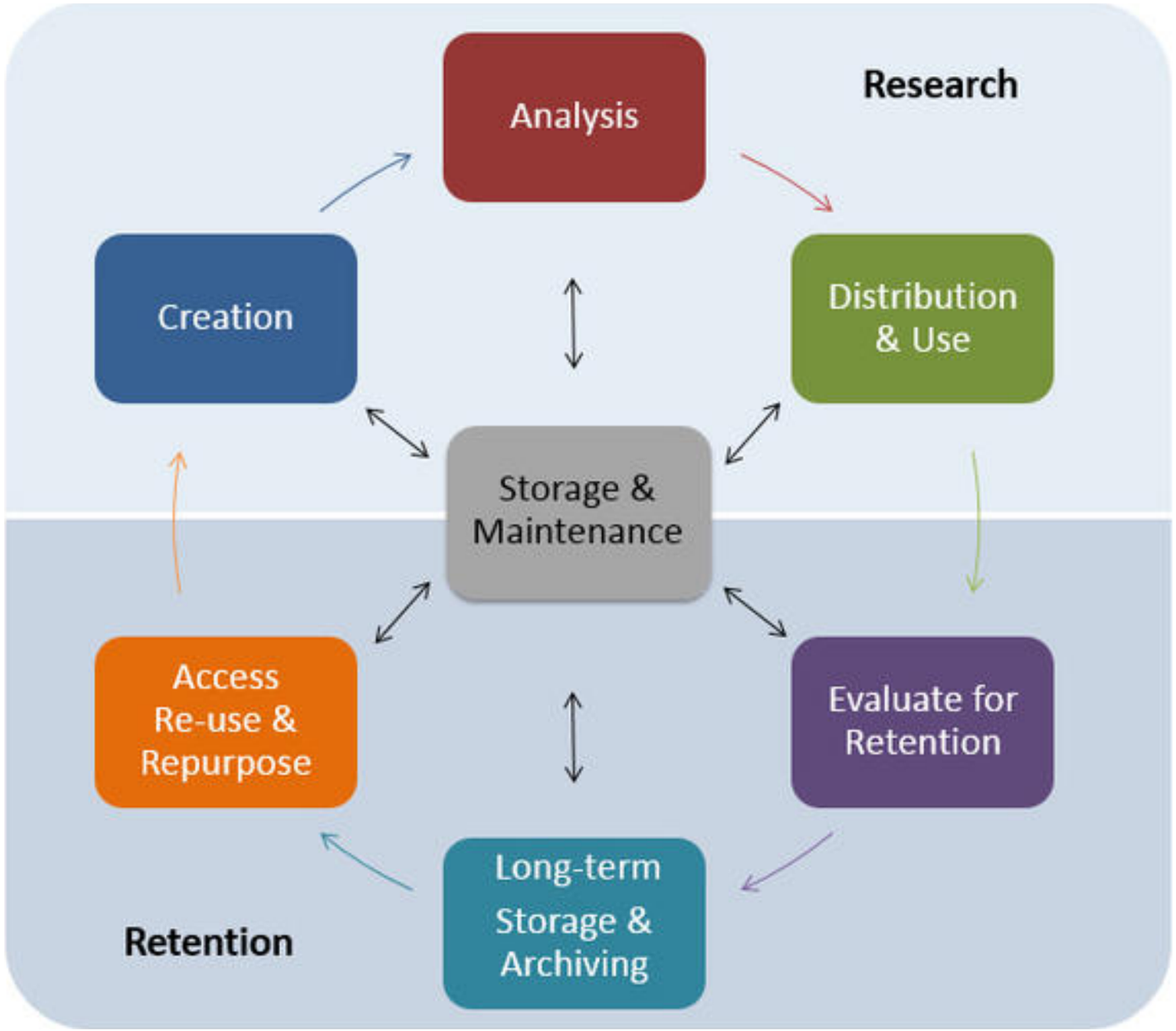
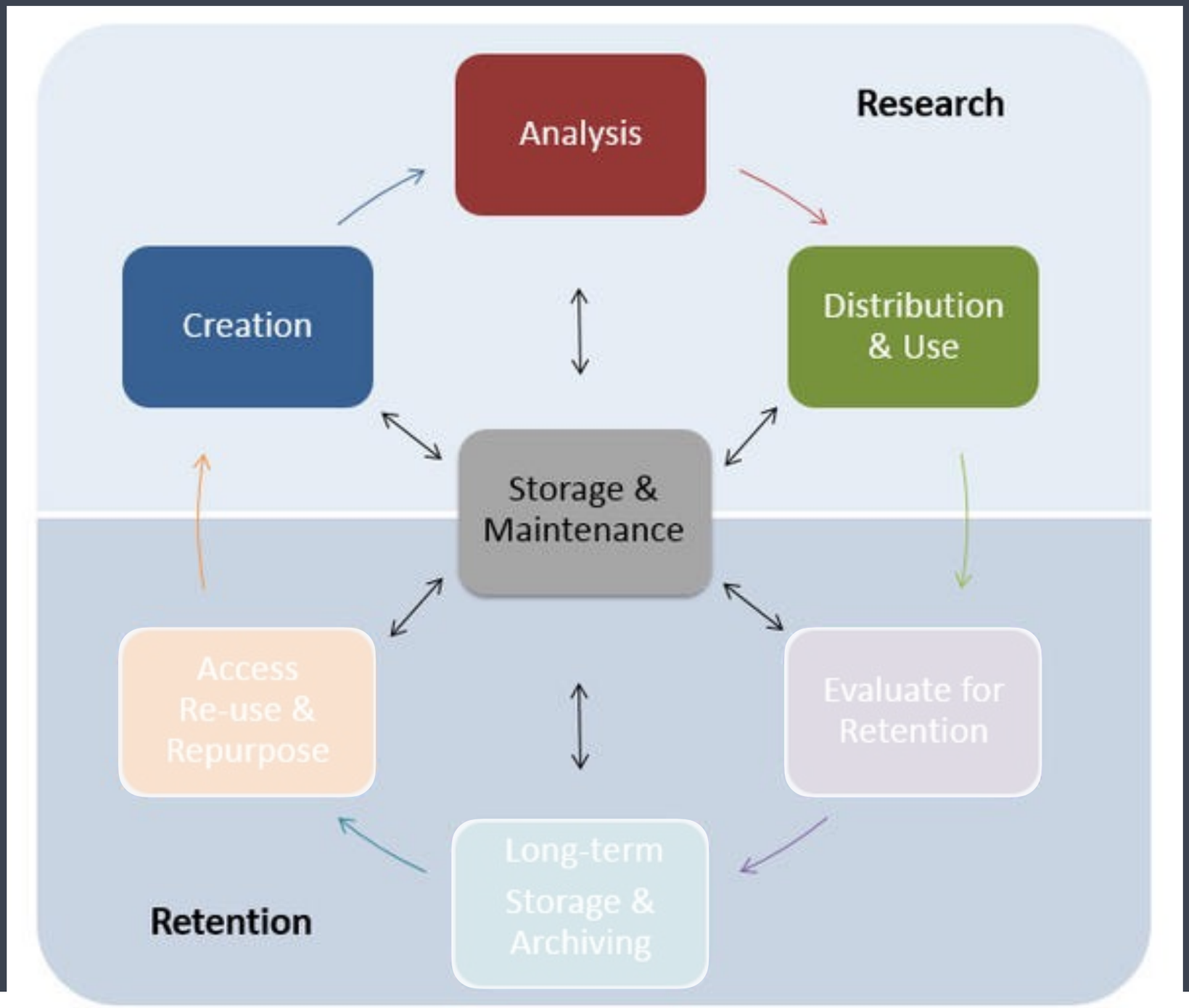
Research Data Management in the context of RNAseq analysis



