Globus 2018: **Beyond File Transfer**

GlobusWorld 2018 Keynote

Steve Tuecke & Ian Foster

globus

Mission

Increase the efficiency and effectiveness of researchers engaged in data-driven science and scholarship through sustainable software

globus

400,075,759,997 MB

TRANSFERRED

Globus by the numbers

1,042
most shared endpoints at a single institution

400 PB
transferred

66 billion
files processed

100,000
users

24
Petabyte+ institutions

15,000
active transfer users

3 months
longest running transfer

20,000
active endpoints

500+
identity providers

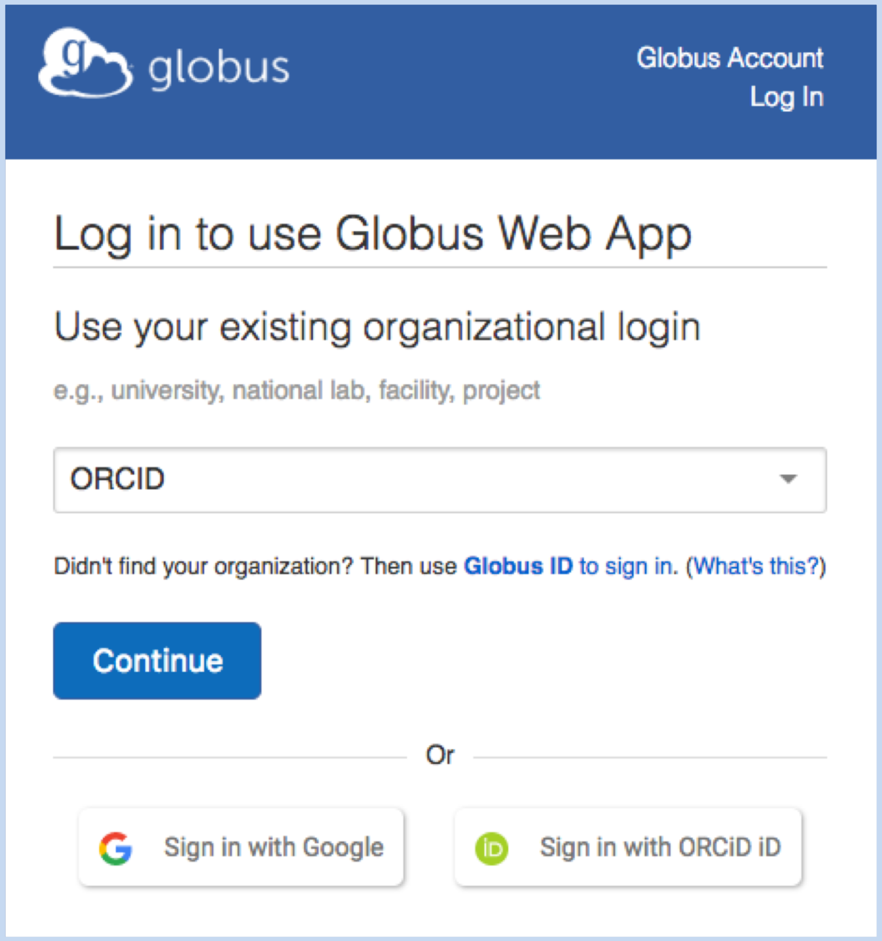
1 PB
largest single transfer to date

8,000
active shared endpoints

99.9%+
availability

Globus Auth adoption

- 100,000 users, 37k new in last year
- 500+ identity providers
- 1,100 registered apps and services
- 103,000 user consents, 35% non-Globus apps
- 99.994% uptime since Feb 2016



Log in to use Globus Web App

Use your existing organizational login

e.g., university, national lab, facility, project

ORCID

Didn't find your organization? Then use Globus ID to sign in. (What's this?)

Progress on sustainability

- **90 subscribers, including 1/3 of R1 universities**
- **½ of our product funding is from subscriptions**
- **Need most R1 and many R2 universities**
- **You can help by encouraging others to subscribe**

Help us get the word out!

- **Do you rely on Globus for your work?**
- **If so, please share your experiences!**
  - **Contribute** to our Usage Brief Library
    globus.org/usage-brief-library
  - **Add a slide** or logo in event talks (we can help!)
  - **Mention Globus** in news articles or interviews
  - **Tag us** in posts about projects that use Globus
  - **Acknowledge Globus** in your journal articles
    globus.org/publications
- **Why?**
  - Give your peers new ideas on how to use Globus
  - Help us grow the user community

"…, and file sharing with Globus."

"…with Globus for data transfer."

"We used Globus for…"

"...and Globus."

"I needed Globus to…"

"#ALCF #ORNL #theNCI #CANDLE #globusonline"

"…using tool x, tool y, Globus, technology z…

Growing outreach

- **GlobusWorld Tour:**
  11 stops, 400+ attendees
  (Contact us to host event)
- **Many innovative user stories and usage briefs**
- **New email options for staying connected**
  (including “real-time” alerts)

nysernet	May 1-2, 2018
NCSA	June 11-12, 2018

OSIRIS

TERRA-REF

DESIGNSAFE-CI

NHERI: A NATURAL HAZARDS ENGINEERING RESEARCH INFRASTRUCTURE

SIMULOCEAN

COSMIC^2

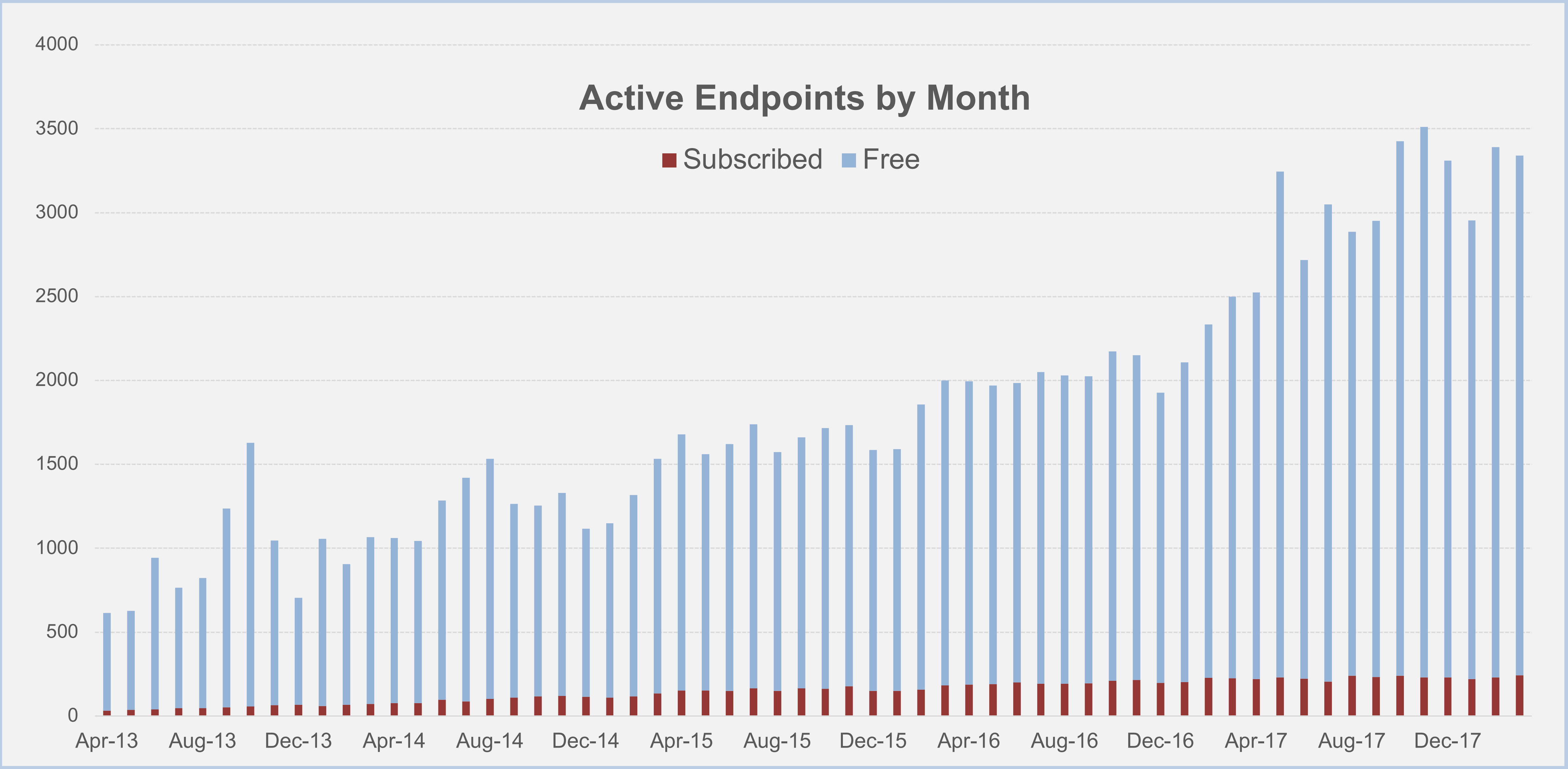
OAK RIDGE
National Laboratory

RICE
Unconventional Wisdom

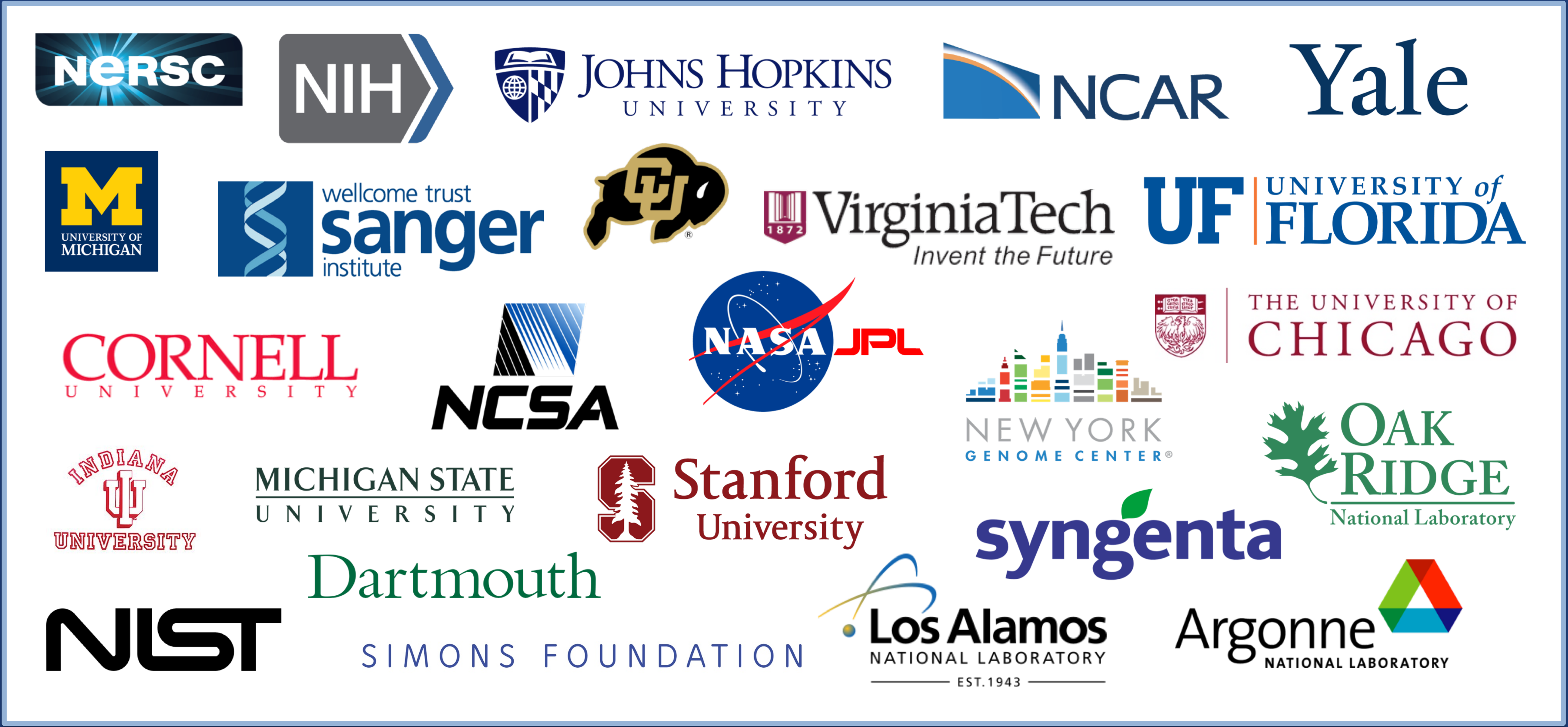
Why subscribe? Go (way) beyond file transfer…

- Remove friction for external collaborators
- Automate/scale research data flows
- Diversify research storage options—with a unified interface
- Gain visibility into research storage utilization
- Integrate robust data management into research apps
- Optimize data transfer performance
- Access expert support resources

Five-year Subscription Growth



THANK YOU, subscribers!



