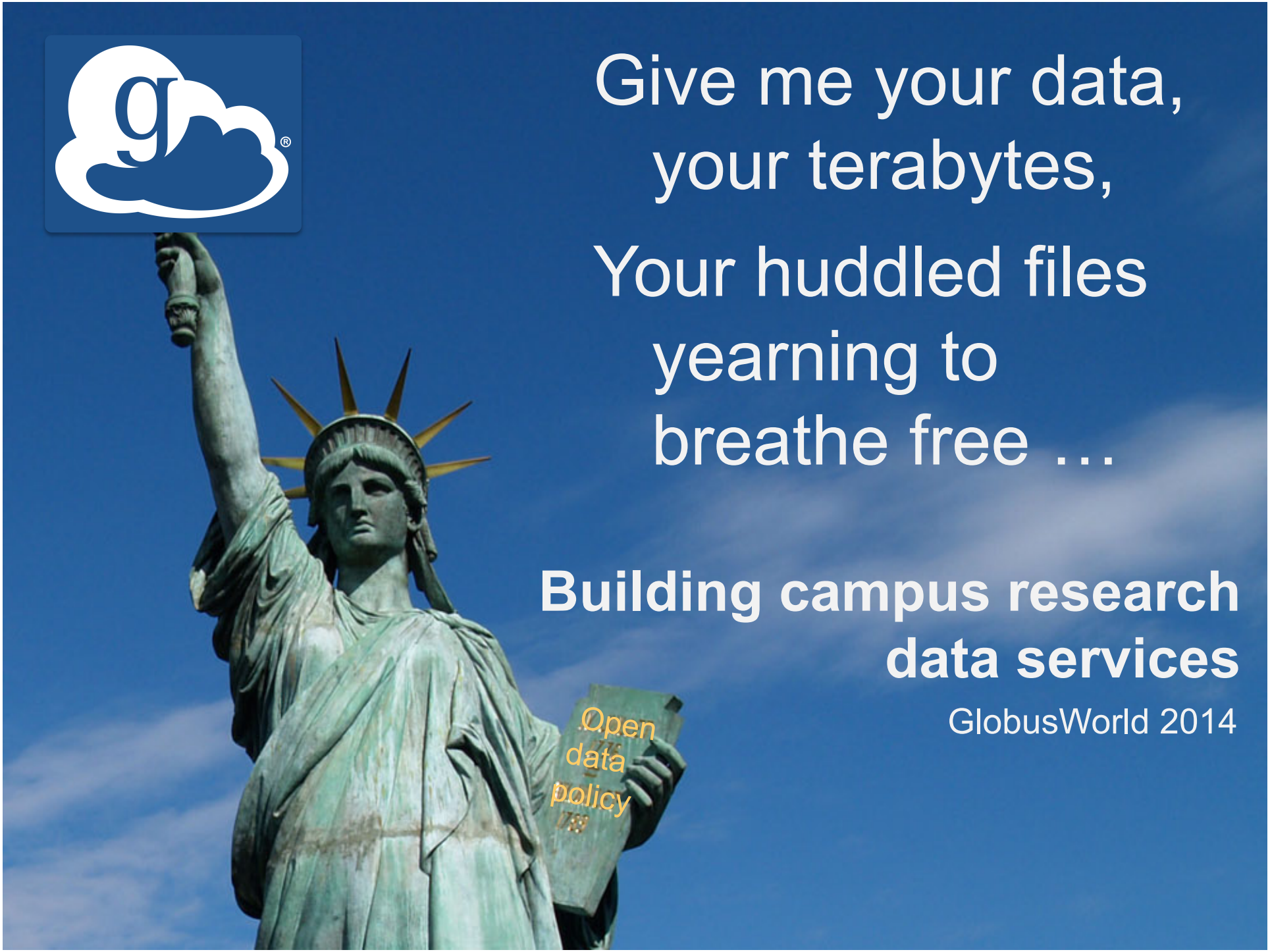
Delivering a Campus Research Data Service with Globus

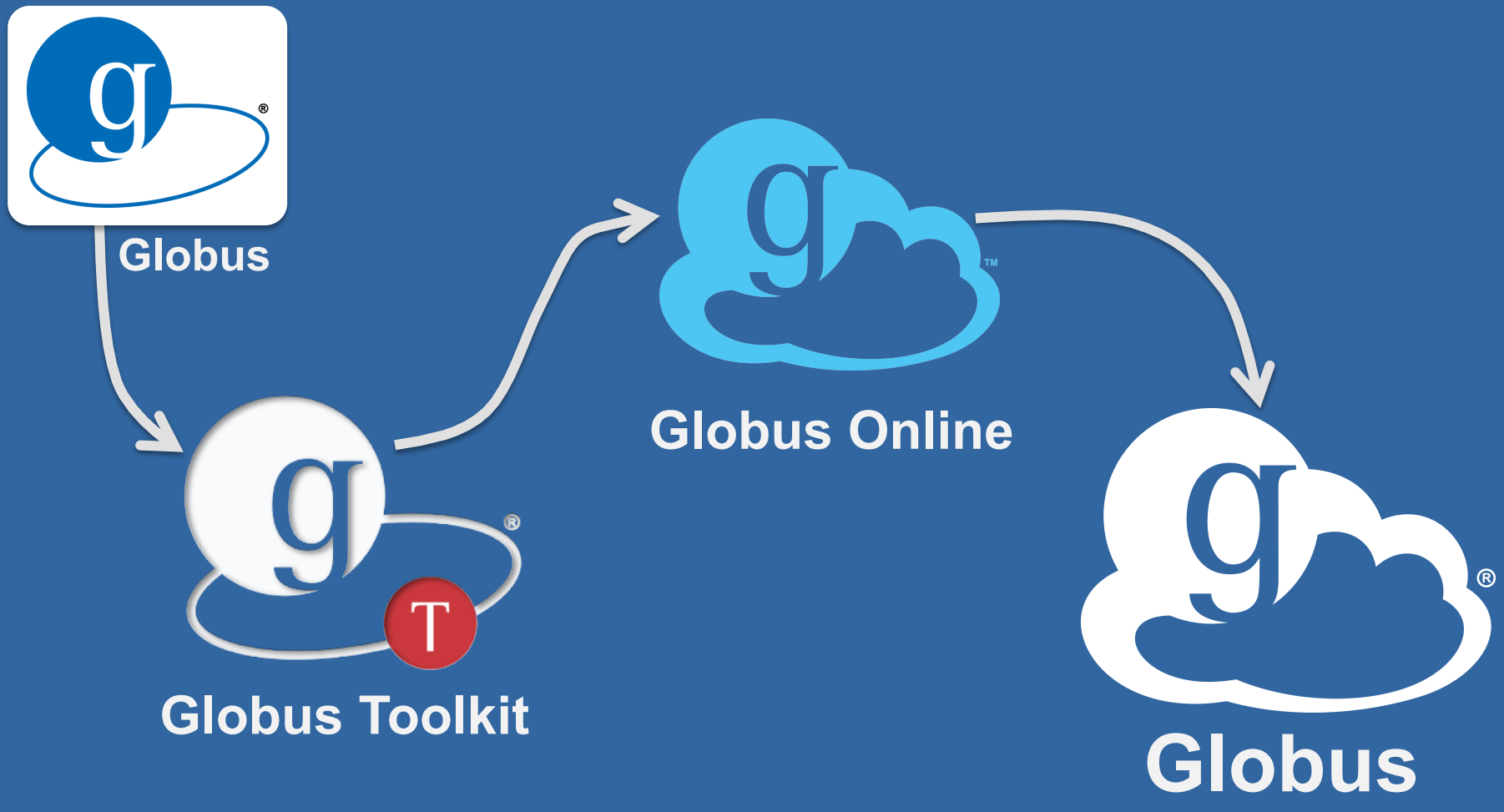
GlobusWorld 2014

Keynote


Give me your data,
your terabytes,
Your huddled files
yearning to
breathe free …
Building campus research data services
GlobusWorld 2014
Open data policy

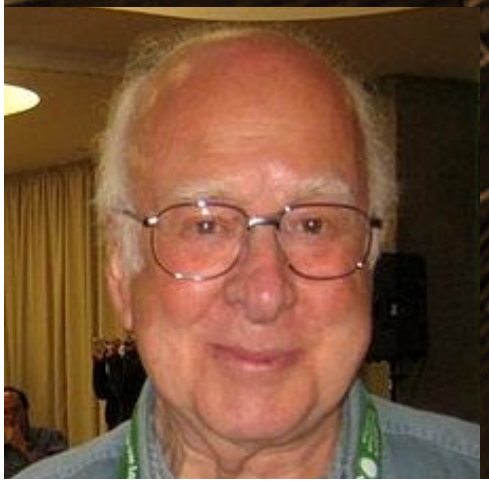
“It’s *deja vu* all over again.”

Yogi Berra



Higgs discovery “only possible because of the **extraordinary achievements** of … **grid computing**”
Rolf Heuer, CERN DG

10s of PB, 100s of institutions,1000s of scientists, 100Ks of CPUs, Bs of tasks



What is Globus (today)?

