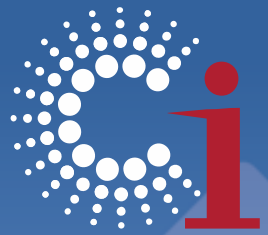
globus online



Computation Institute

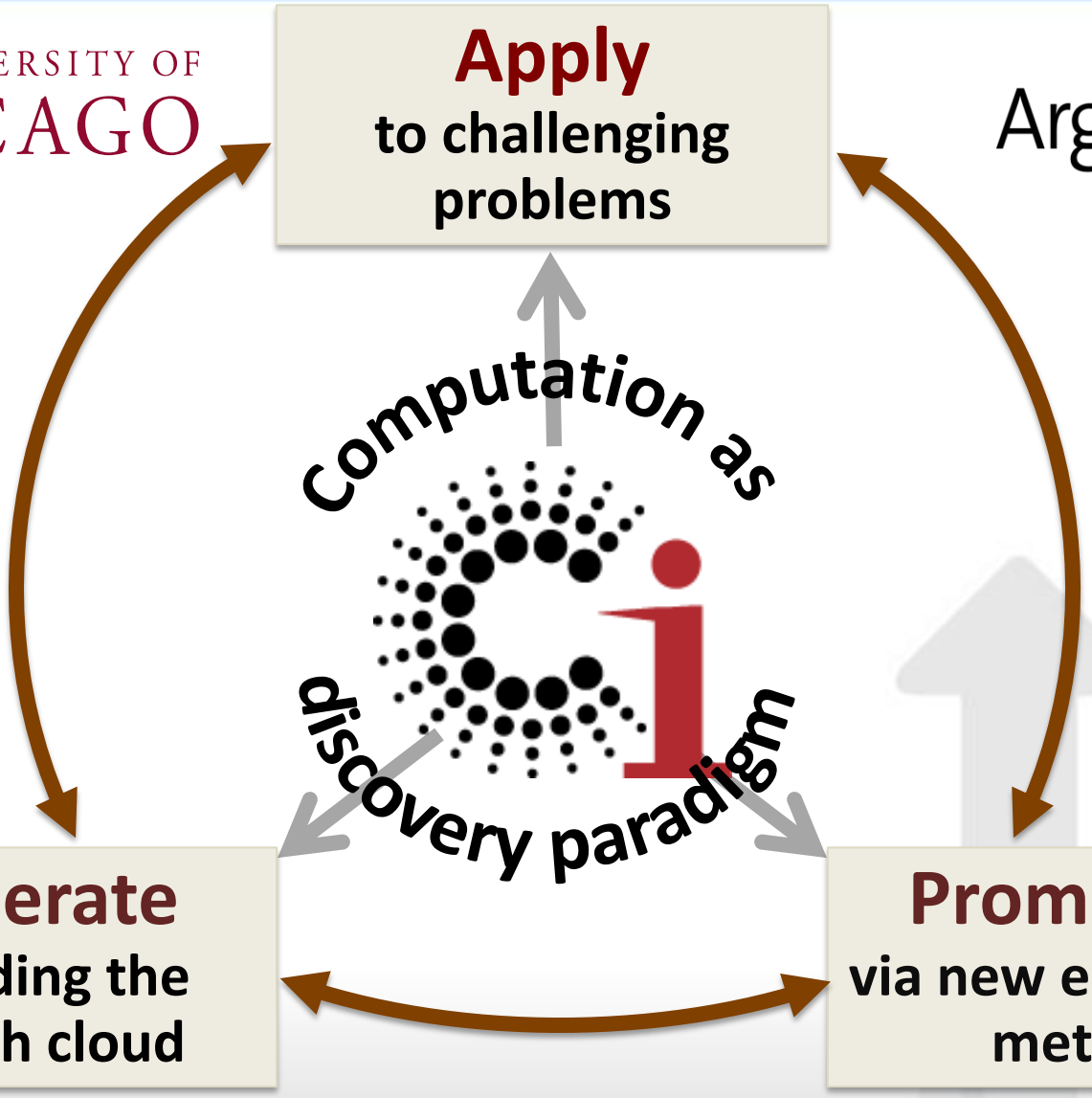
Cloud-based services for (reproducible) science

Ian Foster
Computation Institute
University of Chicago and Argonne National Laboratory

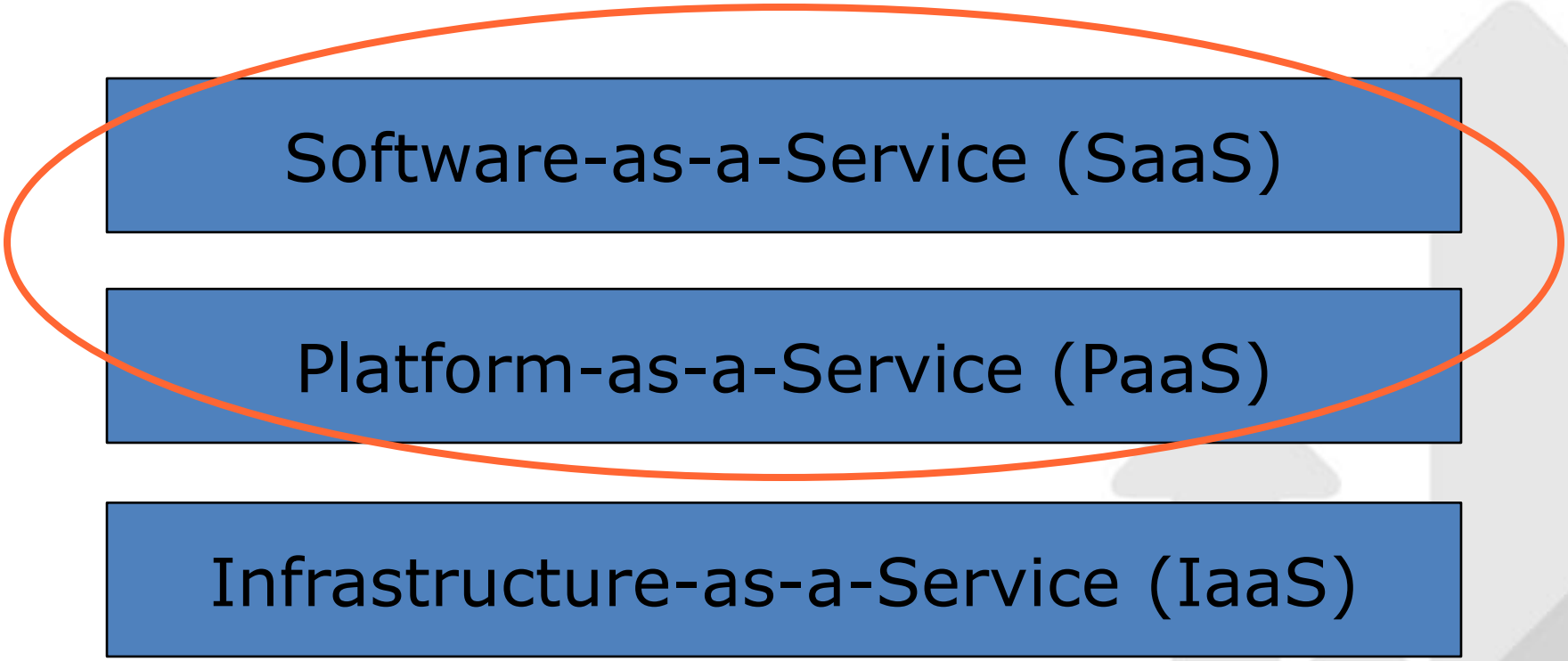
Computation Institute (CI)

