Computation Institute

Science as a service

How on-demand computing can accelerate discovery

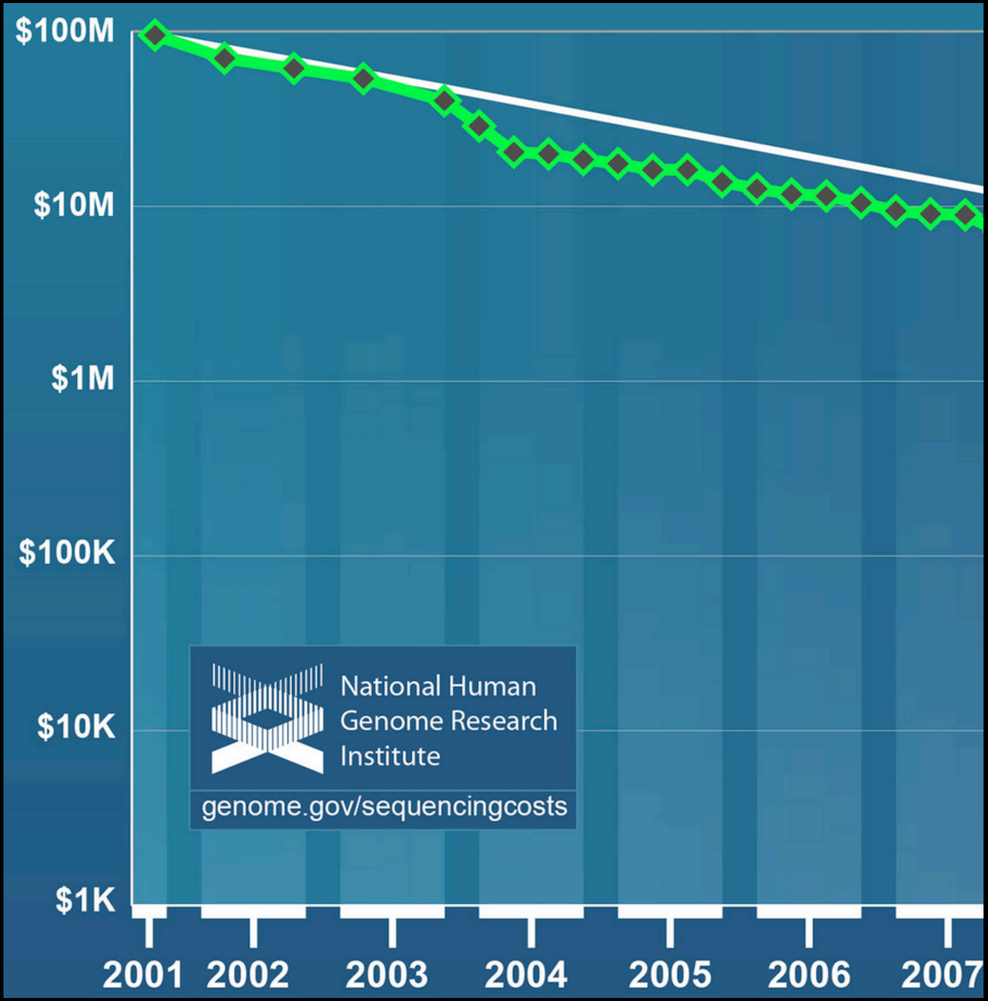
Ian Foster

foster@anl.gov

Argonne
NATIONAL LABORATORY

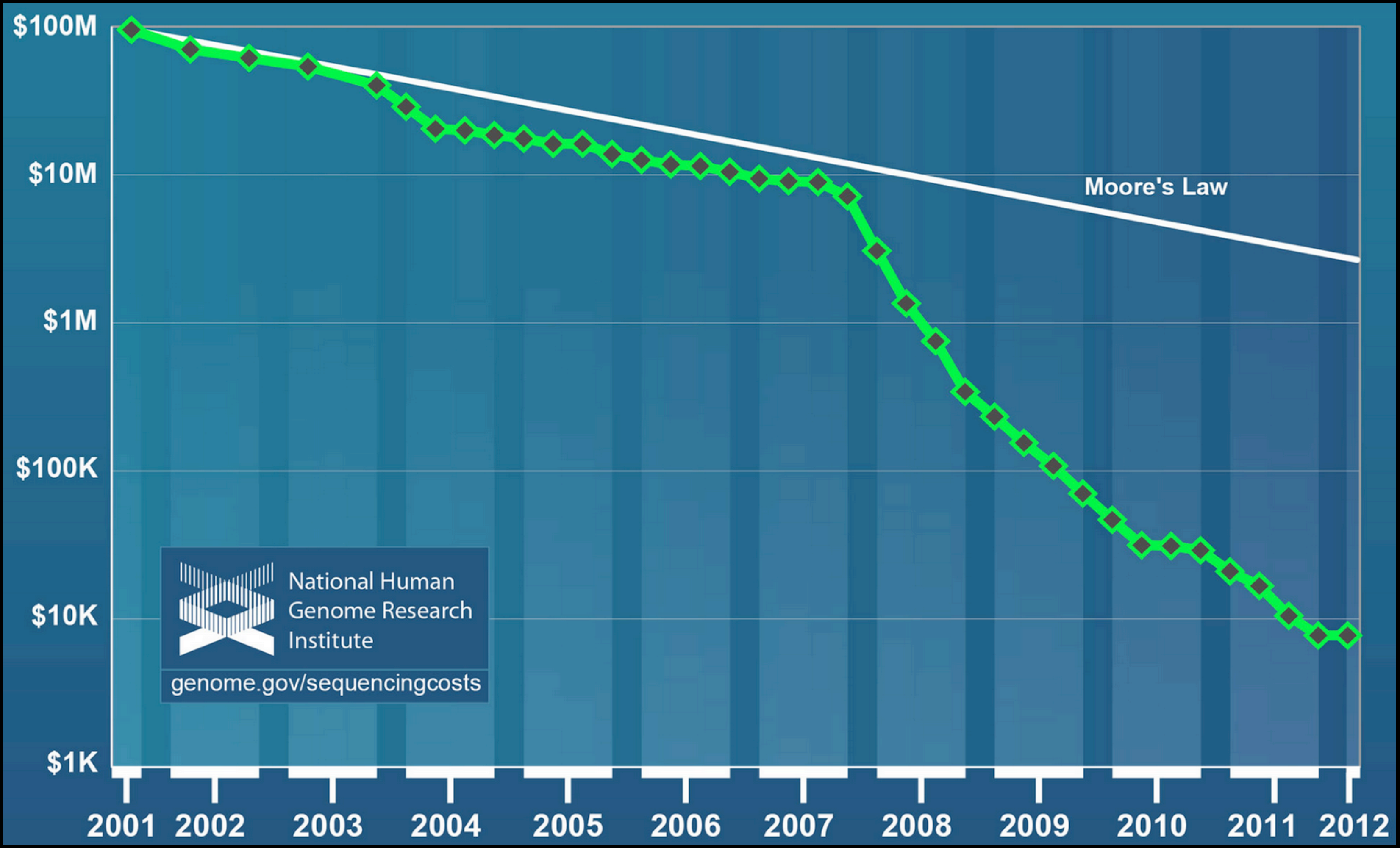
A time of disruptive change

As evidenced by cost per human genome

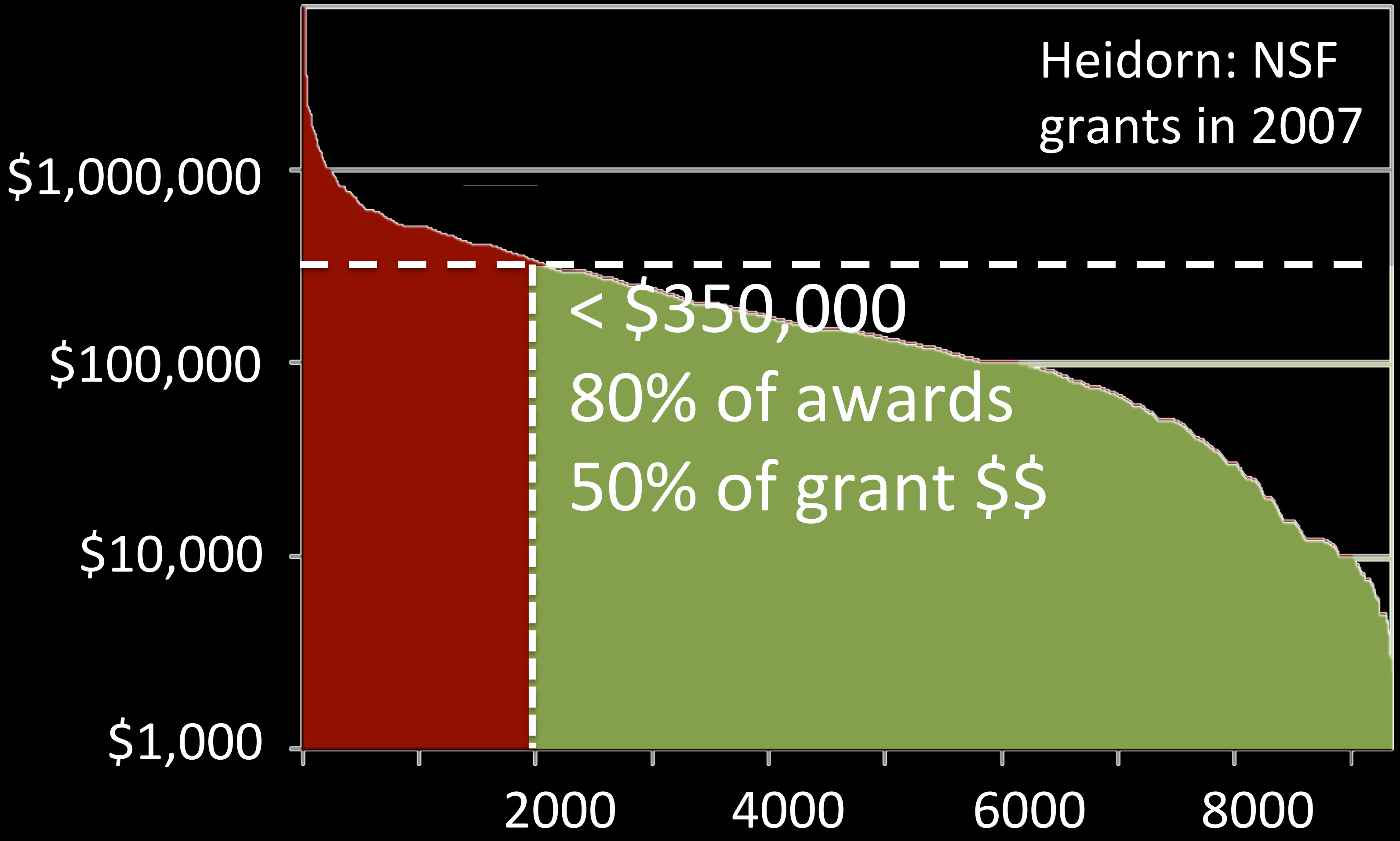


A time of disruptive change

As evidenced by cost per human genome



But most labs have extremely limited resources



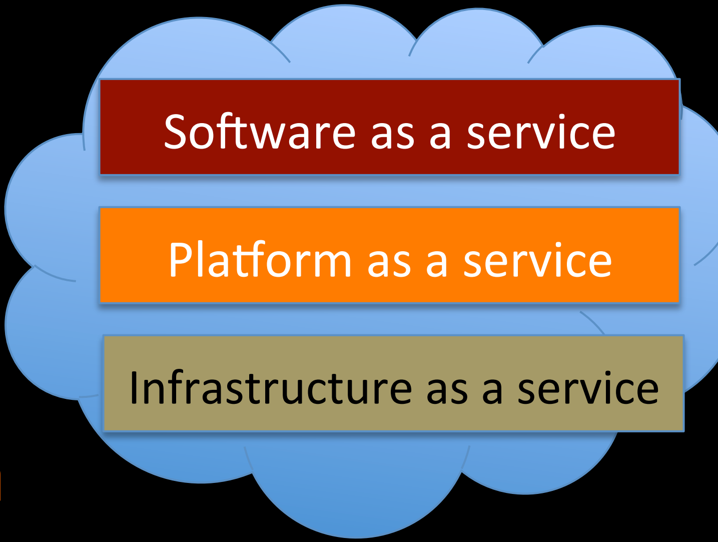
Automation is required to apply more sophisticated methods to far more data

