Data Transfer and Management

University of Wyoming
Advanced Research Computing Center (ARCC)
arcc-info@uwyo.edu
(307) 766-7748

Presented June 23rd , 2016

Agenda

- Data Transfer
- Data Management
- Some Best Practices
- Q&A

Data Transfer and Access

- scp
- sftp
- rsync
- Science DMZ
- OS Specific (CIFS/SMB)
- Globus

sftp

- sftp is a command-line interface client program to transfer files using the SSH File Transfer Protocol (SFTP) as implemented by the sftp-server command by the OpenSSH project, which runs inside the encrypted Secure Shell connection.

- It provides an interactive interface similar to that of traditional FTP clients

- should not be confused with running an FTP client over an SSH connection.

sftp

```
SFTP username@mtmoran.uwyo.edu:/project/bc-201606/DataMangement
```

1)Gets txt file from Big Horn to Local System

```
get rosters_14.csv Dylan_rosters.csv
```

2)Puts txt file from Local System on to Big Horn

```
put source_directory/Dylan_rosters.csv
```

scp

- Secure copy or SCP is a means of securely transferring computer files between a local host and a remote host or between two remote hosts.
  - It is based on the Secure Shell (SSH) protocol

- Normally, a client initiates an SSH connection to the remote host, and requests an SCP process to be started on the remote server. The remote SCP process can operate in one of two modes:
  - source mode, which reads files (usually from disk) and sends them back to the client
  - sink mode, which accepts the files sent by the client and writes them (usually to disk) on the remote host

scp

1) Local

```
scp Dylan_rosters.csv data/perkins_rosters.csv
```

2) MtMoran

```
Scp Dylan_rosters.csv username@mtmoran.uwyo.edu:/project/bc-201606/DataMangement
```

rsync

- rsync is a utility to keep copies of a file on two computer systems
  - functions as both a file synchronization and file transfer program
- The rsync algorithm is a type of delta encoding, and is used to minimize network usage. Zlib may be used for additional compression, and SSH or stunnel can be used for data security
- Rsync is typically used to synchronize files and directories between two different systems.
  - For example, if the command rsync local-file user@remote-host:remote-file is run, rsync will use SSH to connect as user to remote-host

rsync

1)To sync the contents of dir1 to dir2 on the same system,

```
rsync –r Dylan_rosters.csv data/perkins_rosters.csv
```

2) To sync with a remote system

```
rsync -a Dylan_rosters username@mtmoran.uwyo.edu:/project/bc-201606/DataMangement
```