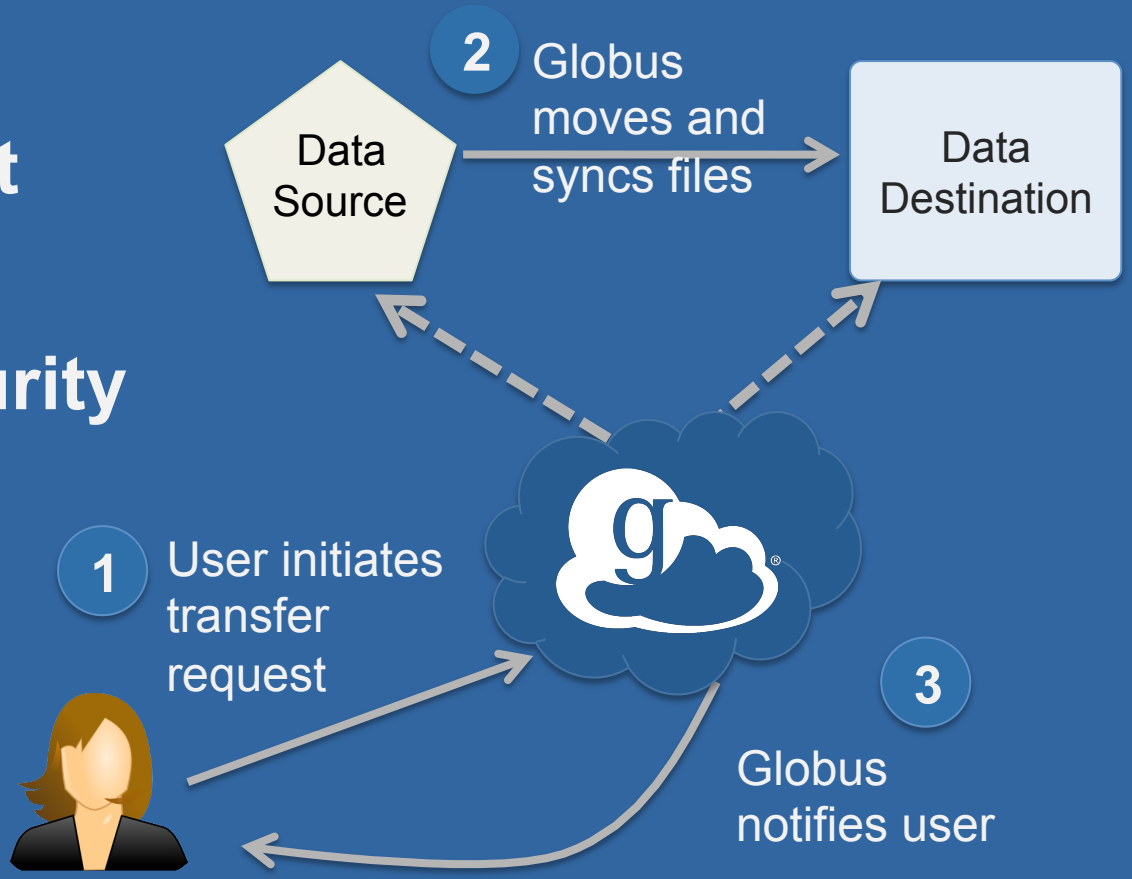
Big data transfer
and sharing…

…with Dropbox-like
simplicity…

…directly from your own
storage systems

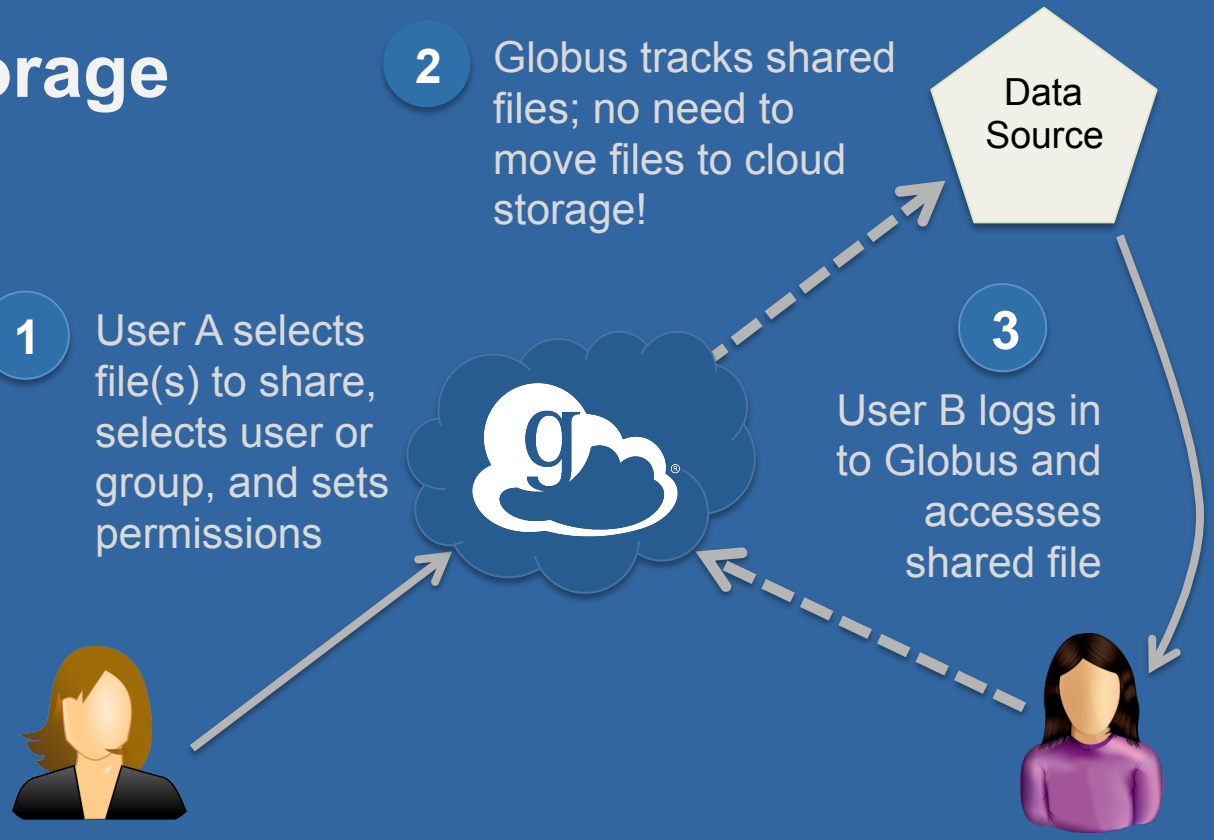
Reliable, secure, high-performance *file transfer* and *synchronization*

- **"Fire-and-forget" transfers**
- **Automatic fault recovery**
- **Seamless security integration**
- **Powerful GUI and APIs**



Simple, secure ***sharing*** off existing storage systems

- **Easily share large data with any user or group**
- **No cloud storage required**



15,000
registered users

8,000

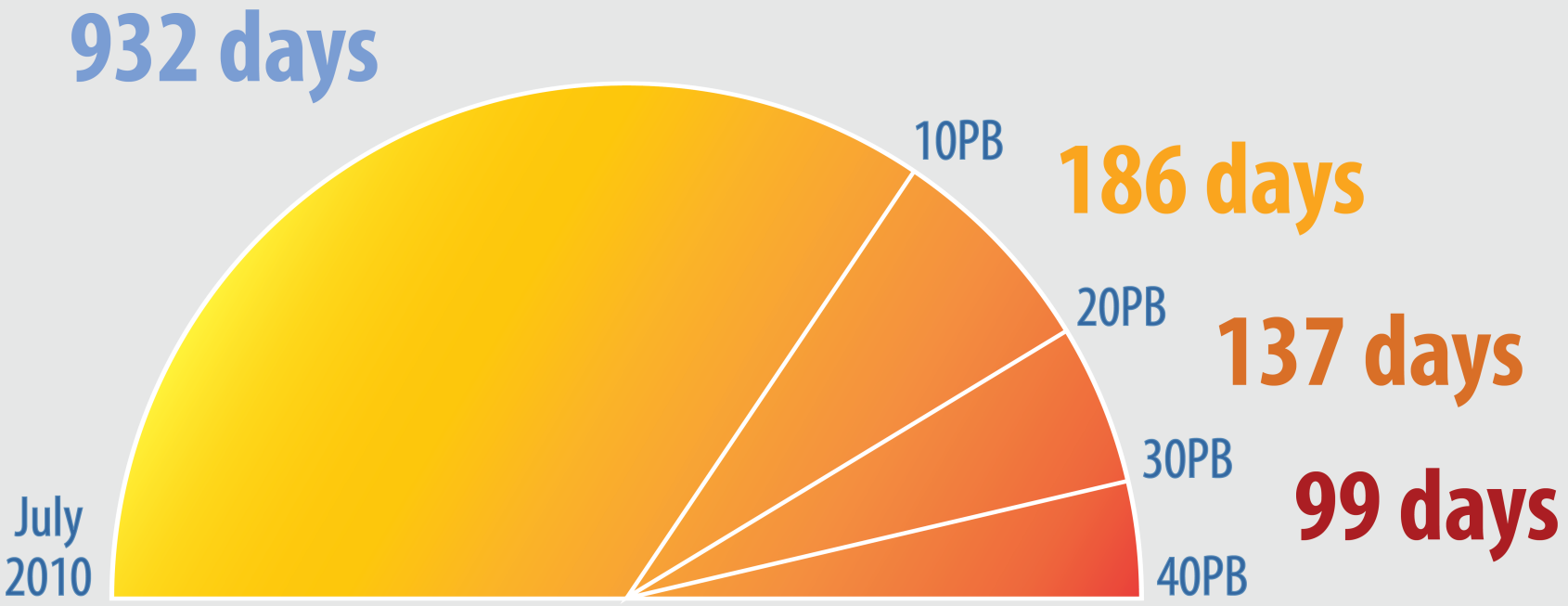
active endpoints

(in the past year)

