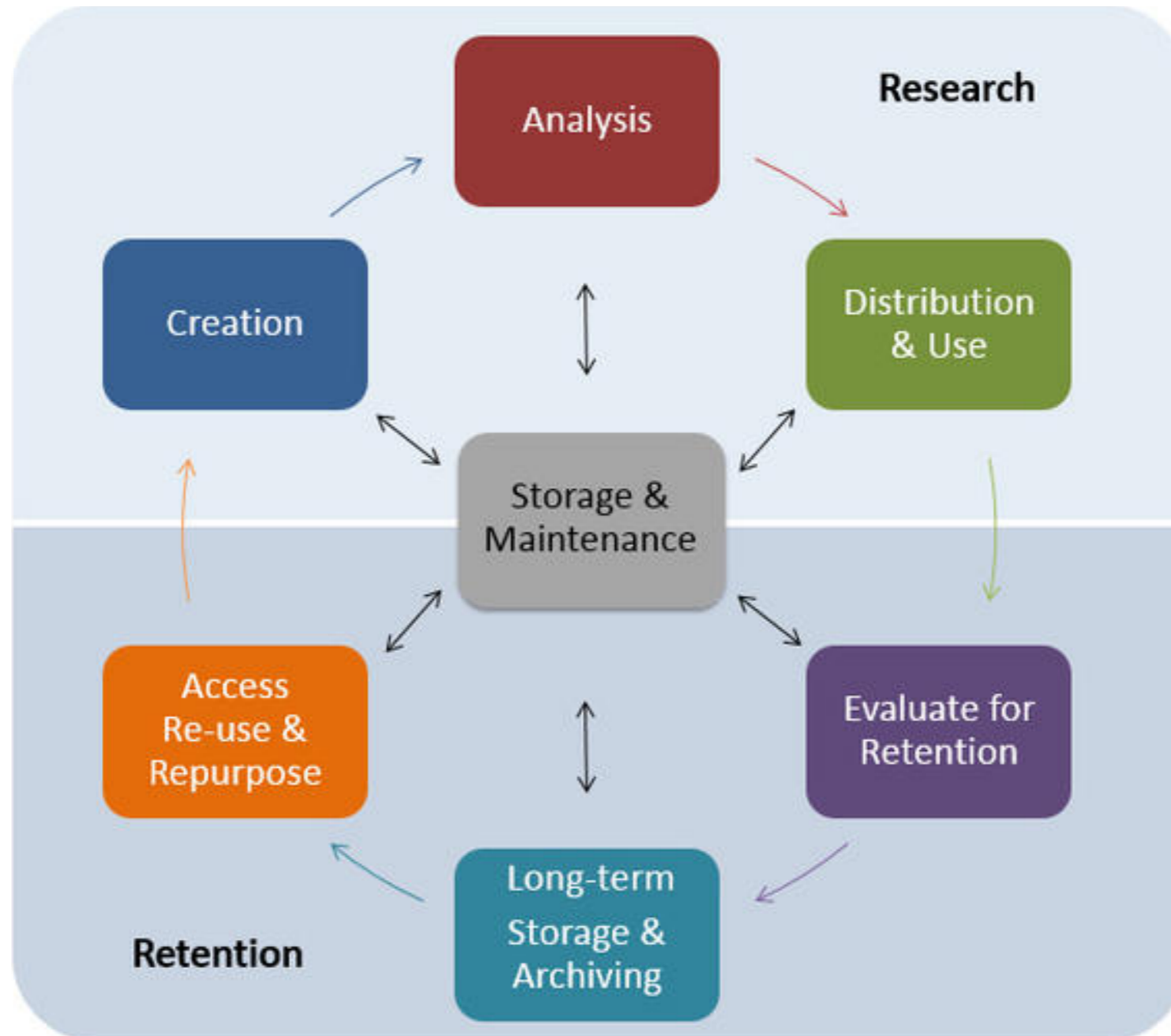