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

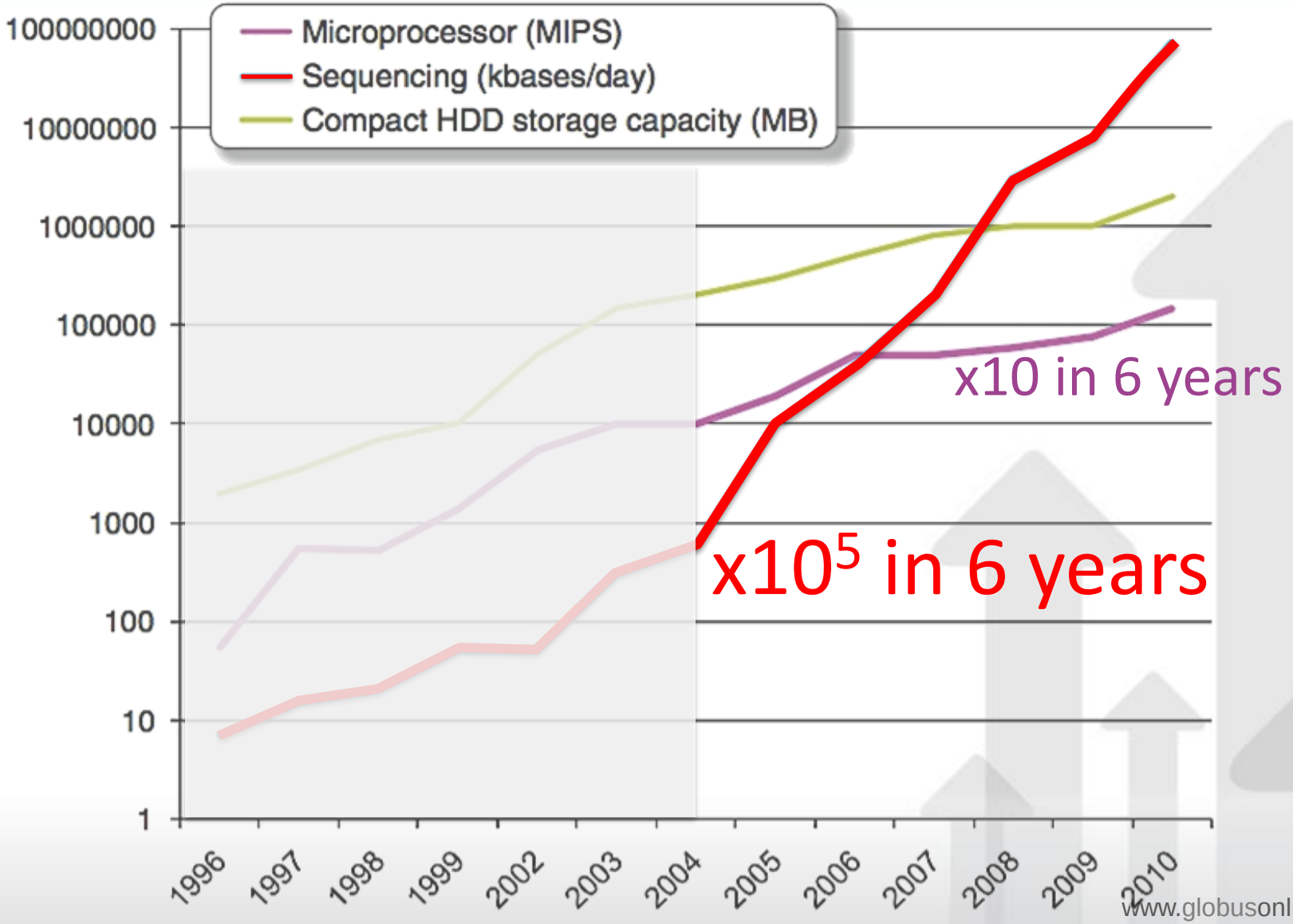


Cloud Layers

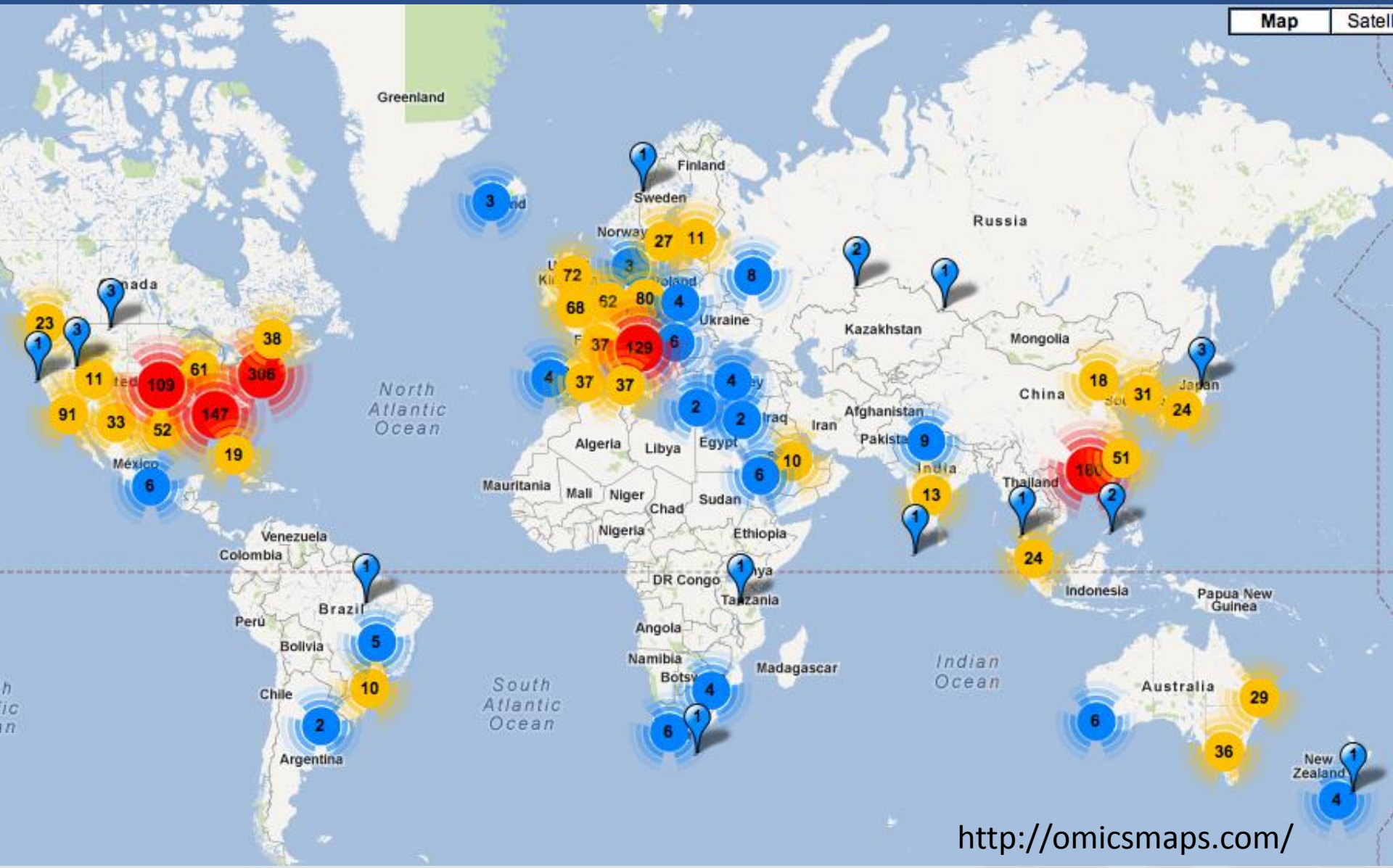


The data deluge

The data deluge in biology



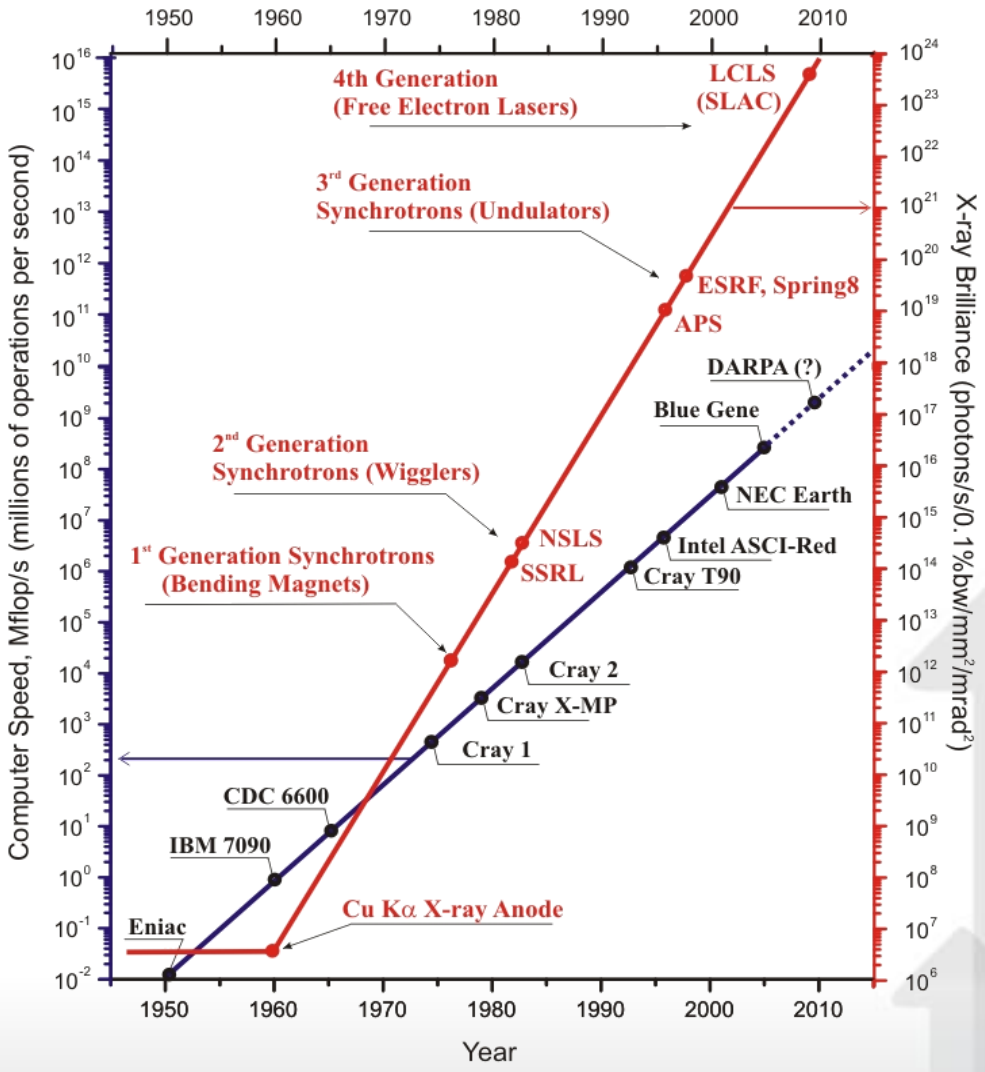
Number of sequencing machines



http://omicsmaps.com/

Moore's Law for X-ray sources

12 orders of magnitude in 6 decades



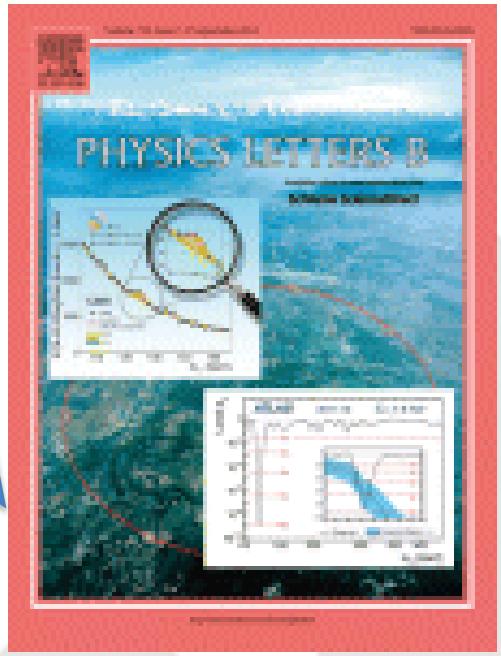
18 orders of magnitude in 5 decades!