3 billion

files transferred

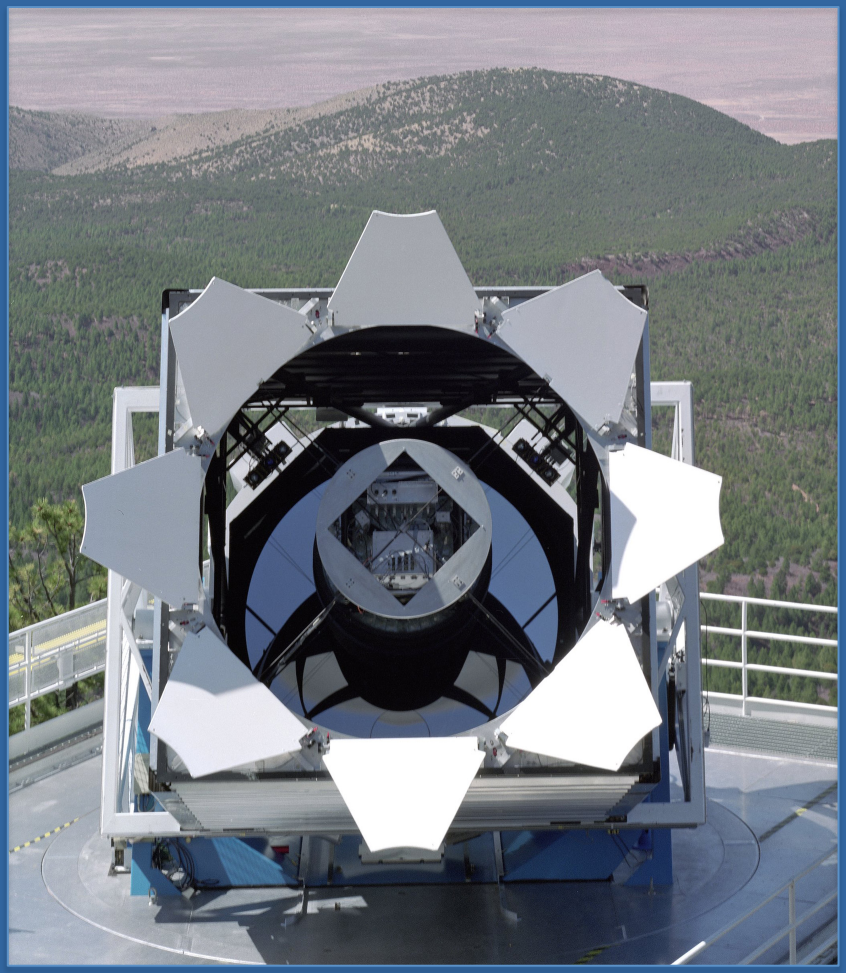
Moving the Needle



42.4PB as of April 15, 2014

Globus is enabling…

Study of the structure and evolution of galaxies, the nature of dark energy, and cosmological history of the universe

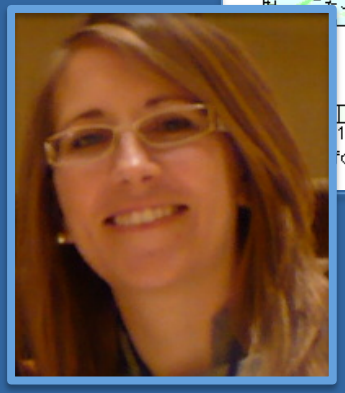


Sloan Digital Sky Survey
Source: University of Utah

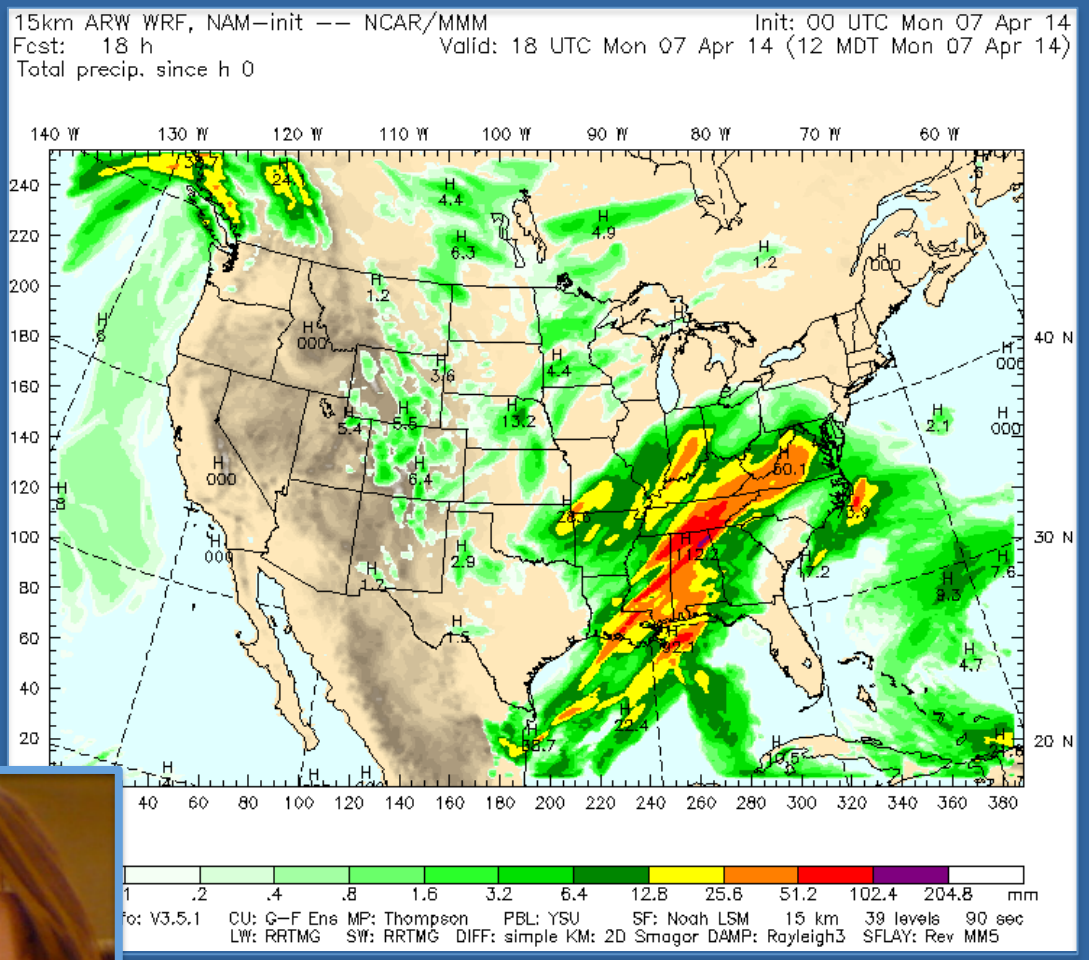
Joel Brownstein
University of Utah

Globus is enabling…

Development of numerical simulations of severe storms for improved responsiveness to weather events