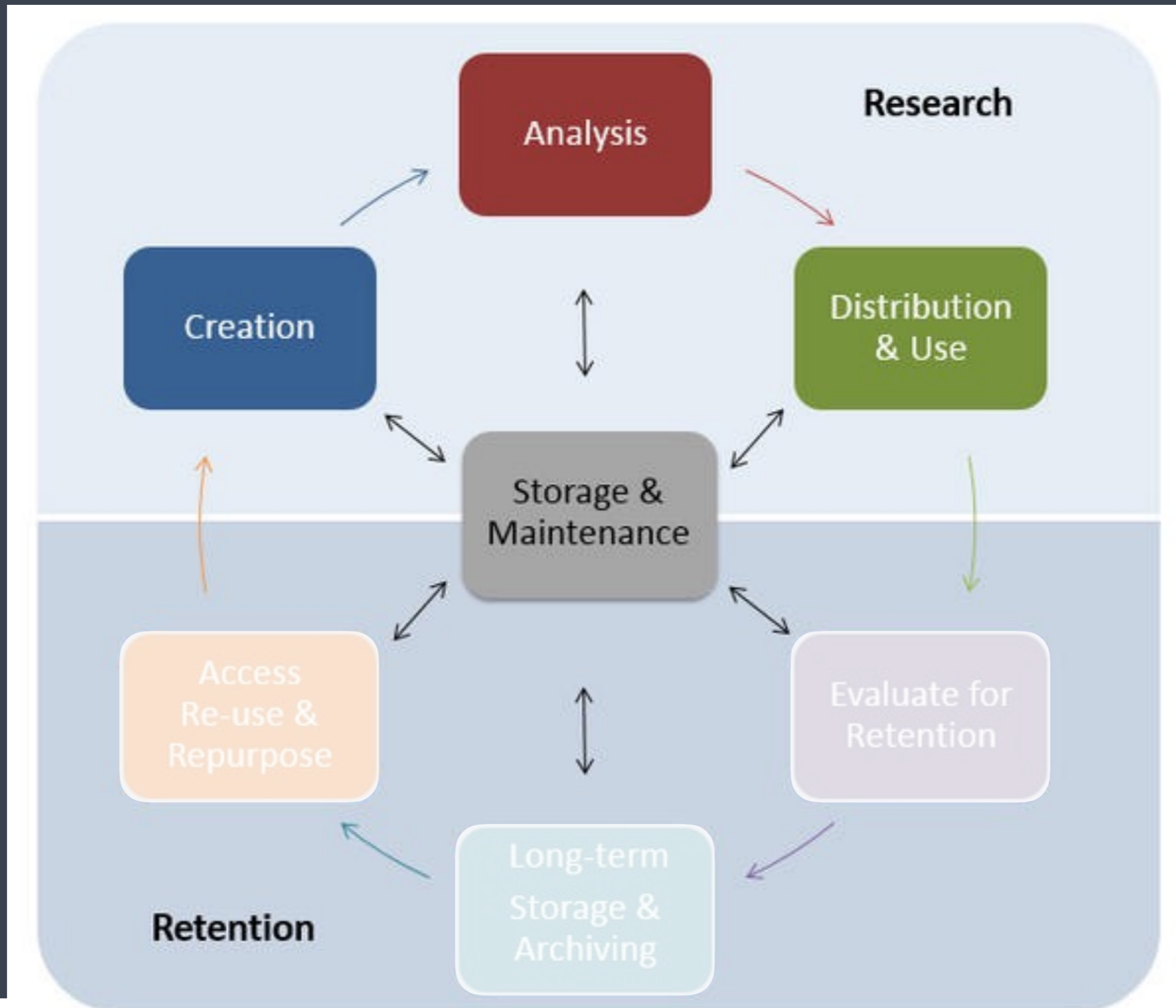


