Big Data Management and Analysis Using Globus Services

Ravi Madduri

Argonne National Laboratory & University of Chicago

2015 Joint Stats Meeting in Seattle, WA

Outline

- Who we are
- Challenges in Big Data Management and Analysis
- Sustainability and Reproducibility
- Globus Research Data Management Service
  - Numbers, Usage Stats
- Globus Genomics
  - Description
  - Novel Pipelines
  - User segments
  - Adoption
  - Economics

Our vision for a 21st century discovery infrastructure

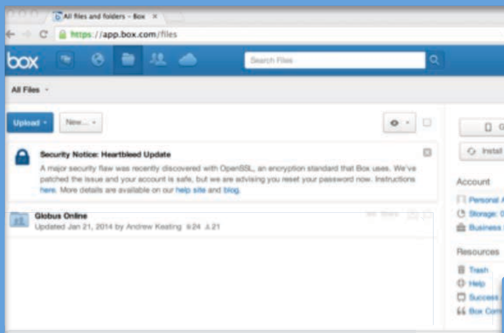
To provide **more** capability for **more** people at **substantially lower cost**

Research data management scenarios and challenges

In Big, Medium *and* Small data

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

Research Computing HPC Cluste



All files and folders - Box

https://app.box.com/files

All Files

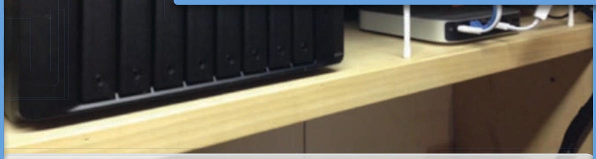
Security Notice: Heartbleed Update

A major security flaw was recently discovered with OpenSSL, an encryption standard that Box uses. We've patched the issue and your account is safe, but we are advising you reset your password now. Instructions here. More details are available on our help site and blog.

Globus Online

Updated Jan 21, 2014 by Andrew Keating 24 21

Campus Home Filesyste



Lab Server

Personal Laptop

Desktop Workstation

XSEDE Resource

Public Cloud

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

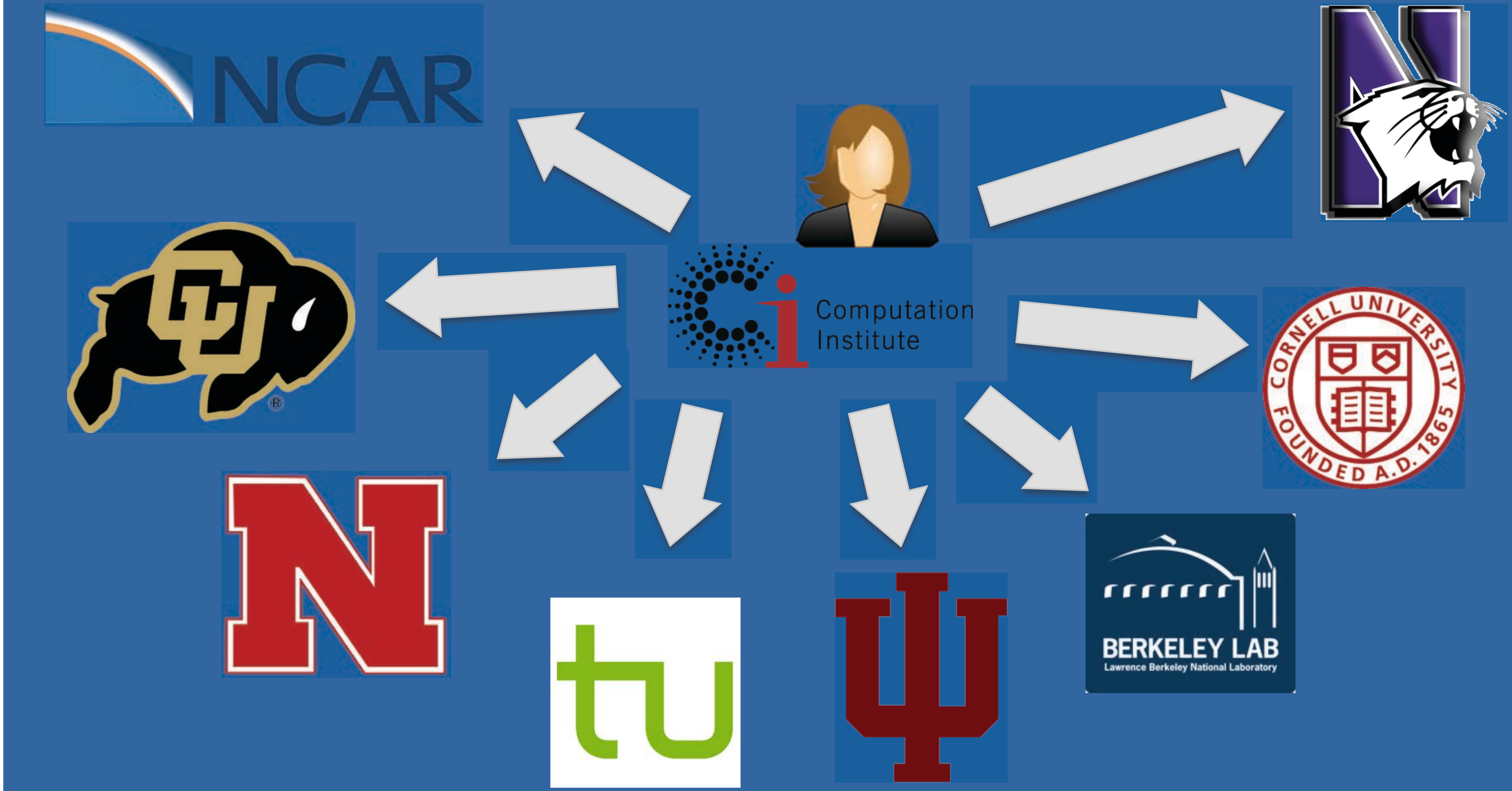
MRI

Advanced Light Source

Next Gen Sequencer

Light Sheet Microscope

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


NCAR
Computation Institute
CORNELL UNIVERSITY FOUNDED A.D. 1865
BERKELEY LAB
Lawrence Berkeley National Laboratory
tu