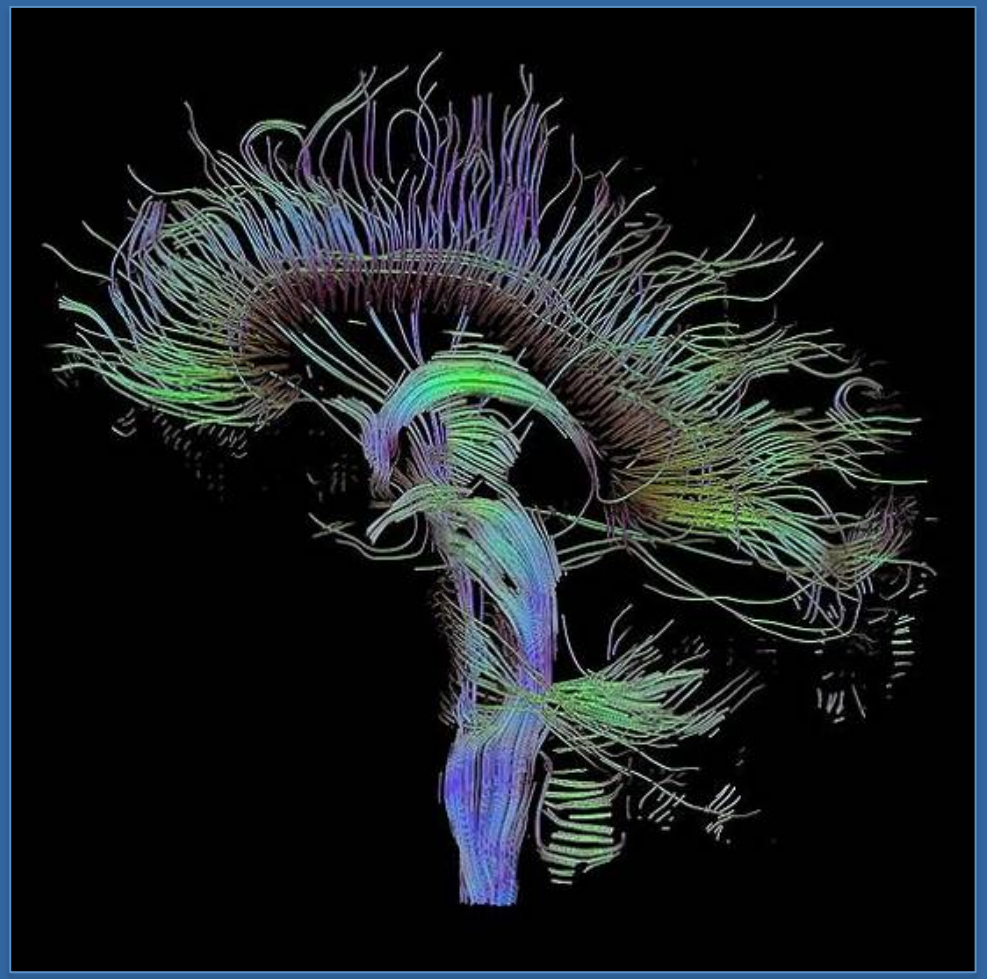
Ann Syrowski
University of Illinois



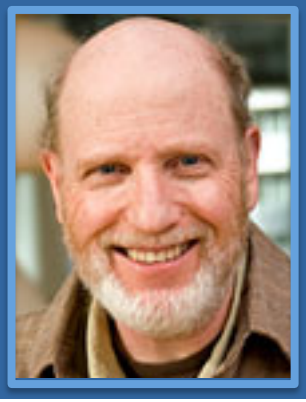
Weather Research and Forecasting Model
Source: UCAR

Globus is enabling…

Pediatric brain research by enhancing analysis of genetic material in pursuit of the underlying cause



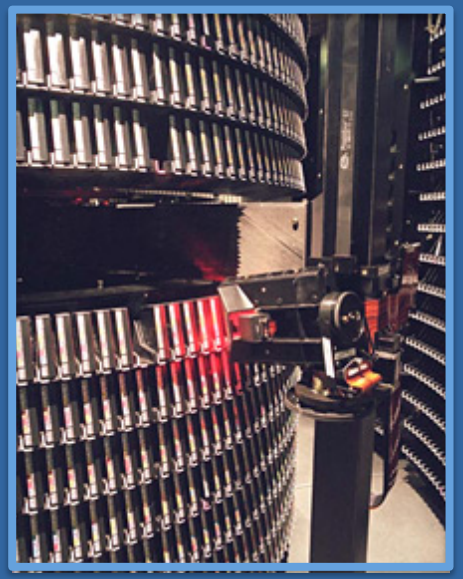
Communication impairment by genetic variants
Source: Wikimedia Commons



William Dobyns
U. Washington

“I need a good place to store / backup / archive my (big) research data, at a reasonable price.”

Campus Store

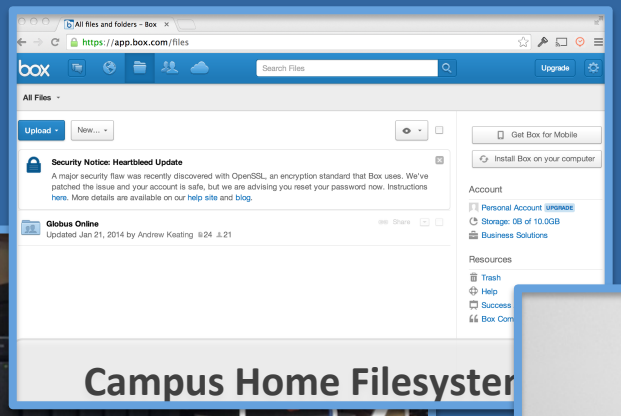


Mass Store

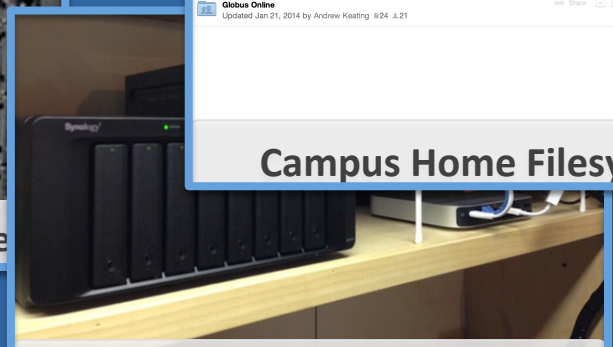
Public Cloud Archive

“I need to easily, quickly, & reliably move or mirror portions of my data to other places.”