Research Data Management in the context of RNAseq analysis



<http://datamanagement.hms.harvard.edu/>

Data Life Cycle



Data Creation, Analysis, Sharing

- ▶ Inextricably linked
- ▶ All contribute to rigor and reproducibility in research
- ▶ Issues with research integrity often stem from these sections of the data life cycle
- ▶ Best practices ensure that appropriate parties/people get credit

Data Creation Best Practices

- ▶ Data generated from scratch?
- ▶ Data generated by sequencing prepared samples?
- ▶ Collecting RNA-seq data from single or multiple existing databases/repositories?

Data Analysis Best Practices

When designing an analysis workflow:

- ▶ Follow guidelines for data use as mandated in any associated DUAs
- ▶ Use appropriate tools and compute environments
- ▶ Keep track of tool versions and parameters used, document everything!
- ▶ Don't reinvent the wheel
- ▶ Stay organized from the start

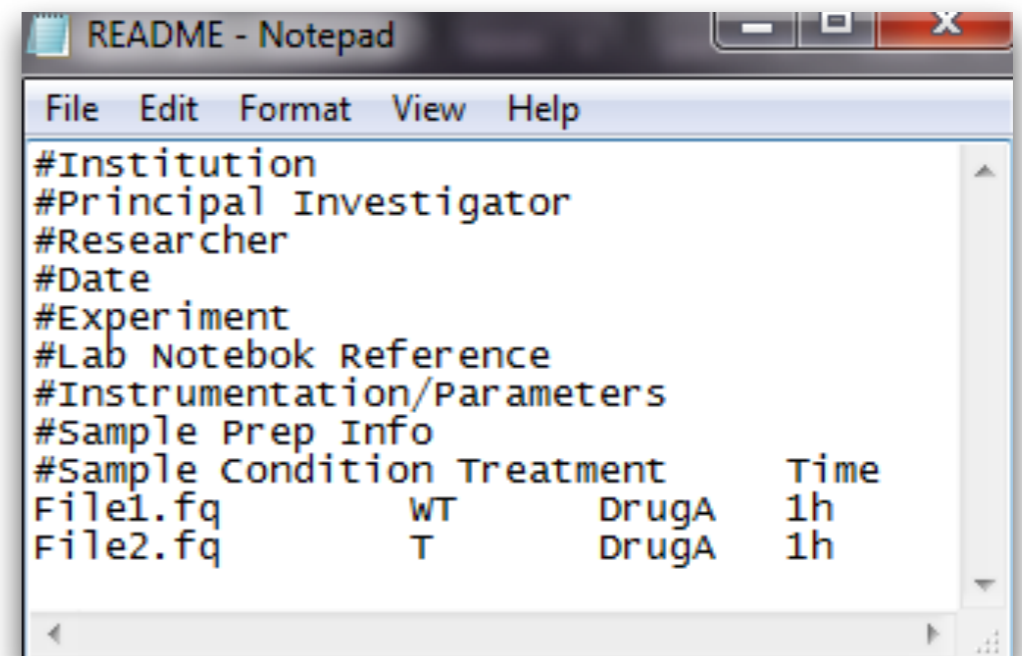
Data types: Metadata

- ▶ Metadata is information about your data (any/all information)
- ▶ Ask yourself:
 - ▶ What experimental & analysis-related information is important to keep track of?
 - ▶ Would a new project member be able to step in and know how the data was created?
 - ▶ Would they be able to reproduce the analysis?
 - ▶ Documenting your metadata is key to reproducible science!!

Metadata: README

- ▶ Create a plain text file (README.txt) to document information about the dataset, things like sample info, naming conventions, abbreviations, codes etc.
- ▶ Precede any comment about the data with “#”s
- ▶ Have a README file for each distinct dataset

<https://www.ersa.edu.au/understanding-metadata/>

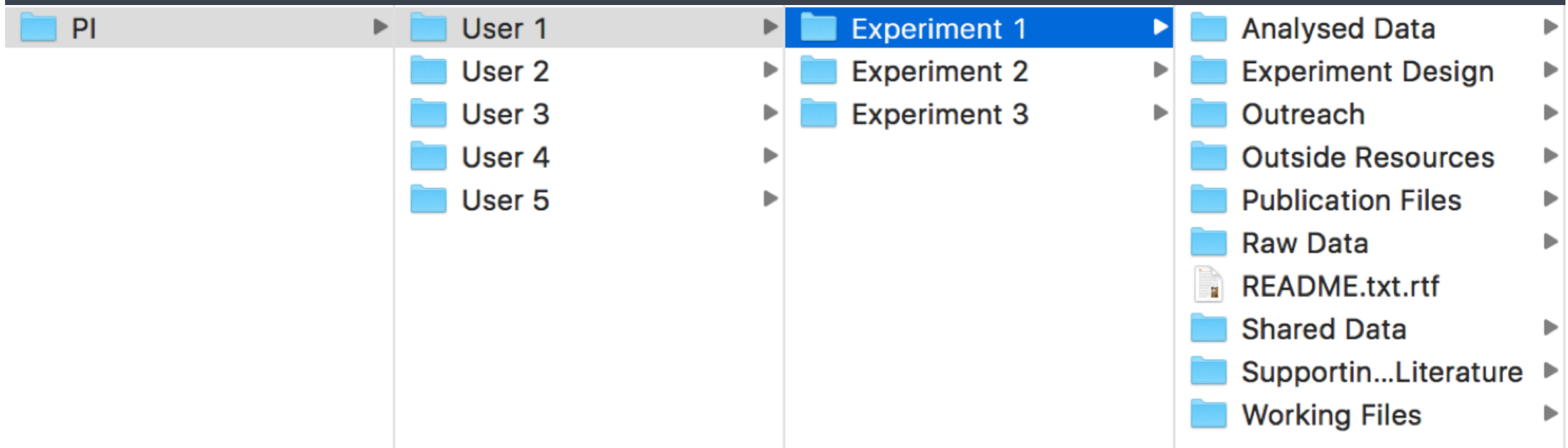


```
README - Notepad
File Edit Format View Help
#Institution
#Principal Investigator
#Researcher
#Date
#Experiment
#Lab Notebook Reference
#Instrumentation/Parameters
#Sample Prep Info
#Sample Condition Treatment Time
File1.fq WT DrugA 1h
File2.fq T DrugA 1h
```