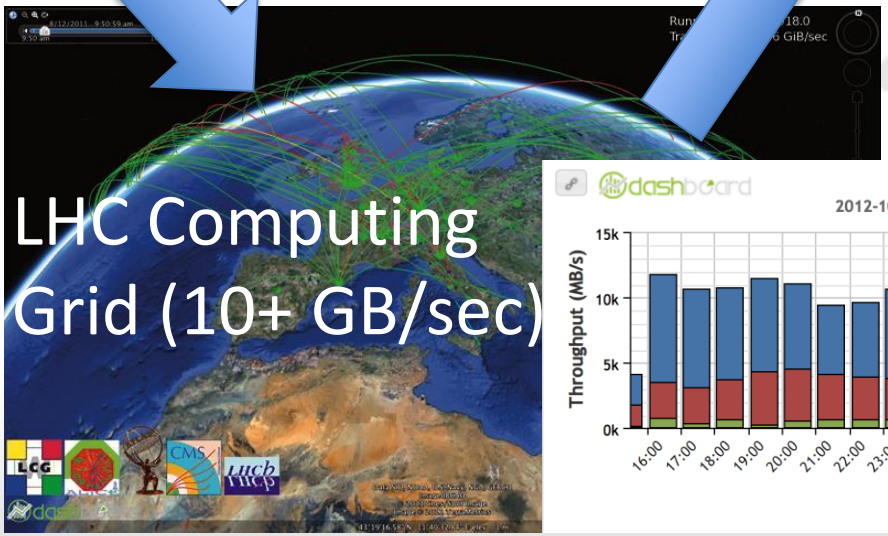
Credit: Linda Young

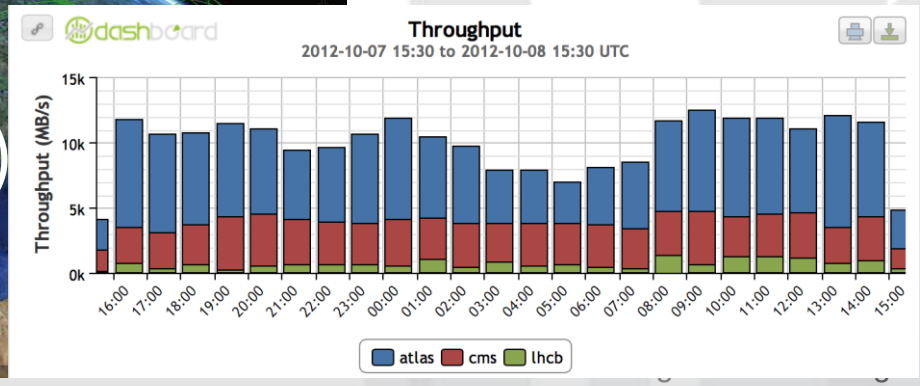
Big science has been successful

15 PB/year
173 TB/day
500 MB/sec

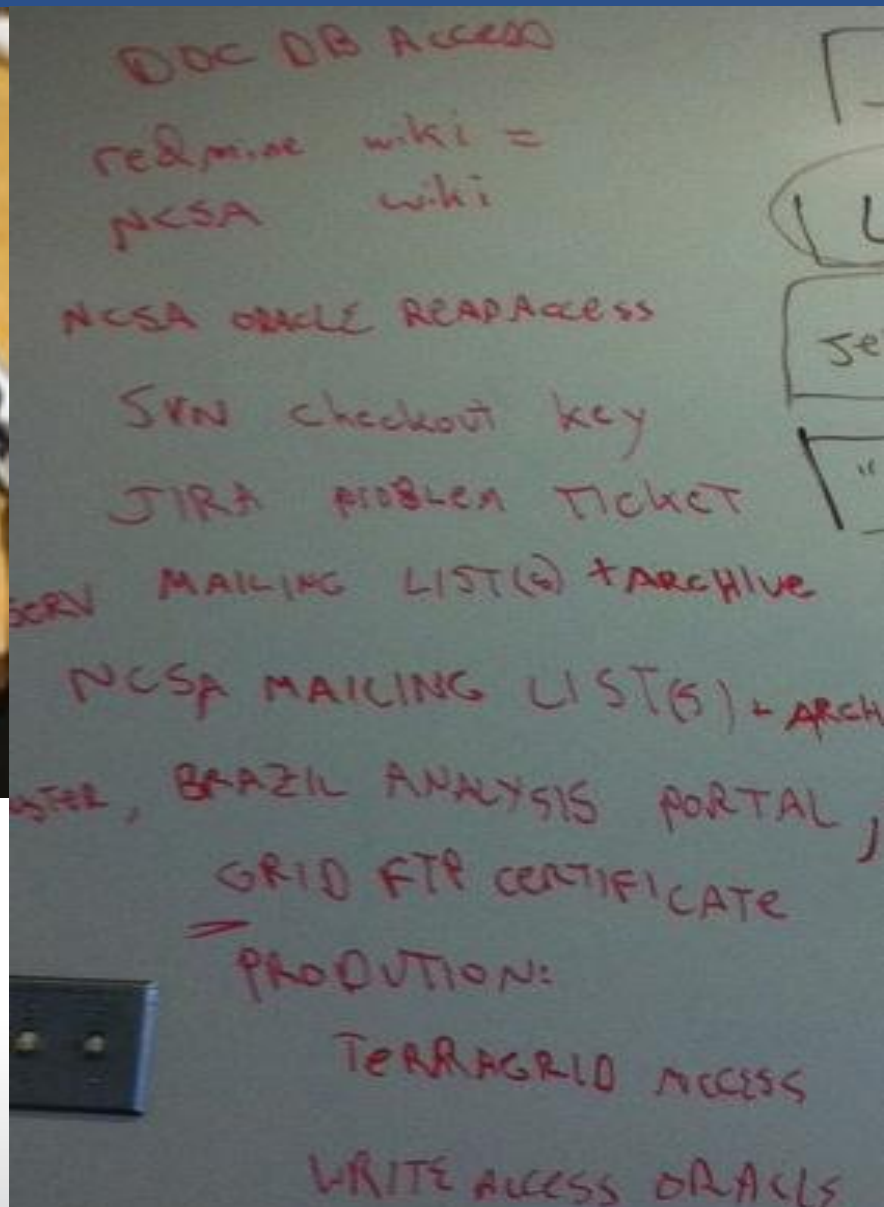
Globus GridFTP
300K files/day
0.3 PB/day
… just for LHC





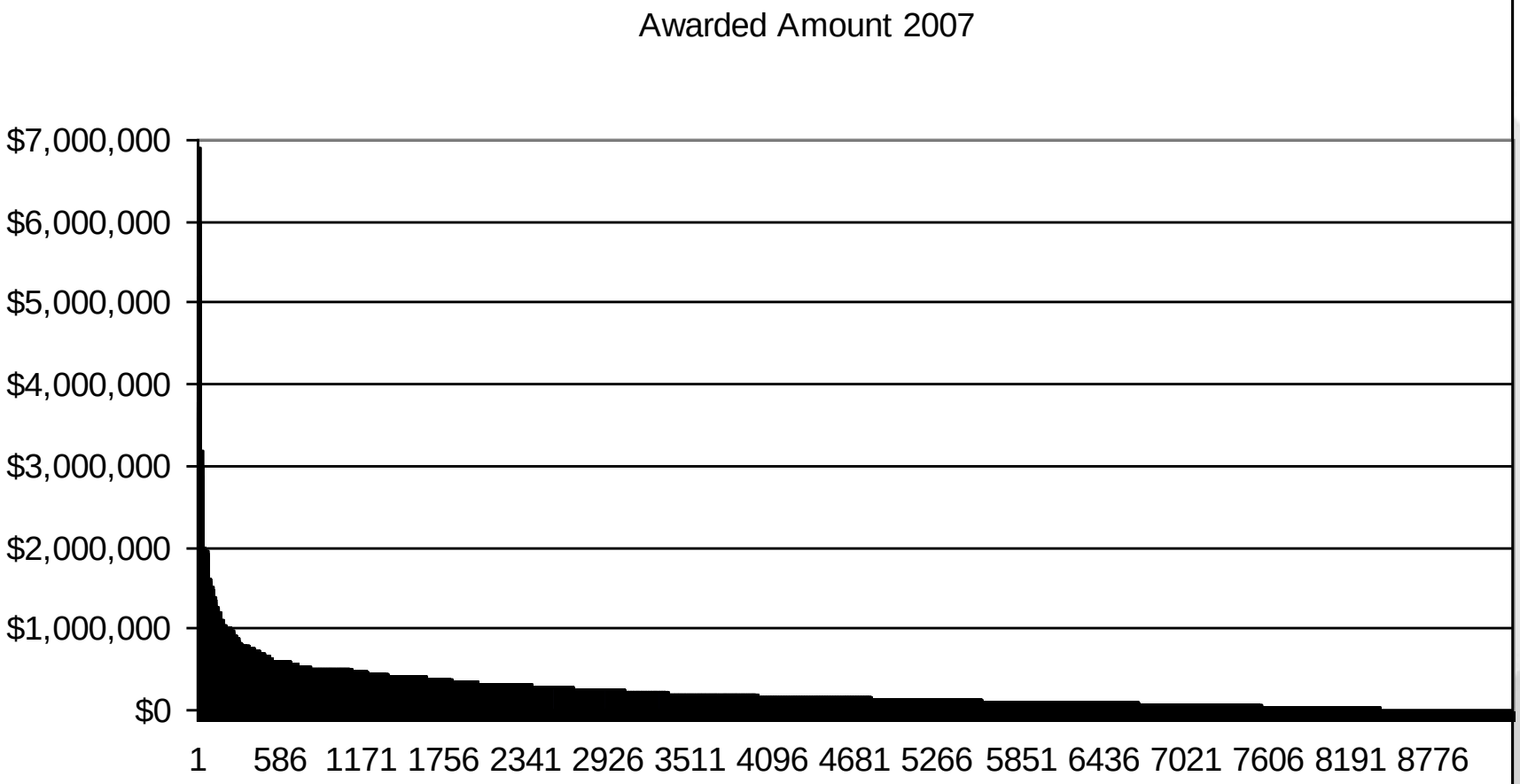


But small and medium science is suffering



- Data deluge
- Ad-hoc solutions
- Inadequate software, hardware & IT staff

Dark data in the long tail of science



NSF grant awards, 2007 (Bryan Heidorn)

A crisis that demands new approaches

- We have **exceptional infrastructure for the 1%** (e.g., supercomputers, Large Hadron Collider, …)
- But **not for the 99%** (e.g., the vast majority of the 1.8M publicly funded researchers in the EU)

We need new approaches to providing research cyberinfrastructure, that:

— Provide advanced functionality

— Reduce barriers to entry

— Are cheaper

— Are sustainable

Run a research project from a coffee shop?

