Computation Institute

# Big process for big data

## Process automation for data-driven science

Ian Foster

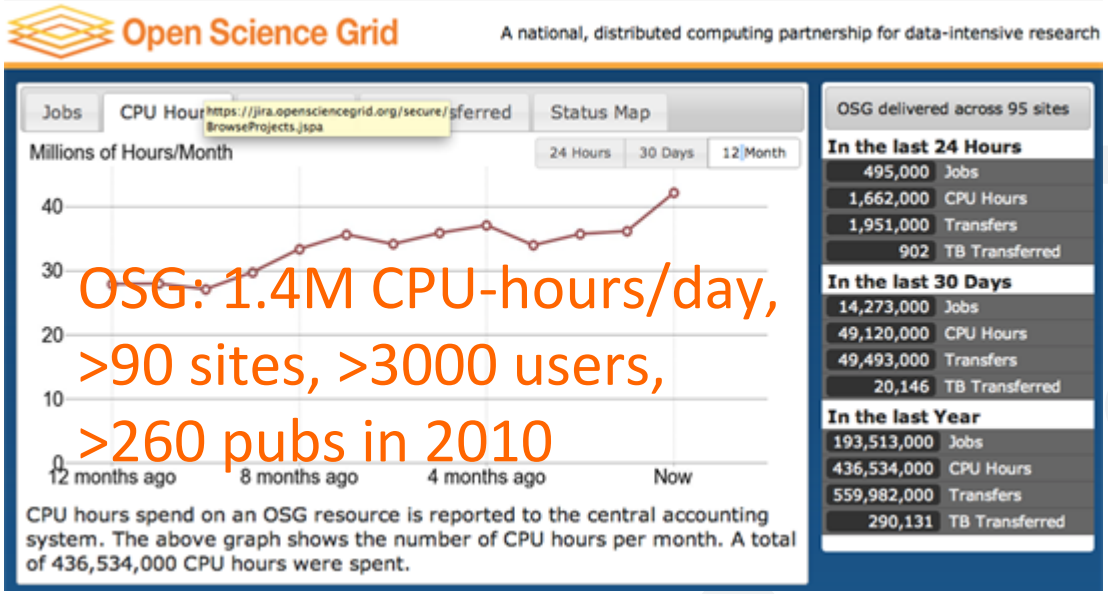
Computation Institute

Argonne National Laboratory & The University of Chicago

Talk at HPC 2012 Conference, Cetraro, Italy, June 25, 2012

# Big science is making it work


LIGO: 1 PB data in last science run, distributed worldwide




OSG: 1.4M CPU-hours/day, >90 sites, >3000 users, >260 pubs in 2010

Robust production solutions
Substantial teams and expense
Sustained, multi-year effort
Application-specific solutions, built on common technology


ESG: 1.2 PB climate data delivered to 23,000 users; 600+ pubs

All build on NSF- & DOE-supported Globus Toolkit software

# But small/medium science is struggling

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

# Arlington, VA, June 2012

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

Edit Trip

**Trip Cost: $1,736.10** [?]

+ Add plans  Export to calendar  More ▸

| | | |
|---|---|---|
| **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 | |

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

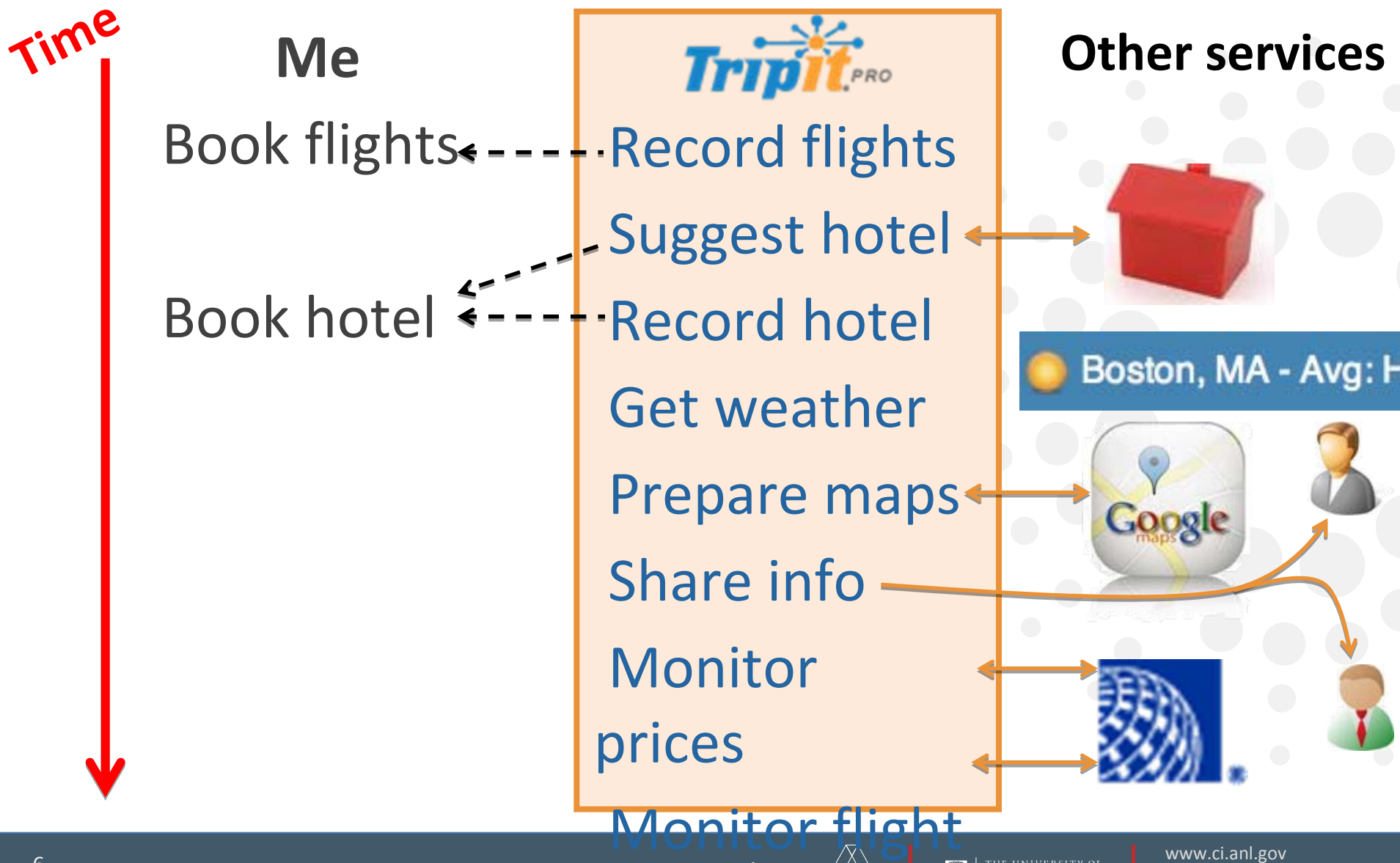
**Booking Information**
Booked on United 5/25/2012
http://www.united.com/