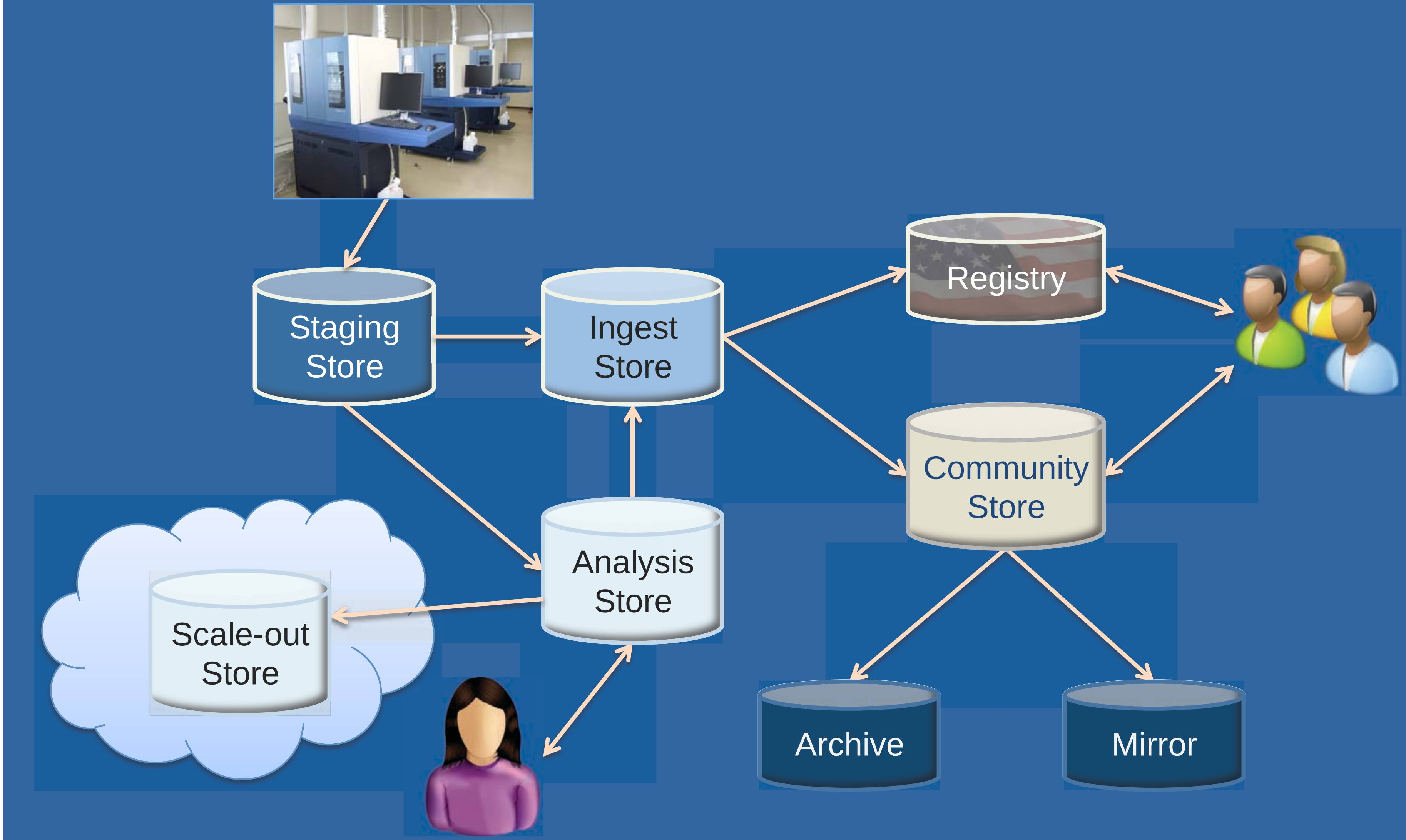
“I need to publish my data so that others can find it and use it.”