Automation is required to apply more sophisticated methods to far more data

Outsourcing is needed to achieve economies of scale in the use of automated methods

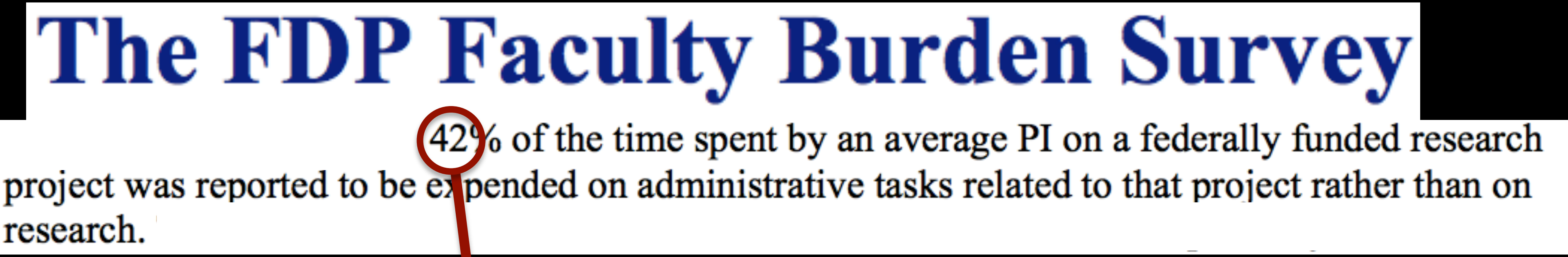
Building a discovery cloud

- Identify **time-consuming activities** that appear amenable to automation and outsourcing
- Implement as high-quality, low-touch **SaaS solution**
- Leverage **commercial IaaS** for reliability, economies of scale
- Extract common elements as a **research automation platform**



Bonus question: Identify methods for delivering Discovery Cloud elements sustainably

Where does time go in research?



The FDP Faculty Burden Survey

42% of the time spent by an average PI on a federally funded research project was reported to be expended on administrative tasks related to that project rather than on research.

42%!!

We aspire (initially) to create a great user experience for **research data management**

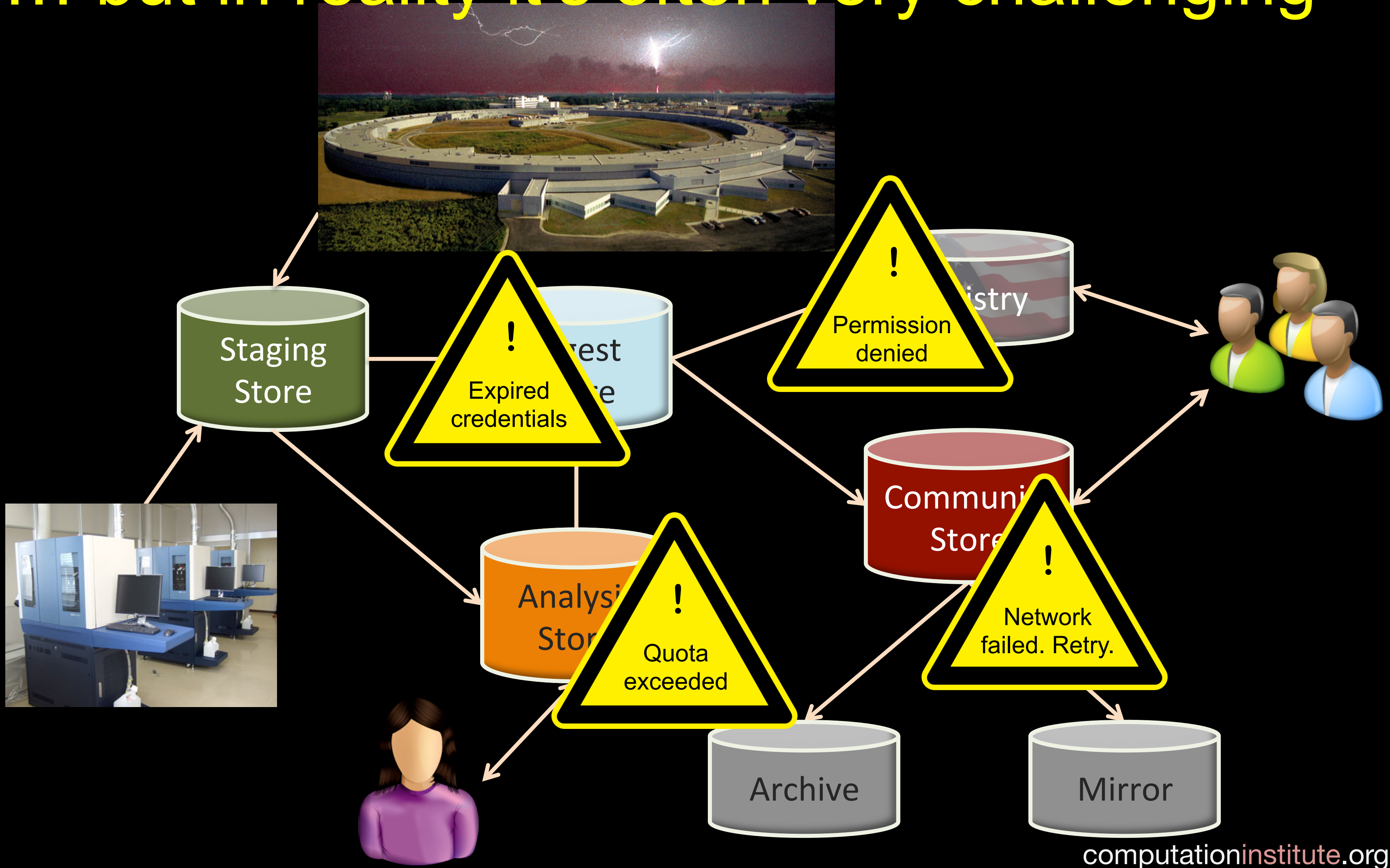
What would a “dropbox for science” look like?

- Collect
- Move
- Sync
- Share
- Analyze
- Annotate
- Publish
- Search
- Backup
- Archive

…for BIG DATA

It should be trivial to **Collect, Move, Sync, Share, Analyze, Annotate, Publish, Search, Backup, & Archive** BIG DATA

… but in reality it’s often very challenging



Staging Store

…est …e

…istry

Communi… Store

Analysi… Stor…

Archive

Mirror

Expired credentials

Permission denied

Quota exceeded

Network failed. Retry.

- Collect
- Move
- Sync
- Share
- Analyze
- Annotate
- Publish
- Search
- Backup
- Archive