Science DMZ

https://arcc.uwyo.edu/guides/uw-science-dmz

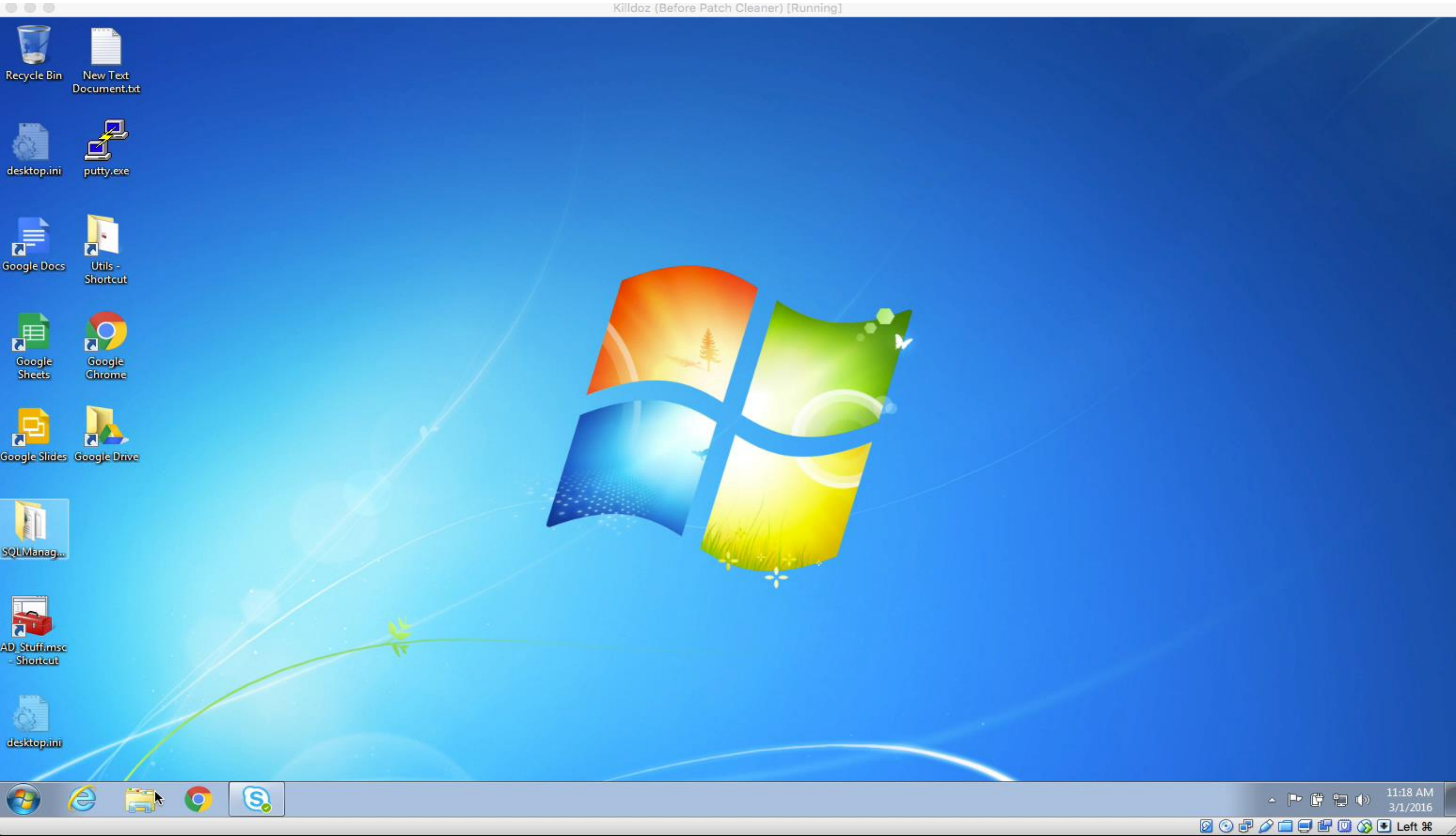
The University of Wyoming (UW) Science Network (UWSN) is the campus implementation of the science DMZ principle. The guiding principle behind UWSN is that the campus research community should be able to optimize their access to other research entities, data stores and computing resources specific to the needs of their individual projects. To accomplish this they must overcome barriers of bandwidth constraint, high-latency, or the restrictions of an active security perimeter and be provisioned with the most efficient network possible.

A "Science DMZ" ("science demilitarized zone") is a portion of a larger network that has been configured and optimized for high-volume bulk data transfer, remote experiment control, and data visualization for high-performance science applications. A science DMZ should be scalable, incrementally deployable, and adaptable to new technologies. In order to achieve the maximum speed and throughput possible, the UW implementation of the science DMZ model avoids the regular campus exit architecture, and is neither restricted nor protected by the campus firewalls.

Connecting with Windows

- Addresses for Windows start with ‘\\’
  - E.g. \\warehouse.uwyo.edu\
- Through Windows Explorer
- Through a mapped drive

