

Computation Institute

# Galaxy and Globus Science as a Service

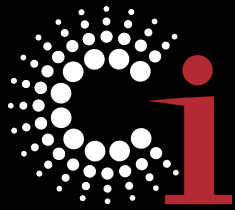
Ravi K Madduri, University of Chicago and Argonne  
National Laboratory

[madduri@uchicago.edu](mailto:madduri@uchicago.edu)

@madduri



[globus.org/genomics](http://globus.org/genomics)



Our vision for a 21st century  
discovery infrastructure

Provide **more** capability for  
**more** people at **lower cost** by  
delivering “**Science as a service**”

[www.globus.org](http://www.globus.org)



# An Old Idea: Service Oriented Science



# Service-Oriented Science

People **create** services (data or functions) ...  
which I **discover** (& decide whether to use) ...  
& **compose** to create a new function ...  
& then **publish** as a new service.

→ I find “someone else” to **host** services,  
so I don’t have to become an expert in  
operating services & computers!

→ I hope that this “someone else” can  
**manage** security, reliability, scalability, ...

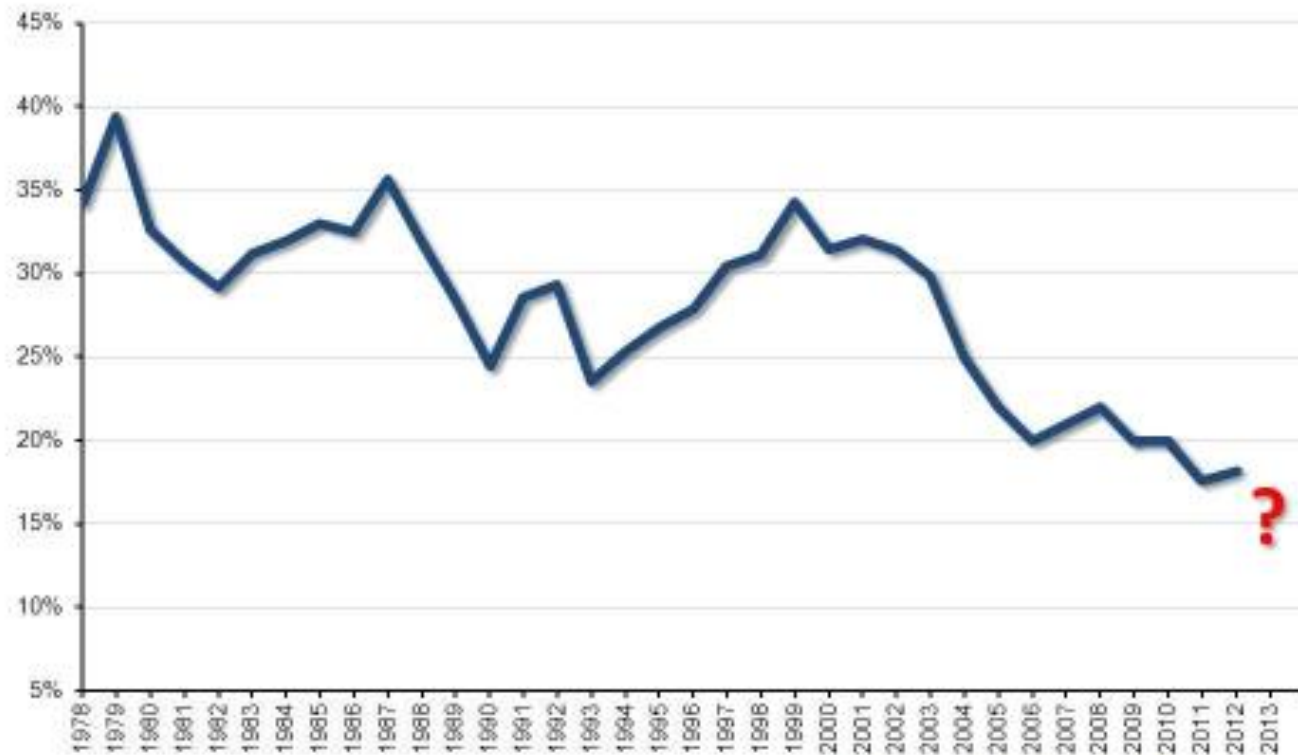




# Two Broader Themes

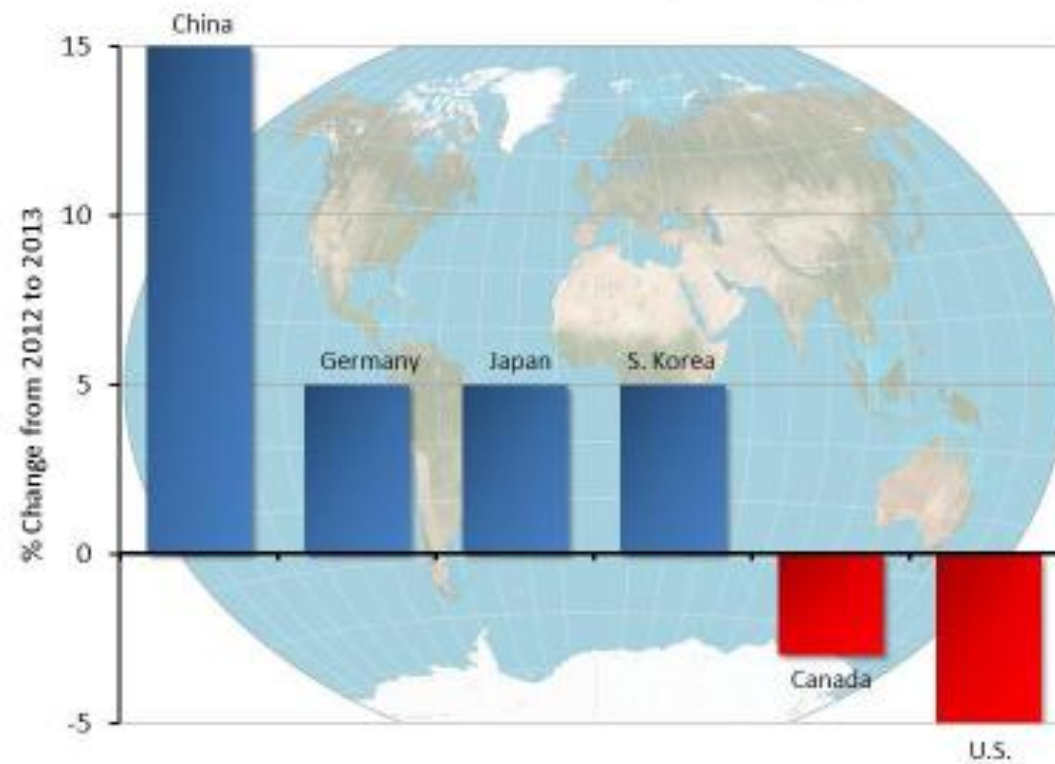
- **Productivity** of Researchers
  - Time spent performing administrative tasks  
Vs time spent doing science
  - **Reproducibility**
- **Sustainability** of scientific software
  - Reduction in funding for science