## 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 »

# Tripit exemplifies process automation



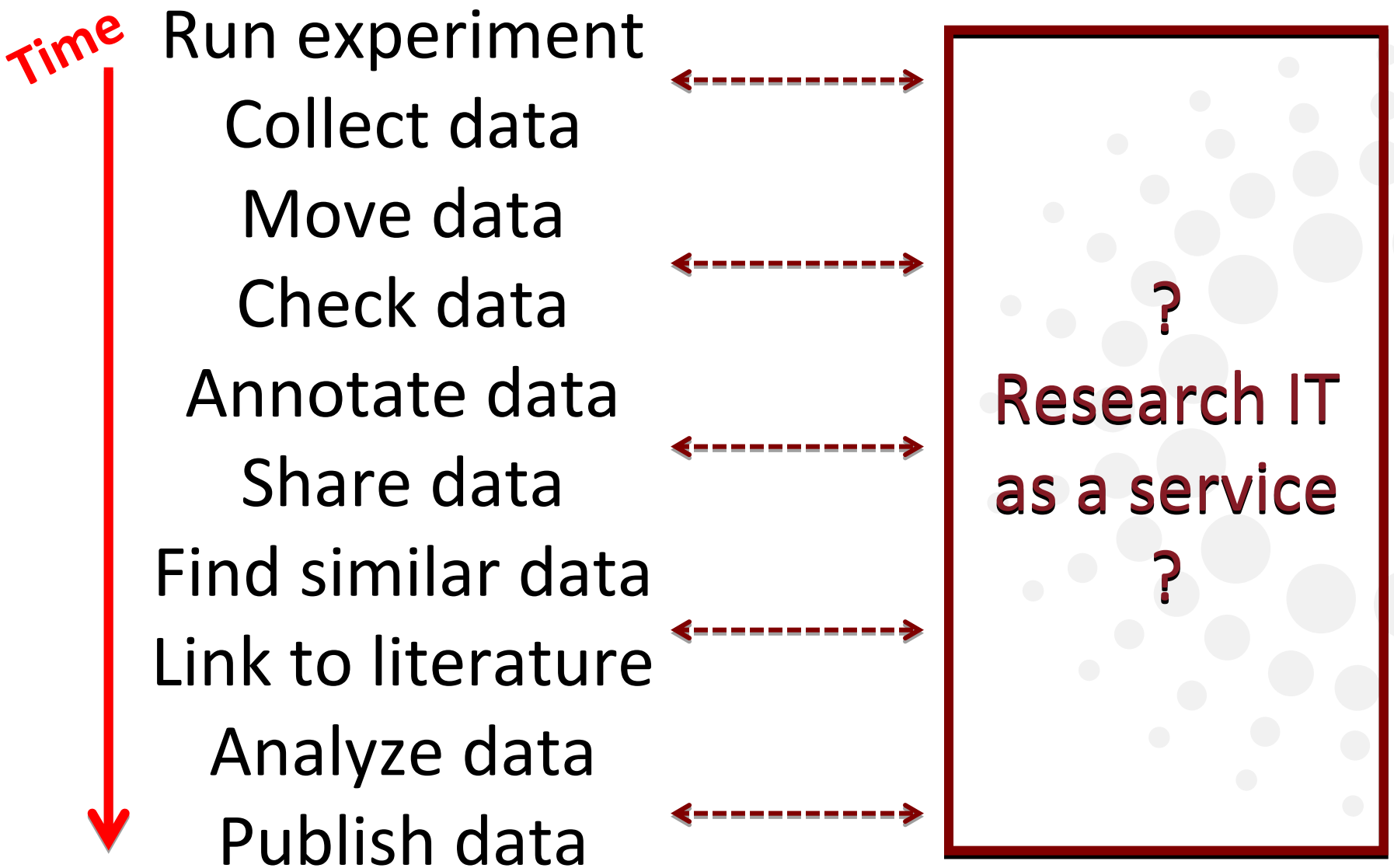
# Complexity is large and growing

Time

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

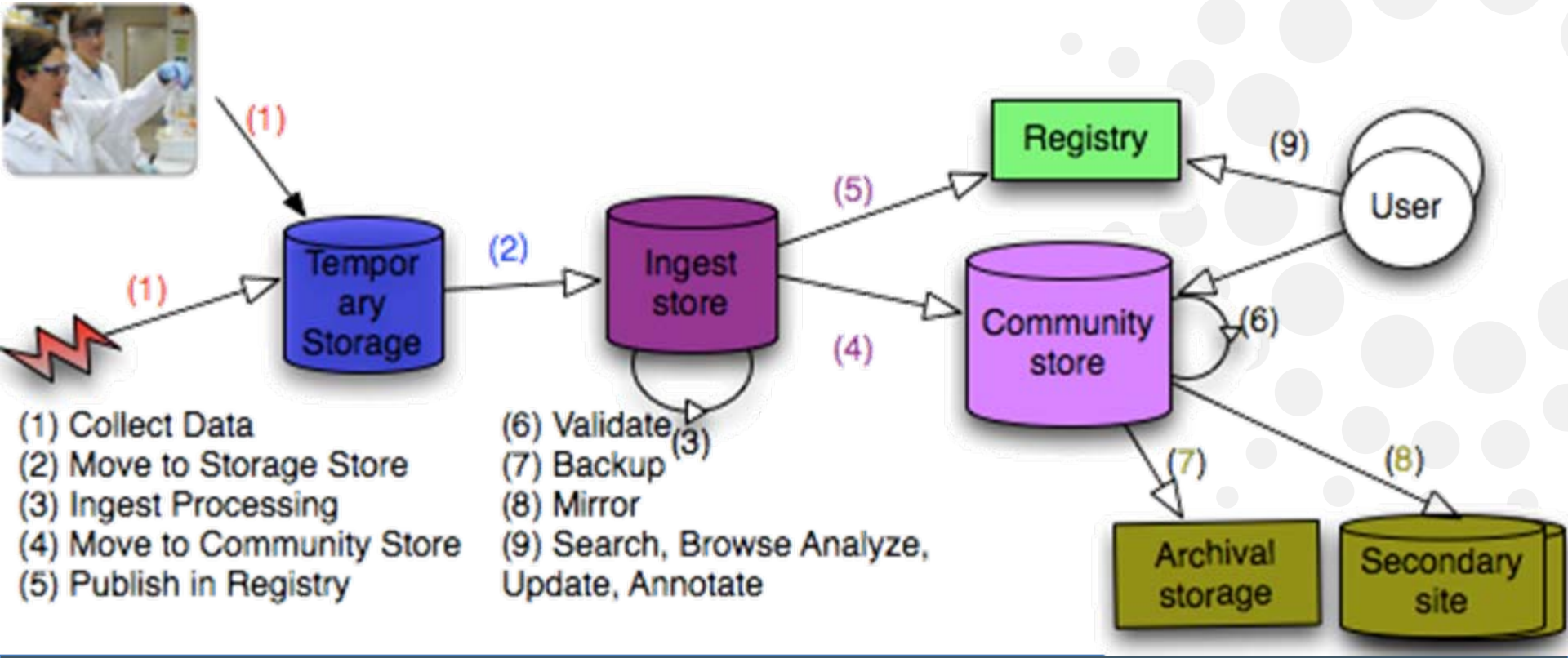
# Can we extract this complexity?

# Process automation for science



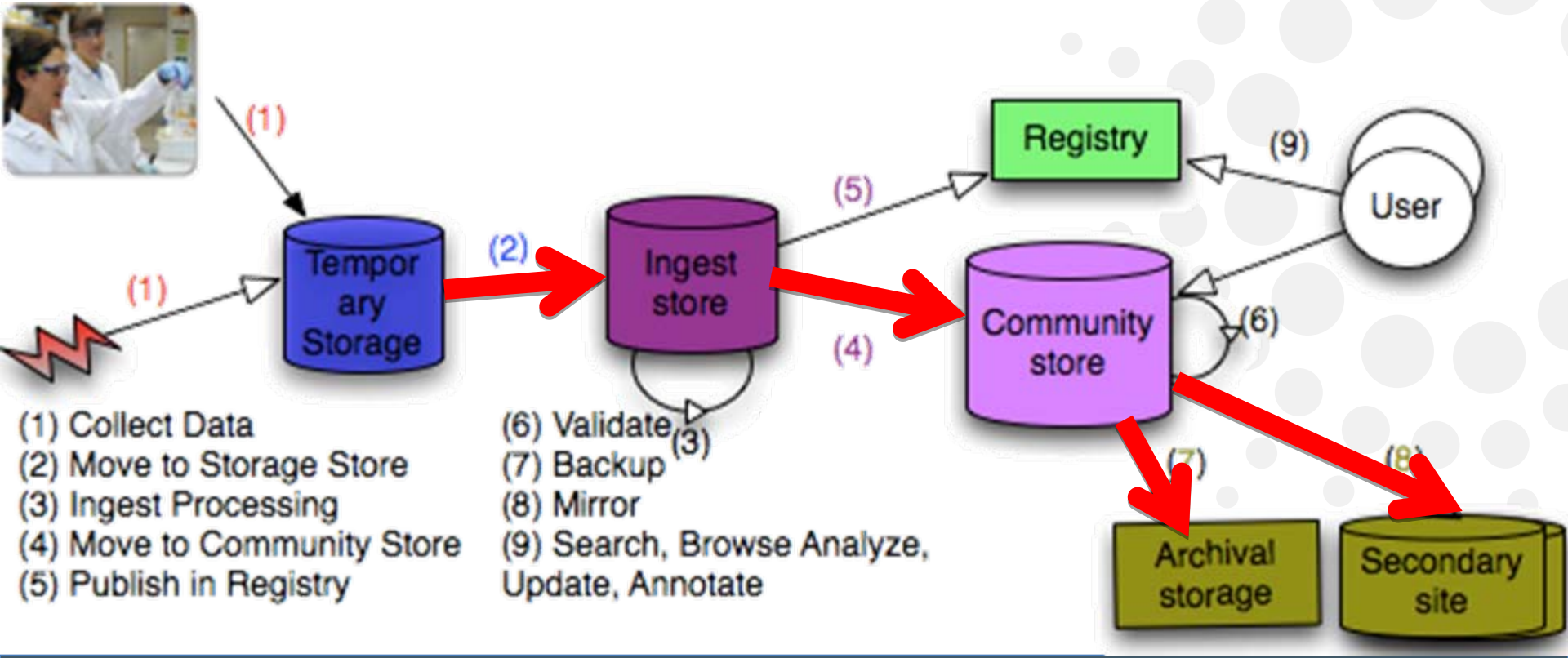
# A first take on “big process for science”

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



# A first take on “big process for science”

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






Argonne NATIONAL LABORATORY | THE UNIVERSITY OF CHICAGO

www.ci.anl.gov
www.ci.uchicago.edu

# Software as a Service (Gartner)

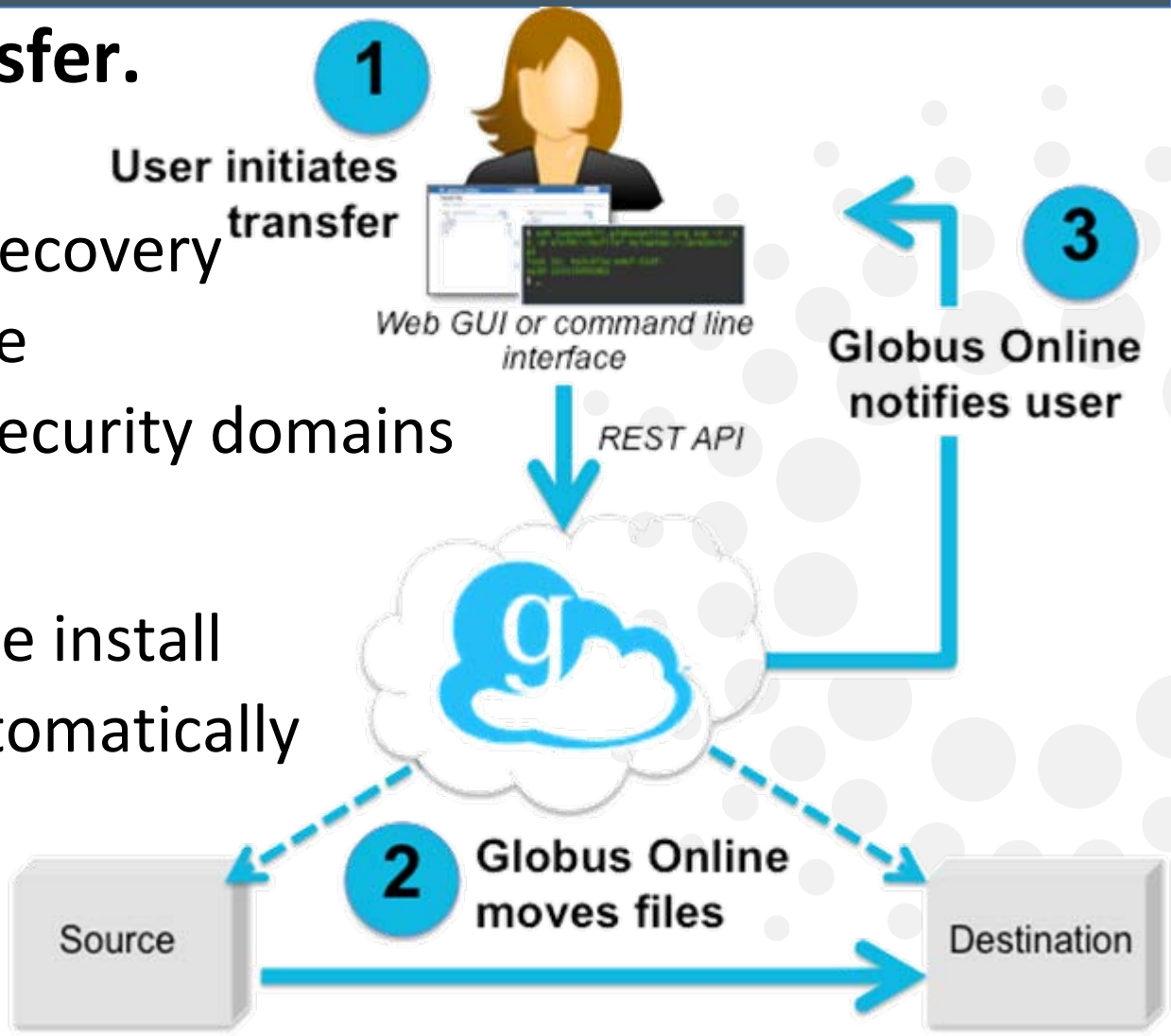
1. The application is owned, delivered, and managed remotely by one or more providers
2. The application is based on a single code base that is consumed in a one-to-many model by all contracted customers at any time
3. The application is licensed on pay-per-use or subscription basis

