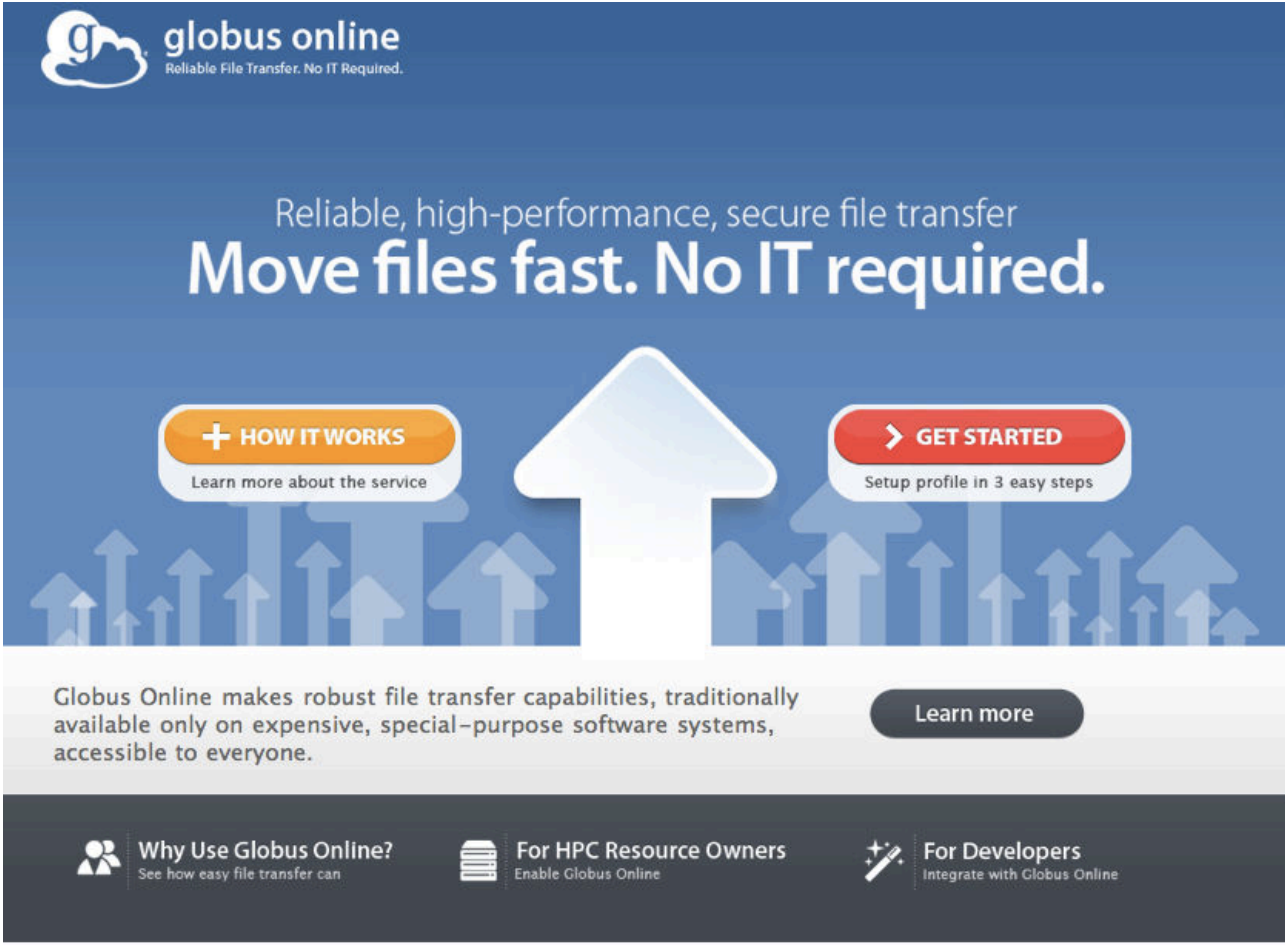
globus online®

The Galaxy Project and Globus Online

Ravi K Madduri
Argonne National Lab
University of Chicago

Outline

- What is Globus Online?
- Globus Online and Sequencing Centers
- What is Galaxy?
- Integrating Galaxy and Globus Online
  - Why ?
  - How ?
- Demo
- Future Work
- Q & A



globus online
Reliable File Transfer. No IT Required.

Reliable, high-performance, secure file transfer

Move files fast. No IT required.

+ HOW IT WORKS
Learn more about the service
› GET STARTED
Setup profile in 3 easy steps

Globus Online makes robust file transfer capabilities, traditionally available only on expensive, special-purpose software systems, accessible to everyone.

Learn more
Why Use Globus Online?
See how easy file transfer can
For HPC Resource Owners
Enable Globus Online
For Developers
Integrate with Globus Online

Benefits of Globus Online

- Reliable file transfer.
  - Easy "fire and forget" file transfers
  - Automatic fault recovery
  - High performance
  - Across multiple security domains
- No IT required.
  - No client software installation
  - New features automatically available
  - Consolidated support and troubleshooting
  - Works with existing GridFTP servers
  - Globus Connect solves "last mile problem"

"I moved 400 GB of files and didn't even have to think about it."

