# Data Literacy and Data Management

HPRC Short Course
January 2019

Rick McMullen, Ph.D.

Associate Director, HPRC
Texas A&M University
mcmullen@tamu.edu

# For More Help…

| | |
|---|---|
| Website: | hprc.tamu.edu |
| Email: | help@hprc.tamu.edu |
| Telephone: | (979) 845-0219 |
| Visit us in person: | Henderson Hall, Room 114A |

**Help us, help you -- we need more info**

- Which Cluster
- UserID/NetID
- Job id(s) if any
- Location of your jobfile, input/output files
- Application used if any
- Module(s) loaded if any
- Error messages
- Steps you have taken, so we can reproduce the problem

# Goals for the session – What are yours?

- Present a conceptual framework for the life cycle of data
- Present a case for attending to managing your data in an organized way
- Learn about the concept of the “life-cycle” of data
- Learn about some tools and systems for managing your data
  - Storing
  - Organizing and finding
  - Moving

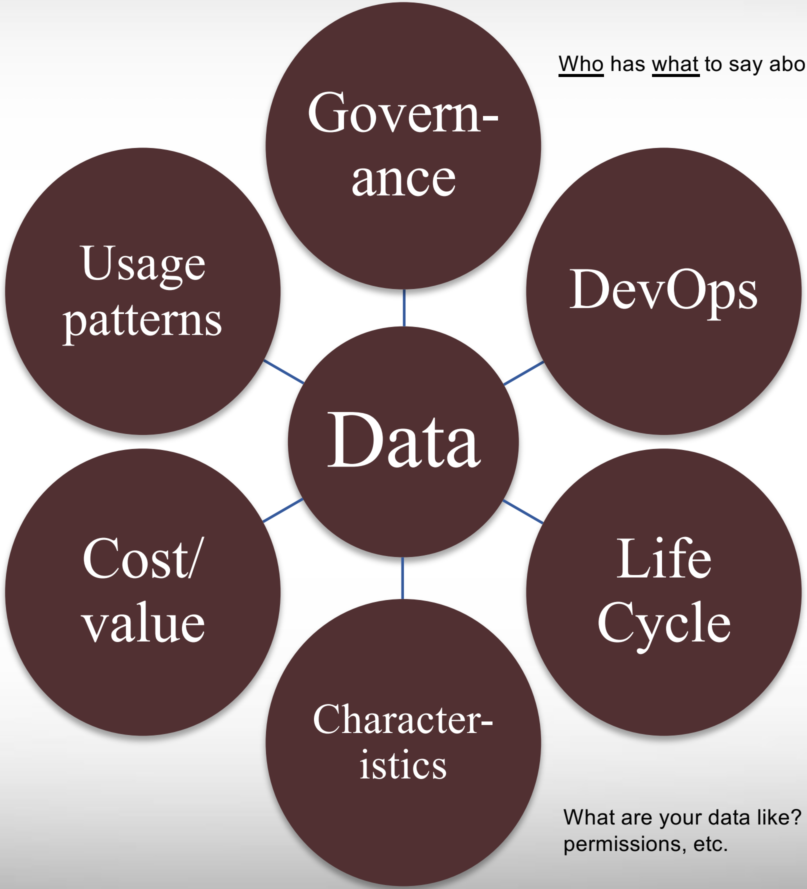
# Outline

- Framework issues
  - Awareness about your data/information assets
  - Conceptual framework for the “life cycle” of your data
- Some specific considerations
  - Kinds of data you might use in your research or scholarship
    - What types do you use now?
  - External drivers of requirements
    - Legal and regulatory controls
    - Institutional processes and requirements
  - Storing your stuff
  - Finding your stuff (metadata, organization and searching)
  - Moving your data from here to there
  - Using git and github as a way to organize, manage and share data

# Questions to think about

- What is your work about?
- Who do you work with? Within your facility? National? International?
- What kind of data do you work with and where does it come from?
- Where do you do your computing? What resources do you use now? Is there a data or computing coordination center?
- What bottlenecks and issues have you identified? Are these process or infrastructure related?

# Lenses for looking at your information assets



Data

Governance

Who has what to say about your data

DevOps

Institutional processes around how your data are produced, accessed and used

Life Cycle

How the form and location of your data change over time

Characteristics

What are your data like? Size, format, applicable standards, permissions, etc.

Cost/ value

Cost to produce, obtain and manage your data, and its value to you and others

Usage patterns

How you (and others) use your data in practice

# External drivers of requirements

Legal and regulatory controls

Privacy laws

Health Insurance Portability and Accountability Act of 1996 (HIPAA, https://www.hhs.gov/hipaa/index.html)

The Family Educational Rights and Privacy Act of 1974 (FERPA, https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html)

The General Data Protection Regulation 2016/679 (GDPR, EU, https://eugdpr.org/)

U.S. and E.U. Intellectual Property Policy

e.g. https://www.uspto.gov/intellectual-property-ip-policy

# External drivers of requirements (2)

## Funding agency requirements

### NSF, NIH, NOAA data sharing and data management plan rules

NSF: https://www.nsf.gov/bfa/dias/policy/dmp.jsp
NIH: https://grants.nih.gov/grants/policy/data_sharing/
NOAA: https://nosc.noaa.gov/EDMC/PD.DSP.php
Others…

### Working with Controlled Unclassified Information

Contractual obligations
For Defense contracts: Defense Federal Acquisition Regulation Supplement (DFARS)
https://www.acq.osd.mil/dpap/dars/dfarspgi/current/index.html
NIH grant/contract rules about management of Protected Health Information (PHI)
https://privacyruleandresearch.nih.gov/pr_02.asp

Infrastructure requirements and solutions to meet these contractual obligations

# External drivers of requirements (3)

## Institutional policies and other issues

Institutional Review Board (IRB) – human subjects
Institutional Animal Care and Use Committee (IACUC) – animal subjects
Responsible Conduct of Research implications for data protection

Intellectual property protection rules
IP: For TAMU, University Rule 17.01.99.M1 “Intellectual Property Management and Commercialization”