Research Computing HPC Cluste



Campus Home Filesyster



Lab Server

Personal Laptop

Desktop Workstation

XSEDE Resource

amazon web services S3

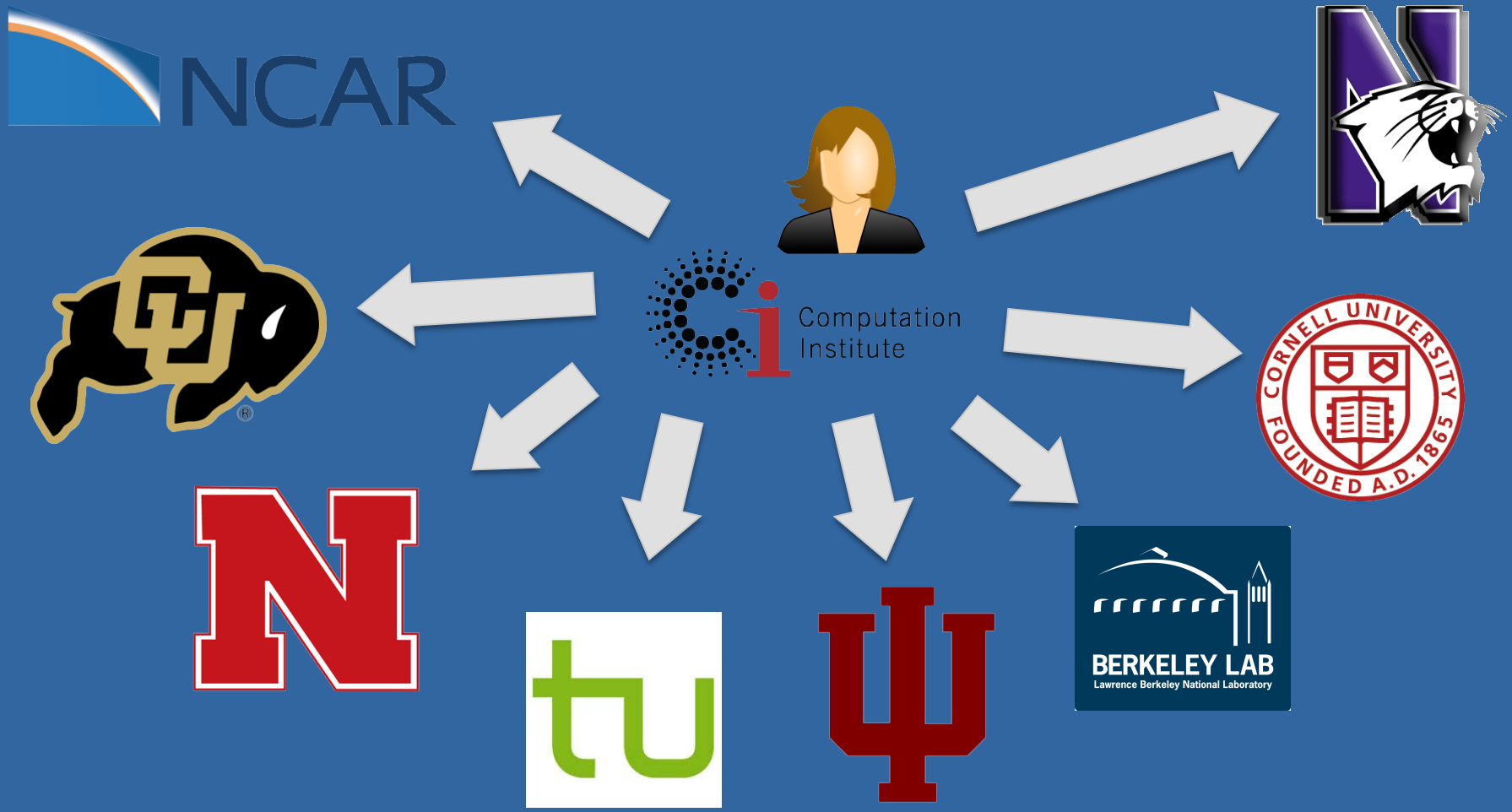
red cloud

Windows Azure

Google Cloud Platform

Public Cloud

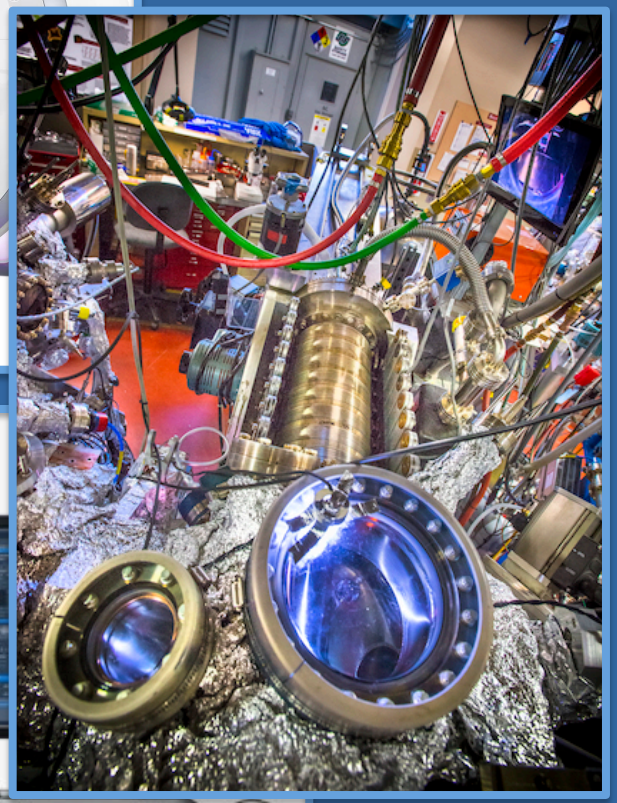
“I need to easily and securely share my data with my colleagues at other institutions.”



“I need to get data from a scientific instrument to my analysis server.”

MRI

Advanced Light Source



Next Gen Sequencer

Light Sheet Microscope

Product highlights

since GlobusWorld 2013

Sharing generally available

News Center — UNIVERSITY COMMUNICATIONS AND PUBLIC AFFAIRS — UC San Diego

All News | News by topic | Calendar | This Week | Subscribe | Search: News Center | Go

April 07, 2014 | By Jan Zverina

General | Research | Science & Engineering | SDSC

SDSC Enables Large-Scale Data Sharing Using Globus

The San Diego Supercomputer Center (SDSC) at the University of California, San Diego, has implemented a new feature of the Globus software that will allow researchers using the Center's computational and storage resources to easily and securely access and share large data sets with colleagues.

In the era of "Big Data"-based science, accessing and sharing of data plays a key role for scientific collaboration and research. Among SDSC users there is a need to share datasets, which can be large, with collaborators who may not have accounts on SDSC resources. The new Globus feature addresses this need.

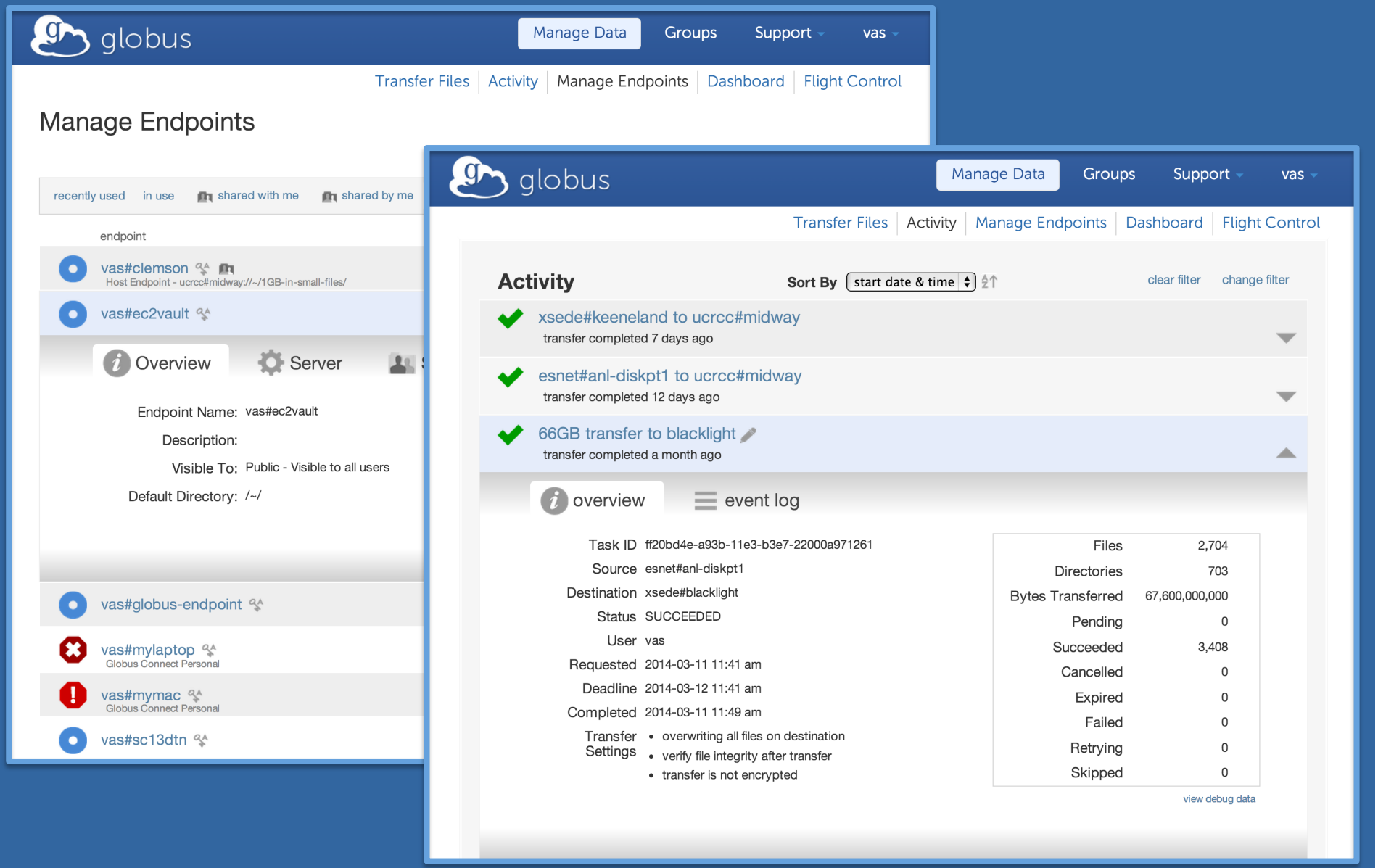
Described as a "dropbox for science", Globus is already widely used by resource providers and users who need a secure and reliable way to transfer files. SDSC is the first supercomputer cente in the National Science Foundation's XSEDE (eXtreme Science and Engineering Discovery Environment) program to offer the new and unique Globus sharing service.

While SDSC has been offering file transfer capability via Globus to users for several years, the Center is now providing a number of Globus Plus accounts via a Globus Provider plan to selected users free of charge so that they can allow their collaborators, including those who don't have an account on SDSC clusters, to access (read and write to their shared file space) data on SDSC resources.