Scholarly Publication

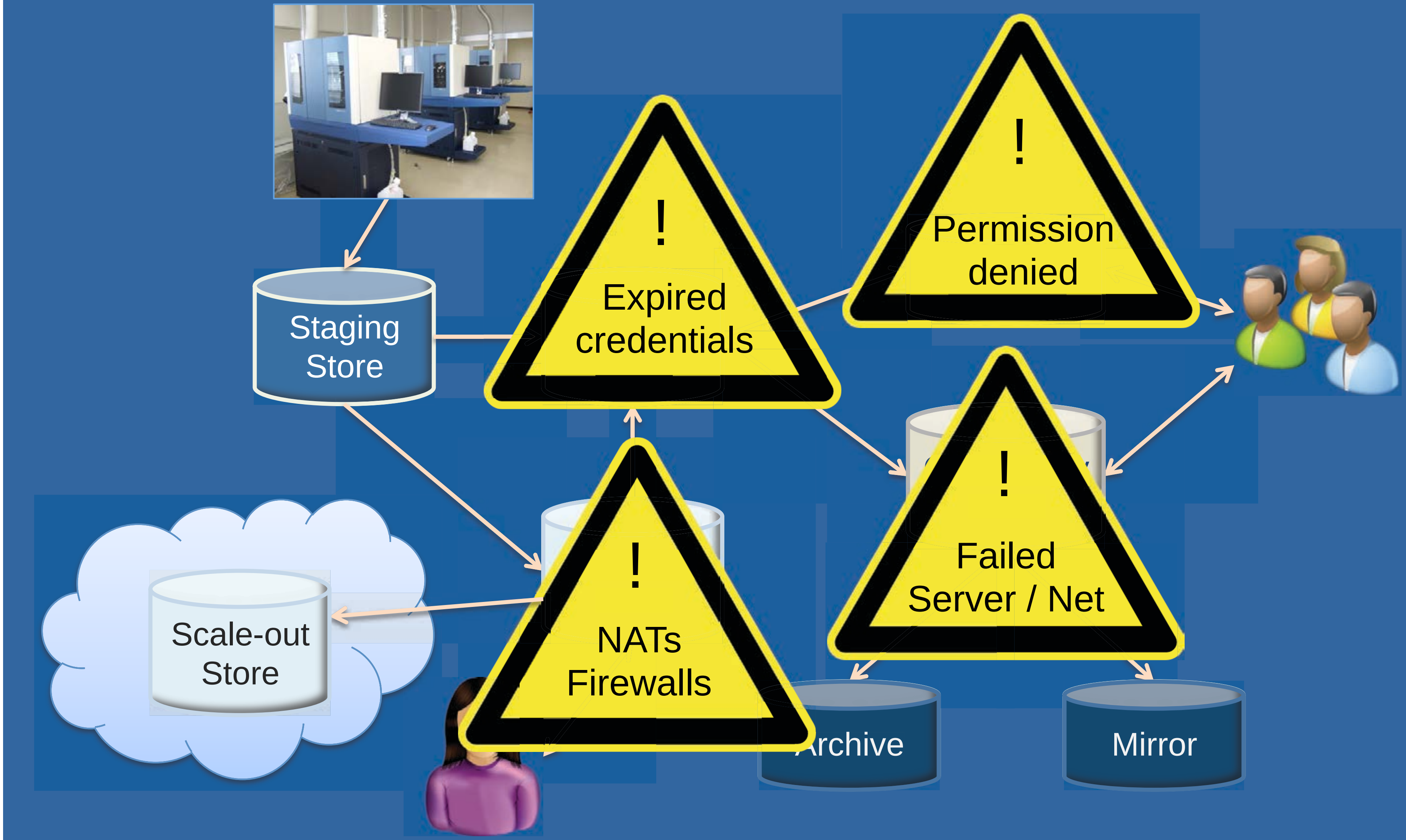
Reference Dataset

Active Research Collaboration

Managing data should be easy …



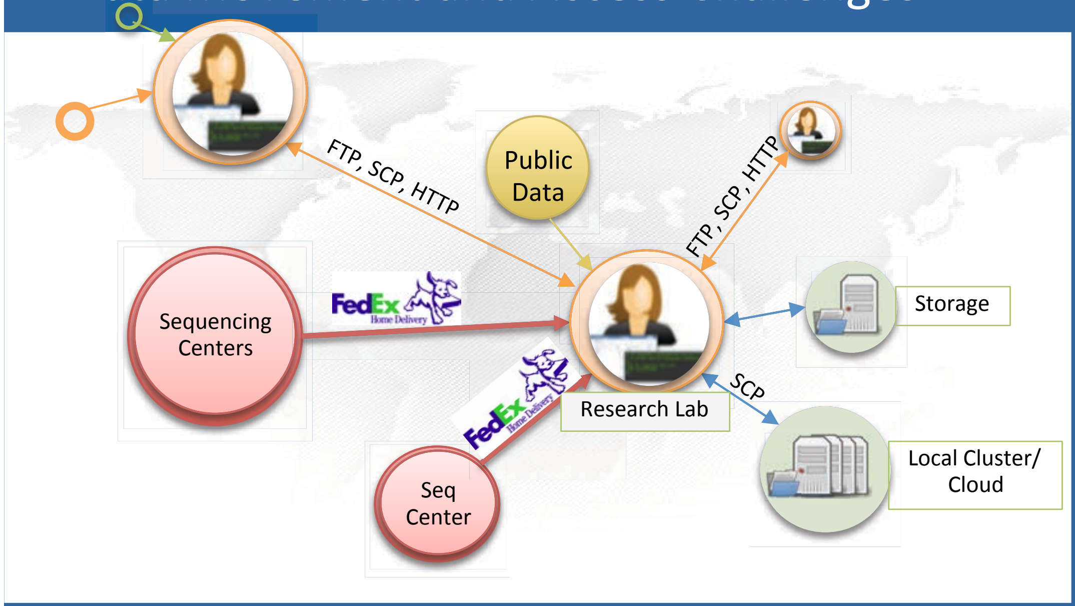
… but it’s hard and frustrating!



In Genomics..

Challenges in Sequencing Analysis

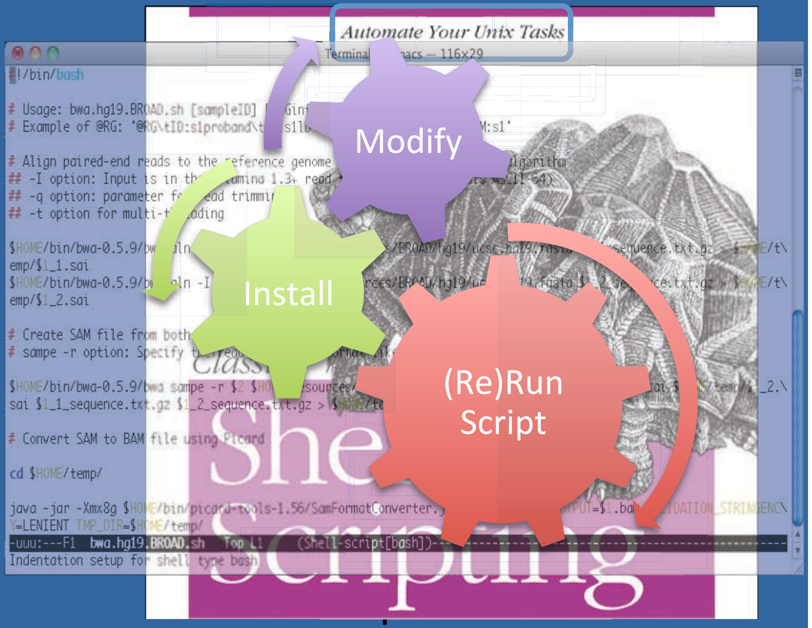
Data Movement and Access Challenges



- Data is distributed in different locations
- Research labs need access to the data for analysis
- Be able to Share data with other researchers/collaborators
  - Inefficient ways of data movement
- Data needs to be available on the local and Distributed Compute Resources
  - Local Clusters, Cloud, Grid

Once we have the Sequence Data

- Manually move the data to the Compute node
- Install all the tools required for the Analysis
  - BWA, Picard, GATK, Filtering Scripts, etc.
- Shell scripts to sequentially execute the tools
- Manually modify the scripts for any change
  - Error Prone, difficult to keep track, messy..
- Difficult to maintain and transfer the knowledge



```
#!/bin/bash

# Usage: bwa.hg19.BROAD.sh [sampleID]   Gin
# Example of @RG: '@RG\tID:s1proband\t   s1lb                 M:s1'

# Align paired-end reads to the reference genome                algorithm
## -I option: Input is in the Illumina 1.3+ read                ASCII 64)
## -q option: parameter for read trimming
## -t option for multi-threading

$HOME/bin/bwa-0.5.9/bwa aln    s/BROAD/hg19/ucsc.hg19.fasta    sequence.txt.gz > $HOME/t\
emp/$1_1.sai
$HOME/bin/bwa-0.5.9/bwa aln -I    rces/BROAD/hg19/ucsc.hg19.fasta $1_2_sequence.txt.gz > $HOME/t\
emp/$1_2.sai

# Create SAM file from both
# sampe -r option: Specify the read    format like

$HOME/bin/bwa-0.5.9/bwa sampe -r $2 $HOME    sources/                   sai $HOME/temp/$1_2.\
sai $1_1_sequence.txt.gz $1_2_sequence.txt.gz > $HOME/te

# Convert SAM to BAM file using Picard

cd $HOME/temp/

java -jar -Xmx8g $HOME/bin/picard-tools-1.56/SamFormatConverter.j    OUTPUT=$1.bam    VALIDATION_STRINGENC\
Y=LENIENT TMP_DIR=$HOME/temp/
-uuu:---F1  bwa.hg19.BROAD.sh   Top L1    (Shell-script[bash])----------------------------------------
Indentation setup for shell type bash
```