– *Lawrence Berkeley National Lab*

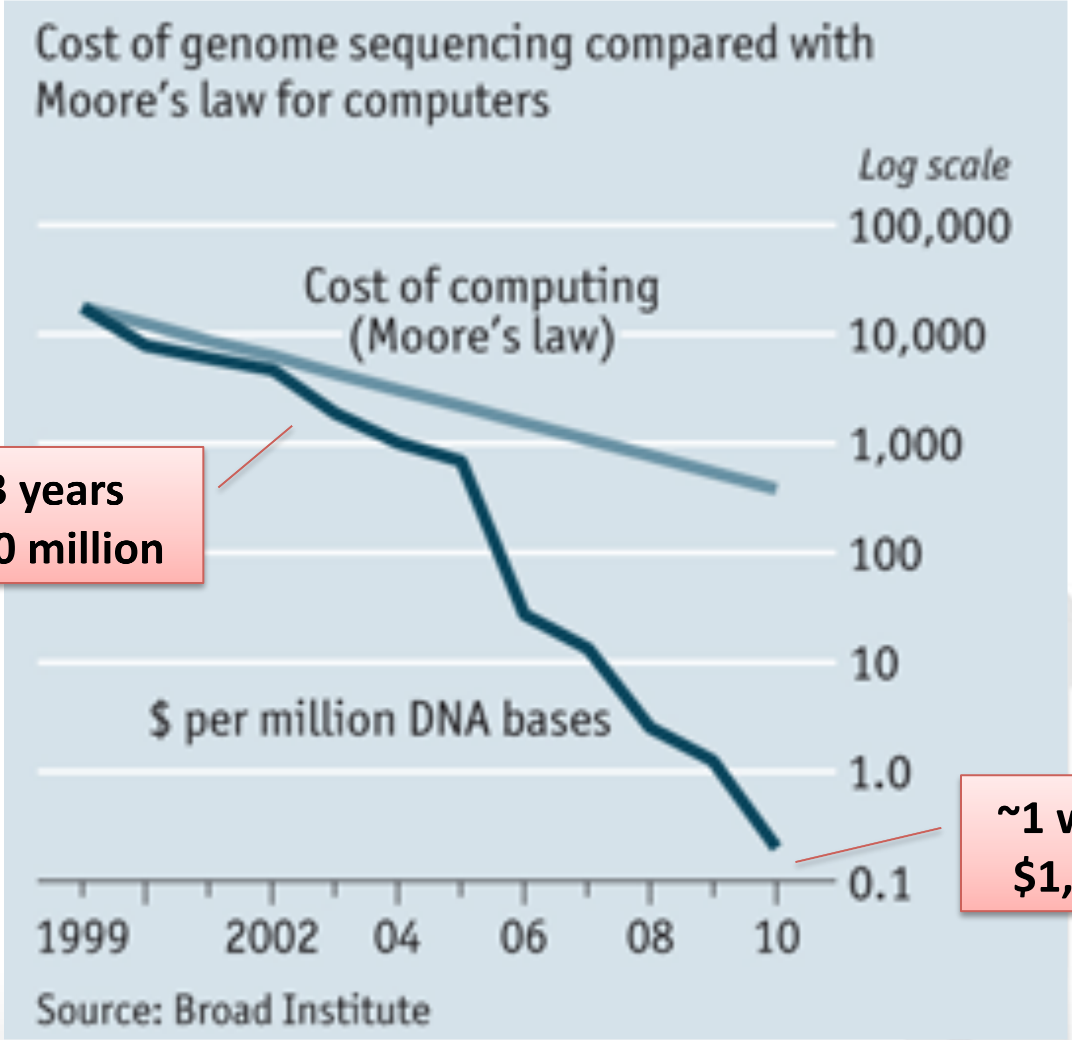
"It's just not a big deal to move big data anymore."

– *Initiative for Biomedical Informatics*

"Fantastic! I have started using globus connect to transfer data, and it only took me 5 minutes to set up. Thank you!"