SDSC staff will issue these accounts based on researchers' needs for sharing data with their collaborators, such as if they are part of a larger collaboration where data sharing becomes crucial. Separately, researchers will be able to purchase a Globus Plus account from Globus directly, with subscriptions currently priced at $7/month or $70/year.

globus plus®

Much improved Web UI



Globus Connect Server

- **Native RPM and Debian packaging**
- **Improved configuration management**
- **Multi-server setup**
- **OAuth support**

Management console: “Flight Control”

globus

Manage Data | Groups | Support | vas

Transfer Files | Activity | Manage Endpoints | Dashboard | Flight Control

Flight Control *Beta*

Search By Task Search By Endpoint

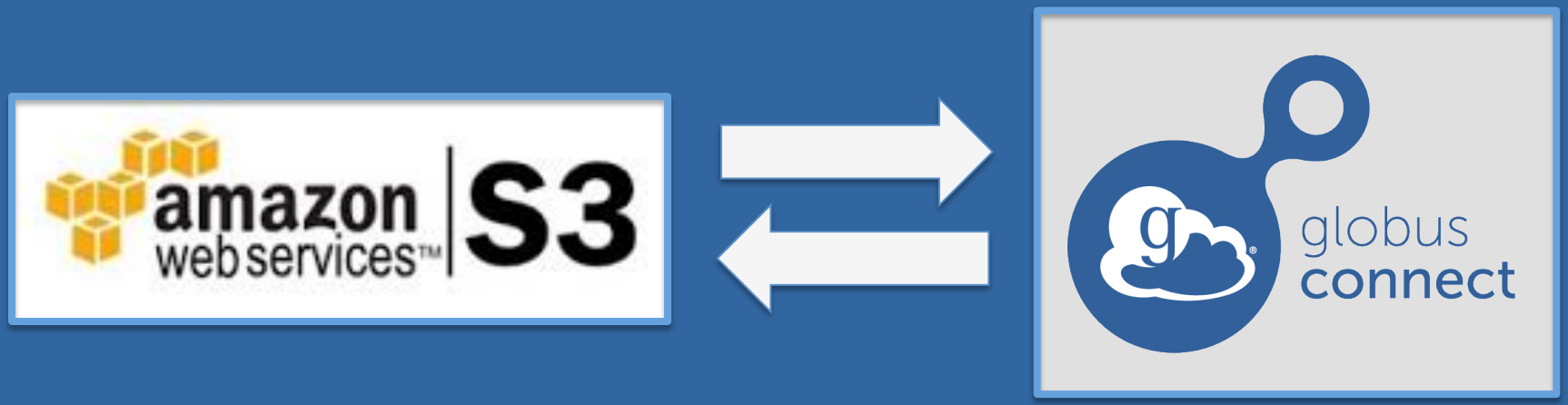
Active Jobs

Status	User	Req. (UTC)	Tasks	OK	Failed	BytesTX	Mbps	Retrying	Duration	Deadline	Faults	Source	Destination	Type	Flags	Task ID
No Jobs Found																

Job History

User	Completed (UTC)	Tasks	OK	Failed	BytesTX	Mbps	Duration	Faults	Source	Destination	Type	Flags	Task ID
vas	2014-04-11 16:16:03	29	2	27	0	0.00	72:00:50	7056	ucrcc#midway	vas#globus-endpoint (S3)	Web	VERIFY	f6bdaf5a...
tuecke	2014-02-18 19:42:09	1	1	0	11.835 MB	4.12	00:00:23	0	vas#projectX (Sharing)	tuecke#bukkitshare (Sharing)	Web	SYNC=3, VERIFY	b2c9b850...
vas	2014-02-18 18:18:13	27	27	0	432.829 MB	40.74	00:01:25	0	xsede#keeneland	vas#macret (GCP)	Web	VERIFY	d478a3d2...
vas	2014-02-18 16:22:43	2	2	0	11.161 MB	17.86	00:00:05	0	xsede#keeneland	vas#macret (GCP)	Web	VERIFY	e1a8ae4a...
vas	2014-02-18 16:18:49	20	20	0	410.500 MB	71.39	00:00:46	0	xsede#keeneland	vas#macret (GCP)	Web	VERIFY	3d723346...
tuecke	2014-02-18 16:18:23	1	1	0	6 B	0.00	00:00:13	0	tuecke#mylaptop (GCP)	vas#projectX (Sharing)	Web	VERIFY	410c8a92...

Amazon S3 Endpoints



Demonstration

85

U.S. campuses

Best practice (fasterdata.es.net)

Create Data Transfer Nodes on existing (or new) storage with Globus Connect Server

…deploy in a Science DMZ…

…use Globus as the interface

We are a non-profit, delivering a production-grade service to the non-profit research community

We are a non-profit, delivering a production-grade service to the non-profit research community

Our challenge:

Sustainability

Globus Provider Subscriptions

- **Managed Endpoints**
  - Priority support
  - Management console
  - Usage reports
  - Mass Storage System optimization
  - Host shared endpoints
  - Integration support

- **Plus Subscriptions**
  - Create and manage shared endpoints
  - Personal transfers
- **Branded Web Site**
- **Alternate Identity Provider (InCommon is standard)**

https://www.globus.org/provider-plans

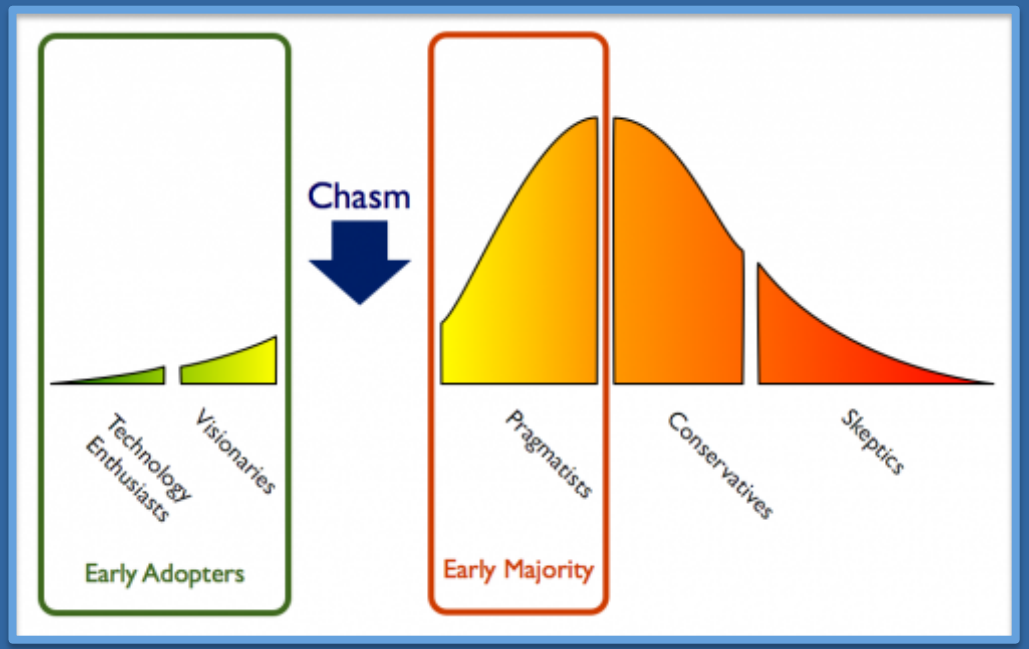
NET+ Globus

- **Internet2 members get discounted Globus Provider subscriptions**
- **Completing "Service Validation" phase**
  - Sponsors: Cornell, U.Michigan, Yale, U.Missouri, and U.Chicago
- **Available to "Early Adopters" soon**