Data Transfer and Management



Connecting with Mac OS X+

- Addresses for Mac’s start with smb:// (or other protocols)
- Finder

Data Transfer and Management

Connecting with Linux

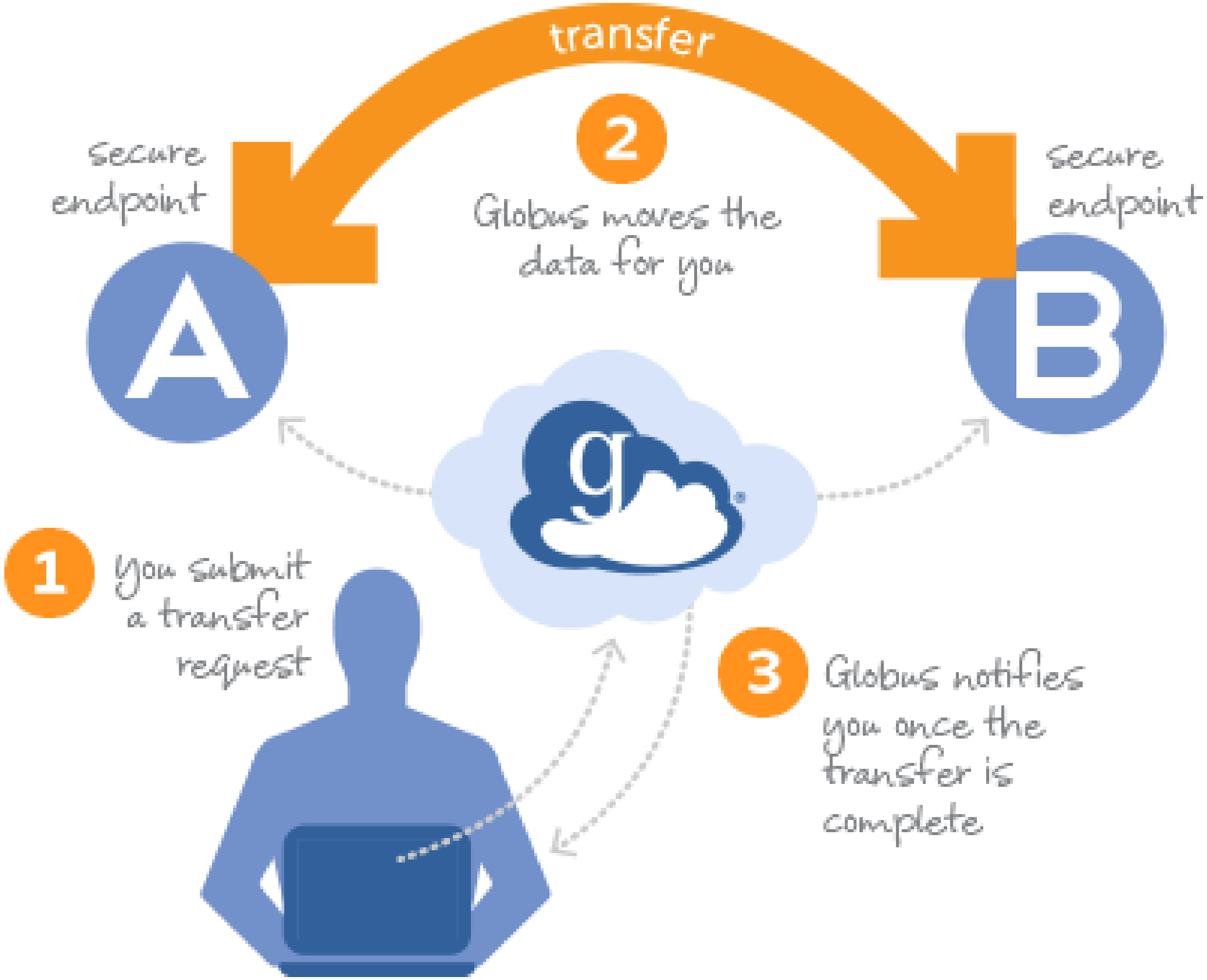
- Addresses for Linux can start with smb://, ”\\”, or ”//”, as well as many others.
- Gnome/nautilus
- Command line mount (Email arcc-info@uwyo.edu if help is required)

Data Transfer and Management

Globus

- ARCC’s preferred method of data transfer
- Web-based, or command line
- Many benefits
  - High performance
  - Simple, intuitive interface
  - Restart Jobs
  - “set and forget”
  - Many research institutions have endpoints
  - Takes advantage of Science DMZ
  - Personal Endpoints/Sharing
- Broad topic - Training available for Globus by ARCC

what is Globus


transfer
secure endpoint
2
Globus moves the data for you
secure endpoint
A
B
1
you submit a transfer request
3
Globus notifies you once the transfer is complete

a few useful Globus Terms

endpoint : a logical address for a GridFTP server, similar to a domain name for a web server. Data is transferred between Globus endpoints.