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

Data Life Cycle



https://datamanagement.hms.harvard.edu/

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:

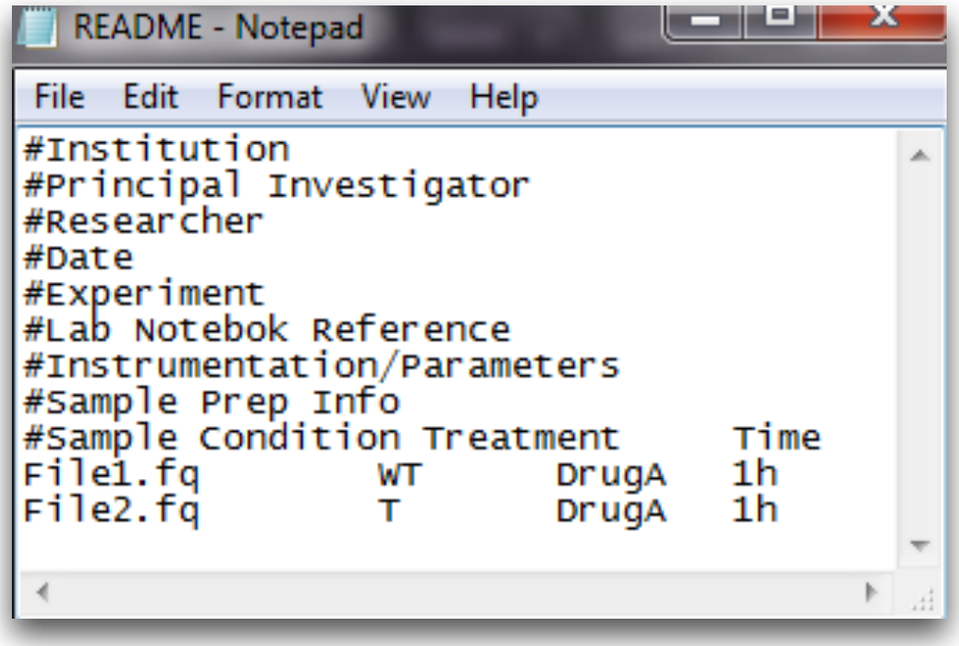
- Follow guidelines for data use as mandated in any associated DUAs
- Use appropriate tools and compute environments
- Keep track of tool versions and parameters used, document everything!
- Don’t reinvent the wheel
- Stay organized from the start