– *NERSC user*

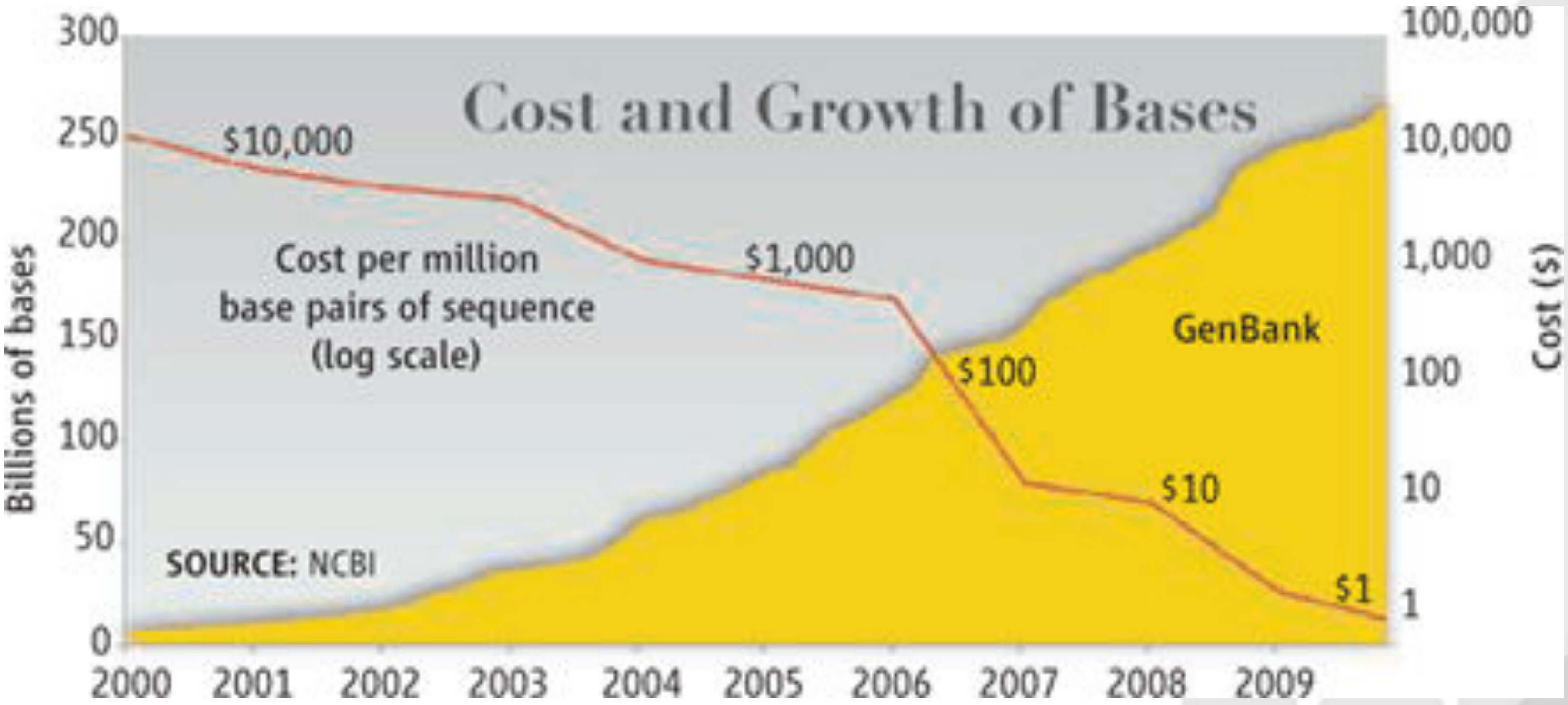
DNA Sequencing Costs are Declining



13 years
$300 million

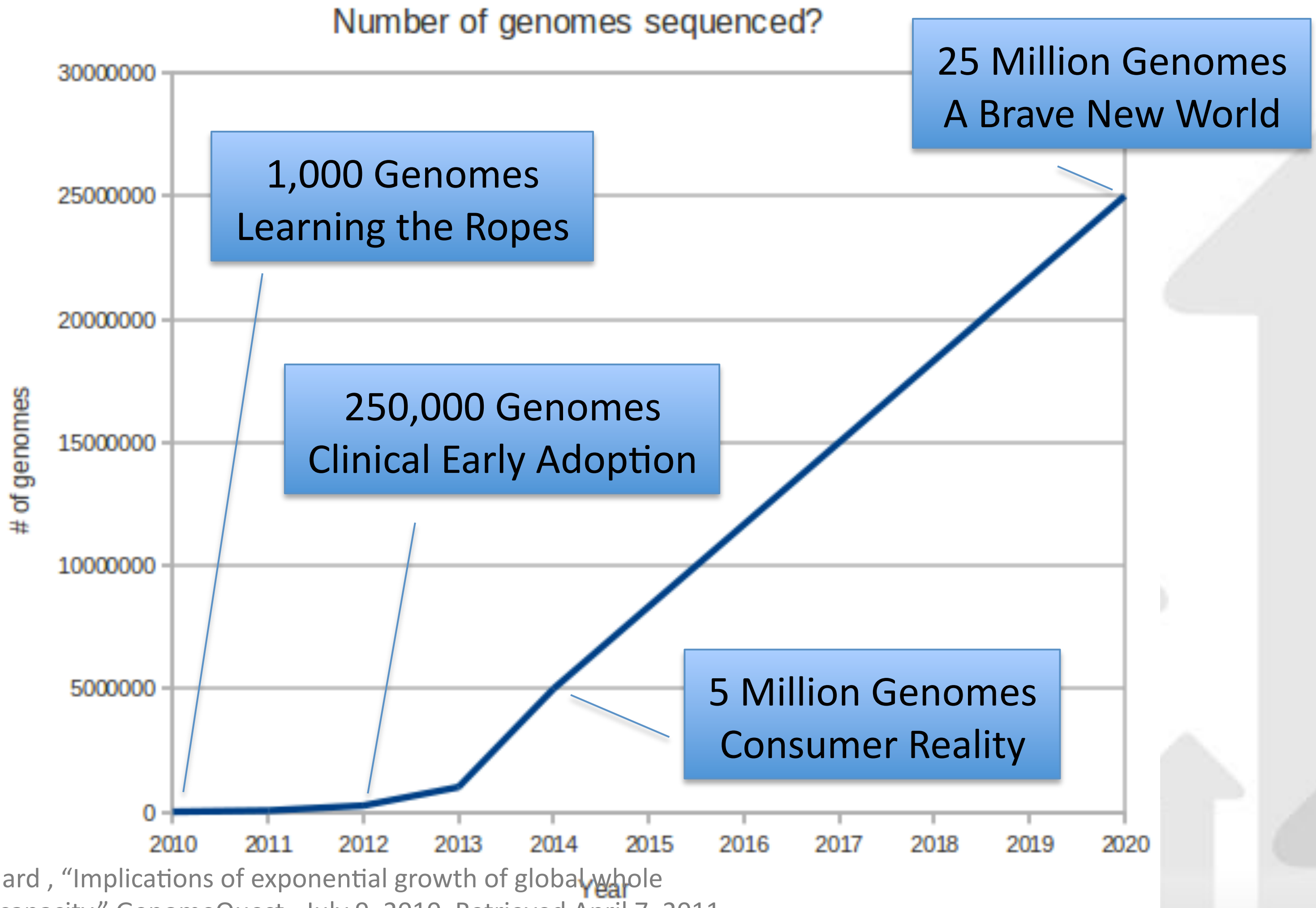
~1 week
$1,000

Sequencing Data is Increasing



Source: Pennisi, E., Science 2011 **331**:6018 pp. 666

Prediction of Number of Sequenced Genomes



1,000 Genomes
Learning the Ropes

250,000 Genomes
Clinical Early Adoption

5 Million Genomes
Consumer Reality

25 Million Genomes
A Brave New World

Source: Resnick, Richard , “Implications of exponential growth of global whole genome sequencing capacity.” GenomeQuest. July 9, 2010. Retrieved April 7, 2011.

Current Computing Infrastructure Won't Support Increasing Data

Will Computers Crash Genomics?