Rethinking 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

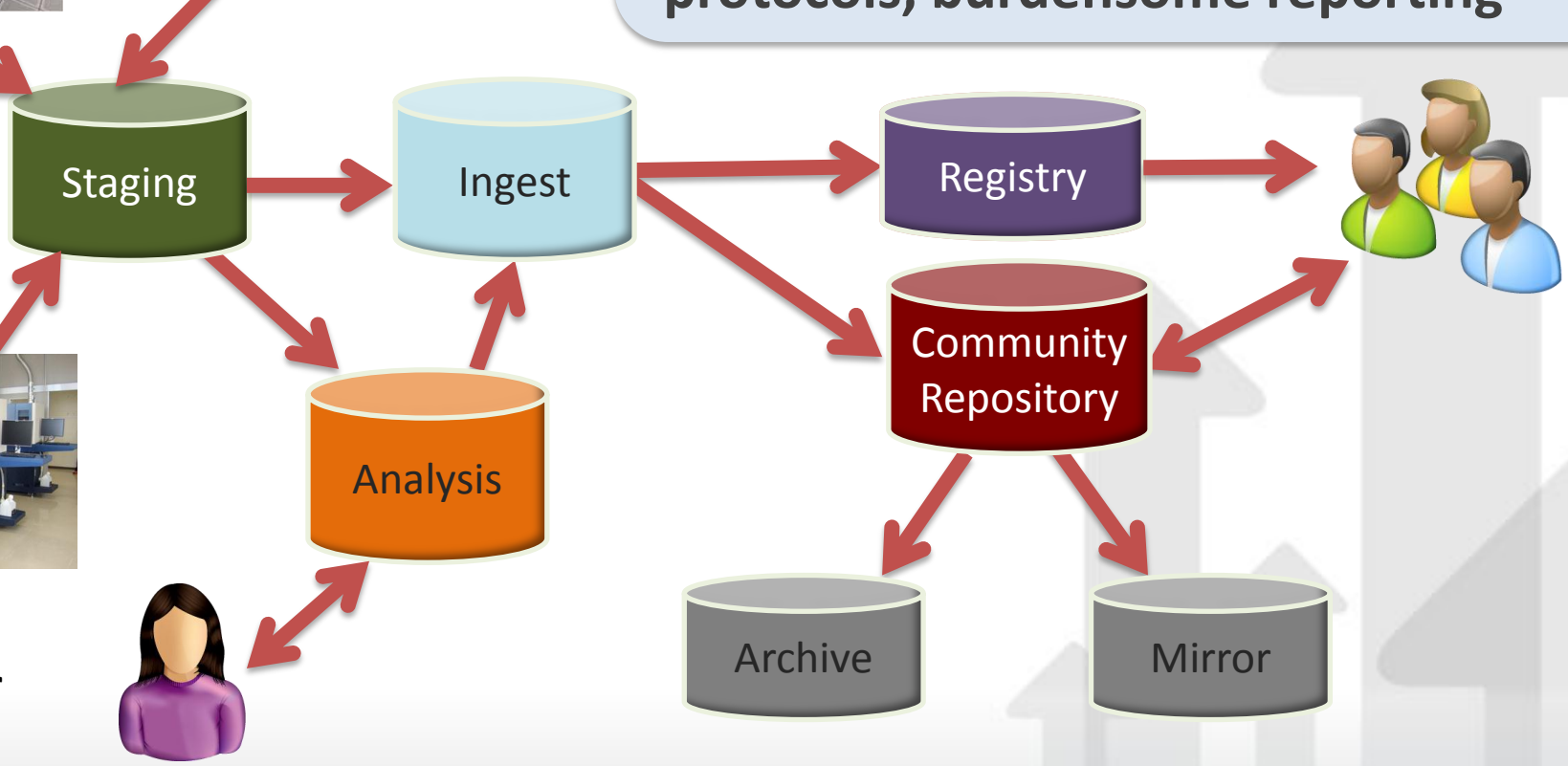
Time-consuming Tasks in Research

- Run experiments
- Collect data
- Manage data
- Run simulations
- Compare experiment with simulation
- Search the literature
- Communicate with colleagues
- Publish papers
- Order supplies
- Write proposals
- Write reports
- …

> *"Civilization advances by extending the number of important operations which we can perform without thinking of them"*
>
> —Alfred North Whitehead , 1911

Addressing the big data challenge

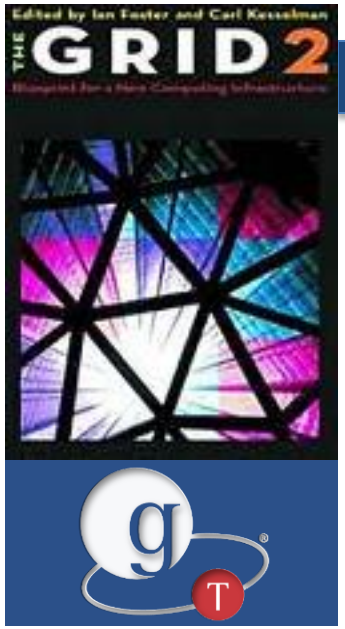
In millions of labs worldwide, researchers struggle with massive data, advanced software, complex protocols, burdensome reporting



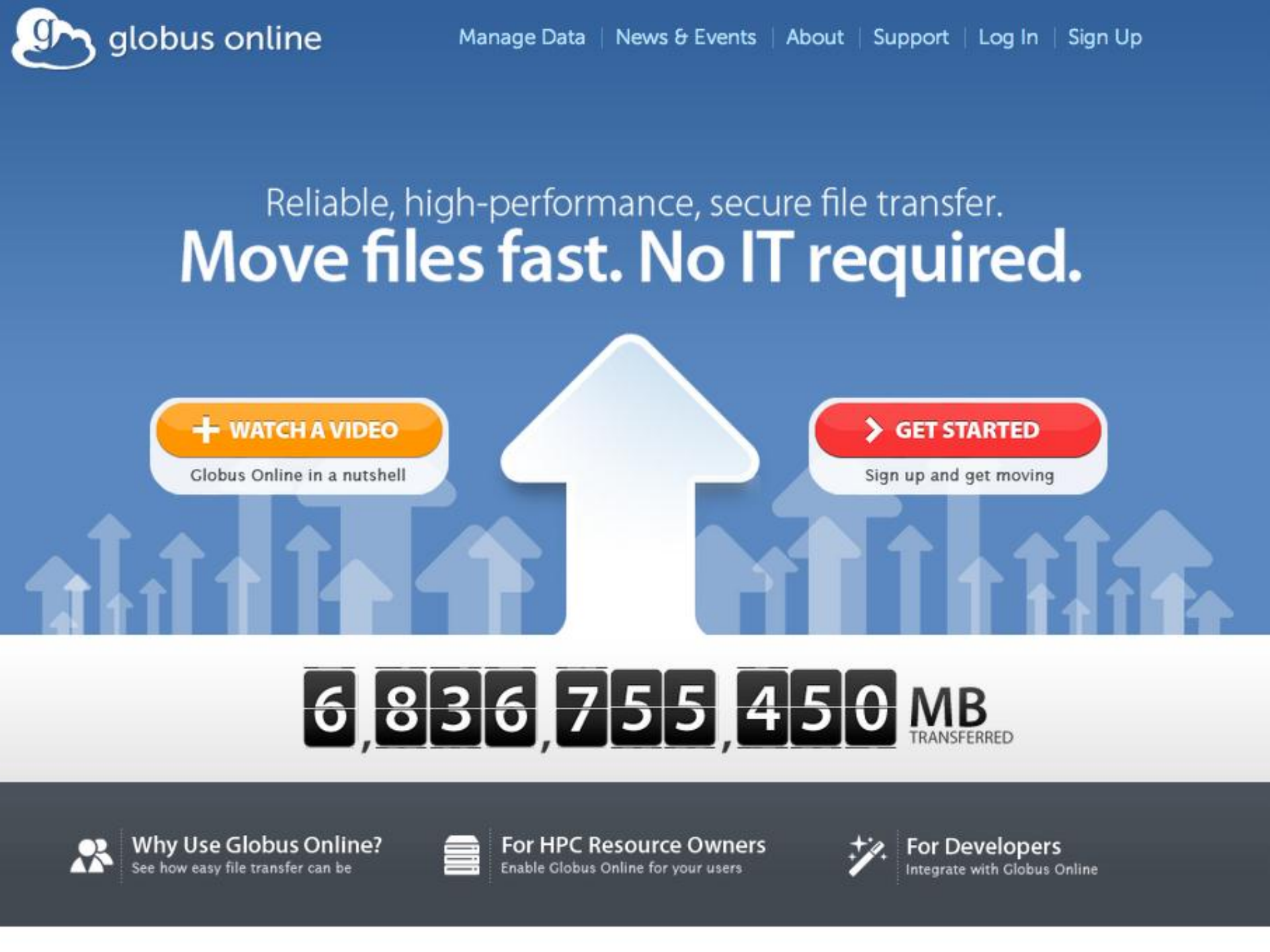
Accelerate discovery and innovation by outsourcing difficult tasks

Grid meets Cloud

Millions of researchers worldwide need advanced IT to tackle important and urgent problems

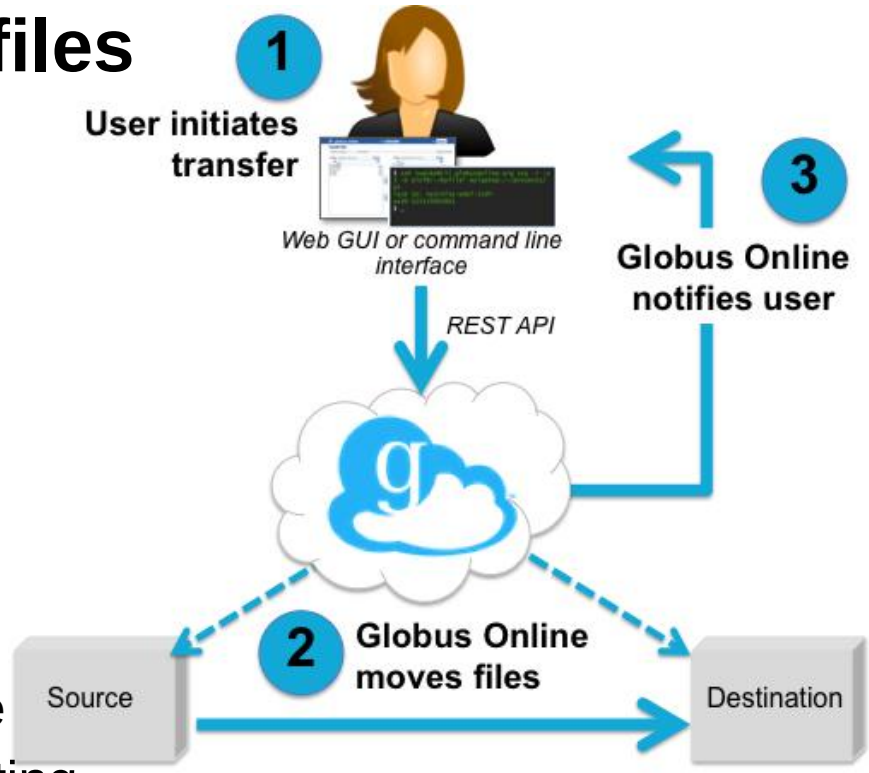


The Research Cloud


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
6,836,755,450 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

What is Globus Online?