Manual Data Analysis

Solutions for data management and analysis at scale

Globus delivers…

Big data transfer, sharing, publication, and discovery…

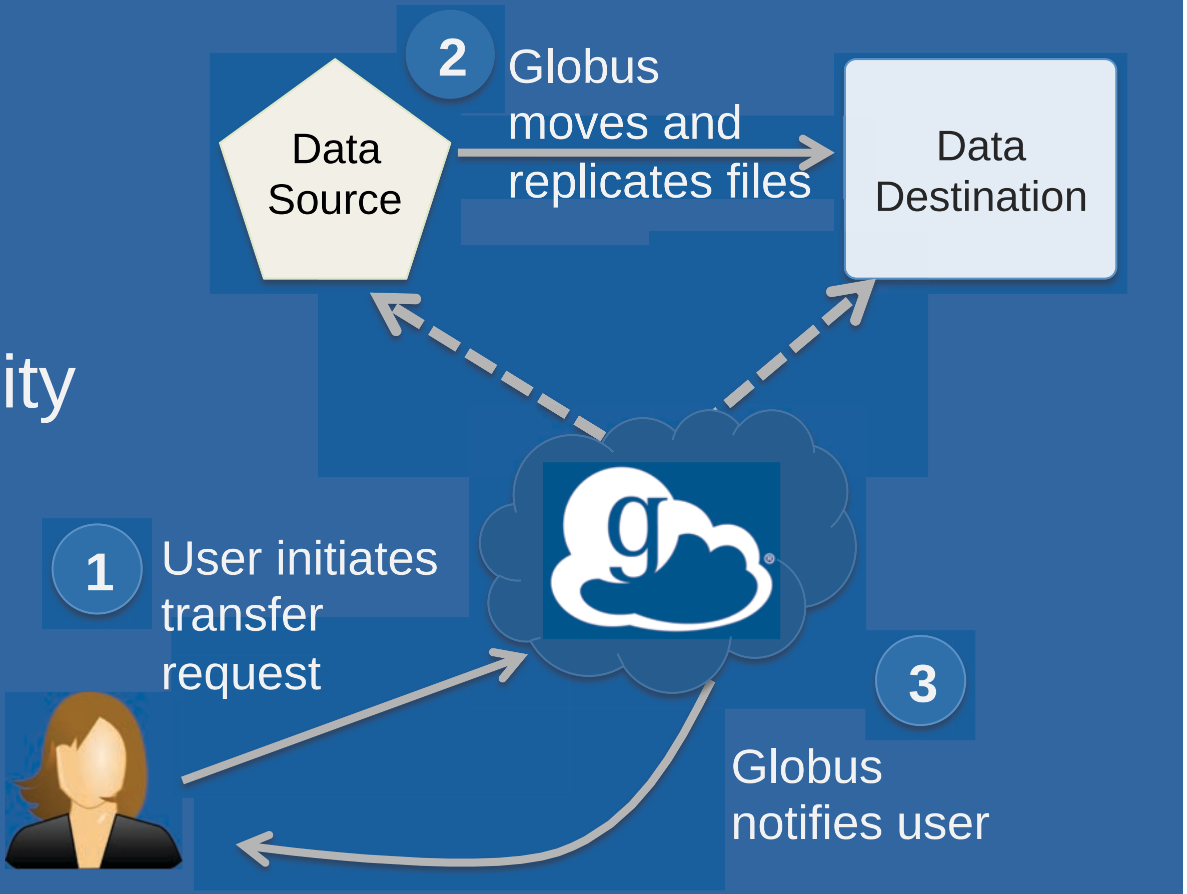
…directly from your own storage systems

Globus is SaaS

- Web, command line, and REST interfaces
- Reduced IT operational costs
- New features automatically available
- Consolidated support & troubleshooting

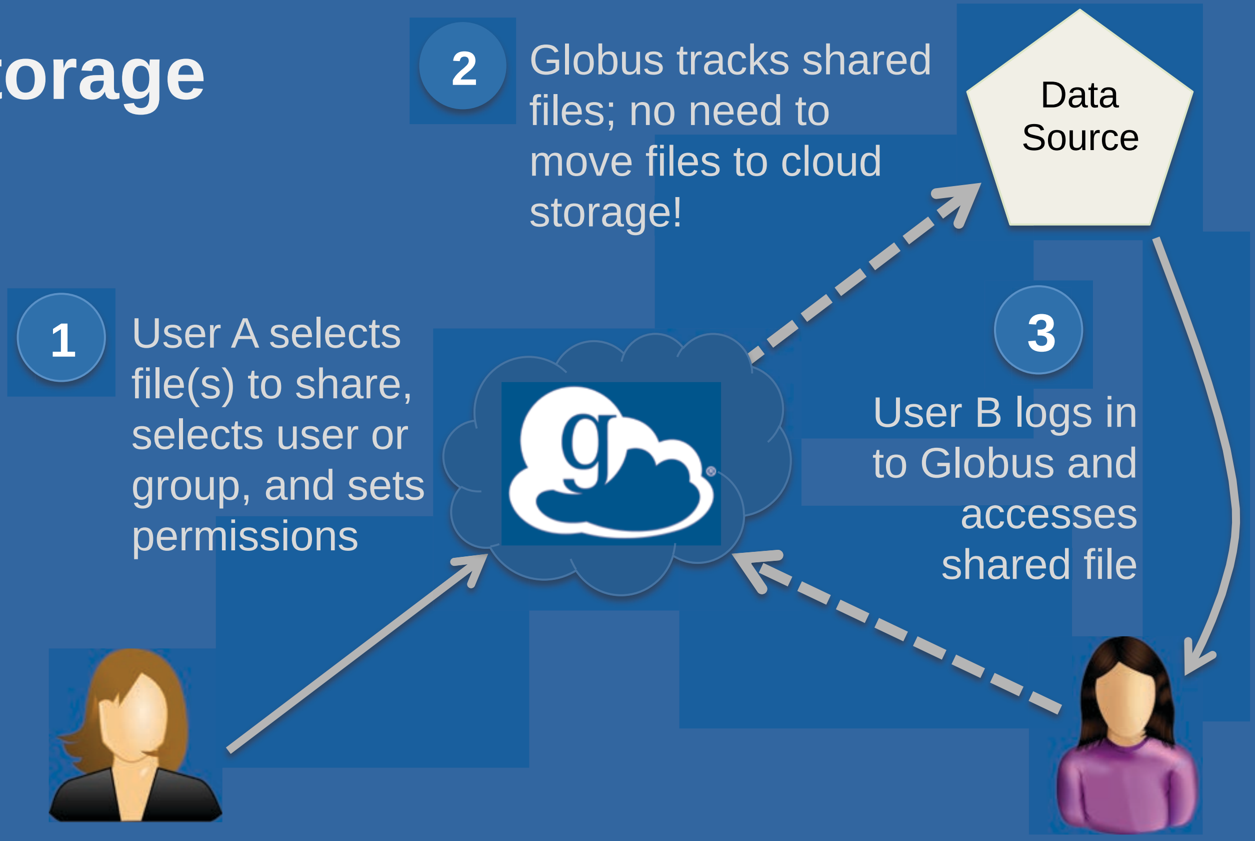
Reliable, secure, high-performance ***file transfer and replication***

- “Fire-and-forget” transfers
- Automatic fault recovery
- Seamless security integration
- Powerful GUI and APIs


Data Source
2
Globus moves and replicates files
Data Destination
1
User initiates transfer request
3
Globus notifies user