Globus Connect Personal : a client for communicating with other GridFTP servers, via your local computer using Globus. Installing Globus Connect Personal on your computer creates your own endpoint that you can use to transfer data to and from your computer.

Globus Connect Server: Globus Connect Server – for multiuser environments – a Linux package that sets up a GridFTP server for use with Globus. Once installed on a server, those with access to the server can move data to and from this location.

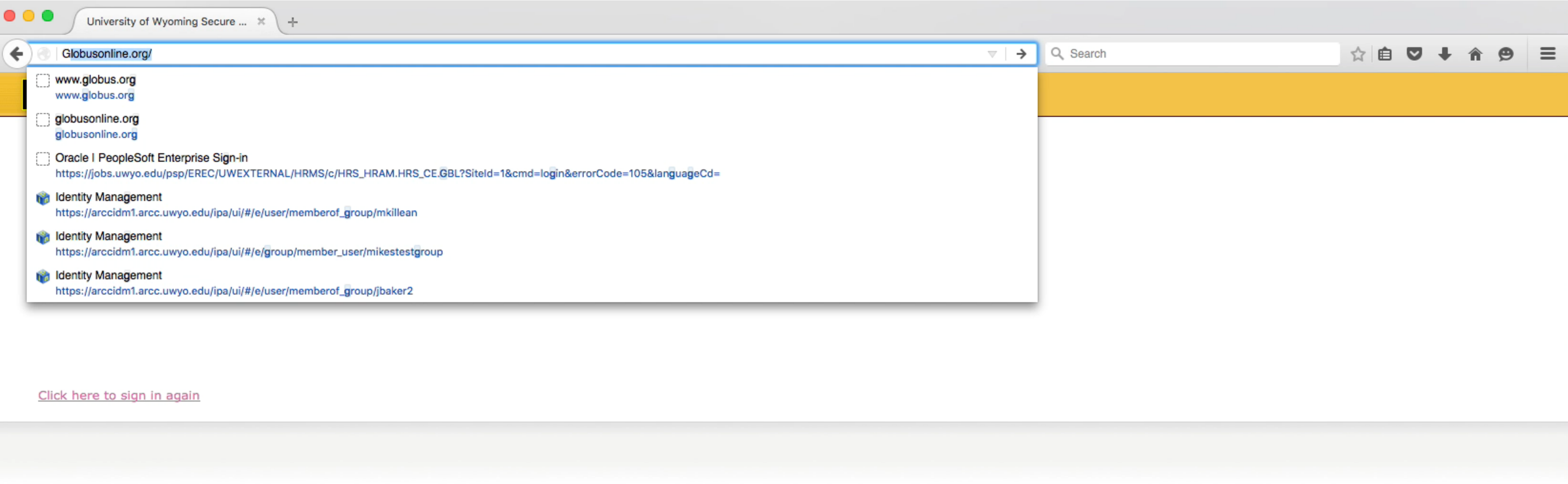
what is GridFTP

GridFTP is an extension of the standard File Transfer Protocol (FTP) for high-speed, reliable, and secure data transfer. Because GridFTP provides a more reliable and high performance file transfer (compared to protocols such as SCP or rsync), it enables the transmission of very large files. GridFTP also addresses the problem of incompatibility between storage and access systems

time to Explore Globus

1. Accessing Globus
2. First simple transfer (National Lab to UWYO)
3. Personal endpoint setup
4. Managing transfers
5. Additional Questions

Data Transfer and Management



University of Wyoming Secure ...

Globusonline.org/

www.globus.org
www.globus.org

globusonline.org
globusonline.org

Oracle I PeopleSoft Enterprise Sign-in
https://jobs.uwyo.edu/psp/EREC/UWEXTERNAL/HRMS/c/HRS_HRAM.HRS_CE.GBL?SiteId=1&cmd=login&errorCode=105&languageCd=

Identity Management
https://arccidm1.arcc.uwyo.edu/ipa/ui/#/e/user/memberof_group/mkillean

Identity Management
https://arccidm1.arcc.uwyo.edu/ipa/ui/#/e/group/member_user/mikestestgroup