## NIH Grant Application Success Rates FY 1978-2013

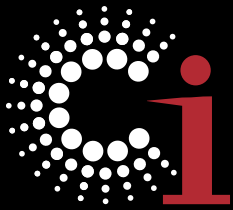


Source: NIH [http://report.nih.gov/success\\_rates/](http://report.nih.gov/success_rates/)

## Scientific R&D Spending



Source: *Cell*, 2013 Jul 3;154(1):16-9.



# Time-consuming tasks in science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature
- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports





# Our Science Stack

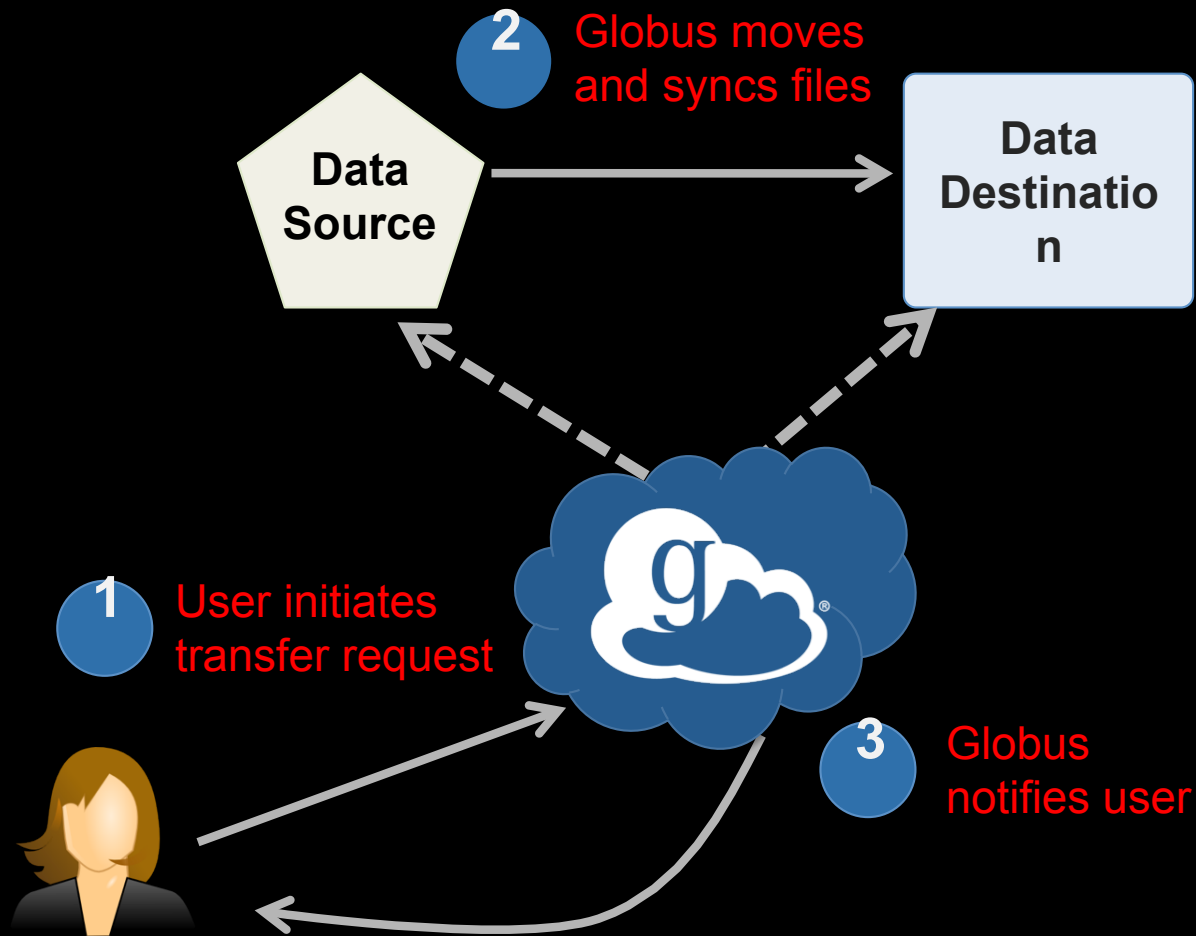
- Galaxy
  - Interactive execution
  - Creation, Execution, Sharing, Discovering Workflows
- Globus
  - Data management
  - Identity Management
- AWS
  - EC2, EBS, S3, SNS,
  - Spot, Route 53, Cloud Formation