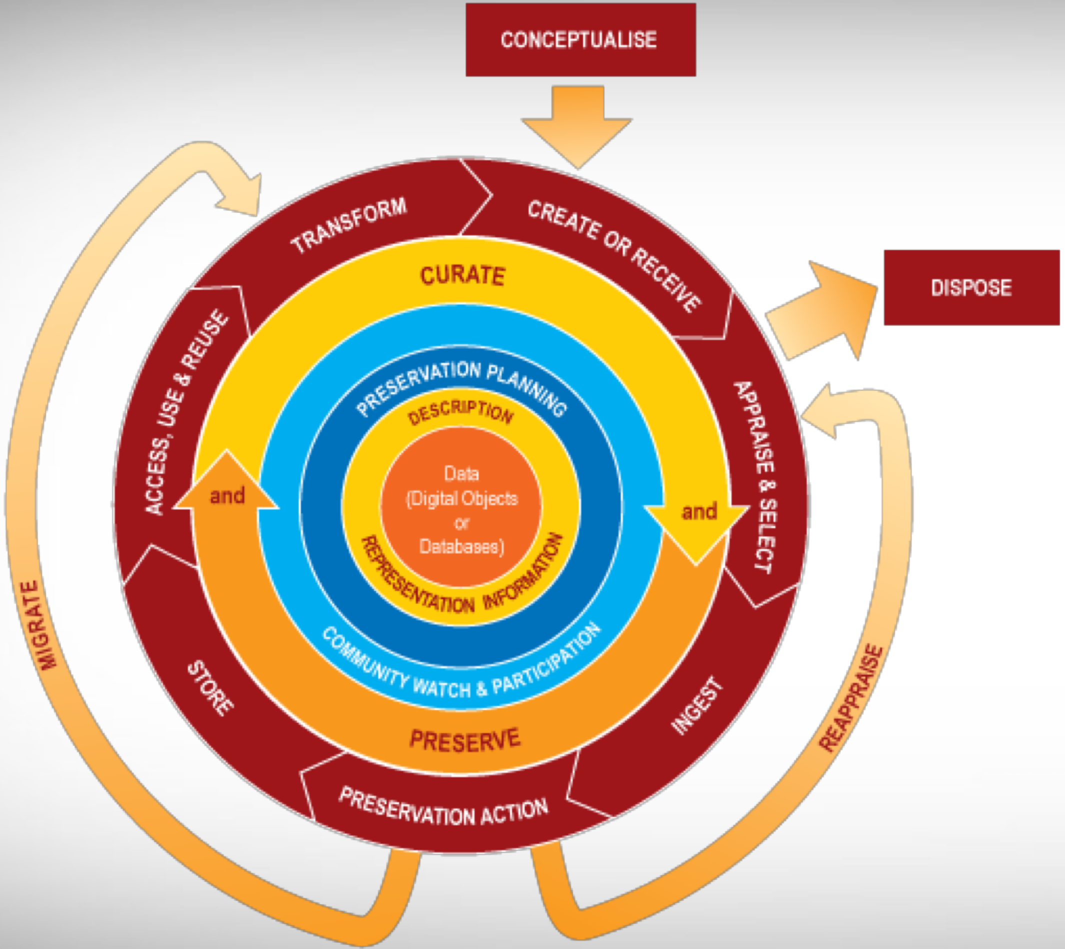
## Domain-specific norms for data sharing and preservation

Repositories (e.g. list at https://library.stanford.edu/research/data-management-services/share-and-preserve-research-data/domain-specific-data-repositories )
File formats and metadata
Quality standards, metrics and tests
Provenance, assurance (e.g. checksums), and change tracking

# It’s Your Data! What’s at risk?

- Your Dissertation or Thesis
- Your research grant
- Your research collaboration and collaborators
- Maintaining compliance with laws, regulations, policies
- Your personal and professional reputation!

# Conceptual framework for the life-cycle of data



# Key elements of the Data Curation Centre Curation Lifecycle Model

1. Conceptualise
2. Create or Receive
3. Appraise and Select
4. Ingest
5. Preservation Action
6. Store
7. Access, Use and Reuse
8. Transform

http://www.dcc.ac.uk/resources/curation-lifecycle-model/

# Let’s focus on these for now

- Storing your data
- Annotating and finding your data (metadata)
- Moving your data

# Basic Linux Shell Commands Review

- Where am I? Print working directory: `pwd`
- Change directory: `cd <new directory path>`
- List directory: `ls`
- List directory with details: `ls –l`
- Create new directory: `mkdir <directory path>`
- Remove a file: `rm <file path>`
- Remove a directory: `rmdir <directory path>`
- Remove a directory  
  and all its files and directories: `rm –r <directory path>`
- Copy a file: `cp <source path> <target path>`
- Move a file: `mv <source path> <target path>`

# Storing your data

# Ways to store and organize your data

- Spreadsheets: Excel, OpenOffice, Google Sheets
  - Convenient, ubiquitous, easy to use
  - Hard to search for specific items (rows) in a large number of spreadsheets
- Databases: “local” SQL, SQL server, NoSQL (key-value, tuple, etc.)
  - Easy to share and keep consistent
  - Someone has to be the database manager
  - Google Datastore, BigTable, Cloud SQL, Azure Data Market databases
- Files and directories -> tar archives
- Structured file formats (e.g. XML, JSON, discipline or vendor-specific)
- Cloud services - AWS, Azure, Google, GitHub;  c.f. `https://github.com/sr320/LabDocs/wiki/Data-Management`

# Places to store your data

- Bad
  - Not-backed-up laptop or desktop
  - RW-CDs and DVDs (bit rot, labeling, filing and tracking)
  - USB external hard drives (usually consumer grade, no backup, easy to drop)
  - USB thumb drives (can be damaged by handling)
  - Your laptop (drop --> damaged hard drive)
- Good
  - File server with RAID and backup (https://en.wikipedia.org/wiki/RAID)
  - Backed-up laptop or desktop as long as it is backed up frequently
  - Cloud storage (Google, Amazon, Azure, GitHub, DropBox) as long as you understand the risk implied in the Service Level Agreement, and if you keep multiple copies
  - Managed database server (also with regular exports if possible for backup)

# Cloud storage…

- Important for data acquisition and exchange from multiple sites in a research collaboration
- Security is really not a problem *at the cloud vendor's end*. Check AWS's compliance and assurance program page: https://aws.amazon.com/compliance/
- Governance needs to cover all data, even in the cloud, and the economics are different to on-premesis
  - Capacity is an ongoing cost
  - No depreciation of infrastructure
  - Metadata may be more difficult to collect, making curation more difficult
  - Life cycle considerations are more concrete – doing nothing costs $$ in real time.

# Other considerations

- Service labs generally won’t save your data for any length of time (e.g. Microscopy, XRF, NMR, Mass spec…)
- Other than file-format-specific metadata, it’s up to you to organize your stuff appropriately
- Cloud services are OK as long as you understand the risks, limits and financial aspects
- Many disciplines offer repositories for specific kinds of data
  - https://www.nature.com/sdata/policies/repositories
- NSF, NIH, NASA and NOAA require a data management plan for ongoing access to publicly funded research data.

# Archiving data

- For moving lots of files, backup and “cold storage”
- Reduces many files and directories to one file
- An archive is easier to handle than a directory hierarchy full of files
- Popular archive tools and formats on all platforms:
  - tar
  - zip
  - cpio

# The `tar` Command - examples

```
cd
```

go to your home directory

Package the temp directory into a file called my_hg19.tar

```
tar -cvf my_hg19.tar temp
```

Package the temp directory into a compressed file called my_hg19.tar.gz

```
tar -cvzf my_hg19.tar.gz temp
```

Show the contents of the compressed tar file

```
tar -tzf my_hg19.tar.gz
```

Change the name of your original temp directory so you don’t overwrite it

```
mv temp temp_orig
```

Extract all contents from the compressed tar file

```
tar -xvzf my_hg19.tar.gz
```

| | | |
|---|---|---|
| -c | = | create |
| -f | = | file |
| -t | = | list contents |
| -v | = | verbose |
| -z | = | gzipped |
| -x | = | extract |

# zip  - Archive and compress

Create an archive:

```
$ zip -r archive5.zip test-dir
```

Check the contents:

```
$ unzip -l archive5.zip
```

Extract it:

```
$ cd /tmp/extract-dir
$ unzip /tmp/archive5.zip
```

# Grep, compression and zip

- Tar can also compress/decompress the archives it creates
  - `tar –c –z –v -f my_archive.tar.gz <source>`
  - `tar –x –z –v -f my_archive.tar.gz`

  - `-c = create archive;`
  - `-x = extract files;`
  - `-z = gzip it;`
  - `-f = file name to create or extract from`
  - `<source> is usually a directory path name`
- `.tar.gz == .tgz`
- the suffix doesn't matter, tar looks at the “-z” switch

# Compression efficiency example

```
$ du -s -h sample_data
117M sample_data
$ ls -l -h sample_data.*
-rw-rw-r-- 1 mcmullen mcmullen 45M Jan 19 11:20 sample_data.tar.gz
-rw-rw-r-- 1 mcmullen mcmullen 45M Jan 19 11:20 sample_data.zip
```

- **45/117 = 38% of the original**
- gzip and zip give comparable results for this data
- **Now, try zip and gzip with [https://marknelson.us/assets/2006-06-20-million-digit-challenge/AMillionRandomDigits.bin](https://marknelson.us/assets/2006-06-20-million-digit-challenge/AMillionRandomDigits.bin)**
- Cf. https://marknelson.us/posts/2006/06/20/million-digit-challenge.html

# Data Cleaning

- Processing pipelines and batch scripts (these preserve methodologies and make them reusable)
- OpenRefine (openrefine.org) for cleaning tabular and relational data

A free, open source, powerful tool for working with messy data

- Save everything!