Identity Management
https://arccidm1.arcc.uwyo.edu/ipa/ui/#/e/user/memberof_group/jbaker2

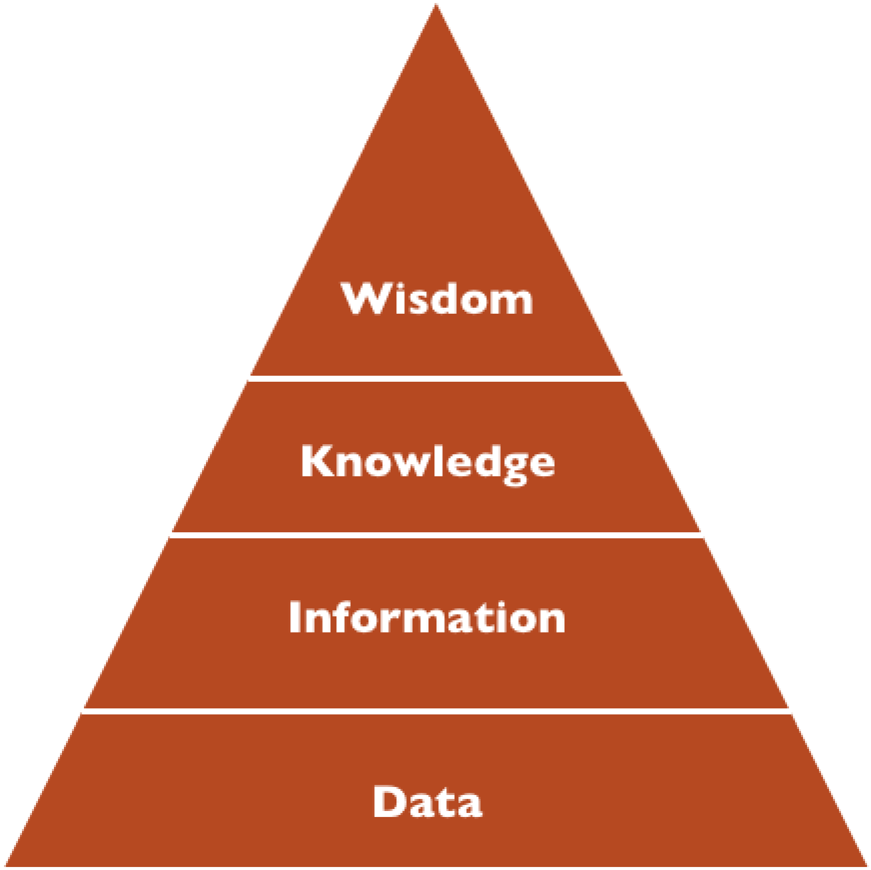
Click here to sign in again

Best Practices

- Shared Systems – Be a friendly neighbor
- Let ARCC know about large transfers (10+ TB)
- If no longer need data remove or archive it.
- If you aren’t using data actively, archive/zip/tar it to reduce load on the servers
    - Data with thousands of files can be problematic
    - thousands of files within a single directory will drastically impact performance
- Backup important data (if not stored on ARCC’s systems).

- Explicit Metadata and Data Management Planning (Dylan’s Class)
- Cite Mt. Moran - https://arcc.uwyo.edu/guides/citing-mount-moran

Why manage Data



Data Management

- **Data Management** is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise.

Data Management

- Data Management is a process of:
  - Efficiently store and organize data
  - Correctly format data
  - Annotation of data (i.e. Metadata)
  - Data archiving and preservation
  - **Data accessibility (researchers & public)**

Data Management

- White House Office of Science and Technology Policy (2013)
  - Ensuring that publications and research data are **<u>accessible to the public</u>**.

- The National Science Foundation (NSF) clearly states in their post award requirements that grant recipients are **<u>expected to share data</u>** within a reasonable time frame (http://www.nsf.gov/).

- Many publishers now **<u>require</u>** data to be shared before manuscripts can be published in a journal.

- Open-data advances science!