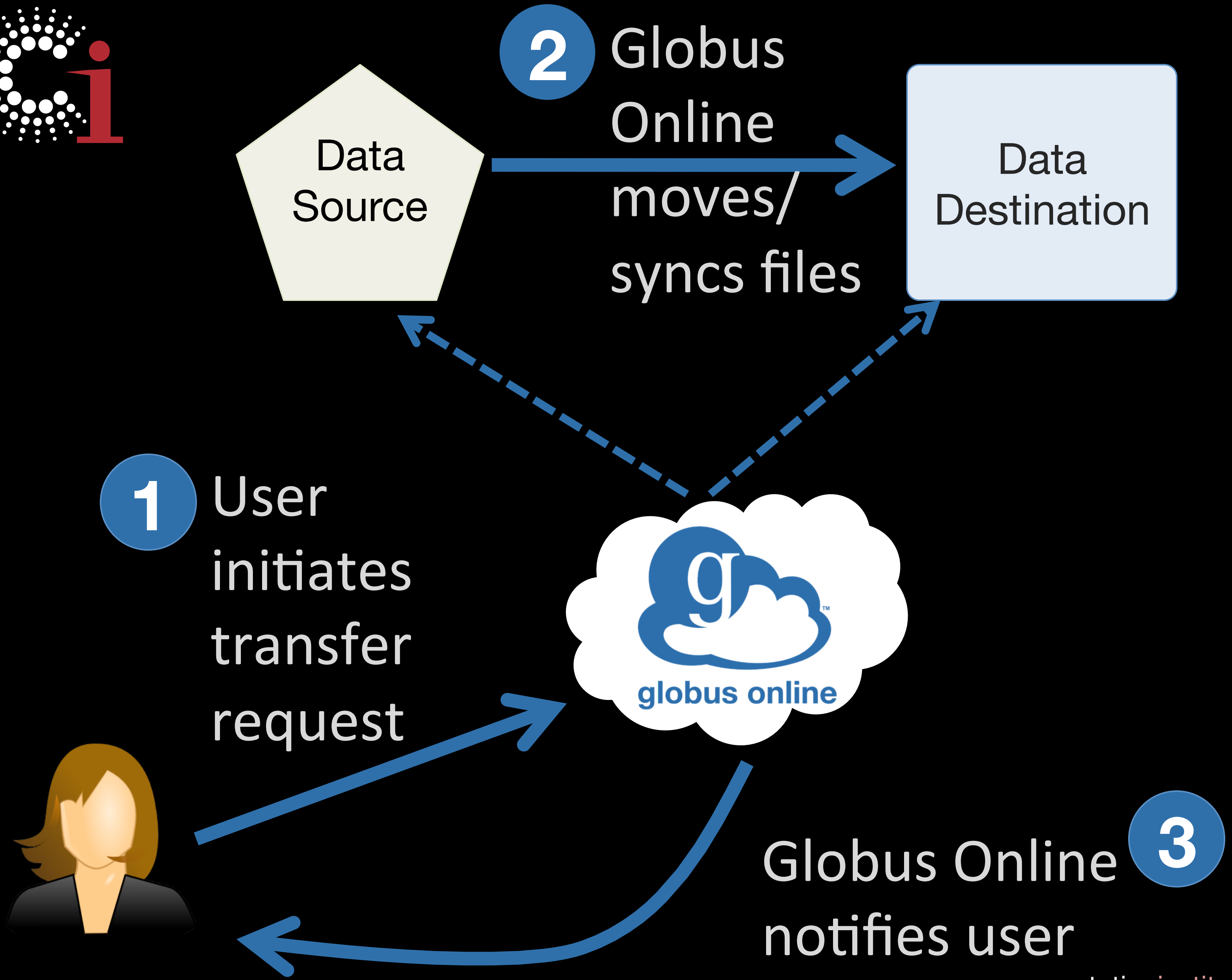
…for **BIG DATA**

- Collect
- Move
- Sync
- Share
- Analyze

- Annotate
- Publish
- Search
- Backup
- Archive

globus online

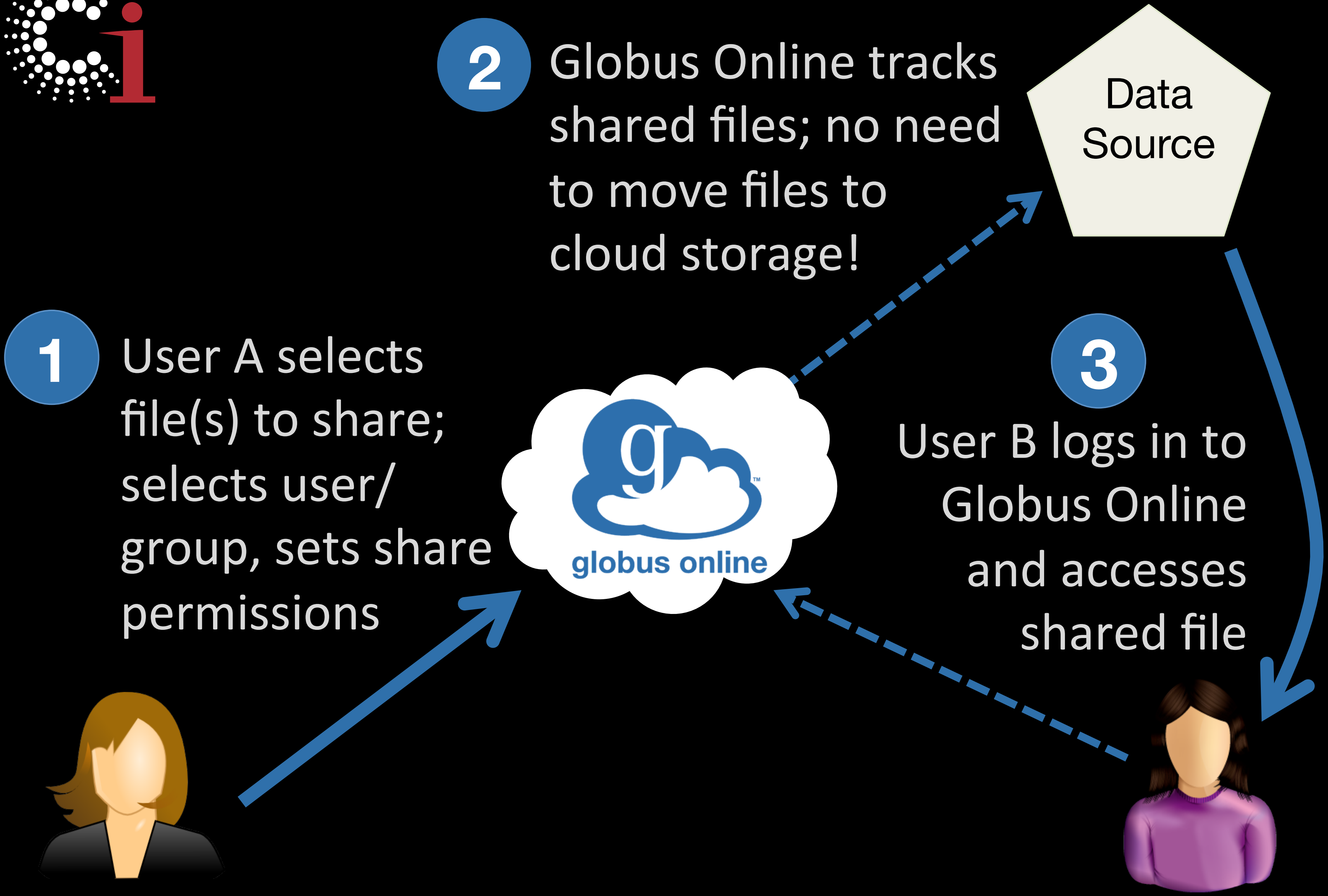
Capabilities delivered using Software-as-Service (SaaS) model



1 User initiates transfer request

2 Globus Online moves/ syncs files

3 Globus Online notifies user



1 User A selects file(s) to share; selects user/ group, sets share permissions

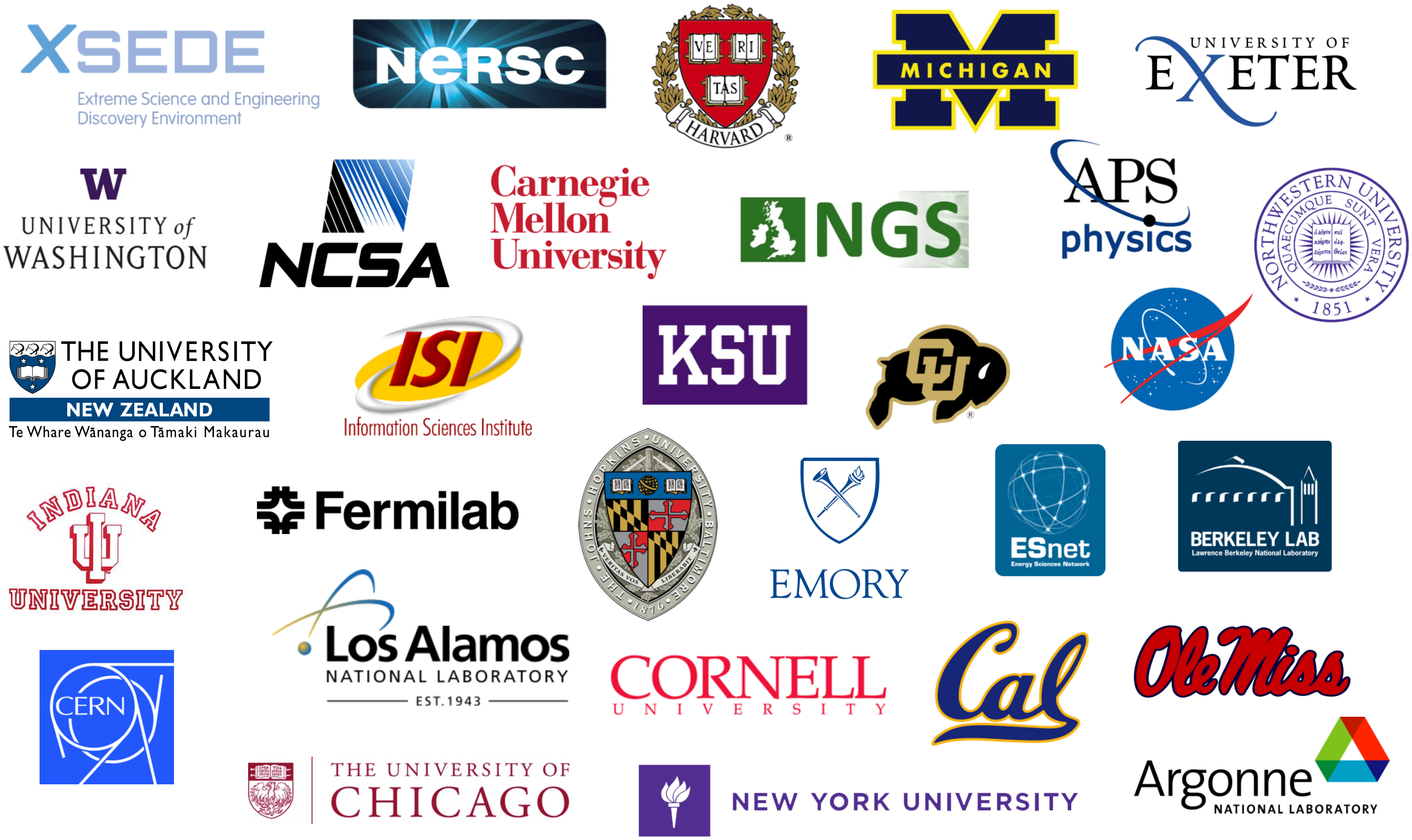
2 Globus Online tracks shared files; no need to move files to cloud storage!

3 User B logs in to Globus Online and accesses shared file

Extreme ease of use

- InCommon, Oauth, OpenID, X.509, …
- Credential management
- Group definition and management
- Transfer management and optimization
- Reliability via transfer retries
- Web interface, REST API, command line
- One-click "Globus Connect" install
- 5-minute Globus Connect Multi User install

Early adoption is encouraging



Early adoption is encouraging

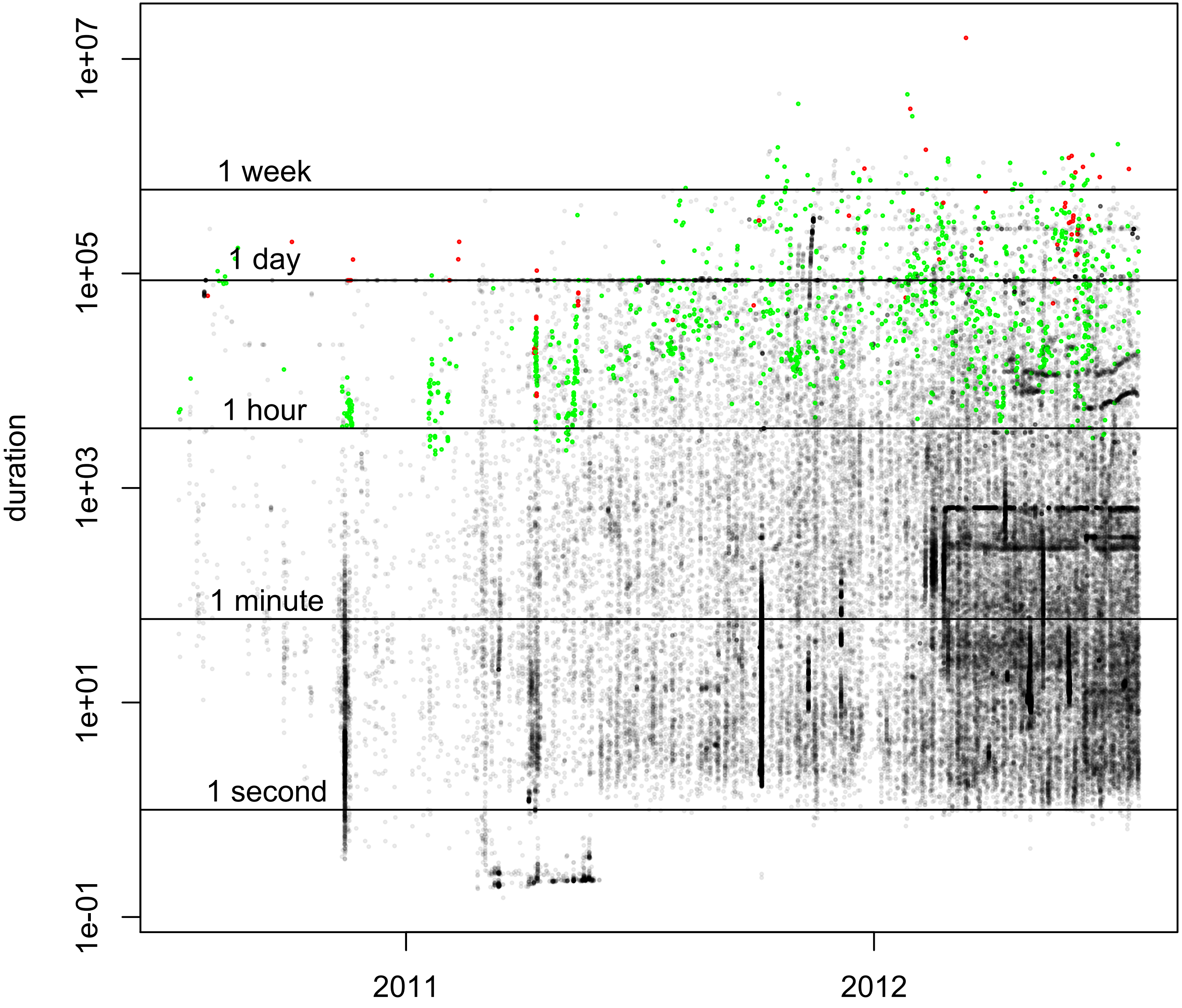
10,000 registered users; >100 daily

~18 PB moved; ~1B files

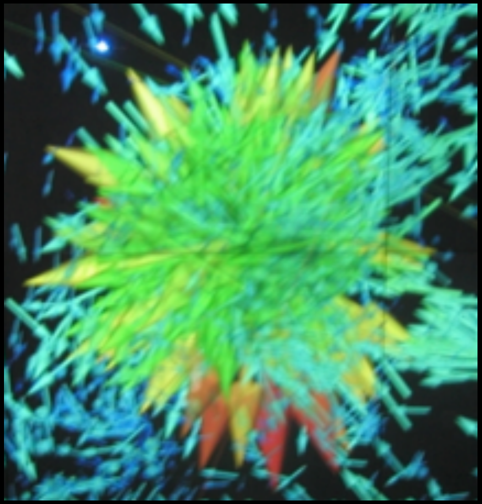
10x (or better) performance vs. scp

99.9% availability

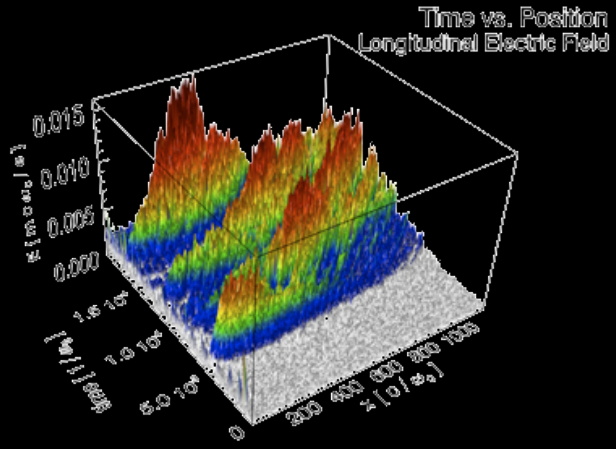
Entirely hosted on Amazon



Duration of runs, in seconds, over time.
Red: >10 TB transfer; green: >1 TB transfer.