---

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

# Globus Online: Data transfer as SaaS

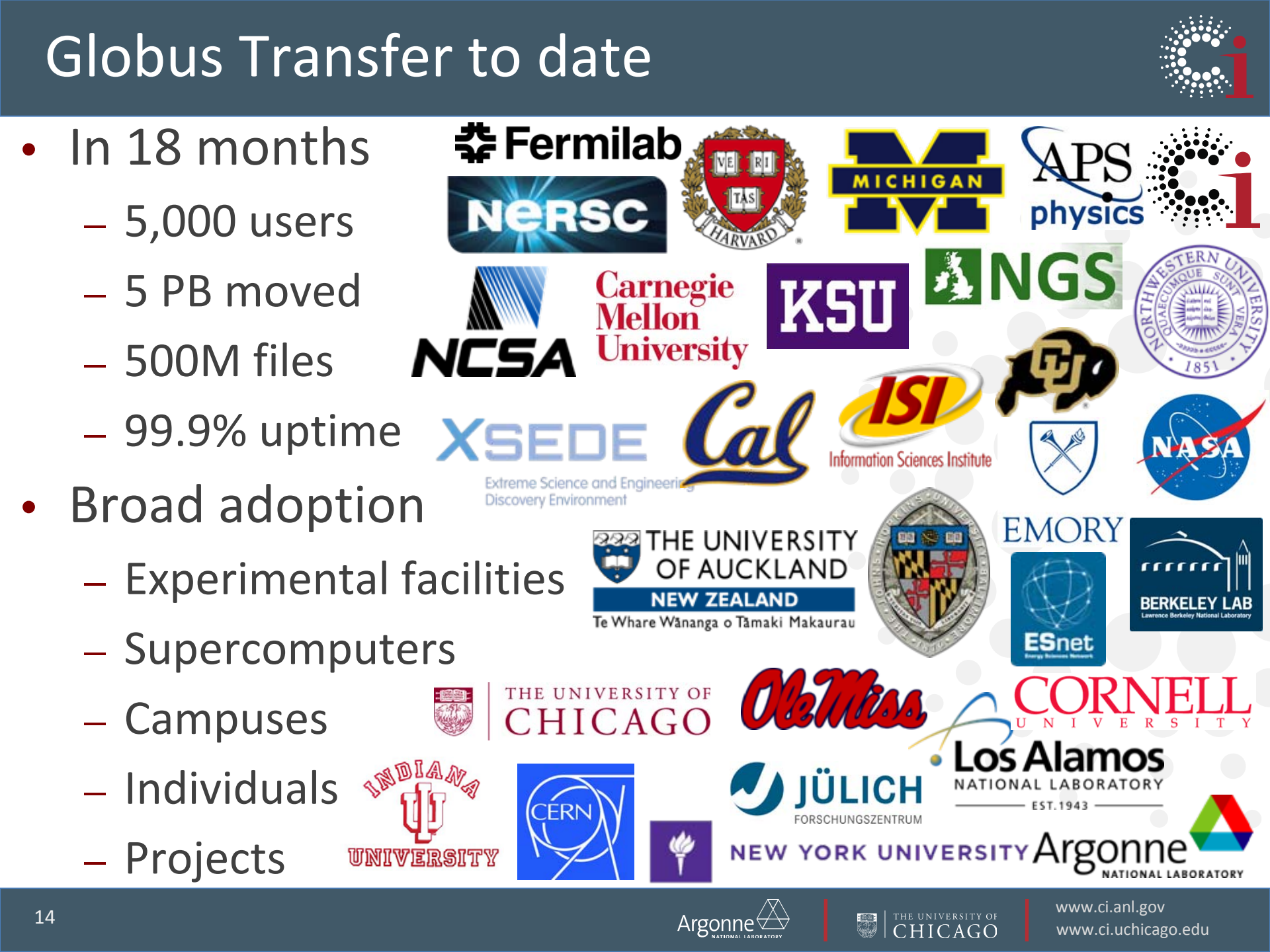
- **Reliable file transfer.**
  - Fire-and-forget
  - Automatic fault recovery
  - High performance
  - Across multiple security domains
- **No IT required.**
  - No client software install
  - New features automatically available
  - Consolidated support and troubleshooting