Directory Structure

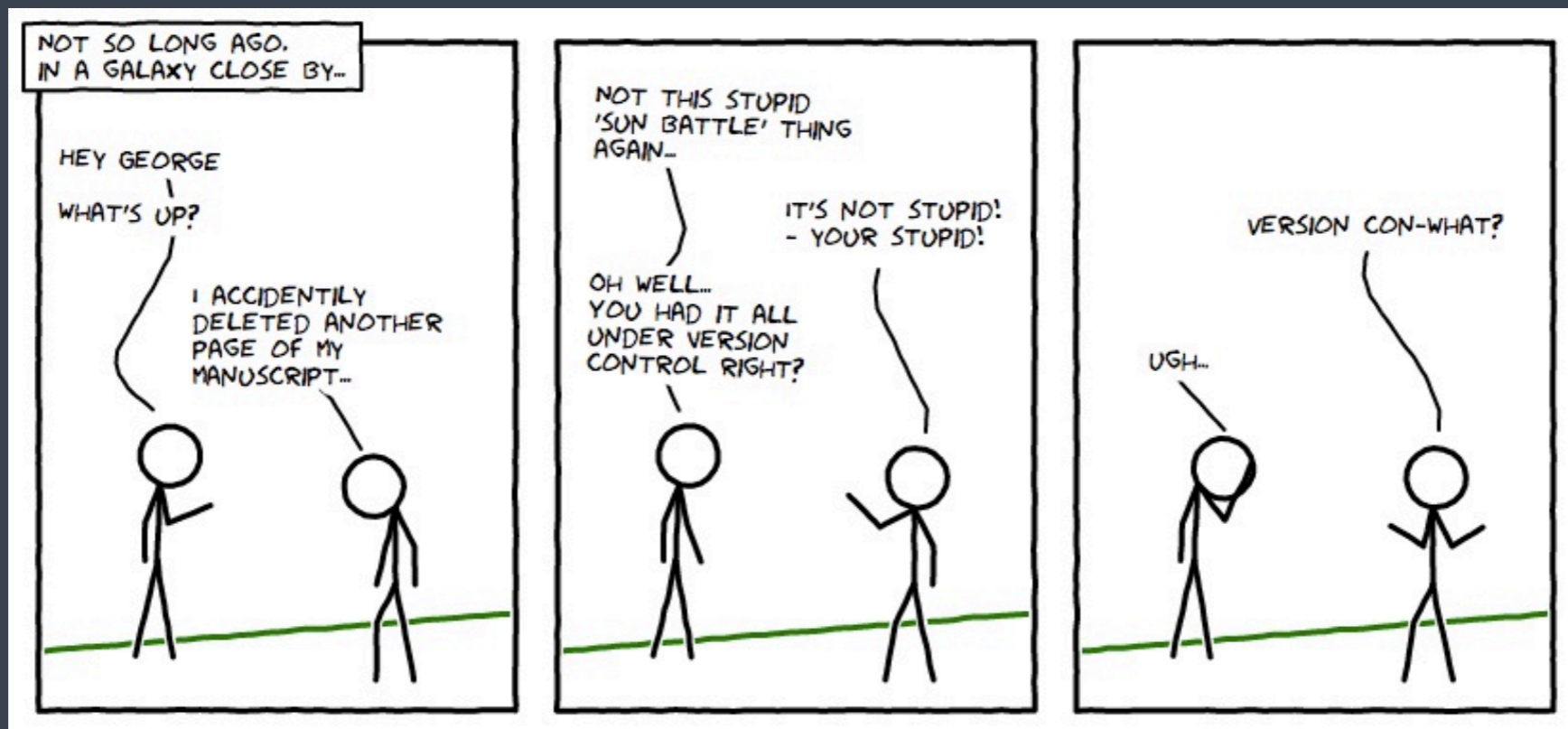
Stay organized from the start, create a directory structure for output files before running the analysis workflow