**SaaS**

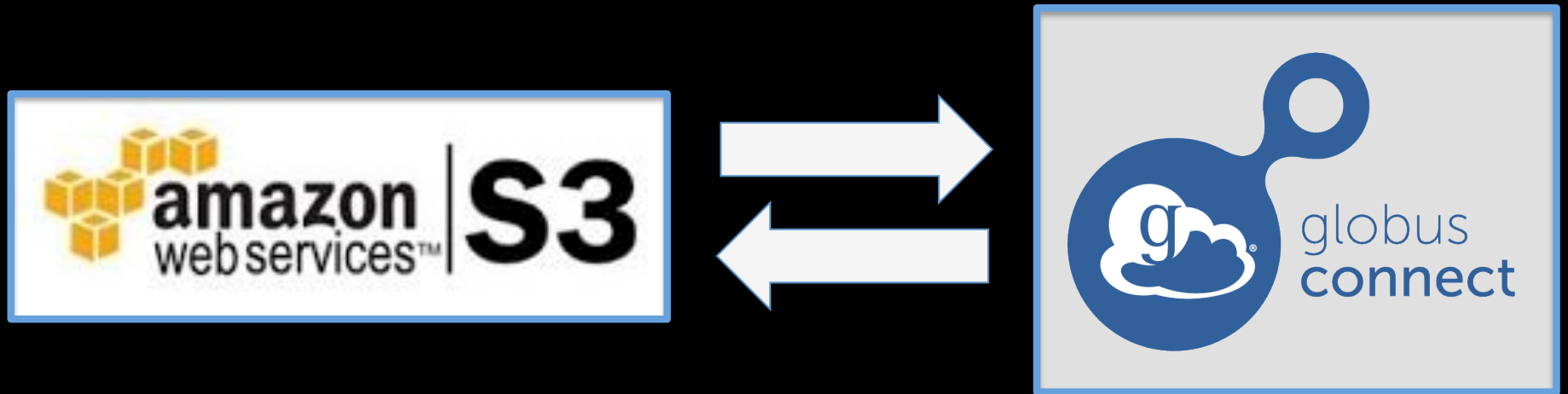
**PaaS**

**IaaS**

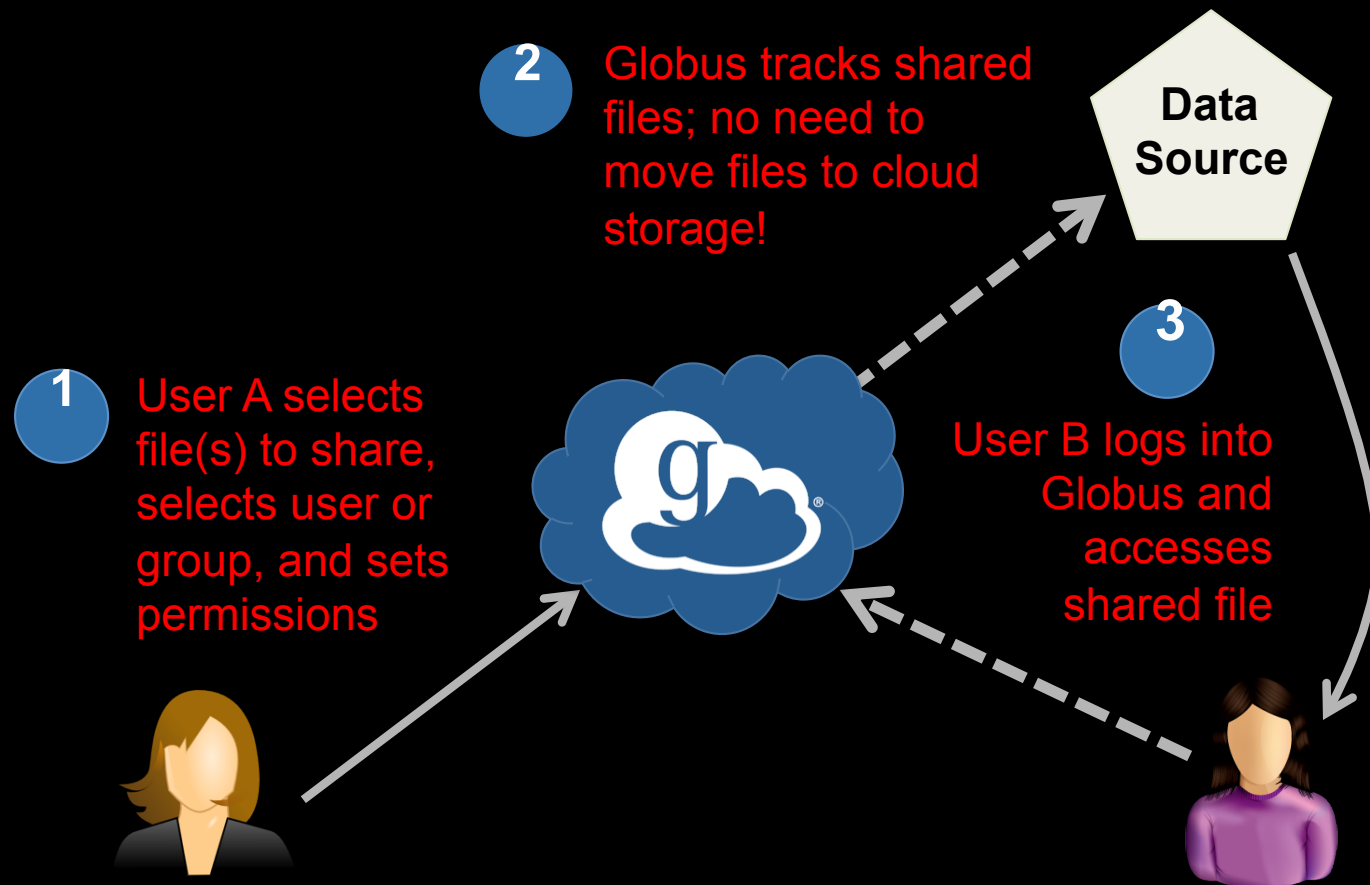
# Globus: Fast, reliable data transfer



# Amazon S3 Endpoints




# Globus: Sharing off existing systems



# Globus: Federated identity

InCommon.



 THE UNIVERSITY OF CHICAGO

---

### Sign In

You are logging in to: `classes.uchicago.edu`  
A Web-Single-Signon protected site

CNetID:

☐ Hospital Employee?