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

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



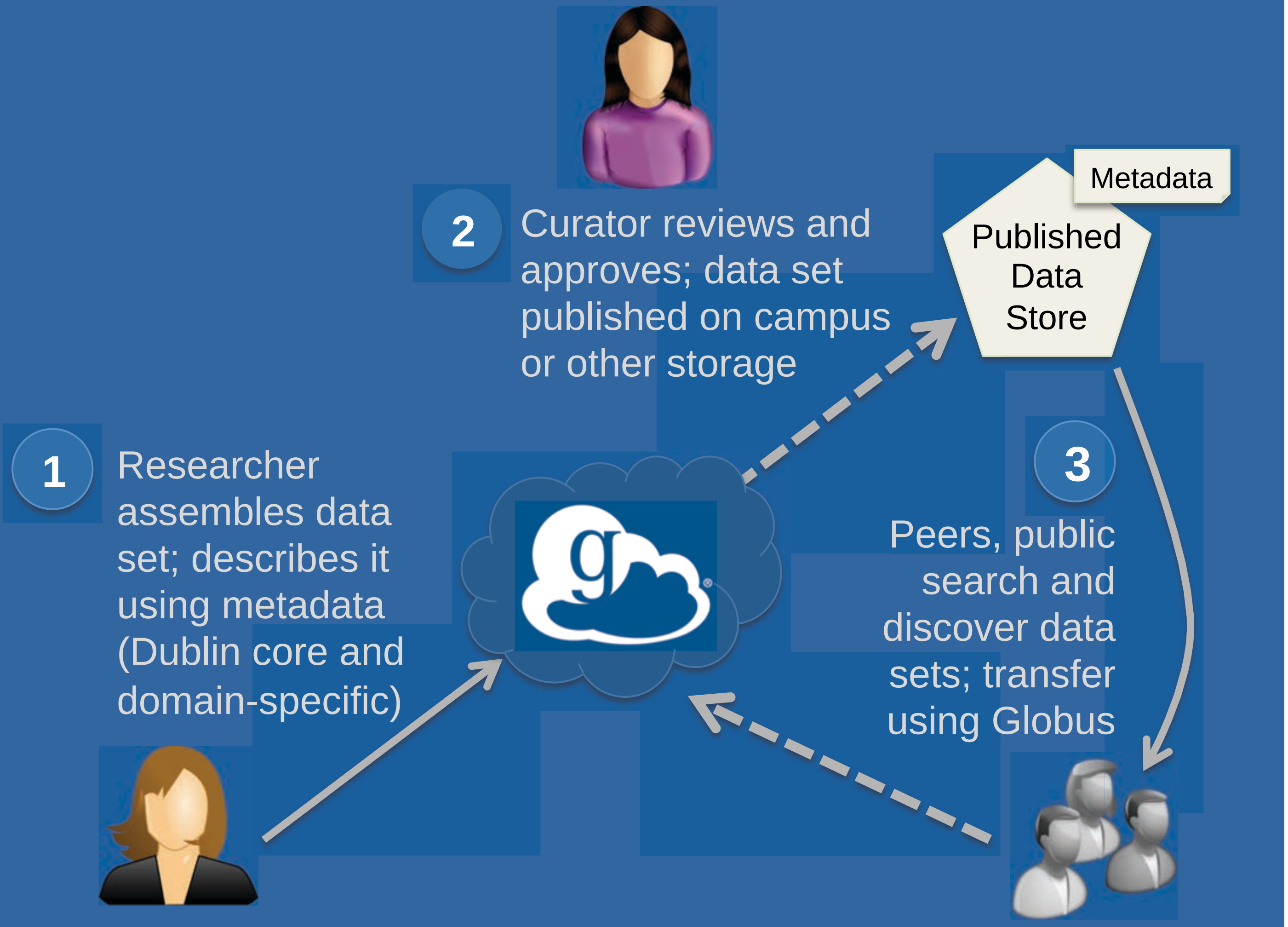
1 User A selects file(s) to share, selects user or group, and sets permissions

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

3 User B logs in to Globus and accesses shared file

Curated ***publication*** of data, with relevant metadata for ***discovery***

- **Identify**
- **Describe**
- **Curate**
- **Verify**
- **Access**
- **Preserve**

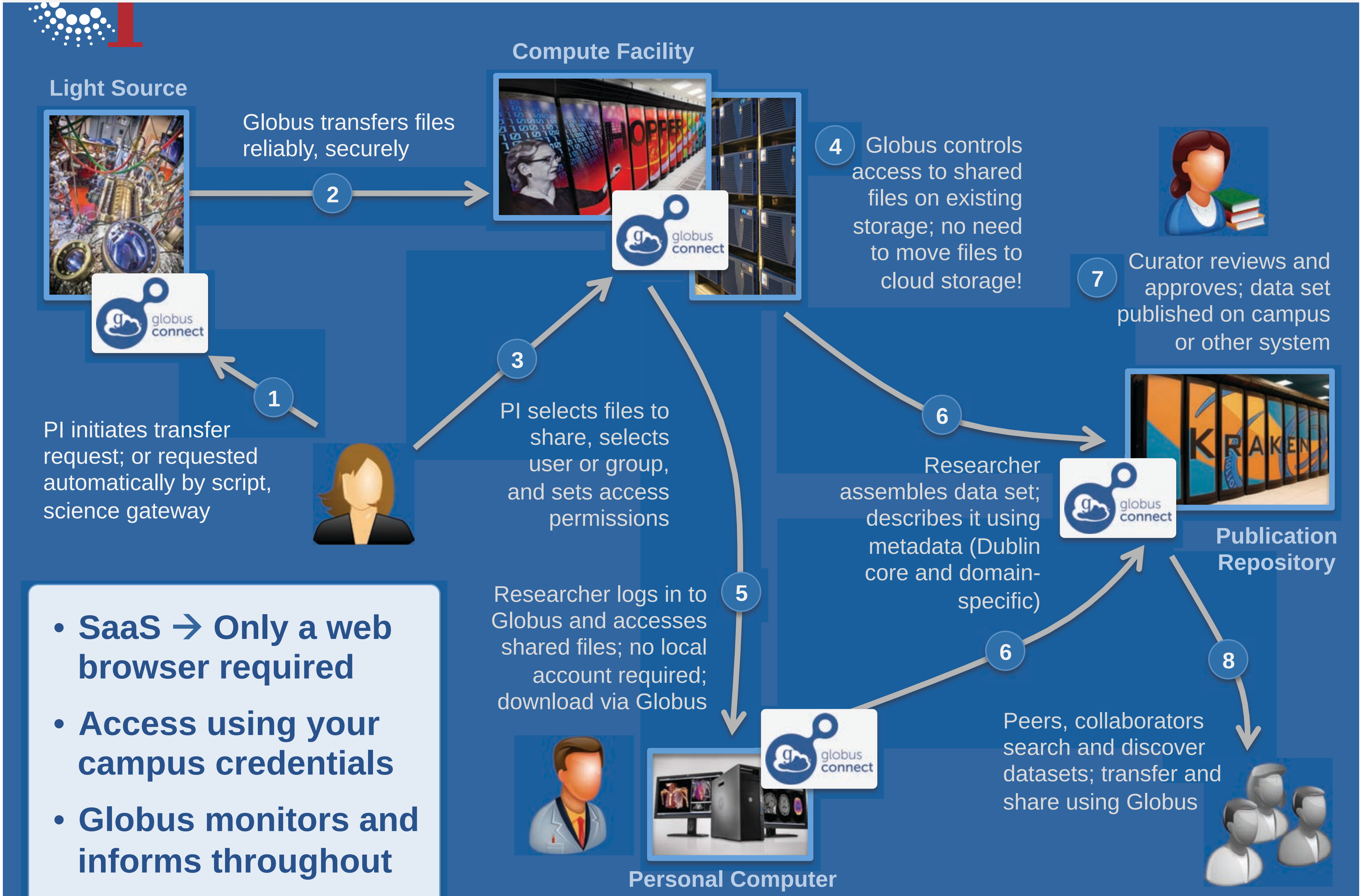


1 Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)

2 Curator reviews and approves; data set published on campus or other storage

3 Peers, public search and discover data sets; transfer using Globus

Managing the research data lifecycle with Globus



Light Source

Compute Facility

Publication Repository

Personal Computer

1 PI initiates transfer request; or requested automatically by script, science gateway

2 Globus transfers files reliably, securely

3 PI selects files to share, selects user or group, and sets access permissions

4 Globus controls access to shared files on existing storage; no need to move files to cloud storage!

5 Researcher logs in to Globus and accesses shared files; no local account required; download via Globus

6 Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)

7 Curator reviews and approves; data set published on campus or other system

8 Peers, collaborators search and discover datasets; transfer and share using Globus

- SaaS → Only a web browser required
- Access using your campus credentials
- Globus monitors and informs throughout