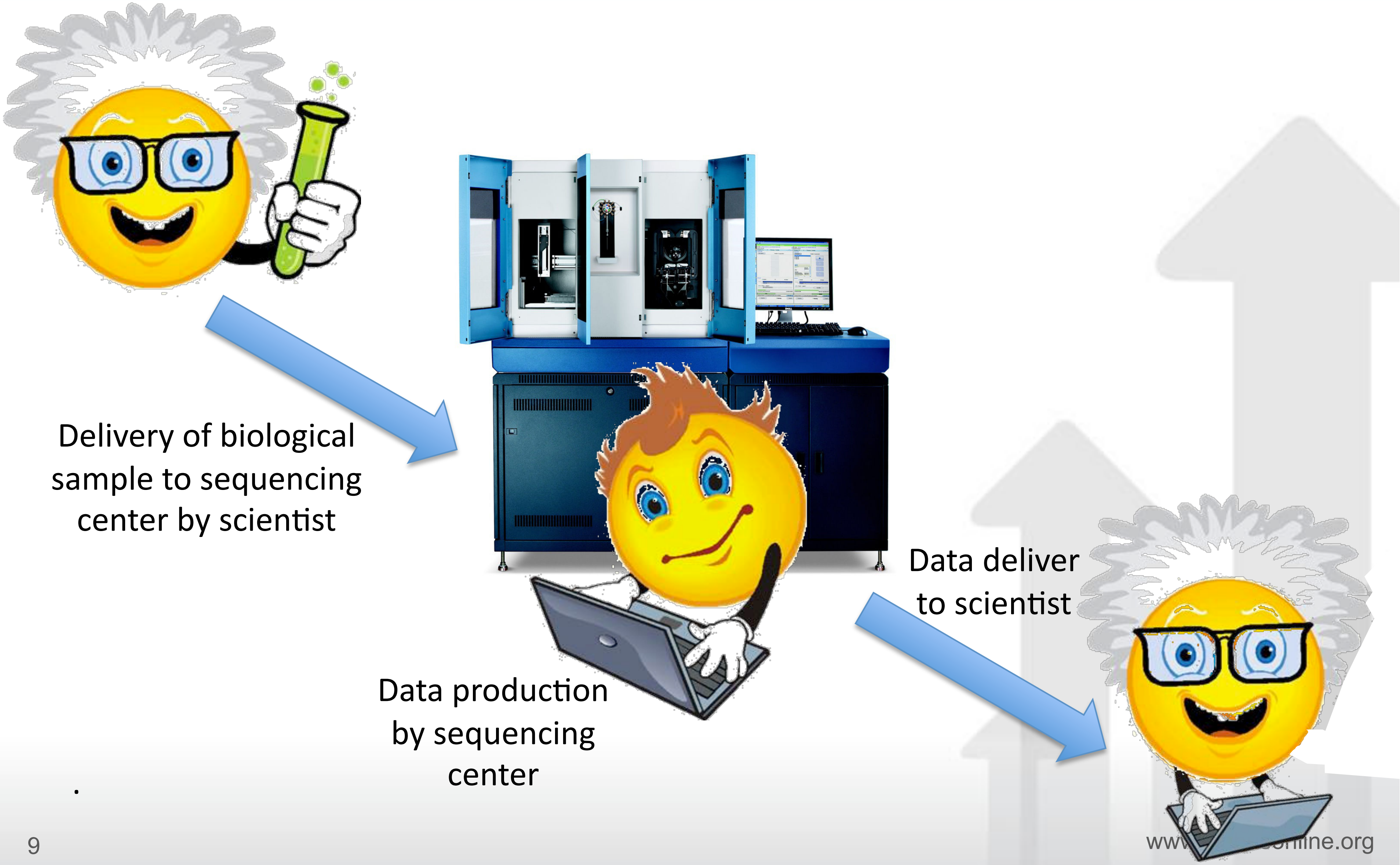
Pennisi, E., Science 2011 **331**:6018 pp. 666

CREDIT: ALVARO ARTEAGA/ALVAREJO.COM

"The various members of the genome informatics ecosystem are now facing a **potential tsunami of genome data** that will swamp our storage systems and crush our compute clusters."

Stein, L.D., *Genome Biology* 2010 **11**:5 pp. 207

Sequencing Center Workflow



Delivery of biological sample to sequencing center by scientist

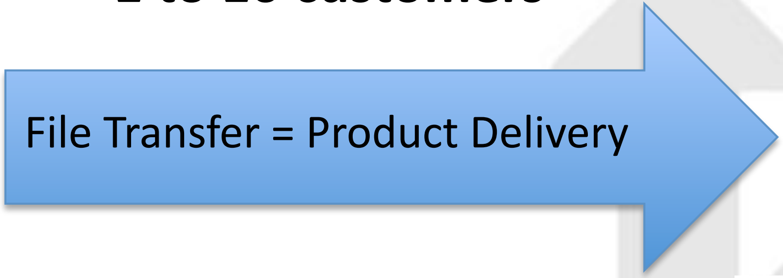
Data production by sequencing center

Data deliver to scientist

.

