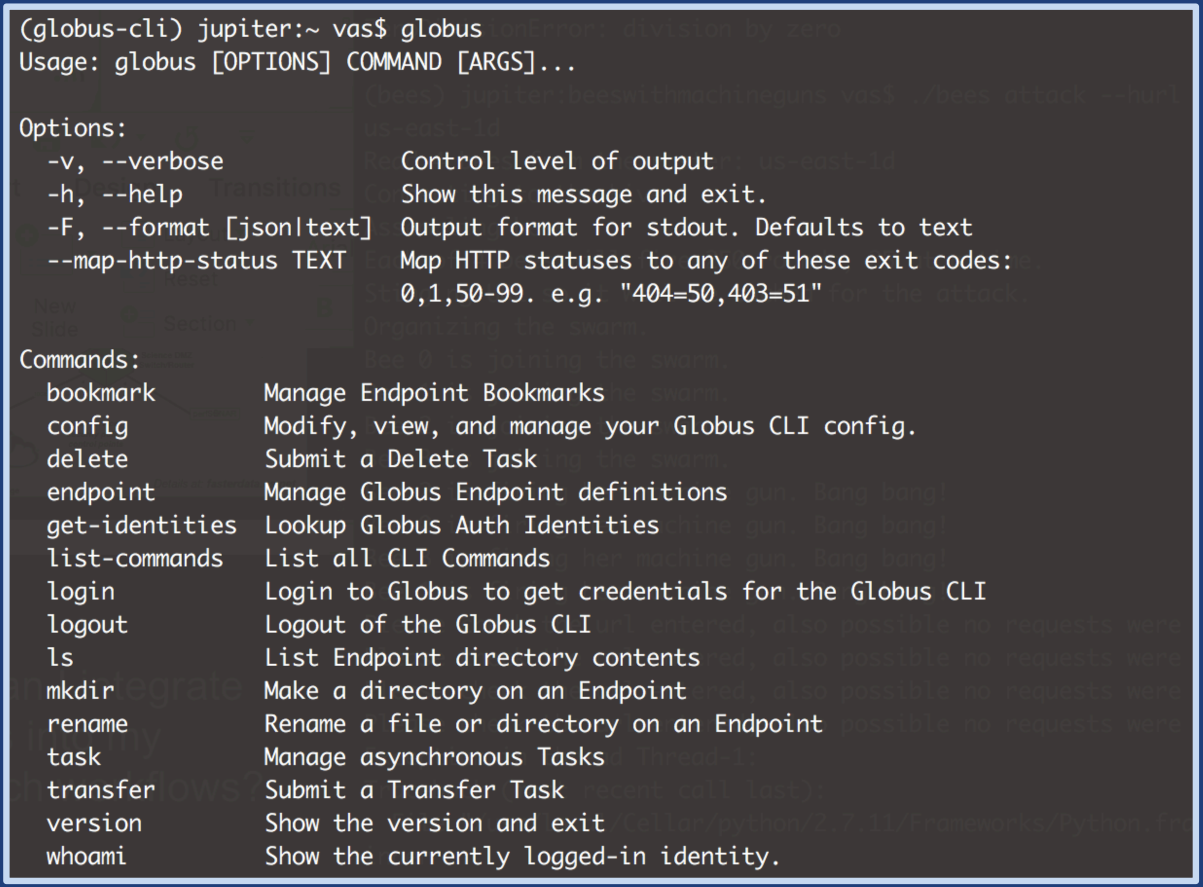
Endpoints, Collections and Globus Connect

- **Globus Connect Server**
  - Multi user Linux Systems
  - https://docs.globus.org/globus-connect-server/
- **Globus Connect Personal**
  - Personal Workstations and Laptops
  - https://www.globus.org/globus-connect-personal
  - OS specific instructions
    - https://docs.globus.org/how-to/

Globus Demo

- Authenticating to Globus
  - Linking in other identities
- The hamburger menu – file / folder management
- Transfer to and from ALCF resources
  - theta, cooley, thetagpu, grand