- Have README.txt files in higher level directories briefly describing their contents
- Have log files for each tool documenting the versions/parameters used



Version control

- ▶ Use a version control system like Git or Subversion to version scripts, READMEs, documentation/metadata files, other text files etc.
- ▶ Essential for reproducible research



High-Performance Computing

*“High Performance Computing **most generally** refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or **business.**”*

<http://insidehpc.com/hpc-basic-training/what-is-hpc/>

High-Performance Computing

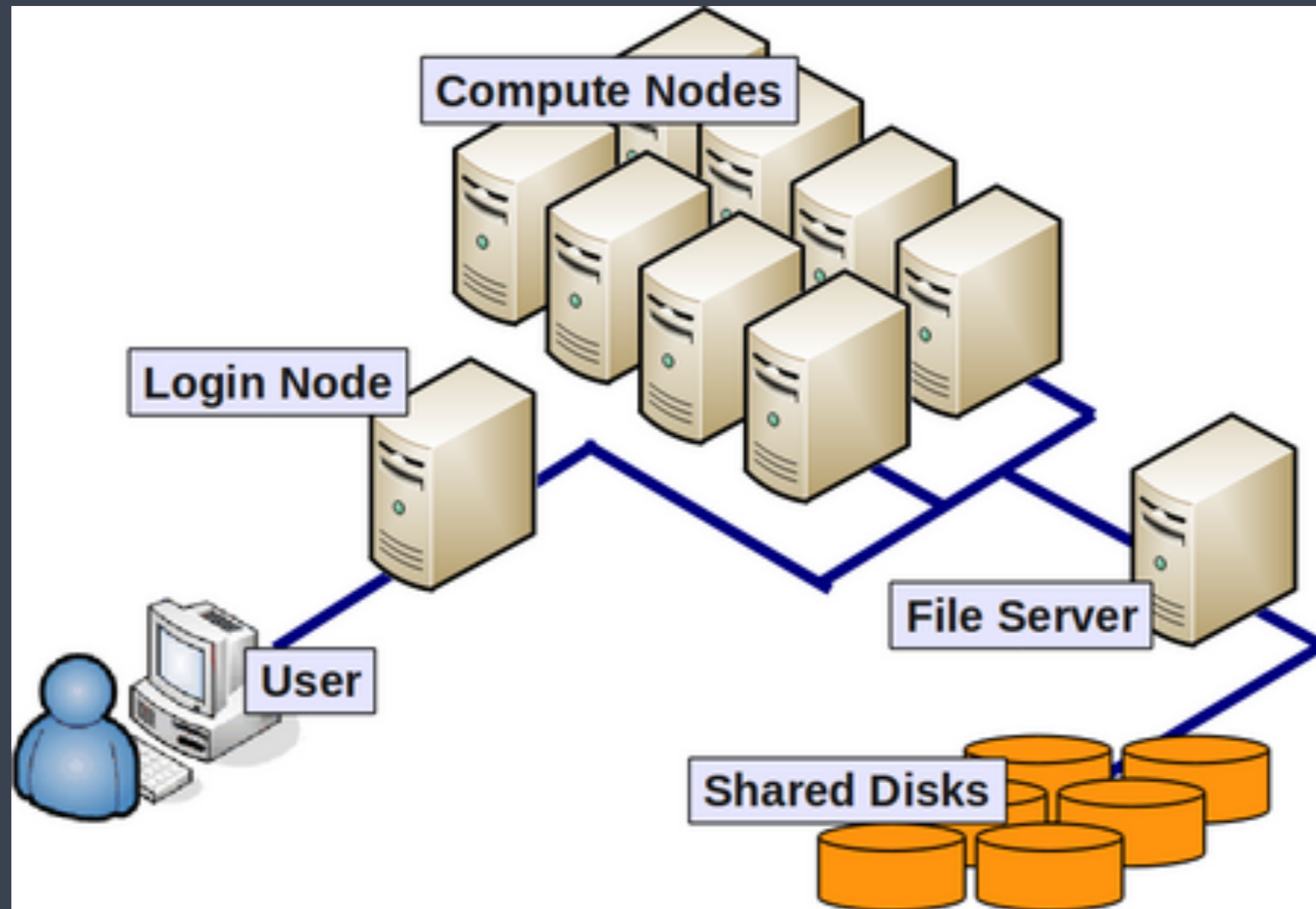
- Provides all the resources to run the desired RNA-seq analysis in one place
- Provides software that is unavailable or unusable on your computer/local system

100s of cores for processing!