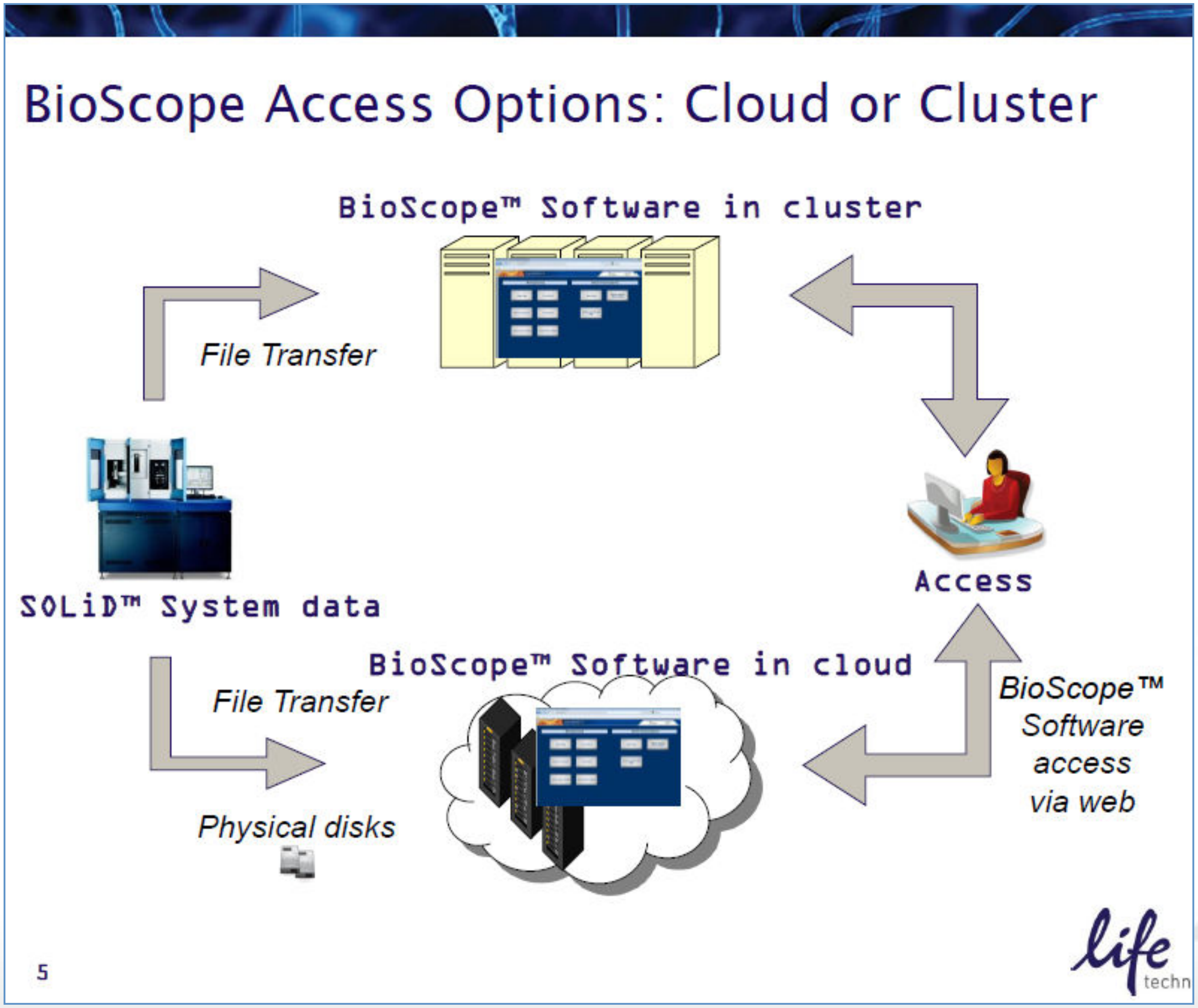
ABI SOLiD Sequencing Instrument

- 1.2 TB of data every 7 to 14 days
- 2 to 1600 files
- 1 to 16 customers



File Transfer = Product Delivery

Current File Transfer Methods



BioScope Access Options: Cloud or Cluster

BioScope™ Software in cluster

File Transfer

SOLiD™ System data

Access

BioScope™ Software in cloud

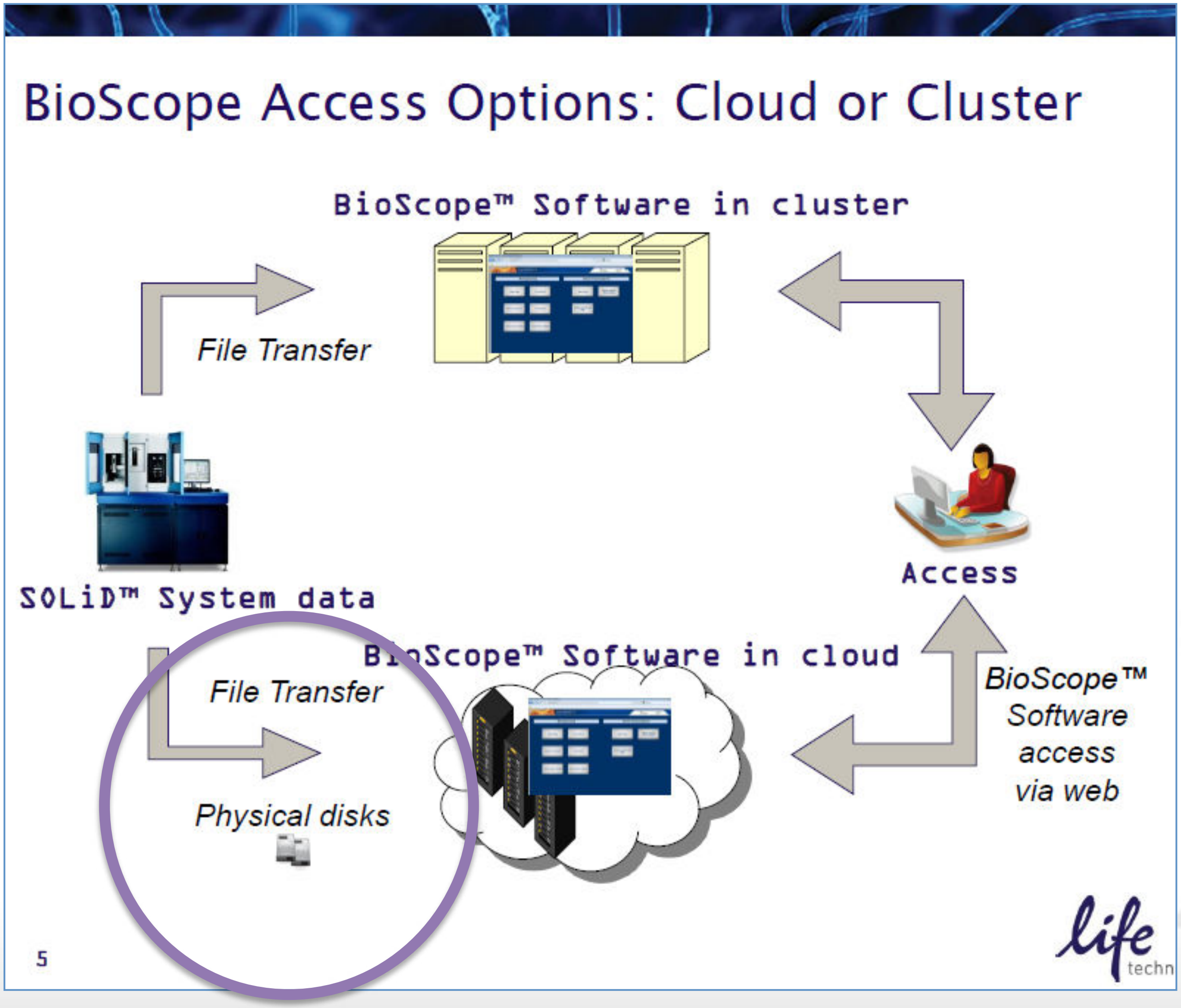
File Transfer

Physical disks

BioScope™ Software access via web

5

Current File Transfer Methods



BioScope Access Options: Cloud or Cluster

BioScope™ Software in cluster

File Transfer

SOLiD™ System data

Access

BioScope™ Software in cloud

File Transfer

Physical disks