# Extract, transform and load tasks (ETL)

Beyond cleaning, transforming data is a part of doing the science

Assuming you have data stored in a database or some specific file format, ETL tasks capture best practices for transforming it.
Capturing ETL tasks in a reproducible way is a critical part of managing your data. Often, unless you preserve the ETL process, your workflow is not reproducible and your results not verifiable.

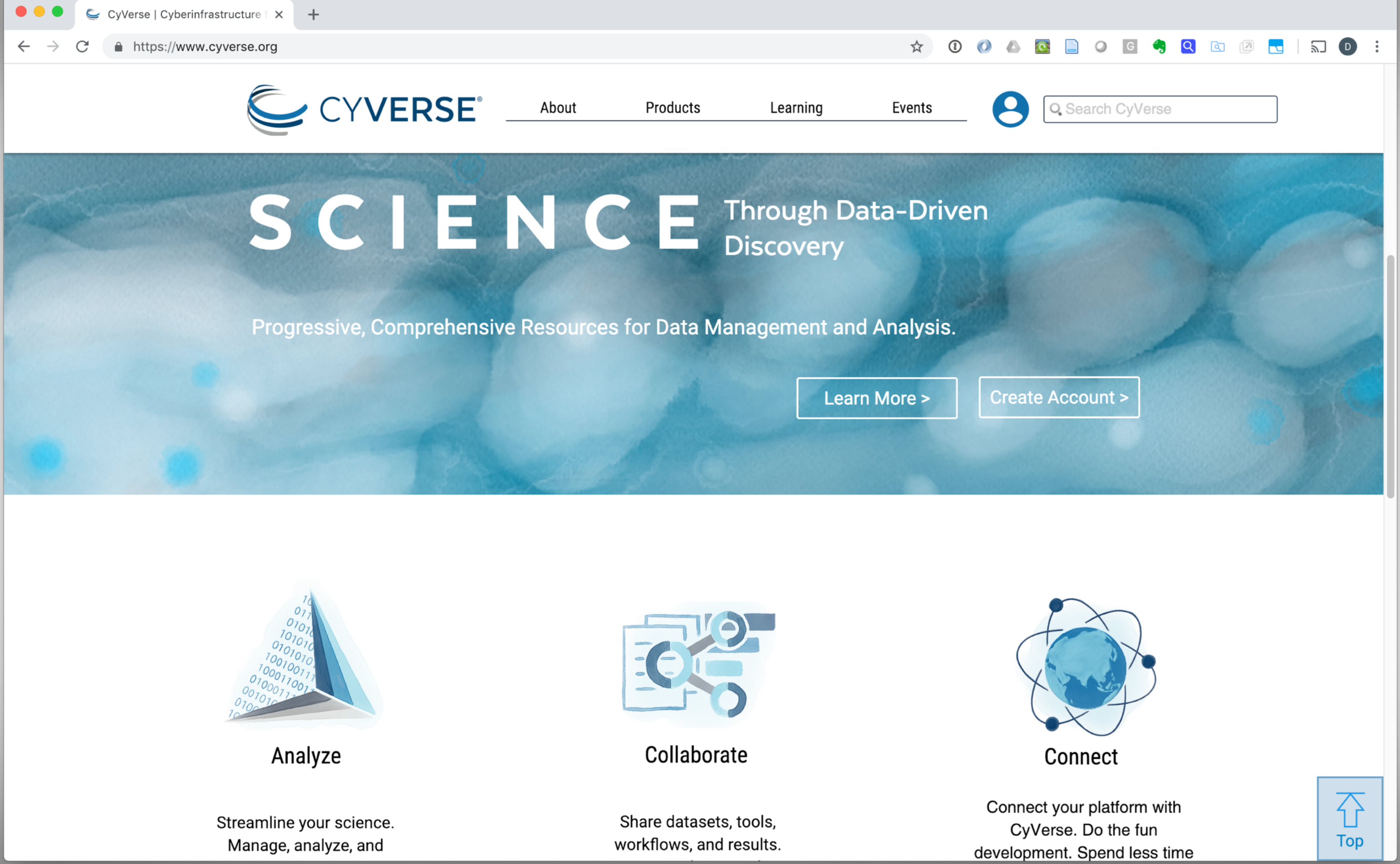
Supporting software:
Shell scripts
SQL scripts
Tool-specific scripts (e.g. Visual Basic for Excel)
Purpose-built systems like Talend Open Studio (https://www.talend.com/download/talend-open-studio/)

# Transformations are computations, VMs are research data

- The computing you do on your data are as important as your data WRT reproducibility and documentation of method
- It is becoming possible and useful to capture your data processing environments and computational research tools as Virtual Machines
- VMs can be saved in a ”portable” format, Open Virtualization Format
- End-user VM systems: VirtualBox, KVM/qemu (free), VMWare (cost)
- Also, “containers” are a more lightweight option for preserving complete work environments. Repositories of lots of containerized software exist.
  - Docker - https://www.docker.com/ (also GUIs Mac and Windows)
  - Singularity https://singularity.lbl.gov/ , https://www.sylabs.io/docs/

# VM Service Examples

- Obviously Amazon Web Services, Google, Azure etc.
- More focused on science and research:
  - Jetstream (NSF) - https://jetstream-cloud.org/
    - Jetstream allows VMs to be catalogued as publications (get a doi, keep in a public repository)
  - CyVerse - http://www.cyverse.org/, “Cyberinfrastructure for Data Management and Analysis”, mostly about biology and bioinformatics



SCIENCE Through Data-Driven Discovery

Progressive, Comprehensive Resources for Data Management and Analysis.