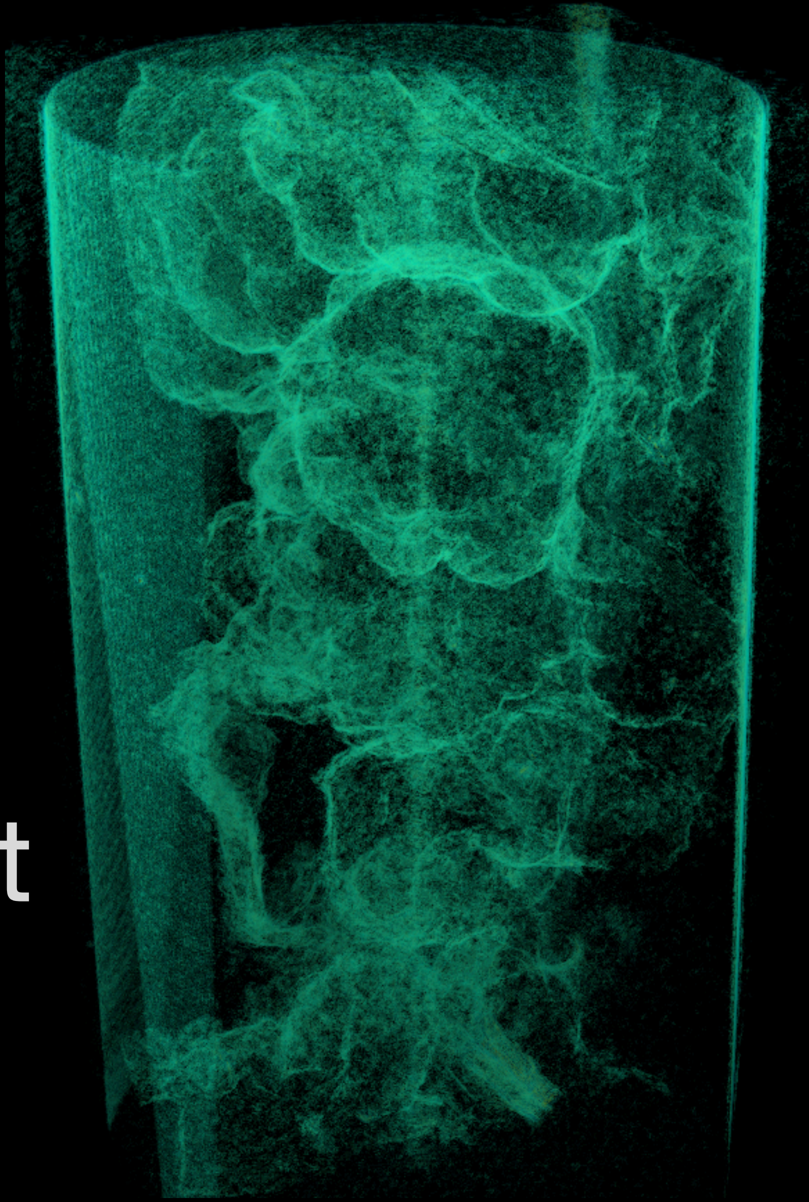
K. Heitmann (Argonne) moves 22 TB of **cosmology** data LANL → ANL at 5 Gb/s



B. Winjum (UCLA) moves 900K-file **plasma physics** datasets UCLA →NERSC

Dan Kozak (Caltech) replicates 1 PB LIGO **astronomy** data for resilience

Erin Miller (PNNL) collects data at Advanced Photon Source, renders at PNNL, and views at ANL



Credit: Kerstin Kleese-van Dam

- Collect
- Move
- Sync
- Share
- Analyze
- Annotate
- Publish
- Search
- Backup
- Archive

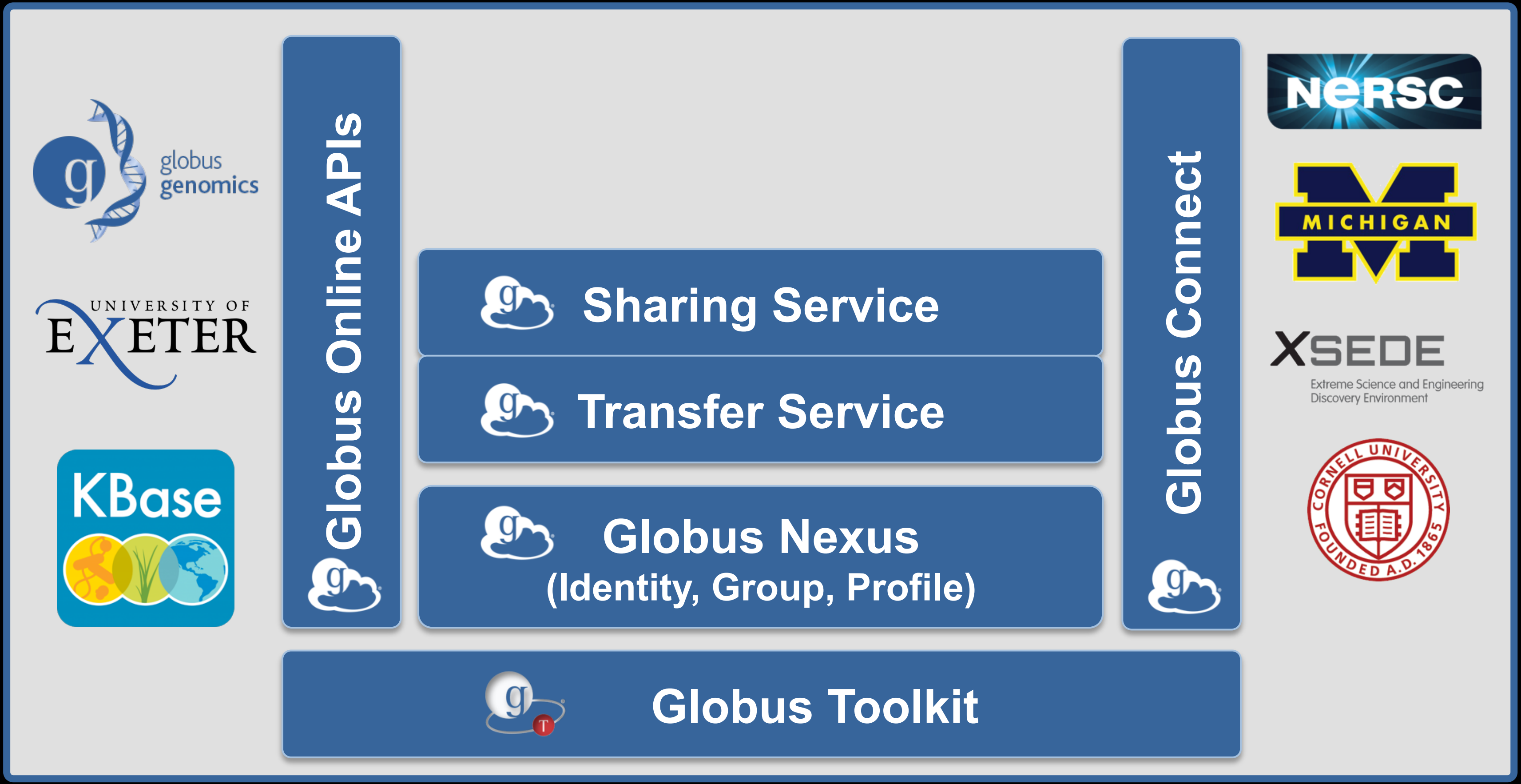
…*for* **BIG DATA**

computationinstitute.org

- Collect
- Move
- Sync
- Share
- Analyze
- Annotate
- Publish
- Search
- Backup
- Archive

…for **BIG DATA**

Globus Online already does a lot



Globus Online APIs

Sharing Service

Transfer Service

Globus Nexus
(Identity, Group, Profile)