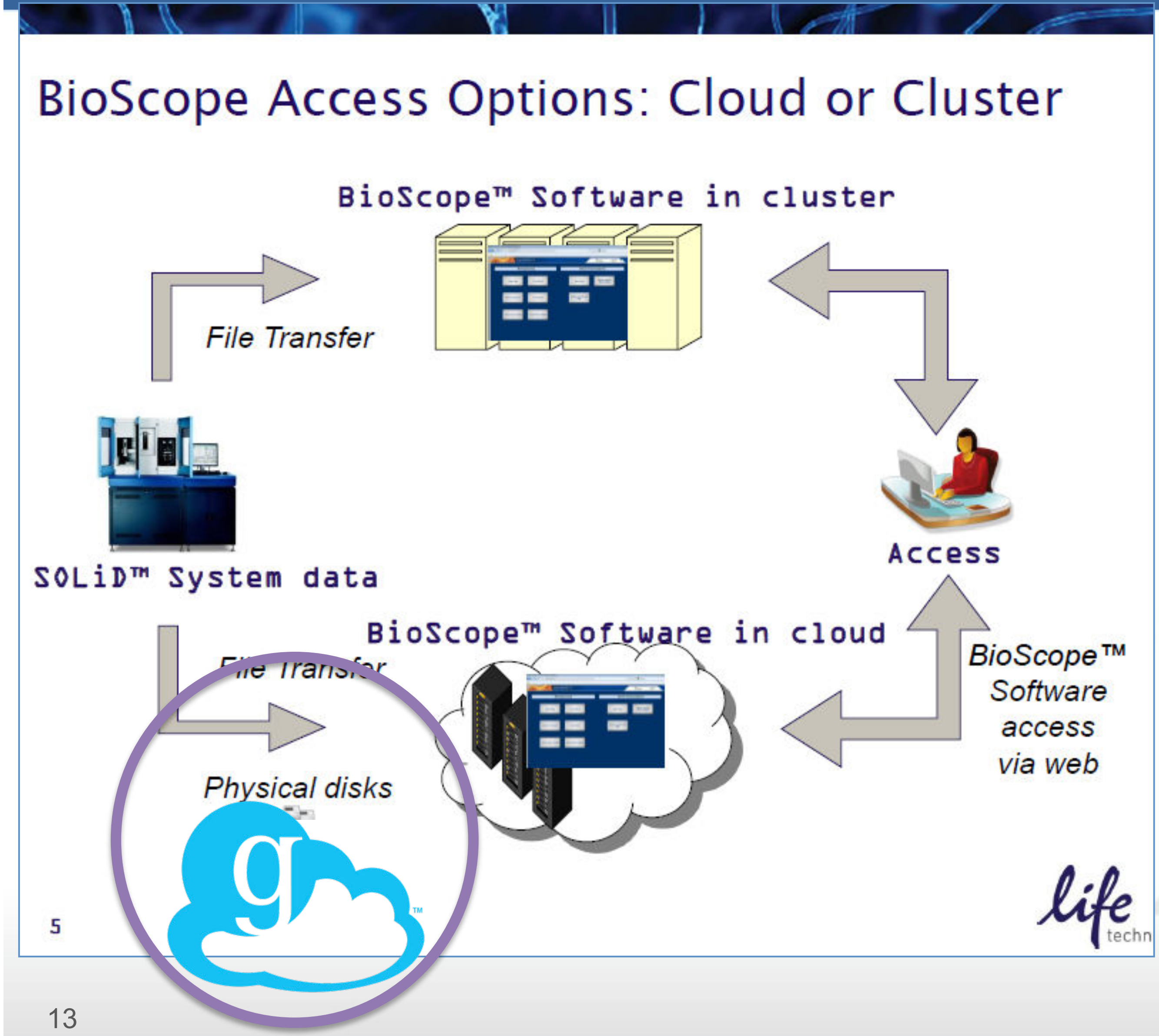
BioScope™ Software access via web

5

"…providers and their customers often resort to the "sneaker net": **overnight shipment of data-laden hard drives**."

Pennisi, E., Science 2011 **331**:6018 pp. 666

Globus Online Improves Current Methods



BioScope Access Options: Cloud or Cluster

BioScope™ Software in cluster

File Transfer

SOLiD™ System data

Access

BioScope™ Software in cloud

File Transfer

Physical disks

BioScope™ Software access via web

5

life techn

"…providers and their customers often resort to the "sneaker net": **overnight shipment of data-laden hard drives**."

Pennisi, E., Science 2011 **331**:6018 pp. 666

Why Globus Online?