**Analyze**

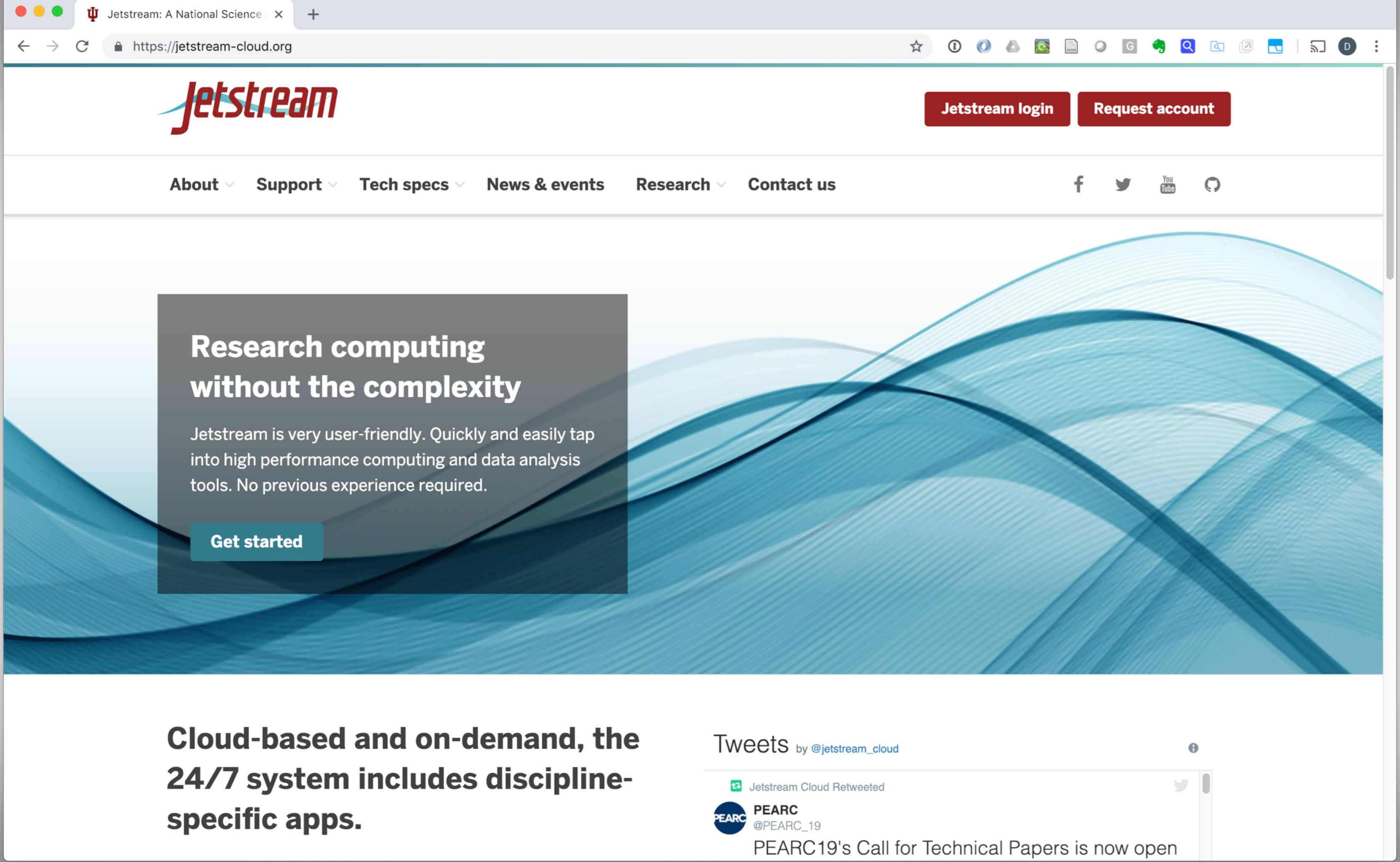
Streamline your science. Manage, analyze, and

**Collaborate**

Share datasets, tools, workflows, and results.

**Connect**

Connect your platform with CyVerse. Do the fun development. Spend less time



## Research computing without the complexity

Jetstream is very user-friendly. Quickly and easily tap into high performance computing and data analysis tools. No previous experience required.

## Cloud-based and on-demand, the 24/7 system includes discipline-specific apps.

Tweets by @jetstream_cloud

Jetstream Cloud Retweeted

PEARC
@PEARC_19
PEARC19's Call for Technical Papers is now open

# Metadata

# Annotating your stuff - Metadata

- “Data about data” – who, what, when, why, how
- ***Critically important if you want to find anything and understand what it means more than a week from now***
- Directory and file naming schemes
- Internal metadata (e.g. TIFF image headers)
- Spreadsheet column names
- Database data directories and field names
- XML Schema (tags and optional values)
- Disk labels, textual documentation

# File types are metadata!

- One kind of typing is a file suffix, e.g. .c, .xlsx, .sql
  - Not 100% reliable as you can rename a file to anything
- External typing using MIME types
  - see IANA list at https://www.iana.org/assignments/media-types/media-types.xhtml
- Files also have internal, sometimes characteristic typing information called “magic numbers”
  - See, e.g., https://asecuritysite.com/forensics/magic
- The Linux “file” command can tell you a lot about a file

# Metadata Standards

Nice thing about standards is there are so many to choose from

Images - Exchangeable image file format (Exif) + TIFF
Instrument data sets – community and vendor standards
…
Pretty good list by discipline at:
http://www.dcc.ac.uk/resources/metadata-standards/list

Point is to understand what standards you should be using, and use it
You can also create your own organization and metadata standards using directory and file naming conventions

# File hierarchy and “name” metadata

```
Growth_rates_enz_1                       - Directory
     Read.me                   - File with description of method
     Experiment_1                        - Directory
          Image_0001_date_time      - File with observations
          …
          Image_9999_date_time      - File with observations
     Experiment_2
     …
```

- Hard to change your mind if you need to modify your metadata schema, e.g. add a location where the experiment took place
- Easy to bundle and export your data at any level of the file tree using Unix “tar” command

# Finding your stuff - searching

- Search for names and types of files
  - Linux/Unix/Mac OS X: “find” command
  - GUI file browser search
- Search for text or text patterns in files
  - Linux/Unix/Mac OS X: “grep” command in a directory
  - Within specific named files: “find … -exec grep … {} \; -print”
- Spreadsheet search box
- Databases – SQL/noSQL queries,
- Web-based information – Google site search, Microformats

# Large scale data management issues

- Getting the right storage system or service for the volume, variety, and velocity of your data
- Tools for automating tasks (metadata extraction, cataloging, tiered storage management)
- Managing risk: security and compliance concerns (e.g. HIPAA, FERPA, licensed data with restrictions and terms, etc.)

# A couple of metadata tools for “big” collections

- Robinhood Policy Engine - https://github.com/cea-hpc/robinhood/wiki
  - Policy Engine: schedule actions on filesystem entries according to admin-defined criteria, based on entry attributes.
  - User/group usage accounting, including file size profiling.
  - Extra-fast 'du' and 'find' clones.
  - Customizable alerts on filesystem entries.
  - Aware of Lustre OSTs and pools.
  - Filesystem disaster recovery tools.
  - Open, LGPL-compatible license.
- Starfish Storage - http://www.starfishstorage.com/
  - Similar to Robinhood but supports cloud-based storage as well as local POSIX FSs
  - Not free.

# Bigger picture questions to think about

- Do you have a business case for using cloud vendors for R&D computing/storage tasks?
- What are your current data governance assumptions and drivers?
- Does your governance strategy work well when your data are in the cloud?
- Do your devops processes work across on-prem and cloud facilities
- Do you have sufficient network capacity (and backup) for working with lots of your data at a cloud vendor?

# Moving your data

# Moving your data

- Typical needs: To/From a service lab; To a colleague; To a repository
- Relatively easy to move files from one server to another, especially on-campus
- Harder to move very large files or a large number of files, especially cross-country or internationally
- Campus bandwidths are 1 to 10 Gbps max. (125 – 1250 MB/s)
- Intercampus can be more but subject to many issues

# Tools to move files and folders

- On the same machine:
  - `mv, cp –R`
- Between machines where you have log-ins:
  - `scp, sftp, rsync, Globus`
  - tar or zip first to reduce copy time
  - Filezilla is a good GUI for moving files
  - rsync can be used to maintain up to date copies of directories and files
- From public sources
  - `wget, git`

# Obvious problems

Size of files
Number of files
Available bandwidth
Firewalls filtering some protocols out
Encryption needed?
Resilience of protocol, i.e. restart capabilities

## NOT so obvious

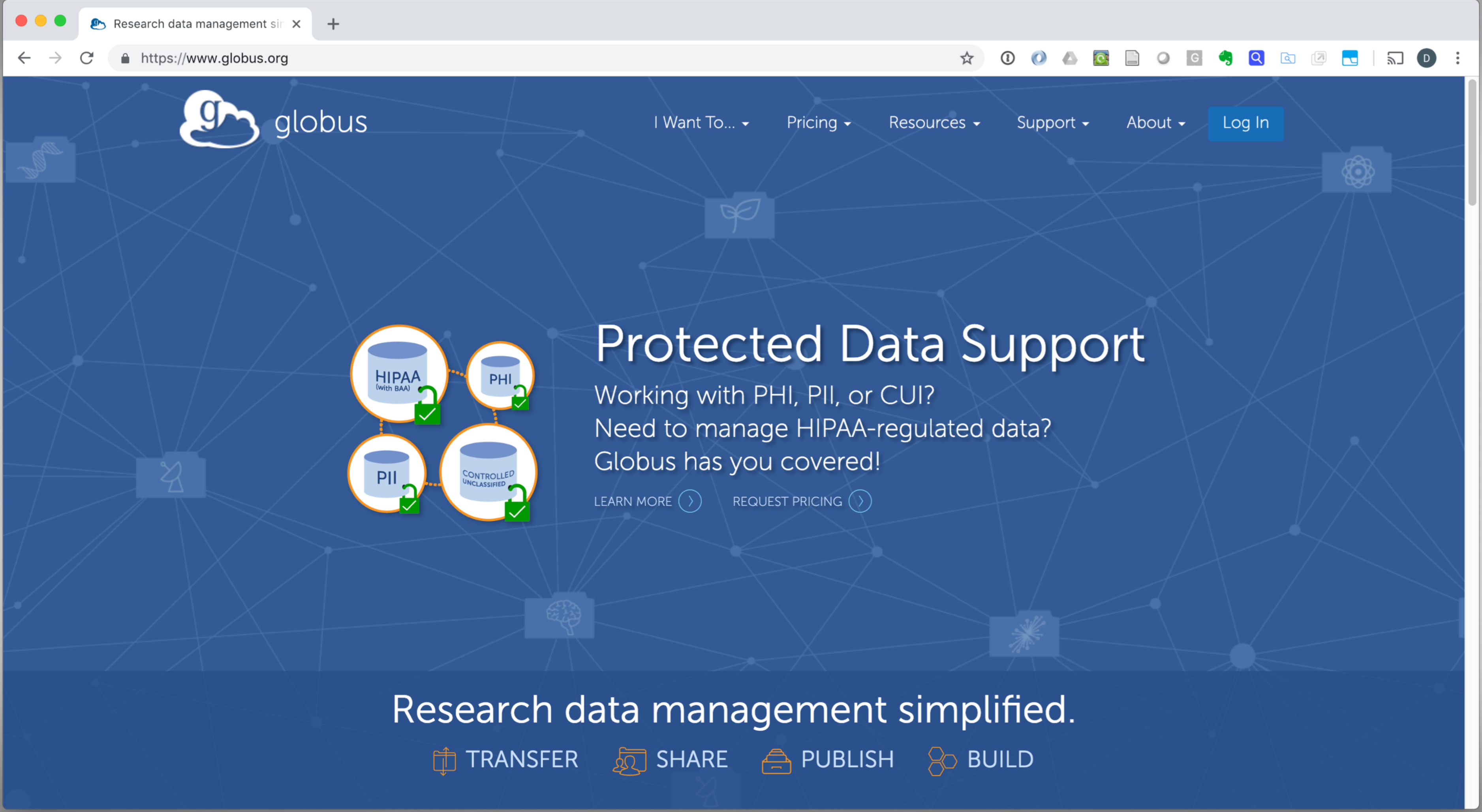
Lifetime of credentials or certificates used during transfer
TCP protocol on local area network vs. long haul network
Firewalls dropping packets to meet some "traffic shaping" policy
One transfer to/from outside the university will cross three administrative domains, at a minimum,
e.g. TAMU to Broad Institute:
TAMU <-> Texas LEARN network <-> Internet2 <-> Broad Institute

# Globus for moving files

- Globus is a service for highly optimized, reliable, and unattended file transfers
- Files are moved between “endpoints” set up by anyone
- File transfers are set up through a web interface
- You must have a globus account (free) and know what endpoints you want to use
- You can set an endpoint up on your own PC/laptop using “globus connect personal”
- Get started at globus.org, use your TAMU netID and login

# Publicly visible TAMU Endpoints

| | |
|---|---|
| tamusc#terra-ftn | owner: tamusc@globusid.org |
| tamusc#ada-ftn1 | owner: tamusc@globusid.org |
| tamusc#ada-ftn2 | owner: tamusc@globusid.org |
| tamu#brazos-dev | owner: tamu@globusid.org |
| tamu#brazos | owner: tamu@globusid.org |
| TAMU terra-ftn | owner: tmarkhuang@globusid.org |
| TAMU ada-ftn2 | owner: tmarkhuang@globusid.org |
| TAMU ada-ftn1 | owner: tmarkhuang@globusid.org |



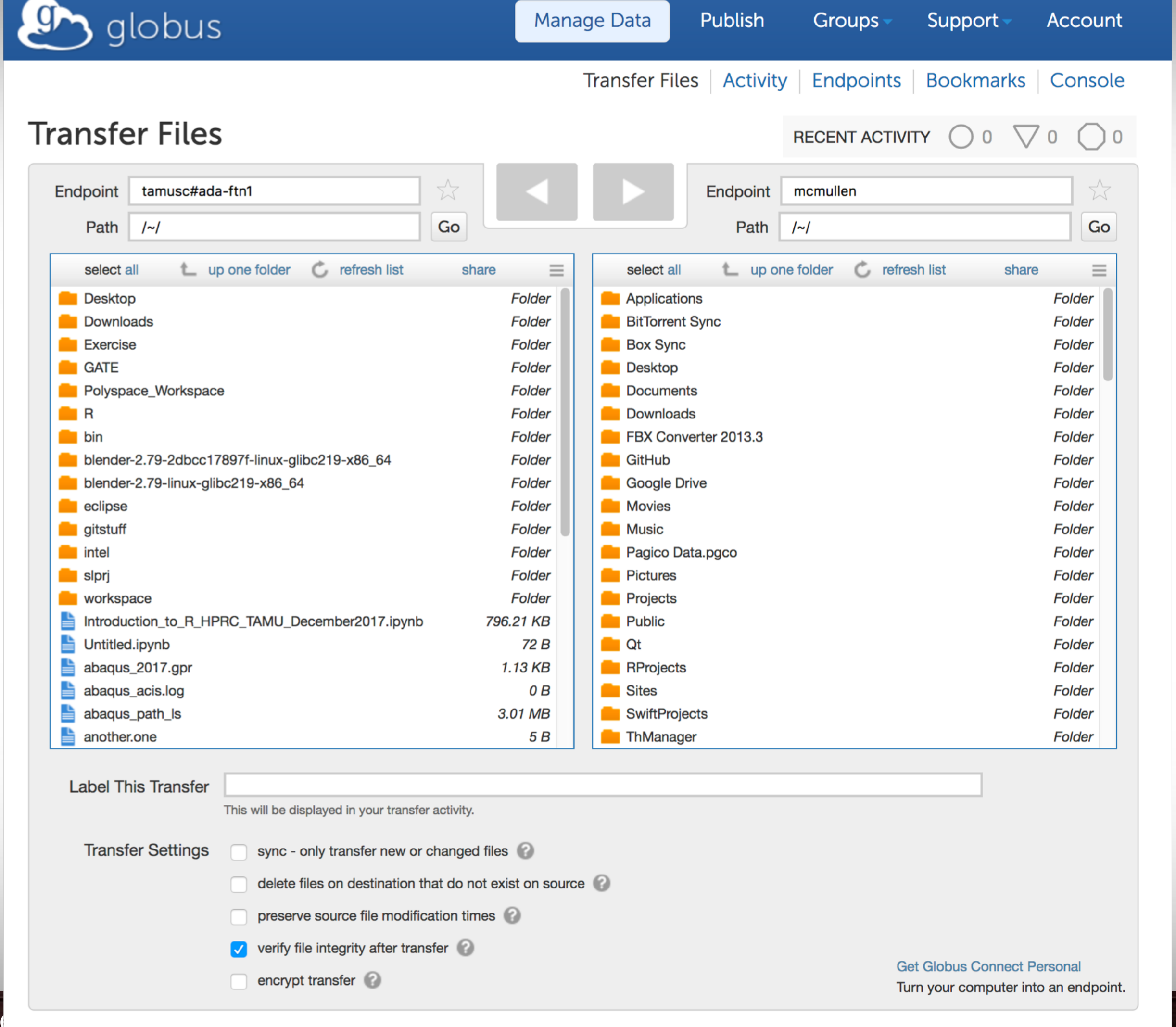
# Protected Data Support

Working with PHI, PII, or CUI?
Need to manage HIPAA-regulated data?
Globus has you covered!

LEARN MORE 〉 REQUEST PRICING 〉

## Research data management simplified.

TRANSFER SHARE PUBLISH BUILD



# Transfer Files

RECENT ACTIVITY ○ 0 ▽ 0 ⬡ 0

Endpoint: tamusc#ada-ftn1 ☆
Path: /~/ Go

Endpoint: mcmullen ☆
Path: /~/ Go

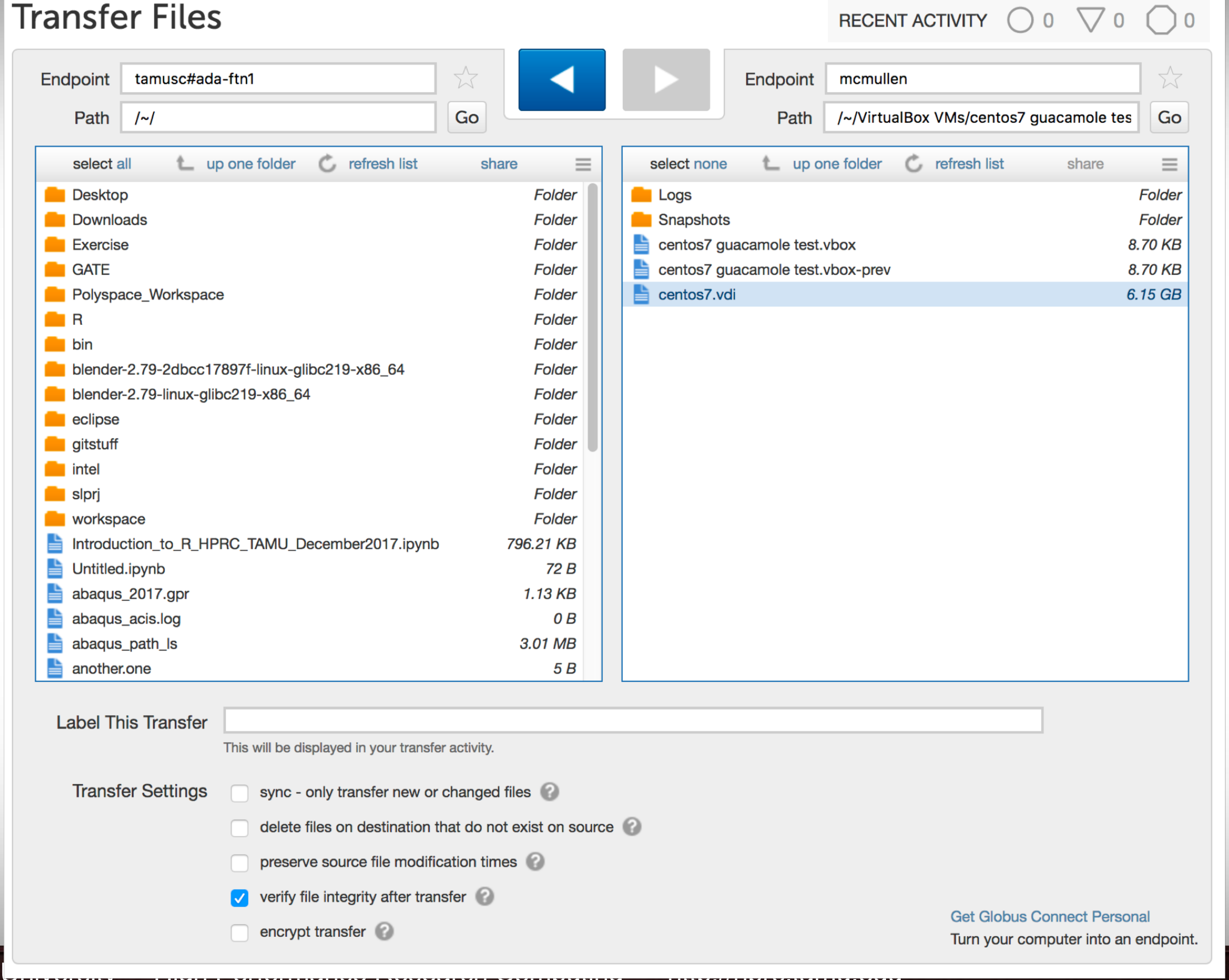
| | |
|---|---|
| Desktop | Folder |
| Downloads | Folder |
| Exercise | Folder |
| GATE | Folder |
| Polyspace_Workspace | Folder |
| R | Folder |
| bin | Folder |
| blender-2.79-2dbcc17897f-linux-glibc219-x86_64 | Folder |
| blender-2.79-linux-glibc219-x86_64 | Folder |
| eclipse | Folder |
| gitstuff | Folder |
| intel | Folder |
| slprj | Folder |
| workspace | Folder |
| Introduction_to_R_HPRC_TAMU_December2017.ipynb | 796.21 KB |
| Untitled.ipynb | 72 B |
| abaqus_2017.gpr | 1.13 KB |
| abaqus_acis.log | 0 B |
| abaqus_path_ls | 3.01 MB |
| another.one | 5 B |

| | |
|---|---|
| Applications | Folder |
| BitTorrent Sync | Folder |
| Box Sync | Folder |
| Desktop | Folder |
| Documents | Folder |
| Downloads | Folder |
| FBX Converter 2013.3 | Folder |
| GitHub | Folder |
| Google Drive | Folder |
| Movies | Folder |
| Music | Folder |
| Pagico Data.pgco | Folder |
| Pictures | Folder |
| Projects | Folder |
| Public | Folder |
| Qt | Folder |
| RProjects | Folder |
| Sites | Folder |
| SwiftProjects | Folder |
| ThManager | Folder |

Label This Transfer: 
This will be displayed in your transfer activity.

Transfer Settings
- □ sync - only transfer new or changed files
- □ delete files on destination that do not exist on source
- □ preserve source file modification times
- ☑ verify file integrity after transfer
- □ encrypt transfer

Get Globus Connect Personal
Turn your computer into an endpoint.



## Transfer Files

RECENT ACTIVITY ○ 0 ▽ 0 ⬡ 0

Endpoint: tamusc#ada-ftn1 ☆

Path: /~/ Go

Endpoint: mcmullen ☆

Path: /~/VirtualBox VMs/centos7 guacamole tes Go

| | |
|---|---|
| Desktop | Folder |
| Downloads | Folder |
| Exercise | Folder |
| GATE | Folder |
| Polyspace_Workspace | Folder |
| R | Folder |
| bin | Folder |
| blender-2.79-2dbcc17897f-linux-glibc219-x86_64 | Folder |
| blender-2.79-linux-glibc219-x86_64 | Folder |
| eclipse | Folder |
| gitstuff | Folder |
| intel | Folder |
| slprj | Folder |
| workspace | Folder |
| Introduction_to_R_HPRC_TAMU_December2017.ipynb | 796.21 KB |
| Untitled.ipynb | 72 B |
| abaqus_2017.gpr | 1.13 KB |
| abaqus_acis.log | 0 B |
| abaqus_path_ls | 3.01 MB |
| another.one | 5 B |

| | |
|---|---|
| Logs | Folder |
| Snapshots | Folder |
| centos7 guacamole test.vbox | 8.70 KB |
| centos7 guacamole test.vbox-prev | 8.70 KB |
| centos7.vdi | 6.15 GB |

Label This Transfer

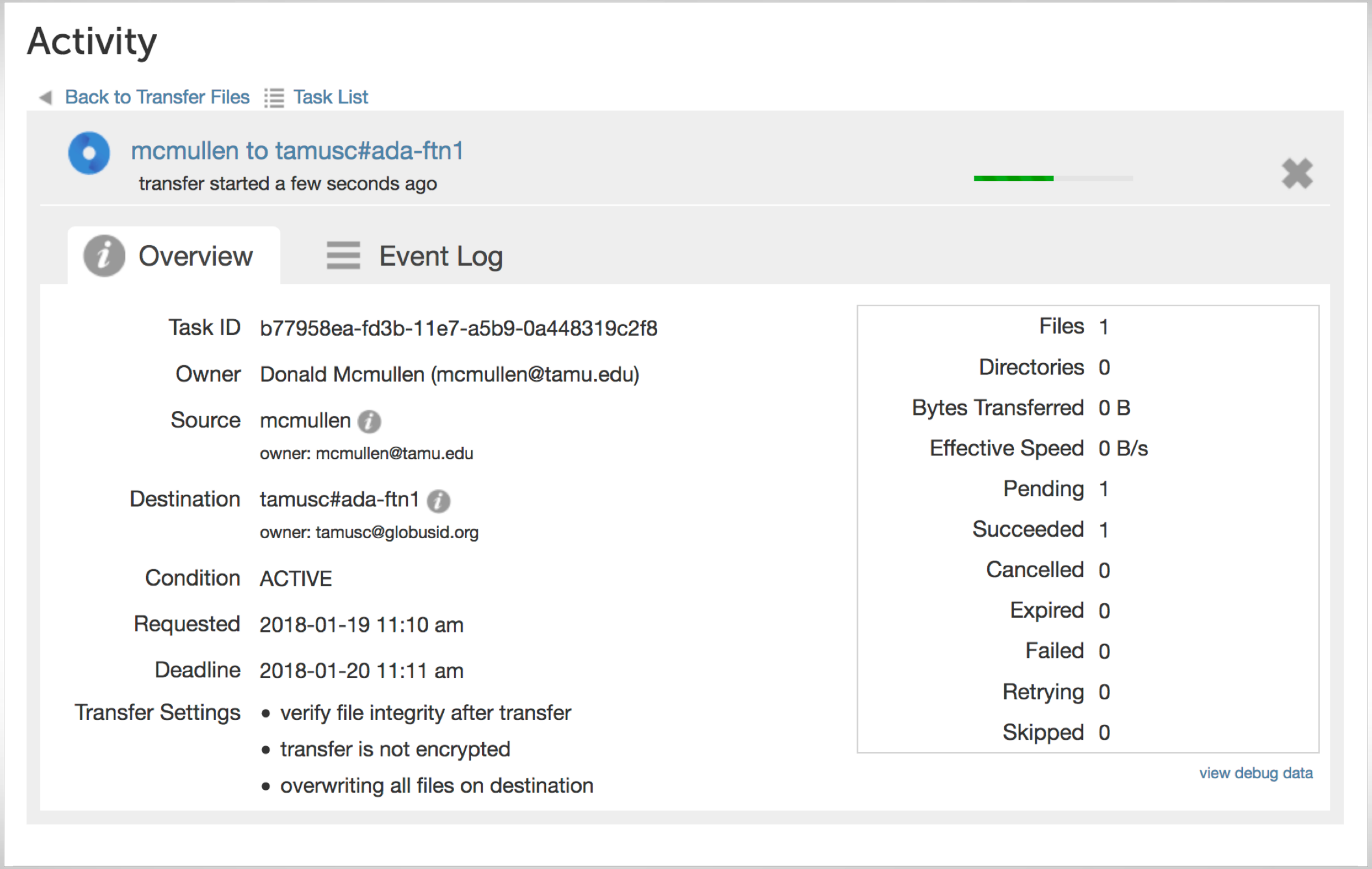
This will be displayed in your transfer activity.

Transfer Settings

- ☐ sync - only transfer new or changed files ?
- ☐ delete files on destination that do not exist on source ?
- ☐ preserve source file modification times ?
- ☑ verify file integrity after transfer ?
- ☐ encrypt transfer ?

Get Globus Connect Personal

Turn your computer into an endpoint.



# Activity

mcmullen to tamusc#ada-ftn1

transfer started a few seconds ago

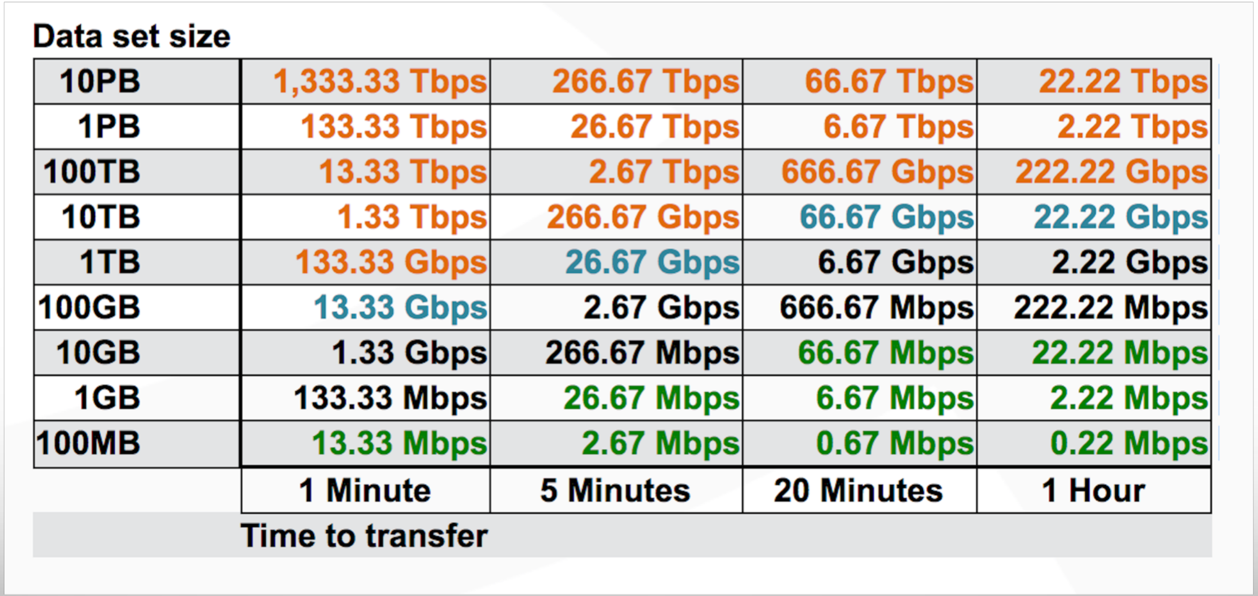
| | |
|---|---|
| Task ID | b77958ea-fd3b-11e7-a5b9-0a448319c2f8 |
| Owner | Donald Mcmullen (mcmullen@tamu.edu) |
| Source | mcmullen<br>owner: mcmullen@tamu.edu |
| Destination | tamusc#ada-ftn1<br>owner: tamusc@globusid.org |
| Condition | ACTIVE |
| Requested | 2018-01-19 11:10 am |
| Deadline | 2018-01-20 11:11 am |
| Transfer Settings | • verify file integrity after transfer<br>• transfer is not encrypted<br>• overwriting all files on destination |