Globus Connect

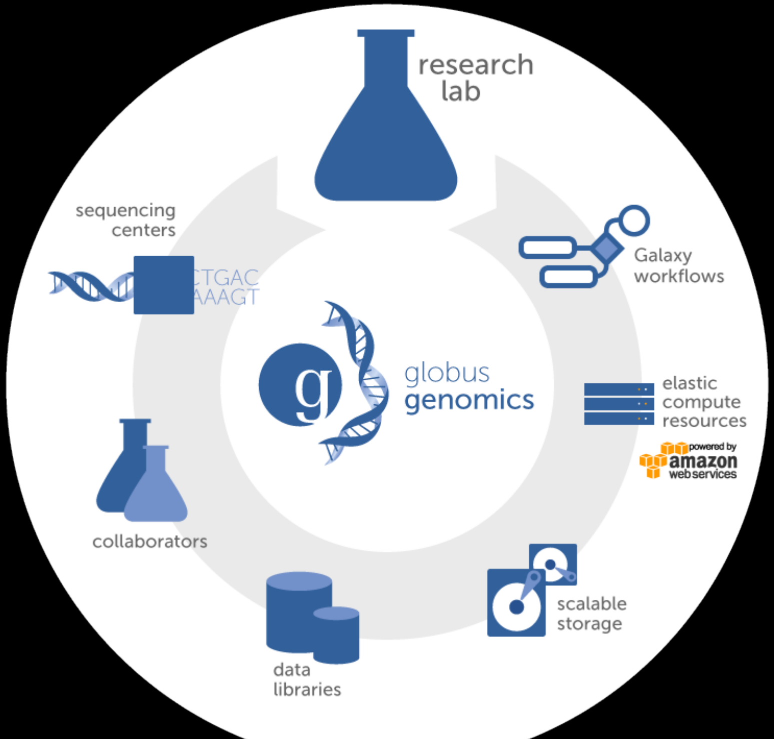
Globus Toolkit

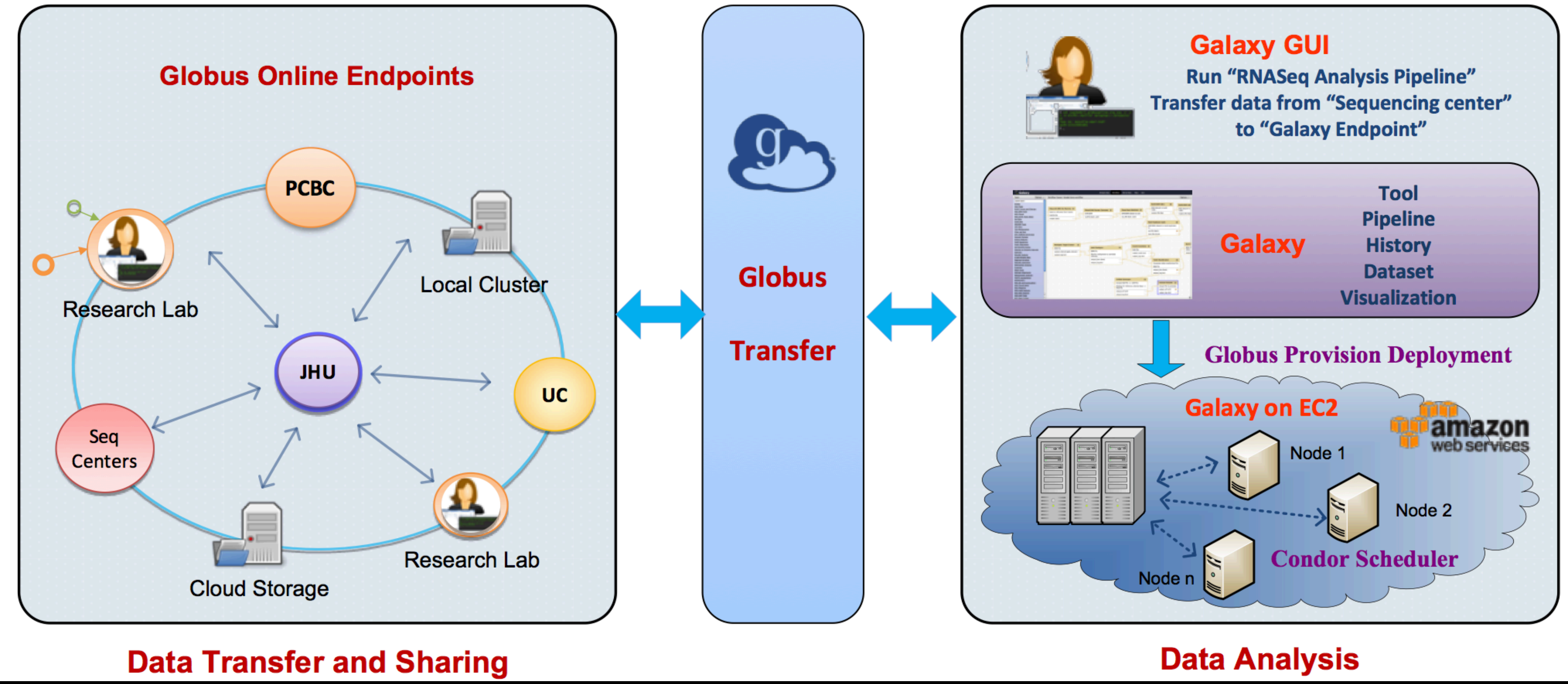
globus genomics
UNIVERSITY OF EXETER
KBase
NERSC
MICHIGAN
XSEDE
Extreme Science and Engineering Discovery Environment
CORNELL UNIVERSITY FOUNDED A.D. 1865

Data management SaaS (Globus) + Next-gen sequence analysis pipelines (Galaxy) + Cloud IaaS (Amazon) =

Flexible, scalable, easy-to-use genomics analysis for all biologists

globus
genomics





Globus Online Endpoints

PCBC

Local Cluster

Research Lab

JHU

UC

Seq Centers

Cloud Storage

Research Lab

Data Transfer and Sharing

Globus Transfer

Galaxy GUI

Run "RNASeq Analysis Pipeline"
Transfer data from "Sequencing center"
to "Galaxy Endpoint"

Galaxy

Tool
Pipeline
History
Dataset
Visualization

Globus Provision Deployment

Galaxy on EC2

amazon web services

Node 1

Node 2

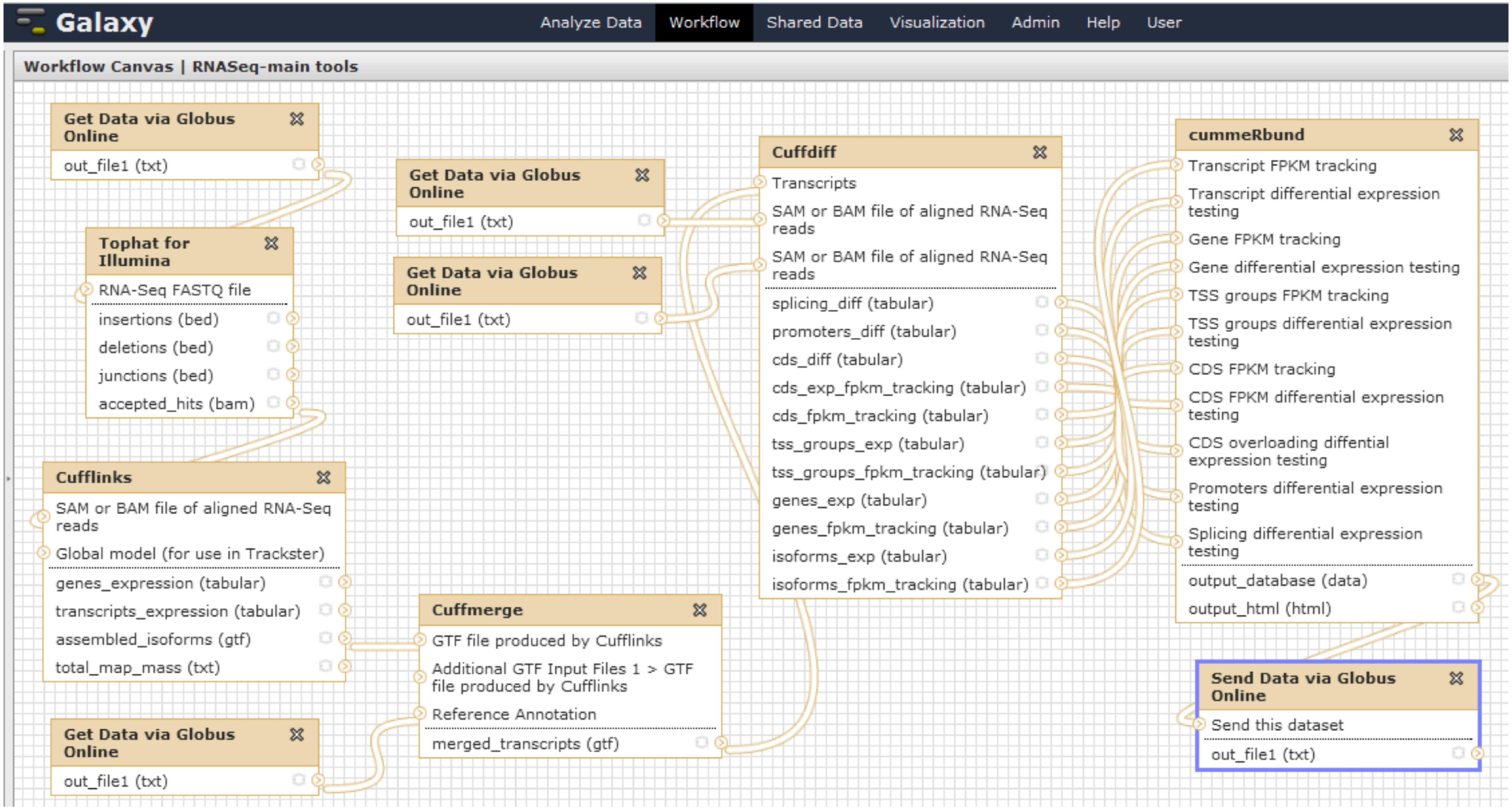
Node n

Condor Scheduler

Data Analysis

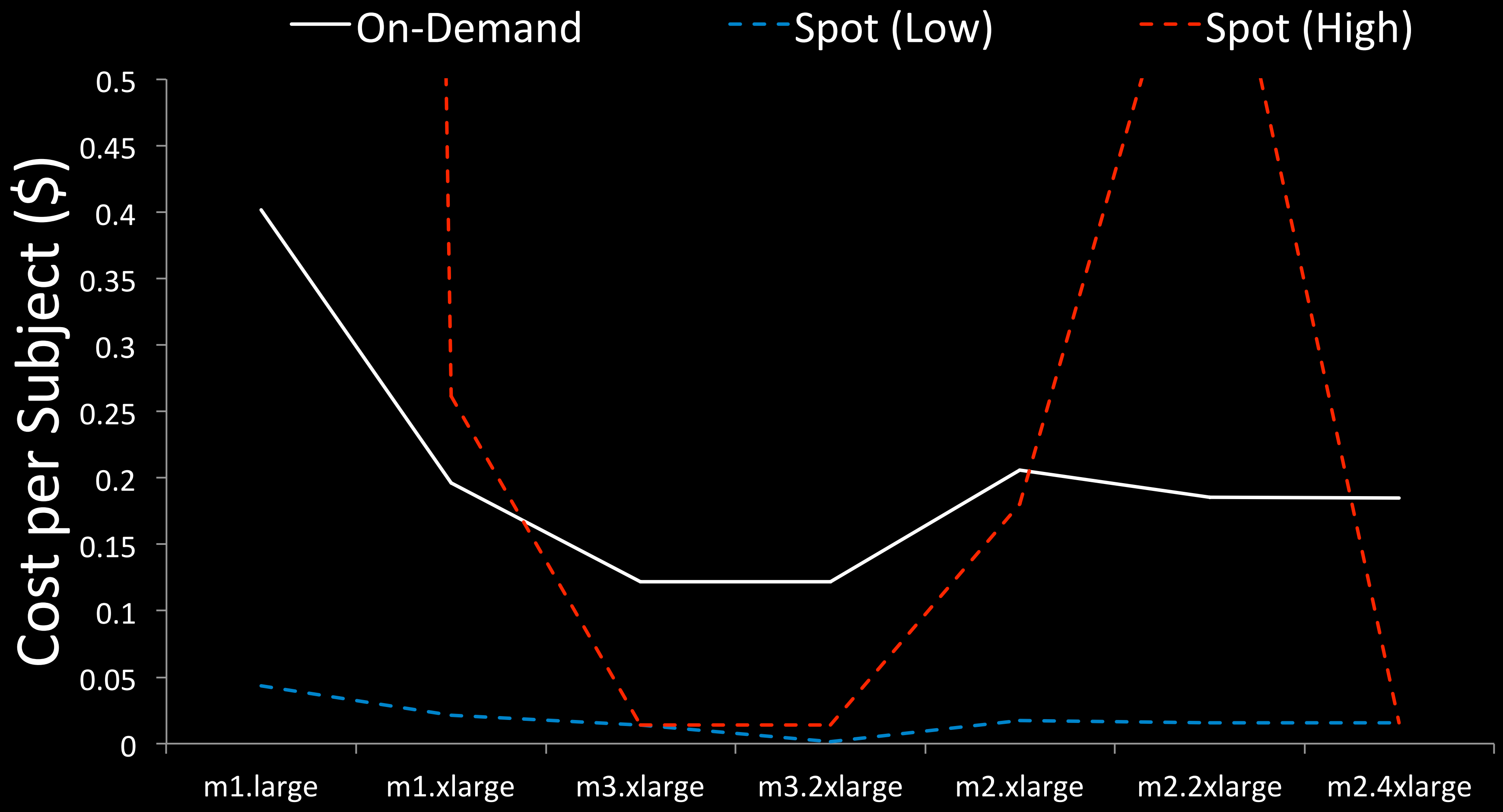
Ravi Madduri, Bo Liu, Paul Davé, et al.