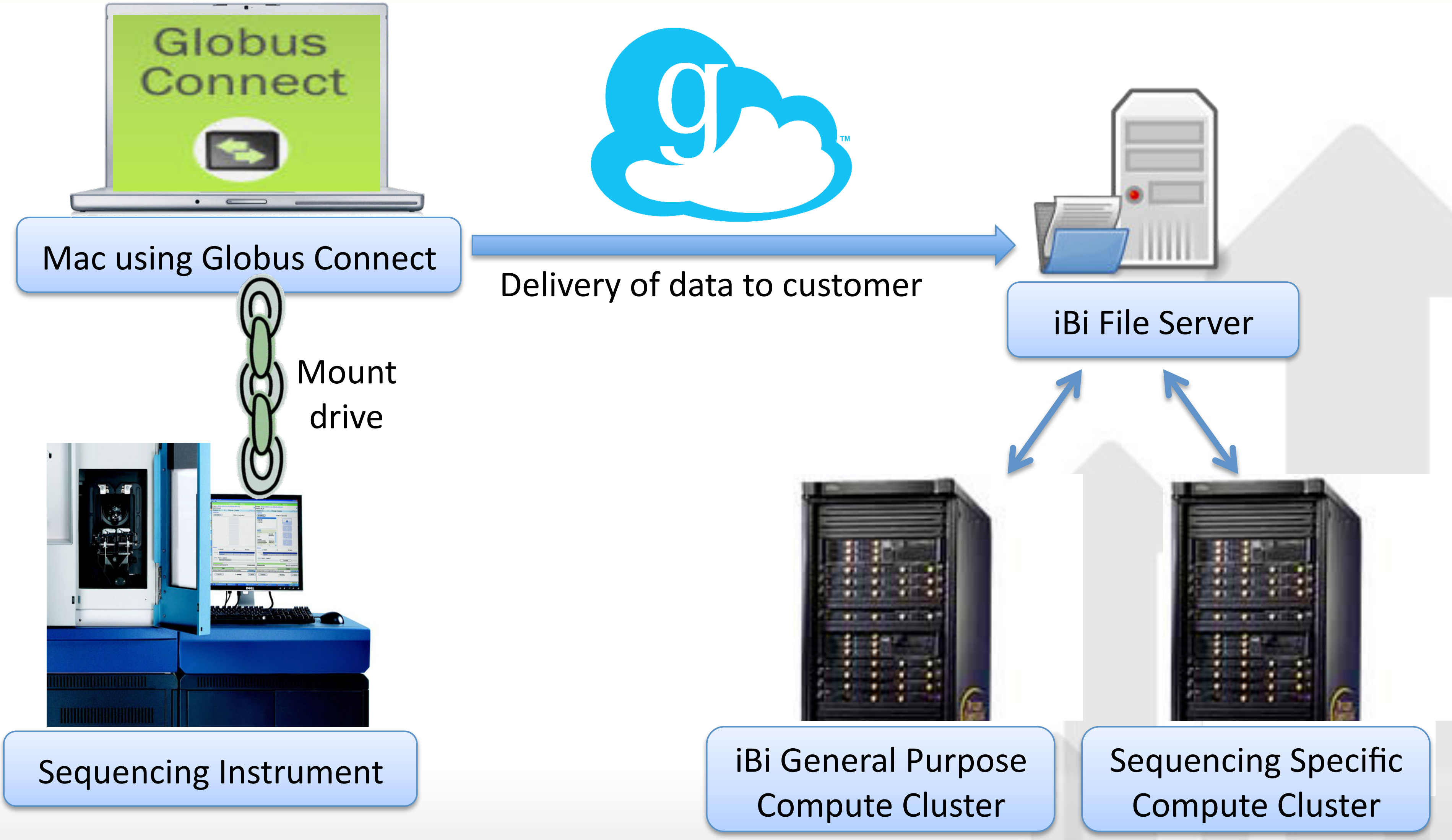
- Fire-and-forget usage
  - Re-trying failed transfers
  - Logs to identify reasons behind failed transfers
- Simplicity
  - Simple logon and authentication
  - Web interface for execution and monitoring
  - Globus Connect for sequencing facility endpoint
- Reliability
  - Checksum option (~$10K/genome)
- Performance
- Secure enablement

"Now, with Globus Online, the process is trivial and our scientists can move data to the right location with just a few clicks."

"File size is no longer a barrier to productivity."

– *Initiative for Biomedical Informatics*

Globus Online at UC Sequencing Facility



Mac using Globus Connect

Delivery of data to customer

iBi File Server

Mount drive

Sequencing Instrument

iBi General Purpose Compute Cluster

Sequencing Specific Compute Cluster

Extend to Other Sequencing Centers

Over 300 public sequencing centers…

Extend Beyond Sequencing Centers

- Future Use Case = Biologists
  - "Moreover, as so-called third generation machines—which promise even cheaper, faster production of DNA sequences (*Science*, 5 March 2010, p. 1190)—become available, **more, and smaller, labs will start genome projects of their own**." Pennisi, E., Science 2011 **331**:6018 pp. 666.

What about Analysis of this Data ?

Enter Galaxy

- A **free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes it easy to integrate your own tools and data and customize your own site
- **Flexible architecture** -> Customizable

Galaxy Adoption

- ~50 deployments of Galaxy
  - Galaxy for MicroArray analysis, Machine Learning, Drug Discovery etc
- ~130,000 jobs a month and growing on the public instance of Galaxy
- 1 TB/week in user uploads
  - 60TB from China
- 150+ attendees in the Galaxy users conference
  - From 6 continents

Globus Online and Galaxy

- Transferring large quantities of data in and out of Galaxy reliably
  - Downloading results to user’s laptop
- Running Galaxy in Production for multiple users
  - Running analysis on a cluster in the user space
- Federated Identity management
- Flexible group management system (VO)
  - Mechanisms to share data and workflows
- Parallel execution
  - Condor and Swift

What works

- Identity Management
  - Used Globus Provision to bootstrap the provisioned ec2 cluster with GO credentials so one can put data from GO-enabled endpoints into galaxy and script transferring data from galaxy to any GO-enabled endpoint
- Data Management
  - Added a GO Transfer task to Galaxy
  - Ability to script the transfer tasks along with the workflow
- Execution Management
  - We know more about this than anything
  - Condor runner for galaxy: The executable along with command line options get passed to a condor runner (runner is galaxy term for execution mechanism)
- Group Management
- Metadata Management