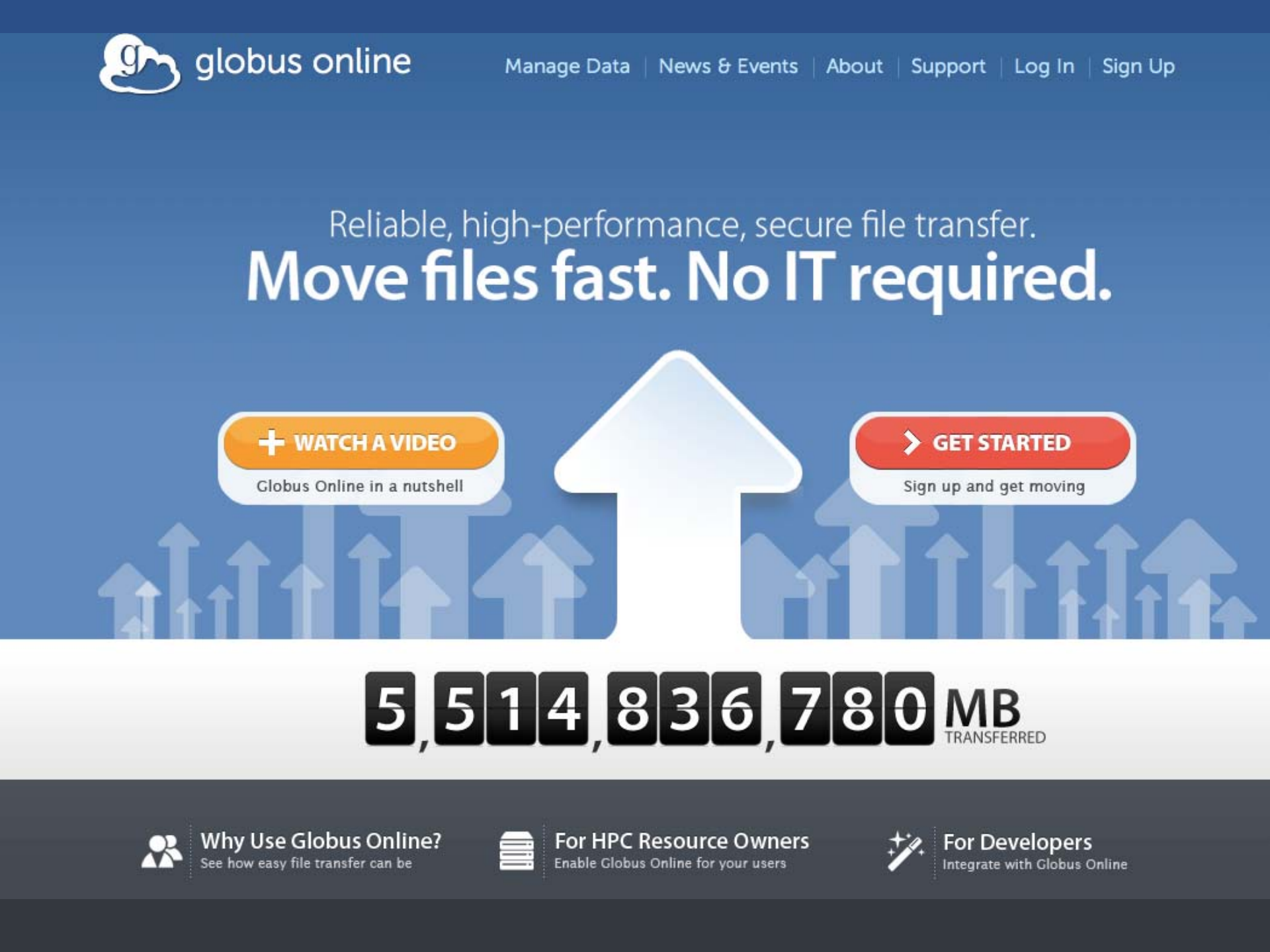


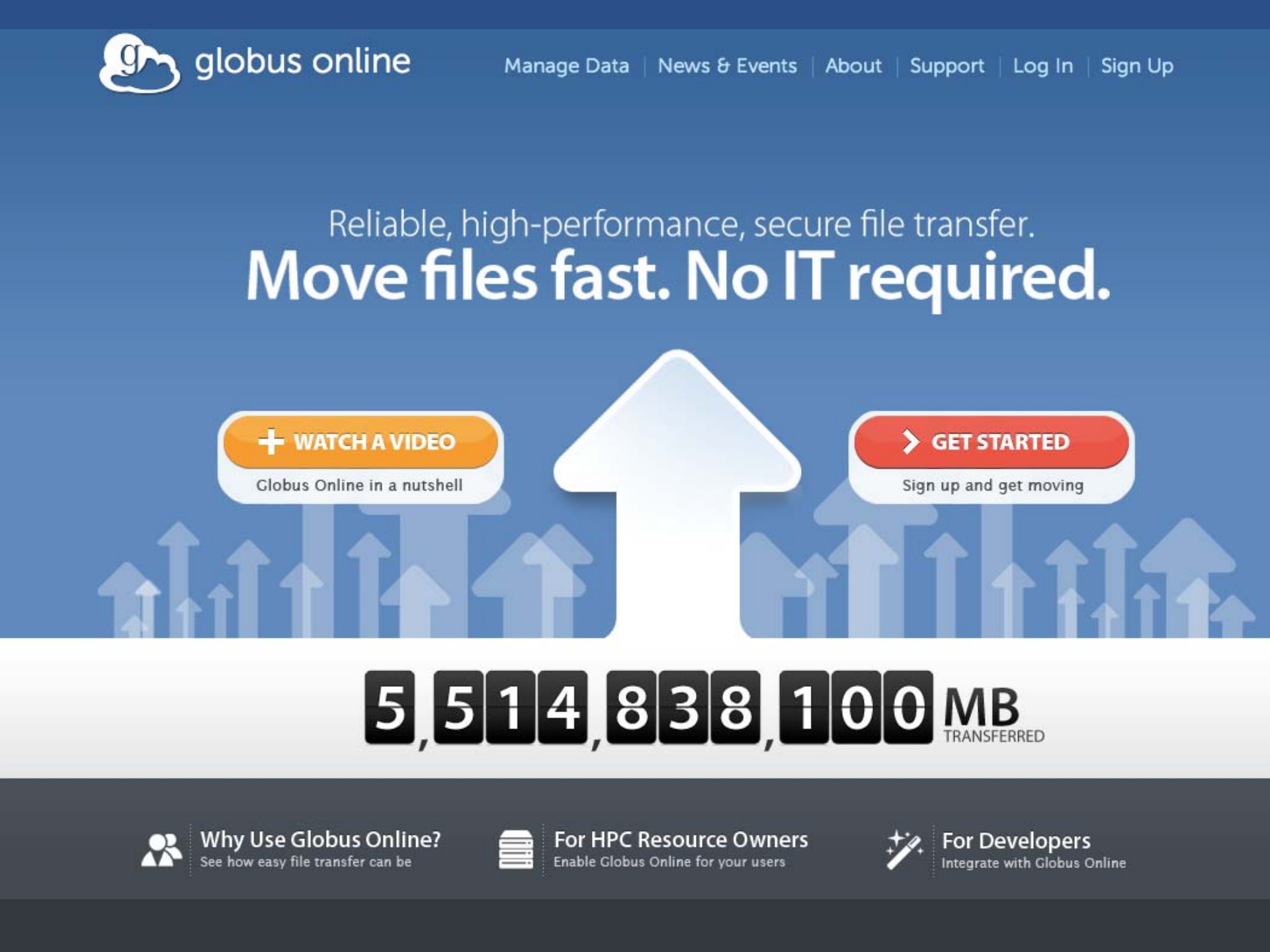

Works with existing GridFTP servers; also Globus Connect

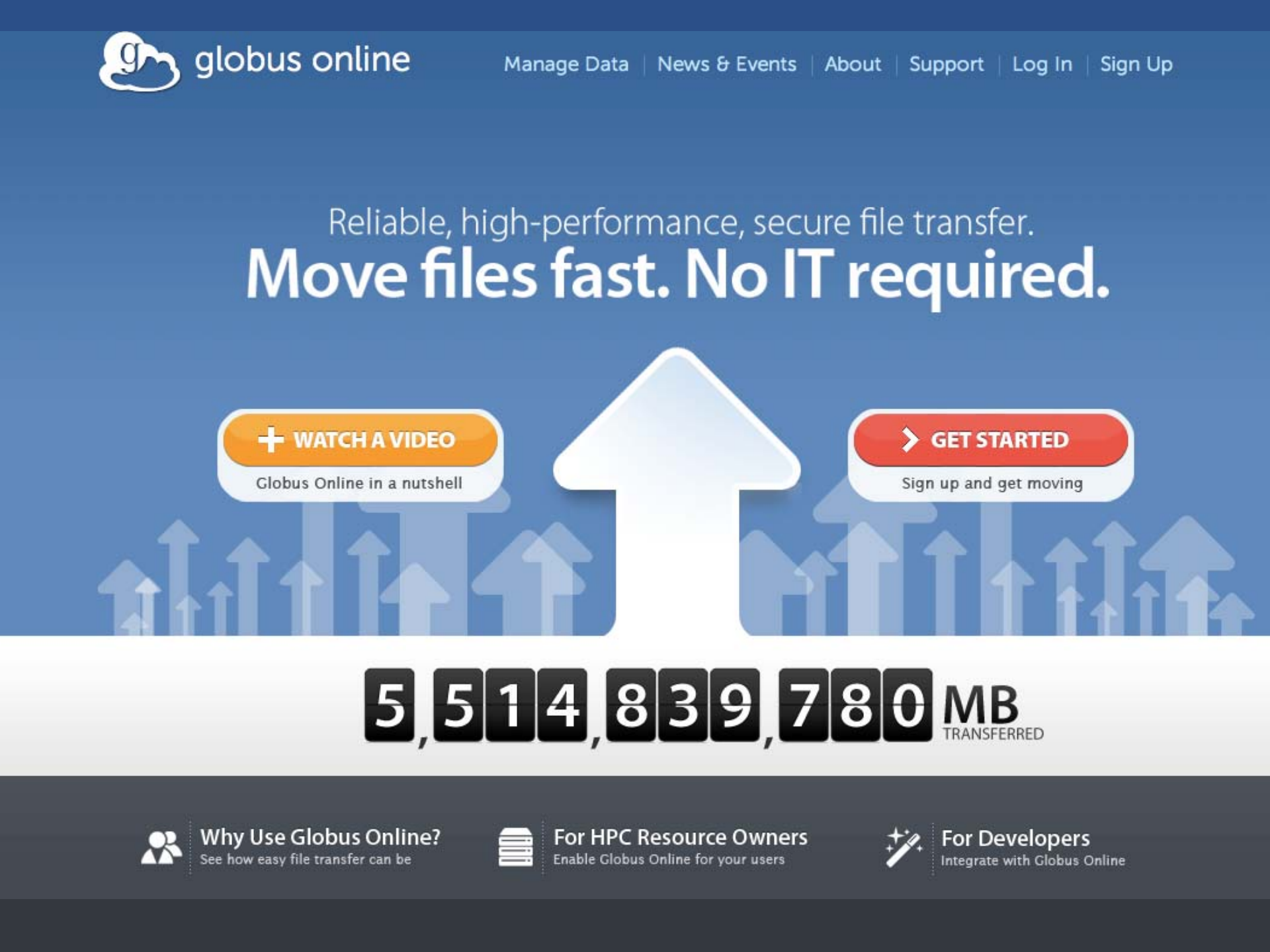
# Globus Transfer to date

- In 18 months
  - 5,000 users
  - 5 PB moved
  - 500M files
  - 99.9% uptime
- Broad adoption
  - Experimental facilities
  - Supercomputers
  - Campuses
  - Individuals
  - Projects




globus online
Manage Data
News & Events
About
Support
Log In
Sign Up
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
5,514,836,780 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 online
Manage Data
News & Events
About
Support
Log In
Sign Up
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
5,514,838,100 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

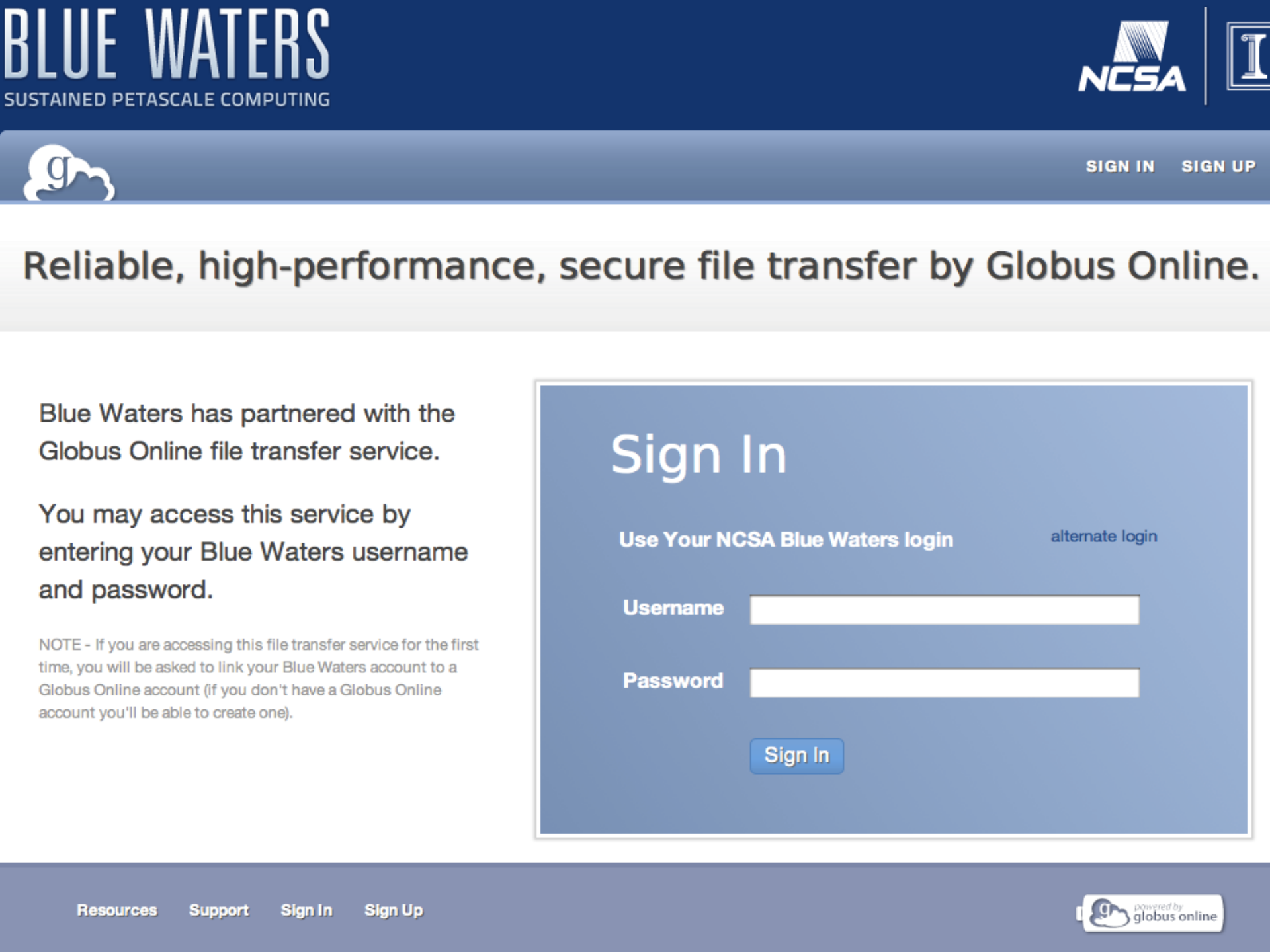

globus online
Manage Data
News & Events
About
Support
Log In
Sign Up
Reliable, high-performance, secure file transfer.
Move files fast. No IT required.
WATCH A VIDEO
Globus Online in a nutshell
GET STARTED
Sign up and get moving
5,514,839,780 MB
TRANSFERRED
Why Use Globus Online?
See how easy file transfer can be
For HPC Resource Owners
Enable Globus Online for your users
For Developers
Integrate with Globus Online