  - University of Chicago RCC Midway
- Globus Connect Personal
  - Installation
  - Transfer

Globus Command Line Interface



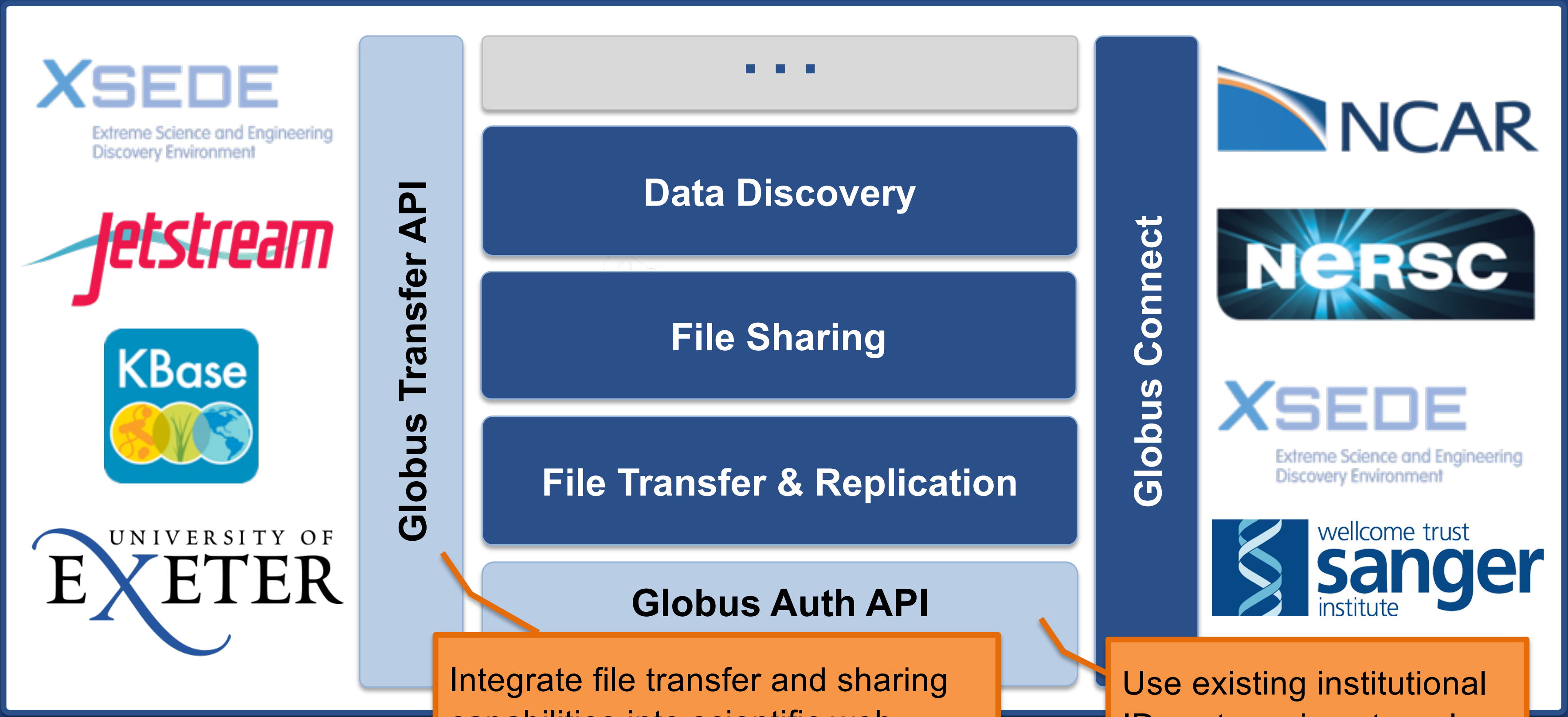
```
(globus-cli) jupiter:~ vas$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

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

Globus Platform-as-a-Service



XSEDE
Extreme Science and Engineering Discovery Environment
Jetstream
KBase
UNIVERSITY OF EXETER
NCAR
NERSC
XSEDE
Extreme Science and Engineering Discovery Environment
wellcome trust sanger institute

Globus Transfer API

. . .