Thank you to our sponsors

U.S. DEPARTMENT OF ENERGY

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

NSF

NATIONAL INSTITUTES OF HEALTH

THE UNIVERSITY OF CHICAGO

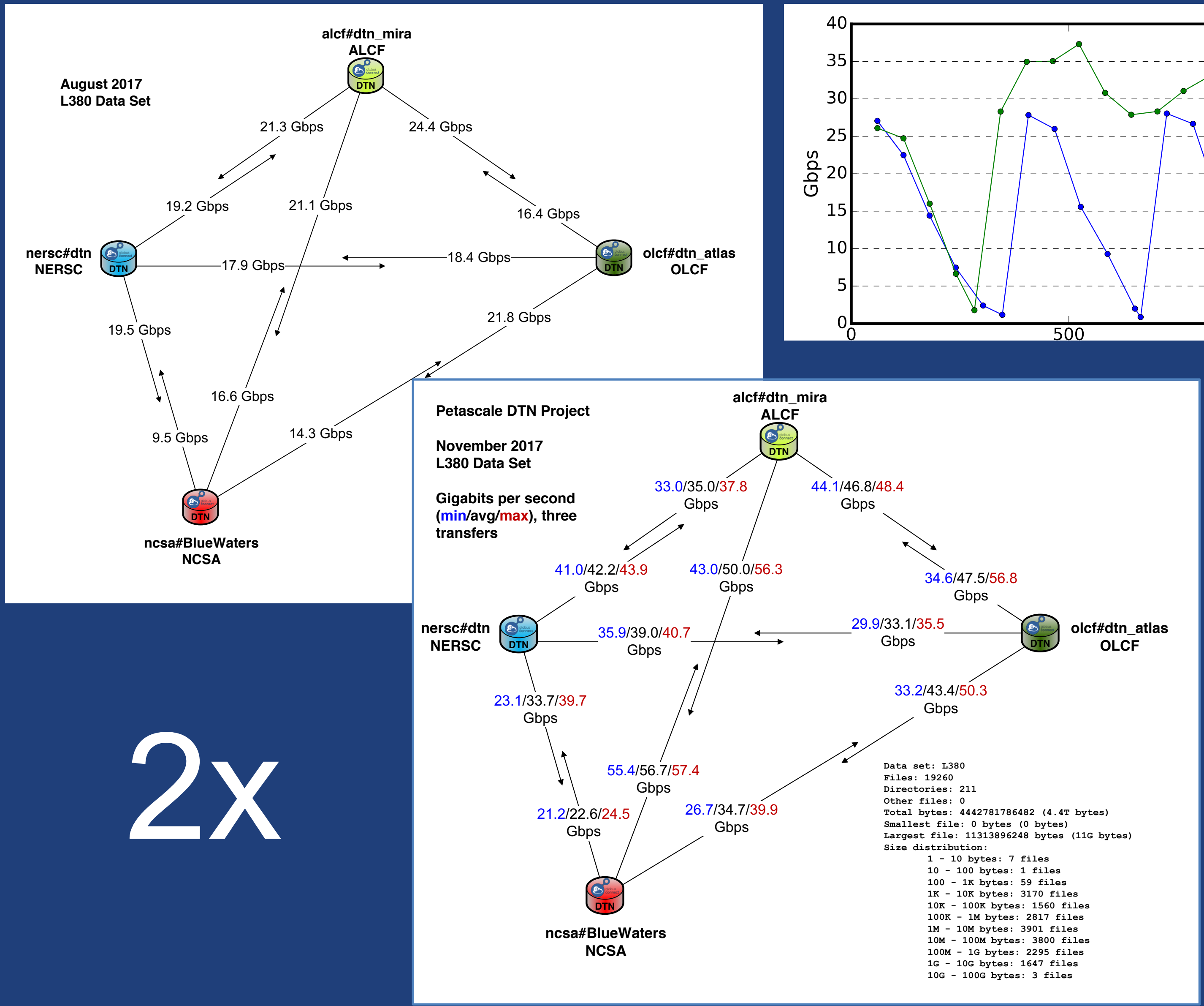
ALFRED P. SLOAN FOUNDATION
1934

Argonne NATIONAL LABORATORY

powered by amazon web services

New Features and Enhancements

Transfer performance improvements



August 2017
L380 Data Set

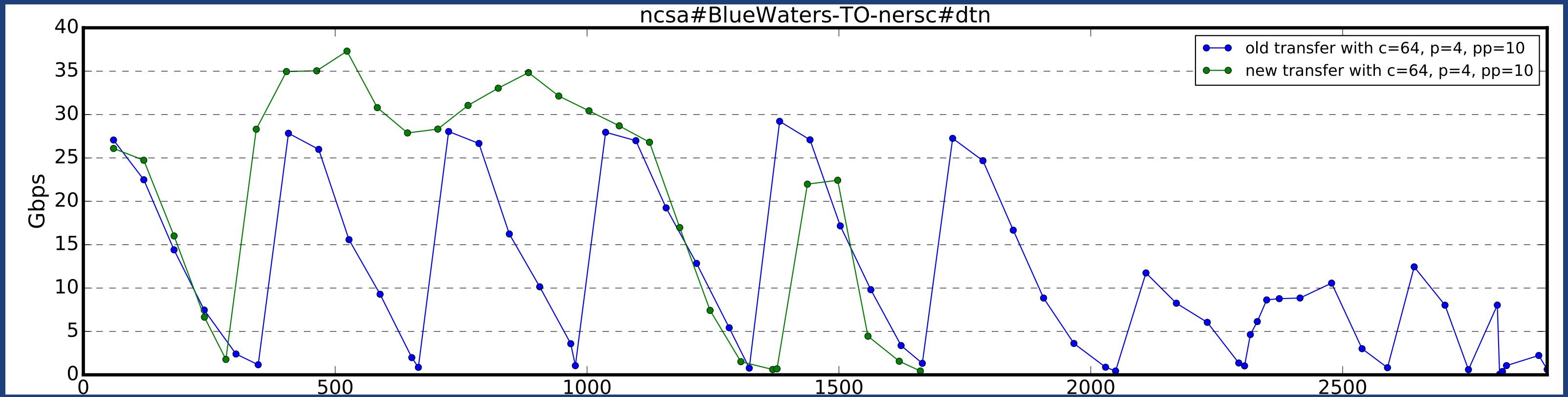
Petascale DTN Project

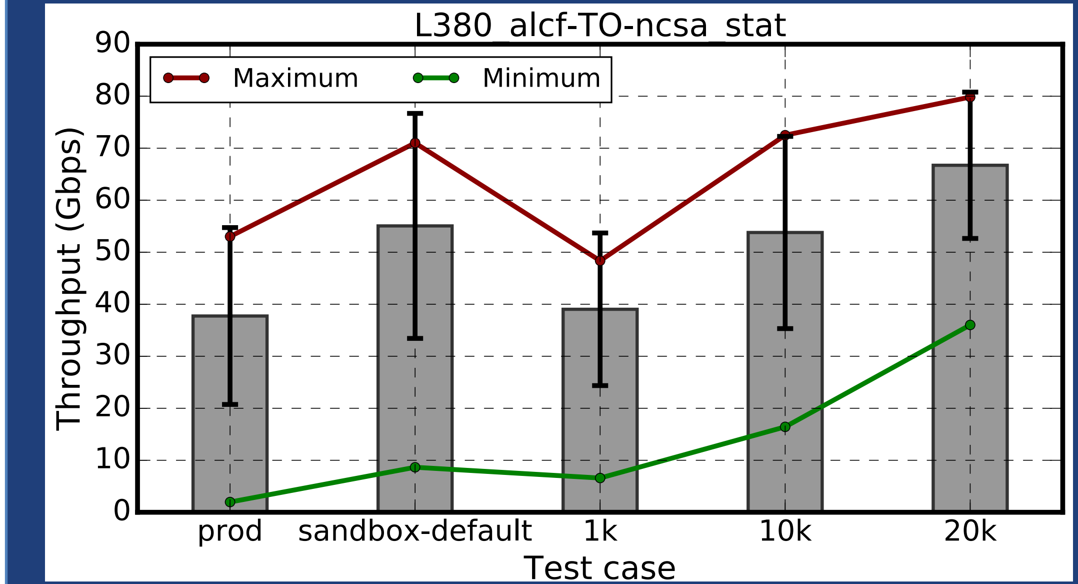
November 2017
L380 Data Set

Gigabits per second (min/avg/max), three transfers

```
Data set: L380
Files: 19260
Directories: 211
Other files: 0
Total bytes: 4442781786482 (4.4T bytes)
Smallest file: 0 bytes (0 bytes)
Largest file: 11313896248 bytes (11G bytes)
Size distribution:
        1 - 10 bytes: 7 files
        10 - 100 bytes: 1 files
        100 - 1K bytes: 59 files
        1K - 10K bytes: 3170 files
        10K - 100K bytes: 1560 files
        100K - 1M bytes: 2817 files
        1M - 10M bytes: 3901 files
        10M - 100M bytes: 3800 files
        100M - 1G bytes: 2295 files
        1G - 10G bytes: 1647 files
        10G - 100G bytes: 3 files
```

2x





Storage connectors - globus.org/connectors

HPSS

amazon web services™ | S3

Lustre®

ceph

SPECTRA®

globus connect

Google Drive

hadoop HDFS

IBM Spectrum Scale

Western Digital®

HGST ActiveScale

Western Digital®

globus connect

- **Turnkey on-premise object storage**
- **Globus connector using S3 API**
- **Low TCO:**
  - – Manufactures own drives
  - – Erasure coding
  - – BitDynamics: background data integrity checks with self-healing
  - – Cloud-based systems management tools
  - – Data Forever: automatic migration to new tech

https://docs.globus.org/premium-storage-connectors/wd-activescale/

Connectors for S3 ”compatible” systems

- **S3 API is de-facto standard API for object storage**
- **Make it easier to validate and support connectors for S3 “compatible” object storage systems**
  - Functionality and performance test suite
  - Improving connector robustness and performance
  - E.g., Ceph, ActiveScale, SwiftStack, Wasabi, IBM Cloud Object Storage System (CleverSafe)
- **Also requires vendor engagement and market interest**

Upcoming Webinar: May 22

Simplifying large-scale data management and lowering total cost of storage with Globus and Spectra

- May 22, 2018 at 11 a.m. EDT / 8 a.m. PDT
- Guest speaker from UMN / MSI
- Topics include:
  - MSI's use of the Spectra® BlackPearl® solution with Globus premium connector
  - Cost model for Spectra® BlackPearl®

https://globus.org/events/webinar-tco-spectra-msi

"Spectra Logic's T950 and BlackPearl are important components in our strategic and comprehensive storage plan for hundreds of terabytes of critical research data."

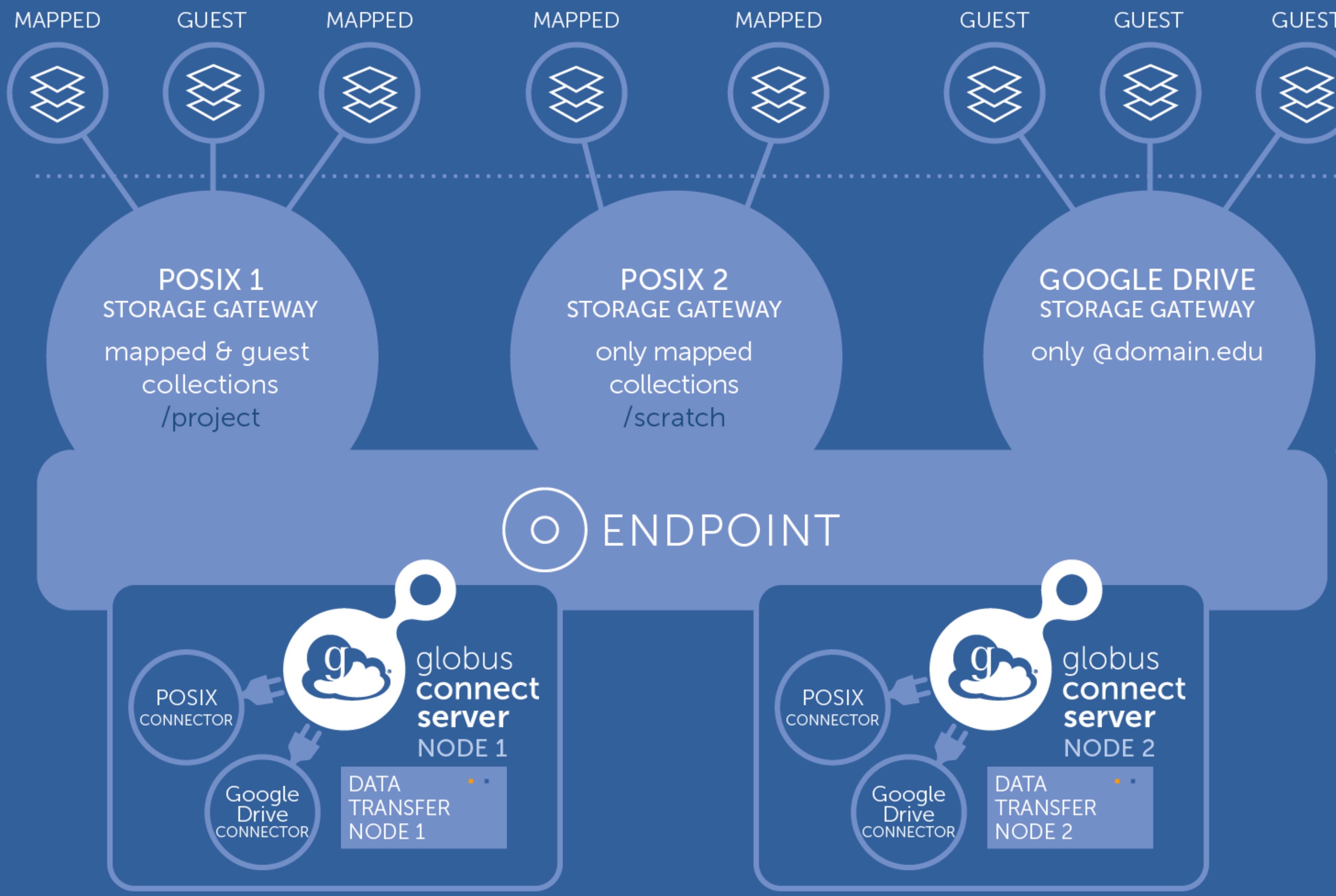
--Jeff McDonald, Assistant Director for HPC Operations, MSI

HPSS

- Community has agreed on sustainability model
- NERSC & ORNL investing in enhancements
- Premium storage connector subscription

Globus Connect Server v5 update

- **Limited production releases:**
  - V5.0: Google Drive (Fall 2017)
  - V5.1: HTTPS and Posix shared endpoints (next week)
  - Subsequent v5.x building up to full functionality
- **Year-end target for full release**
  - End-to-end Globus Auth
  - Multi-DTN
  - Remaining connectors
- Globus Connect Community Source License



MAPPED GUEST MAPPED MAPPED MAPPED GUEST GUEST GUEST

POSIX 1
STORAGE GATEWAY
mapped & guest collections
/project

POSIX 2
STORAGE GATEWAY
only mapped collections
/scratch

GOOGLE DRIVE
STORAGE GATEWAY
only @domain.edu

ENDPOINT

POSIX CONNECTOR

Google Drive CONNECTOR

globus connect server
NODE 1

DATA TRANSFER NODE 1

POSIX CONNECTOR

Google Drive CONNECTOR

globus connect server
NODE 2

DATA TRANSFER NODE 2

COLLECTIONS

Data access interface
GridFTP & HTTPS

STORAGE GATEWAYS

Policies & configuration

ENDPOINT

Management & config interface

DATA TRANSFER NODES

Network & storage connected servers in ScienceDMZ

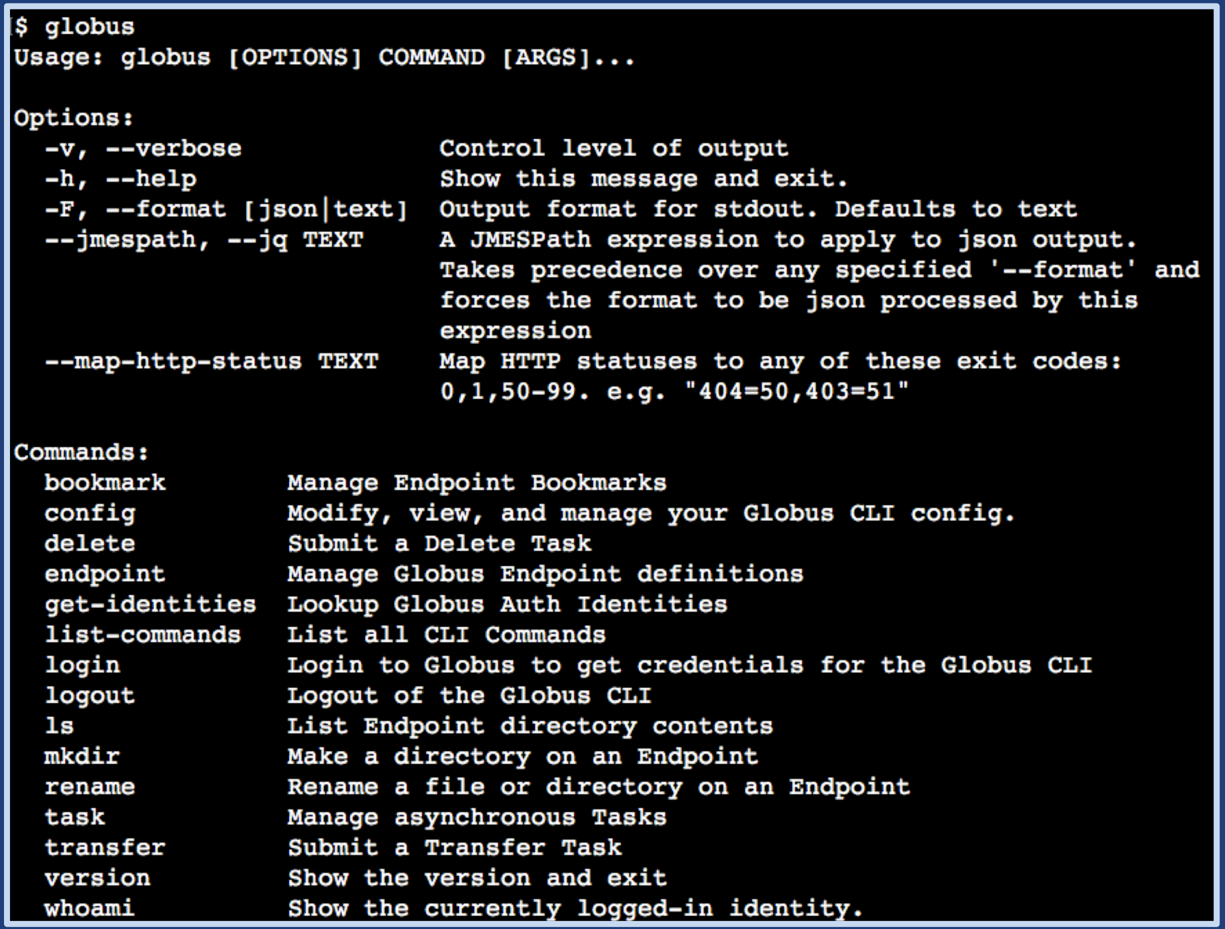
Globus Toolkit end of support

- **General support for open source Globus Toolkit has ended**
  - Does not effect Globus service or Globus Connect
- **Customers using the GT GridFTP with Globus service will continue to be supported until GCSv4 is discontinued**
  - Security patches continue for GridFTP, MyProxy, GSI-OpenSSH
- **GCSv4 and GT GridFTP will be discontinued 6 months after GCSv5 full release**

https://github.com/globus/globus-toolkit/blob/globus_6_branch/support-changes.md

Command Line Interface

- **New Globus CLI is generally available**
  - Fully functional
  - Many enhancements
  - Simple updater
- **Deprecating old hosted SSH CLI**
  - Will be turned off August 1

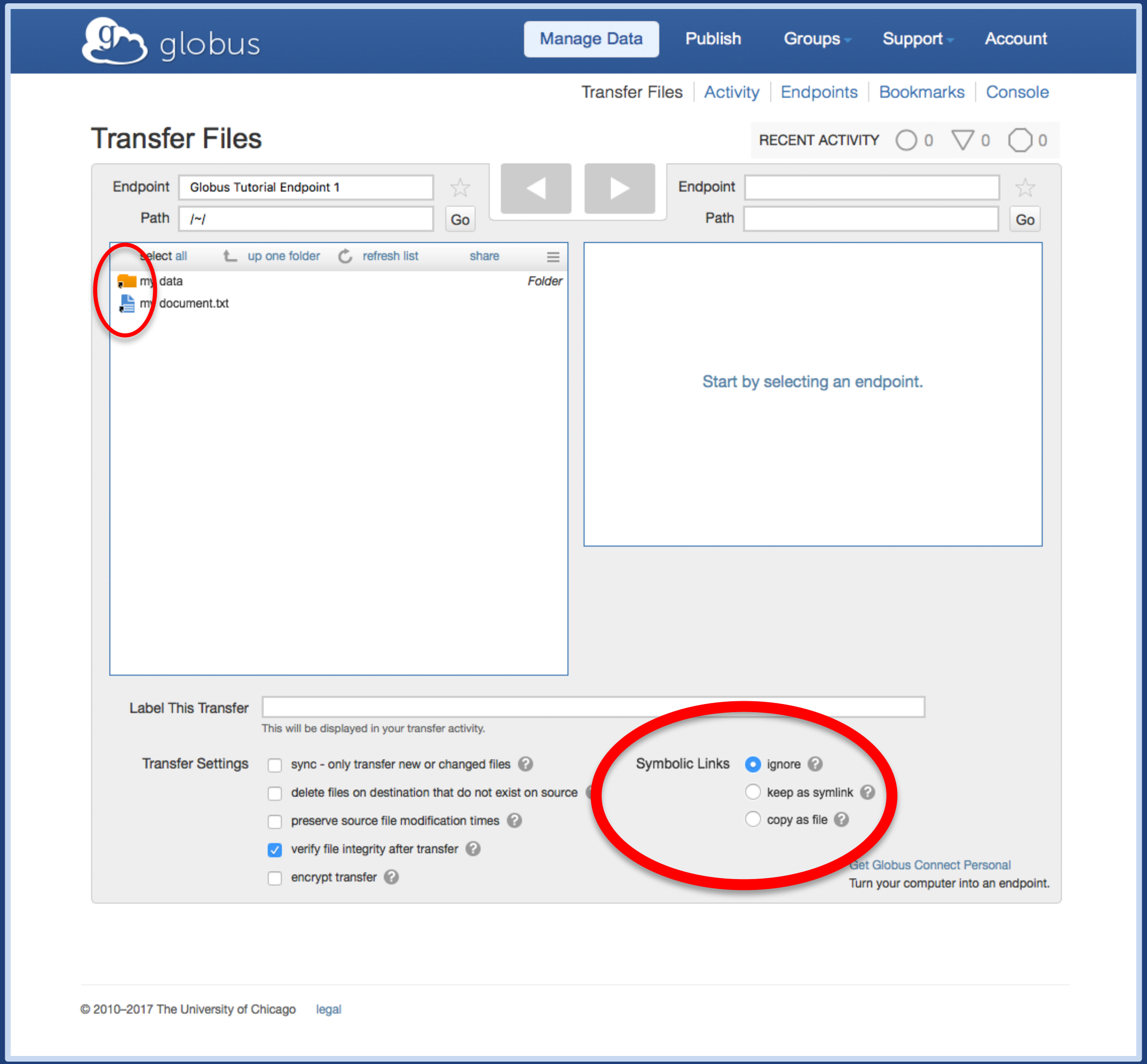


```
$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

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
  --jmespath, --jq TEXT       A JMESPath expression to apply to json output.
                              Takes precedence over any specified '--format' and
                              forces the format to be json processed by this
                              expression
  --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.
```

https://docs.globus.org/cli

Symbolic links delayed

- **On transfer and sync**
- **Options:**
  - – Ignore
  - – Keep as symlinks
  - – Copy as files
- **Target later this year**



Transfer Files

Endpoint: Globus Tutorial Endpoint 1
Path: /~/

my data	Folder
my document.txt	

Endpoint:
Path:

Start by selecting an endpoint.