Funding Requirements

- Many funding sources require a data management plan **(DMP)**
  - Including the sharing of primary data, software, and other products produced under the grant(s)
  - Often incentives are offered for data cleanup, documentation, dissemination, and storage

NATIONAL ENDOWMENT FOR THE Humanities

NASA

U.S. DEPARTMENT OF ENERGY

NSF

NIH National Institutes of Health
Office of Extramural Research

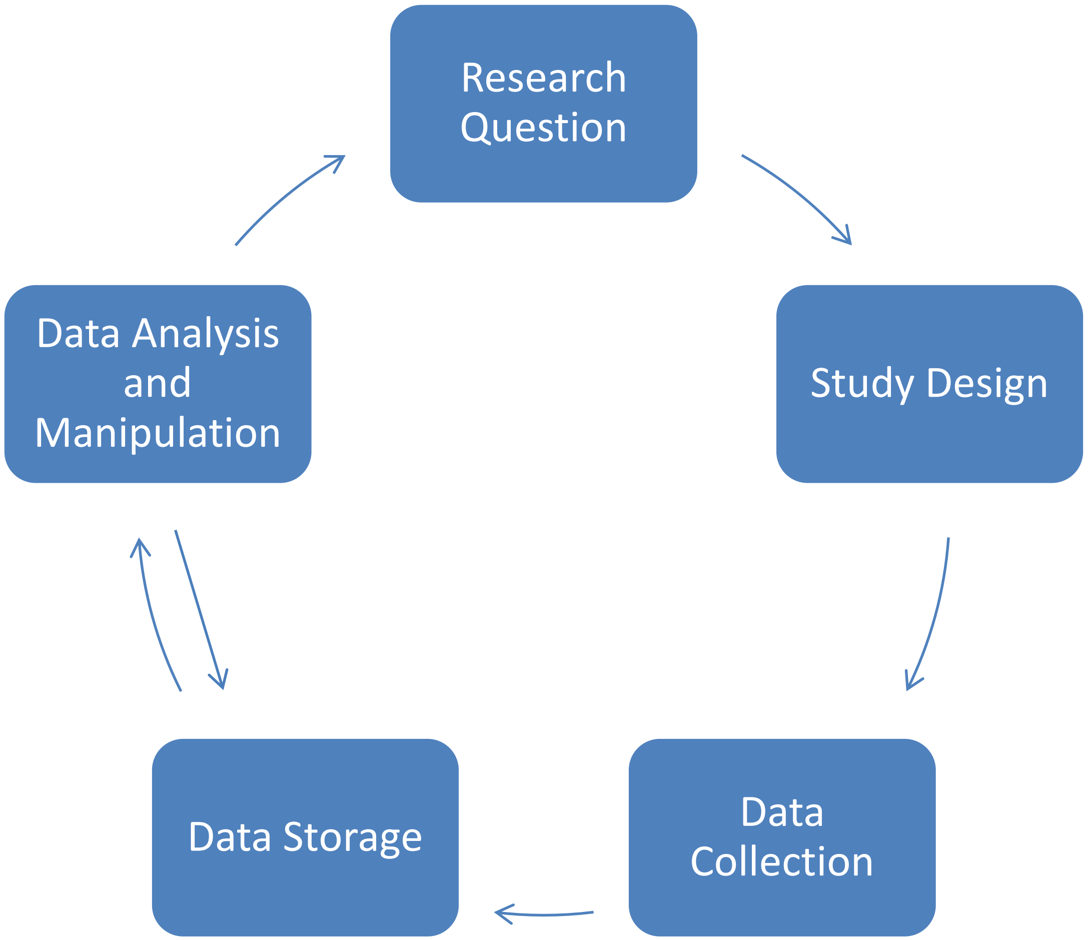
Data Management

- Define the contents of your data
- Assign descriptive data set titles
- Use consistent data organization
- Preserve the raw data
- Protect your data
- Provide documentation and metadata
- Perform basic quality assurance

DMPTool.org

- UW has recently become a member intuition of DMPTool.org
  - The original DMPTool was a grassroots effort, beginning in January 2011 with eight institutions partnering to provide in-kind contributions of personnel and development.
  - https://dmptool.org/video