# Dark Energy Survey use of Globus Online

- Dark Energy Survey receives 100,000 files each night in Illinois
- They transmit files to Texas for analysis … then move results back to Illinois
- Process must be reliable, routine, and efficient
- They outsource this task to Globus Online

Blanco 4m on Cerro Tololo

Image credit: Roger Smith/NOAO/AURA/NSF


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

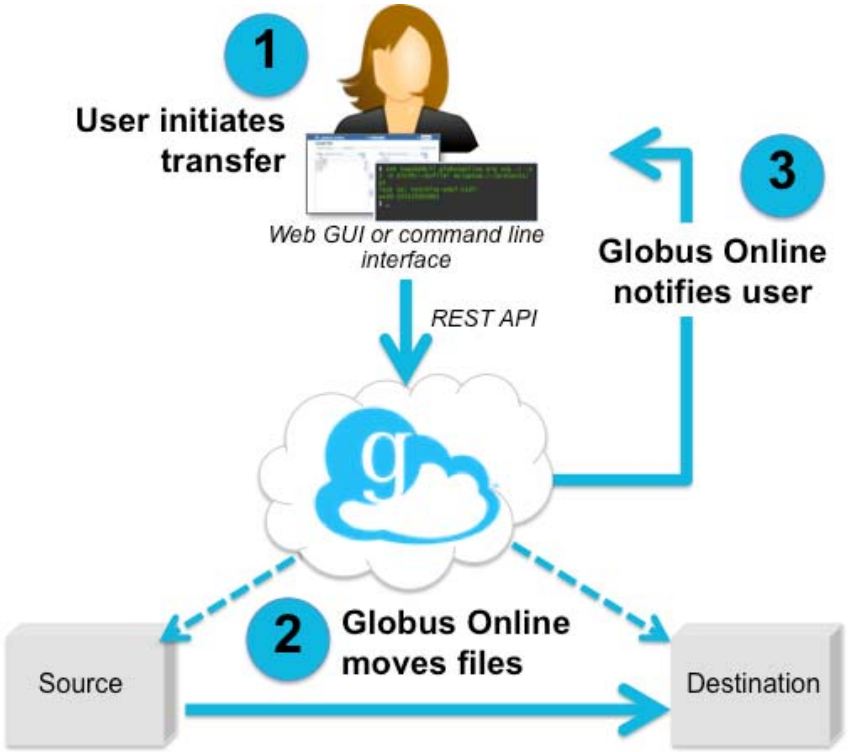
# Genome sequence analysis pipelines

Amazon EC2 computing

Amazon S3 storage

Commercial sequencing center

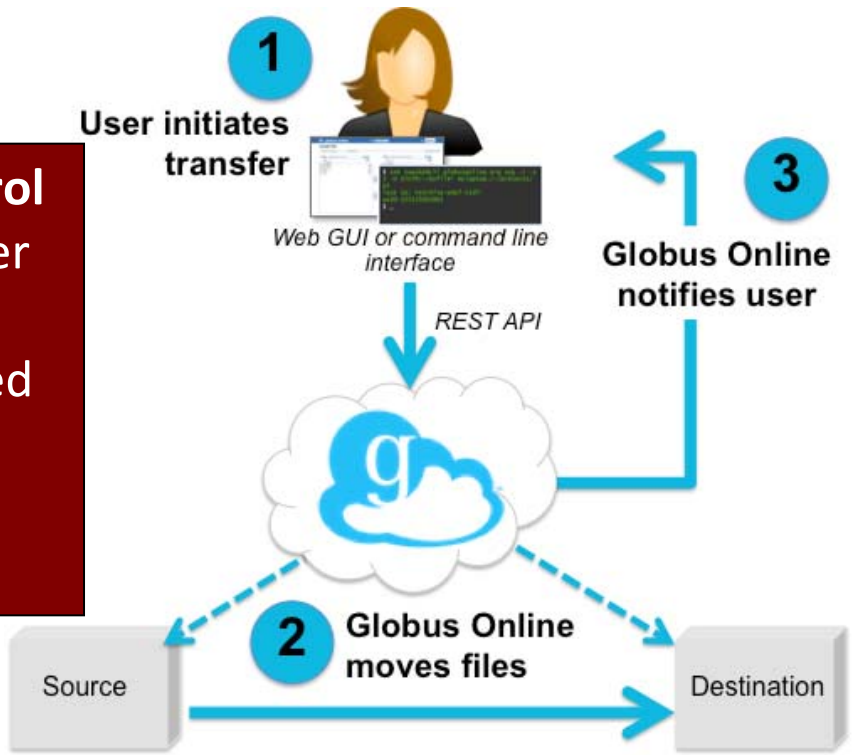
# Globus Online under the covers




**Globus Nexus** is used to manage
-- user identities
-- user profiles
-- groups and policies
-- resource definitions

# Globus Online under the covers

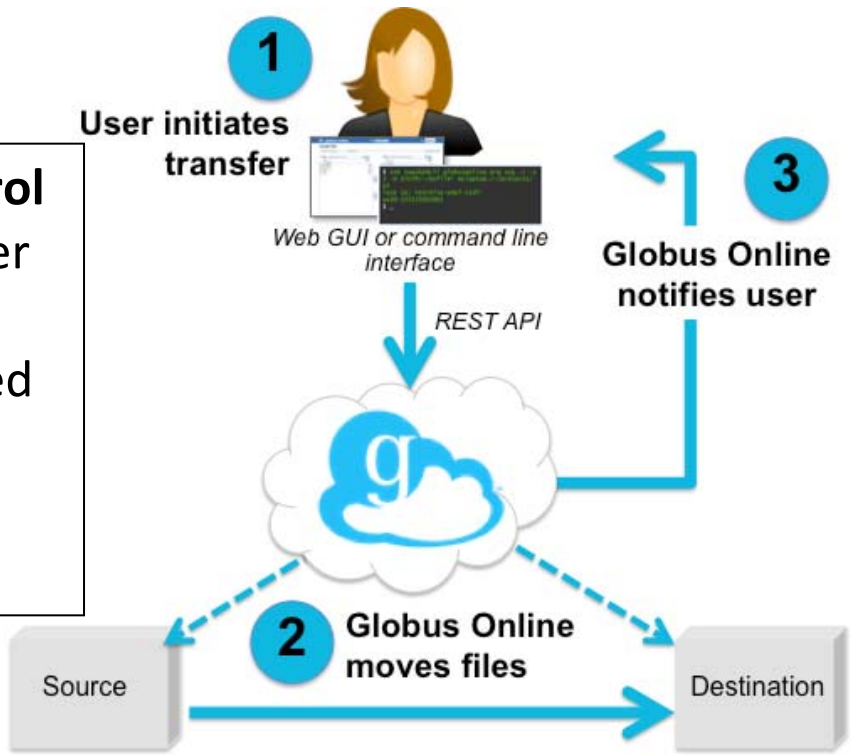
**Monitoring and control**
Auto-tuning of transfer parameters
Detection & attempted correction of errors
Manual intervention when required




**Globus Nexus** is used to manage
-- user identities
-- user profiles
-- groups and policies
-- resource definitions

# Globus Online under the covers

**Monitoring and control**
Auto-tuning of transfer parameters
Detection & attempted correction of errors
Manual intervention when required

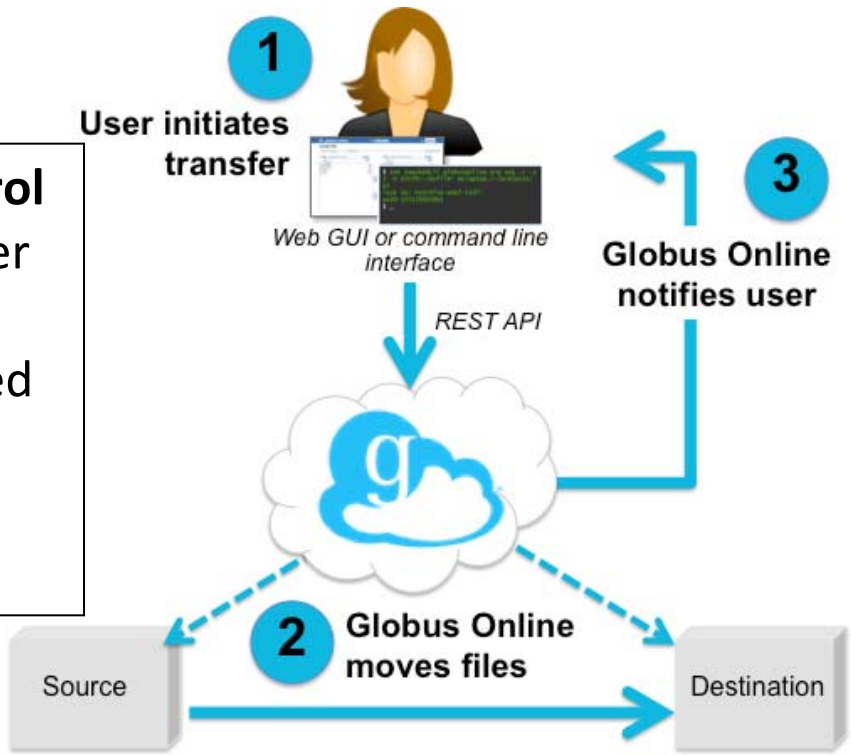



**Globus Nexus** is used to manage
-- user identities
-- user profiles
-- groups and policies
-- resource definitions

**Reliable cloud-based infrastructure**
EC2 for transfer management
S3 for system state
SimpleDB for lock management
Replication across availability zones