Globus Adoption and Usage

- 166,449 active Globus endpoints
- 27,961 users registered
- Biggest transfer: 500.42TB
- Longest running transfer: 182 days.
- Fastest transfer: 58.5Gbps (average)
- 55TB moved per day, on average, since the service was launched in November 2010
- Average throughput: 637.7Mbps (since service launch)

globus
genomics

Flexible, scalable, affordable genomics analysis for all biologists

Challenges in Scaling Up

- Rapidly evolving state-of-the-art in tools
- Things work reasonably well for small-scale
  - Local and on cloud
- Large-scale analysis requires
  - A computationally gifted postdoc or two
  - Co-location with a large compute facility hungry for justifying purchase
  - Understanding different kinds of parallelism
    - Tool level
    - Workflow level
  - And relate it to science
    - Chromosome level
    - Sample level

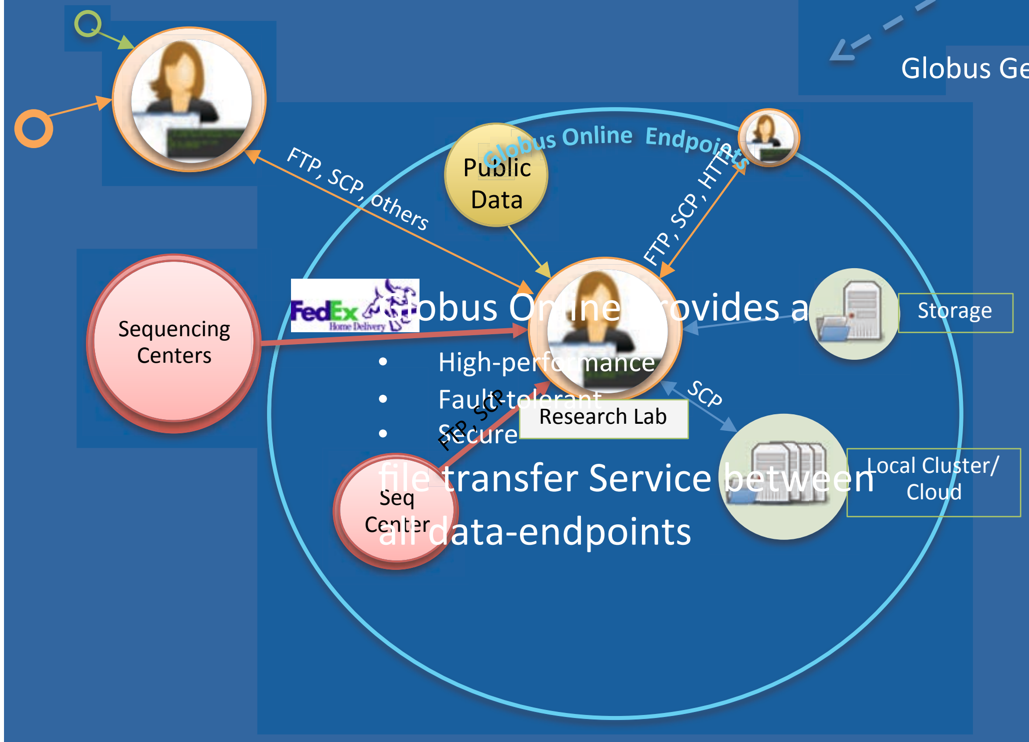
Challenges in Scaling Up

- Doing it right once
- Doing it again for the same dataset or a new dataset
- Reproducing the results
- Sharing results, process
- Publishing
- Economics
- Expertise

Globus Genomics

Globus Genomics

Galaxy Based Workflow Management System



Globus Online provides a

- High-performance
- Fault-tolerant
- Secure

file transfer Service between all data-endpoints

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 Management

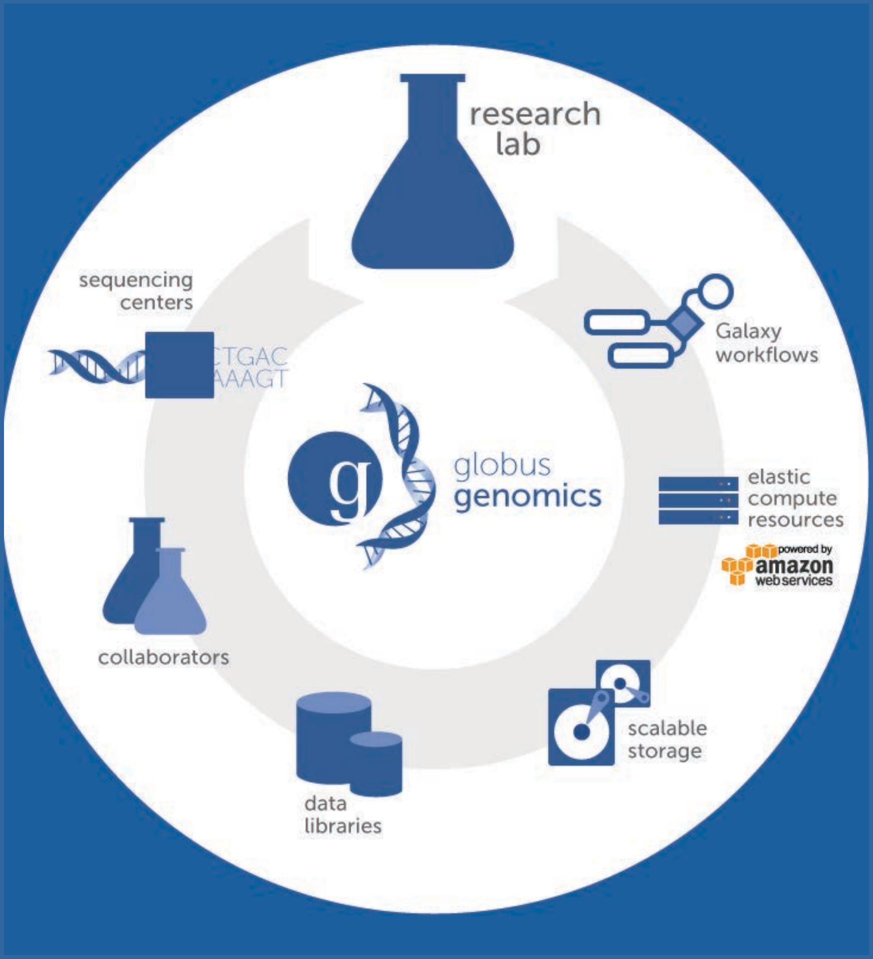
Data Analysis

Globus Genomics

- Workflows can be easily defined and automated with integrated Galaxy Platform capabilities

- Data movement is streamlined with integrated Globus file-transfer functionality

- Resources can be provisioned on-demand with Amazon Web Services cloud based infrastructure