Research Workflow


Research Question

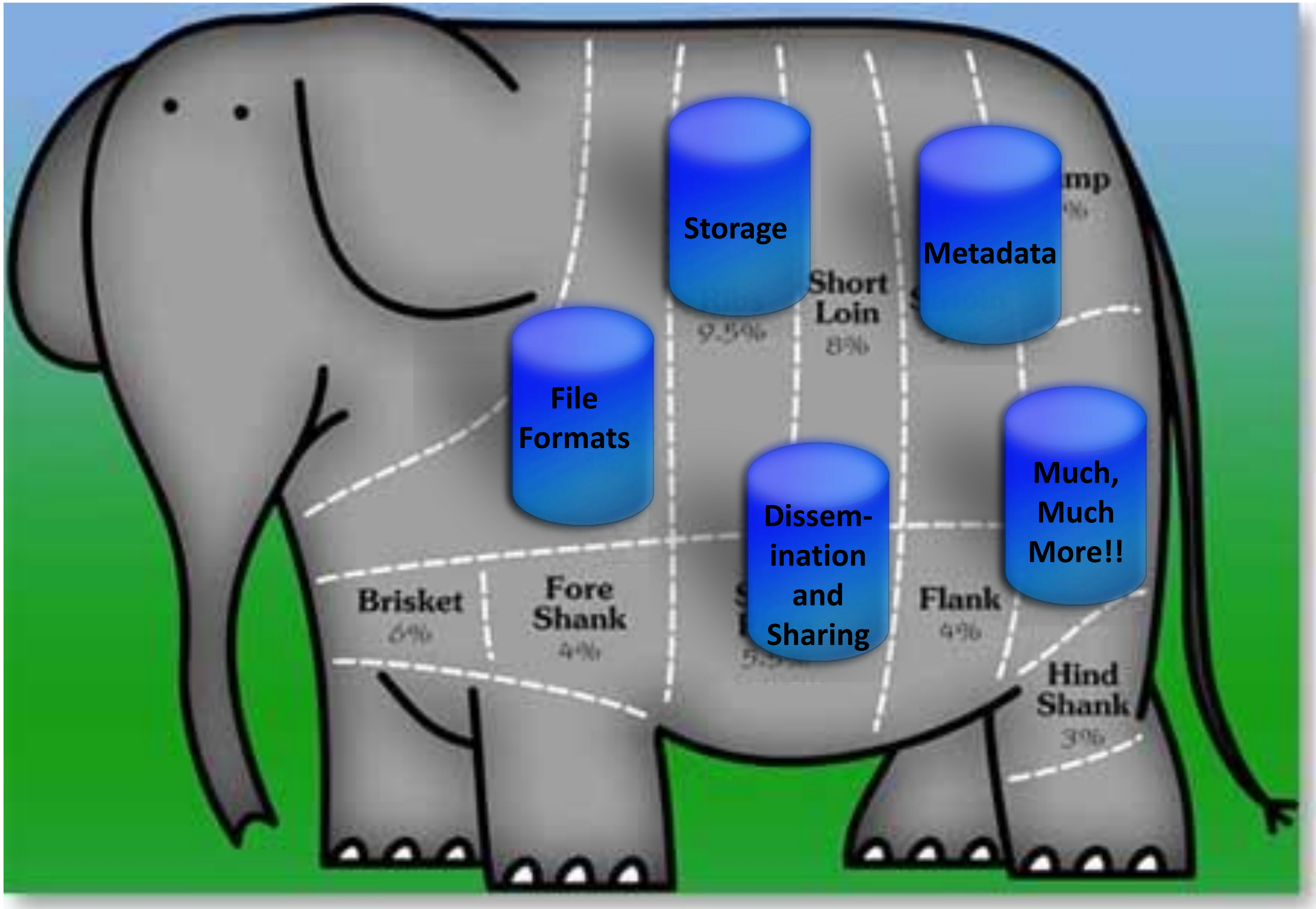
Study Design

Data Collection

Data Storage

Data Analysis and Manipulation

Data Management



Storage

Metadata

File Formats

Dissemination and Sharing

Much, Much More!!

