Computation Institute

Big process for big data

Process automation for data-driven science

Ian Foster

Computation Institute

Argonne National Laboratory & The University of Chicago

Talk at Astroinformatics 2012, Redmond, September 10, 2012

Researchers struggle with data

More data, more complex data
Ad-hoc solutions
Inadequate software, hardware
Data plan mandates

Complexity is large and growing

Time

Run experiment
Collect data
Move data
Check data
Annotate data
Share data
Find similar data
Link to literature
Analyze data
Publish data

The challenge of staying competitive

"Well, in our country," said Alice … "you'd generally get to somewhere else — if you run very fast for a long time, as we've been doing."

"A slow sort of country!" said the Queen. "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"

Arlington, VA, June 2012

Jun 10 - Jun 14, 2012 - Arlington, VA
Arlington, VA; Boston, MA

Edit Trip

Travelers	Ian Foster	+ Add Travelers
Non-Travelers	Brigitte Raumann , asmyth1	Share | Manage
Visible to [?]	TripIt connections, TripIt Groups	Privacy settings
Who's close	Daniel S Katz, Jennifer M Schopf (+3 others)	View all
Trip Description	Add a Description	

Trip Cost: $1,736.10 [?]

+ Add plans Export to calendar More ▸

Offers for Your Trip

- Cleveland Park: $35 for a deep tissue massage at Facials by Camille
- Wine Tasting Pedicab Tour For Two People
- One ($24) or 12 ($300) Tickets to North End Pizza Tour
- One ($39) or Two ($69) Acupuncture Sessions with Consultation
- Entry for One, Two, or Four to the CitySolve Urban Race on [...]

See more offers »

Itinerary: **Expand** | Collapse **Details** | Map

Sun, Jun 10

Boston, MA - Avg: Hi 79°F / Lo 57°F + Add Plans

4:03 PM CDT

Arrived - On Time

Chicago (ORD) to Boston (BOS)

Options

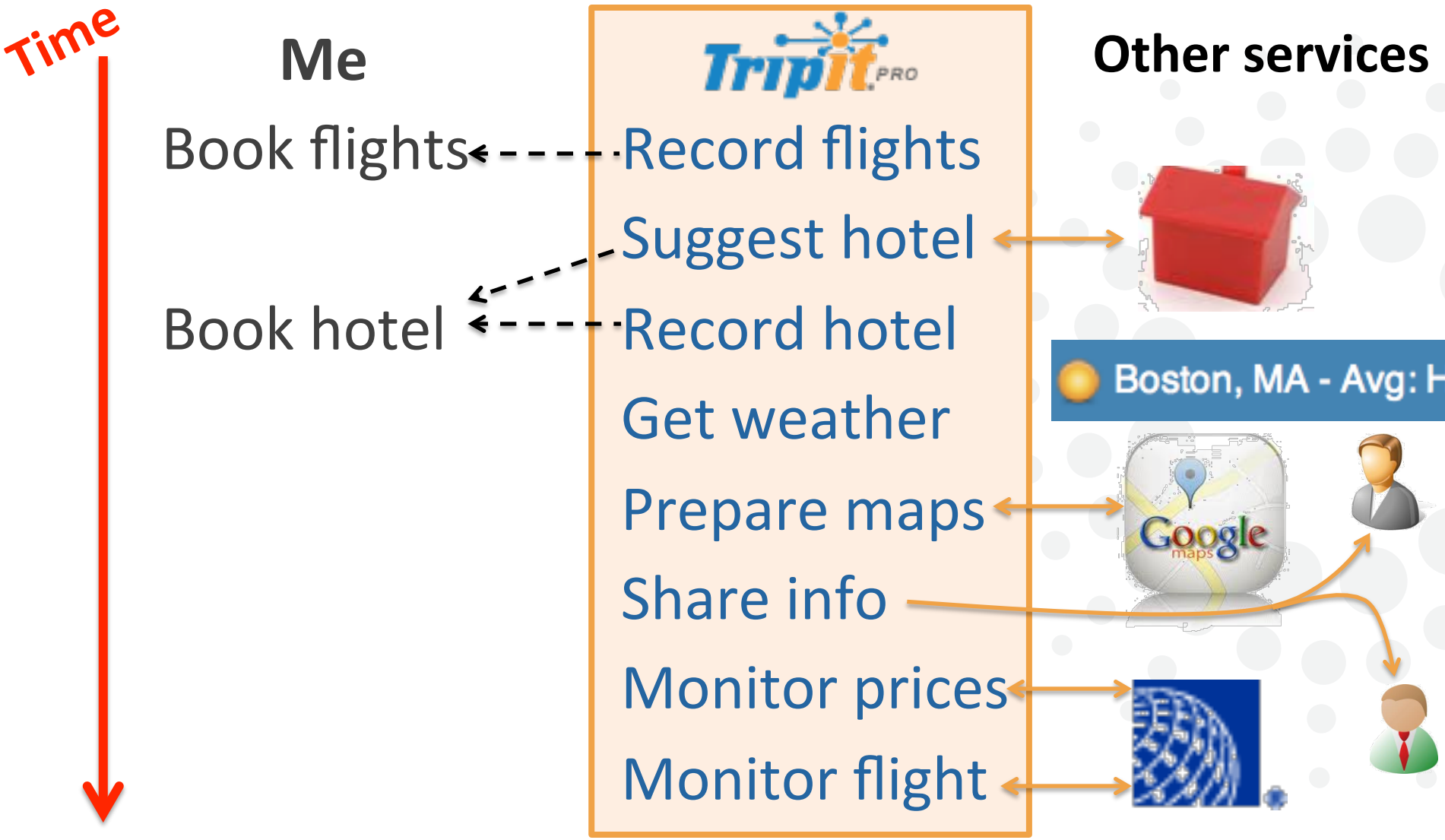
United Airlines 349 - Conf # NZVRFM
Aircraft Airbus A319
nonstop 2h, 21m 864 mi E
Purchase

Depart: Chicago (ORD), 4:03pm CDT, terminal 1, gate C17
Arrive: Boston (BOS), **7:22pm EDT** *(orig arr time: 7:24pm)*, terminal C, gate C17

Passenger
Iant Foster **FF #**ABL3XXXX **Ticket #**0162328680876

Booking Information
Booked on United 5/25/2012
http://www.united.com/

Tripit exemplifies process automation



Software as a Service (Gartner)

1. The application is owned, delivered, and managed remotely by one or more providers
2. The application is based on a single code base that is consumed in a one-to-many model by all contracted customers at any time
3. The application is licensed on pay-per-use or subscription basis

---

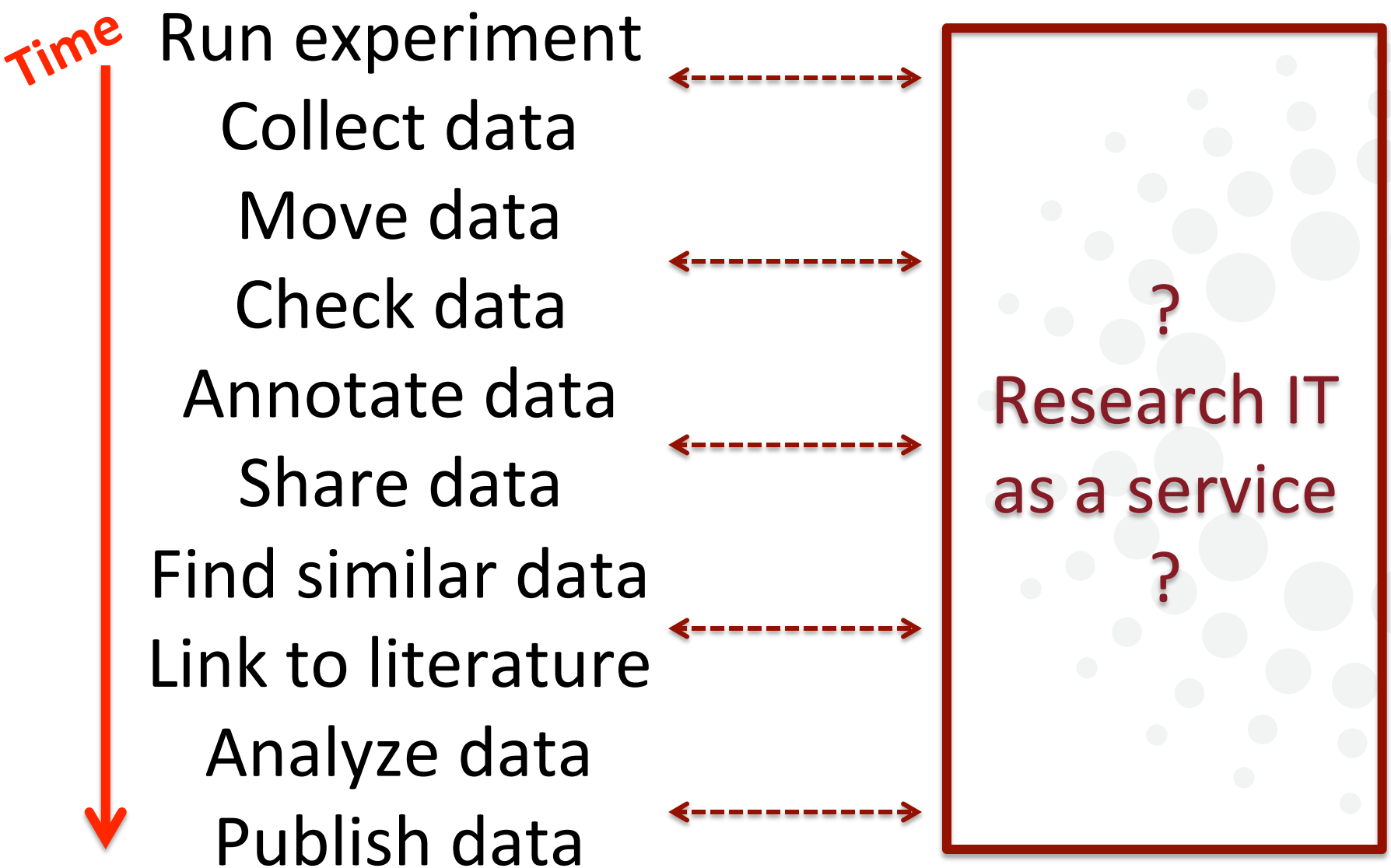
4. The application behind the service is properly web architected—not an existing application web enabled [D. Terrar]