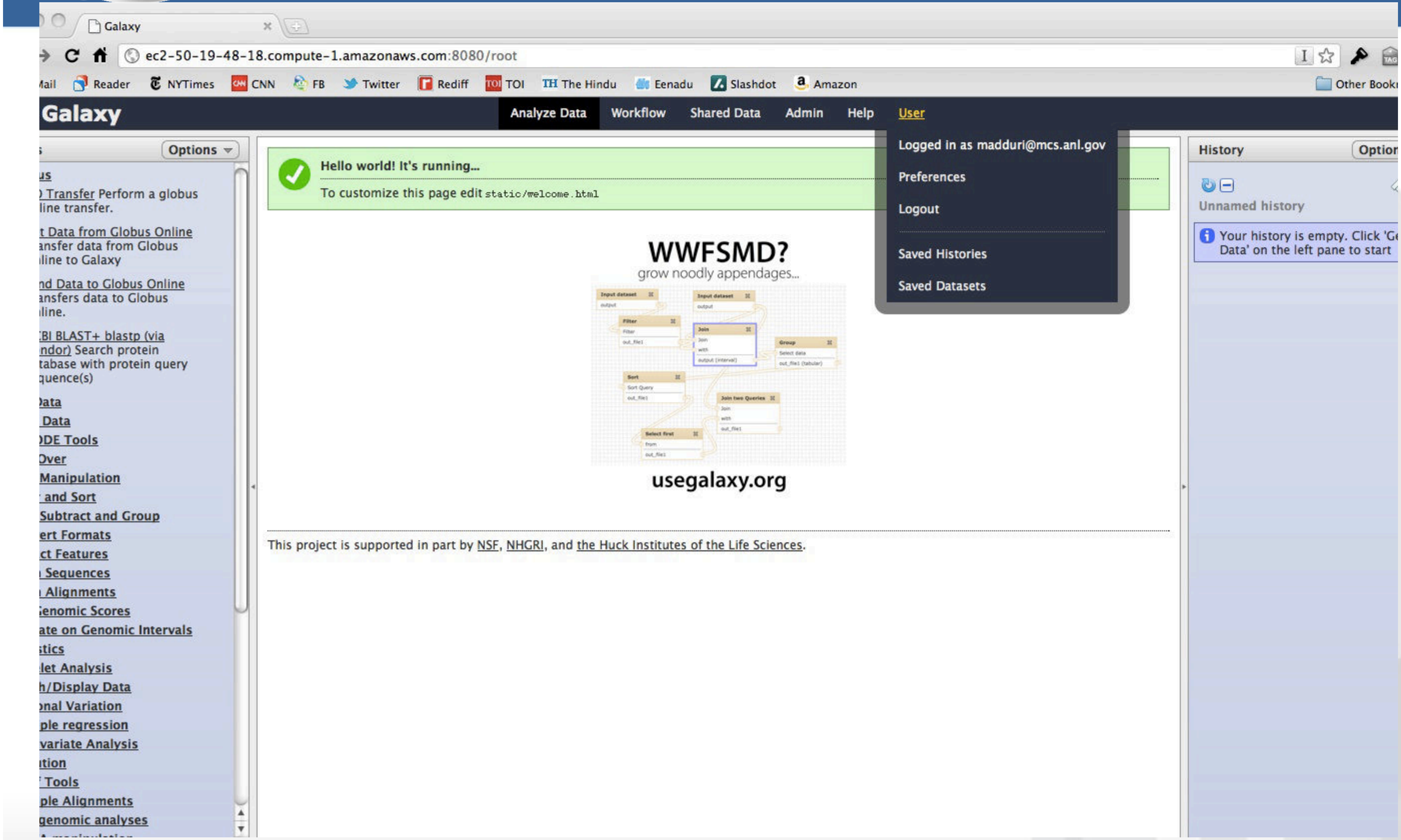
Automated deployment of Galaxy clusters on EC2

- Uses Globus Provision
- An NFS server providing global directories for home directories, software, and scratch space.
- An NIS server providing an authentication domain within the cluster. The list of users must be specified when the cluster is created
- A GridFTP server, which is automatically registered as a GO endpoint.
- A Condor pool. The number of worker nodes is fixed at the time the cluster is created.
- A Galaxy server with the modifications outlined above.
- Each user in the cluster has a user certificate signed by the Globus Provision CA (this CA is trusted by GO). If a GO username matches a cluster username, and that user's Globus Provision certificate is uploaded to GO as a trusted certificate, then the user will be able to transfer files to/from the cluster's endpoint (and will also be able to use Galaxy's GO transfer tasks; this further requires that the user's "public name" in Galaxy matches his GO and cluster username).

GO Galaxy Description

- The deployment of clusters on EC2 is completely automated, and we can deploy fully-functional Galaxy clusters on EC2 within minutes
- We created galaxy runners for Condor and Swift
  - So the applications are run on a worker node instead of the galaxy node
- Reusable CHEF recipes to provision production galaxy instances on demand on EC2(-ish) clusters

Integration of Globus Online with Galaxy



Logged in as madduri@mcs.anl.gov

Hello world! It's running...

To customize this page edit `static/welcome.html`

Your history is empty. Click 'Ge Data' on the left pane to start

This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

Demo of Current Capabilities

Globus Online: Part of the Computing Infrastructure to Support Genomics

CREDIT: ALVARO ARTEAGA/ALVAREJO.COM

Globus Online: Part of the Computing Infrastructure to Support Genomics

CREDIT: ALVARO ARTEAGA/ALVAREJO.COM

GO Galaxy Future Capabilities

- Integrate flexible Globus Online identity management, group management with Galaxy
- Ability to create a VO with the above capabilities (User/Groups management, File system, Transfer, HTC, Parallel execution, EC2 Cluster, data sharing) with click of a button
- Integrate with Globus Online on-demand storage solution
- Ability to easily share data, workflows within a Virtual Organization

References

- More details about Globus Online can be found here: http://www.globusonline.org
- Details on Galaxy: http://usegalaxy.org
- Details on work we did integrating Globus Online and Galaxy: bit.ly/qDOHhW

globus online

Thanks !

Questions?

(or later: rm@anl.gov, @madduri on twitter)