Password:


[Forgot your password?](#)

Signing in allows you to access multiple University of Chicago web applications while entering your CNetID and password only once. To end your session, simply close your browser.

**Questions?** Contact the IT Services Service Desk by phone at 2-5800 (773-702-5800), via email at [itservices@uchicago.edu](mailto:itservices@uchicago.edu), or get walk-in help at the TECHB@R on the first floor of Regenstein Library during reference desk hours <http://hours.lib.uchicago.edu/>.

**Alumni** account holders may contact [alumni-support@uchicago.edu](mailto:alumni-support@uchicago.edu) or call 1-877-292-3945 between 9 AM and 3 PM CST with any questions.

Authentication powered by Shibboleth™ 

 Cornell University

---

### CUWebLogin

ApplicantID:

Password:

What is this?  
I forgot my password!  
I don't have a NetID, now what?

MyProxy

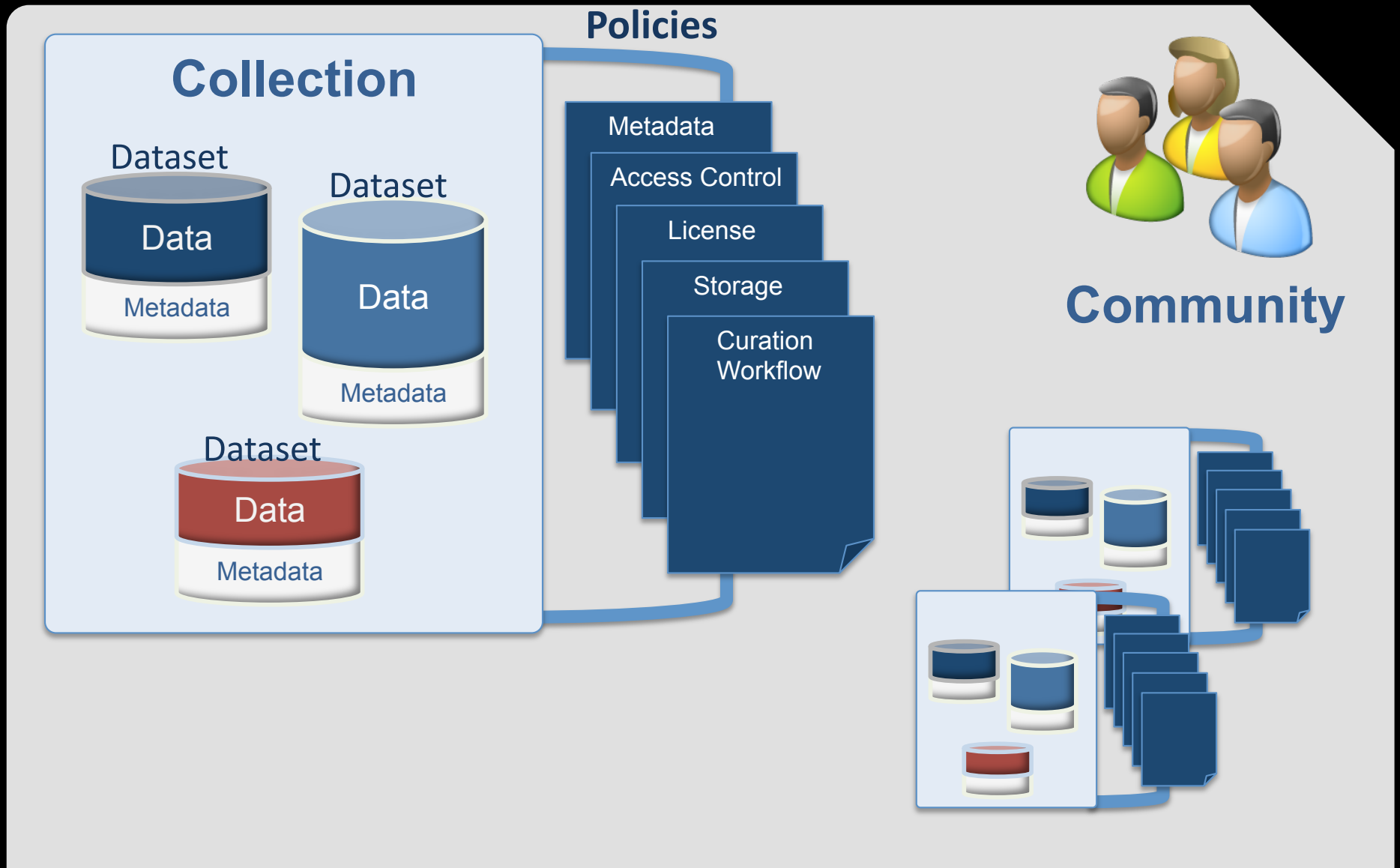


[globus.org/genomics](http://globus.org/genomics)

# Globus Transfer

**>25,000 registered users; >150 daily  
50 PB moved; >1B files  
10x (or better) performance vs. scp  
99.9% availability  
Entirely hosted on Amazon**

# Globus: Data publication service



User Information																
User:	jtxt															
Email:	james@taylorlab.org															
Created:	2013-12-14 19:29:48															
Job Count:	30															
Bytes Transferred:	45644557967406 (45.645 TB)															
Files:	10538680															
Files Skipped:	4175915															
DN:	/DC=org/DC=cilogon/C=US/O=Johns Hopkins/CN=James Taylor A12216															
DN:	/C=US/O=Pittsburgh Supercomputing Center/CN=James Taylor/CN=1717207743															
SSH Public Key:	ssh-dss AAAAB3NzaC1kc3MAAACBAOSZONptSidEU3Uy4B3UvLWwyda4q65bXtK0zOtgzWovh5hQSG/RWCfiUulUGvQeVrIK34iU...															
Endpoints:	<a href="#">Link</a>															