Data Discovery

File Sharing

File Transfer & Replication

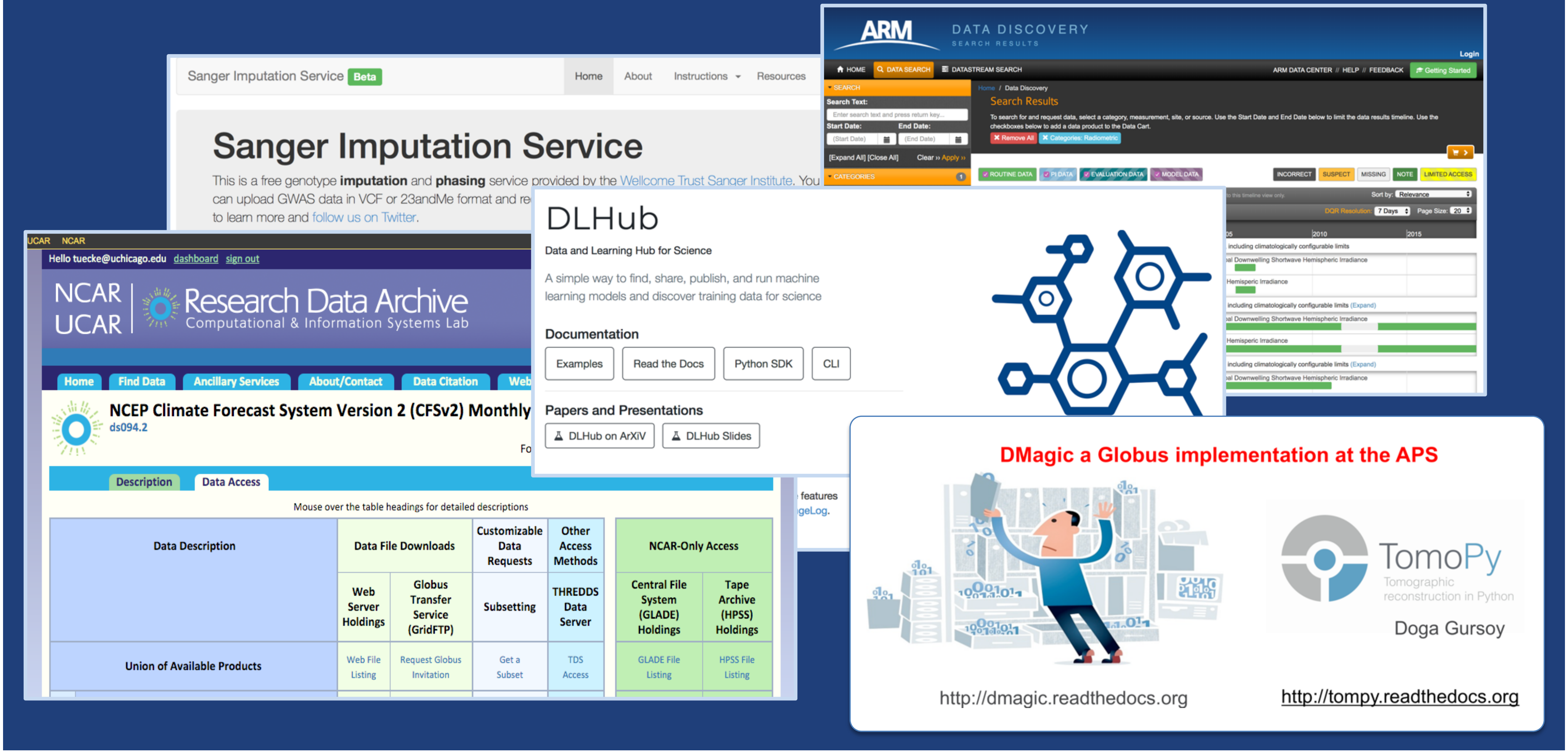
Globus Auth API

Globus Connect

Integrate file transfer and sharing capabilities into scientific web apps, portals, gateways, etc...