| | |
|---|---|
| Files | 1 |
| Directories | 0 |
| Bytes Transferred | 0 B |
| Effective Speed | 0 B/s |
| Pending | 1 |
| Succeeded | 1 |
| Cancelled | 0 |
| Expired | 0 |
| Failed | 0 |
| Retrying | 0 |
| Skipped | 0 |

# *Minimum* Time to Transfer Data

***Minimum*** time needed to to transfer **1 Terabyte of data** across various speed networks:

| | |
|---|---|
| **10 Mbps** network | **300 hrs (12.5 days)** |
| **100 Mbps** network | **30 hrs** |
| **1 Gbps** network | **3 hrs** |
| **10 Gbps** network | **20 minutes** |



**Data set size**

| | | | | |
|---|---|---|---|---|
| 10PB | 1,333.33 Tbps | 266.67 Tbps | 66.67 Tbps | 22.22 Tbps |
| 1PB | 133.33 Tbps | 26.67 Tbps | 6.67 Tbps | 2.22 Tbps |
| 100TB | 13.33 Tbps | 2.67 Tbps | 666.67 Gbps | 222.22 Gbps |
| 10TB | 1.33 Tbps | 266.67 Gbps | 66.67 Gbps | 22.22 Gbps |
| 1TB | 133.33 Gbps | 26.67 Gbps | 6.67 Gbps | 2.22 Gbps |
| 100GB | 13.33 Gbps | 2.67 Gbps | 666.67 Mbps | 222.22 Mbps |
| 10GB | 1.33 Gbps | 266.67 Mbps | 66.67 Mbps | 22.22 Mbps |
| 1GB | 133.33 Mbps | 26.67 Mbps | 6.67 Mbps | 2.22 Mbps |
| 100MB | 13.33 Mbps | 2.67 Mbps | 0.67 Mbps | 0.22 Mbps |
| | 1 Minute | 5 Minutes | 20 Minutes | 1 Hour |

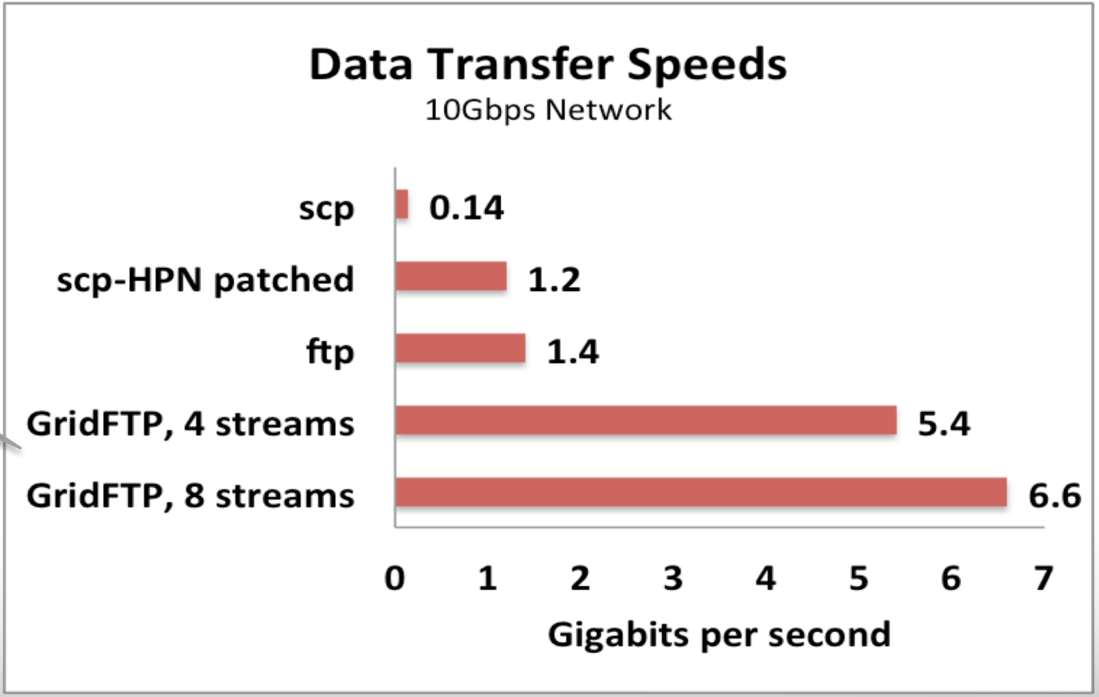
**Time to transfer**

Source: http://fasterdata.es.net/home/requirements-and-expectations

SCP and FTP variants all use TCP, which places a premium on being a “good neighbor” and not hogging the network. The result is reliability, not performance. Encryption also reduces the overall throughput for transfers.

# Use the right tool…

Berkeley, CA ←→ Argonne, IL; RTT=53ms



Globus uses GridFTP and tries to optimize using multiple TCP streams

# Some lessons learned and observations about storage systems

- Regardless of where you work, all those file types will follow
  - Small files
  - Large files (>600GB)
  - Directories with millions of files
  - Spreadsheets
  - “Structured” flat files
  - Very large binary files
  - Very large text files
- But a given storage system will usually only handle a few of these well
- OK, then what about Metadata? Keeping track of your stuff will need attention, thought, planning and automation. “Storage is cheap, Metadata are precious.”    (The next thing, may be big.)

Source: Riffing on Chris Dadigian, Bioteam.

# More observations

- Shipping disk drives is dangerous for your data, though Amazon will come and pick it up for you (https://aws.amazon.com/snowmobile/). Great bandwidth, terrible latency.

- Using the network (under the right conditions) is still the better option.
- There is no easy way to determine a file's "goodness" except hashes or checksums, although these can be automated to an extent, e.g. Globus (globus.org), during network copies.

# Summary

- Who knew data management was so complicated?
- In research, data management is critical to success, or lack of attention can lead to trouble
- Three main aspects of data management are
  - How/where you store your data
  - How you annotate your data for understanding and findability
  - Moving your data has some non-trivial aspects if you have a lot of it
- An emerging part of data management is saving computational methods and code; can be done now by saving fully-configured virtual machines and containers.

# Questions?