Label This Transfer:
This will be displayed in your transfer activity.

Transfer Settings
- □ sync - only transfer new or changed files
- □ delete files on destination that do not exist on source
- □ preserve source file modification times
- ☑ verify file integrity after transfer
- □ encrypt transfer

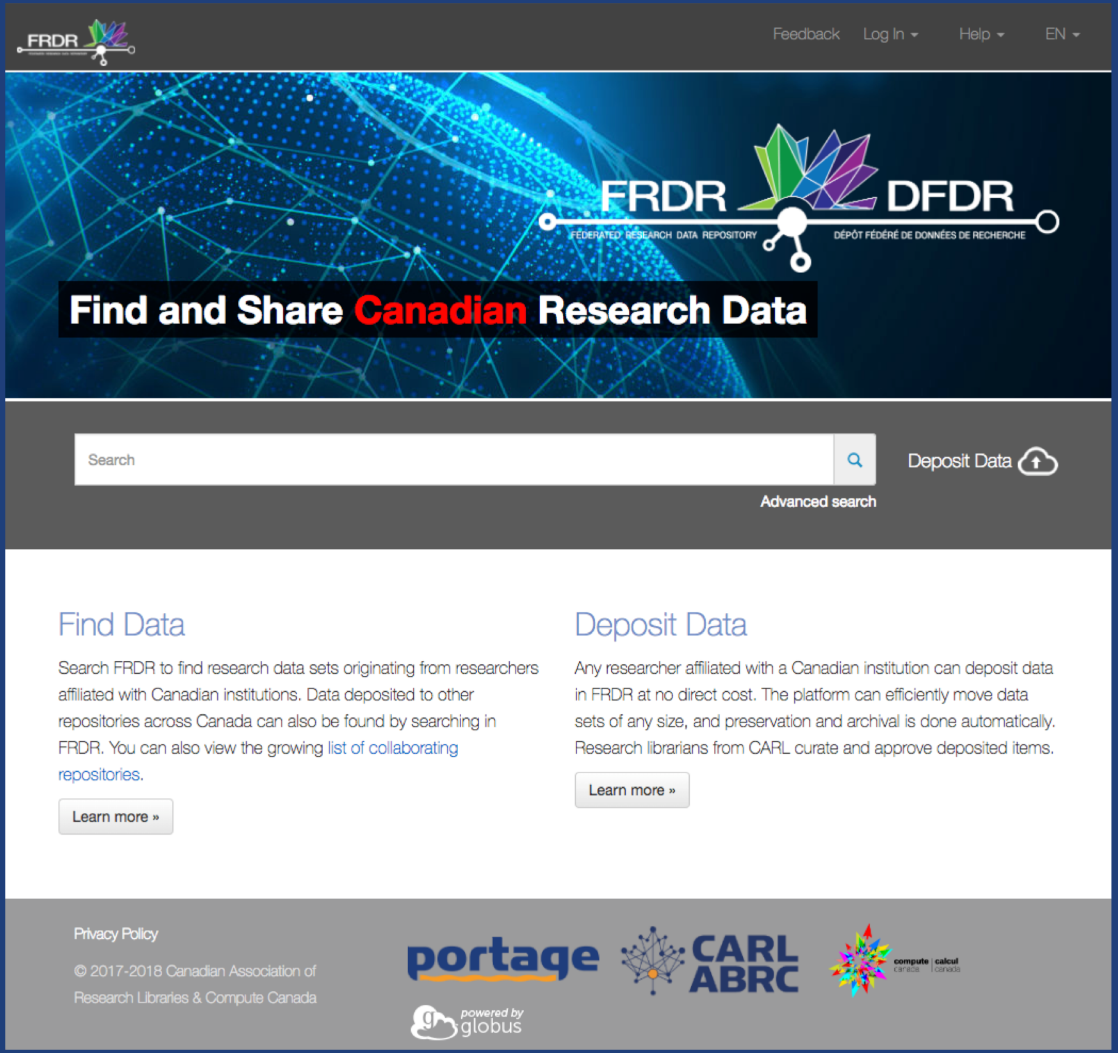
Symbolic Links
- ● ignore
- ○ keep as symlink
- ○ copy as file

Get Globus Connect Personal
Turn your computer into an endpoint.

© 2010–2017 The University of Chicago legal

Publication v1

- **Publication v1 app**
  - Publish datasets to Globus Search
  - Internationalization
- **Canadian Federated Research Data Repository**
  - https://frdr.ca/
  - Uses v1 open source and Globus Search



Find and Share Canadian Research Data

Find Data

Search FRDR to find research data sets originating from researchers affiliated with Canadian institutions. Data deposited to other repositories across Canada can also be found by searching in FRDR. You can also view the growing list of collaborating repositories.

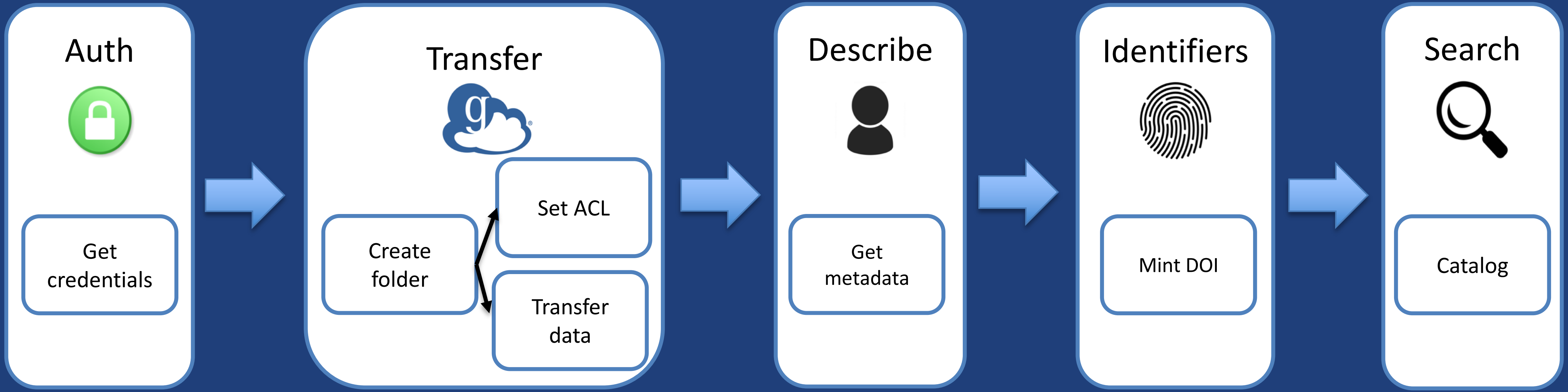
Deposit Data

Any researcher affiliated with a Canadian institution can deposit data in FRDR at no direct cost. The platform can efficiently move data sets of any size, and preservation and archival is done automatically. Research librarians from CARL curate and approve deposited items.

© 2017-2018 Canadian Association of Research Libraries & Compute Canada

Publication v2 platform

- **Decompose Publication v1 into platform components**
- **Allow flexible re-composition & adaptation by customers**



Auth

Get credentials

Transfer

Create folder

Set ACL

Transfer data

Describe

Get metadata

Identifiers

Mint DOI

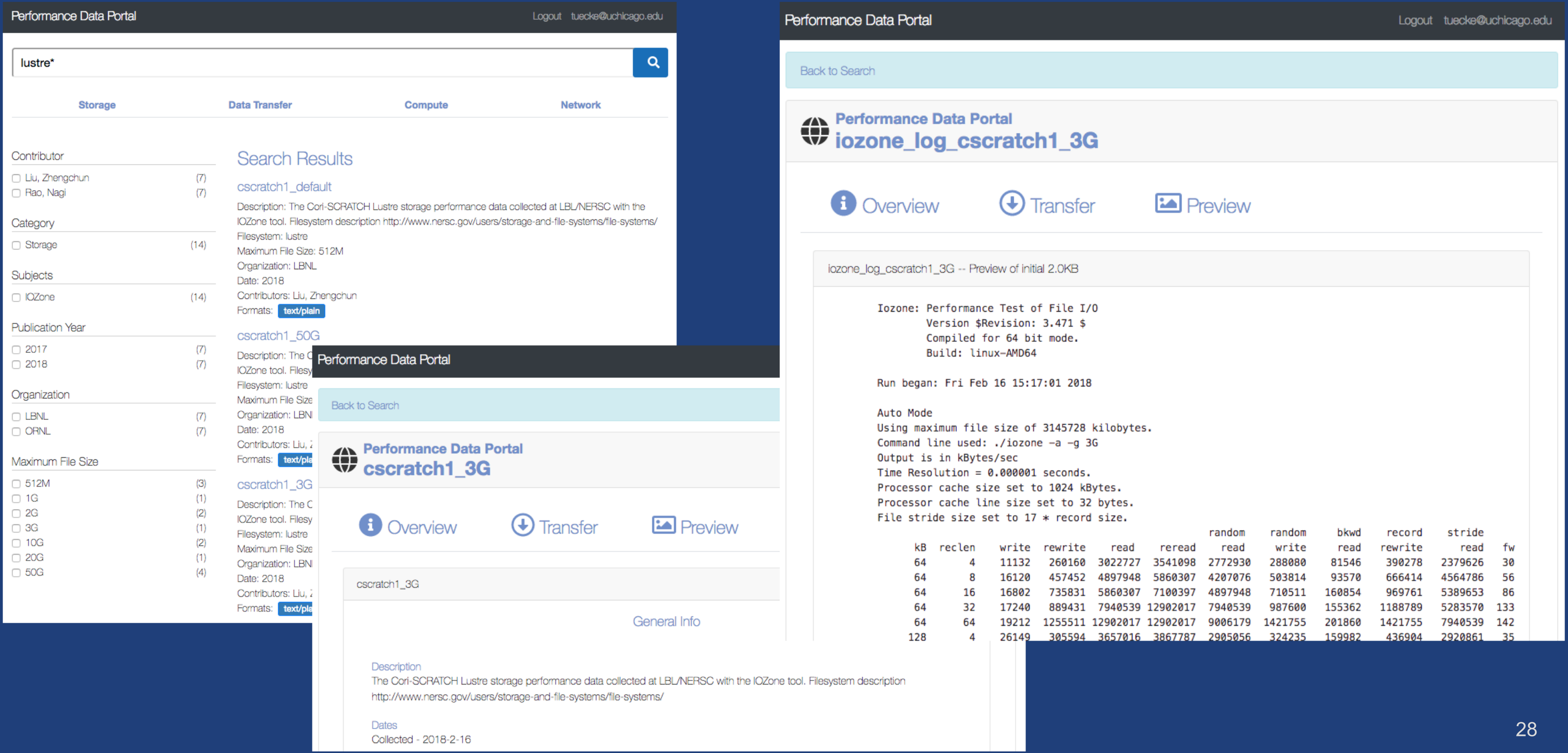
Search

Catalog

Globus Search platform service

- **Search service:**
  - **Scalable**: to billions of entries
  - **Schema agnostic**: can use standard (e.g., DataCite) or custom metadata
  - **Fine grain access control**: only returns results that are visible to user
  - **Plain text search**: ranked results
  - **Faceted search**: for data discovery
  - **Rich query language**: ranges, expressions, regex, fuzzy, stemming, etc.
- **Limited production, generally available target year end**
- **Tutorial: Data Publication and Discovery with Globus**

Django Globus Portal App



Contributor

- □ Liu, Zhengchun (7)
- □ Rao, Nagi (7)

Category

- □ Storage (14)

Subjects

- □ IOZone (14)

Publication Year

- □ 2017 (7)
- □ 2018 (7)

Organization

- □ LBNL (7)
- □ ORNL (7)

Maximum File Size

- □ 512M (3)
- □ 1G (1)
- □ 2G (2)
- □ 3G (1)
- □ 10G (2)
- □ 20G (1)
- □ 50G (4)

Search Results

cscratch1_default

Description: The Cori-SCRATCH Lustre storage performance data collected at LBL/NERSC with the IOZone tool. Filesystem description http://www.nersc.gov/users/storage-and-file-systems/file-systems/
Filesystem: lustre
Maximum File Size: 512M
Organization: LBNL
Date: 2018
Contributors: Liu, Zhengchun
Formats: text/plain

cscratch1_50G

Description: The C
IOZone tool. Filesy
Filesystem: lustre
Maximum File Size
Organization: LBN
Date: 2018
Contributors: Liu, Z
Formats: text/pla

cscratch1_3G

Description: The C
IOZone tool. Filesy
Filesystem: lustre
Maximum File Size
Organization: LBN
Date: 2018
Contributors: Liu, Z
Formats: text/pla

cscratch1_3G

cscratch1_3G

General Info

Description
The Cori-SCRATCH Lustre storage performance data collected at LBL/NERSC with the IOZone tool. Filesystem description http://www.nersc.gov/users/storage-and-file-systems/file-systems/

Dates
Collected - 2018-2-16

iozone_log_cscratch1_3G

iozone_log_cscratch1_3G -- Preview of initial 2.0KB

```
	Iozone: Performance Test of File I/O
	        Version $Revision: 3.471 $
	        Compiled for 64 bit mode.
	        Build: linux-AMD64

	Run began: Fri Feb 16 15:17:01 2018

	Auto Mode
	Using maximum file size of 3145728 kilobytes.
	Command line used: ./iozone -a -g 3G
	Output is in kBytes/sec
	Time Resolution = 0.000001 seconds.
	Processor cache size set to 1024 kBytes.
	Processor cache line size set to 32 bytes.
	File stride size set to 17 * record size.
                                                              random    random      bkwd    record    stride
              kB  reclen    write  rewrite     read    reread     read     write      read   rewrite      read   fw
              64       4    11132   260160  3022727   3541098  2772930    288080     81546    390278   2379626   30
              64       8    16120   457452  4897948   5860307  4207076    503814     93570    666414   4564786   56
              64      16    16802   735831  5860307   7100397  4897948    710511    160854    969761   5389653   86
              64      32    17240   889431  7940539  12902017  7940539    987600    155362   1188789   5283570  133
              64      64    19212  1255511 12902017  12902017  9006179   1421755    201860   1421755   7940539  142
             128       4    26149   305594  3657016   3867787  2905056    324235    159982    436904   2920861   35
```

Django Globus Portal App Demo

Globus Identifiers platform service

- **Issue persistent identifiers**
  - DOI, ARK, Handle, Globus
  - E.g., https://identifiers.globus.org/doi:10.1145/2076450.2076468
- **Within a namespace**
  - E.g., Your University's DataCite namespace
  - Control which identities and groups can create identifiers in your namespace
- **Each identifier has:**
  - **Link to data**: one or more https URLs, to file, folder or manifest
  - **Landing page**: provided by service, or by user
  - **Visibility**: which identities and groups can see identifier
  - **Checksum**: of the file or manifest
  - **Metadata**: as required by identifier (e.g., DataCite), extensible
  - **Replaces / Replaced-by**: for versioning
- **Limited beta available now, generally available year end**
- **Tutorial: Data Publication and Discovery with Globus**

Jupyter Integration

- **Authenticate to JupyterHub with Globus Auth**
  - Passes tokens into notebooks as environment variable

- **Use Globus data management platform from notebooks**
  - With Globus Python SDK