Complexity is large and growing

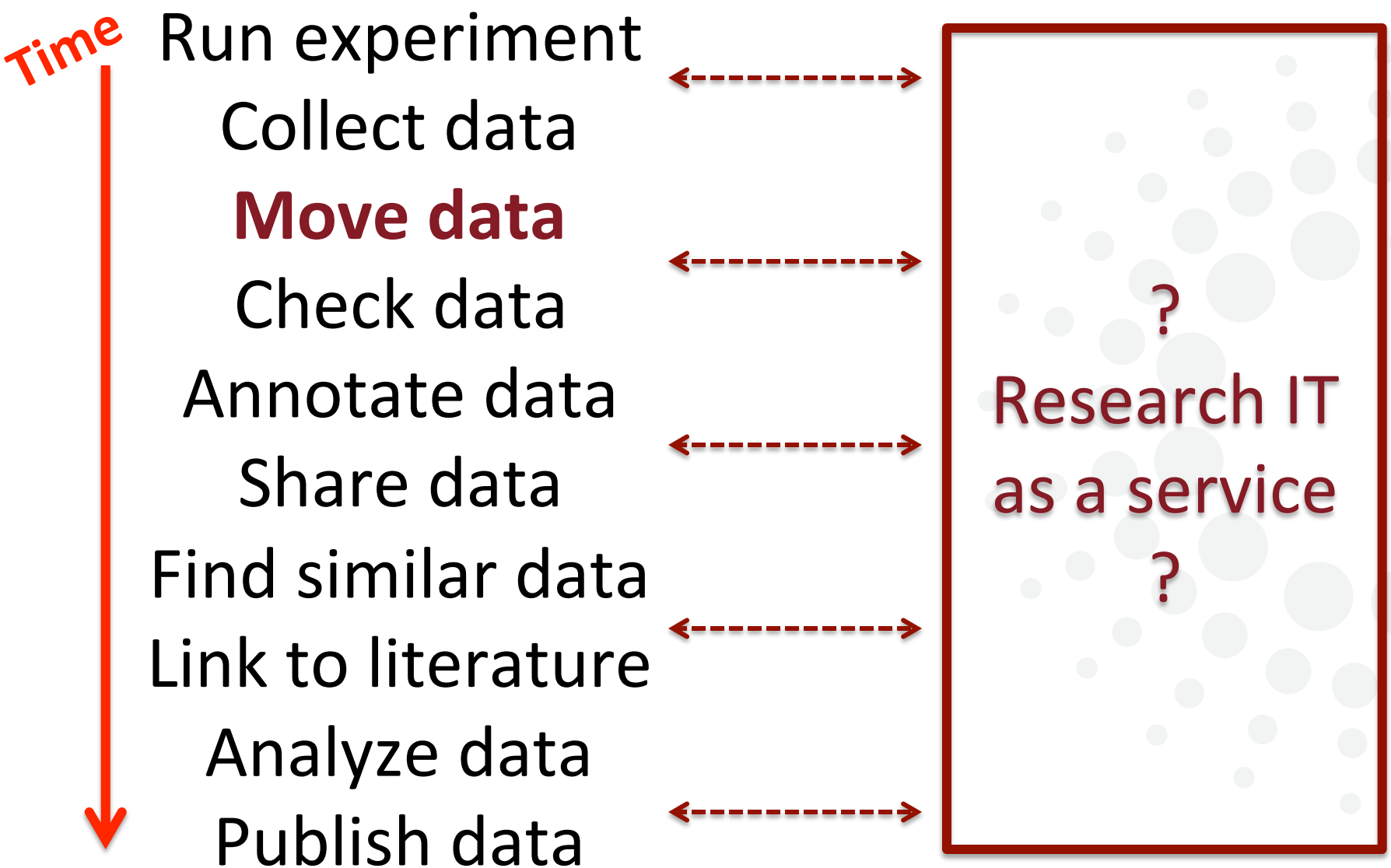
Time

Run experiment
Collect data
Move data
Check data
Annotate data
Share data
Find similar data
Link to literature
Analyze data
Publish data

Process automation for science

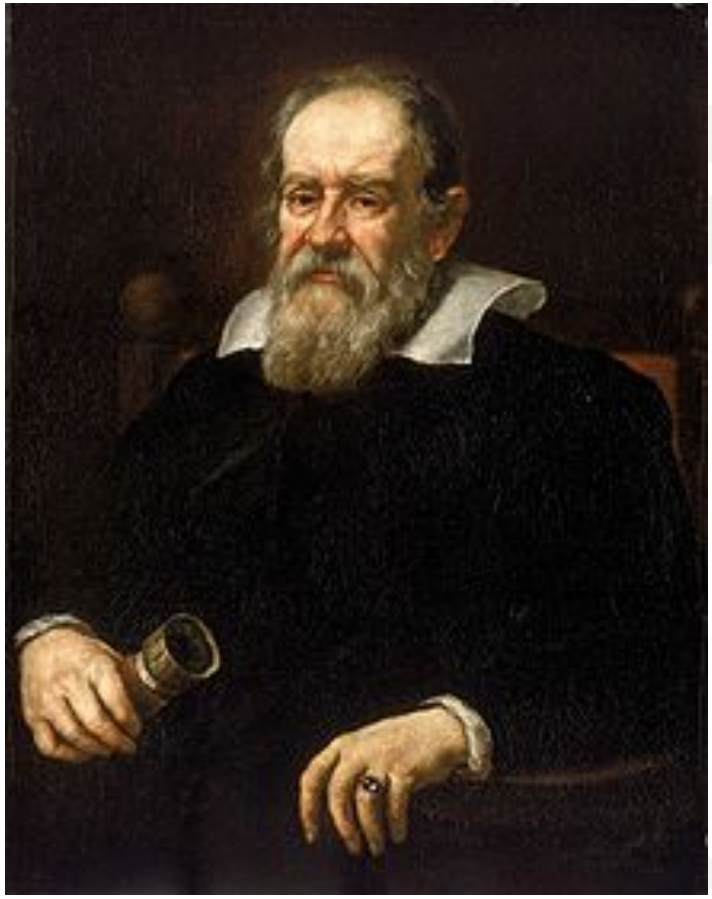


Process automation for science



Eppur si muove ("and yet it moves") …

Galileo said this about the **earth** (the earth is not the center of the universe)



Same observation holds about **data**, which is often in the "wrong place" for various reasons

Reliable, high-performance, secure file transfer.

Move files fast. No IT required.

+ WATCH A VIDEO

Globus Online in a nutshell

GET STARTED

Sign up and get moving

6,723,492,472 MB TRANSFERRED

Why Use Globus Online?