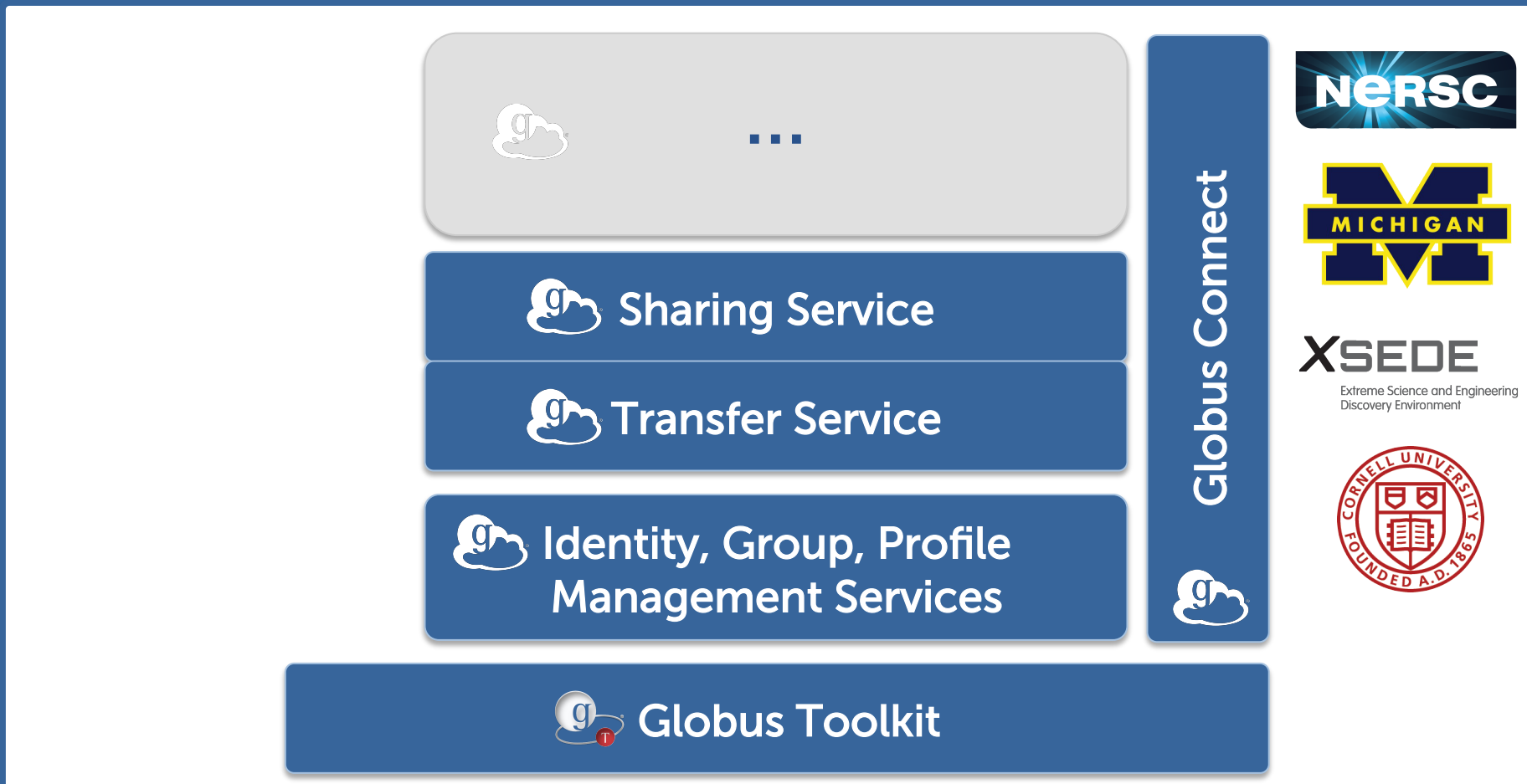
Internet2 NET+

Bridging the gap to sustainability

- **$500,000 from Sloan Foundation**
- **Recognition of what it takes to "cross the chasm"**
- **Funds non-R&D activities**
  - User Support
  - Operations
  - Marketing



Globus Under the Covers



Globus Platform-as-a-Service

globus genomics

Flexible, scalable, affordable genomics analysis for all biologists

Next-gen sequence analysis SaaS

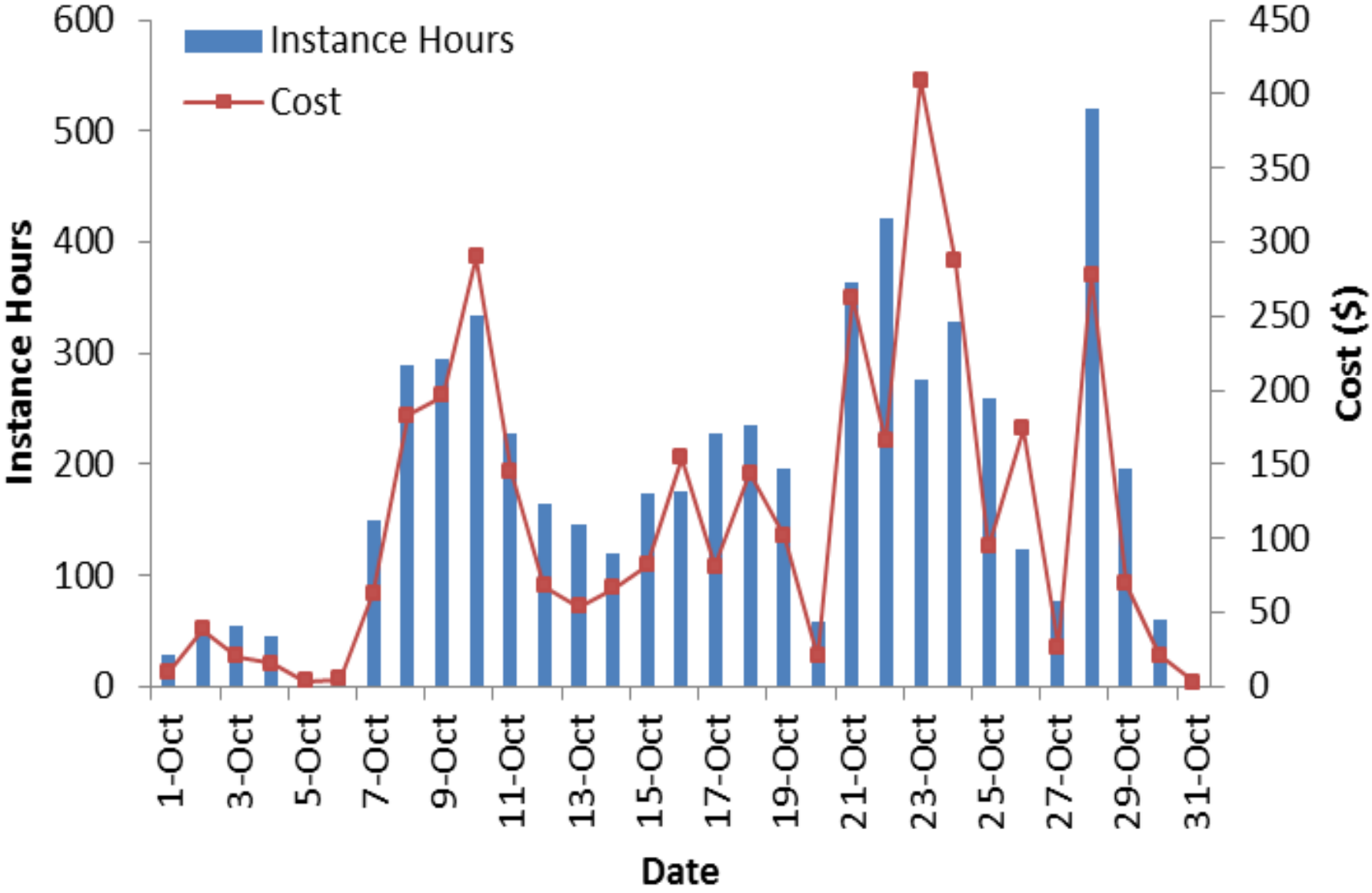
+

Data management PaaS

+

Scalable IaaS

Globus Genomics on AWS



Exome: $3 – $20
Whole Genome: $20 – $50
RNA-Seq: <$5

Alternatives are at 10-20x

UNIVERSITY OF WASHINGTON

Dobyns Lab
Exome analysis
20x speed-up
Next: 50x

THE UNIVERSITY OF CHICAGO

Cox Lab

Consensus variant calling

134 samples; 4 days

<0.01% Mendel error rate

Next: 13,000 samples

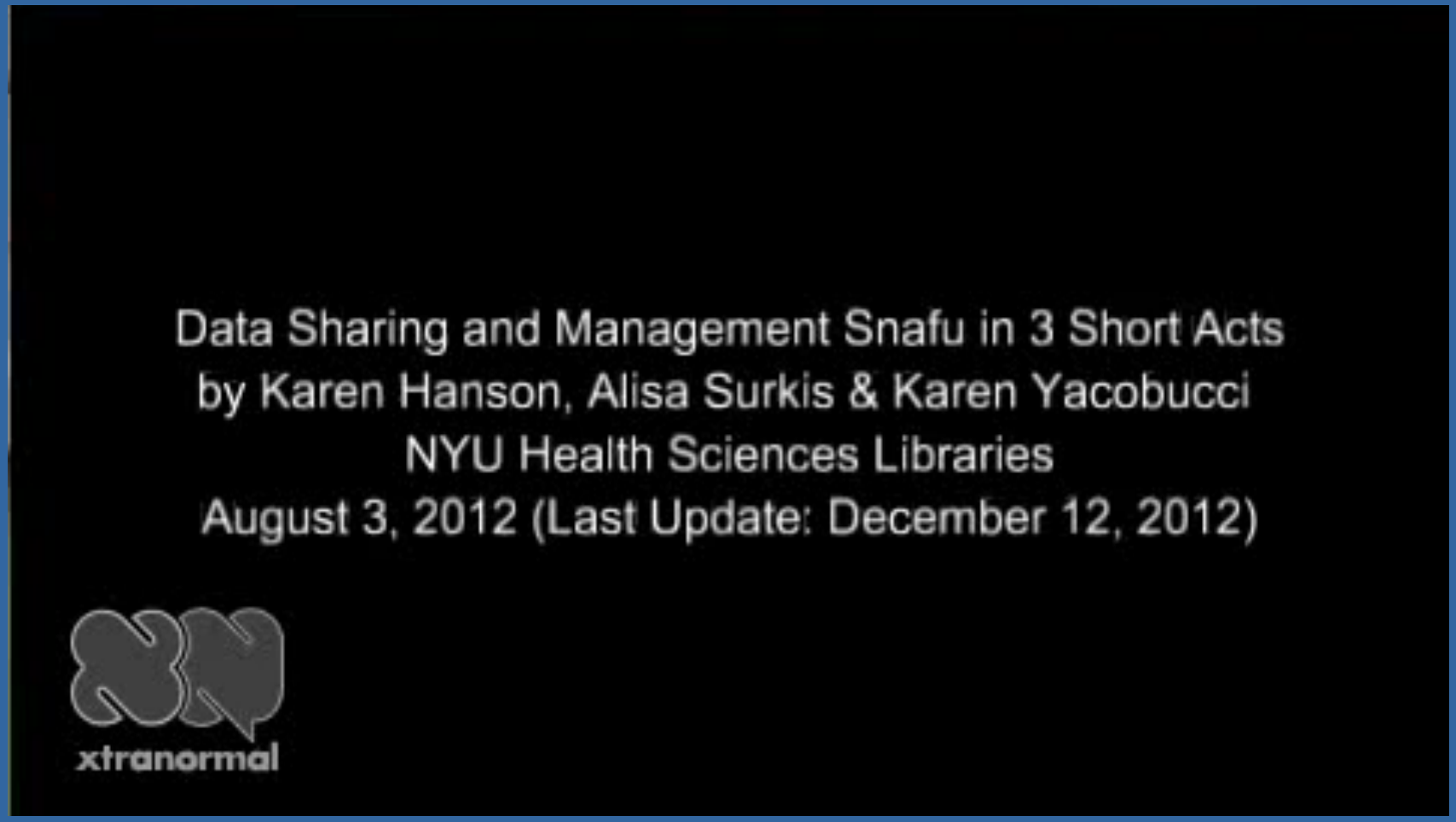
Campus Data Service User Stories

- **“I need a good place to store / backup / archive my (big) research data, at a reasonable price.”**
- **“I need to easily, quickly, and reliably move or mirror portions of my data to other places.”**
- **“I need a way to easily and securely share my data with my colleagues at other institutions.”**