Data Formats

- Questions to address when writing this section:
  - What data will be generated in the research?
  - What data types will you be creating or capturing? (e.g. experimental measures, qualitative, raw, processed)
  - How will you capture or create the data? (This should cover content selection, instrumentation, technologies and approaches chosen, methods for naming, versioning, meeting user needs, etc, and should be sensitive to the location in which data capture is taking place.)
  - If you will be using existing data, state that fact and include where you got it. What is the relationship between the data you are collecting and the existing data?
  - Which file formats will you use for your data and why?
  - What is the relationship between the data you are collecting and any existing data?
  - How will the data be processed?
  - What quality assurance & quality control measures will you employ?

Data Formats

- Flat files (i.e. .csv, .txt, spreadsheet, etc.) on disk space
  - Pros
    - Easily read by most software
    - Easily shareable
  - Cons
    - Little to no built-in functionality
    - Can be difficult to maintain (larger datasets)
    - Limited scalability
    - Limited data type storage

- Other formats might include HDF5, NetCDF, JSON etc.

Data Storage

- Questions to address when writing this section:
  - What is the long-term strategy for maintaining, curating and archiving the data?
  - Which archive/repository/database have you identified as a place to deposit data?
  - What procedures does your intended long-term data storage facility have in place for preservation and backup?
  - How long will/should data be kept beyond the life of the project?
  - What data will be preserved for the long-term?
  - What transformations will be necessary to prepare data for preservation / data sharing?
  - What related information will be deposited?

ARCC UW

Data Storage

- NoSQL/unstructured databases (MongoDB, CouchDB, Allegro, Hadoop, etc.)
  - Pros
    - Mostly open source
    - Very fast
    - Horizontal scalability
    - Many data types
  - Cons
    - Immature?
    - Little to no indexing
    - Bad reporting
    - Absence of Standardization

Data Storage

- Relational/structured databases (Oracle, MS SQL Server, PostgreSQL, MySQL, etc.)
  - Pros
    - Very mature (over 40 years of development)
    - Reduced data redundancy and errors
    - Cater for future requirements
    - Improved security
  - Cons
    - Perhaps outdated
    - Potentially expensive
      - Hardware and software
    - Time consuming to design

Research Metadata

- Questions to address when writing this section:
  - What contextual details (metadata) are needed to make the data you capture or collect meaningful?
  - What form will the metadata describing/documenting your data take?
  - How will you create or capture these details?
  - Which metadata standards will you use and why have you chosen them? (e.g. accepted domain-local standards, widespread usage)
  - What metadata/documentation will be submitted alongside the data or created on deposit/transformation in order to make the data reusable?

- Metadata standards
  - Examples
    - Readme file
    - Dublin Core
    - DataCite
    - XML
    - JSON-LD
    - Etc.

Research Metadata

- When
  - Temporal data
    - Very important to know what time zone that your data are in
- Where
  - Spatial information
    - MUST include coordinate reference system and datum!
- How
  - Instruments
    - Units!
  - Derived data
    - process

Data Sharing

- Questions to address when writing this section:
  - How, when, and where will you make the data available? (Include resources needed to make the data available: equipment, systems, expertise, etc.)
  - What resources will be needed to reuse the data? Examples include software or equipment.
  - What is the process for gaining access to the data?
  - Who are the foreseeable data users?
  - How should your data be cited?
  - How long will the original data collector/creator/principal investigator retain the right to use the data before opening it up to wider use?

ARCC UW

Data Sharing

- Generally, the best course of action for sharing data is using a discipline specific external data repository.

NSIDC
National Snow and Ice Data Center

neon®
National Ecological Observatory Network

CZO

CUAHSI
UNIVERSITIES ALLIED FOR WATER RESEARCH

AMERIFLUX

Coming Attractions

- EZID
  - UW Libraries have recently purchased a license to EZID for the creation of DOIs and ARKs cite datasets and other resources
- ORCID
  - UW Libraries is in the process of implementing ORCID to reference individuals

EZID
Long-term identifiers made easy

ORCID
Connecting Research
and Researchers

Q&A

- What would you like this class to cover?
- Email comments/feedback to arcc-info@uwyo.edu

University of Wyoming
Advanced Research Computing Center (ARCC)
arcc-info@uwyo.edu
(307) 766-7748