# Globus Online under the covers

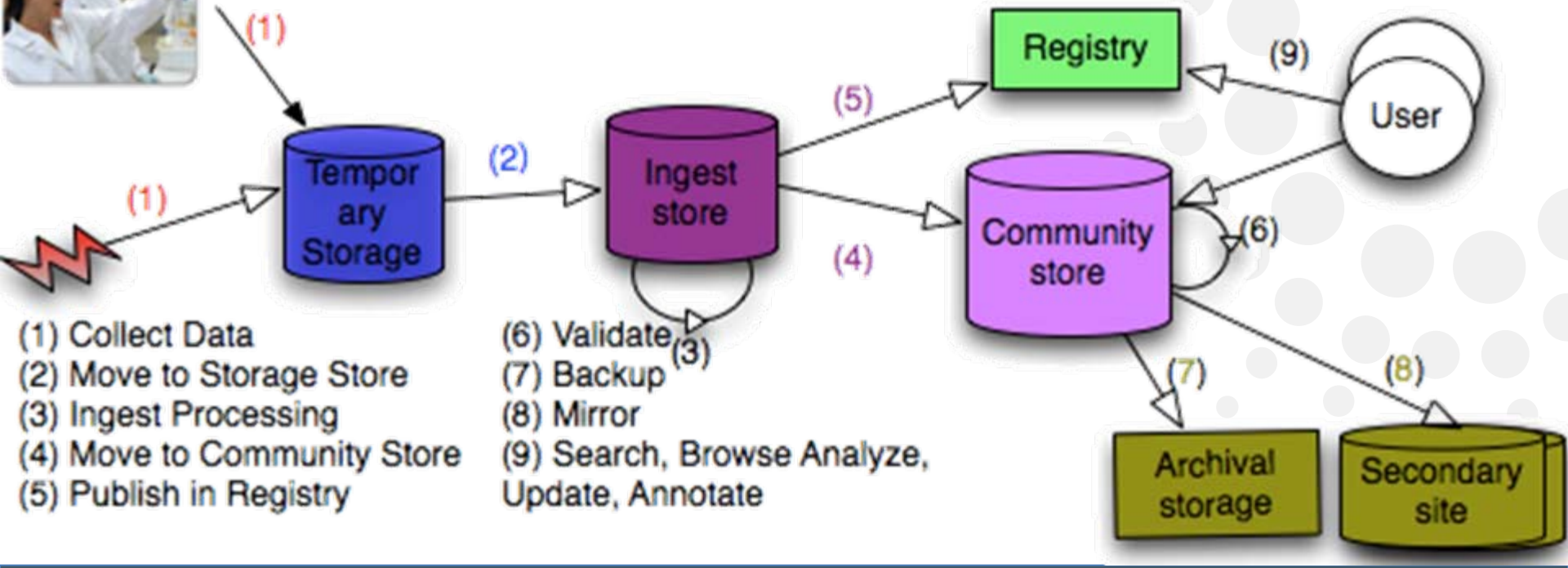
**Monitoring and control**
Auto-tuning of transfer parameters
Detection & attempted correction of errors
Manual intervention when required



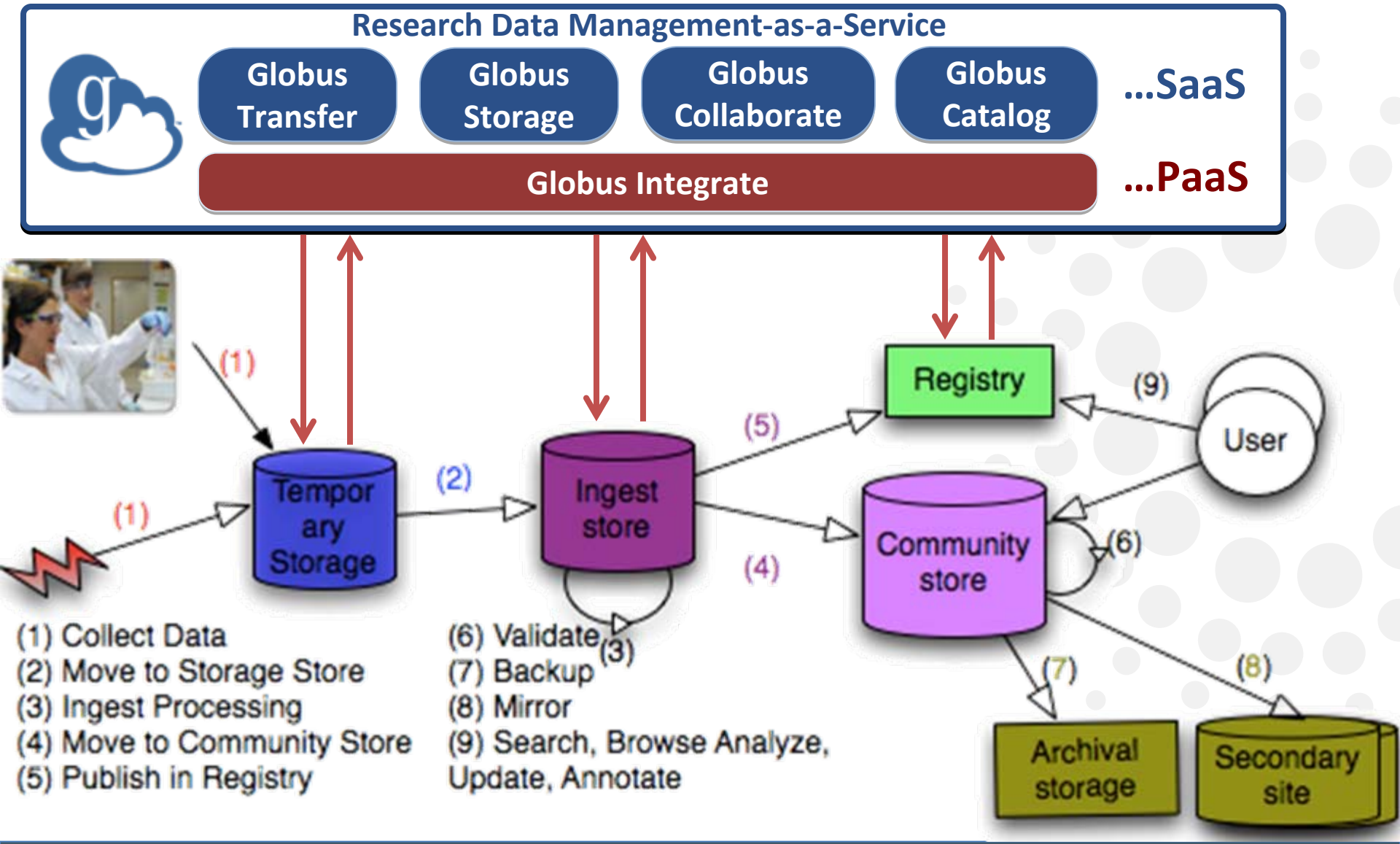

**Globus Nexus** is used to manage
-- user identities
-- user profiles
-- groups and policies
-- resource definitions

**Reliable cloud-based infrastructure**
EC2 for transfer management
S3 for system state
SimpleDB for lock management
Replication across availability zones