See how easy file transfer can be

For HPC Resource Owners

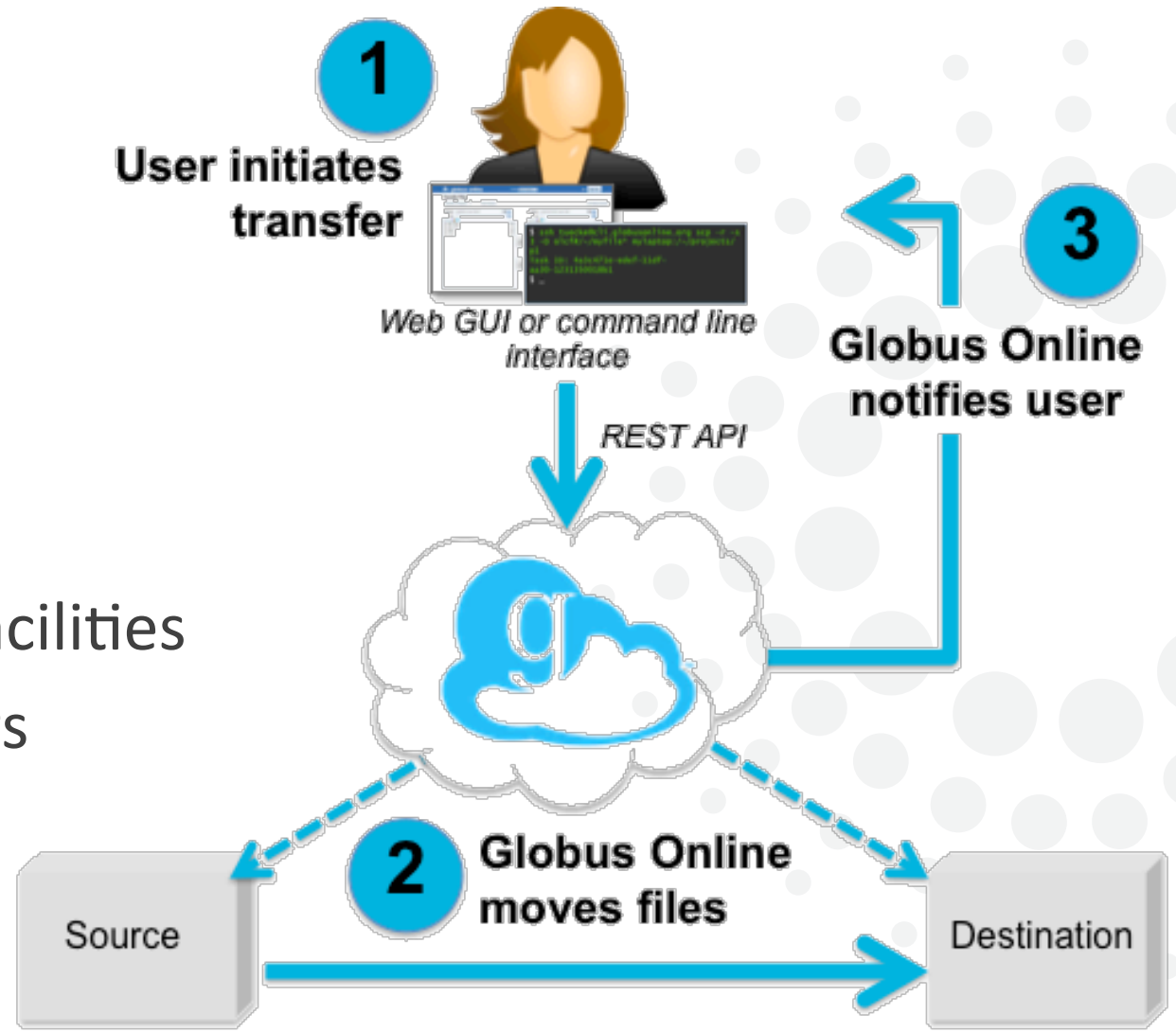
Enable Globus Online for your users

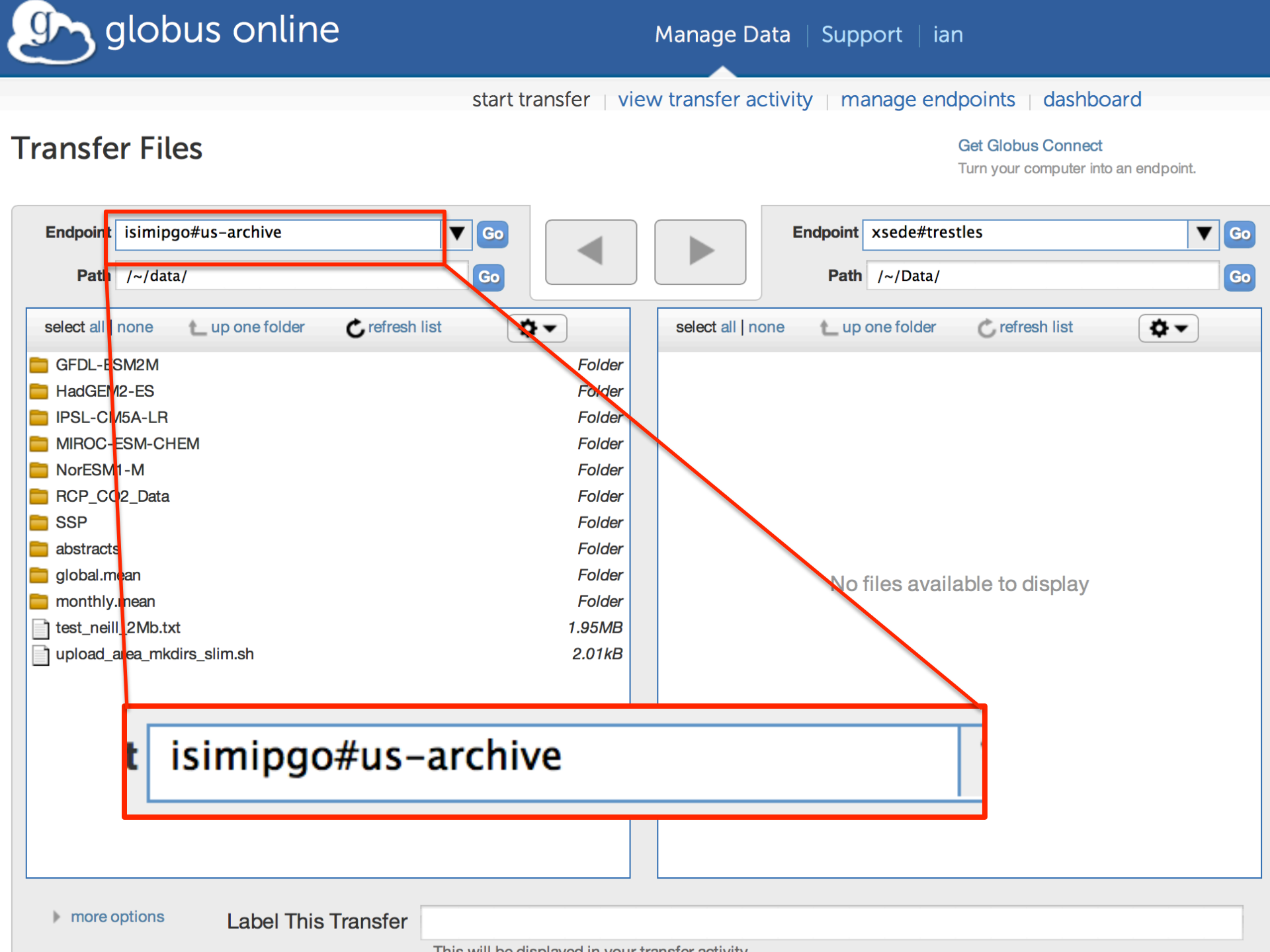
For Developers

Integrate with Globus Online

Globus Transfer details

- In 20 months
  - 6,000 users
  - 6 PB moved
  - 500M files
  - 99.9% uptime
- Broad adoption
  - Experimental facilities
  - Supercomputers
  - Campuses
  - Individuals
  - Projects




globus online
Manage Data | Support | ian
start transfer | view transfer activity | manage endpoints | dashboard
Transfer Files
Get Globus Connect
Turn your computer into an endpoint.
Endpoint isimipgo#us-archive Go
Path /~/data/ Go
Endpoint xsede#trestles Go
Path /~/Data/ Go
select all | none
up one folder
refresh list
GFDL-ESM2M Folder
HadGEM2-ES Folder
IPSL-CM5A-LR Folder
MIROC-ESM-CHEM Folder
NorESM1-M Folder
RCP_CO2_Data Folder
SSP Folder
abstracts Folder
global.mean Folder
monthly.mean Folder
test_neill_2Mb.txt 1.95MB
upload_area_mkdirs_slim.sh 2.01kB
select all | none
up one folder
refresh list
No files available to display
isimipgo#us-archive
more options
Label This Transfer

Transfer Files

Get Globus Connect
Turn your computer into an endpoint.

Endpoint isimipgo#us-archive Go

Path /~/data/ Go

select all | none up one folder refresh list

GFDL-ESM2M	*Folder*
HadGEM2-ES	*Folder*
IPSL-CM5A-LR	*Folder*
MIROC-ESM-CHEM	*Folder*
NorESM1-M	*Folder*
RCP_CO2_Data	*Folder*
SSP	*Folder*
abstracts	*Folder*
global.mean	*Folder*
monthly.mean	*Folder*
test_neill_2Mb.txt	*1.95MB*
upload_area_mkdirs_slim.sh	*2.01kB*

Endpoint xsede#trestles Go

Path /~/Data/ Go

select all | none up one folder refresh list

No files available to display

more options

Label This Transfer

This will be displayed in your transfer activity.

globus online

Manage Data | Support | ian

start transfer | view transfer activity | manage endpoints | dashboard

Transfer Request Submitted Successfully. Task ID:01fab246-dfe8-11e1-bf56-1231380b8963

Transfer Files

Get Globus Connect

Turn your computer into an endpoint.

Endpoint isimipgo#us-archive Go

Path /~/data/ Go

select all | none up one folder refresh list

GFDL-ESM2M	*Folder*
HadGEM2-ES	*Folder*
IPSL-CM5A-LR	*Folder*
MIROC-ESM-CHEM	*Folder*
NorESM1-M	*Folder*
RCP_CO2_Data	*Folder*
SSP	*Folder*
abstracts	*Folder*
global.mean	*Folder*
monthly.mean	*Folder*
test_neill_2Mb.txt	*1.95MB*
upload_area_mkdirs_slim.sh	*2.01kB*

Endpoint xsede#trestles

Path /~/Data/

select all | none up one folder refresh list

No files available to display

Transfer Activity

Cancel

1 of 6

View 25 Records

	Status	Label	Task Progress	Completion Time	Request Time
☐	⟳	Task Id:5cbcf6bc-dfe3-11e1-bf56-1231380b8963	197 / 1026		08/06/2012 11:25 AM
☐	✓	Task Id:1fe43dd4-df1d-11e1-bf56-1231380b8963	4362 / 4362	08/06/2012 03:34 AM	08/05/2012 11:46 AM
☐	✓	Task Id:d3618606-df1c-11e1-bf56-1231380b8963	1 / 1	08/05/2012 11:45 AM	08/05/2012 11:44 AM
☐	!	Task Id:b3c3dba0-defe-11e1-bf56-1231380b8963	0 / 1 / 1	08/05/2012 11:42 AM	08/05/2012 08:08 AM
☐	✓	Task Id:a37697c4-defe-11e1-bf56-1231380b8963	624 / 624	08/05/2012 10:08 AM	08/05/2012 08:08 AM
☐	✓	Task Id:d370fd7c-deac-11e1-bf56-1231380b8963	3738 / 3738	08/05/2012 07:51 AM	08/04/2012 10:22 PM
☐	✓	Task Id:278af9f2-dea9-11e1-bf56-1231380b8963	624 / 624	08/05/2012 12:12 AM	08/04/2012 09:56 PM
☐	!	Task Id:6d8aa43c-dcfd-11e1-bf56-1231380b8963	299 / 717 / 1016	08/02/2012 07:00 PM	08/02/2012 06:54 PM
☐	!	Task Id:c150b78e-dcfb-11e1-bf56-1231380b8963	96 / 528 / 624	08/02/2012 06:49 PM	08/02/2012 06:42 PM
☐	✓	Task Id:e84f6718-dcf6-11e1-bf56-1231380b8963	11 / 11	08/02/2012 06:08 PM	08/02/2012 06:08 PM
☐	✓	Task Id:784d8008-dcf6-11e1-bf56-1231380b8963	11 / 11	08/02/2012 06:04 PM	08/02/2012 06:04 PM
☐	✓	Task Id:33f4206a-dcf6-11e1-bf56-1231380b8963	11 / 11	08/02/2012 06:03 PM	08/02/2012 06:02 PM
☐	!	Task Id:60950aae-dcf5-11e1-bf56-1231380b8963	9 / 2 / 11	08/02/2012 06:01 PM	08/02/2012 05:57 PM
☐	✓	Task Id:a4273bfc-dcf0-11e1-bf56-1231380b8963	11 / 11	08/02/2012 09:13 PM	08/02/2012 05:23 PM
☐	✓	Task Id:77993940-d2c4-11e1-bf56-1231380b8963	1339 / 1339	07/21/2012 06:18 PM	07/20/2012 06:41 PM
☐	✓	Task Id:bff65a14-d135-11e1-bf56-1231380b8963	1556 / 1556	07/19/2012 06:27 PM	07/18/2012 07:07 PM
☐	✓	Sync test	1078 / 1078	07/18/2012 10:17 AM	07/18/2012 10:16 AM
☐	✓	Task Id:be22047c-d0ea-11e1-bf56-1231380b8963	1077 / 1077	07/18/2012 10:13 AM	07/18/2012 10:11 AM
☐	✓	Task Id:d927feec-d0e6-11e1-bf56-1231380b8963	1 / 1	07/18/2012 10:14 AM	07/18/2012 09:44 AM
☐	✓	Task Id:19b4e54a-d0da-11e1-bf56-1231380b8963	1 / 1	07/18/2012 08:12 AM	07/18/2012 08:11 AM