100s of Gigabytes or even Petabytes of storage!

100s of Gigabytes of memory!

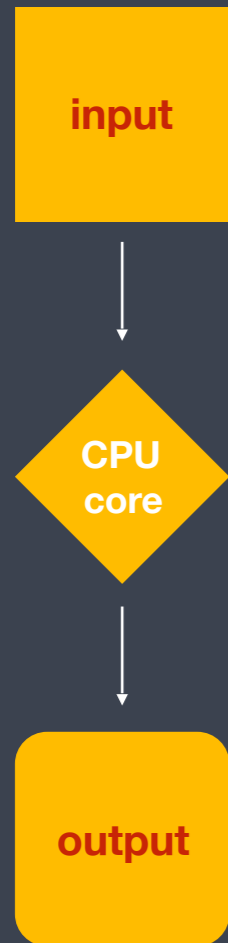
High-Performance Computing



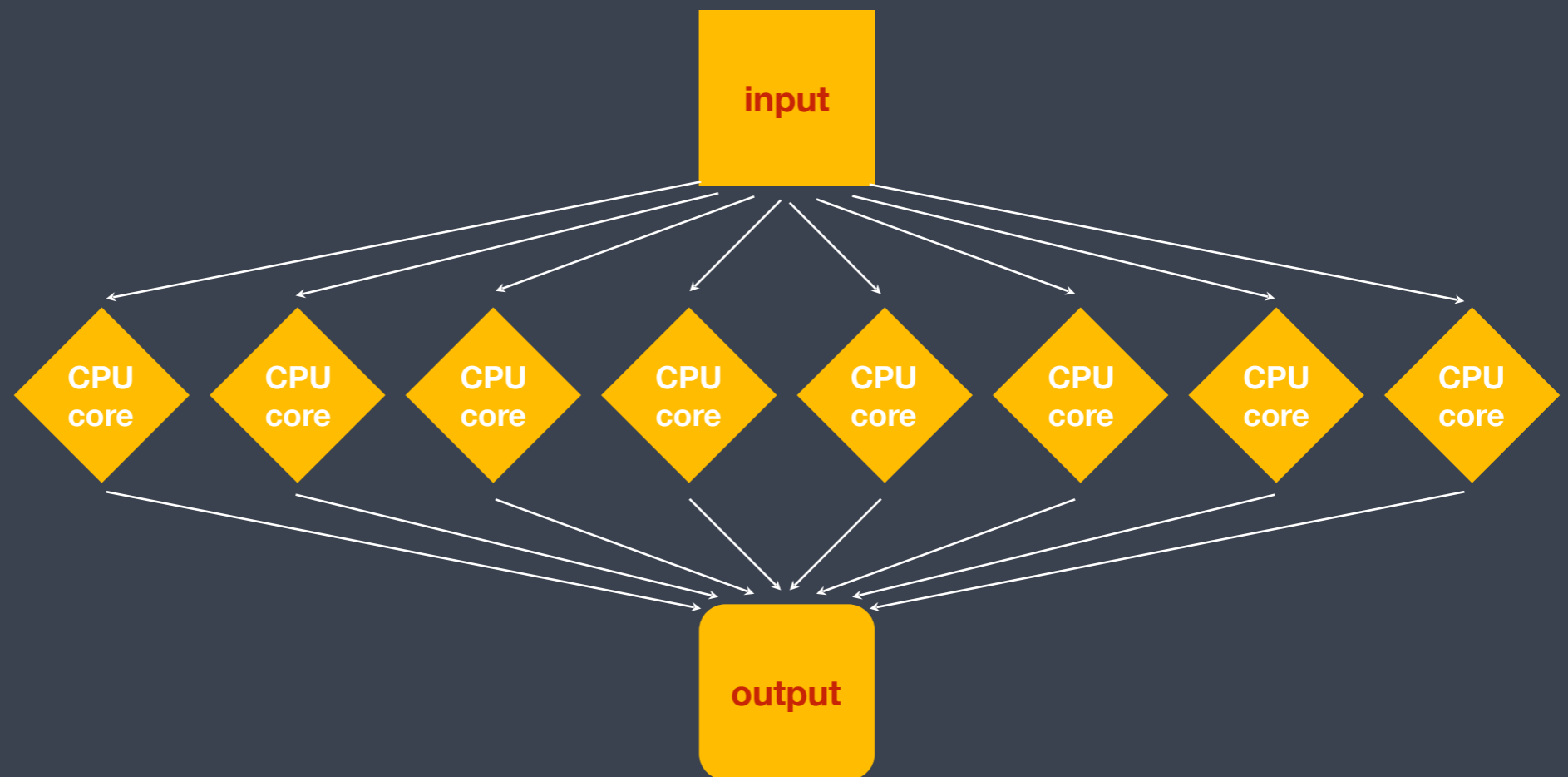
HPC == efficiency

For 1 sample

Serial



Multithreaded



Faster and more efficient...

NGS data analysis is very amenable to this strategy

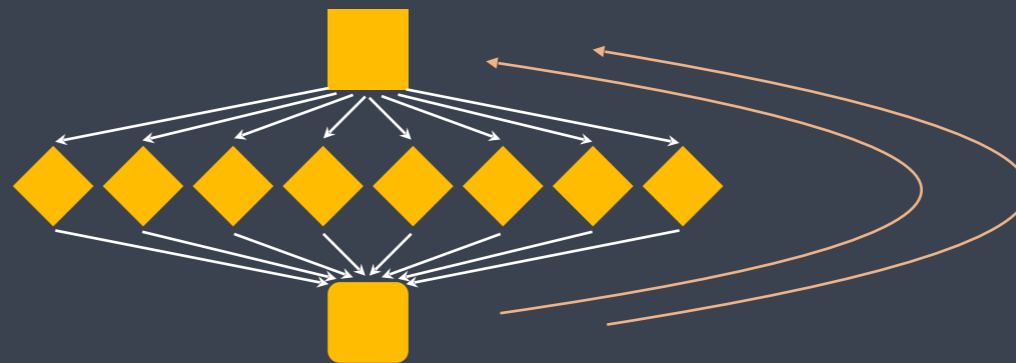
HPC == efficiency

For 3 samples

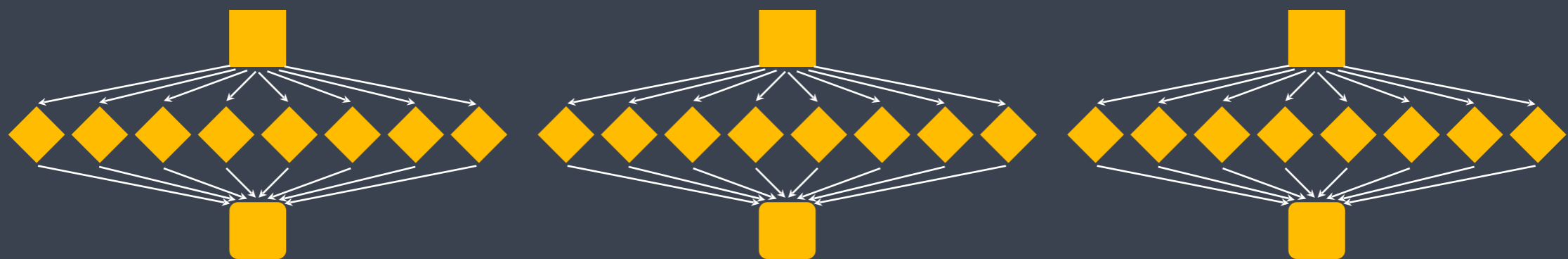
Serial



Multithreaded & Serial



Multithreaded and Parallel



Data Analysis Best Practices

When combining/comparing datasets from multiple sources

- ▶ Analysis should take into account any differences in dataset metadata (e.g. microarray expression data \neq RNA-seq data)
- ▶ Use appropriate analysis tools to counter the differences (*don't reinvent the wheel*)

Data Sharing Best Practices

- ▶ Share appropriate metadata with the raw & processed data
- ▶ Note that funding agencies often require deposition of data into public repositories when a study ends
- ▶ Examples of data sharing policies:
 - <https://www.ncbi.nlm.nih.gov/sra/docs/submit/>
 - https://grants.nih.gov/grants/policy/data_sharing/
 - https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html
 - <https://www.nsf.gov/bfa/dias/policy/dmp.jsp>
 - <https://science.energy.gov/funding-opportunities/digital-data-management>

COMMENT | [OPEN ACCESS](#)

Gene name errors are widespread in the scientific literature

[Mark Ziemann](#), [Yotam Eren](#) and [Assam El-Osta](#) 

Genome Biology 2016 17:177 | DOI: [10.1186/s13059-016-1044-7](https://doi.org/10.1186/s13059-016-1044-7) | © The Author(s). 2016

Published: 23 August 2016

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Be careful with Excel!