Jupyter Integration Demo

What’s Coming Next

Protected data

- **High assurance endpoints**
  - User must authenticate with specific identity within a specified time period, with browser session and native app device instance isolation
  - Audit logging
  - Multi-factor authentication
- **For data that requires additional security**
  - HIPAA Personal Health Information (PHI) w/ BAA
  - Personally Identifiable Information (PII)
  - Sensitive but unclassified
- **NIST 800-171 Low**
- **Two additional subscription tiers**
  - **High assurance tier**: for all added security features
  - **BAA tier**: high assurance features plus BAA with Uchicago
- **Available this Summer**
  - Transfer, sharing, web app, CLI only. Excludes publish, search, identifiers, hosted CLI, GlobusID

SSH with Globus Auth

- **Securely access resource using SSH with federated identity**
  - Leverage same security model as rest of data infrastructure
  - Facilitates automation
  - Eliminate need to manage SSH key lifecycle and provisioning
- **Replaces GSI SSH**
- **Client side wrapper around local SSH client (globus-ssh …)**
- **No changes to the SSH server (PAM module)**
- **Status:**
  - Prototype complete, early customer feedback
  - GA by end of year
- **Lightning talk: SSH with Globus Auth**

New Storage Connectors

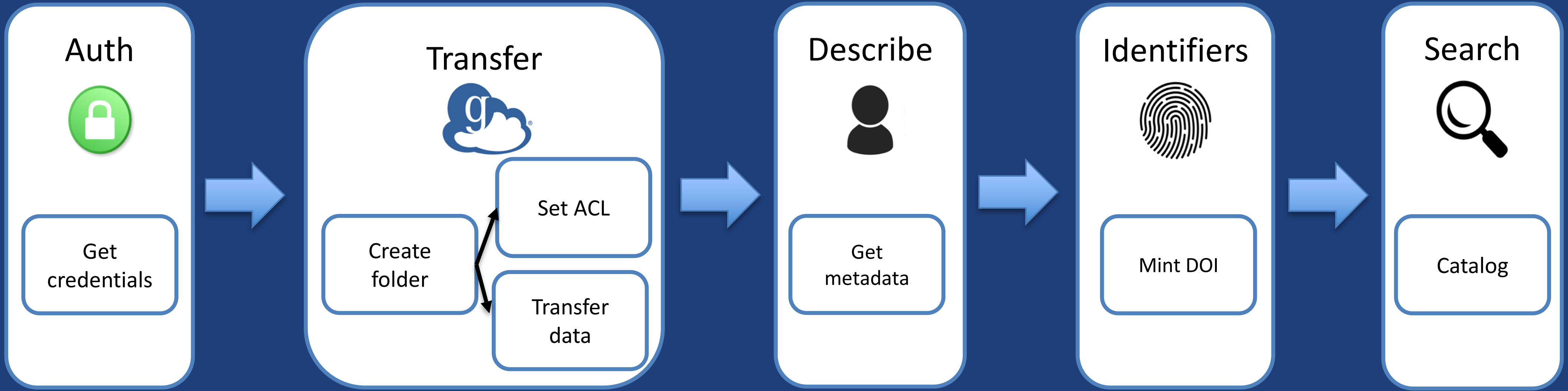
- **We continue to grow our connector set**
- **On near-term radar**
  - Box
  - Google Cloud Storage
- **What else do you want?**
  - Microsoft Azure Blob Storage
  - Wasabi
  - …

Groups

- **Generally available in web app**
- **REST API has been in limited production**
- **Plan on opening some portion to general availability**
  - Please tell us your use cases

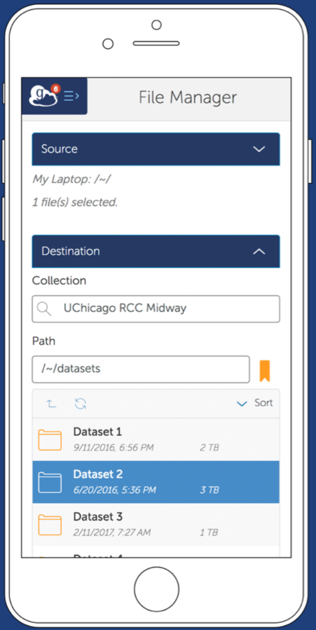
Publication v2 platform

- **Form-based metadata entry**
- **Automate**



New web app

- **Complete file manager for any research storage**
- **Improved browser experience**
  - **Accessibility**: WCAG 2.0 AA
  - **Responsiveness**: from large desktop to small phone
  - **Touch support**: for phones and pads
- **Leverage Globus Connect HTTPS**
  - E.g., Preview, download
- **Beta available now: https://app.globus.org**



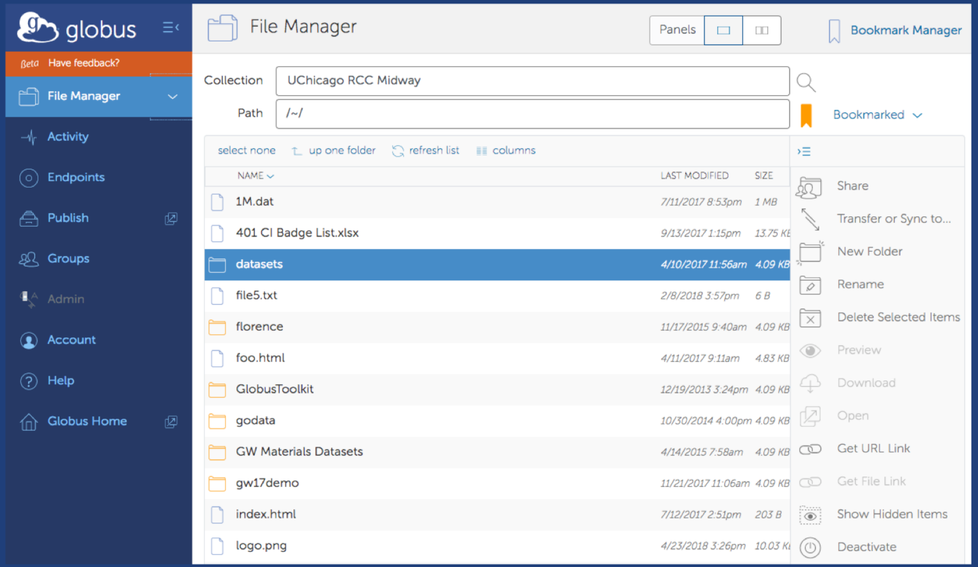
My Laptop: /~/

1 file(s) selected.

Collection: UChicago RCC Midway

Path: /~/datasets

Dataset 1	*9/11/2016, 6:56 PM*	*2 TB*
Dataset 2	*6/20/2016, 5:36 PM*	*3 TB*
Dataset 3	*2/11/2017, 7:27 AM*	*1 TB*



Collection: UChicago RCC Midway

Path: /~/

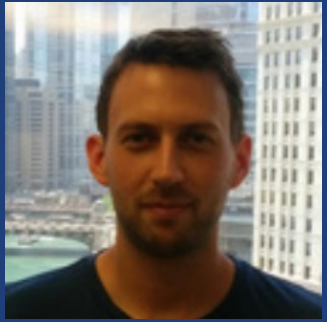
NAME	LAST MODIFIED	SIZE
1M.dat	*7/11/2017 8:53pm*	*1 MB*
401 CI Badge List.xlsx	*9/13/2017 1:15pm*	*13.75 K*
datasets	*4/10/2017 11:56am*	*4.09 KB*
file5.txt	*2/8/2018 3:57pm*	*6 B*
florence	*11/17/2015 9:40am*	*4.09 KB*
foo.html	*4/11/2017 9:11am*	*4.83 KB*
GlobusToolkit	*12/19/2013 3:24pm*	*4.09 KB*
godata	*10/30/2014 4:00pm*	*4.09 KB*
GW Materials Datasets	*4/14/2015 7:58am*	*4.09 KB*
gw17demo	*11/21/2017 11:06am*	*4.09 KB*
index.html	*7/12/2017 2:51pm*	*203 B*
logo.png	*4/23/2018 3:26pm*	*10.03 K*

New Web App Demo

Globus Labs

globus labs -- labs.globus.org

Ian Foster


Ryan Chard

Tyler Skluzacek

Zhi Hong

Logan Ward

Yulie Zamora

Ben Blaiszik

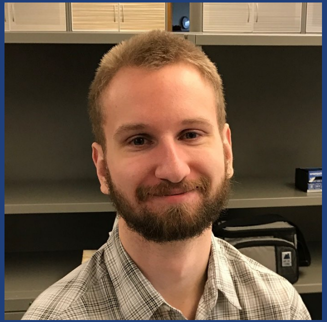
Ricardo Lourenco

Roselyne Tchoua

Anna Woodward

Kyle Chard

Sam Nickolay


Jonathan Gaff

Steve Tuecke

NATIONAL INSTITUTES OF HEALTH

NSF

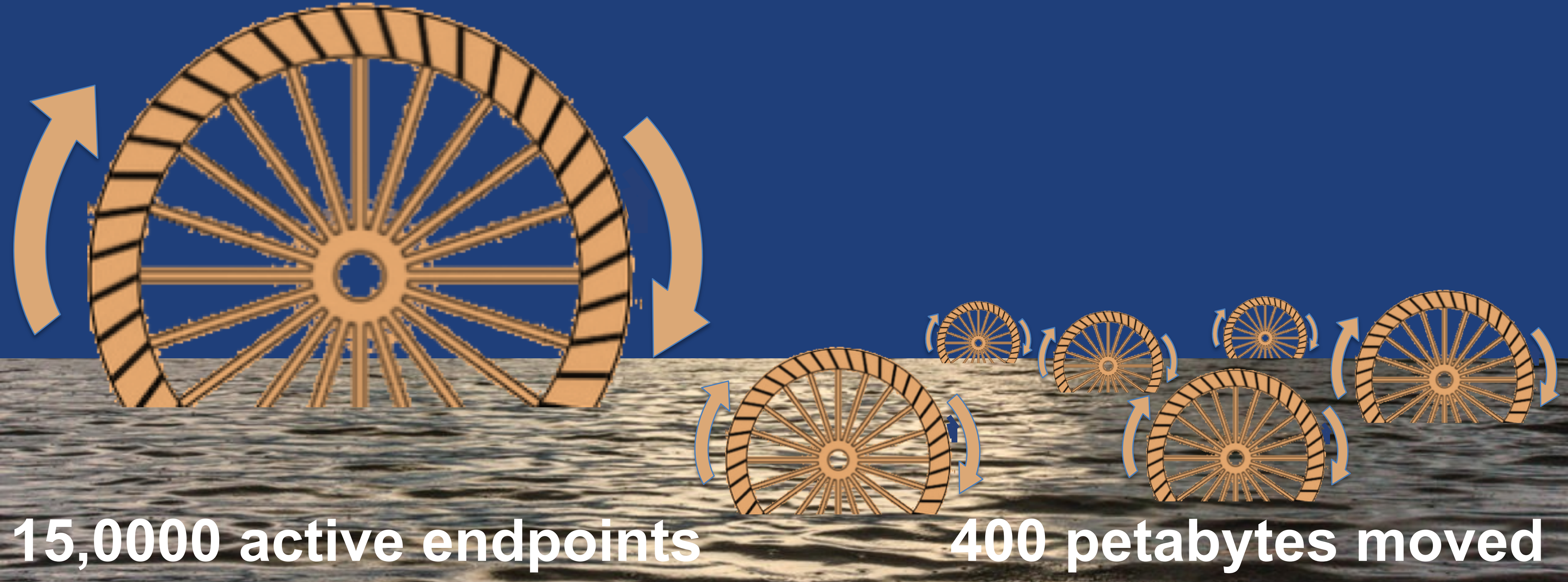
NIST

U.S. DEPARTMENT OF ENERGY

“Make all research data reliably, rapidly, and securely accessible, discoverable, and usable”

- Address computer science & domain science challenges
- Contribute to Globus with improved performance, new features, product directions, exploratory prototypes
- Leverage advanced Globus features

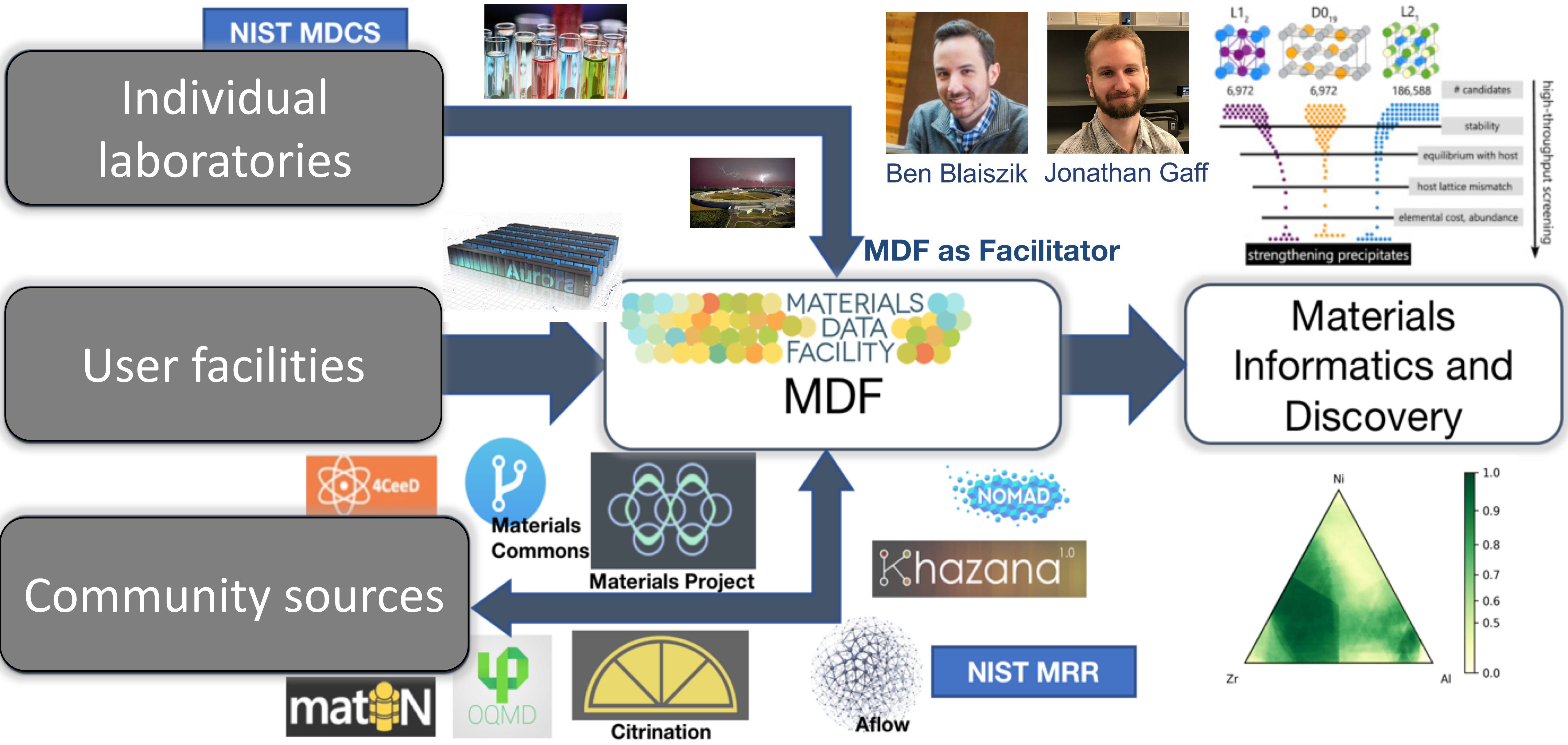
Building "data turbines"?