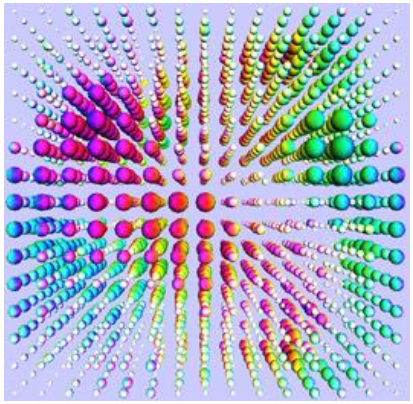
- **Transfer and synchronize files**
  - Easy “fire-and-forget” transfers
  - Automatic fault recovery
  - High performance
  - Across multiple security domains
- **Minimize IT costs**
  - Software as a Service (SaaS)
    - No client software installation
    - New features automatically available
  - Consolidated support & troubleshooting
  - Simple endpoint installation with Globus Connect and GridFTP
- **Recommended by XSEDE, Blue Water, NERSC, ALCF, ESnet, many Universities**



Case Study: Lattice QCD

NICS

- **Fast**: Reduced transfer times
- **Easy**: Fire-and-forget transfers
  - Automated retry
  - No file pre-staging
  - No complex infrastructure
  - Convenient CLI and GUI interfaces

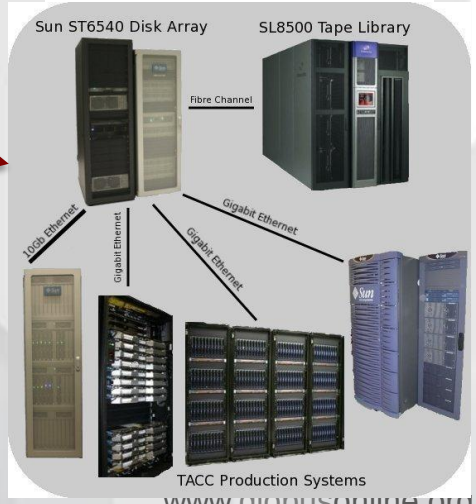


"Globus Online frees up my time to do more creative work than typing scp commands or devising scripts to initiate and monitor progress to move many files."

Indiana University researcher moved ~6 TB from Tennessee to Texas in 2 days

"I moved 100 7.3 GB files tonight in about 1.5 hours. I am very impressed. I also like the new commands and help system."

TACC

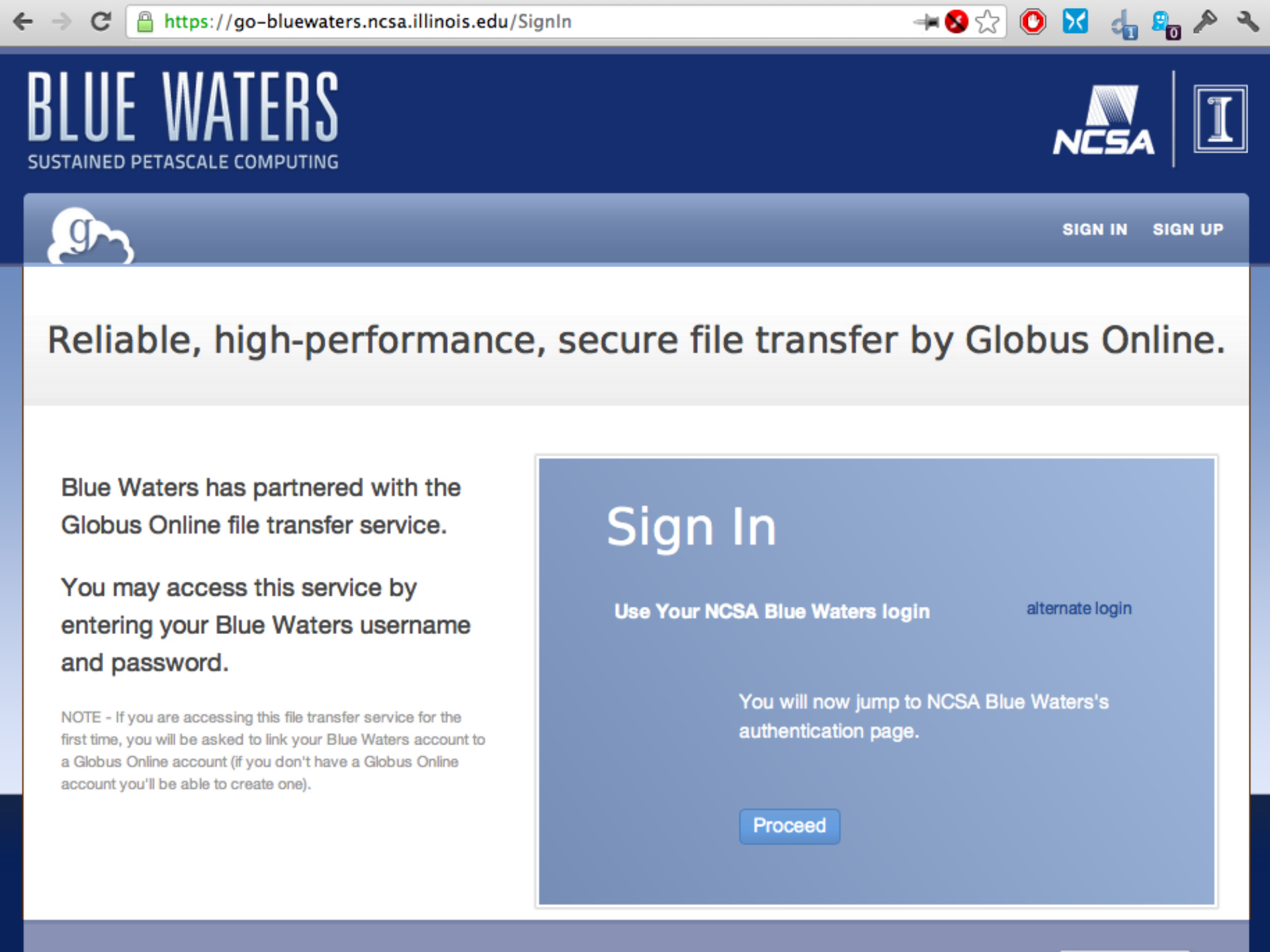


Case Study: Enabling Users @ NERSC

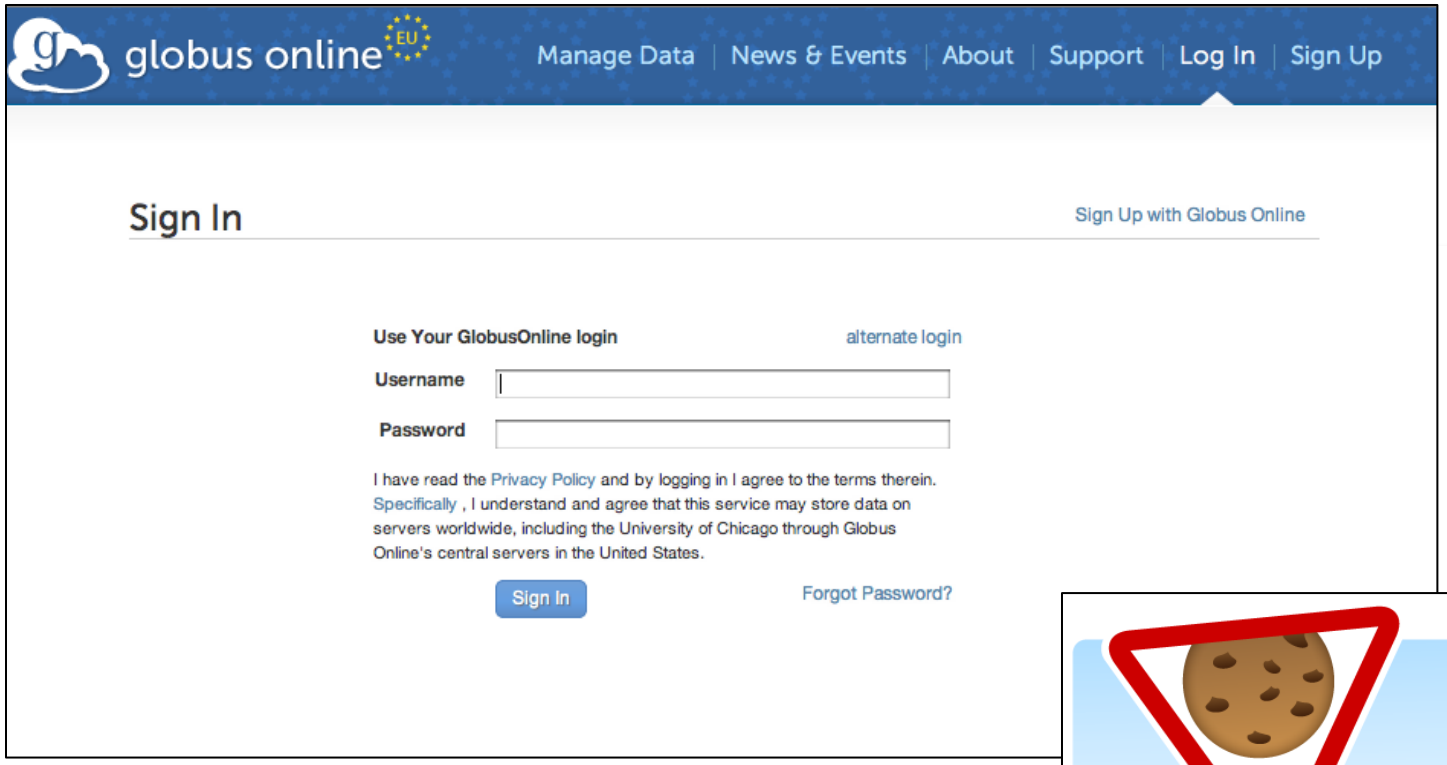
- **Challenge**
  - *"We need to provide web-based ways to accomplish computing tasks – it's what our scientists expect. And it will make them more productive."*
- **Solution**
  - Globus Online endpoints maintained by NERSC
  - Globus = recommended transfer method
- **Benefits for NERSC users**
  - Drag and drop archiving
  - Easy to use
  - Users can focus on their research (not on IT)
- **Benefits for NERSC**
  - Operations and support outsourced to Globus
  - Fast and easy to make endpoints available
  - Automated authentication
  - Reliable performance and support

"Fantastic! I have already started using Globus Connect to transfer data, and it only took me 5 minutes to set up. Thank you!" – NERSC user

Hopper, Franklin and HPSS are among the NERSC resources leveraged by Globus Online.


https://go-bluewaters.ncsa.illinois.edu/SignIn
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
You will now jump to NCSA Blue Waters's authentication page.
Proceed

globusonline.eu



Partners

Initiative for Globus in Europe

EGI

- EU-friendly identity providers
- EU relevant endpoints
- Integrated support with EGI & IGE
- EU Cookie Law compliant
- Consent to store log data in US

Why SaaS?