RNA-Seq pipeline


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Workflow Canvas | RNASeq-main tools
Get Data via Globus Online
out_file1 (txt)
Tophat for Illumina
RNA-Seq FASTQ file
insertions (bed)
deletions (bed)
junctions (bed)
accepted_hits (bam)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)
Get Data via Globus Online
out_file1 (txt)
Get Data via Globus Online
out_file1 (txt)
Get Data via Globus Online
out_file1 (txt)
Cuffmerge
GTF file produced by Cufflinks
Additional GTF Input Files 1 > GTF file produced by Cufflinks
Reference Annotation
merged_transcripts (gtf)
Cuffdiff
Transcripts
SAM or BAM file of aligned RNA-Seq reads
SAM or BAM file of aligned RNA-Seq reads
splicing_diff (tabular)
promoters_diff (tabular)
cds_diff (tabular)
cds_exp_fpkm_tracking (tabular)
cds_fpkm_tracking (tabular)
tss_groups_exp (tabular)
tss_groups_fpkm_tracking (tabular)
genes_exp (tabular)
genes_fpkm_tracking (tabular)
isoforms_exp (tabular)
isoforms_fpkm_tracking (tabular)
cummeRbund
Transcript FPKM tracking
Transcript differential expression testing
Gene FPKM tracking
Gene differential expression testing
TSS groups FPKM tracking
TSS groups differential expression testing
CDS FPKM tracking
CDS FPKM differential expression testing
CDS overloading diffential expression testing
Promoters differential expression testing
Splicing differential expression testing
output_database (data)
output_html (html)
Send Data via Globus Online
Send this dataset
out_file1 (txt)

Ravi Madduri, Bo Liu, Paul Davé, et al.

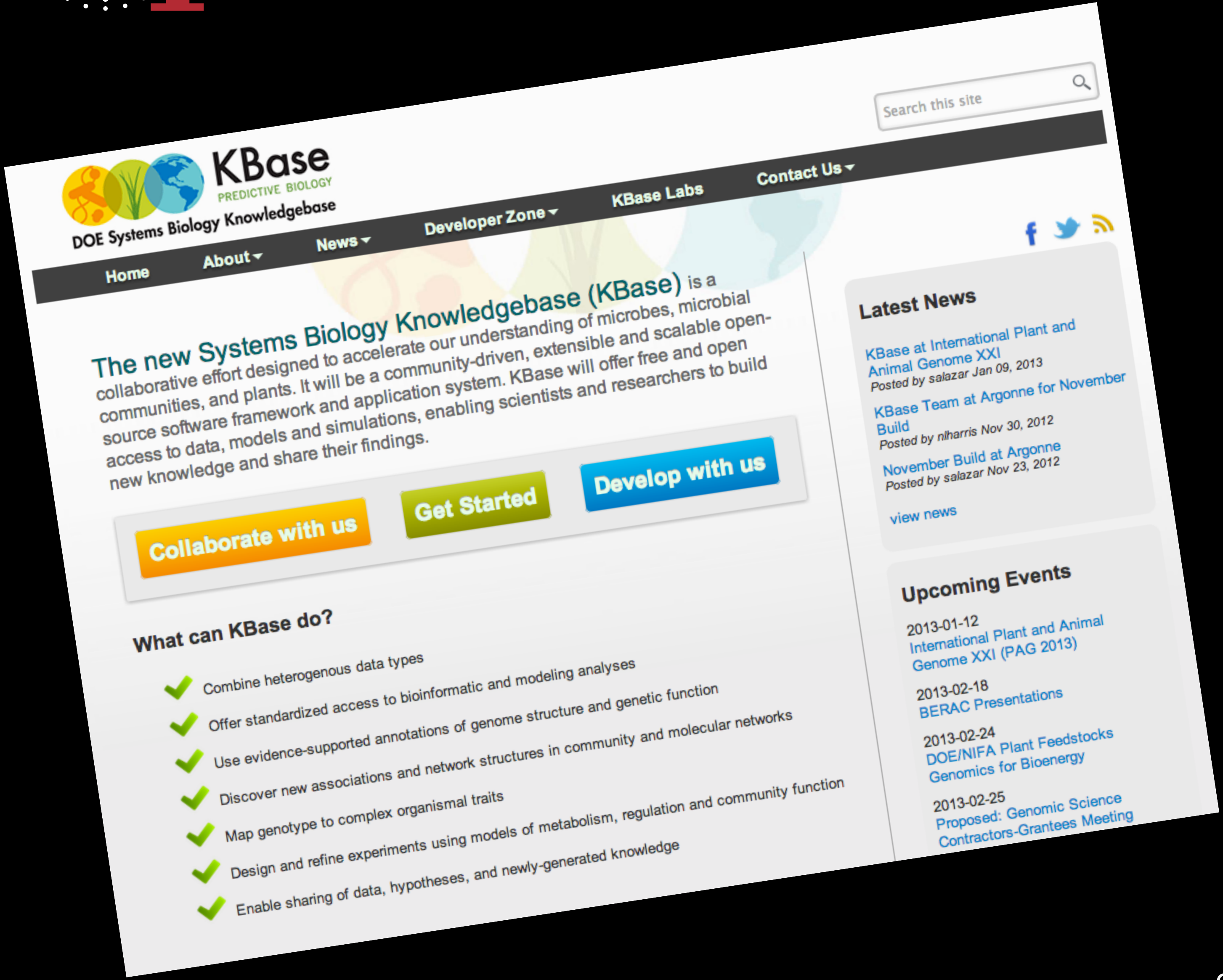
Amazon pricing for Diffusion Tensor Imaging pipeline



Credit: Kyle Chard

computationinstitute.org

A platform for integration



The new Systems Biology Knowledgebase (KBase) is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It will be a community-driven, extensible and scalable open-source software framework and application system. KBase will offer free and open access to data, models and simulations, enabling scientists and researchers to build new knowledge and share their findings.

What can KBase do?

- ✓ Combine heterogeneous data types
- ✓ Offer standardized access to bioinformatic and modeling analyses
- ✓ Use evidence-supported annotations of genome structure and genetic function
- ✓ Discover new associations and network structures in community and molecular networks
- ✓ Map genotype to complex organismal traits
- ✓ Design and refine experiments using models of metabolism, regulation and community function
- ✓ Enable sharing of data, hypotheses, and newly-generated knowledge

Latest News

KBase at International Plant and Animal Genome XXI
Posted by salazar Jan 09, 2013

KBase Team at Argonne for November Build
Posted by nlharris Nov 30, 2012

November Build at Argonne
Posted by salazar Nov 23, 2012

view news

Upcoming Events

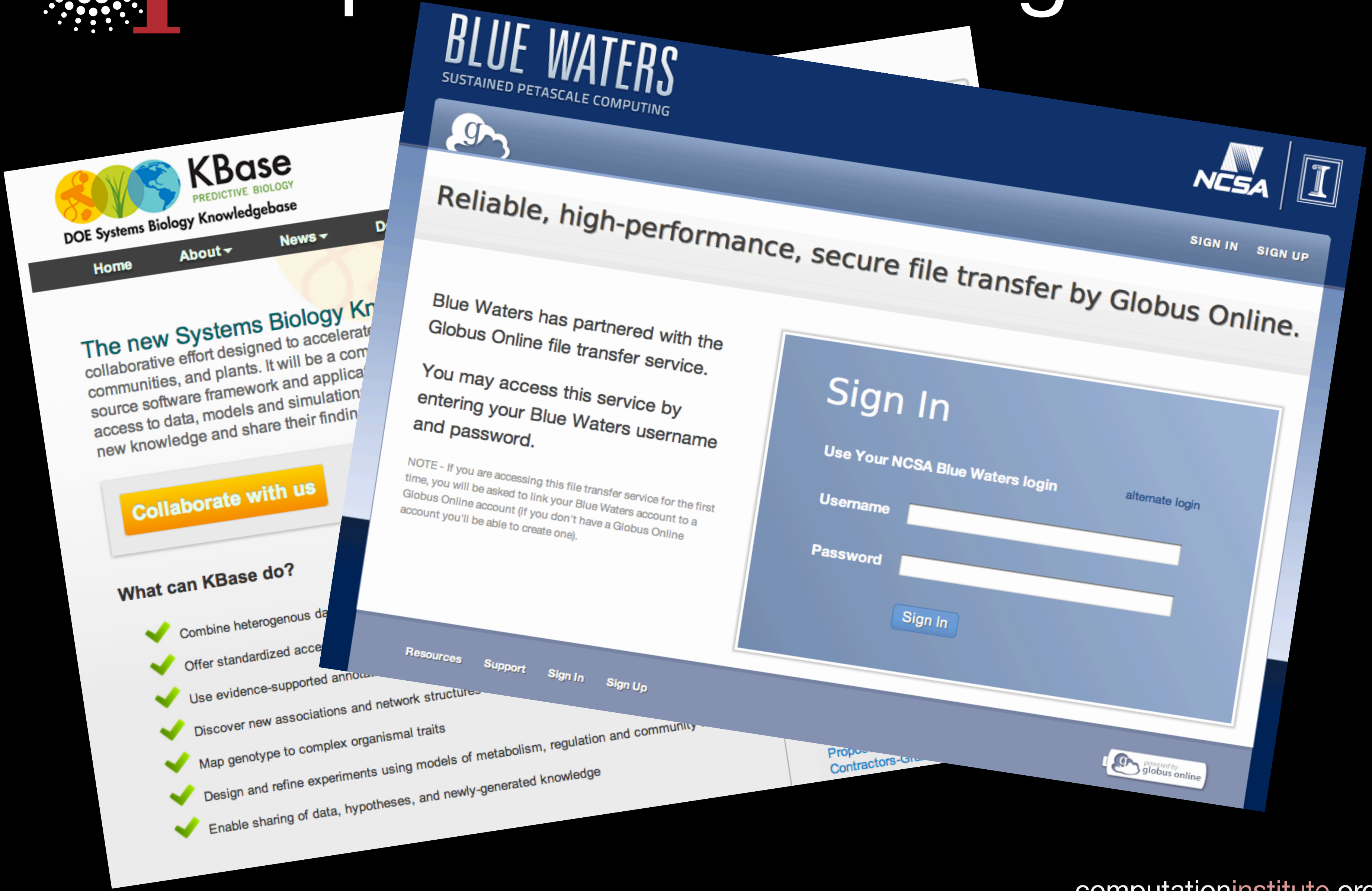
2013-01-12
International Plant and Animal Genome XXI (PAG 2013)

2013-02-18
BERAC Presentations

2013-02-24
DOE/NIFA Plant Feedstocks Genomics for Bioenergy

2013-02-25
Proposed: Genomic Science Contractors-Grantees Meeting

A platform for integration



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

The new Systems Biology Kn
collaborative effort designed to accelerat
communities, and plants. It will be a com
source software framework and applica
access to data, models and simulation
new knowledge and share their findin

What can KBase do?