Active Jobs																
Status	User	Req. (UTC)	Tasks	OK	Failed	BytesTX	Mbps	Retrying	Duration	Deadline	Faults	Source	Destination	Type	Flags	Task ID
CONNECT_FAILED	jtxt	06-26 15:24	106488	106439	0	1.745 TB	30.01	49	129:13:57	67:46:30	964	jtxt#rhyolite	vincent#dslogin01	Web	VERIFY	02ead438-fd46-11e3-b575-12

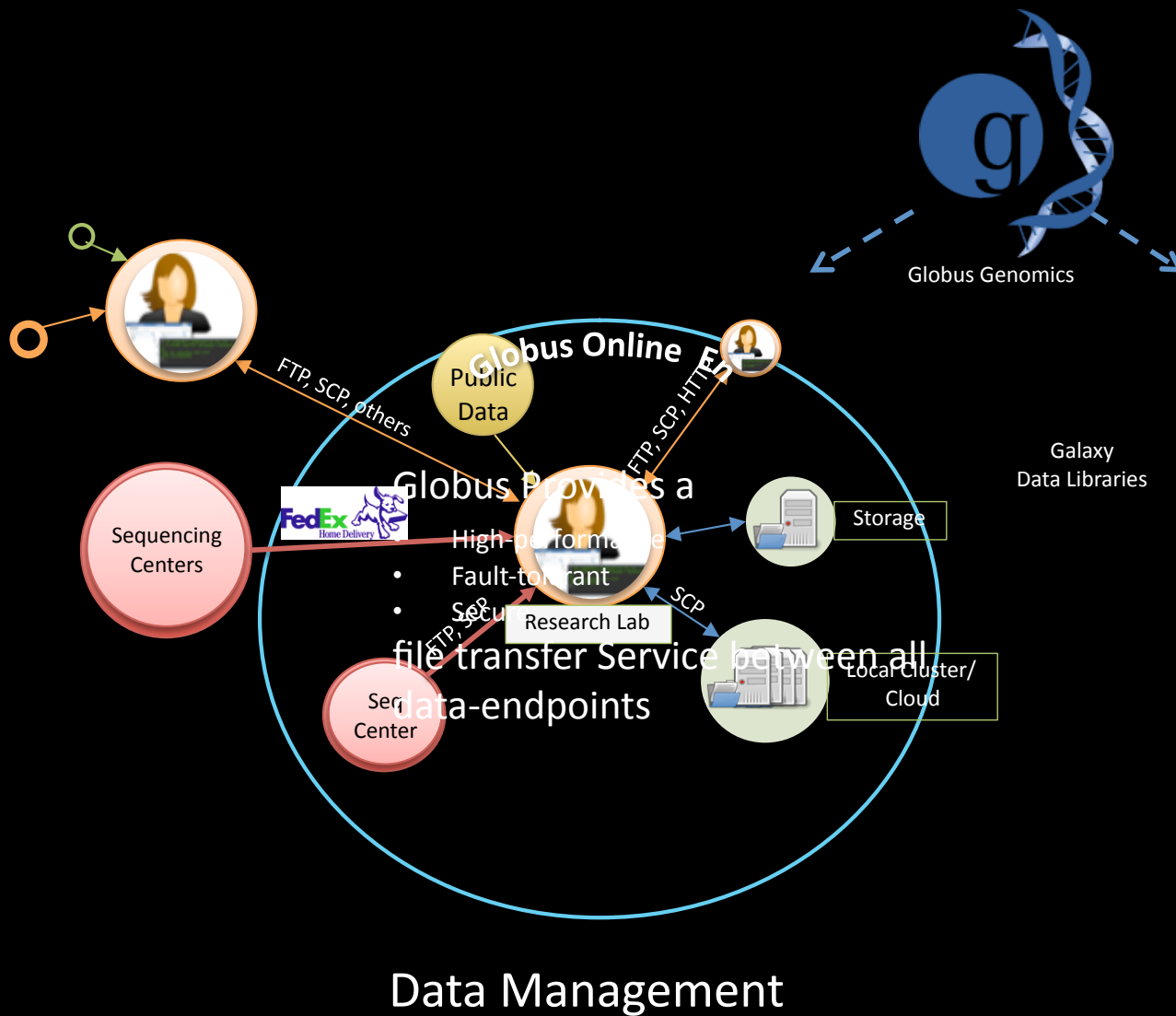
Job History																
User	Completed (UTC)	Tasks	OK	Failed	BytesTX	Mbps	Duration	Faults	Source	Destination	Type	Flags		Task ID		
jtxt	2014-07-01 20:20:36	83469	83469	0	673.296 GB	11.98	124:55:59	822	jtxt#rhyolite	vincent#dslogin01	Web	VERIFY		fbef6bbc-fd45-11e3-b575-123139403		
jtxt	2014-06-26 15:05:07	380204	380204	0	42.262 GB	2.29	41:02:44	370	jtxt#rhyolite	vincent#dslogin01	Web	VERIFY, MTIME		380fa2bc-fbeb-11e3-b574-123139403		
jtxt	2014-06-24 21:42:14	260190	260190	0	7.709 TB	384.15	44:35:41	7	jtxt#rhyolite	vincent#dslogin01	Web	SYNC=3, VERIFY		9dbe3742-fa72-11e3-b570-123139403		
jtxt	2014-06-23 01:03:04	8664	8664	0	547.805 GB	435.54	02:47:42	0	jtxt#rhyolite	vincent#dslogin01	Web	SYNC=3, VERIFY		b3fa0c06-fa5a-11e3-b56d-123139403		
jtxt	2014-06-21 18:21:23	272732	272732	0	1.773 TB	150.46	26:11:06	3	jtxt#rhyolite	vincent#dslogin01	Web	SYNC=3, VERIFY		5e1c6f98-f895-11e3-b56b-123139403		
jtxt	2014-06-20 16:08:34	856561	856561	0	246.727 KB	0.00	04:07:12	0	jtxt#rhyolite	vincent#dslogin01	Web	SYNC=3, VERIFY, MTIME		98e4a9a6-f872-11e3-b56b-123139403		
jtxt	2014-06-20 04:12:10	78312	78312	0	373.587 GB	254.77	03:15:31	2	jtxt#rhyolite	vincent#dslogin01	Web	SYNC=3, VERIFY, MTIME		bc5c18b8-f815-11e3-b56b-123139403		
jtxt	2014-06-20 00:56:23	1207605	996956	210649	6.915 TB	315.39	48:43:23	31	jtxt#rhyolite	vincent#dslogin01	Web	SYNC=3, VERIFY		4eb13116-f67d-11e3-b56a-123139403		