15,0000 active endpoints

400 petabytes moved

Materials Data Facility: materialsdatafacility.org



Individual laboratories

User facilities

Community sources

MDF as Facilitator

MATERIALS DATA FACILITY

MDF

Materials Informatics and Discovery

Ben Blaiszik

Jonathan Gaff

Materials Data Facility adoption

Publication

61 Total datasets

29 Institutions

22 CHiMaD datasets

150 Authors

>18 TB Data Volume

MDF Index

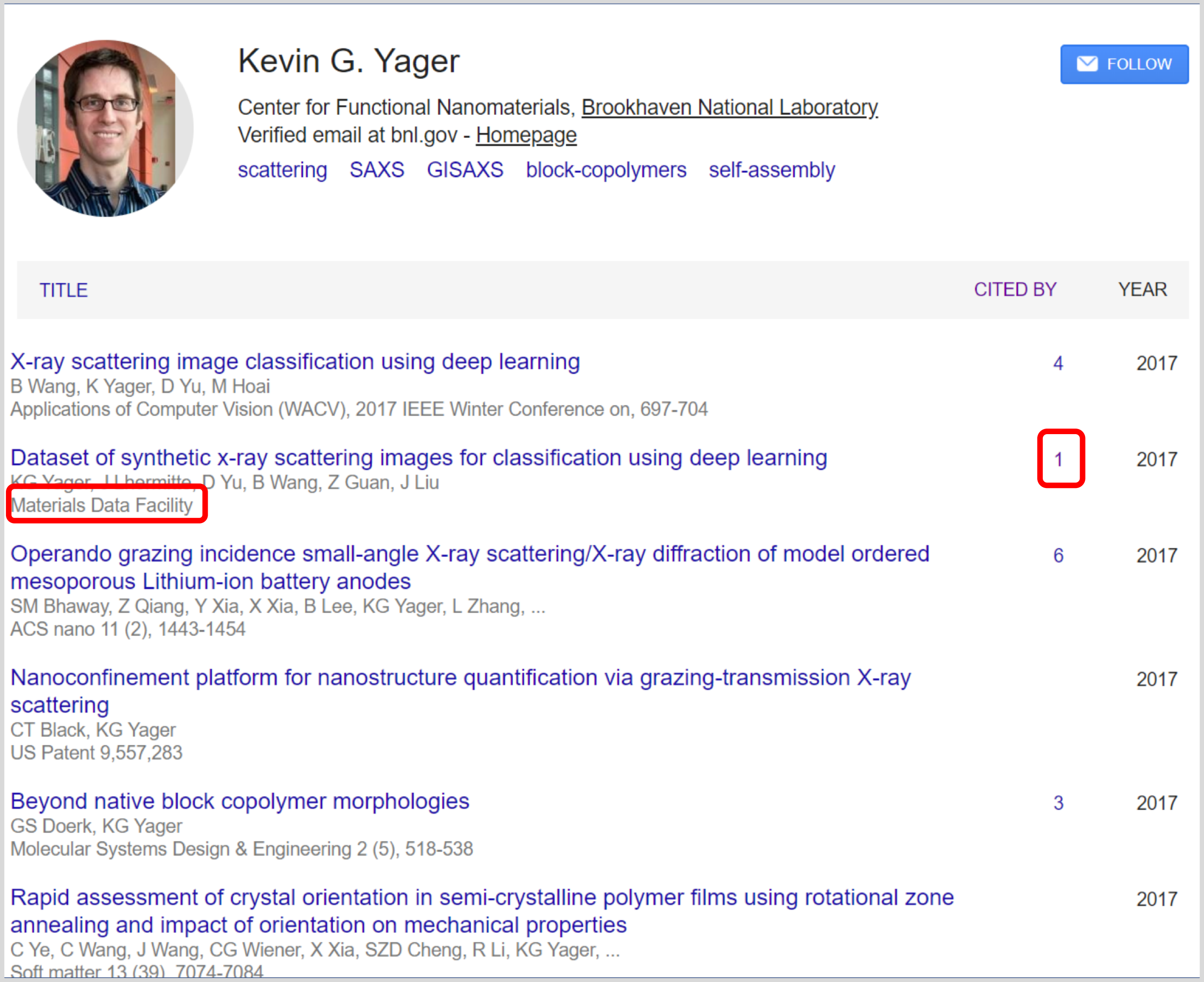
117 Data resources indexed

>3.4M Records

8 Repositories harvested

~ 200 Datasets

~ 300 TB Made discoverable



Kevin G. Yager

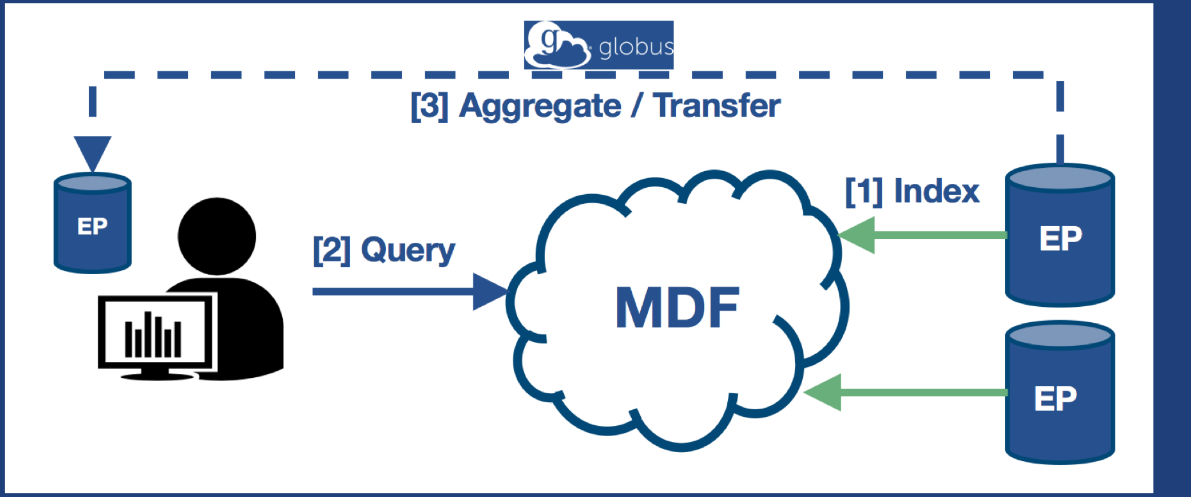
Center for Functional Nanomaterials, Brookhaven National Laboratory
Verified email at bnl.gov - Homepage

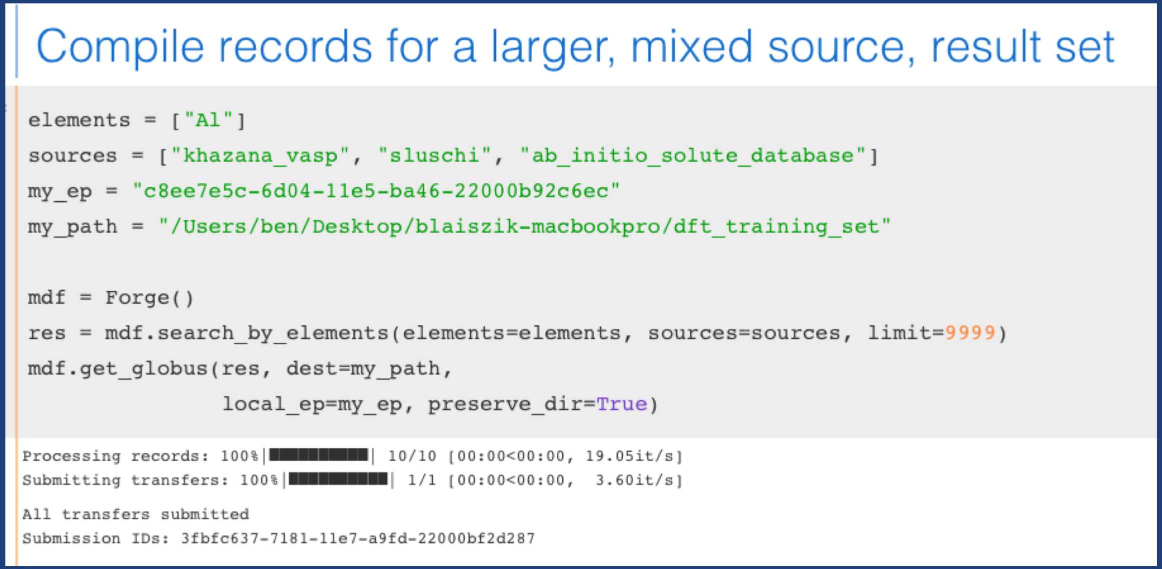
scattering SAXS GISAXS block-copolymers self-assembly

TITLE	CITED BY	YEAR
X-ray scattering image classification using deep learning B Wang, K Yager, D Yu, M Hoai Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, 697-704	4	2017
Dataset of synthetic x-ray scattering images for classification using deep learning KG Yager, J Lhermitte, D Yu, B Wang, Z Guan, J Liu Materials Data Facility	1	2017
Operando grazing incidence small-angle X-ray scattering/X-ray diffraction of model ordered mesoporous Lithium-ion battery anodes SM Bhaway, Z Qiang, Y Xia, X Xia, B Lee, KG Yager, L Zhang, ... ACS nano 11 (2), 1443-1454	6	2017
Nanoconfinement platform for nanostructure quantification via grazing-transmission X-ray scattering CT Black, KG Yager US Patent 9,557,283		2017
Beyond native block copolymer morphologies GS Doerk, KG Yager Molecular Systems Design & Engineering 2 (5), 518-538	3	2017
Rapid assessment of crystal orientation in semi-crystalline polymer films using rotational zone annealing and impact of orientation on mechanical properties C Ye, C Wang, J Wang, CG Wiener, X Xia, SZD Cheng, R Li, KG Yager, ... Soft matter 13 (39), 7074-7084		2017

MDF is enabling new science: Dataset mixing

Assemble Training Set





Compile records for a larger, mixed source, result set

```
elements = ["Al"]
sources = ["khazana_vasp", "sluschi", "ab_initio_solute_database"]
my_ep = "c8ee7e5c-6d04-11e5-ba46-22000b92c6ec"
my_path = "/Users/ben/Desktop/blaiszik-macbookpro/dft_training_set"

mdf = Forge()
res = mdf.search_by_elements(elements=elements, sources=sources, limit=9999)
mdf.get_globus(res, dest=my_path,
               local_ep=my_ep, preserve_dir=True)
```

```
Processing records: 100%|██████████| 10/10 [00:00<00:00, 19.05it/s]
Submitting transfers: 100%|██████████| 1/1 [00:00<00:00,  3.60it/s]

All transfers submitted
Submission IDs: 3fbfc637-7181-11e7-a9fd-22000bf2d287
```

Dataset Mixing

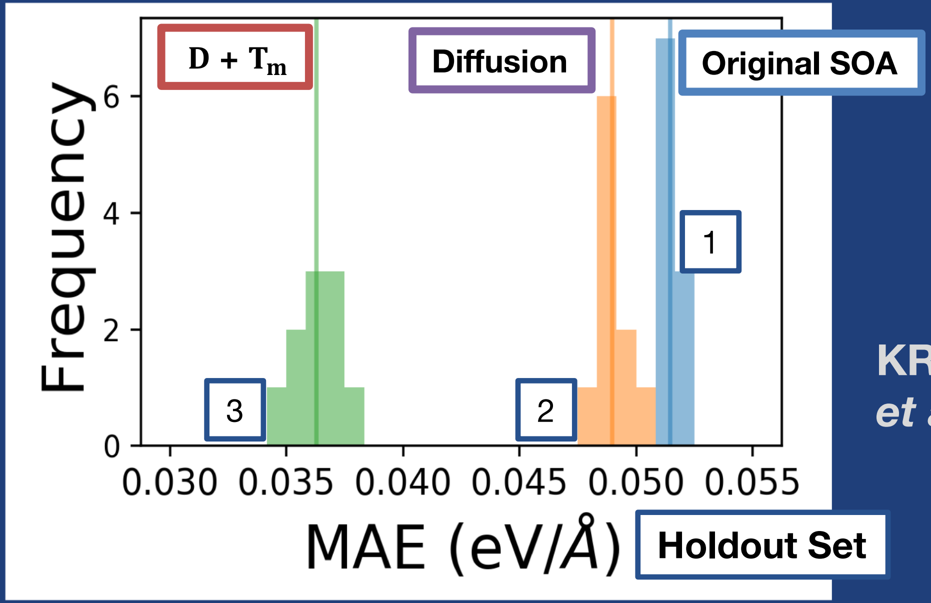
Logan Ward

Build force-field potentials from different datasets

Data resources: 3 Aluminum DFT datasets

1 dataset from khazana.uconn.edu, 2 from materialsdata.nist.gov

Result: Improved performance by integrating data sources

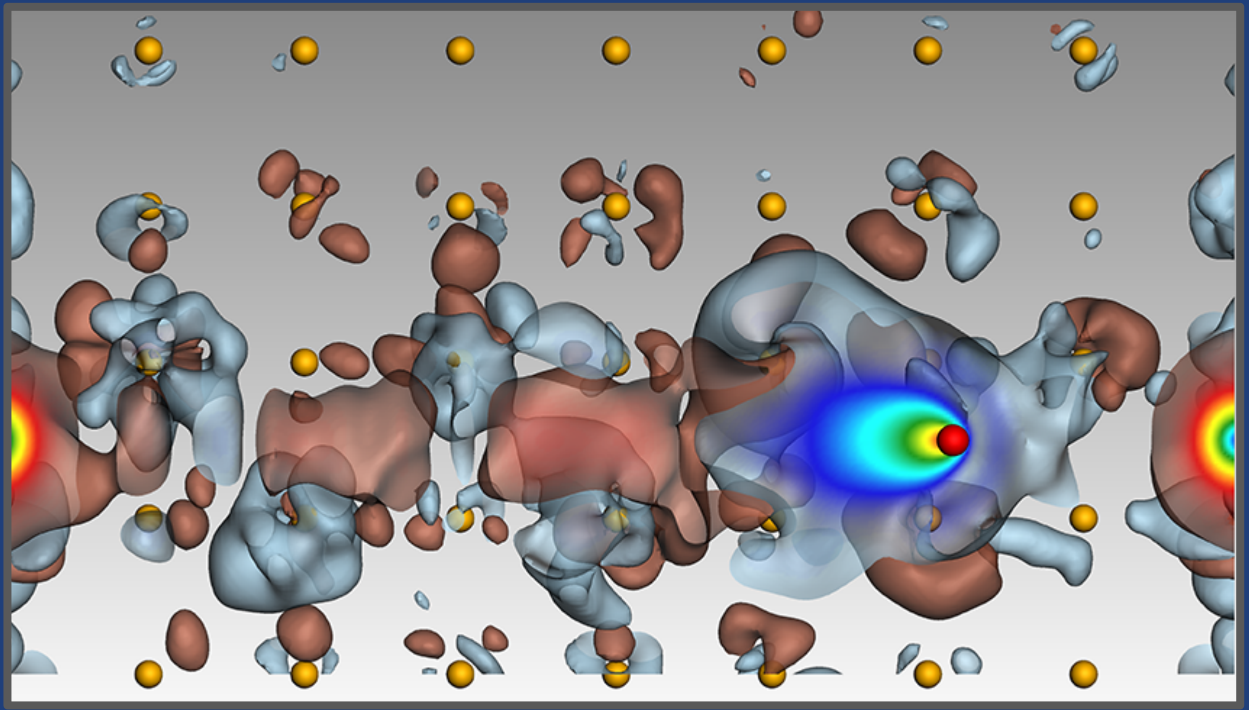


KRR Method: Botu *et al.* JPCC. (2017)

Better performance in original application: No new DFT calculations

MDF enables new science: Stopping power

Stopping Power: A “drag” force experienced by high speed protons, electrons, or positrons in a material



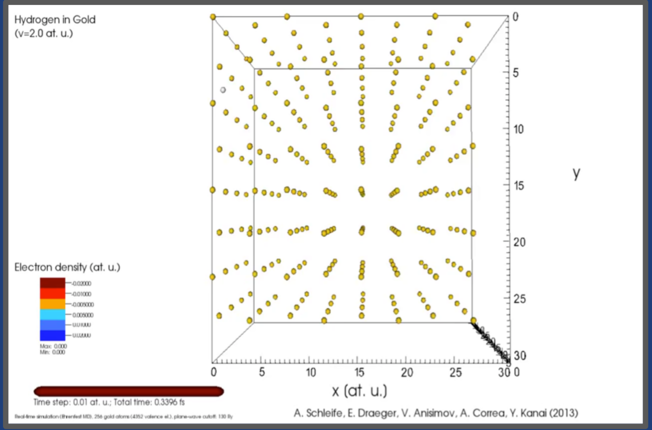
Areas of Application

- Nuclear reactor safety
- Magnetic confinement / inertial containment for nuclear fusion
- Solar cell surface adsorption
- Medicine (e.g., proton therapy treatment)
- Critical to understanding material radiation damage

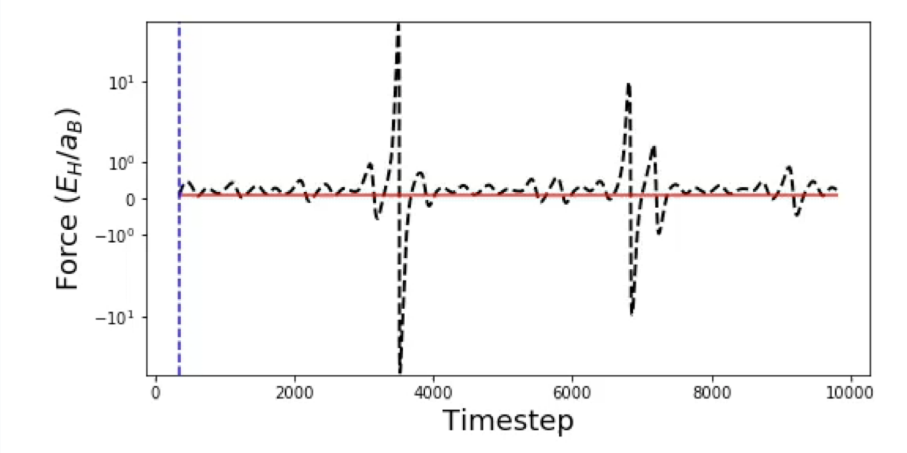
André Schleife and Cheng-Wei Lee (UIUC)
2016 ALCF INCITE Project
“Electronic Response to Particle Radiation in Condensed Matter”

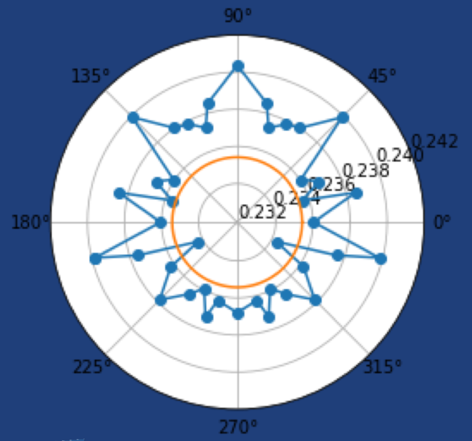
Connecting MDF and HPC

Index Data → **Existing Data** → **Direction-Dependent Stopping Power**



Hydrogen in Gold
(v=2.0 at. u.)
Electron density (at. u.)
-0.02000
-0.01000
-0.005000
0.005000
0.01000
0.02000
Max: 0.000
Min: 0.000
0
5
10
15
20
25
30
x (at. u.)
y
Time step: 0.01 at. u.; Total time: 0.3396 fs
Real-time simulation (Ehrenfest MD), 256 gold atoms (4352 valence el.), plane-wave cutoff: 130 Ry
A. Schleife, E. Draeger, V. Anisimov, A. Correa, Y. Kanai (2013)