- ✓ Combine heterogenous da
- ✓ Offer standardized acce
- ✓ Use evidence-supported annota
- ✓ Discover new associations and network structures
- ✓ Map genotype to complex organismal traits
- ✓ Design and refine experiments using models of metabolism, regulation and community
- ✓ Enable sharing of data, hypotheses, and newly-generated knowledge

Propo
Contractors-Gra

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

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

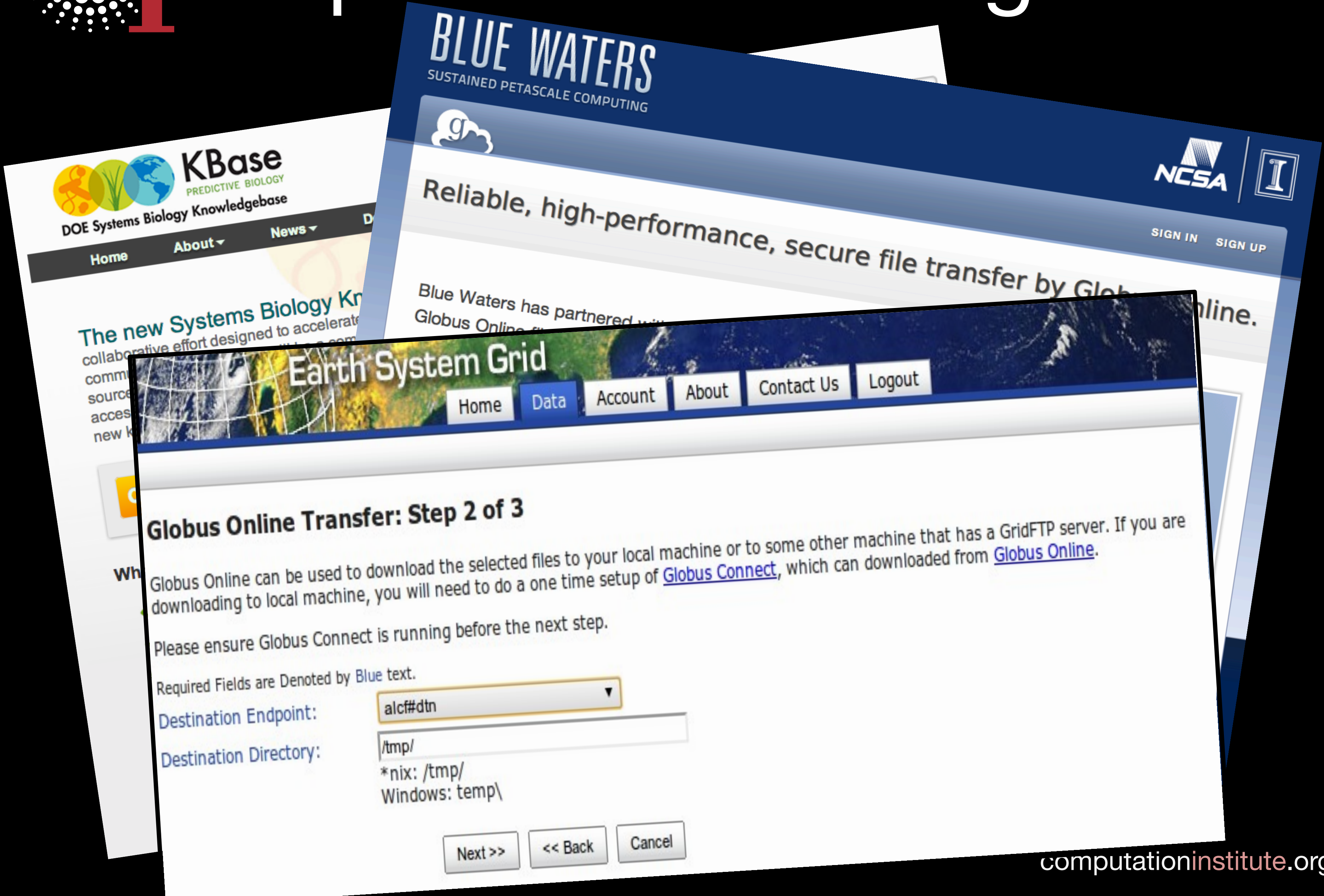
Use Your NCSA Blue Waters login

alternate login

Username

Password

A platform for integration



The new Systems Biology Kn
collaborative effort designed to accelerat
comm
source
acces
new k

Reliable, high-performance, secure file transfer by Glob nline.

Blue Waters has partnered w
Globus Onlin

Globus Online Transfer: Step 2 of 3

Globus Online can be used to download the selected files to your local machine or to some other machine that has a GridFTP server. If you are downloading to local machine, you will need to do a one time setup of Globus Connect, which can downloaded from Globus Online.

Please ensure Globus Connect is running before the next step.

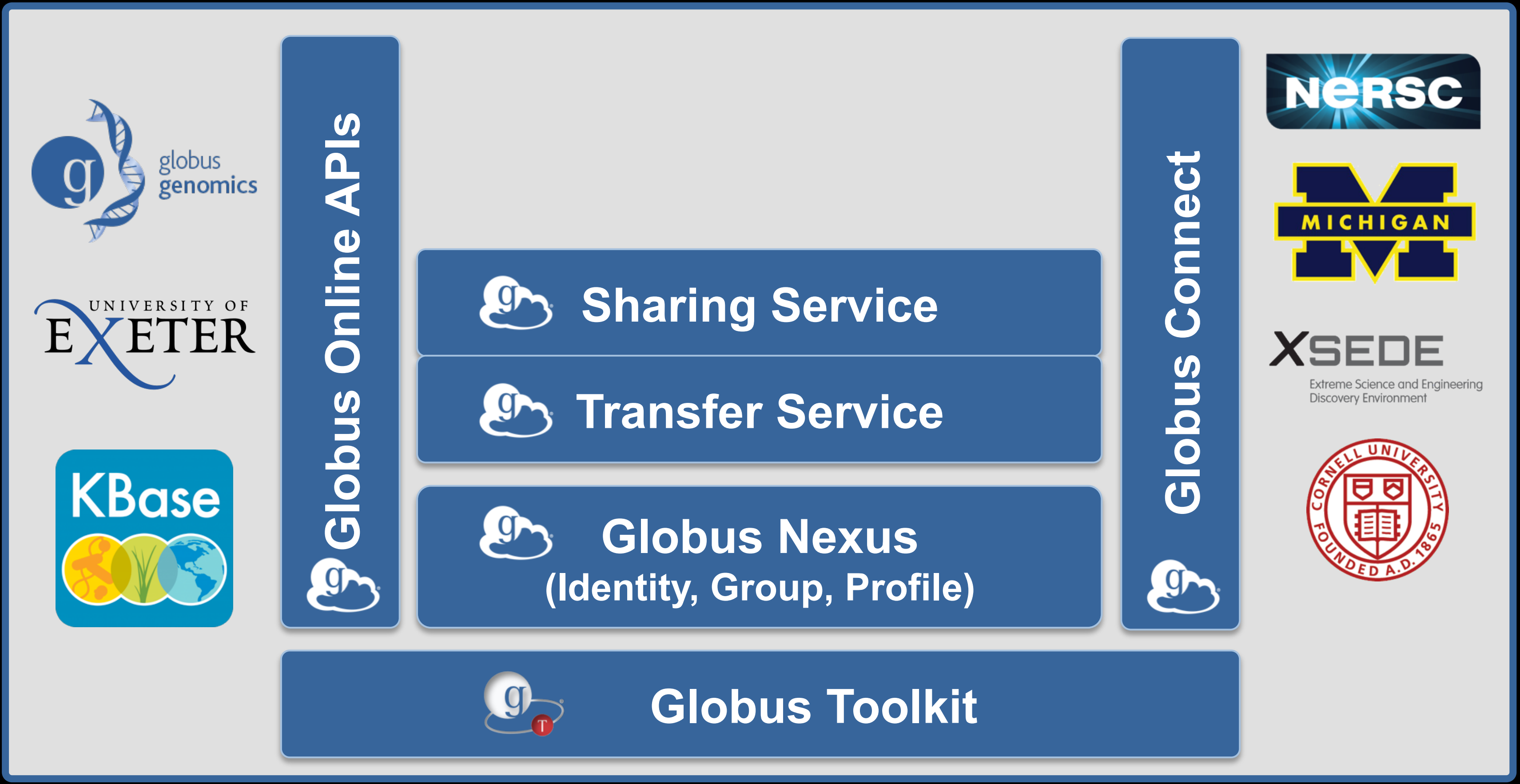
Required Fields are Denoted by Blue text.

Destination Endpoint: alcf#dtn ▾

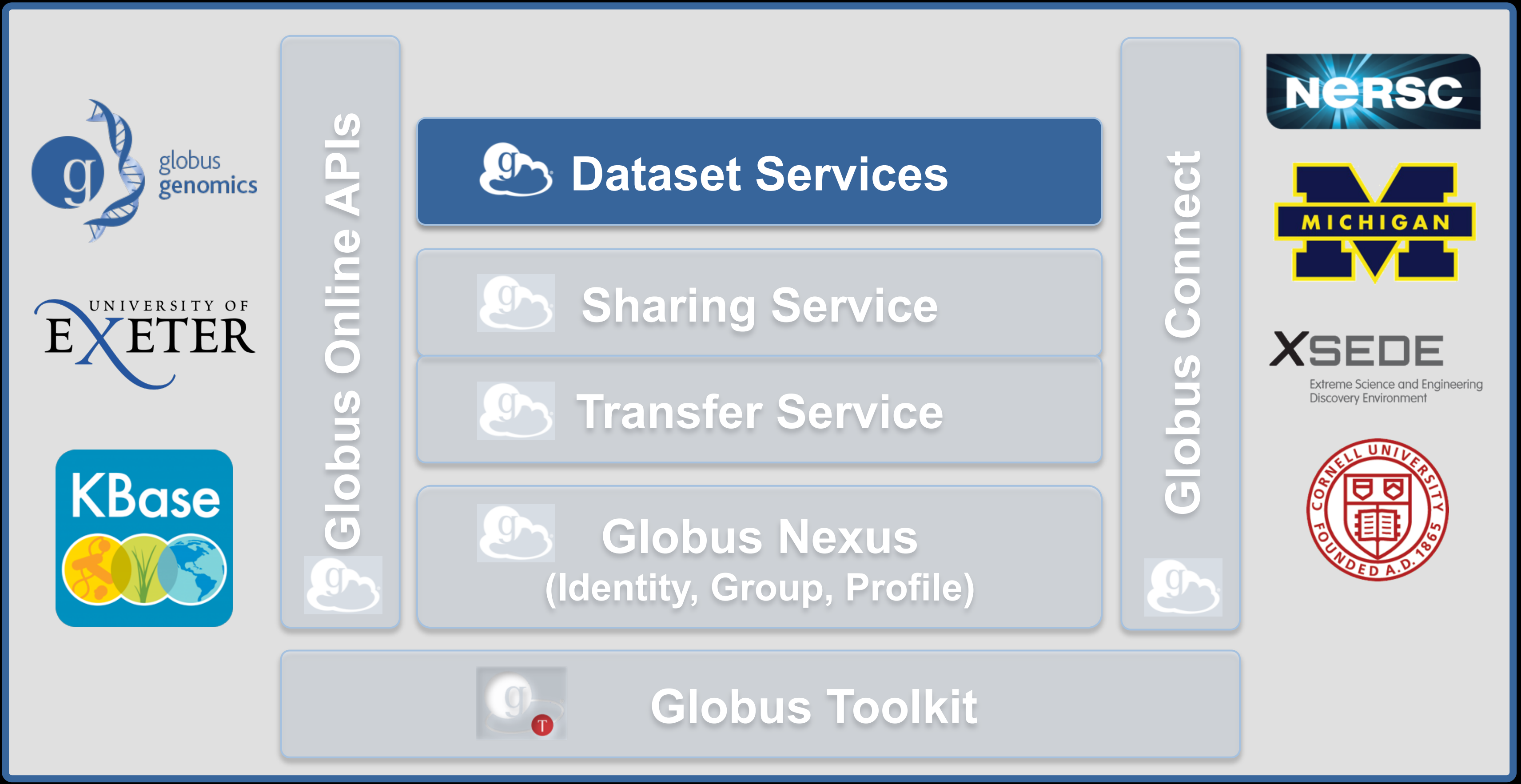
Destination Directory: /tmp/

*nix: /tmp/
Windows: temp\

We are also adding capabilities



More capabilities underway …



Expanding Globus Online services

- Ingest and publication
  - Imagine a DropBox that not only replicates, but also extracts metadata, catalogs, converts
- Cataloging
  - Virtual views of data based on user-defined and/or automatically extracted metadata
- Computation
  - Associate computational procedures, orchestrate application, catalog results, record provenance