Campus Data Service User Stories

- **“I need a good place to store / backup / archive my (big) research data, at a reasonable price.”**
- **“I need to easily, quickly, and reliably move or mirror portions of my data to other places.”**
- **“I need a way to easily and securely share my data with my colleagues at other institutions.”**
- **“I want to publish my data.”**
- **“I want to discover published data.”**

An all-too familiar tale …



What does it mean to **publish**?

Data is:

Identified

Described

Curated

Verifiable

Accessible

Preserved

What does it mean to **discover**?

I can:

Search

Browse

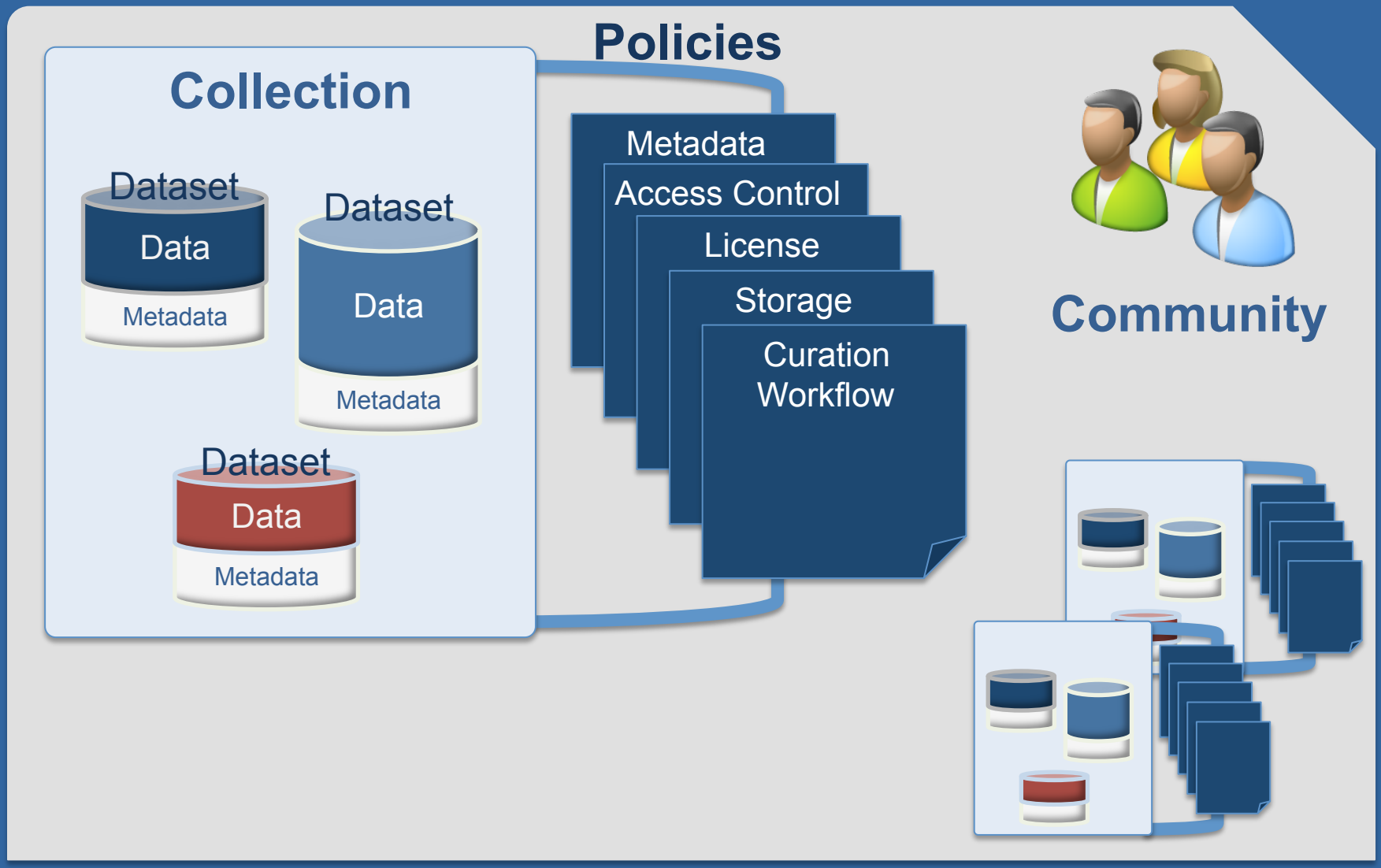
Access

the data

Announcing…

Globus data publication services

Teeing Up a Few Terms …



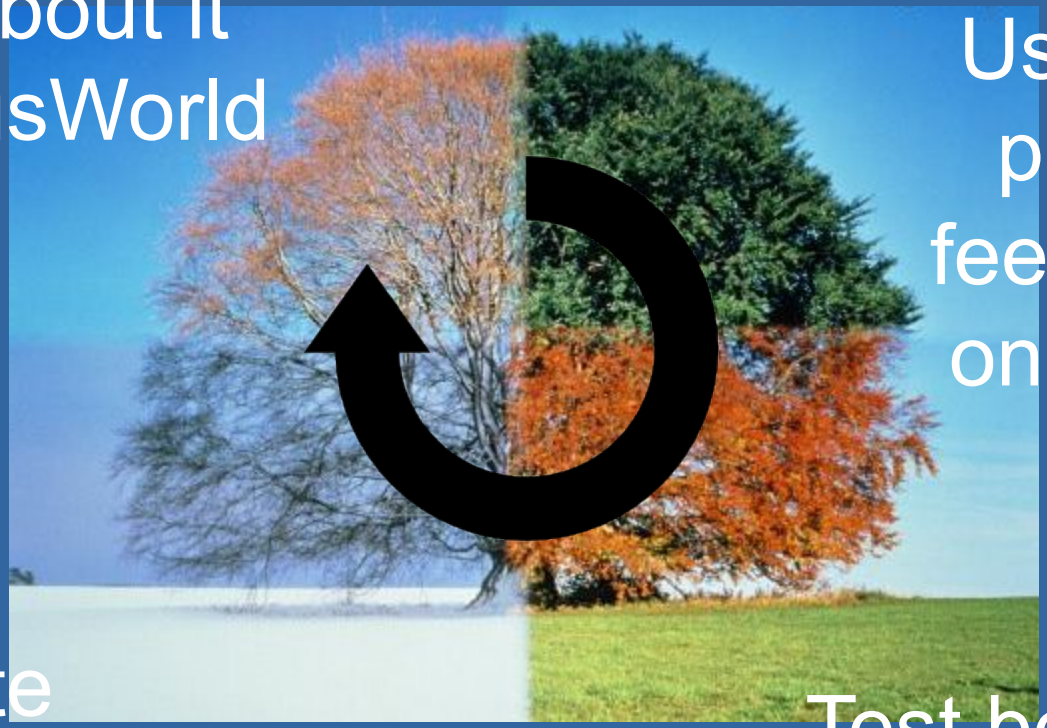
Demonstration

Recap: Globus Data Publication

- **SaaS for publishing large research data**
- **Bring your own storage**
- **Extensible metadata**
- **Publication and curation workflows**
- **Public and restricted collections**
- **Rich discovery model**

Looking for 3-5 early adopters

Spring:
Tell us about it at GlobusWorld 2015!



Summer:
Use and provide feedback on alpha

Winter:
Celebrate General Availability

Fall:
Test beta on your campus

Looking for 3-5 early adopters

Spring:
Tell us about it at GlobusWorld 2015!

Summer:
Use and provide feedback on alpha

Winter:
Celebrate General Availability

Fall:
Test beta on your campus

Our vision for 21st century research data management

To provide **affordable**, **advanced** capabilities for **all** researchers, delivering **sustainable** services that **aggregate** and **federate** existing resources

Thank you to our sponsors!

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY