Introduction to Data Literacy and Data Management

DIVISION OF RESEARCH
TEXAS A&M UNIVERSITY

v1.0 18/09/2017
mcmullen@tamu.edu

Contact the HPRC Helpdesk

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

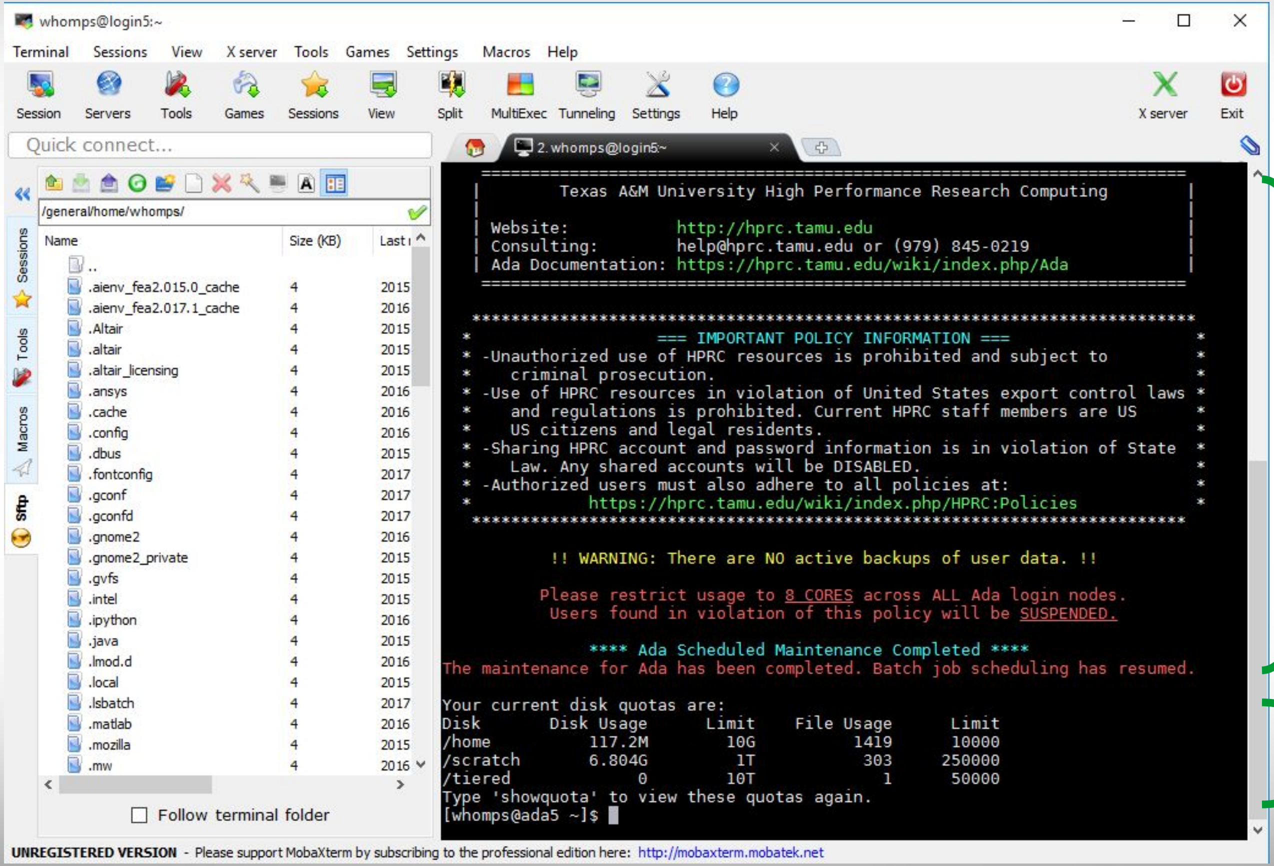
Logging in to the system

- SSH (secure shell)
  - The only program allowed for remote access; encrypted communication; freely available for Linux/Unix and Mac OS X hosts;
- For Microsoft Windows PCs, use *MobaXterm*
    - **https://hprc.tamu.edu/wiki/HPRC:MobaXterm**
      - You are able to view images and use GUI applications with MobaXterm
  - or *Putty*
    - **https://hprc.tamu.edu/wiki/HPRC:Access#Using_PuTTY**
      - You can not view images or use GUI applications with PuTTY

Your Login Password

- Both state of Texas law and TAMU regulations prohibit the sharing and/or illegal use of computer passwords and accounts
- Don’t write down passwords
- Don’t choose easy to guess/crack passwords
- Change passwords frequently

Using SSH - MobaXterm (on Windows)



message of the day

your quotas

Using SSH (on a Linux/Unix Client)

https://hprc.tamu.edu/wiki/Ada:Access

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

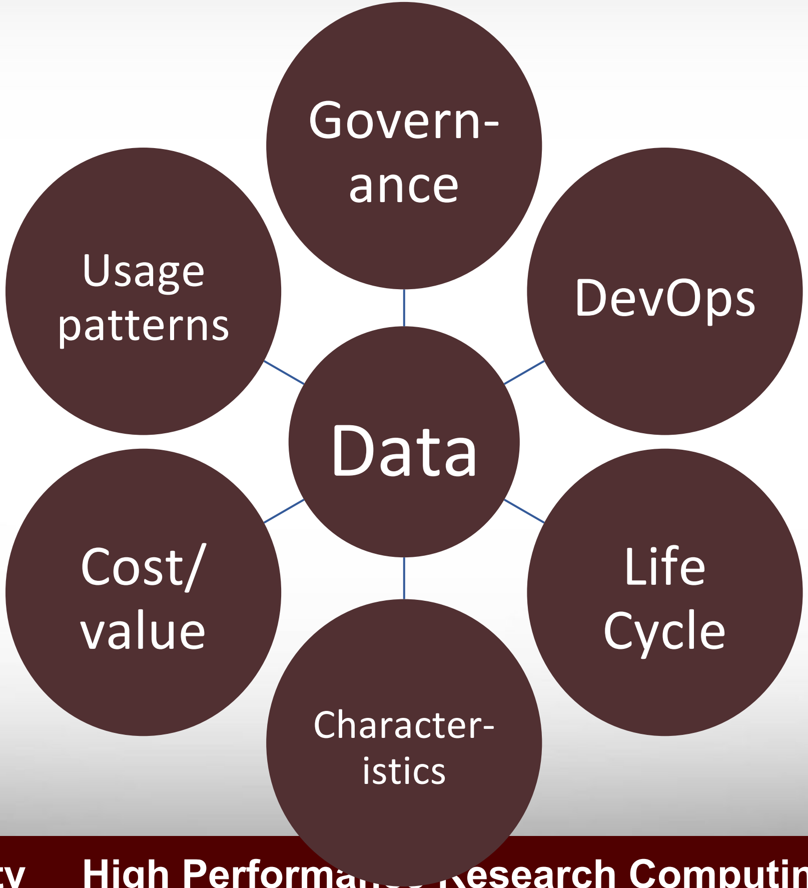
Goals for the hour – What are yours?

- 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

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?

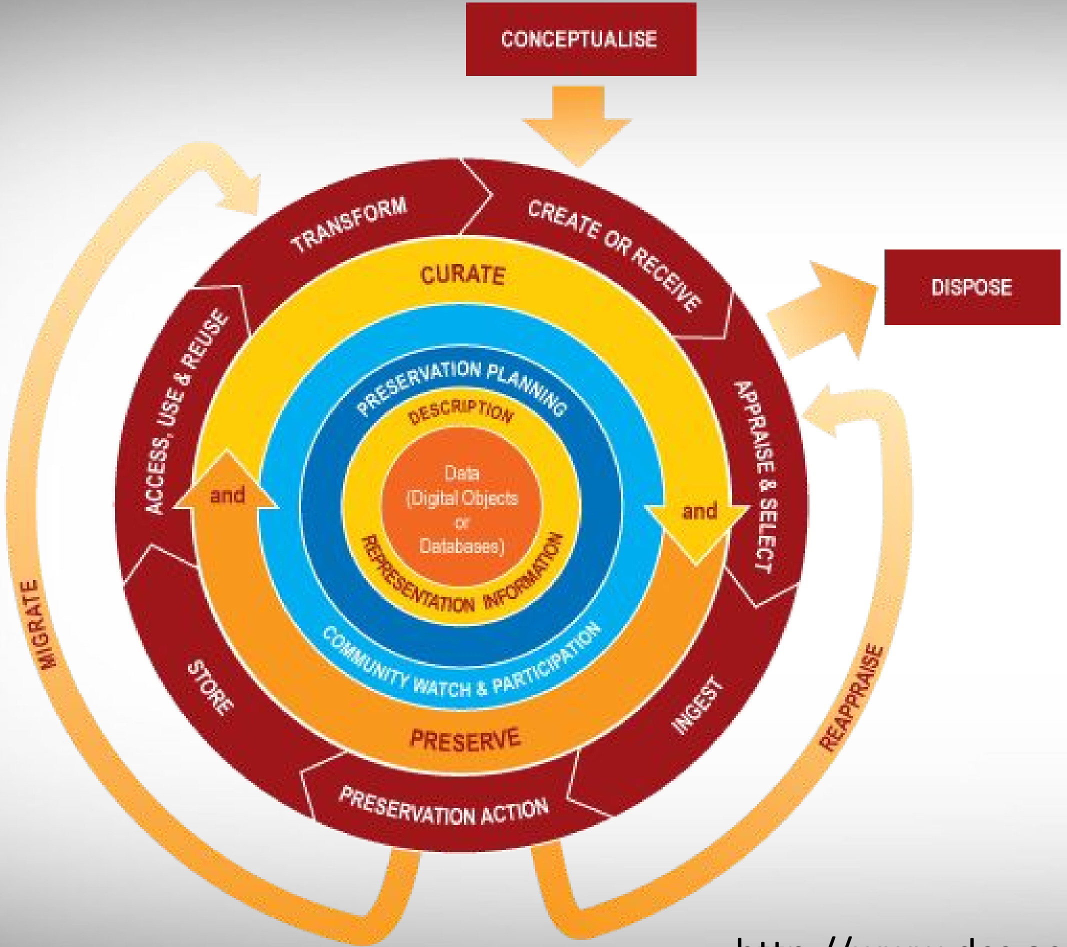
Lenses for looking at your information assets



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 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
- Moving your data

Storing your data

Ways to store and organize your data

- Spreadsheets: Excel, OpenOffice, Google Sheets
  - Convenient, ubiquitous, easy to use
  - Easy to lose track of
- 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
- 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-prem
    - 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.

Data Cleaning

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

OPEN Refine

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

and performing extract-transform-load (ETL) tasks.
Also Talend Open Studio (https://www.talend.com/download/talend-open-studio/)

- Save everything!

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
- Jetstream (NSF) - https://jetstream-cloud.org/
  - Jetstream allows VMs to be catalogued as publications (get a doi, keep in a repository)
- CyVerse - http://www.cyverse.org/
- VMs can be saved in a ”portable” format, Open Virtualization Format
- End-user VM systems: VirtualBox, KVM/qemu (free), VMWare (cost)

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 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

- 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

Expected 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**

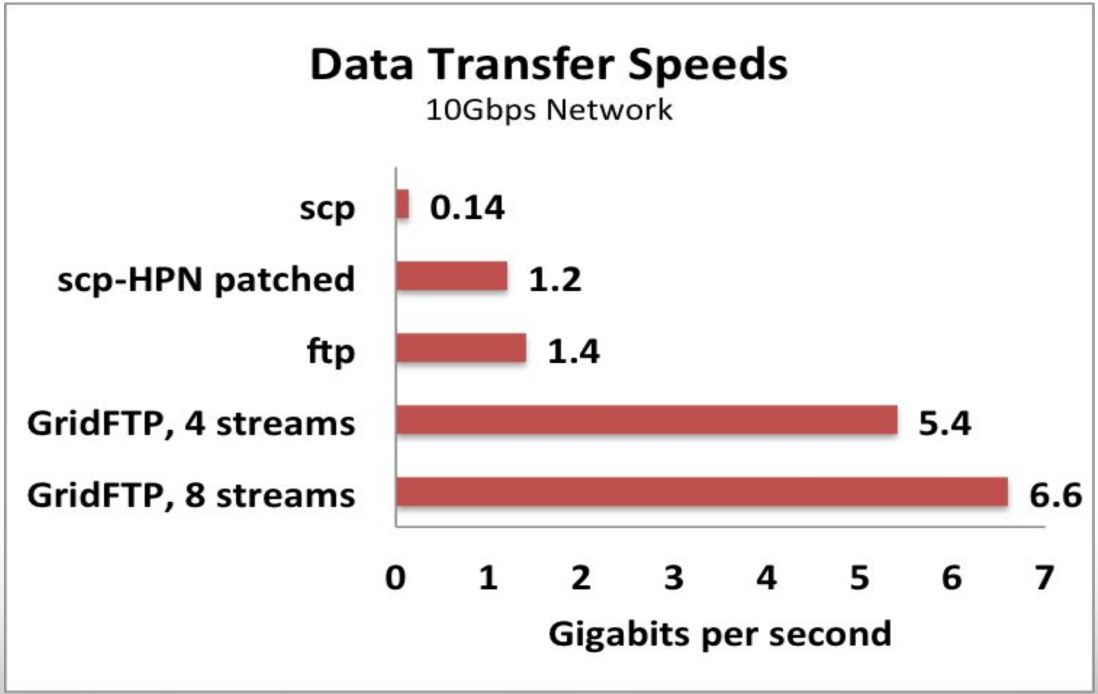
Data set size				
10PB	1,333.33 Tbps	266.67 Tbps	66.67 Tbps	22.22 Tbps
1PB	133.33 Tbps	26.67 Tbps	6.67 Tbps	2.22 Tbps
100TB	13.33 Tbps	2.67 Tbps	666.67 Gbps	222.22 Gbps
10TB	1.33 Tbps	266.67 Gbps	66.67 Gbps	22.22 Gbps
1TB	133.33 Gbps	26.67 Gbps	6.67 Gbps	2.22 Gbps
100GB	13.33 Gbps	2.67 Gbps	666.67 Mbps	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

Use the right tool…

Berkeley, CA ←→ Argonne, IL RTT=53



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.

Questions?