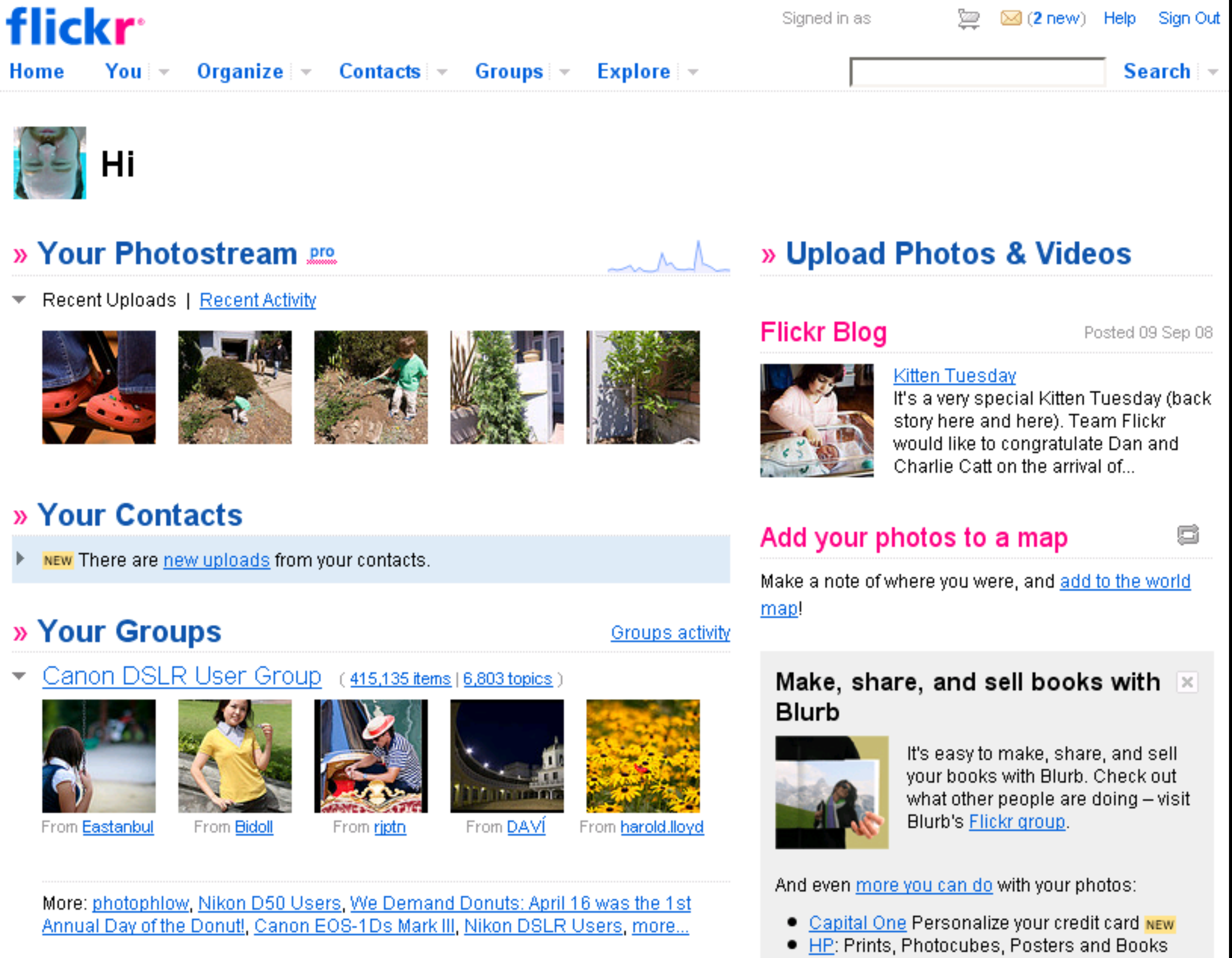
Looking deeply at how researchers use data

- A single research question often requires the integration of many data elements, that are:
  - – In different locations
  - – In different formats (Excel, text, CDF, HDF, …)
  - – Described in different ways
- Best grouping can vary during investigation
  - – Longitudinal, vertical, cross-cutting
- But always needs to be operated on as a unit
  - – Share, annotate, process, copy, archive, …

How do we manage data today?

- Often, a curious mix of ad hoc methods
  - Organize in directories using file and directory naming conventions
  - Capture status in README files, spreadsheets, notebooks
- Time-consuming, complex, error prone

Why can’t we manage our data like we manage our pictures and music?



Hi

» Your Photostream pro

Recent Uploads | Recent Activity

» Your Contacts

NEW There are new uploads from your contacts.

» Your Groups

Groups activity

Canon DSLR User Group (415,135 items | 6,803 topics)

From Eastanbul   From Bidoll   From rjptn   From DAVÍ   From harold.lloyd

More: photophlow, Nikon D50 Users, We Demand Donuts: April 16 was the 1st Annual Day of the Donut!, Canon EOS-1Ds Mark III, Nikon DSLR Users, more...

» Upload Photos & Videos

Flickr Blog

Posted 09 Sep 08

Kitten Tuesday

It's a very special Kitten Tuesday (back story here and here). Team Flickr would like to congratulate Dan and Charlie Catt on the arrival of...

Add your photos to a map

Make a note of where you were, and add to the world map!

Make, share, and sell books with Blurb

It's easy to make, share, and sell your books with Blurb. Check out what other people are doing – visit Blurb's Flickr group.

And even more you can do with your photos:

- Capital One Personalize your credit card NEW
- HP: Prints, Photocubes, Posters and Books

Introducing the **dataset**

- **Group** data based on use, not location
  - Logical grouping to organize, reorganize, search, and describe usage
- **Tag** with characteristics that reflect content …
  - Capture as much existing information as we can
- …or to reflect current status in investigation
  - Stage of processing, provenance, validation, ..
- **Share** data sets for collaboration
  - Control access to data and metadata
- **Operate** on datasets as units
  - Copy, export, analyze, tag, archive, …

Builds on catalog as a service

Approach

- Hosted user-defined catalogs
- Based on tag model

  <subject, name, value>
- Optional schema constraints
- Integrated with other Globus services

Three REST APIs

/query/

- Retrieve subjects

/tags/

- Create, delete, retrieve tags

/tagdef/

- Create, delete, retrieve tag definitions

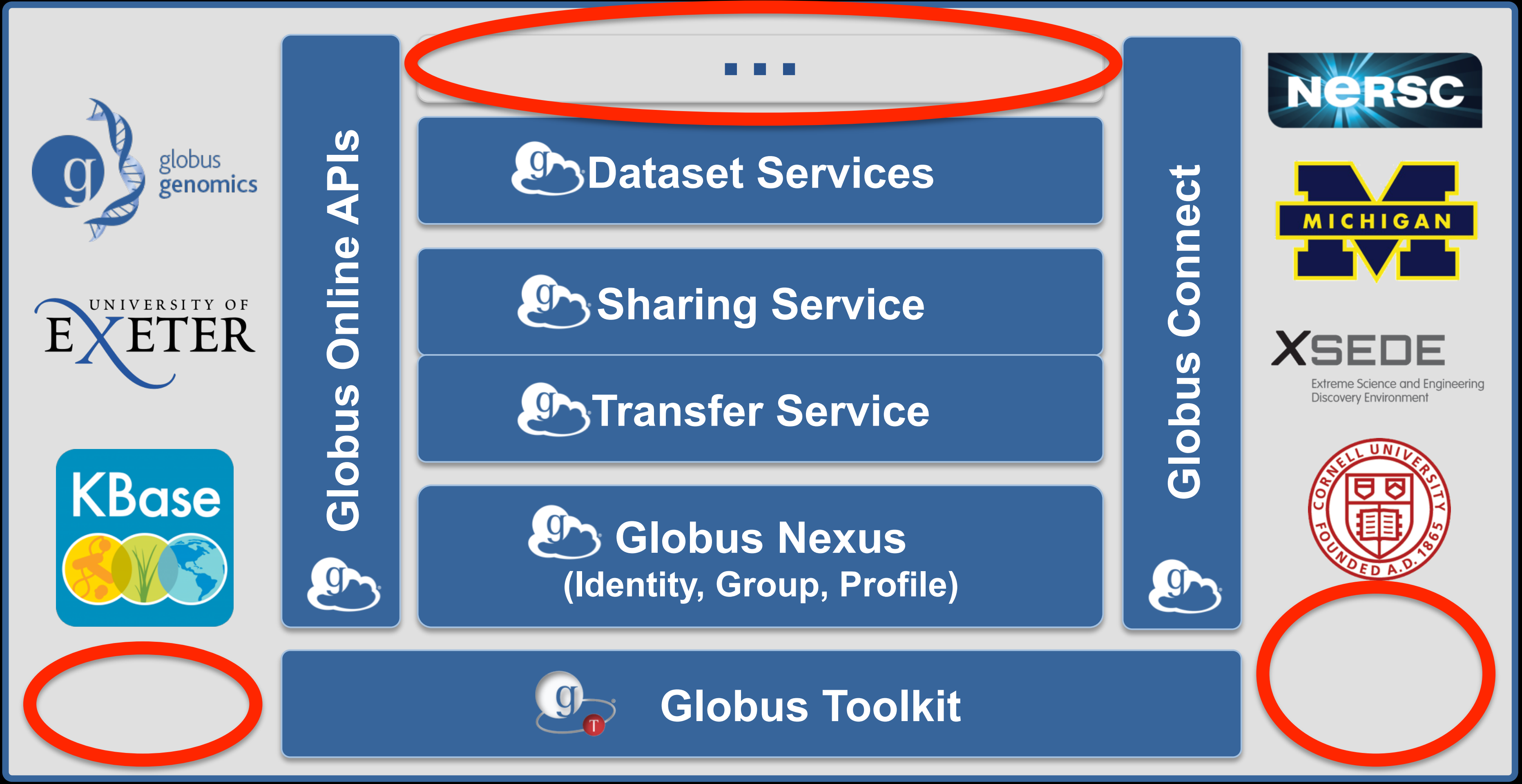
Builds on USC Tagfiler project (C. Kesselman et al.)

Our vision for a 21st century discovery infrastructure

Provide **more** capability for **more** people at **lower cost** by building a **“Discovery Cloud”**

Delivering “Science as a service”

It's a time of great opportunity … to develop and apply Science aaS



...

Dataset Services

Sharing Service

Transfer Service

Globus Nexus (Identity, Group, Profile)

Globus Online APIs

Globus Connect

Globus Toolkit

globus genomics
UNIVERSITY OF EXETER
KBase
NERSC
MICHIGAN
XSEDE
Extreme Science and Engineering Discovery Environment
CORNELL UNIVERSITY FOUNDED A.D. 1865

Thanks to great colleagues and collaborators

- Steve Tuecke, Rachana Ananthakrishnan, Kyle Chard, Raj Kettimuthu, Ravi Madduri, Tanu Malik, and many others at Argonne & Uchicago
- Carl Kesselman, Karl Czajkowski, Rob Schuler, and others at USC/ISI
- Francesco de Carlo, Chris Jacobsen, and others at Argonne

Thank you to our sponsors!

U.S. DEPARTMENT OF ENERGY

NSF

THE UNIVERSITY OF CHICAGO

NATIONAL INSTITUTES OF HEALTH

Argonne NATIONAL LABORATORY