Use existing institutional ID systems in external web applications

Data centric applications leveraging Globus



Sanger Imputation Service Beta

Home | About | Instructions | Resources

Sanger Imputation Service

This is a free genotype **imputation** and **phasing** service provided by the Wellcome Trust Sanger Institute. You can upload GWAS data in VCF or 23andMe format and re

to learn more and follow us on Twitter.

features

geLog.

UCAR NCAR

Hello tuecke@uchicago.edu dashboard sign out

NCAR UCAR Research Data Archive Computational & Information Systems Lab

Home | Find Data | Ancillary Services | About/Contact | Data Citation | Web

NCEP Climate Forecast System Version 2 (CFSv2) Monthly

ds094.2

Fo

Description | Data Access

Mouse over the table headings for detailed descriptions

Data Description	Data File Downloads		Customizable Data Requests	Other Access Methods	NCAR-Only Access	
	Web Server Holdings	Globus Transfer Service (GridFTP)	Subsetting	THREDDS Data Server	Central File System (GLADE) Holdings	Tape Archive (HPSS) Holdings
Union of Available Products	Web File Listing	Request Globus Invitation	Get a Subset	TDS Access	GLADE File Listing	HPSS File Listing

DLHub

Data and Learning Hub for Science

A simple way to find, share, publish, and run machine learning models and discover training data for science

Documentation

Examples | Read the Docs | Python SDK | CLI

Papers and Presentations

DLHub on ArXiV | DLHub Slides

ARM DATA DISCOVERY SEARCH RESULTS

Login

HOME | DATA SEARCH | DATASTREAM SEARCH | ARM DATA CENTER // HELP // FEEDBACK | Getting Started

SEARCH

Search Text: Enter search text and press return key...

Start Date: (Start Date) End Date: (End Date)

[Expand All] [Close All] Clear ›› Apply ››

CATEGORIES 1

Home / Data Discovery

Search Results

To search for and request data, select a category, measurement, site, or source. Use the Start Date and End Date below to limit the data results timeline. Use the checkboxes below to add a data product to the Data Cart.

✖ Remove All | ✖ Categories: Radiometric

☑ ROUTINE DATA | ☑ PI DATA | ☑ EVALUATION DATA | ☑ MODEL DATA

INCORRECT | SUSPECT | MISSING | NOTE | LIMITED ACCESS

to this timeline view only. Sort by: Relevance

DQR Resolution: 7 Days Page Size: 20

including climatologically configurable limits

al Downwelling Shortwave Hemispheric Irradiance

Hemisperic Irradiance

including climatologically configurable limits (Expand)

al Downwelling Shortwave Hemispheric Irradiance

Hemisperic Irradiance

including climatologically configurable limits (Expand)

al Downwelling Shortwave Hemispheric Irradiance

DMagic a Globus implementation at the APS

TomoPy Tomographic reconstruction in Python

Doga Gursoy

http://dmagic.readthedocs.org

http://tompy.readthedocs.org

Developer references

- Globus documentation: docs.globus.org
  - Command Line Interface: docs.globus.org/cli/
  - Transfer API : docs.globus.org/api/transfer/
  - SDK: globus-sdk-python.readthedocs.io/en/stable/
- Globus GitHub: github.com/globus/
  - Jupyter Notebooks
    - Stand alone notebooks and hub integrations that walk through much of the functionality of our SDK
    - https://github.com/globus/globus-jupyter-notebooks
  - Automation Examples
    - Shell scripted CLI and Python module examples of common research data management use cases
    - https://github.com/globus/automation-examples

Support resources

- Globus documentation: docs.globus.org
- YouTube channel: youtube.com/user/GlobusOnline
- Helpdesk and issue escalation: support@globus.org
- Mailing Lists
  - globus.org/mailing-lists
- Customer engagement team