Software-as-a-Service (SaaS)

Platform-as-a-Service (PaaS)

Infrastructure-as-a-Service (IaaS)

- **Deliver advanced functionality that:**
  - Requires no user software installation or operation
    - Minimal IT proficiency required
  - Can be cheaply and incrementally adopted
    - Usage-based subscription pricing; no big up-front costs
  - Consolidates troubleshooting and support
    - An expert group can proactively detect and correct problems
  - Utilizes an efficient software delivery lifecycle
    - Updates developed, tested and deployed quickly
- **Dominates commercial & consumer markets**
  - What about the research market?

Software as a Service (SaaS) vs. traditional software delivery

- **SaaS changes assumptions and approach throughout the software lifecycle**
  - Architecture and design
    - Designed for specific environment (e.g., AWS)
  - Software development
    - No porting or cross platform testing. Focus on functionality.
  - Operations
    - Continuous update cycle. Nobody else will operate.
    - Focus on availability, automation, monitoring
  - Support
    - Tightly integrated with operations
- **We are delivering a *service*, not *software***

Potential economies of scale

Small laboratories

- PI, postdoc, technician, grad students
- Estimate 10,000 across US research community
- Average ill-spent/unmet need of 0.5 FTE/lab?

+ Medium-scale projects

- Multiple PIs, a few software engineers
- Estimate 1,000 across US research community
- Average ill-spent/unmet need of 3 FTE/project?

= Total 8,000 FTE: at ~$100K/FTE => $800M/yr

(If we could even find 8,000 skilled people)

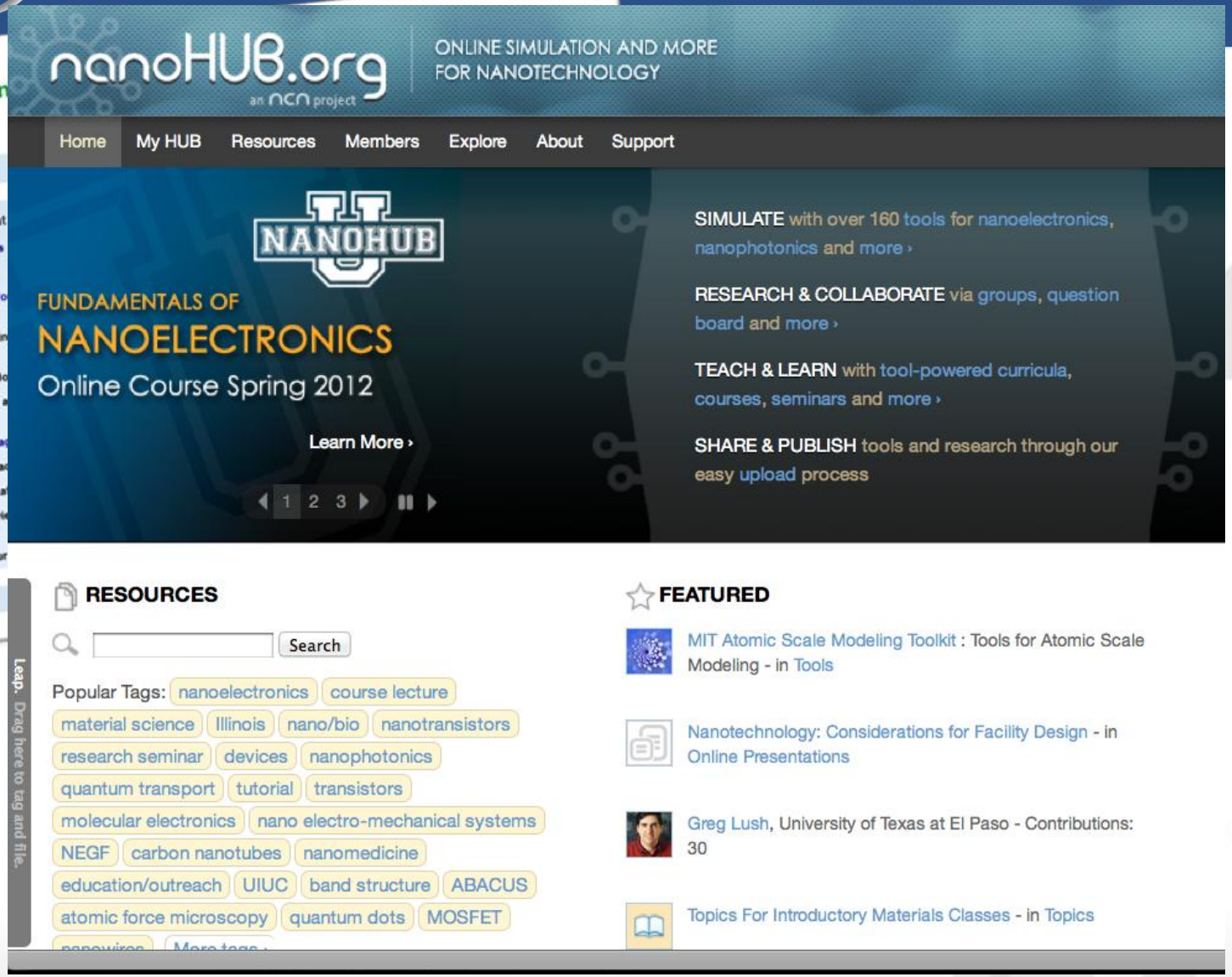
Plus computers, storage, opportunity costs, …

SaaS and reproducible publications

- **Hosted services can contribute to relieving author burden**
- **Store data, code, and workflows**
- **Maintain bindings among data, code, executions, workflows, provenance**
- **Etc.**

