Planned training: *Best Practices in HPC Data Management*

- **Purpose:**
  - Maximize use of existing storage resources
  - Minimize disruption of file system / compute nodes
  - Prevent accidental data loss
  - Increase overall productivity
- **Format:**
  - 1.5 to 2 hrs with student engagement
  - In a room equipped with computers and internet access
    - 24-seat room available in BSRB, larger if students bring laptops
  - Students should have cluster accounts / access
- **Target audience:**
  - Beginner to intermediate familiarity with CLI
  - HPC users requesting large quota increases (?)

Email: wallen@tacc.utexas.edu

Planned training: *Best Practices in HPC Data Management*

- **Module 1: Overview**
  - Why do we need data management?
  - Types of filesystems in HPC
  - Specifics of storage infrastructures
  - Active vs. inactive data
  - Staging data for compute / analysis / storage
- **Module 2: Navigating a file system**
  - File system limitations
  - Determining the age / size of a file
  - Determining the size of a directory
  - Total amount of free disk space
  - Checking your quota
- **Module 3: Moving and backing up data**
  - When is it time to archive data?
  - Zipping and archiving files and directories
  - Transferring data from point to point (e.g. rsync)
  - Transferring data to LTFS for archiving
- **Module 4: Maximize file system usage**
  - Best practices in directory organization
  - Removing duplicate copies of data
  - Sharing files instead of duplicating files (e.g. ACL)
  - Don't install software in your home directory
- **Module 5: Advanced topics**
  - How to care for sensitive data (HIPAA, ITAR, FIZMA)
  - Complicated / large data transfers (e,g, globus)
  - Data repositories
  - Metadata, provenance, standards
  - iRODS data management (Corral)

Email: wallen@tacc.utexas.edu