ARGONNE LEADERSHIP COMPUTING FACILITY

PETREL
Data Management and Sharing Pilot

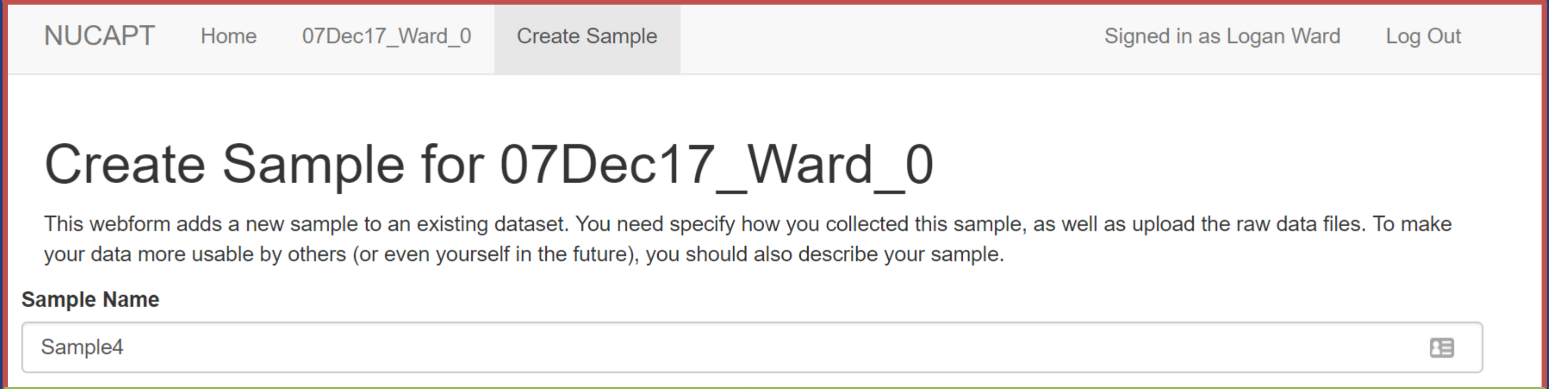
MATERIALS DATA FACILITY

Parsl

Existing Data → **Data Infrastructure** → **New Capabilities**

Northwestern University Center for Atom-Probe Tomography (NUCAPT) Integration

Simple webforms for data / metadata collections

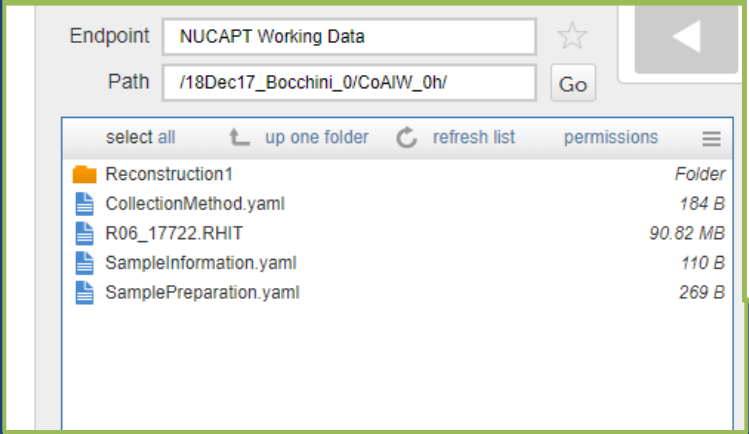


Create Sample for 07Dec17_Ward_0

This webform adds a new sample to an existing dataset. You need specify how you collected this sample, as well as upload the raw data files. To make your data more usable by others (or even yourself in the future), you should also describe your sample.

Sample Name: Sample4

Data accessible via Globus



Endpoint: NUCAPT Working Data

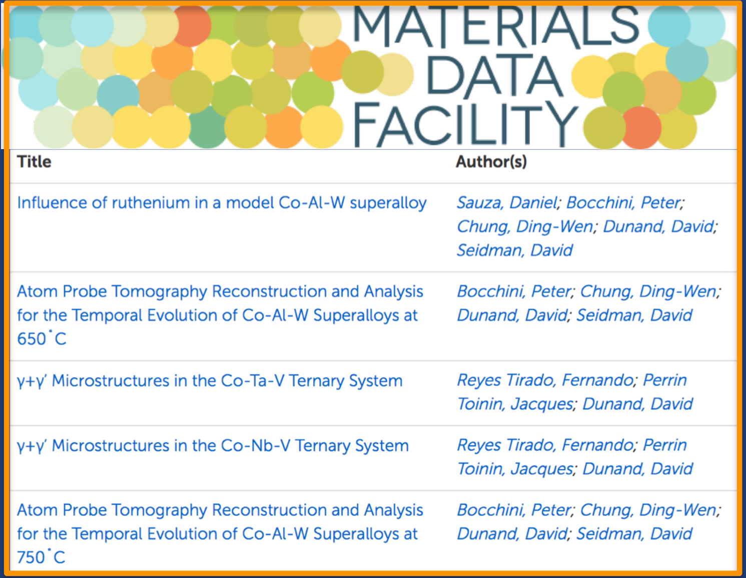
Path: /18Dec17_Bocchini_0/CoAlW_0h/

Reconstruction1	Folder
CollectionMethod.yaml	184 B
R06_17722.RHIT	90.82 MB
SampleInformation.yaml	110 B
SamplePreparation.yaml	269 B

Once on a Globus endpoint, data are…

- sharable to anyone with Globus account
- can be published to MDF with one click
- usable by automation tools

Publish



Title	Author(s)
Influence of ruthenium in a model Co-Al-W superalloy	*Sauza, Daniel; Bocchini, Peter; Chung, Ding-Wen; Dunand, David; Seidman, David*
Atom Probe Tomography Reconstruction and Analysis for the Temporal Evolution of Co-Al-W Superalloys at 650˚C	*Bocchini, Peter; Chung, Ding-Wen; Dunand, David; Seidman, David*
γ+γ′ Microstructures in the Co-Ta-V Ternary System	*Reyes Tirado, Fernando; Perrin Toinin, Jacques; Dunand, David*
γ+γ′ Microstructures in the Co-Nb-V Ternary System	*Reyes Tirado, Fernando; Perrin Toinin, Jacques; Dunand, David*
Atom Probe Tomography Reconstruction and Analysis for the Temporal Evolution of Co-Al-W Superalloys at 750˚C	*Bocchini, Peter; Chung, Ding-Wen; Dunand, David; Seidman, David*

CHiMaD

NUCAPT
Northwestern University Center for Atom-Probe Tomography

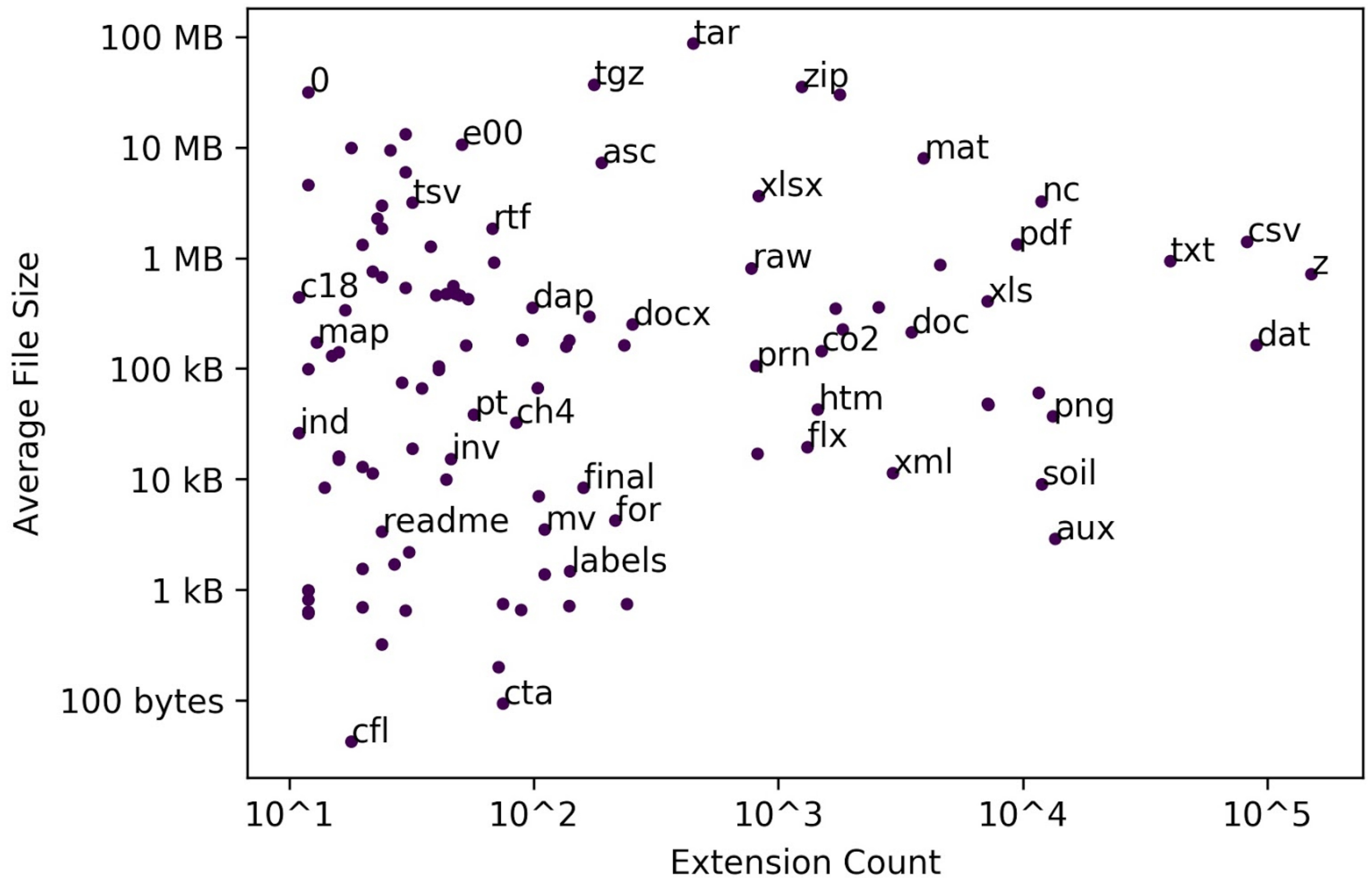
“Draining the data swamp”

File systems and data repositories often inconsistent & disorganized

Many file types:

- **Tabular**
- **Structured (JSON/XML)**
- **Images: photos, maps, plots**
- **Text: READMEs, papers, abstracts**
- **Not useful (?): Hadoop error logs, Windows installers, desktop shortcuts**

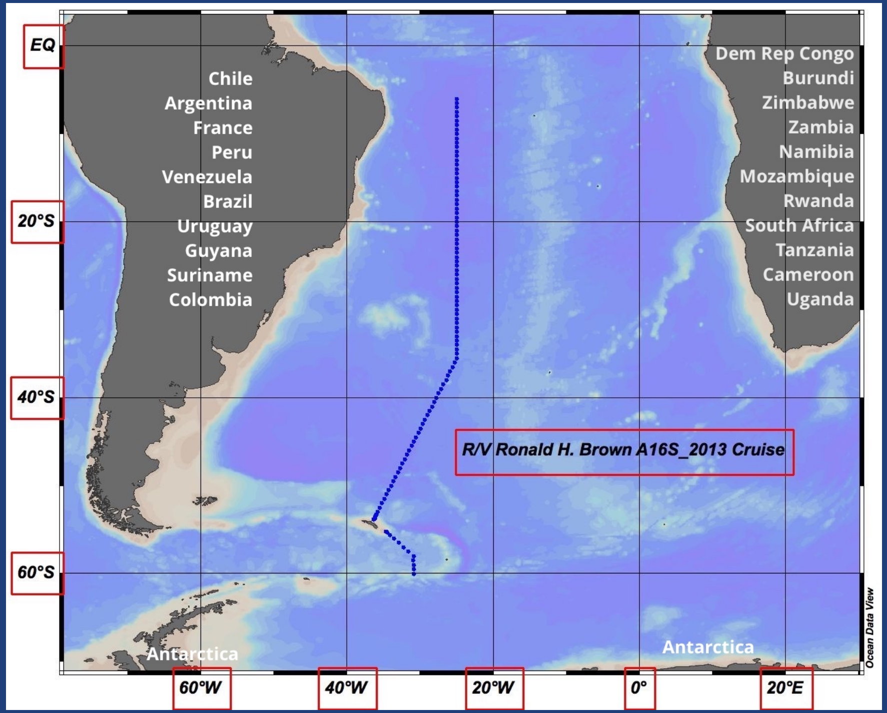
Assumption: extension =/= content

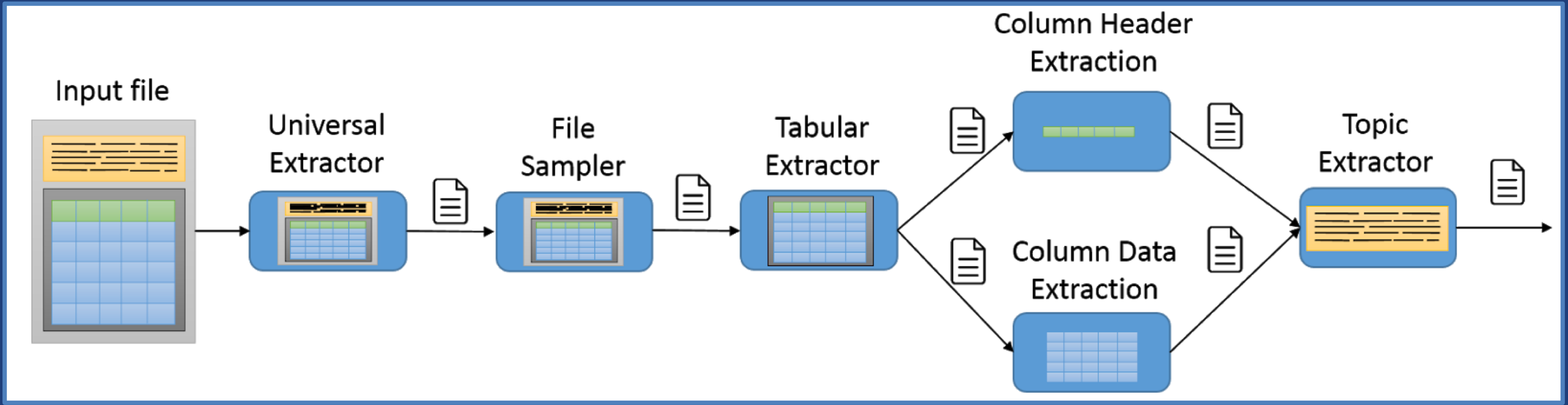


Carbon Dioxide Information Analysis Center (CDIAC) file extension frequency vs size for 500,001 files

Skluma: automated metadata extraction pipelines

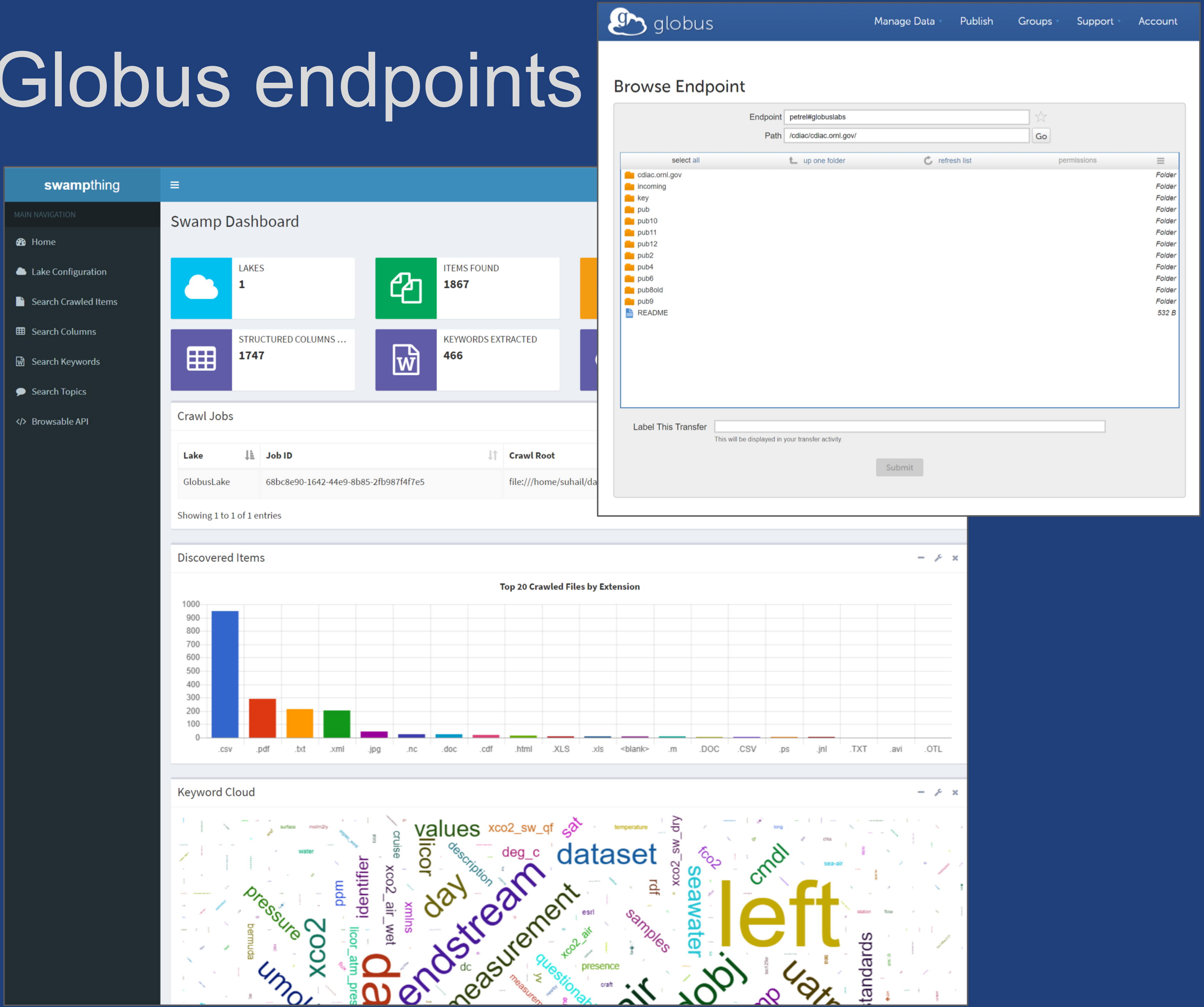
- Automatically crawls arbitrary file systems
- Constructs individualized extraction pipelines for each file
- Uses ML methods to determine extractors
- Handles *hybrid* files





Skluma and Globus endpoints

- **Input: Globus endpoint on a file system or repository**
- **Output: Globus search index of extracted metadata**
- **User interface for visualizing and managing the data swamp**



Browse Endpoint

Endpoint: petrel#globuslabs

Path: /cdiac/cdiac.ornl.gov/

cdiac.ornl.gov	Folder
incoming	Folder
key	Folder
pub	Folder
pub10	Folder
pub11	Folder
pub12	Folder
pub2	Folder
pub4	Folder
pub6	Folder
pub8old	Folder
pub9	Folder
README	532 B

Label This Transfer

This will be displayed in your transfer activity.

Swamp Dashboard

LAKES 1

ITEMS FOUND 1867

STRUCTURED COLUMNS ... 1747

KEYWORDS EXTRACTED 466

Crawl Jobs

Lake	Job ID	Crawl Root
GlobusLake	68bc8e90-1642-44e9-8b85-2fb987f4f7e5	file:///home/suhail/da

Showing 1 to 1 of 1 entries

Discovered Items

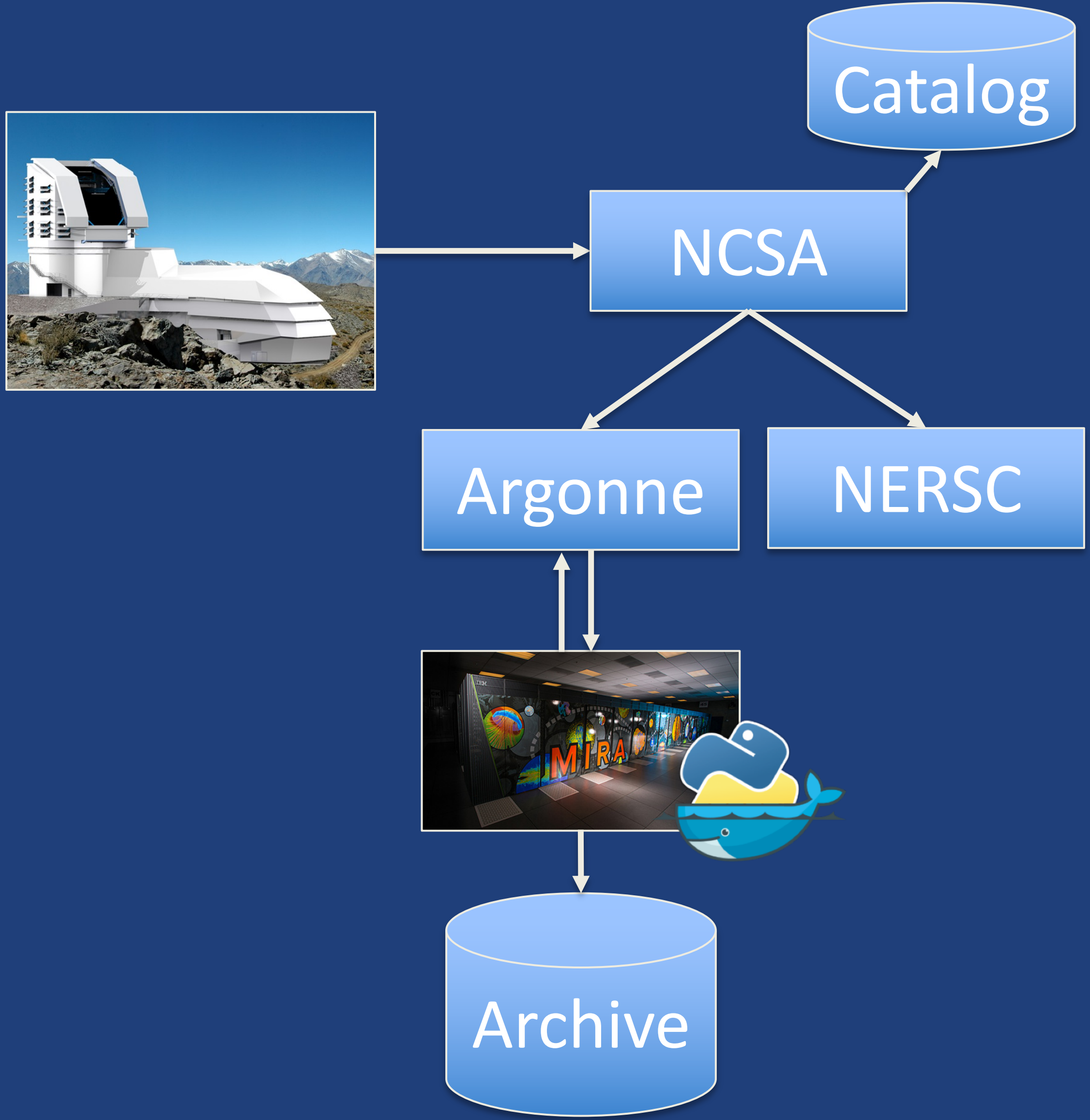
Top 20 Crawled Files by Extension

Keyword Cloud

Distributed research automation

Automation of scientific lifecycles:

- Data acquisition at different locations/times/instruments
- Analysis execution on distributed/heterogeneous resources
- Cataloging of descriptive metadata and provenance
- Dynamic collaborations around data and analysis



Catalog

NCSA

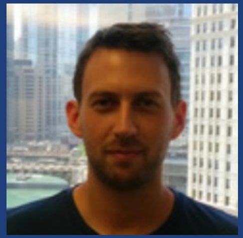
Argonne

NERSC

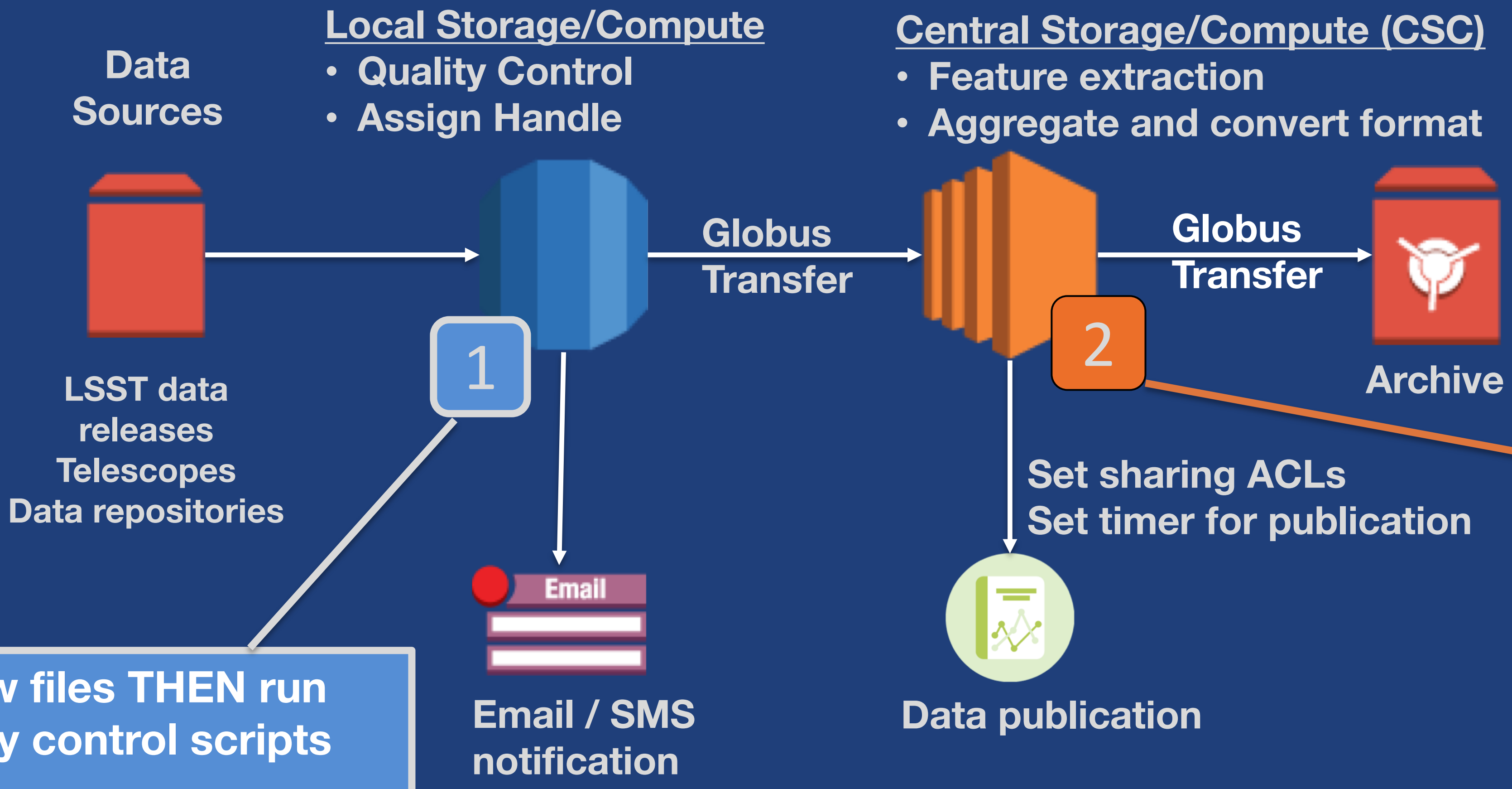
Archive

LSST data distribution and analysis pipeline

Encoding automation flows using trigger-action programming



Ryan Chard



Data Sources

LSST data releases
Telescopes
Data repositories

Local Storage/Compute

- Quality Control
- Assign Handle

Globus Transfer

Central Storage/Compute (CSC)

- Feature extraction
- Aggregate and convert format

Globus Transfer

Archive

1

2

Email / SMS notification

Set sharing ACLs
Set timer for publication

Data publication

- IF new files THEN run quality control scripts
- IF quality is good THEN send email AND transfer data to CSC

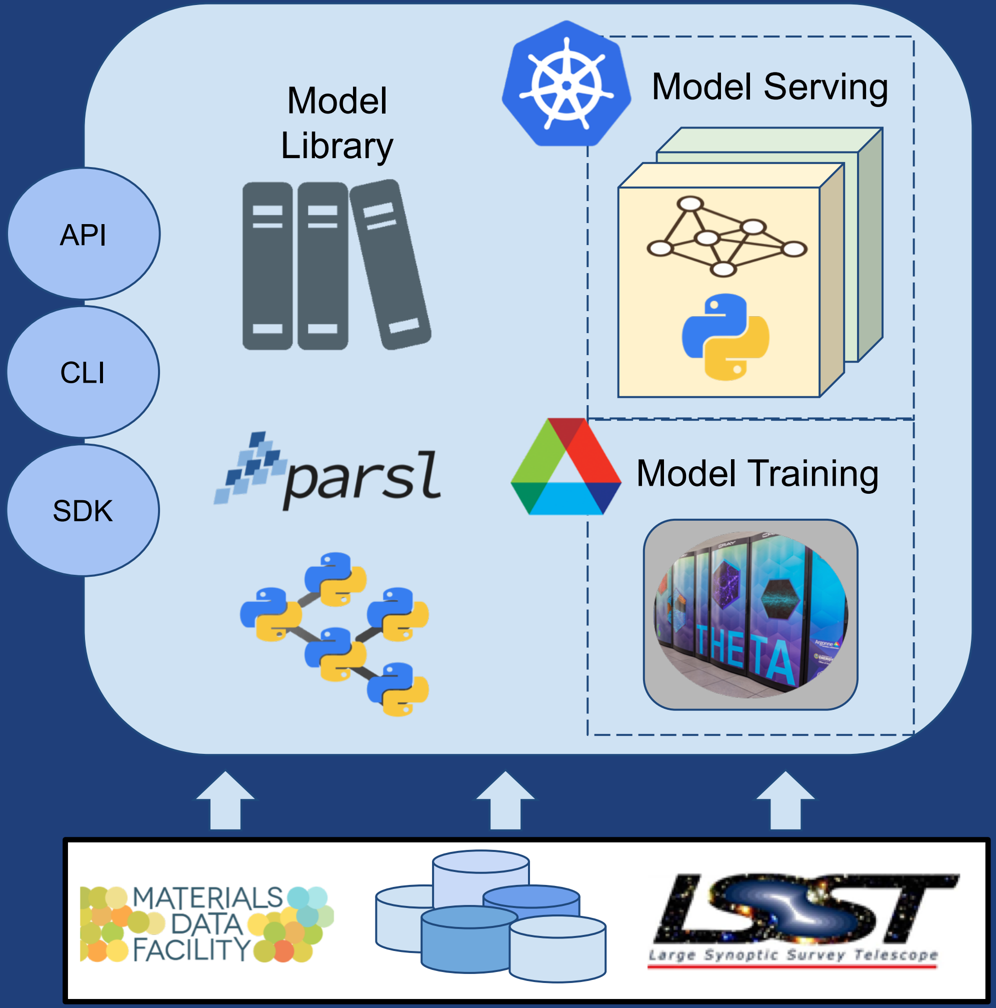
- IF new files THEN run feature extraction
- IF feature detected THEN transfer data to archival storage
- IF time since ingest > 6 months THEN publish dataset to MDF

Deep Learning Hub (DLHub)

Framework for finding, training, and serving machine learning and deep learning models

Scalable, container-based execution model

Globus integration to connect external repositories



Parallel Scripting in Python

Parsl

Parallel programming library for Python

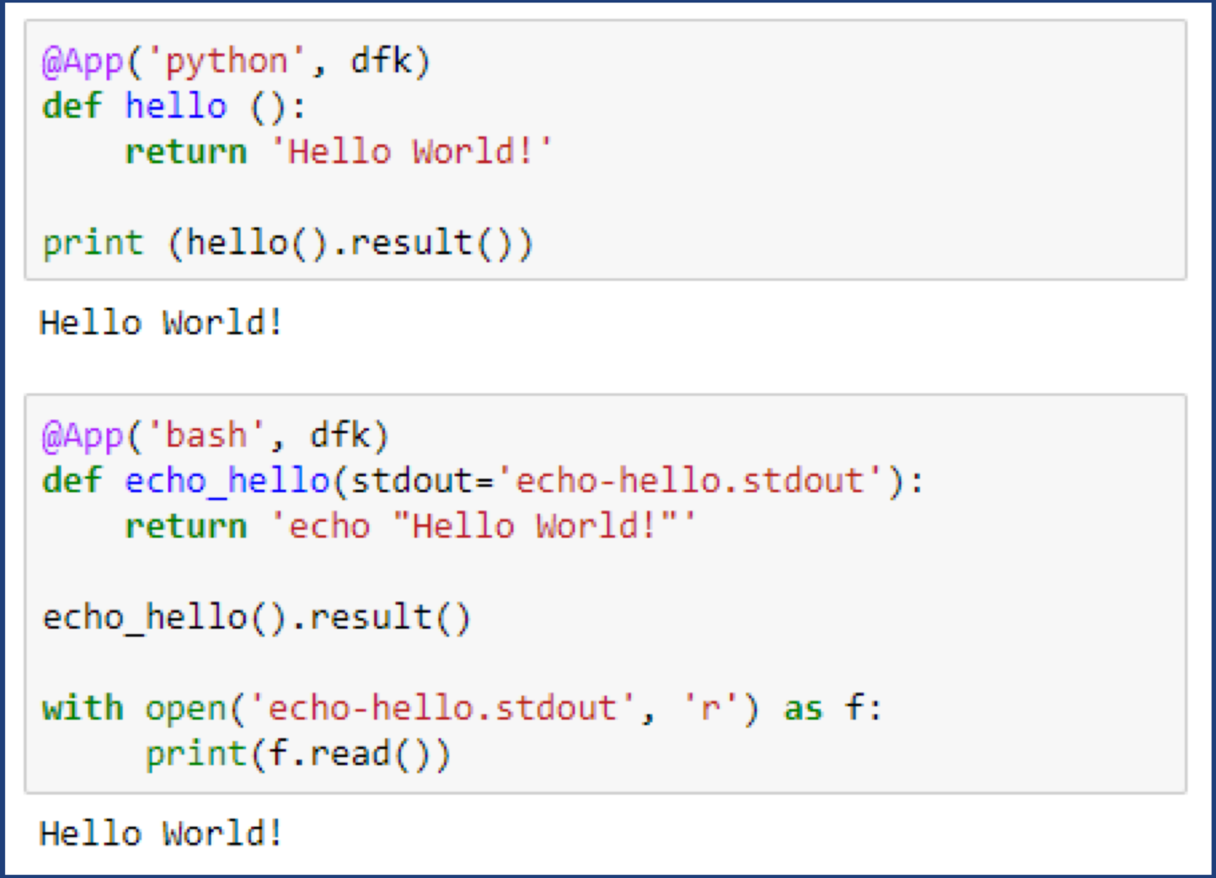
```
pip install parsl
```

Annotate functions to define Parsl *apps*

- Bash apps call external applications
- Python apps call Python functions

Apps run concurrently respecting data dependencies via futures. Natural parallel programming!

Parsl scripts are independent of where they run. Write once run anywhere!



```
@App('python', dfk)
def hello ():
    return 'Hello World!'

print (hello().result())
```

```
Hello World!
```

```
@App('bash', dfk)
def echo_hello(stdout='echo-hello.stdout'):
    return 'echo "Hello World!"'

echo_hello().result()

with open('echo-hello.stdout', 'r') as f:
     print(f.read())
```

```
Hello World!
```

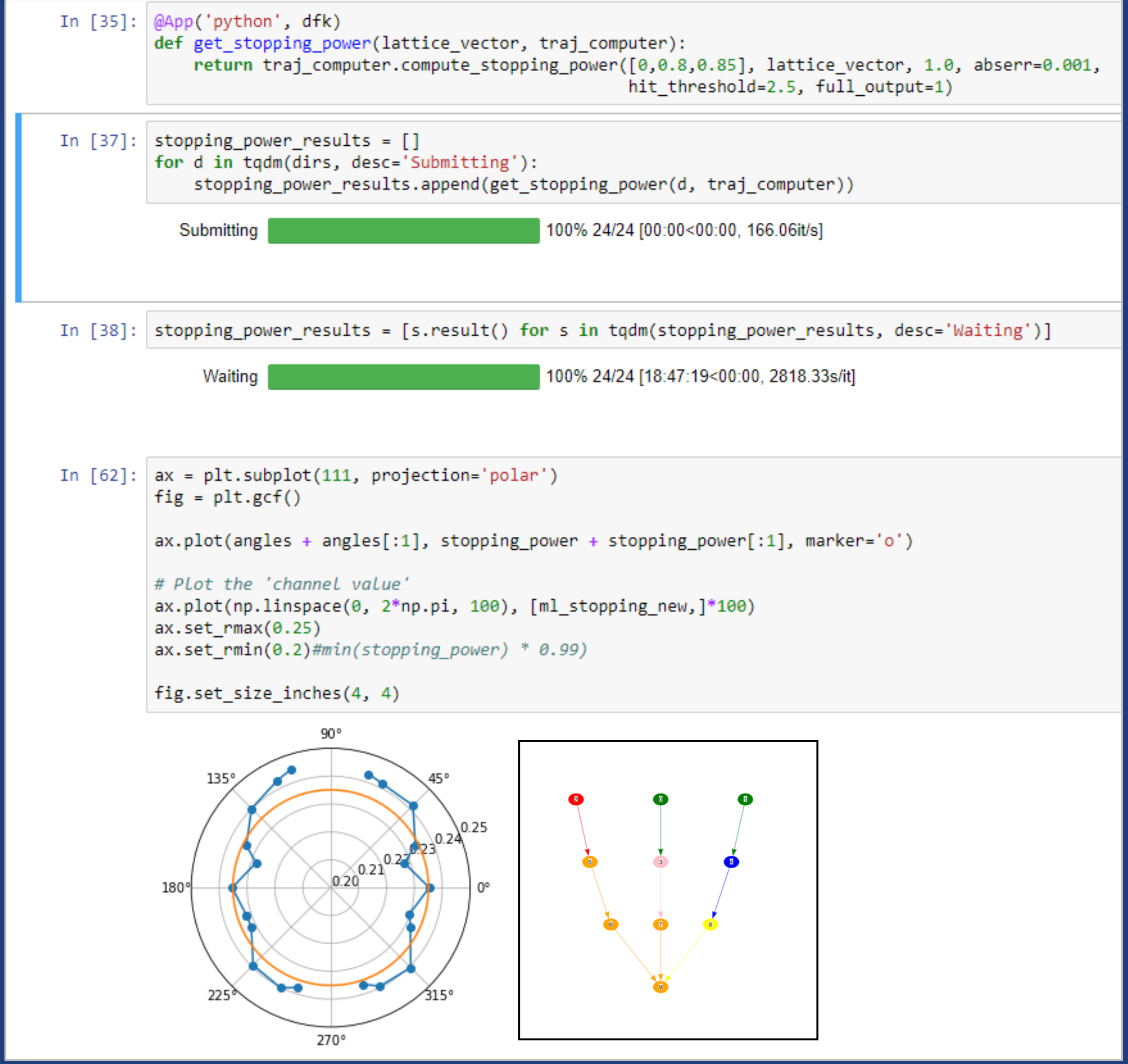
Interactive scalable computing in Jupyter

Large Synoptic Survey Telescope

PETREL
Data Management and Sharing Pilot

MATERIALS DATA FACILITY

Data servers



```
In [35]: @App('python', dfk)
         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))
```

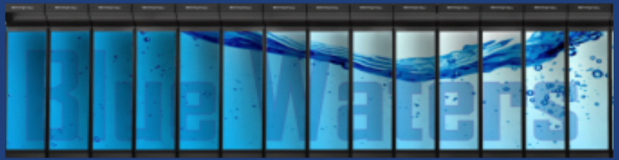
```
Submitting 100% 24/24 [00:00<00:00, 166.06it/s]
```

```
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)
```

- Globus integration (Auth and transfer)
- Multi-site execution
- Exec provider/model independent
- Automated elasticity
- Container support
- Fault tolerance
- Python/Jupyter

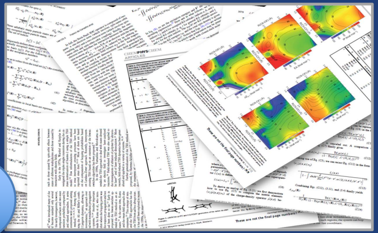


amazon web services

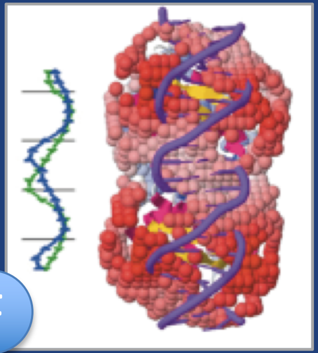
XSEDE
Extreme Science and Engineering Discovery Environment

Parsl adoption across sciences

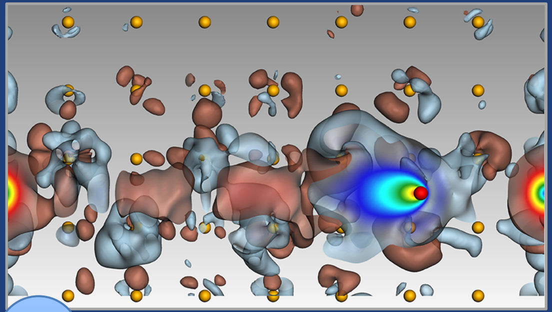
A Simulating galaxy formation using sky surveys



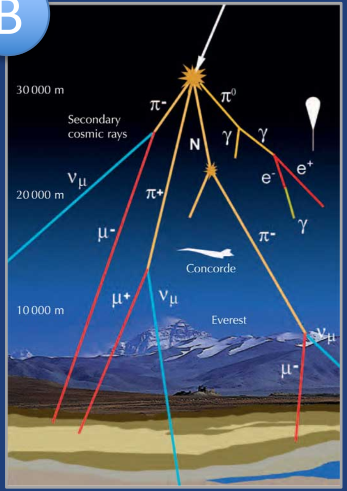
E Information extraction



F Biomolecule modelling

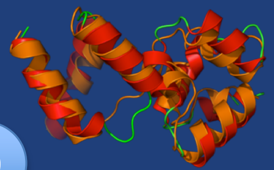


H Computational calculation of stopping power



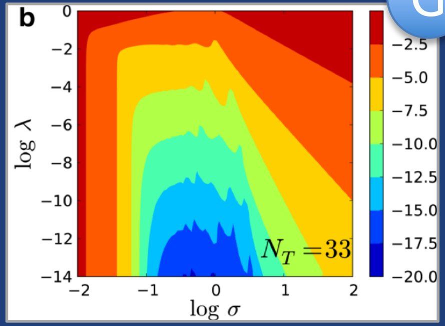
B Analysis of Cosmic ray showers

C Real-time computing (e.g., APS image reconstruction)

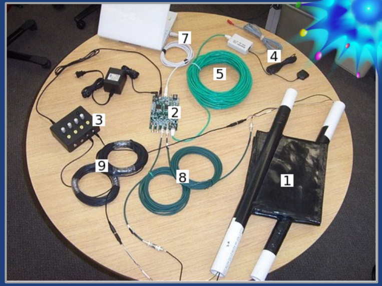


D Protein docking

Machine learning and data analytics



G



I Education programs in Jupyter notebooks

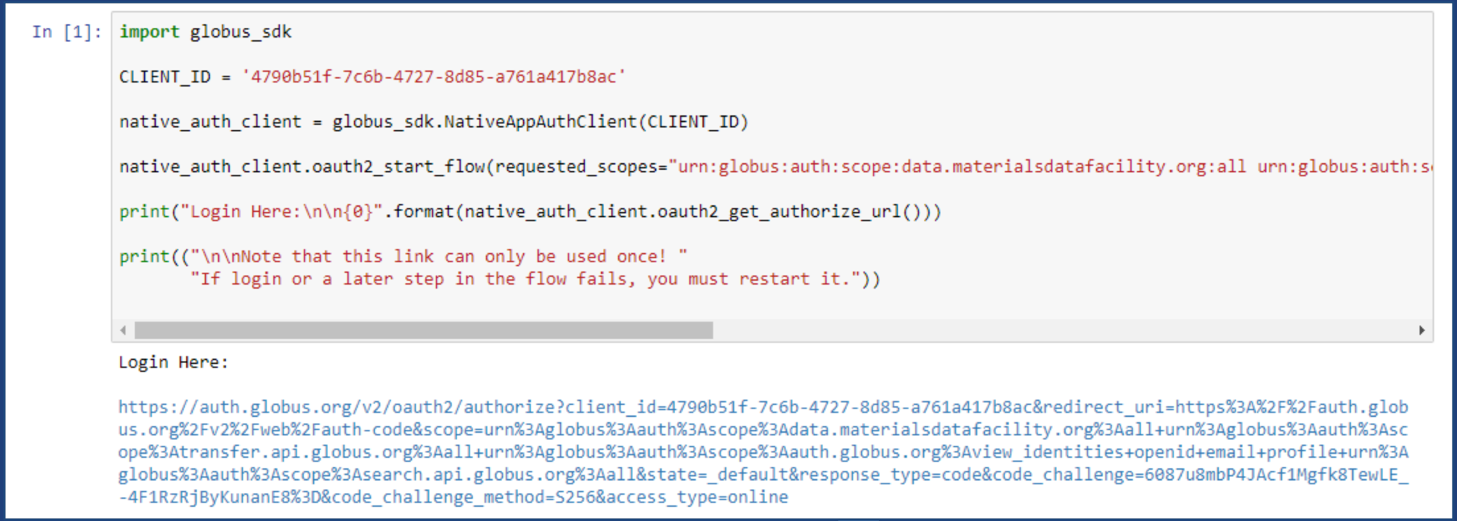
Parsl + Globus + Jupyter

Native app integration to provide embedded access to Globus (and other) services

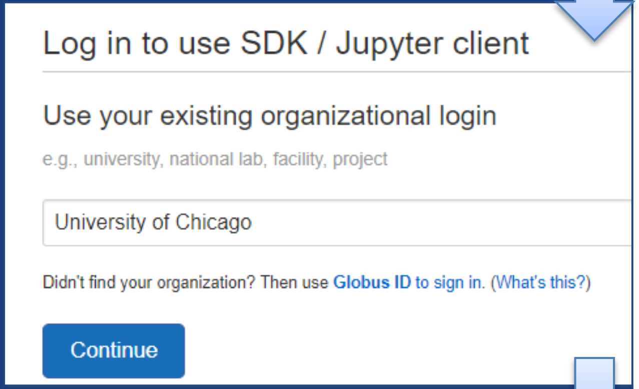
Transparent SSH-based authentication to compute resources (soon)

High-speed data access and **reliable data movement** to/from repositories, laptops, supercomputers, …

Staging to/from DTNs



```
In [1]: import globus_sdk

        CLIENT_ID = '4790b51f-7c6b-4727-8d85-a761a417b8ac'

        native_auth_client = globus_sdk.NativeAppAuthClient(CLIENT_ID)

        native_auth_client.oauth2_start_flow(requested_scopes="urn:globus:auth:scope:data.materialsdatafacility.org:all urn:globus:auth:s

        print("Login Here:\n\n{0}".format(native_auth_client.oauth2_get_authorize_url()))

        print(("\n\nNote that this link can only be used once! "
               "If login or a later step in the flow fails, you must restart it."))

        Login Here:

        https://auth.globus.org/v2/oauth2/authorize?client_id=4790b51f-7c6b-4727-8d85-a761a417b8ac&redirect_uri=https%3A%2F%2Fauth.glob
        us.org%2Fv2%2Fweb%2Fauth-code&scope=urn%3Aglobus%3Aauth%3Ascope%3Adata.materialsdatafacility.org%3Aall+urn%3Aglobus%3Aauth%3Asc
        ope%3Atransfer.api.globus.org%3Aall+urn%3Aglobus%3Aauth%3Ascope%3Aauth.globus.org%3Aview_identities+openid+email+profile+urn%3A
        globus%3Aauth%3Ascope%3Asearch.api.globus.org%3Aall&state=_default&response_type=code&code_challenge=6087u8mbP4JAcf1Mgfk8TewLE_
        -4F1RzRjByKunanE8%3D&code_challenge_method=S256&access_type=online
```



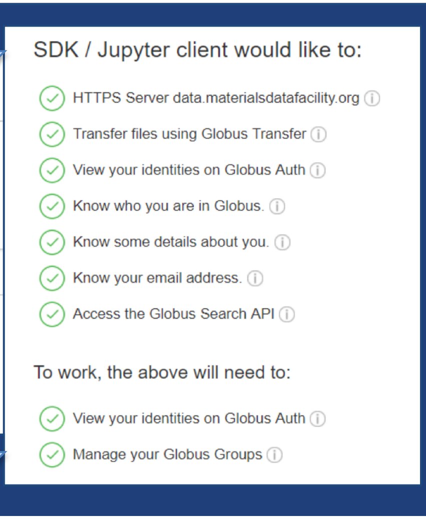
Log in to use SDK / Jupyter client

Use your existing organizational login

e.g., university, national lab, facility, project

University of Chicago

Didn't find your organization? Then use Globus ID to sign in. (What's this?)

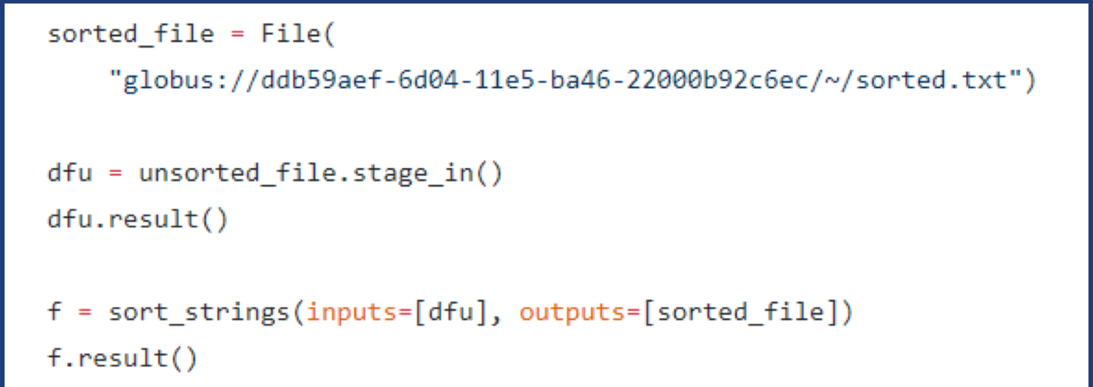


SDK / Jupyter client would like to:

- ✓ HTTPS Server data.materialsdatafacility.org ⓘ
- ✓ Transfer files using Globus Transfer ⓘ
- ✓ View your identities on Globus Auth ⓘ
- ✓ Know who you are in Globus. ⓘ
- ✓ Know some details about you. ⓘ
- ✓ Know your email address. ⓘ
- ✓ Access the Globus Search API ⓘ

To work, the above will need to:

- ✓ View your identities on Globus Auth ⓘ
- ✓ Manage your Globus Groups ⓘ



```
sorted_file = File(
    "globus://ddb59aef-6d04-11e5-ba46-22000b92c6ec/~/sorted.txt")

dfu = unsorted_file.stage_in()
dfu.result()

f = sort_strings(inputs=[dfu], outputs=[sorted_file])
f.result()
```

globus labs -- labs.globus.org

thank you

Program Preview

globusworld.org/conf/program

- **Today**
  - Lightning talks
  - Guest keynote: Alex Szalay, Building the Open Storage Network
  - Reception
- **Tomorrow**
  - Tutorials
  - Office Hours
- **Friday morning**
  - Customer forum

#globus2018

@globus