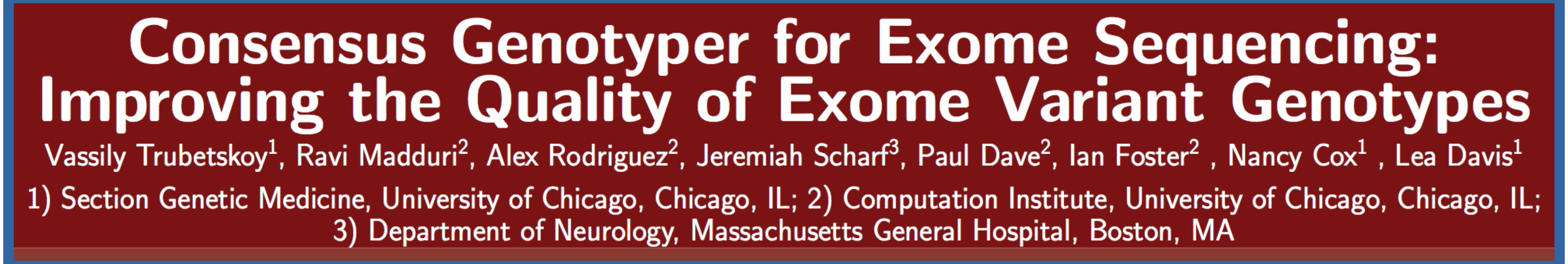
Additional Capabilities

- Professionally managed and supported platform
- Best practice pipelines
  - Whole Genome, Exome, RNA-Seq, ChIP-Seq, …
- Enhanced workbench with breadth of analytic tools
- Technical support and bioinformatics consulting
- Access to pre-integrated end-points for reliable and high-performance data transfer (e.g. Broad Institute, Perkin Elmer, university sequencing centers, etc.)
- Cost-effective solution with subscription-based pricing

Adoption of Globus Genomics

- Individual Research Groups
- Informatics cores at various universities
- Health Care providers
- Sequencing Service Providers

Cox lab, UChicago



Consensus Genotyper for Exome Sequencing: Improving the Quality of Exome Variant Genotypes

Vassily Trubetskoy[1], Ravi Madduri[2], Alex Rodriguez[2], Jeremiah Scharf[3], Paul Dave[2], Ian Foster[2] , Nancy Cox[1] , Lea Davis[1]

1) Section Genetic Medicine, University of Chicago, Chicago, IL; 2) Computation Institute, University of Chicago, Chicago, IL; 3) Department of Neurology, Massachusetts General Hospital, Boston, MA

- 134 samples and 4 workflows
- 4 TB data
- 2200 core hours in 6 days

Olopade lab, UChicago

A profile of inherited predisposition to breast cancer among Nigerian women

Y. Zheng, T. Walsh, F. Yoshimatsu, M. Lee, S. Gulsuner,
S. Casadei, A. Rodriguez, T. Ogundiran, C. Babalola,
O. Ojengbede, D. Sighoko, R. Madduri, M.-C. King, O. Olopade

- 200 targeted exomes
- 200 GB data
- 76,920 core hours in 1.25 days

Innovation Center for Biomedical Informatics - Georgetown

A case study for high throughput analysis of NGS data for translational research using Globus Genomics

D. Sulakhe, A. Rodriguez, K. Bhuvaneshwar, Y. Gusev, R. Madduri, L. Lacinski, U. Dave, I. Foster, S. Madhavan

- 78 exomes from lung cancer study
- 2 TB data
- 125,936 core hours in 1.7 days

Globus Genomics at a glance

30
institutions, groups

2 PBs
raw sequences analyzed

1000s
genomes processed

5 days
longest running workflow

10s
million core hours labs

>1500
analysis tools

>50
workflows

99%
uptime over the past two years

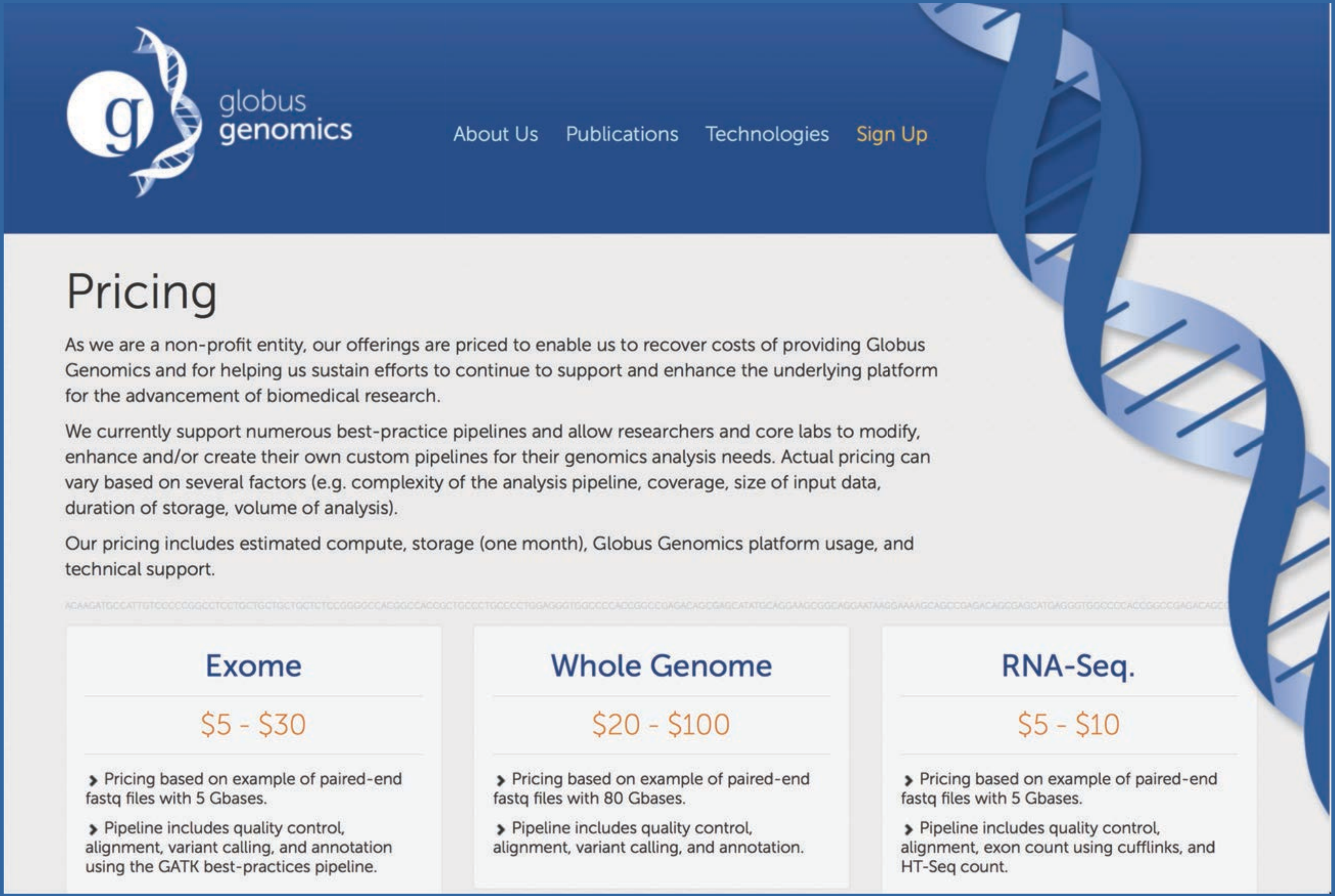
1000s
genomes processed

1 PB
largest single transfer to do

5 days
longest running workflow

100s
different species

Globus Genomics Pricing



Pricing

As we are a non-profit entity, our offerings are priced to enable us to recover costs of providing Globus Genomics and for helping us sustain efforts to continue to support and enhance the underlying platform for the advancement of biomedical research.