# A first take on “big process for science”

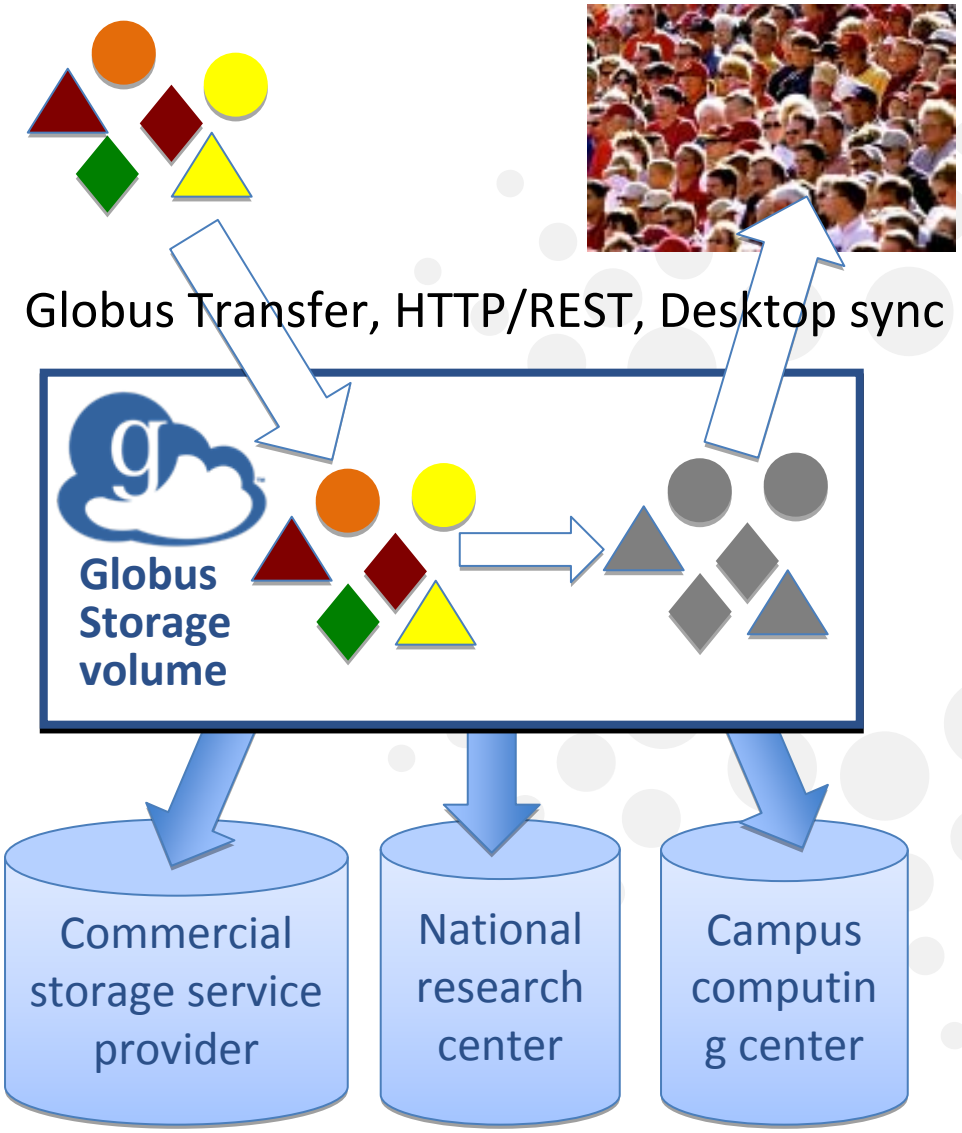


# A first take on "big process for science"

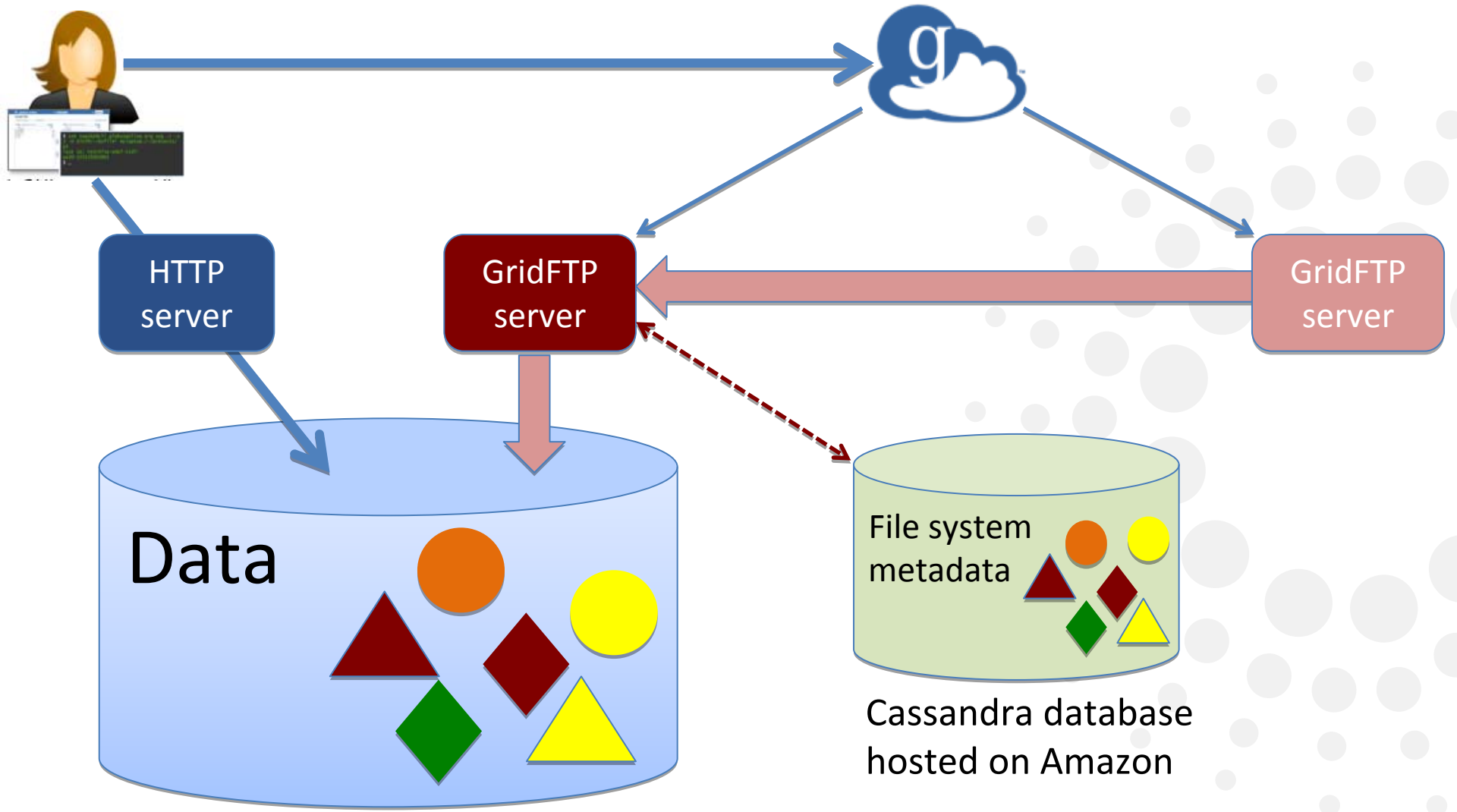


# Globus Storage: For when you want to …

- **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



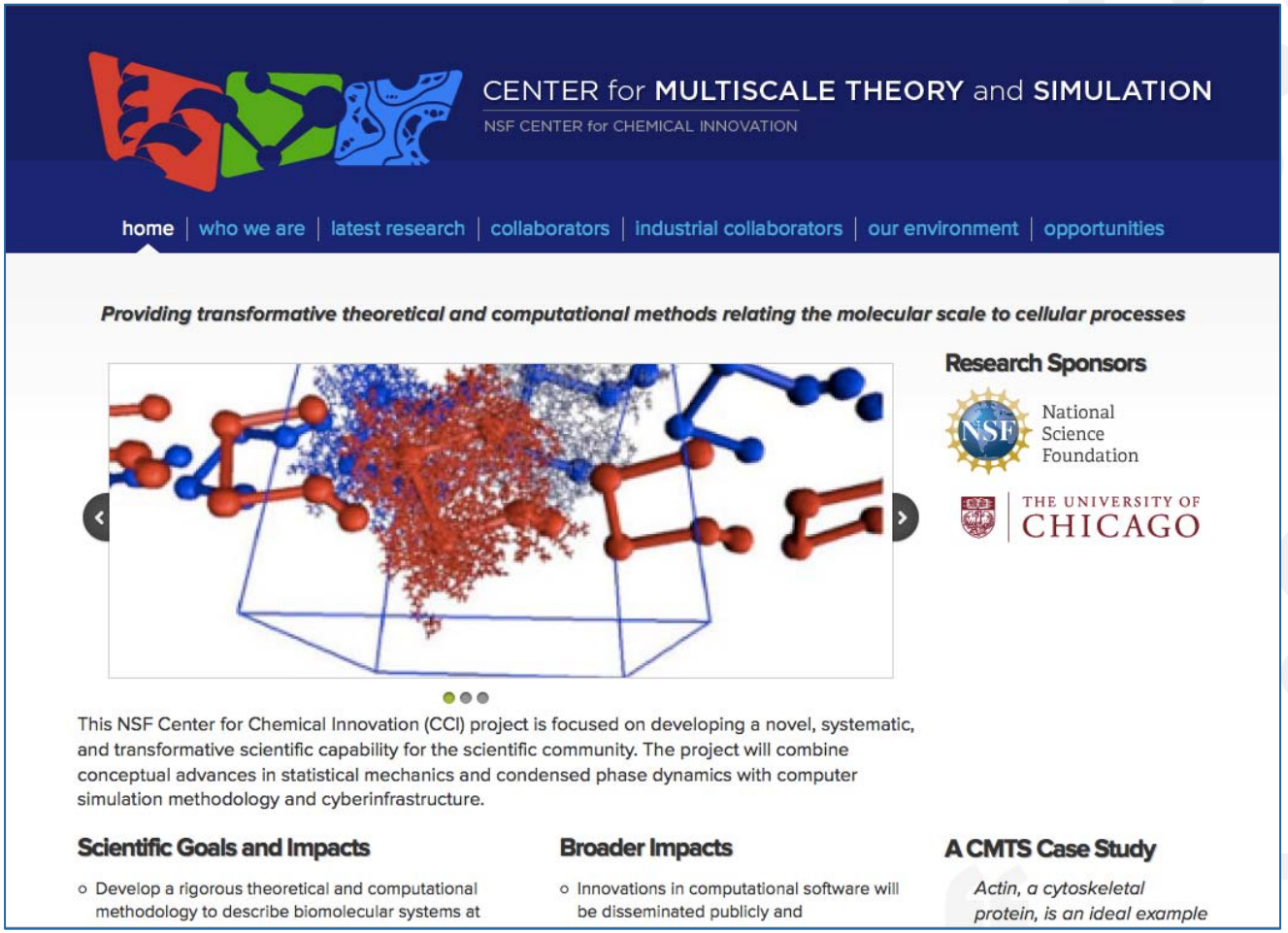
# Globus Collaborate: For when you want to

**Join with a few or many people to:**

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

**With:**

- Common groups
- Delegated management



# Globus Storage & Collaborate in action


**Globus Connect**
Move DTI results to Bryce's laptop


DTI Group
- Kyle
- **Bryce**

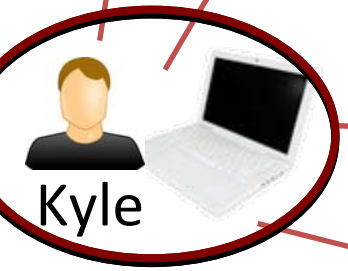
**Globus Storage**
Create snapshot to share with group

**Globus Transfer**
Copy TBI data to compute cluster

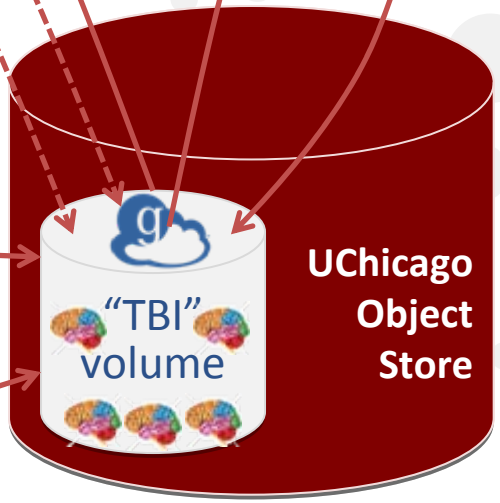
**Globus Nexus**
Add Bryce to TBI collaboration

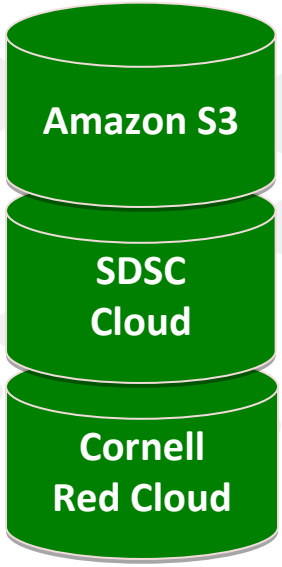
**Globus Transfer**
Move DTI results to shared volume