Transfer Activity

Cancel

	Status	Label	Task Progress	Completion Time
☐		Task Id:5c		
☐	✓	Task Id:1f		
☐	✓	Task Id:d3		
☐	!	Task Id:b3		
☐	✓	Task Id:a3		
☐	✓	Task Id:d3		
☐	✓	Task Id:27		
☐	!	Task Id:6c		
☐	!	Task Id:c1		
☐	✓	Task Id:e8		
☐	✓	Task Id:78		
☐	✓	Task Id:33		
☐	!	Task Id:60		
☐	✓	Task Id:a4273bfc-dcf0-11e1-bf56-1231380b8963	11 / 11	08/02/2012 09:13 PM
☐	✓	Task Id:77993940-d2c4-11e1-bf56-1231380b8963	1339 / 1339	07/21/2012 06:18 PM
☐	✓	Task Id:bff65a14-d135-11e1-bf56-1231380b8963	1556 / 1556	07/19/2012 06:27 PM
☐	✓	Sync test	1078 / 1078	07/18/2012 10:17 AM
☐	✓	Task Id:be22047c-d0ea-11e1-bf56-1231380b8963	1077 / 1077	07/18/2012 10:13 AM

Transfer Details

Task ID: 5cbcf6bc-dfe3-11e1-bf56-1231380b8963

Status	ACTIVE	**Request Time**	08/06/2012 11:25 AM
Origin	isimipgo#us-archive	**Deadline**	08/07/2012 11:25 AM
Destination	xsede#trestles	**Completion Time**	
User	ian	**Bytes Transferred**	18583448810
Directories	10	**Files**	3731

Task Statistics

Pending	3342	**Succeeded**	399
Cancelled	0	**Expired**	0
Failed	0	**Retrying**	0

View Event Log

Globus Online Notification

To: Ian Foster

Task a37697c4-defe-11e1-bf56-1231380b8963: SUCCEEDED

=== Task Details ===
Task ID : a37697c4-defe-11e1-bf56-1231380b8963
Task Type : TRANSFER
Parent Task ID : n/a
Status : SUCCEEDED
Request Time : 2012-08-05 13:08:24Z
Deadline : 2012-08-08 14:10:32Z
Completion Time : 2012-08-05 15:08:31Z
Total Tasks : 624
Tasks Successful : 624
Tasks Expired : 0
Tasks Canceled : 0
Tasks Failed : 0
Tasks Pending : 0
Tasks Retrying : 0
Command : API 0.10 GO
Label : n/a
Data Encryption : No
Checksum Verification: No
Sync Level : 0
Files : 623
Files Skipped : 0
Directories : 1
Bytes Transferred : 440658434897
Bytes Checksummed : 0
MBits/sec : 489.213
Faults : 1

2 hours

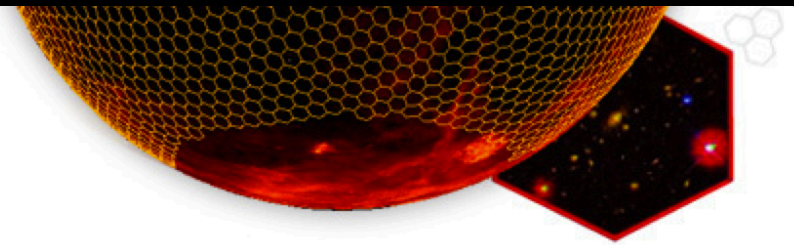
440 GB @ 60 MB/sec

Dark Energy Survey

- Receives 100,000 files each night in Illinois
- Transmit files to Texas for analysis … then move results back to Illinois
- Process must be reliable, routine, and efficient
- Use of Globus Transfer avoids need for custom software solution

Blanco 4m on Cerro Tololo

Image credit: Roger Smith/NOAO/AURA/NSF



Start Transfer | View Transfers | Manage Endpoints | Sign In | Sign Up

Sign In

Sign Up with Globus Online

Use Your GlobusOnline login alternate login

Username

Password

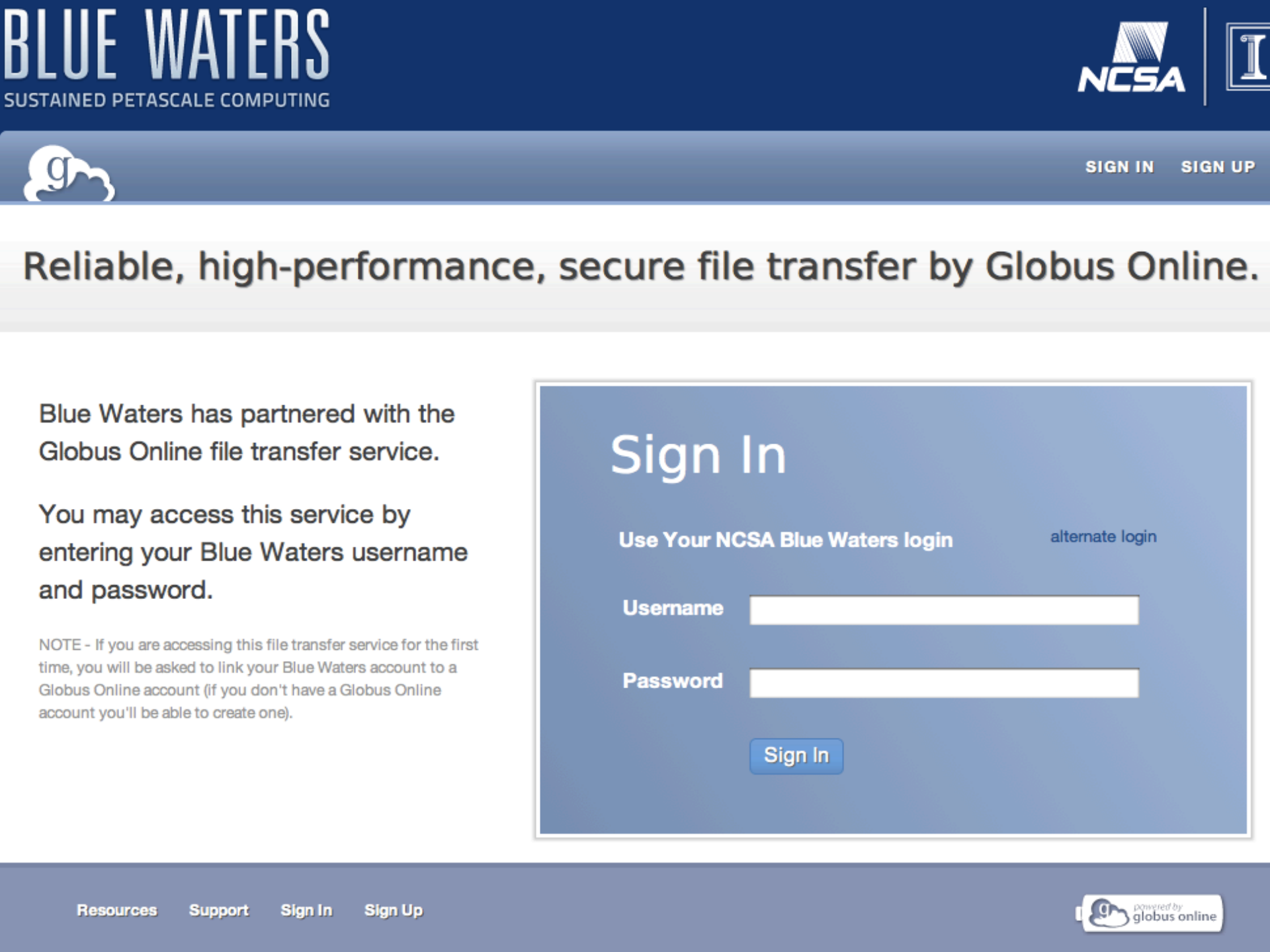
Sign In

Forgot Password?