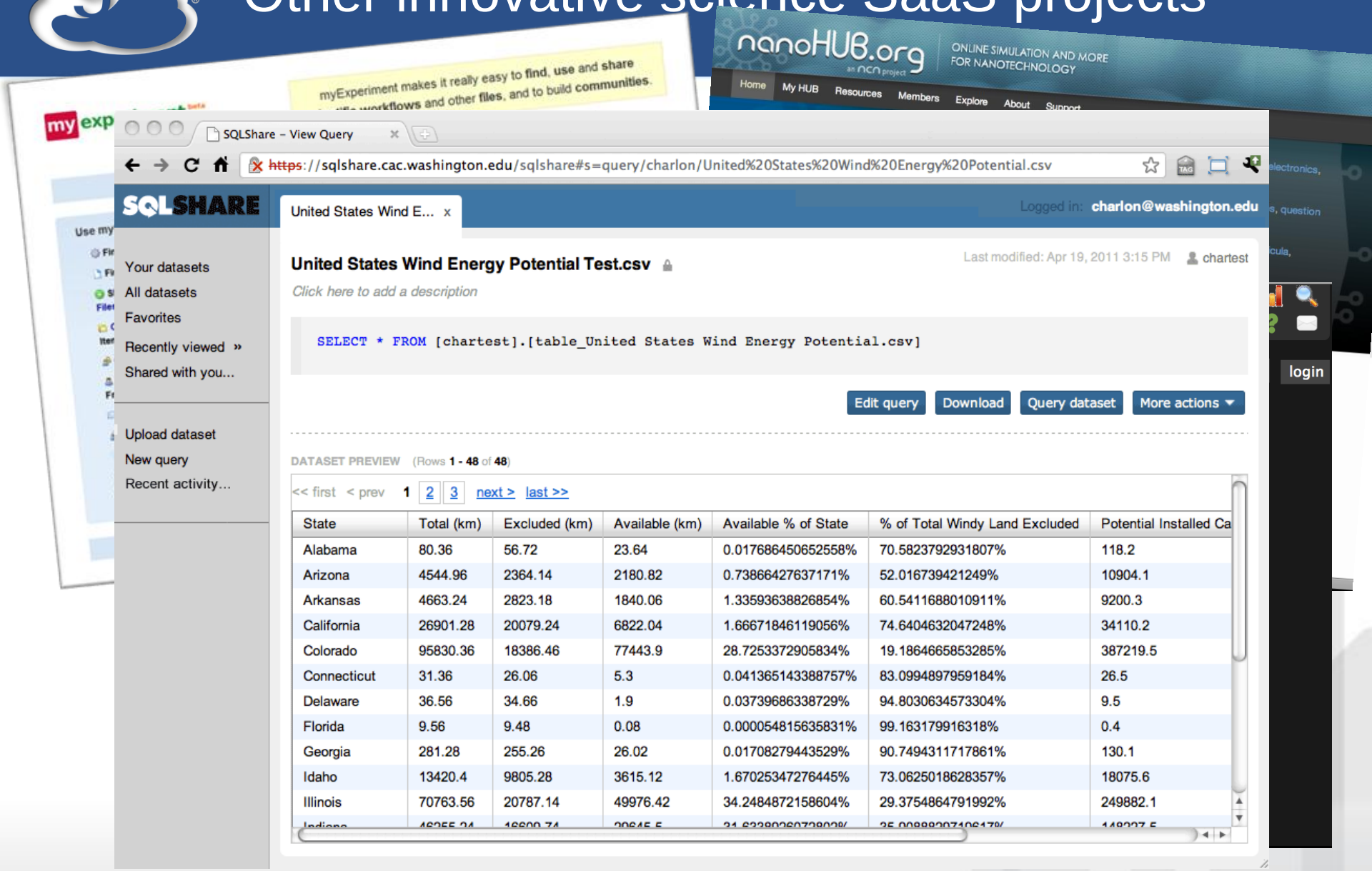
Other innovative science SaaS projects

Other innovative science SaaS projects



Other innovative science SaaS projects

Other innovative science SaaS projects

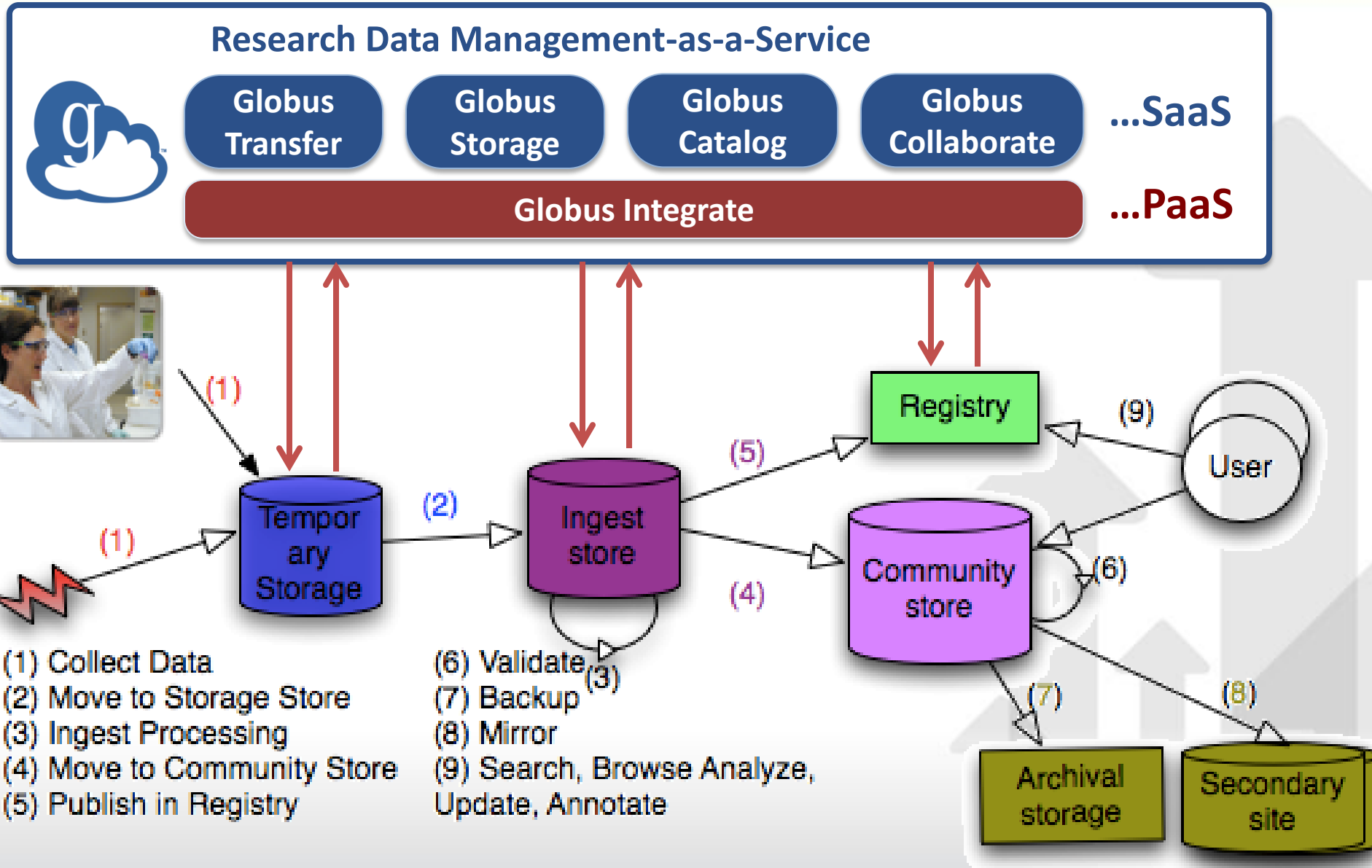


State	Total (km)	Excluded (km)	Available (km)	Available % of State	% of Total Windy Land Excluded	Potential Installed Ca
Alabama	80.36	56.72	23.64	0.017686450652558%	70.582379293180 7%	118.2
Arizona	4544.96	2364.14	2180.82	0.73866427637171%	52.016739421249%	10904.1
Arkansas	4663.24	2823.18	1840.06	1.33593638826854%	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

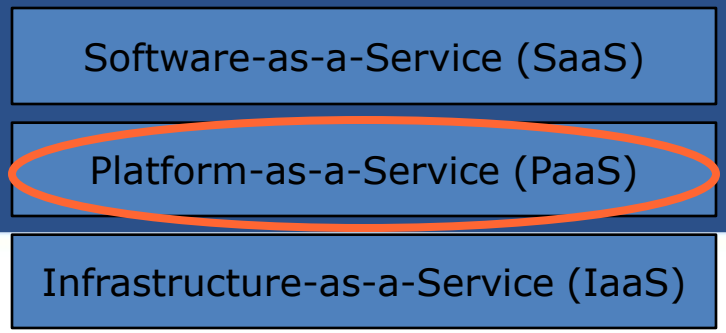
Other innovative science SaaS projects

Other innovative science SaaS projects

Towards “research IT as a service”



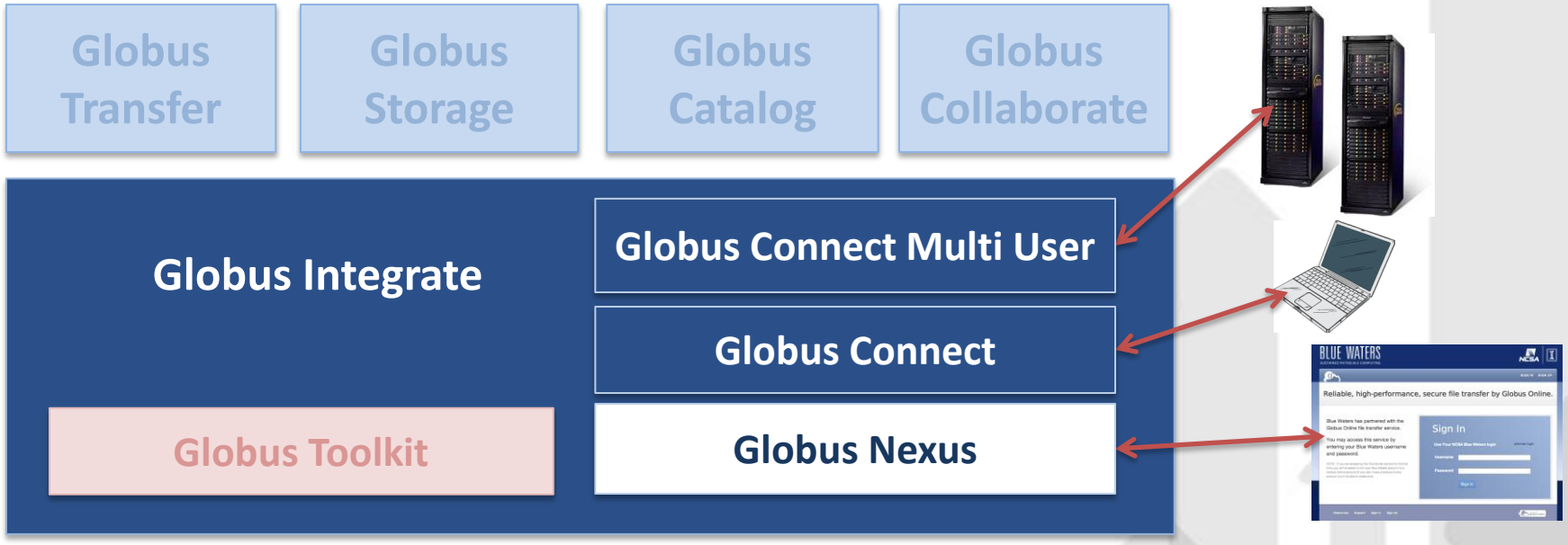
PaaS for Research



- **No one SaaS provider can deliver it all**
- **Must create ecosystem that:**
  - Allows any SaaS provider to easily participate
  - Dramatically reduces the cost of creating and operating services within the ecosystem
  - Provides seamless user experience across services
  - Agnostic to / works across any cloud IaaS provider
- **Ecosystem requires Platform as a Service**
  - Target the unique needs of the research community

Globus Integrate: For when you want to…

- **Integrate with the Globus research cloud ecosystem**
- **Write programs that leverage:**
  - (federated) user identities, profiles, groups
  - data, compute and collaboration



… via REST APIs and command line programs

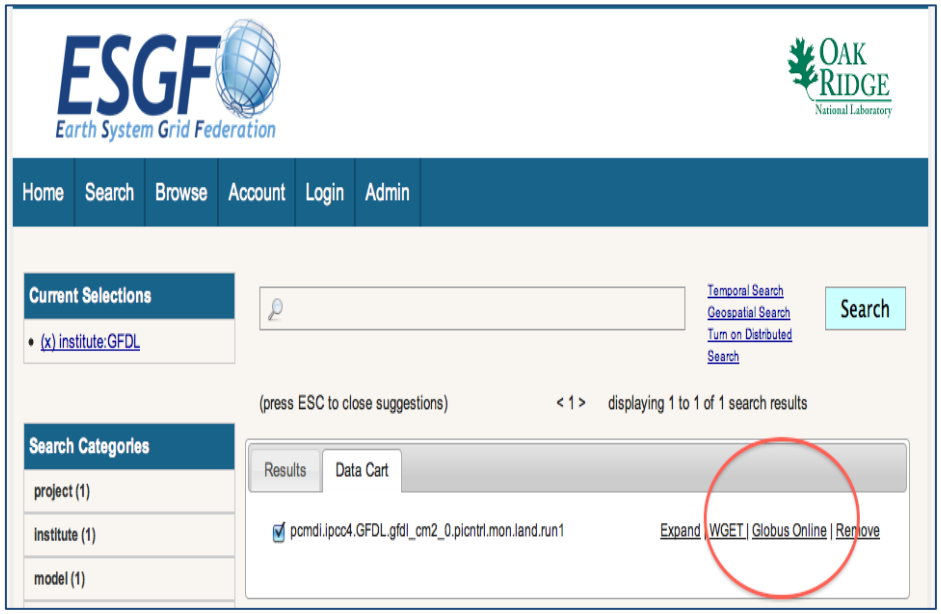
Dark Energy Survey – Leveraging transfer