**Globus Collaborate**
Publish DTI data to TBI web site




**Globus Storage**
Create volume and share with TBI group

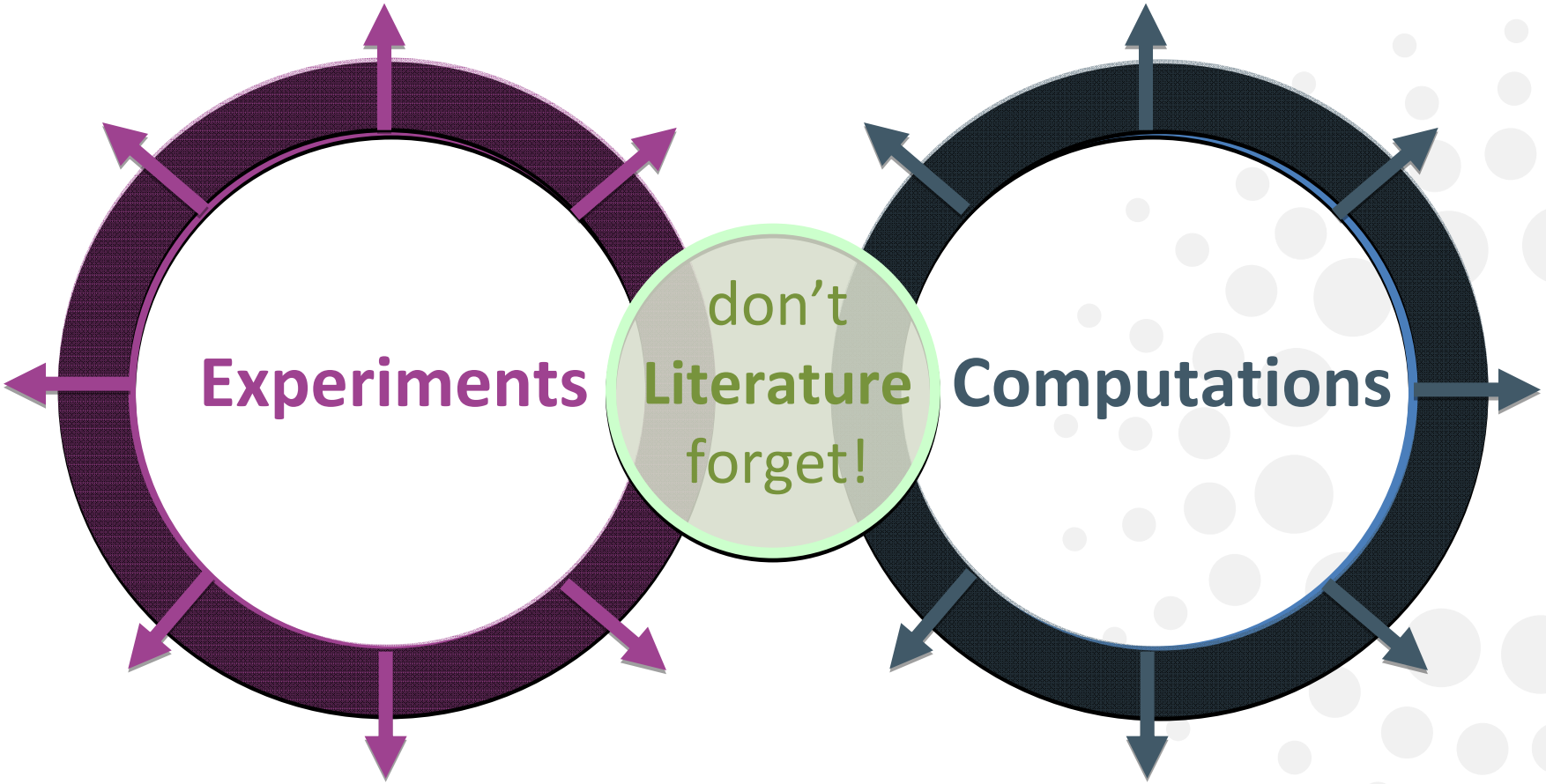



**Globus Connect**
Move MRI files to TBI shared volume

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

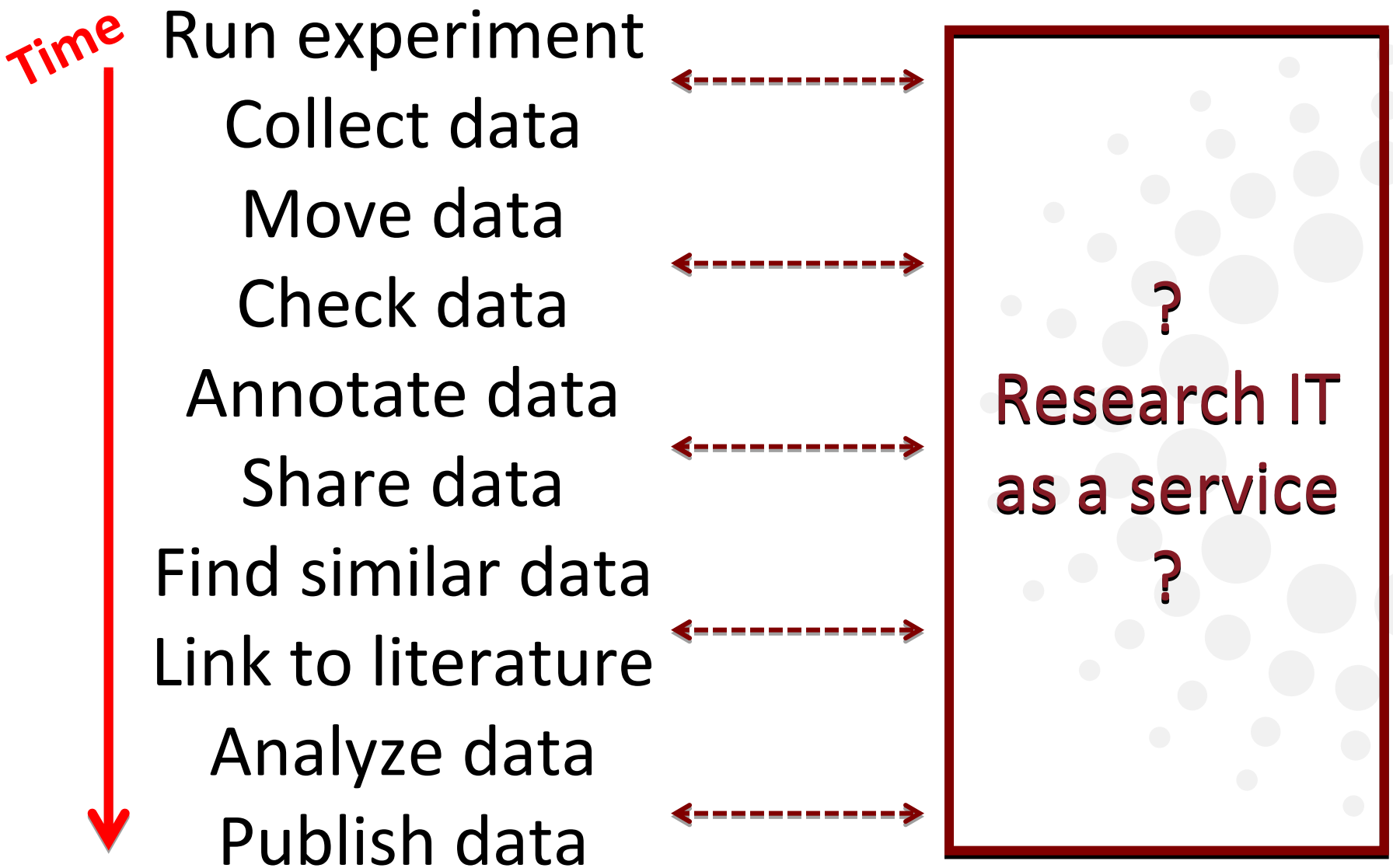
# Let's rethink how we provide **research IT**

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

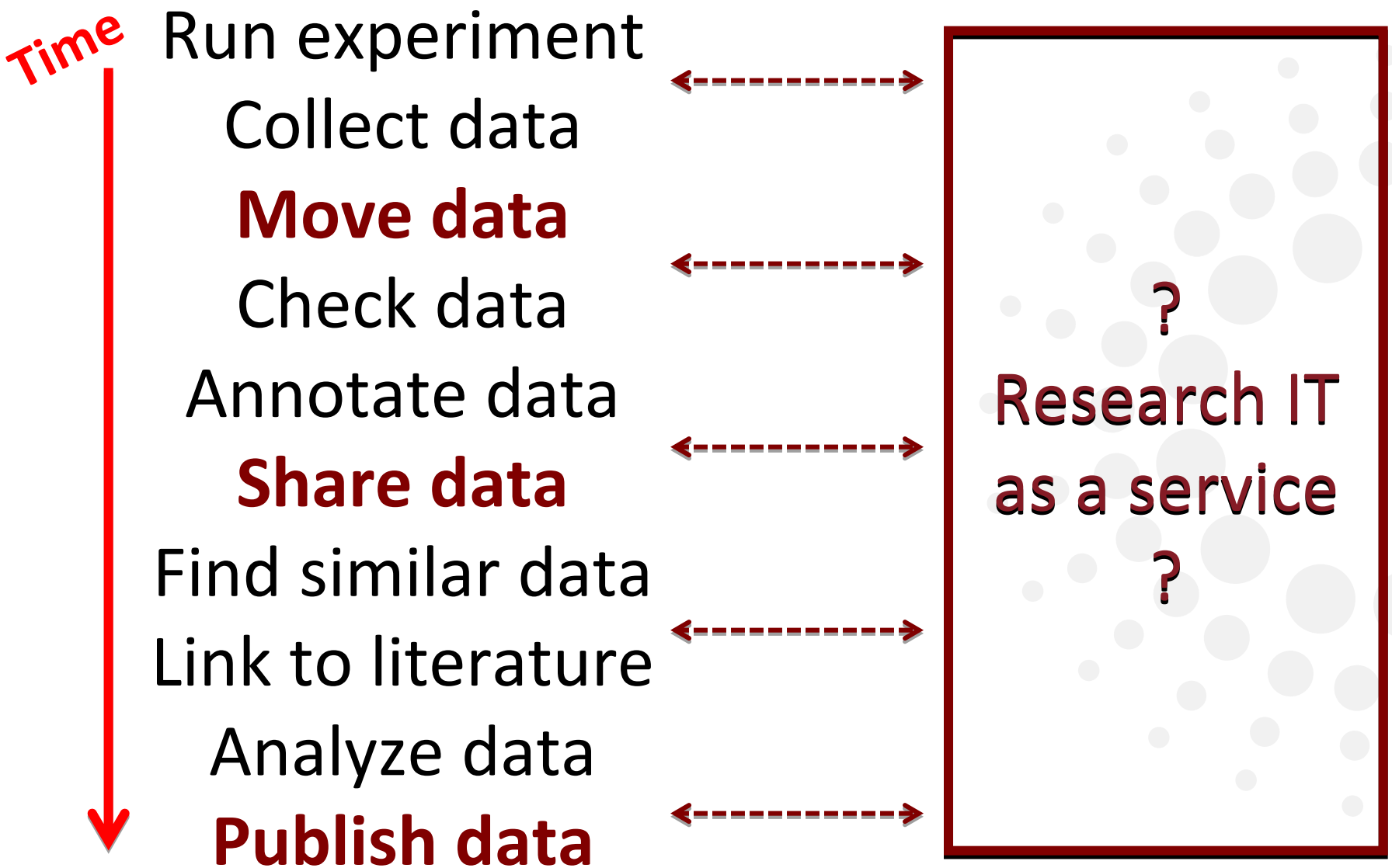
Leverage the cloud to

- 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

# Process automation for science



# Process automation for science



# Acknowledgements


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

Computation Institute

# Thank you!

globusonline.org

@globusonline

foster@anl.gov
foster@uchicago.edu