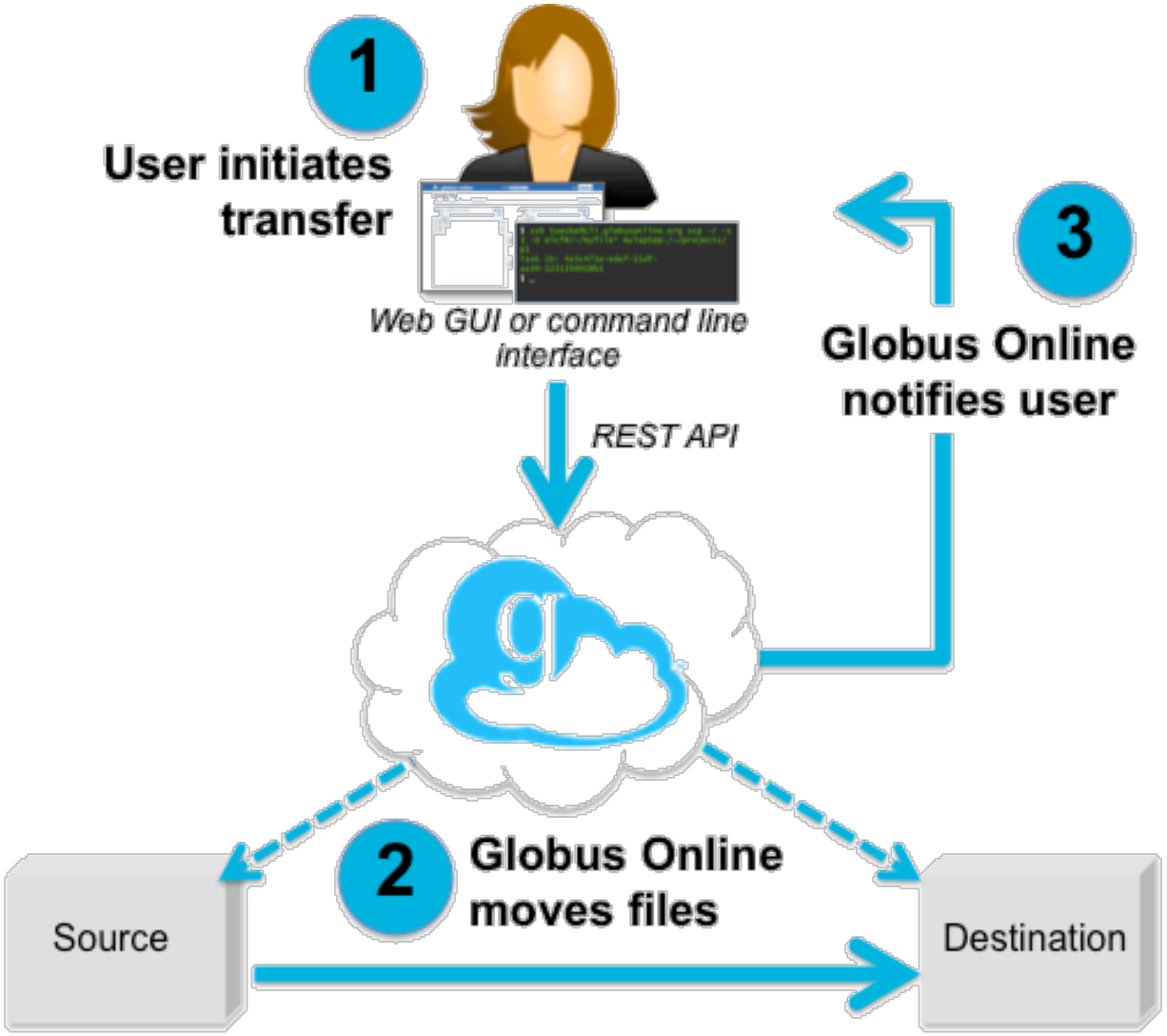




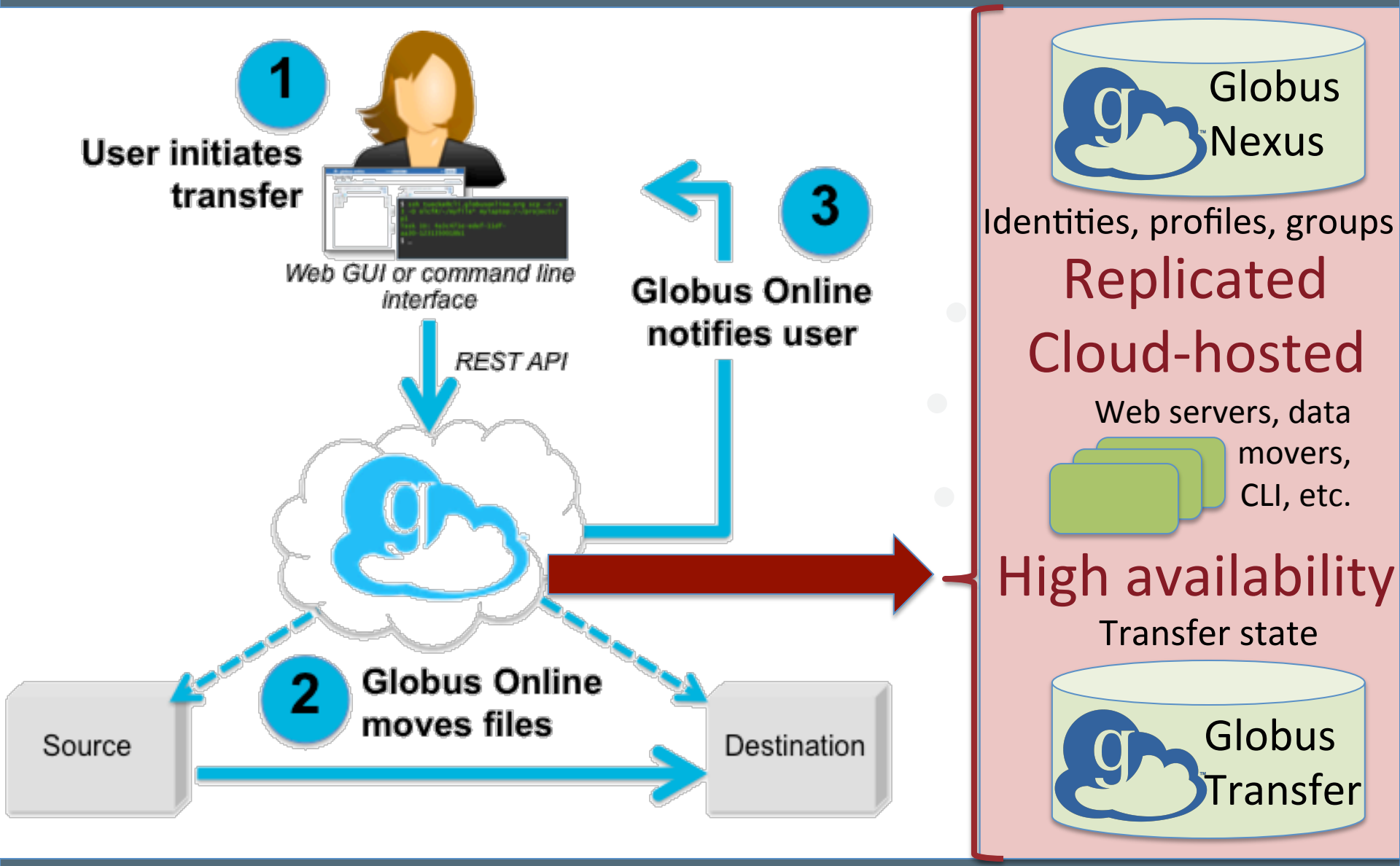
powered by globus online


BLUE WATERS
SUSTAINED PETASCALE COMPUTING
NCSA
SIGN IN
SIGN UP
Reliable, high-performance, secure file transfer by Globus Online.
Blue Waters has partnered with the Globus Online file transfer service.
You may access this service by entering your Blue Waters username and password.
NOTE - If you are accessing this file transfer service for the first time, you will be asked to link your Blue Waters account to a Globus Online account (if you don't have a Globus Online account you'll be able to create one).
Sign In
Use Your NCSA Blue Waters login
alternate login
Username
Password
Sign In
Resources
Support
Sign In
Sign Up
powered by globus online

Globus Transfer under the covers

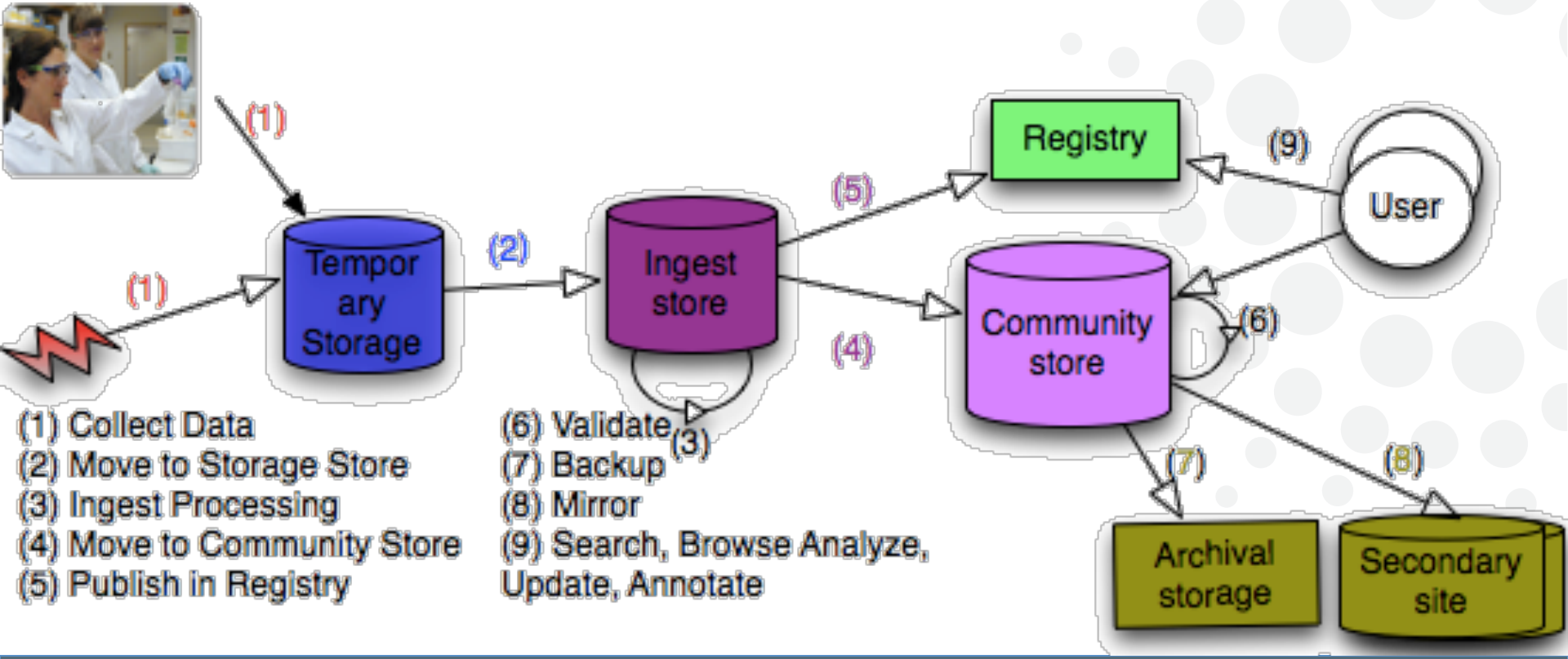


Globus Transfer under the covers

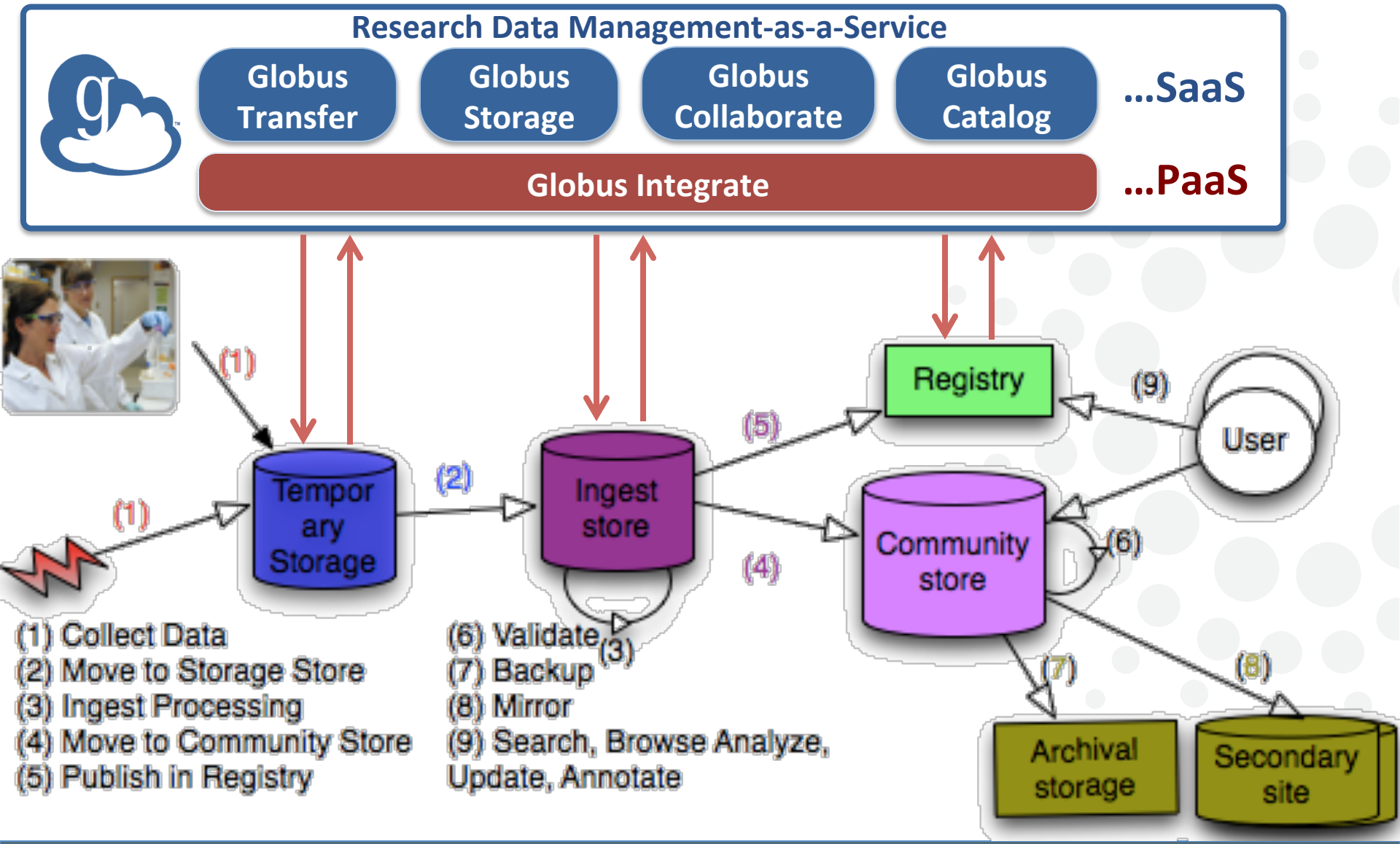


Moving beyond data movement

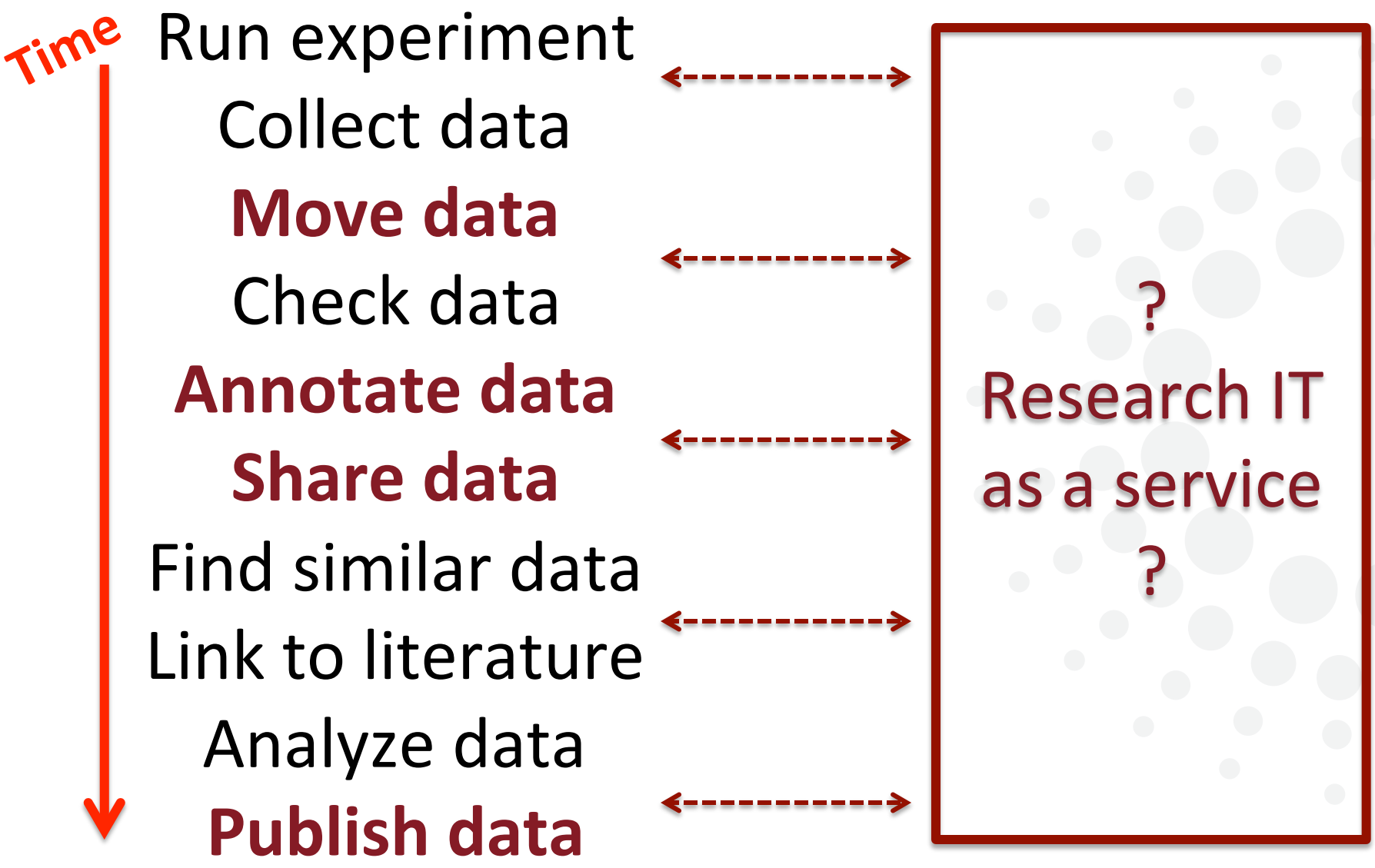
Dark Energy Survey Metagenomics Climate science
Genomics Land use change X-ray source data
Biomedical imaging High energy physics Nielsen data



Moving beyond data movement

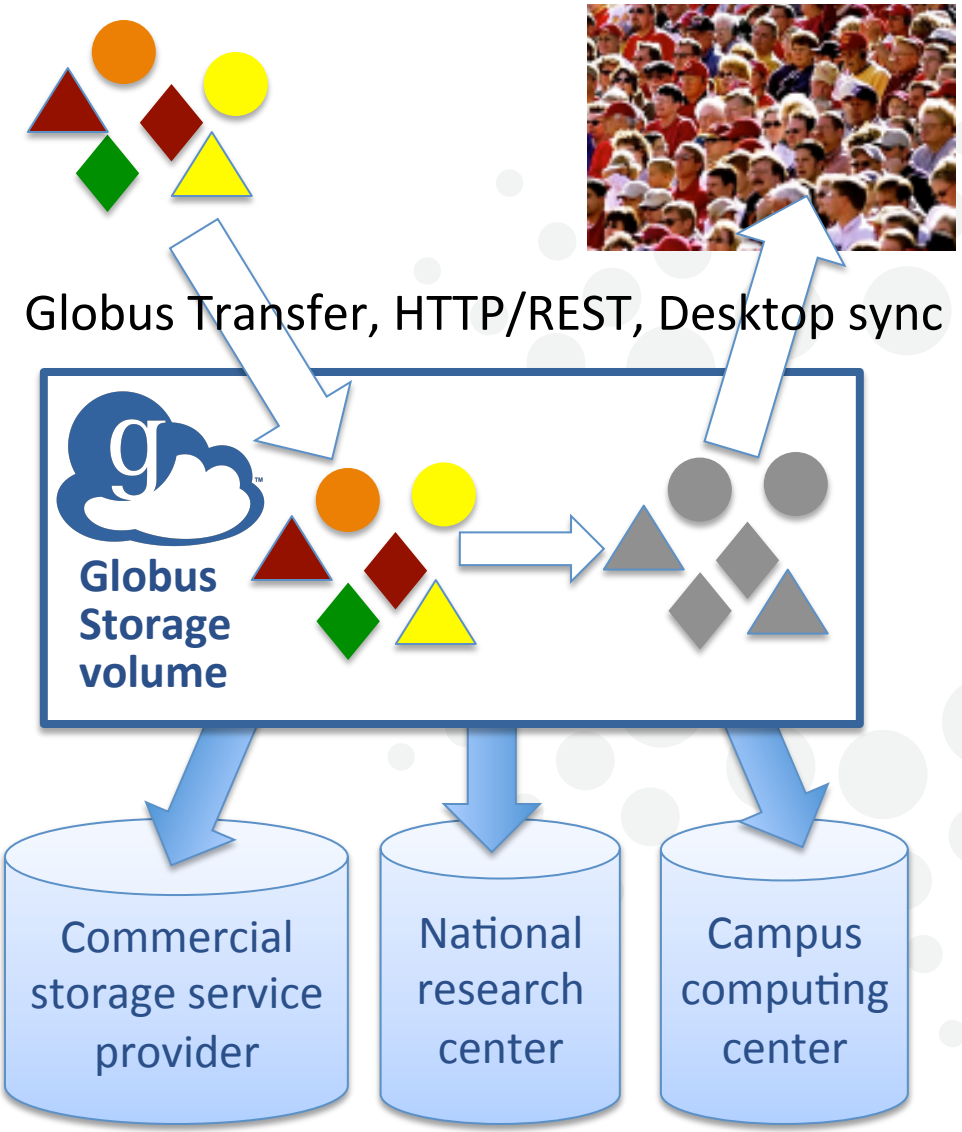


Process automation for science

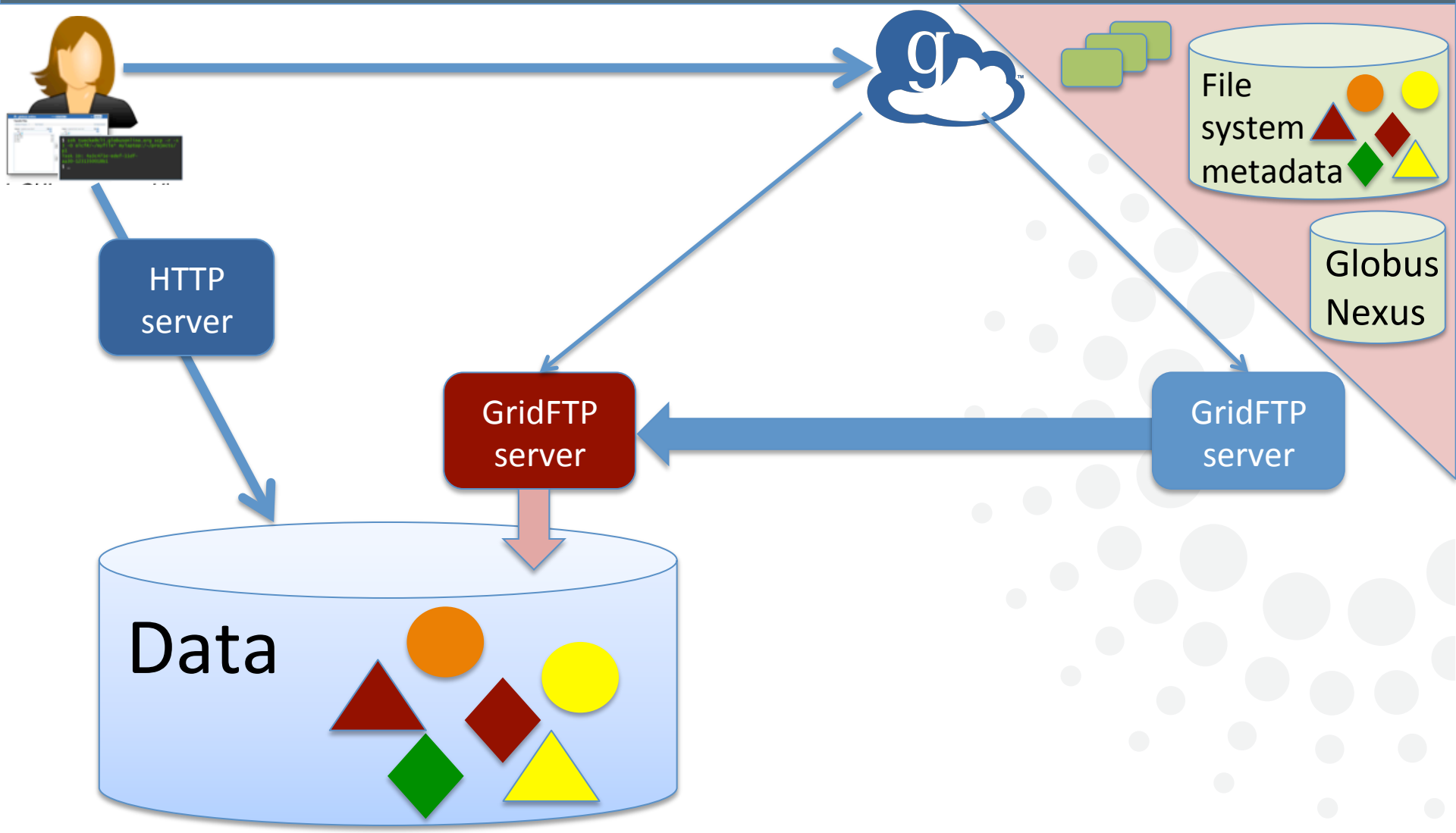


Globus Storage: For when you want to …

- **Place** your data where you want
- **Access** it from anywhere via different protocols
- Update it, version it, and take snapshots
- **Share** versions with who you want
- **Synchronize** among locations



Globus Storage under the covers



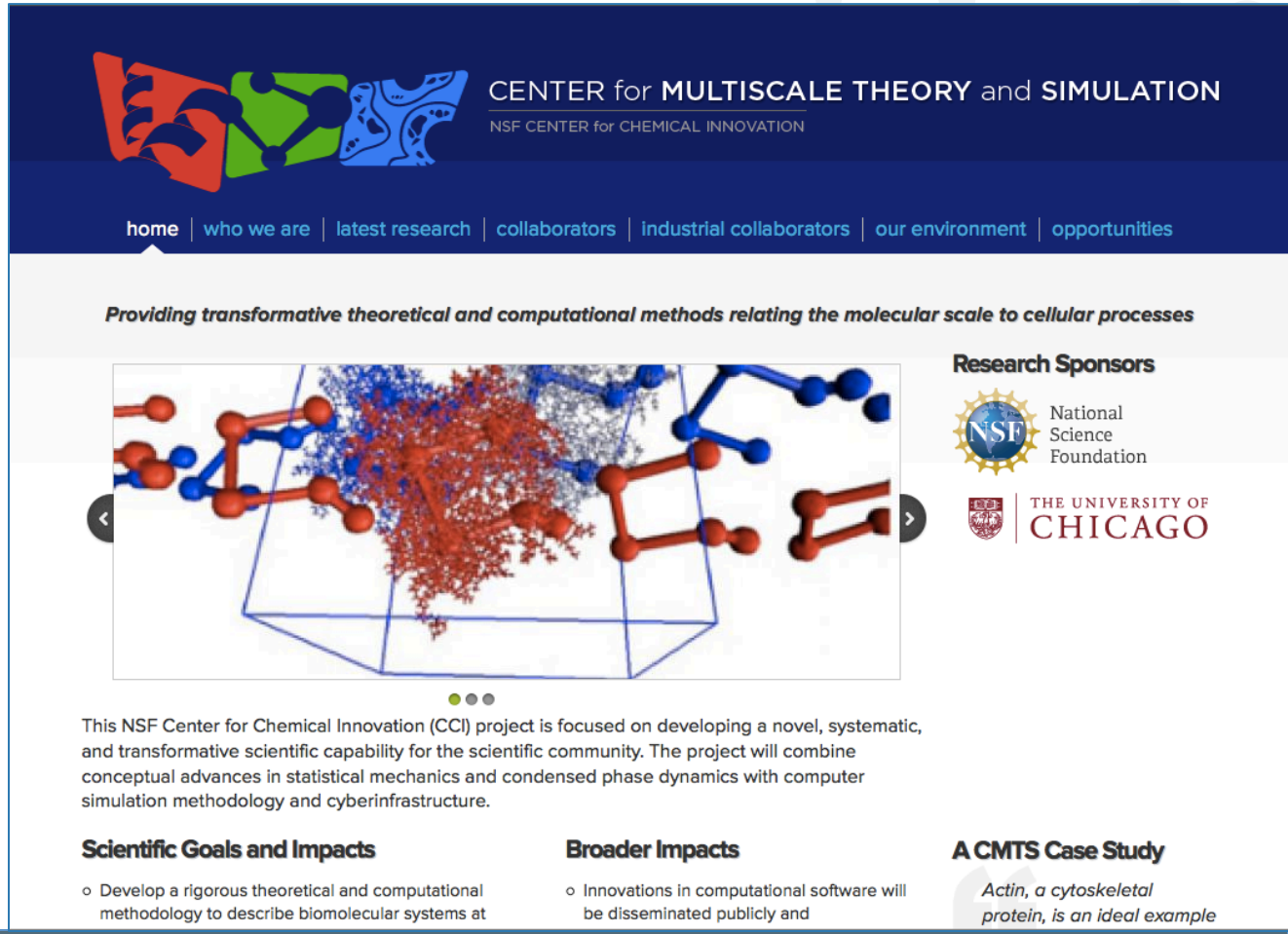
Globus Collaborate: For when you want to

Join with a few or many people to:

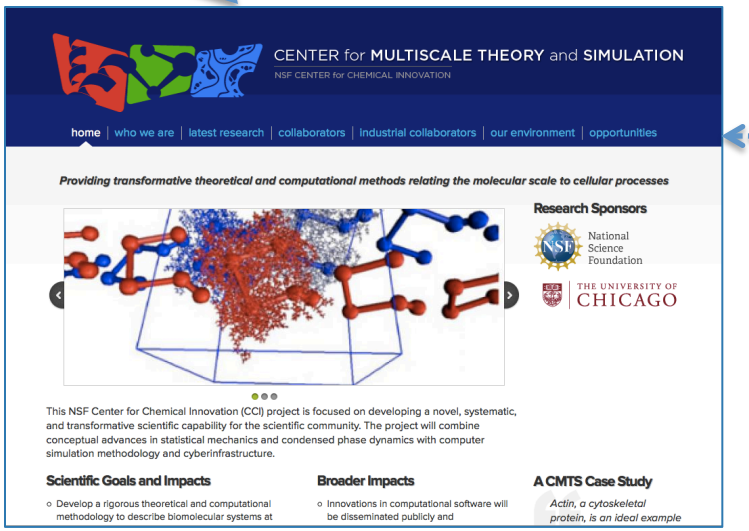
- Share documents
- Track tasks
- Send email
- Share data
- Do whatever

With:

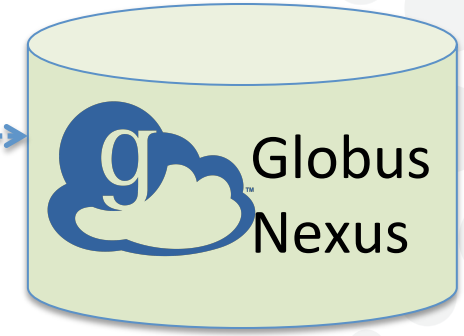
- Common groups
- Delegated management



Globus Collaborate under the covers



Collaborative tool



Identity, profile, group database hosted on Amazon

Globus Storage & Collaborate in action

Bryce

Globus Connect
Move DTI results to Bryce's laptop

PADS Compute Cluster

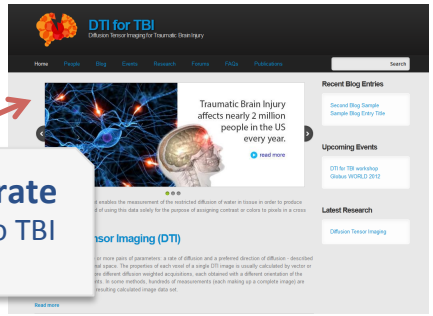
DTI Group
- Kyle
- **Bryce**

Globus Storage
Create snapshot to share with group

Globus Transfer
Copy TBI data to compute cluster

Globus Nexus
Add Bryce to TBI collaboration

Globus Transfer
Move DTI results to shared volume



Globus Collaborate
Publish DTI data to TBI web site

Kyle

Globus Storage
Create volume and share with TBI group

"TBI" volume

UChicago Object Store

Amazon S3

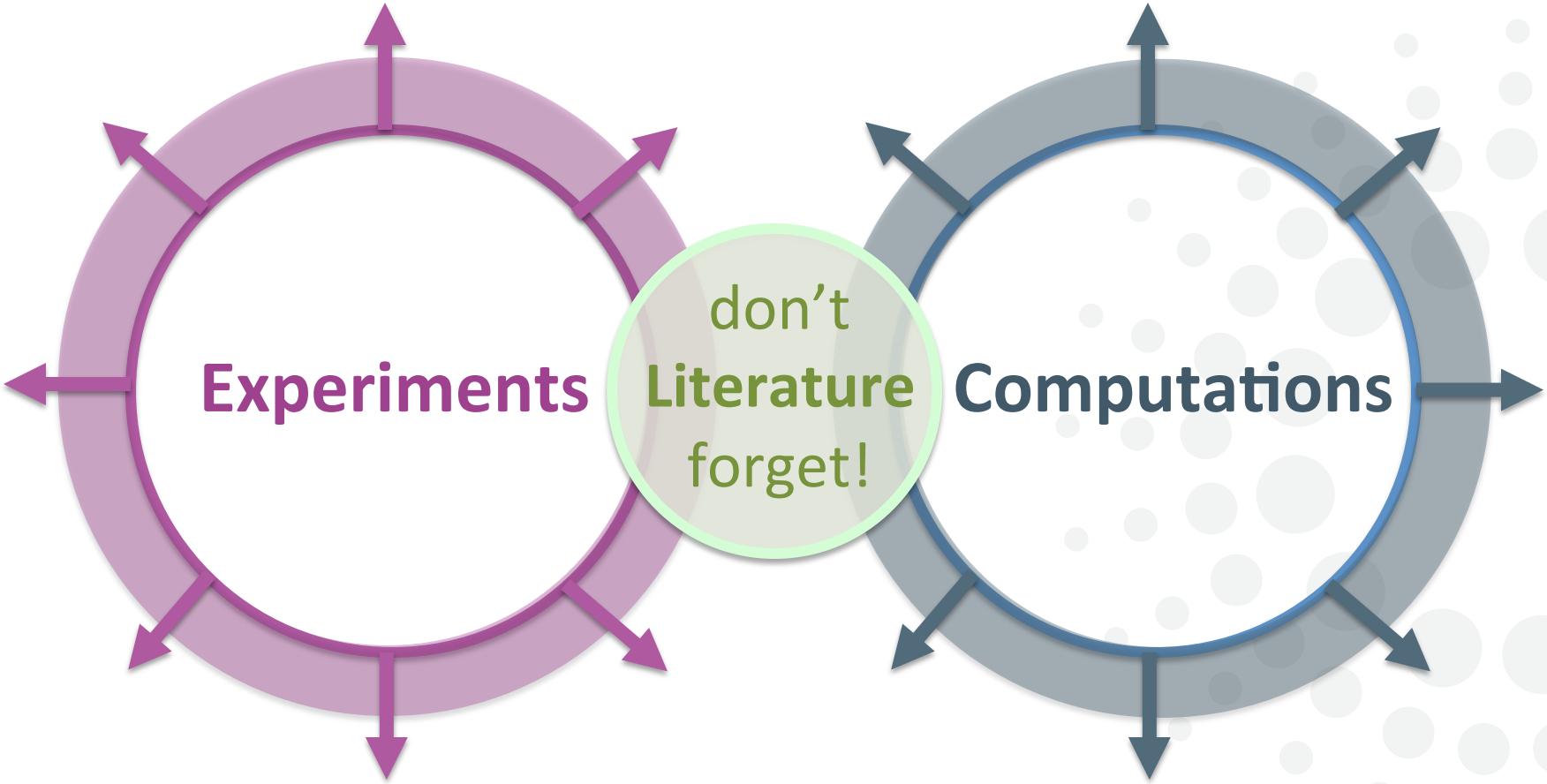
SDSC Cloud

Cornell Red Cloud

Globus Connect
Move MRI files to TBI shared volume

TBI=Traumatic Brain Injury
DTI=Diffusion Tensor Imaging
MRI=Magnetic Resonance Imaging

Data acquisition, management, analysis



Big Data (volume, velocity, variety, variability)
… demands **Big Process** in order for discovery to scale

Let's rethink how we provide **research IT**

Accelerate discovery and innovation worldwide by providing **research IT as a service**

Leverage the cloud to

- provide millions of researchers with unprecedented access to powerful tools;
- enable a massive shortening of cycle times in time-consuming research processes; and
- reduce research IT costs dramatically via economies of scale

MG-RAST

metagenomics analysis server

LOG IN REGISTER PASSWO

Browse Metagenomes

search for metagenomes

Register Contact Help Upload * News

About

MG-RAST (the Metagenomics RAST) server is an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data.

# of metagenomes	40,895
# base pairs	11.22 Tbp
# of sequences	104.3 billion
# of public metagenomes	7,391

The server provides web based upload, quality control, automated annotation and analysis for samples up to 10GBp. Comparison between large numbers of samples is enabled via pre-computed abundance profiles.

* login required

This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900040C.

NANOHUB

FUNDAMENTALS OF
NANOELECTRONICS
Online Course Spring 2012

Learn More ›

1 2 3

SIMULATE with over 160 tools for nanoelectronics, nanophotonics and more ›

RESEARCH & COLLABORATE via groups, question board and more ›

TEACH & LEARN with tool-powered curricula, courses, seminars and more ›

SHARE & PUBLISH tools and research through our easy upload process

Leap. Drag here to tag and file.

RESOURCES

Search

Popular Tags: nanoelectronics | course lecture | material science | Illinois | nano/bio | nanotransistors | research seminar | devices | nanophotonics | quantum transport | tutorial | transistors | molecular electronics | nano electro-mechanical systems | NEGF | carbon nanotubes | nanomedicine | education/outreach | UIUC | band structure | ABACUS | atomic force microscopy | quantum dots | MOSFET

FEATURED

MIT Atomic Scale Modeling Toolkit : Tools for Atomic Scale Modeling - in Tools

Nanotechnology: Considerations for Facility Design - in Online Presentations

Greg Lush, University of Texas at El Paso - Contributions: 30

Topics For Introductory Materials Classes - in Topics

SQLShare – View Query

https://sqlshare.cac.washington.edu/sqlshare#s=query/charlon/United%20States%20Wind%20Energy%20Potential.csv

SQLSHARE

United States Wind E... x

Logged in: charlon@washington.edu

- Your datasets
- All datasets
- Favorites
- Recently viewed »
- Shared with you...

- Upload dataset
- New query
- Recent activity...

United States Wind Energy Potential Test.csv

Last modified: Apr 19, 2011 3:15 PM chartest

Click here to add a description

```
SELECT * FROM [chartest].[table_United States Wind Energy Potential.csv]
```

Edit query | Download | Query dataset | More actions

DATASET PREVIEW (Rows **1 - 48** of **48**)

<< first < prev **1** 2 3 next > last >>

State	Total (km)	Excluded (km)	Available (km)	Available % of State	% of Total Windy Land Excluded	Potential Installed Ca
Alabama	80.36	56.72	23.64	0.017686450652558%	70.5823792931807%	118.2
Arizona	4544.96	2364.14	2180.82	0.73866427637171%	52.016739421249%	10904.1
Arkansas	4663.24	2823.18	1840.06	1.33593638826854%	60.5411688010911%	9200.3
California	26901.28	20079.24	6822.04	1.66671846119056%	74.6404632047248%	34110.2
Colorado	95830.36	18386.46	77443.9	28.7253372905834%	19.1864665853285%	387219.5
Connecticut	31.36	26.06	5.3	0.041365143388757%	83.0994897959184%	26.5
Delaware	36.56	34.66	1.9	0.03739686338729%	94.8030634573304%	9.5
Florida	9.56	9.48	0.08	0.000054815635831%	99.163179916318%	0.4
Georgia	281.28	255.26	26.02	0.01708279443529%	90.7494311717861%	130.1
Idaho	13420.4	9805.28	3615.12	1.67025347276445%	73.0625018628357%	18075.6
Illinois	70763.56	20787.14	49976.42	34.2484872158604%	29.3754864791992%	249882.1
Indiana	46255.24	16600.74	29645.5	21.6238926072802%	35.9088207406417%	148227.5

Acknowledgements

- Thanks for vital and much appreciated support:
  - NSF Office of Cyberinfrastructure (OCI)
  - DOE Office of Advanced Scientific Computing Research (ASCR)
  - National Institutes of Health
  - The University of Chicago
- Thanks to the amazing Globus Online team at the University of Chicago and Argonne. See www.globusonline.org/about/goteam/

Computation Institute

Thank you!

globusonline.org

@globusonline

foster@anl.gov
foster@uchicago.edu