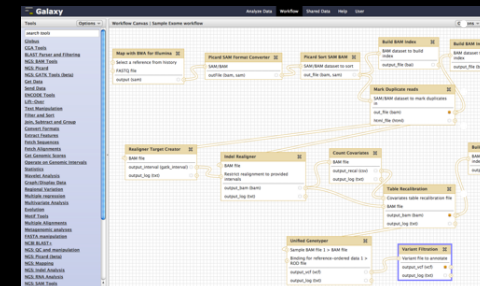


**Flexible, scalable,  
affordable  
genomics analysis  
for all biologists**

# Globus Genomics



## Galaxy Based Workflow Management System



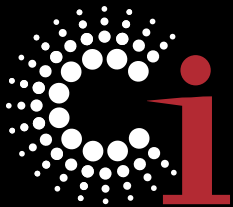
Globus Integrated with Galaxy  
Web-based UI  
Drag-Drop workflow creations  
Easily modify Workflow with new tools



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

Globus Genomics on Amazon EC2

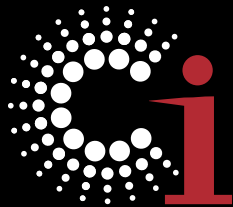
## Data Analysis



# Core Capabilities



- Computational profiles for various analysis tools to provide optimal performance
- Resources can be provisioned on-demand with Amazon Web Services cloud based infrastructure
- High performance, Reliable Data movement is streamlined with integrated Globus file-transfer functionality
- Integrated Globus endpoints and Campus login



# Diversity of collaborations

Dobyns  
Lab



Seattle Children's  
HOSPITAL • RESEARCH • FOUNDATION



UPMC  
University of Pittsburgh  
Medical Center

KU MEDICAL  
CENTER  
The University of Kansas



BROAD  
INSTITUTE

Berkeley  
UNIVERSITY OF CALIFORNIA



Nagarajan Lab

Washington  
University  
in St. Louis



THE UNIVERSITY OF  
CHICAGO

Cox Lab  
Volchenbom Lab  
Olopade Lab



Wexner Medical Center



INOVA®  
Join the future of health.



GEORGETOWN UNIVERSITY



UCSF



Genome  
Science  
Institute



Boston University Medical Center

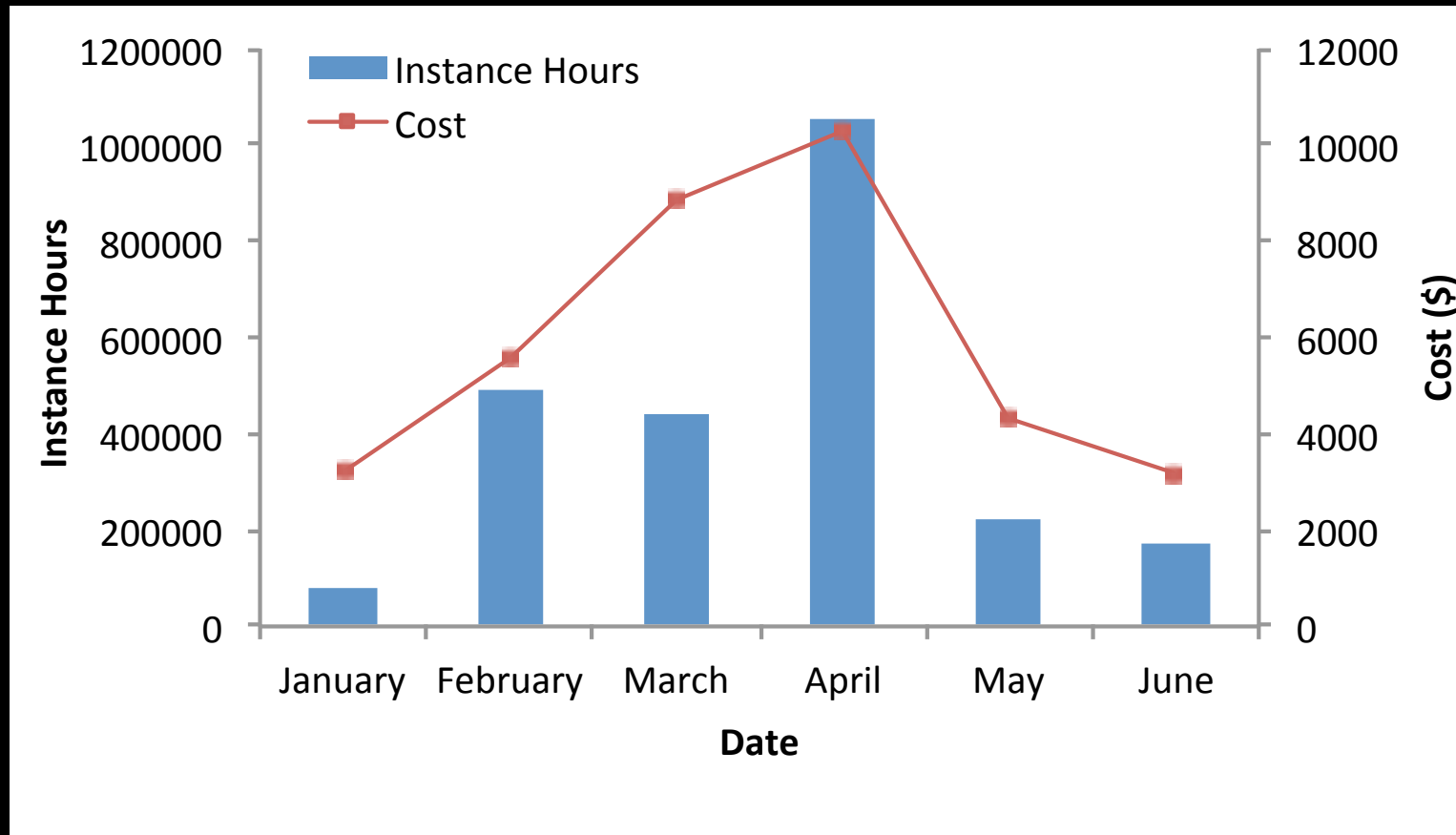


PerkinElmer®  
For the Better



CEDARS-SINAI MEDICAL CENTER.