We currently support numerous best-practice pipelines and allow researchers and core labs to modify, enhance and/or create their own custom pipelines for their genomics analysis needs. Actual pricing can vary based on several factors (e.g. complexity of the analysis pipeline, coverage, size of input data, duration of storage, volume of analysis).

Our pricing includes estimated compute, storage (one month), Globus Genomics platform usage, and technical support.

Exome

$5 - $30

- Pricing based on example of paired-end fastq files with 5 Gbases.
- Pipeline includes quality control, alignment, variant calling, and annotation using the GATK best-practices pipeline.

Whole Genome

$20 - $100

- Pricing based on example of paired-end fastq files with 80 Gbases.
- Pipeline includes quality control, alignment, variant calling, and annotation.

RNA-Seq.

$5 - $10

- Pricing based on example of paired-end fastq files with 5 Gbases.
- Pipeline includes quality control, alignment, exon count using cufflinks, and HT-Seq count.

Diversity of Collaborations

Seattle Children's
HOSPITAL • RESEARCH • FOUNDATION

LABioMed
Los Angeles
Biomedical
Research Ir
at Harbor-UCLA Me

KU MEDICAL CENTER
The University of Kansas

Washington University in St.Louis

UNIVERSITY OF HAWAI'I
MĀLAMALAMA
1907
UA MAU KE EA O KA 'ĀINA I KA PONO

EuPathDB
Eukaryotic Pathogen Database Resources
http://EuPathDB.org

THE UNIVERSITY OF CHICAGO

Cox Lab
Volchenboum Lab
Olopade Lab

Avera

OHIO STATE
Wexner Medical Center

INOVA®
Join the future of health.

GEORGETOWN UNIVERSITY
1789

M®
UNIVERSITY OF MINNESOTA

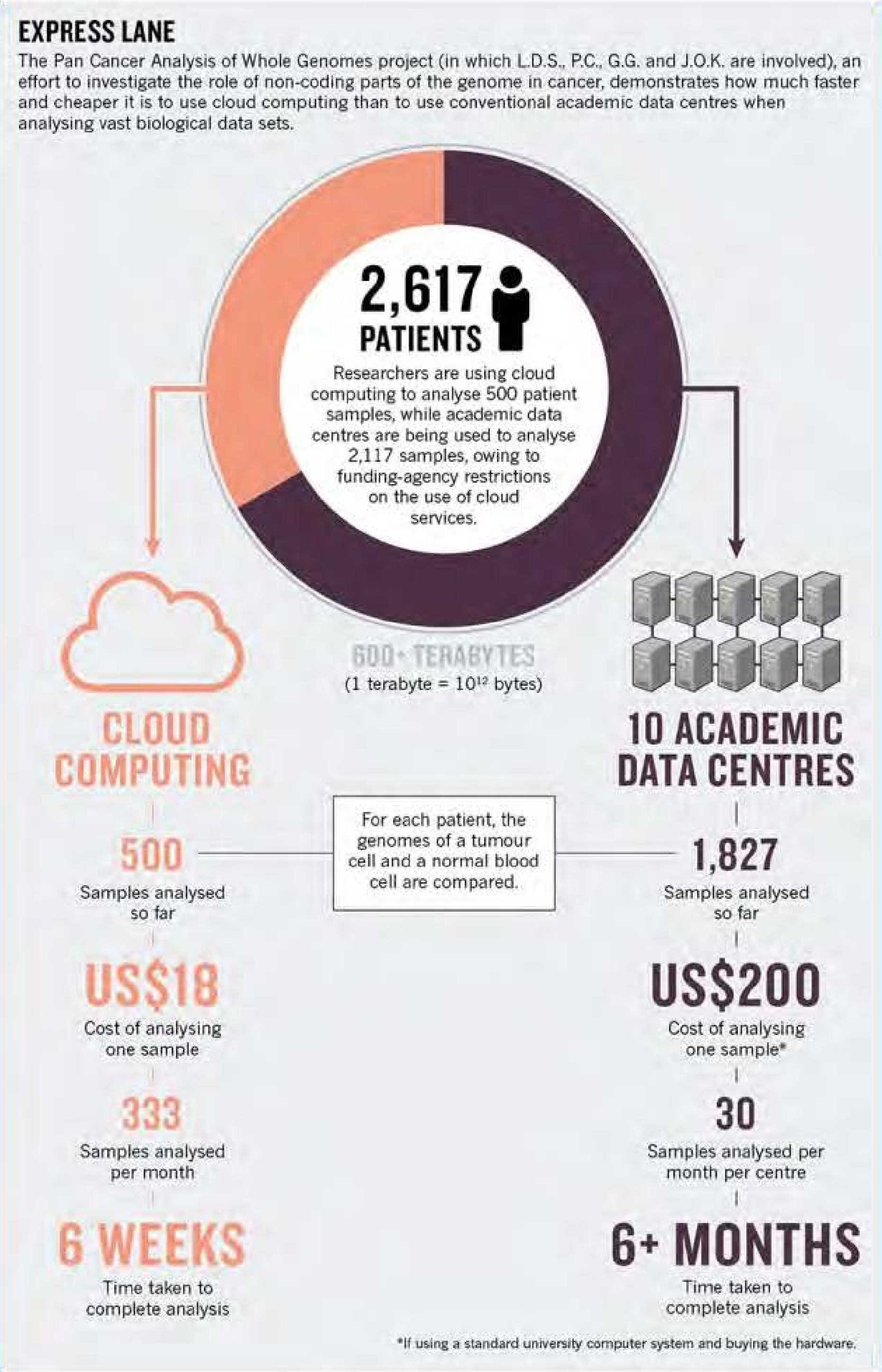
Cal

ngxbio

BU
Genome
Science
Institute
Boston University Medical Center

PerkinElmer®
For the Better

CS
CEDARS-SINAI MEDICAL CENTER®



EXPRESS LANE

The Pan Cancer Analysis of Whole Genomes project (in which L.D.S., P.C., G.G. and J.O.K. are involved), an effort to investigate the role of non-coding parts of the genome in cancer, demonstrates how much faster and cheaper it is to use cloud computing than to use conventional academic data centres when analysing vast biological data sets.

2,617 PATIENTS

Researchers are using cloud computing to analyse 500 patient samples, while academic data centres are being used to analyse 2,117 samples, owing to funding-agency restrictions on the use of cloud services.

600+ TERABYTES
(1 terabyte = 10^{12} bytes)

For each patient, the genomes of a tumour cell and a normal blood cell are compared.

CLOUD COMPUTING

500
Samples analysed so far

US$18
Cost of analysing one sample

333
Samples analysed per month

6 WEEKS
Time taken to complete analysis

10 ACADEMIC DATA CENTRES

1,827
Samples analysed so far

US$200
Cost of analysing one sample*

30
Samples analysed per month per centre

6+ MONTHS
Time taken to complete analysis

*If using a standard university computer system and buying the hardware.

- More information on Globus Genomics and to sign up for a free trial : www.globus.org/genomics

- More information on Globus: www.globus.org

Thank you to our sponsors!

U.S. DEPARTMENT OF ENERGY

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA ·

NSF

NATIONAL INSTITUTES OF HEALTH

THE UNIVERSITY OF CHICAGO

Crescat scientia vita excolatur

Argonne NATIONAL LABORATORY

powered by amazon web services