

Data Management Practices Primer

Windows Users:

To follow along, please download Mobaxterm
(a free SSH client)

Google Search “Mobaxterm” or navigate to:
<https://mobaxterm.mobatek.net/download.html>

Download the “Installer edition” (green button)

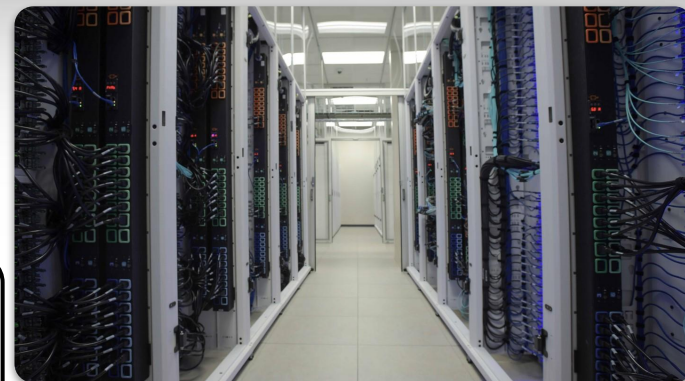


MobaXterm Home Edition v20.2
(Installer edition)

HPRC's Newest Cluster

Grace is a 925-node Intel cluster from Dell with an InfiniBand HDR-100 interconnect, A100 GPUs, RTX 6000 GPUs and T4 GPUs. There are 925 nodes based on the Intel Cascade Lake processor.

Grace
3TB Large Memory-80 cores/nodes
Other Login Nodes-48 cores/node



Login Nodes	5
384GB memory general compute nodes	800
GPU - A100 nodes with 384GB memory	100
GPU - RTX 6000 nodes with 384GB memory	9
GPU - T4 nodes with 384GB memory	8
3TB Large Memory	8

For more information:

<https://hprc.tamu.edu/wiki/Grace:Intro>

Logging in to the system

- SSH (secure shell)
 - freely available for Linux/Unix and Mac OS X hosts

For Microsoft Windows PCs, use *MobaXterm*

- or *Putty*

VPN needed for off campus access

- https://u.tamu.edu/VPN_help

Using SSH (on a Linux/Unix Client)

<https://hprc.tamu.edu/wiki/Grace:Access>

```
ssh user_NetID@grace.tamu.edu
```

You may see something like the following the first time you connect to the remote machine from your local machine:

```
Host key not found from the list of known hosts.  
Are you sure you want to continue connecting (yes/no)?
```

Type yes, hit enter and you will then see the following:

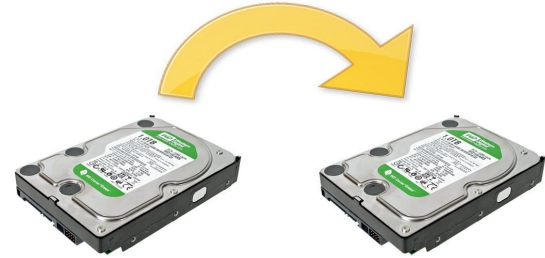
```
Host 'grace.tamu.edu' added to the list of known hosts.  
user_NetID@grace.tamu.edu's password:
```

Good Data Practice

Rule of thumb:

1 is none

2 is one



Keep multiple copies of important data!

Having just one copy is not enough

Backup Backup Backup

Data on Our Clusters: Grace

There are limits on data on our clusters → AKA quota
The limits are on *Disk Space & File Usage*

showquota

View your current quota with this command

Your current disk quotas are:

Disk	Disk Usage	Limit	File Usage	Limit
/home	416.1M	10G	4489	10000
/scratch	18.64G	1T	122616	250000

Need more space?

Submit a ***Quota Increase Request***

Contact help@hprc.tamu.edu

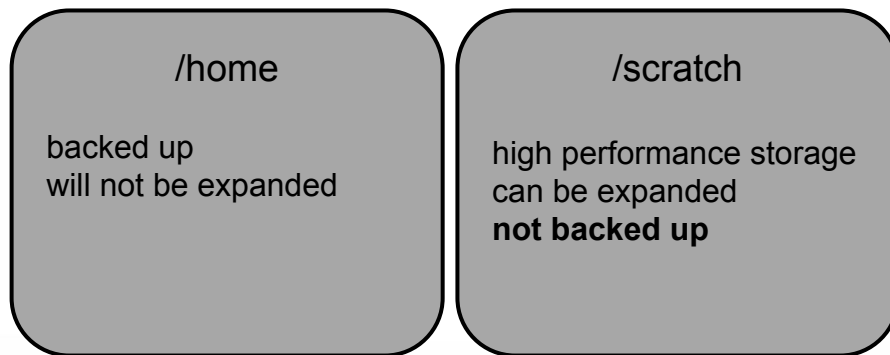
Data on Our Clusters: Grace

Default Limits

/home	10G / 10,000 files
/scratch	1T / 250,000

Data on Our Clusters: Grace

What's the difference between these filesystems?



Need more space?

Submit a ***Quota Increase Request***

Contact help@hprc.tamu.edu

Data Transfer: Grace

Grace's login nodes have 10 Gigabit Ethernet to the TAMU Network
scp - sftp - *rsync* are all available

Login nodes have a 60 minute process limit

rsync is preferred → supports intermittent transfer

```
rsync source_file destination
```

GUI transfer programs are easiest for new users
WIN SCP, MobaXterm, HPRC Web Portal

Data Transfer: Grace DTN

Grace has 2 nodes dedicated to data transfer → Data Transfer Nodes
Accessible via Grace's login nodes

SSH to either DTN from any Grace login node:

```
ssh netID@grace-dtn1.tamu.edu  
ssh netID@grace-dtn2.tamu.edu
```

Large Transfers should use the Data Transfer Nodes

Both nodes have 40 gigabit capability

No programming environment installed → these are for transfers only!

These nodes have access to all of Grace's filesystem

/home

/scratch

Data on Our Clusters: Terra

There are limits on data on our clusters → AKA quota
The limits are on *Disk Space* & *File Usage*

showquota

View your current quota with this command

Your current disk quotas are:

Disk	Disk Usage	Limit	File Usage	Limit
/home	416.1M	10G	4489	10000
/scratch	18.64G	1T	122616	250000

Data on Our Clusters: Terra

Default Limits

/home	10G / 10,000 files
/scratch	1TB / 250,000

Need more space?

Submit a ***Quota Increase Request***

Contact help@hprc.tamu.edu

Data on Our Clusters: Terra

What's the difference between these filesystems?

/home

backed up nightly
will not be expanded

/scratch

'high performance storage'
can be expanded
not backed up

Data Transfer: Terra

Terra's login nodes have 10 Gigabit Ethernet to the TAMU Network
scp - sftp - *rsync* are all available

Login nodes have a 60 minute process limit

rsync is preferred → supports intermittent transfer

```
rsync source_file destination
```

GUI transfer programs are easiest for new users
WIN SCP, MobaXterm, HPRC Web Portal

Data Transfer: Terra FTN

Terra has 1 node dedicated to data transfer → Fast Transfer Node
No process time limit

SSH to Terra's FTN directly:

```
ssh netID@terra-ftn.hprc.tamu.edu
```

Large Transfers should use the Data Transfer Nodes

The node has 10 gigabit capability

No programming environment installed → these are for transfers only!

These nodes have access to all of Terra's filesystem

/home

/scratch

Pop Quiz

What is the process limit for the login node?

- A. 45 minutes
- B. 75 minutes
- C. 60 minutes
- D. No process limit

Pop Quiz

What is the process limit for the login node?

- A. 45 minutes
- B. 75 minutes
- C. 60 minutes
- D. No process limit

Command Line Tools

1. `cp` -- copy
2. `rm` -- remove
3. `scp` -- secure copy (remote copy)
4. `sftp` -- secure file transfer
5. `tar` -- archiving

Command Line Tools: cp

Copy

Makes a copy of a file

```
cp source_file new_fileName
```

Easy solution for copying a file onto the *same machine*

How about moving data between machines?

Command Line Tools: rm

Remove

Deletes a file

```
rm some_file
```

Completely deletes a file

There is no “trash bin” on the command line
add the -i flag to be prompted prior to file deletion

```
rm -i some_file
```

Command Line Tools: scp

Secure copy

Copies files between hosts on a network -- uses ssh for data transfer (hence "Secure")

```
scp source_file netId@grace.tamu.edu:/home/netID
```

Can be used to copy:

- from local to remote
- from remote to local
- between 2 remote systems from local system

Command Line Tools: sftp

Secure file transfer program

interactive file transfer program -- uses ssh (again so hence “secure”)

```
sftp netID@grace.tamu.edu
```

Connects and logs into specified host, enters command mode

- cd - change directory
- get - download file
- put - upload file
- bye - quit sftp

Command Line Tools: tar

Archiving files

saves many files together into a single file (archive)

```
tar -cvf archive.tar source
```

create a compressed archive

```
tar -czvf archive.tar.gz source
```

extract an archive

```
tar -xvf archive.tar
```

Useful for consolidating (and compressing) files prior to transfer

Important flags

-cf	create archive
-xf	extract archive
-v	verbose
-z	compress with gzip

GUI Clients

There are many GUI solutions for file transfer:

1. MobaXterm
2. WinSCP
3. FileZilla
4. Cyberduck

Globus Connect

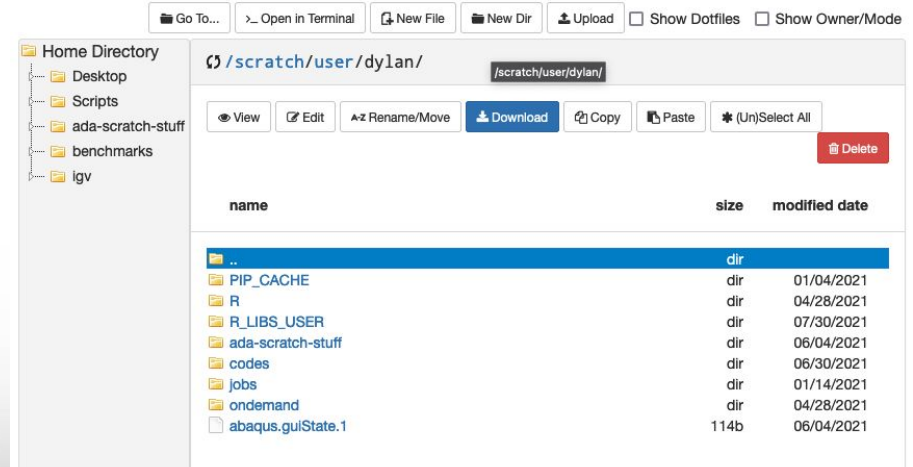
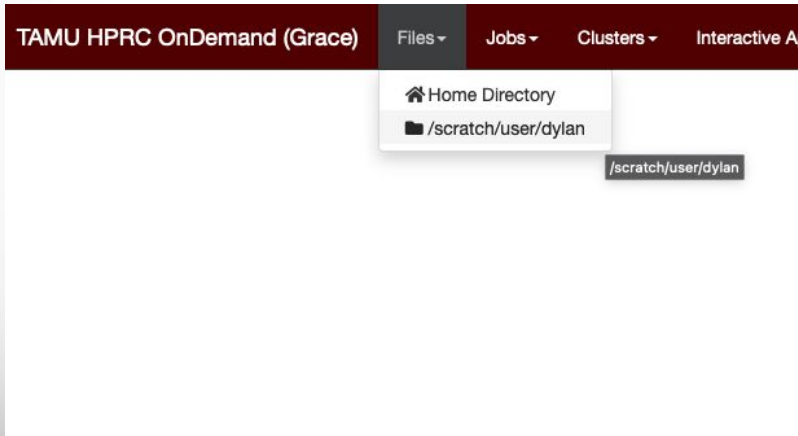


GUI Clients: HPRC Portal

Access your files through (almost)
any web browser

View, Edit, Upload, Download, Remove
through the Portal

<https://portal.hprc.tamu.edu>

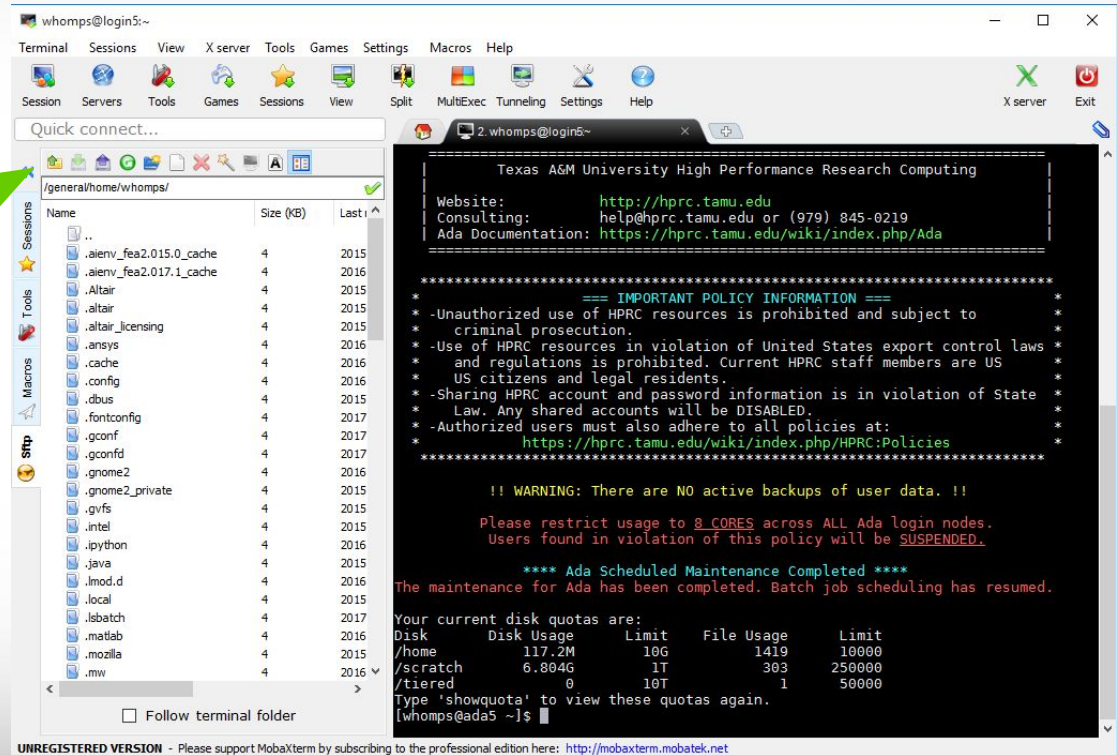


GUI Clients: MobaxTerm

Available on Windows machines

SFTP side panel in
MobaxTerm

Can download, upload
files with a few clicks
from the CLI



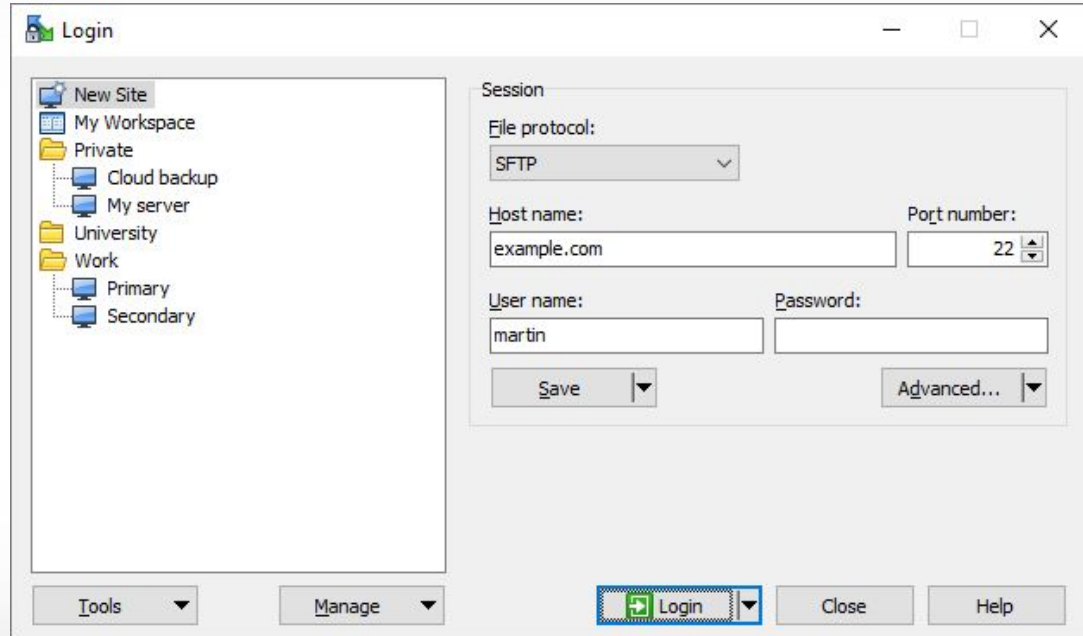
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <http://mobaxterm.mobatek.net>

GUI Clients: WinSCP

Available on Windows machines

Connects to host directly with SFTP

Allows for transfers through the GUI



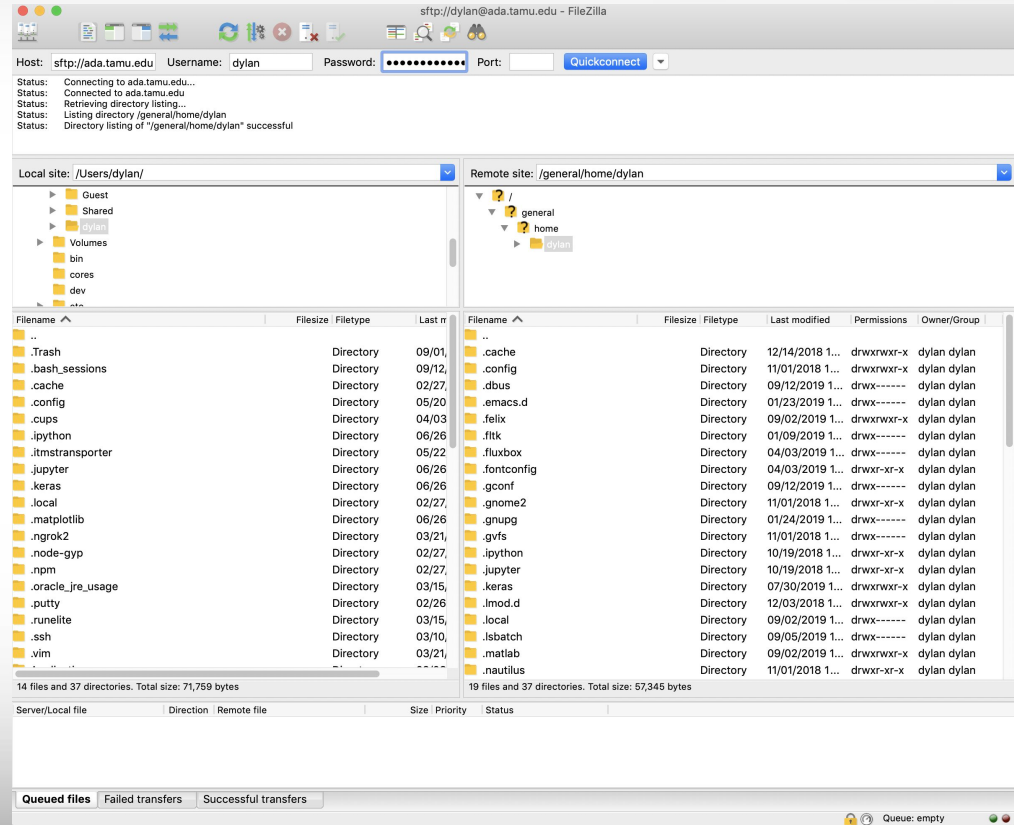
GUI Clients: FileZilla

Open source - available on all platforms

2 Factor Authentication
makes FileZilla hard to use

Connects to host directly with SFTP

Allows for transfers through the GUI

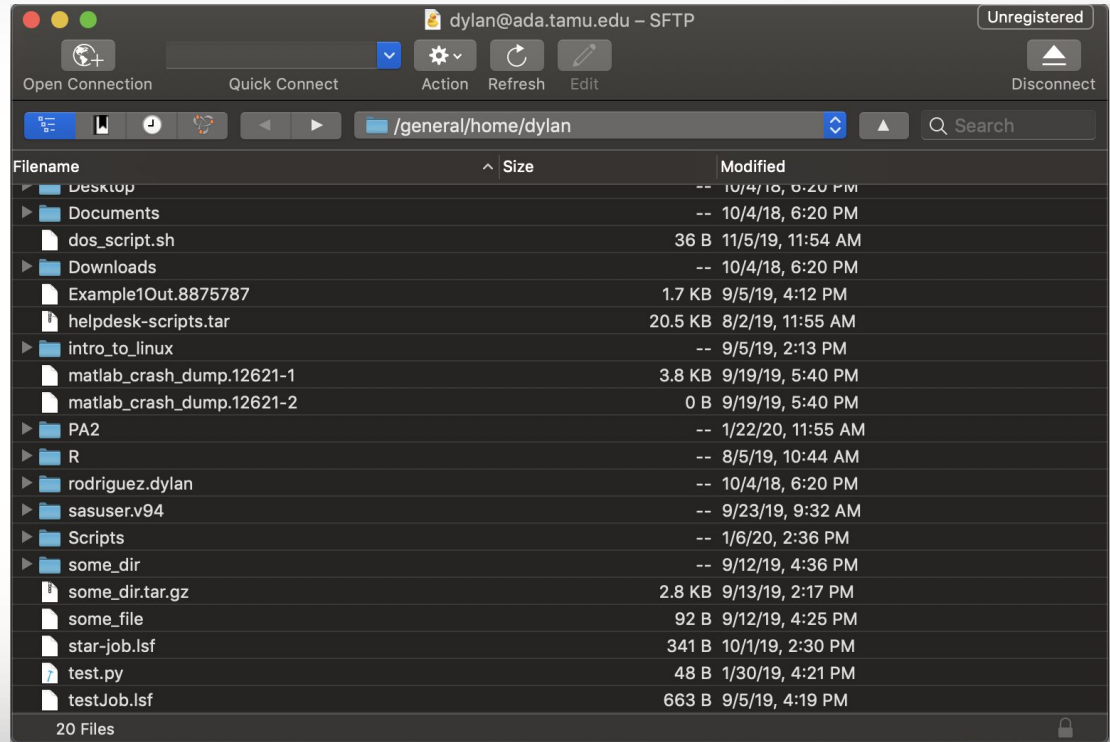


GUI Clients: CyberDuck

Available on Windows & MacOS

Connects to host directly with SFTP

Allows for transfers through the GUI



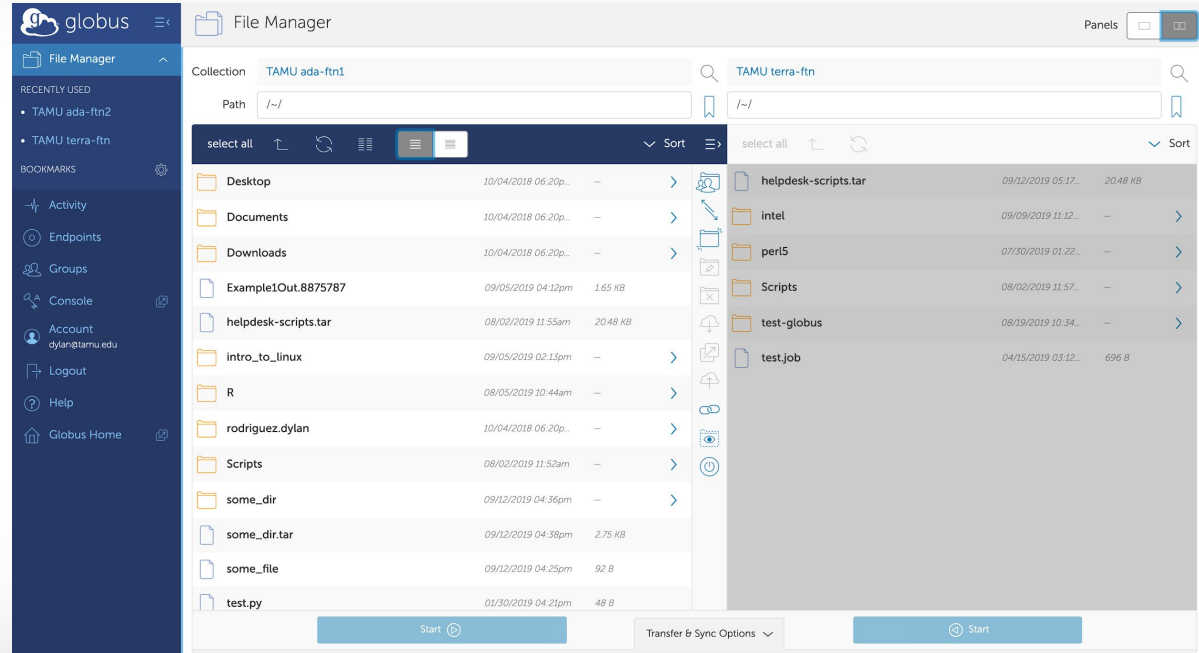
GUI Clients: Globus

Web based, with application
you can download

Grace endpoints:
grace-dtn1
grace-dtn2

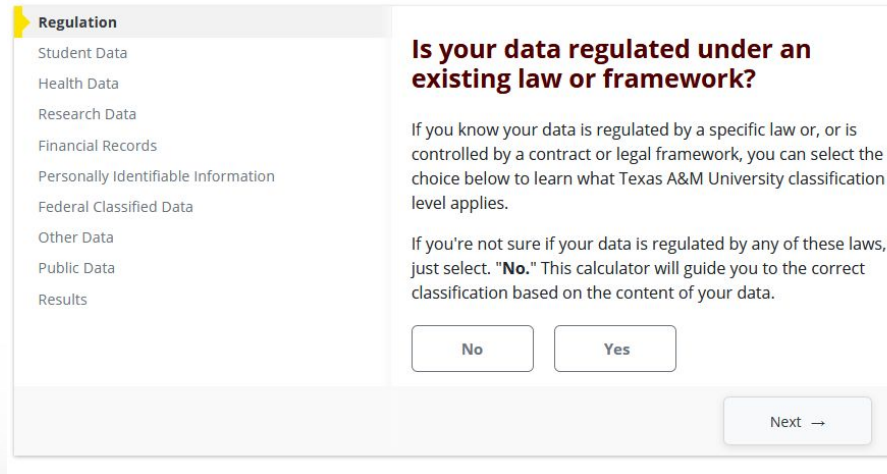
Terra endpoint
terra-ftn

<https://www.globus.org/>



Data Classification Tool

The process of sorting and categorizing data based on the sensitivity of information and the impact of potential loss



The screenshot shows a web interface for the Data Classification Tool. On the left is a sidebar with a yellow folder icon at the top, followed by a list of categories: Regulation (highlighted), Student Data, Health Data, Research Data, Financial Records, Personally Identifiable Information, Federal Classified Data, Other Data, Public Data, and Results. The main content area is titled "Is your data regulated under an existing law or framework?" in bold. Below the title is a paragraph explaining that if data is regulated by a specific law or contract, the user should select a choice to determine the Texas A&M University classification level. Another paragraph states that if the user is unsure, they should select "No," as the calculator will guide them to the correct classification. At the bottom of the main area are two buttons: "No" and "Yes". A "Next →" button is located at the bottom right of the interface.

Can be found at:

<https://it.tamu.edu/community/tools/data-classification.php>

Questions?

Continued Learning

[Intro to HPRC Video Tutorial Series](#)

[HPRC's Wiki Page](#)