# Usage has been promising



**2.5 Million Core hours**

# iFACE-IT (Globus + Galaxy->Climate)

Galaxy

faceit-portal.org

https://ais...fo.com/it-it globus geno...arth System ?level=picture&id=37 Adafruit Ind...ics and kits Valutazione ...i di Sistemi iPhone SDK Articles

globus | Galaxy / Earth System Analyze Data Workflow Shared Data Visualization Admin Help User Using 493.0 MB

Tools

search tools

Tool Installer

**DATA TRANSFER**

Globus Data Transfer

Text Manipulation

**FACE-IT**

Face-IT

- Upload File from your computer
- Global Marine Network  
Download one or more datasets from Global Marine Net
- Grib2 to GrADS idx Create an index file form a grib2
- Plot CSV
- Get Place
- Grib2 to GrADS cti Create a control file form a grib2
- Sdf Query File Query info on a self description NetCDF file
- Get CDL Get the CDL metadata from a NetCDF file
- Get NcXML Get the NcXML metadata from a NetCDF file
- Structure Info On Extract information about the file structure
- Info On Extract information about a file
- DayMet Daymet Demo Browser

AgMIP

Easy-SIM

WORKFLOWS

- All workflows
- Batch Submit

Map Satellite

Latitude: 30.6742  
Longitude: -83.3206

Google

Imagery ©2014 TerraMetrics Terms of Use

Latitude: 30.6742  
Longitude: -83.3206

Variable:

ALL  
TMAX  
TMIN  
DAYL  
PRCP  
SRAD  
SWE  
VP

Year:

ALL  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988

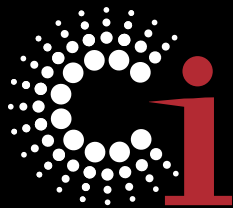
Get Data Visualize Data Send To Galaxy

Your History

Unnamed history  
485.8 MB

70: 200_204.psims.nc	eye icon	delete icon
69: 200_204.psims.nc	eye icon	delete icon
39: DSSAT Output (DSSAT1)	eye icon	delete icon
37: DSSAT Output Reference	eye icon	delete icon
36: DSSAT Output	eye icon	delete icon
35: Acmo Meta (AcmoMeta1)	eye icon	delete icon
34: DSSAT Input (DSSATInput1)	eye icon	delete icon
32: Acmo Meta Reference	eye icon	delete icon
31: DSSAT Input Reference	eye icon	delete icon
30: Acmo Meta	eye icon	delete icon
29: DSSAT Input	eye icon	delete icon
28: Regions Data	eye icon	delete icon
27: Strategy CSV Data Embu	eye icon	delete icon
26: Field CSV Data Embu	eye icon	delete icon
25: Survey CSV Data Embu	eye icon	delete icon
24: Region Data Embu	eye icon	delete icon
23: Strategy CSV Data Embu	eye icon	delete icon
22: Field CSV Data Embu	eye icon	delete icon
21: Survey CSV Data Embu	eye icon	delete icon
20: Region Data Embu	eye icon	delete icon
6: 200_204.psims.nc	eye icon	delete icon





# Galaxy+Globus-> Materials

globus genomics | Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Using 60.3 KB

Tools

Workflow Canvas | Material Science - Workflow 2

Details

search tools

MATERIAL SCIENCE

MOOSE Tools

- Upload Input from your computer
- Upload Mesh from your computer
- Generate Input usable by MOOSE tools
- Generate Mesh usable by MOOSE tools
- MOOSE test for selected input and mesh files
- Ferret for selected input and mesh files

Moose Tools

DATA TRANSFER

Globus Data Transfer

Get Data

Send Data

Generate Input

generated\_input (txt)

Generate Mesh

generated\_mesh (txt)

MOOSE test

Input file

Mesh file

mesh\_out (exodusII)

Generate Input

generated\_input (txt)

Ferret

Input file

Mesh file

mesh\_out (exodusII)

Tool: Generate Input

Version: 0.0.1

Edit Step Actions

Rename Dataset

generated\_input

Create

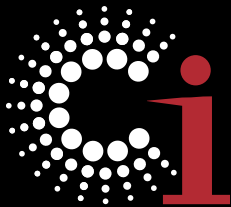
Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