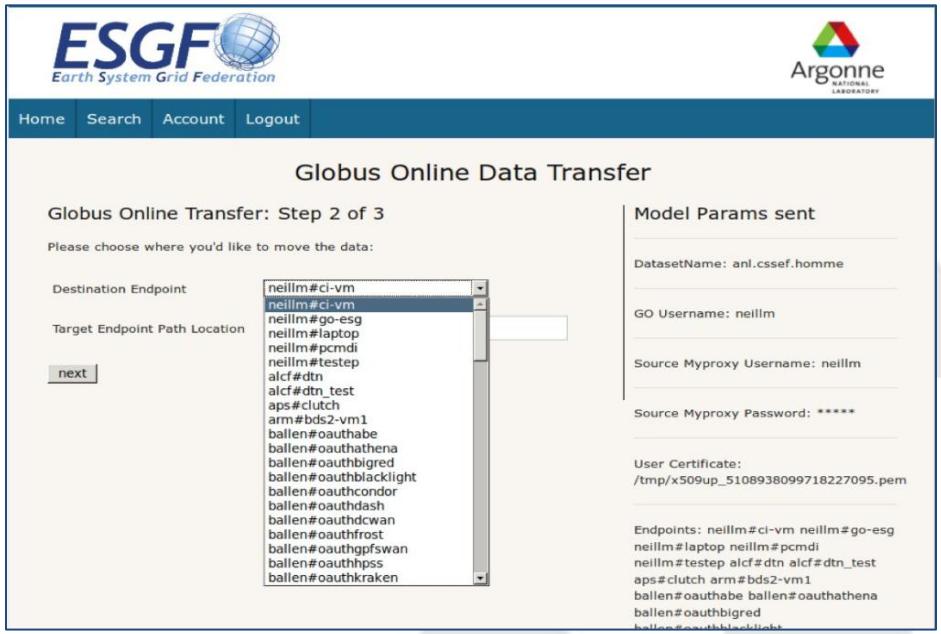
- **Every night, they receive 100,000 files in Illinois**
- **They transmit files to Texas for analysis … then move results back to Illinois … and make them available to users**
- **Process must be reliable, routine, and efficient**
- **The cyberinfrastructure team is not large!**

Blanco 4m on Cerro Tololo

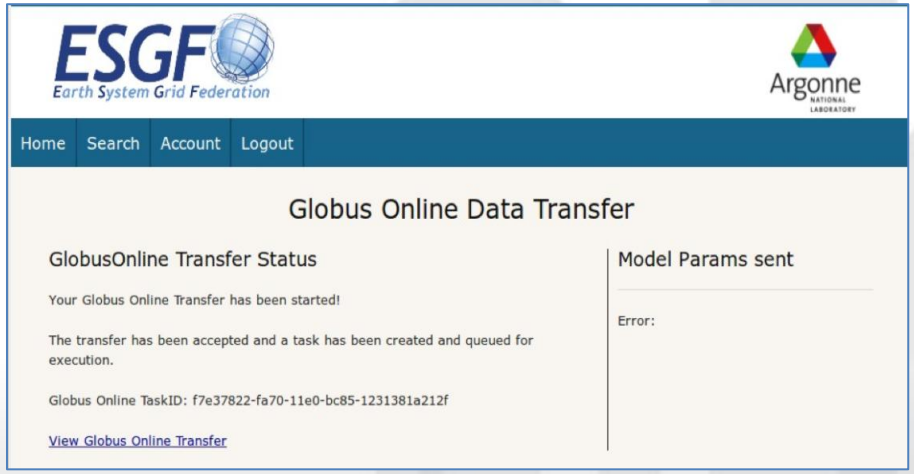
Image credit: Roger Smith/NOAO/AURA/NSF

Earth System Grid – Portal integration

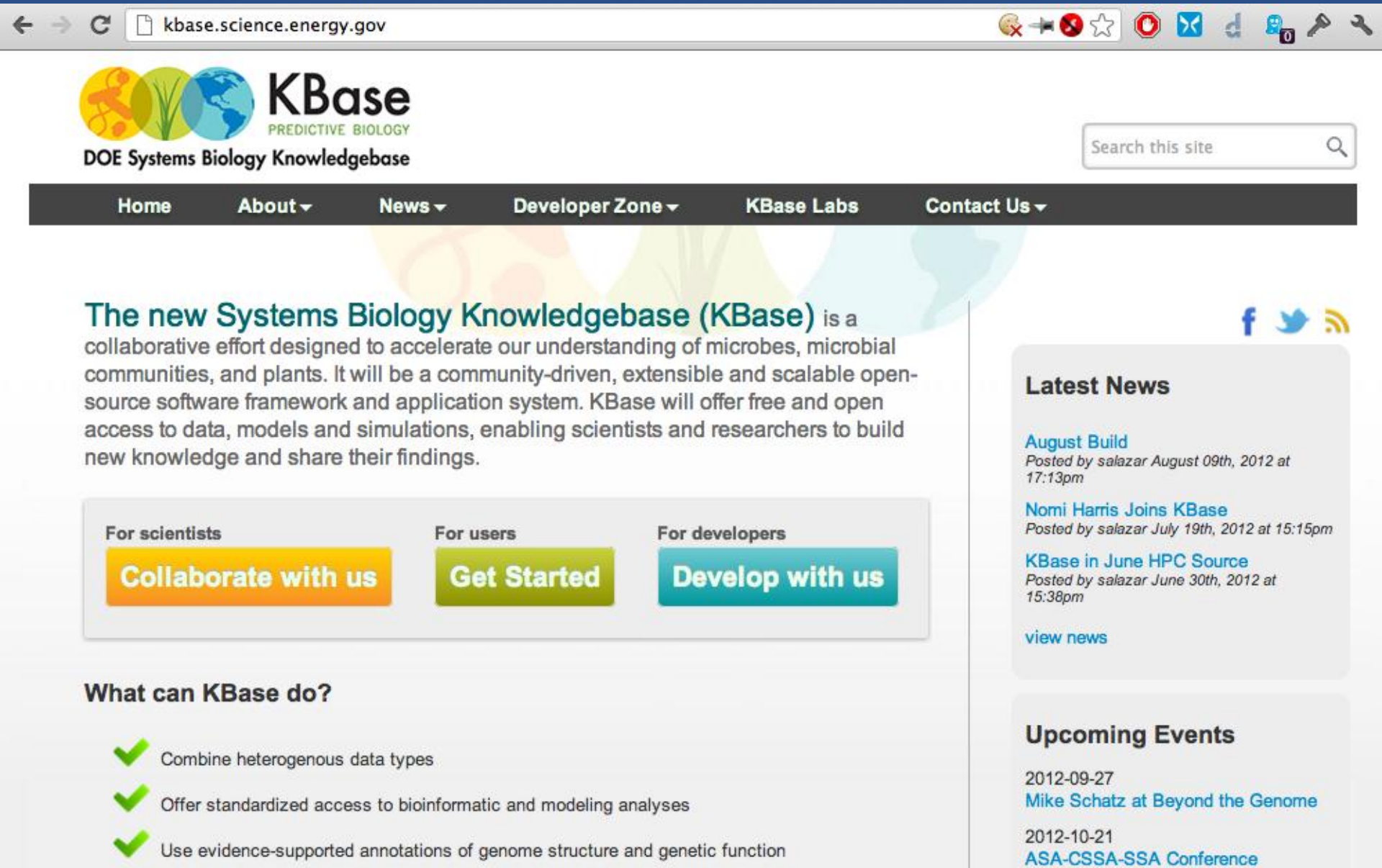




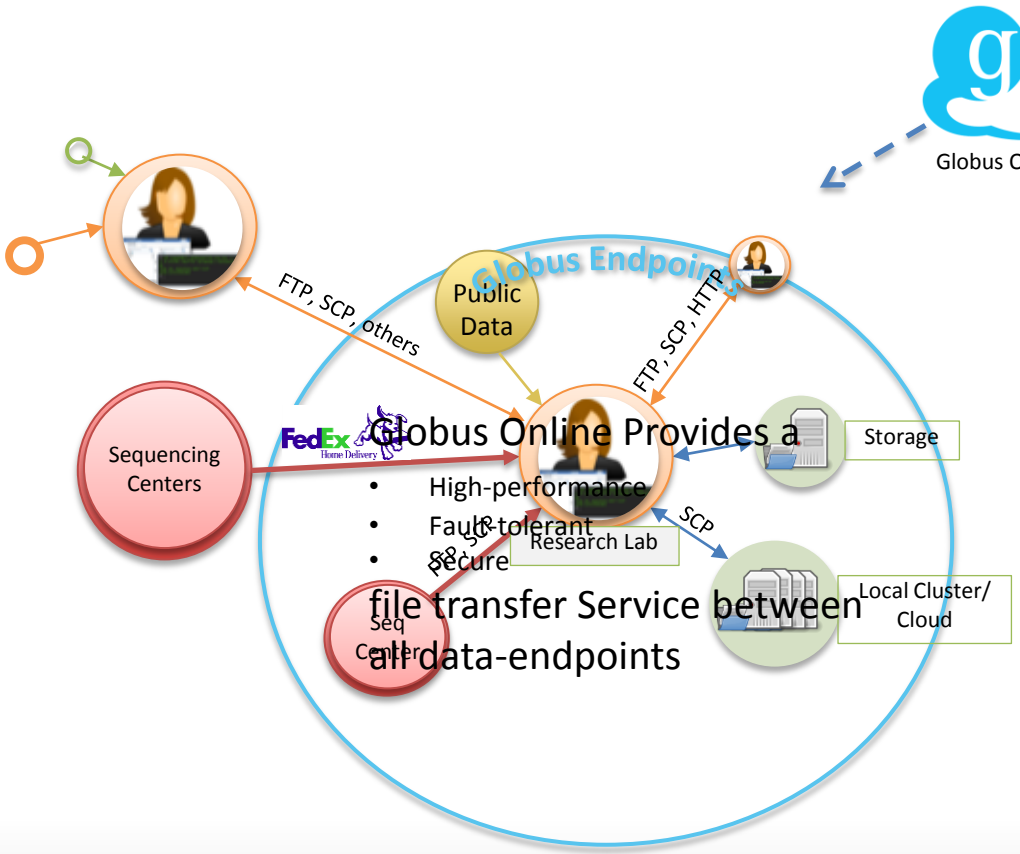
- Outsource data transfer to Globus
  - Data download to user machine from search
  - Data transfer to another server by user
  - Replication of data between sites by administrator
- No ESGF client software needed



KBase – Leveraging the full platform



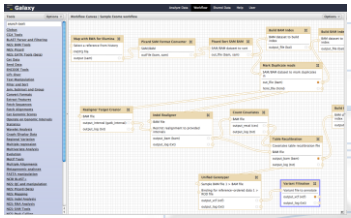
Genomics Analysis using Globus Online and Galaxy



Data Management

Galaxy Based Workflow Management System

Galaxy Data Libraries



- Globus Online Integrated within Galaxy
- Web-based UI
- Drag-Drop workflow creations
- Easily modify Workflows with new tools

- Analytical tools are automatically run on the scalable compute resources when possible

Data Analysis

Thank you!