Upcoming workshops!

<i>Setting up for success: Everything you need to know when planning for an (bulk) RNA-seq analysis Part II</i>	October 25th	1 PM	HSPH Kresge G1	None
<i>Setting up for Success: Everything you need to know to make your data analysis reproducible</i>	November 15th	1 PM	HSPH Kresge G2	None
<i>Setting up for Success: Introduction to Version Control (Git)</i>	December 13th	1 PM	HSPH Kresge G2	None

Upcoming workshops!

<i>Setting up for success: Everything you need to know when planning for an (bulk) RNA-seq analysis Part II</i>	October 25th	1 PM	HSPH Kresge G1	None
<i>Setting up for Success: Everything you need to know to make your data analysis reproducible</i>	November 15th	1 PM	HSPH Kresge G2	None
<i>Setting up for Success: Introduction to Version Control (Git)</i>	December 13th	1 PM	HSPH Kresge G2	None

2019 SEP 23

Data Management Onboarding: Procedures for Research Consistency

2:00pm to 4:00pm

Location: Modell 100A Fred S. Rosen Lecture Hall, Harvard Longwood Campus

Audience: Faculty, Staff, Lab Managers.

This class is part of the Research Data Management Seminar Series... [Read more](#)

Upcoming workshops!

<i>Setting up for success: Everything you need to know when planning for an (bulk) RNA-seq analysis Part II</i>	October 25th	1 PM	HSPH Kresge G1	None
<i>Setting up for Success: Everything you need to know to make your data analysis reproducible</i>	November 15th	1 PM	HSPH Kresge G2	None
<i>Setting up for Success: Introduction to Version Control (Git)</i>	December 13th	1 PM	HSPH Kresge G2	None

2019
SEP
23

Data Management Onboarding: Procedures for Research Consistency

2:00pm to 4:00pm

Location: Modell 100A Fred S. Rosen Lecture Hall, Harvard Longwood Campus

Audience: Faculty, Staff, Lab Managers.

This class is part of the Research Data Management Seminar Series... [Read more](#)

2019
OCT
07

Data Management Offboarding: Essential Steps for a Smooth Transition

10:00am to 12:00pm

Location: Countway Library, Ware Room, Harvard Longwood Campus

Audience: Faculty, Staff, Lab Managers.

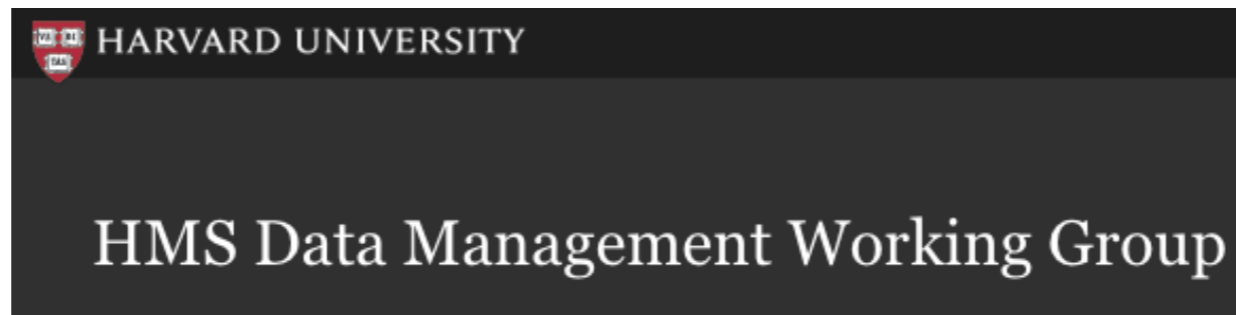
This class is part of the Research Data Management Seminar Series... [Read more](#)

Data Management Resources @Harvard

- ▶ <https://datamanagement.hms.harvard.edu/>
- ▶ <https://researchdatamanagement.harvard.edu/>
- ▶ <https://hlrdm.library.harvard.edu/>

Acknowledgements!

These materials were adapted from existing materials created by members of the Data Management Working Group at HMS, specifically Jessica Pierce from RITS, & Julie Goldman and Meghan Kerr from HMS' Countway library



These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