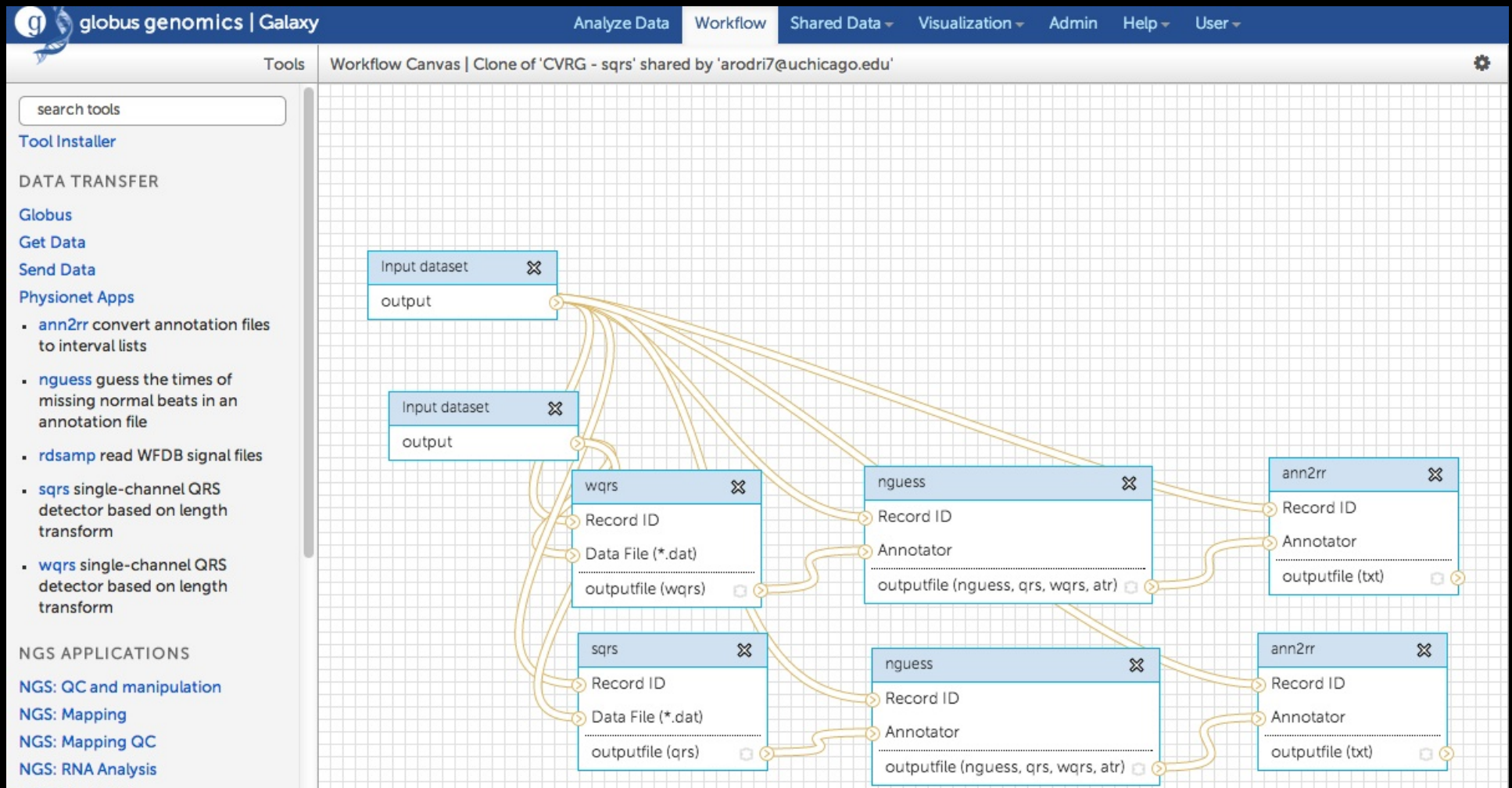
Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

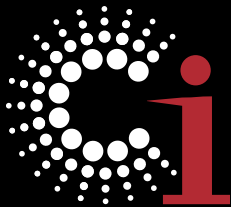
Press "Execute" to generate a sample input file (based on /moose/test/tests/splits/splitting\_additive\_ and add it to your history. Then this file



# Cardio Vascular Research Grid – Dr. Winslow @ Hopkins







# Cosmology - PDACS

Galaxy / PDACS

Workflow Canvas | Unnamed workflow

Tools

search tools

[Globus](#)

[Get Data](#)

[Halos - Simulation Data Analysis Tools](#)

[Halos - Predictors](#)

[2-point Functions - Simulation Data Analysis Tools](#)

[2-point Functions - Predictors](#)

[Conversion Tools](#)

[Graph/Display Data](#)

[Workflow control](#)

[Inputs](#)

Select Snapshot

SelectedSnapshot (dbm)

FOF Mass Function

Input Snapshot

o (csv)

Halo Finder

Input Snapshot

FOFProperties (binary)

SOProperties (binary)

HaloParticles (binary)

ParticleHaloTagFile (binary)

SOHaloProfiles (binary)

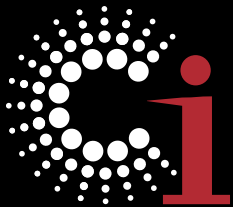
Output (dbi)

**Edit Workflow Attributes**

**Name:**  
Unnamed workflow

**Tags:**  
Apply tags to make it easy to search for and find items with the same tag.

**Annotation / Notes:**  
Describe or add notes to workflow  
Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.



# Affordability/TCO

## 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

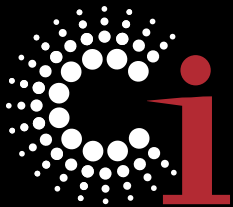
- Pricing based on example of paired-end fastq files with 5 Gbases.
- Pipeline includes quality control, alignment, exon count using cufflinks, and HT-Seq count.

- Pricing includes
  - Estimated compute
  - Storage (one month)
  - Globus Genomics platform usage
  - **Support**



# Sustainability

- Our goal is to build service that lives beyond a funded proposal
- Two pricing options and multiple usage tiers.
  - Targeted users include individual research groups and bioinformatics cores
  - Platform pricing (includes only subscription to the Globus Genomics platform)
  - Bundled pricing (includes Globus Genomics platform subscription and AWS usage costs)



- More information on Globus Genomics and to sign up for a **free** trial :  
[www.globus.org/genomics](http://www.globus.org/genomics)
- More information on Globus:  
[www.globus.org](http://www.globus.org)
- We are hiring!

# Our work is supported by:



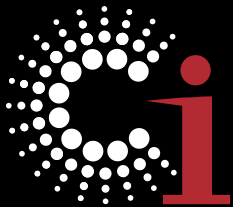
U.S. DEPARTMENT OF  
**ENERGY**



THE UNIVERSITY OF  
**CHICAGO**

**Argonne**  
NATIONAL LABORATORY





Thank you!

@madduri