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

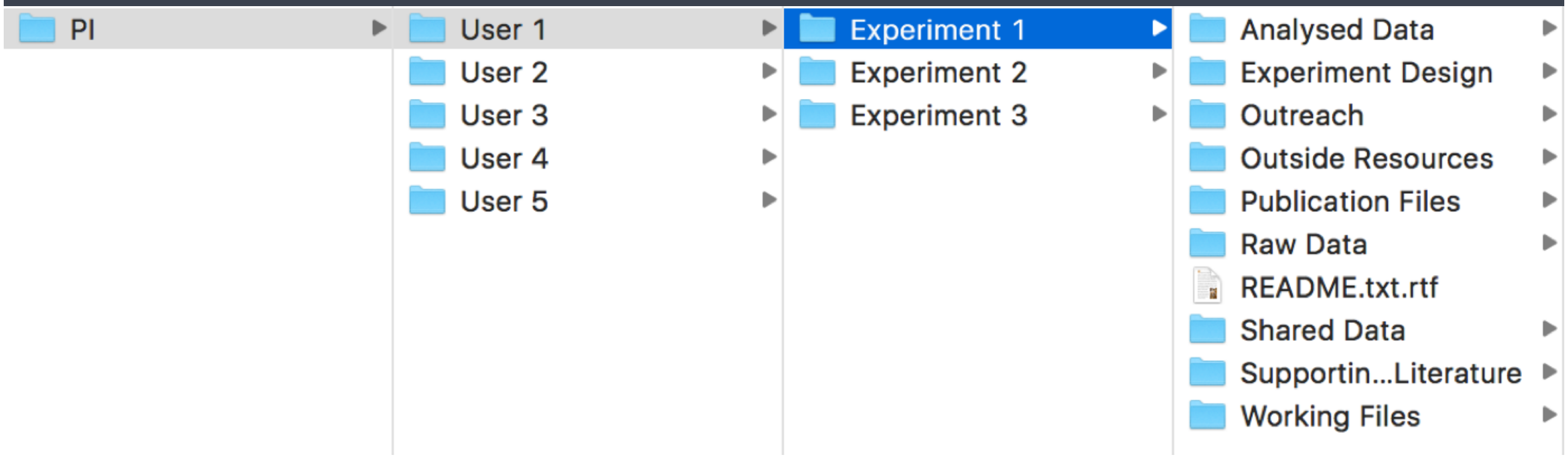


Directory Structure

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

https://xkcd.com/1597/

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

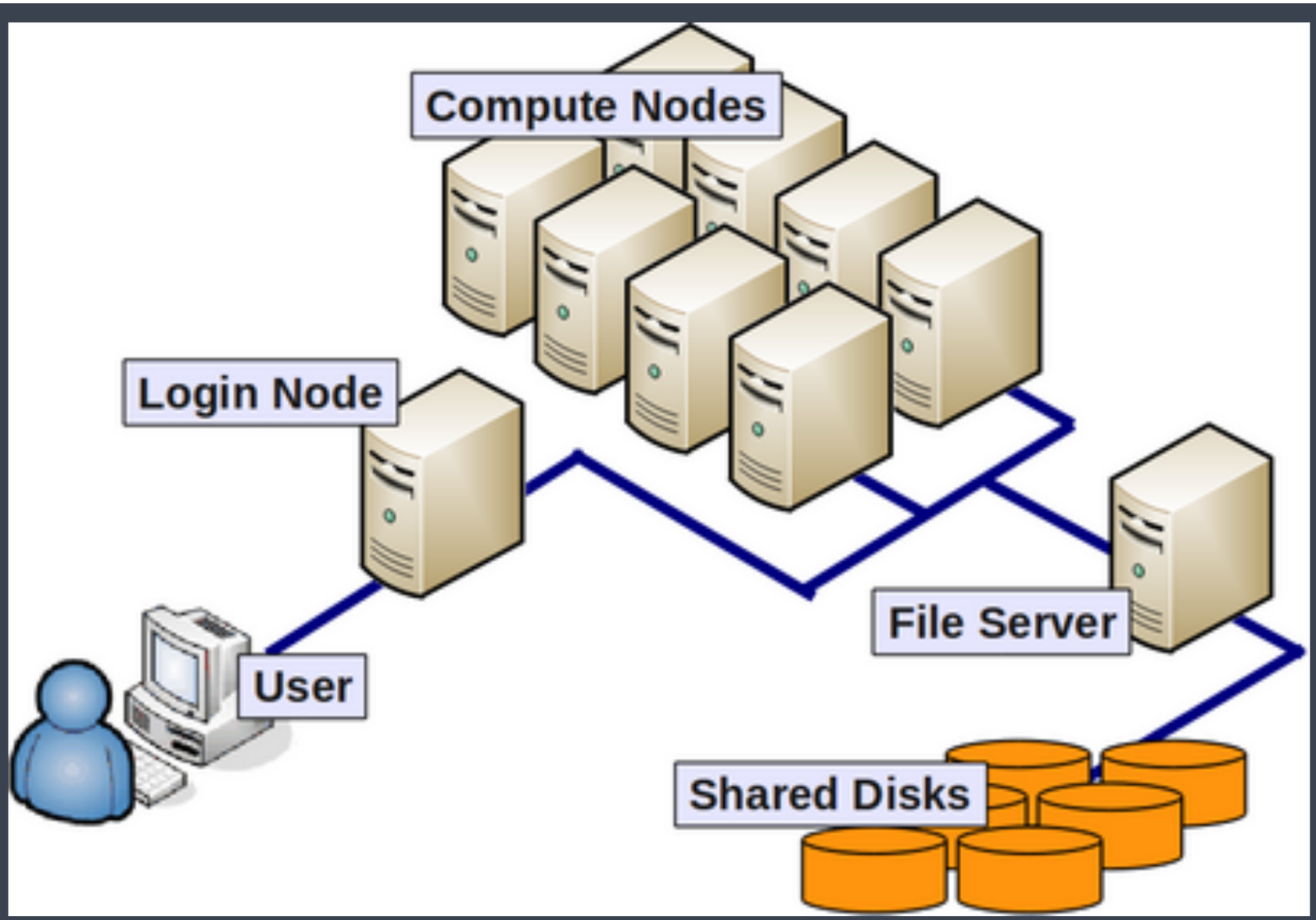
- Provides all the resources to run the desired RNA-seq analysis in one place
- Provides software that is unavailable or unusable on your computer/local system

100s of cores for processing!

100s of Gigabytes or even Petabytes of storage!

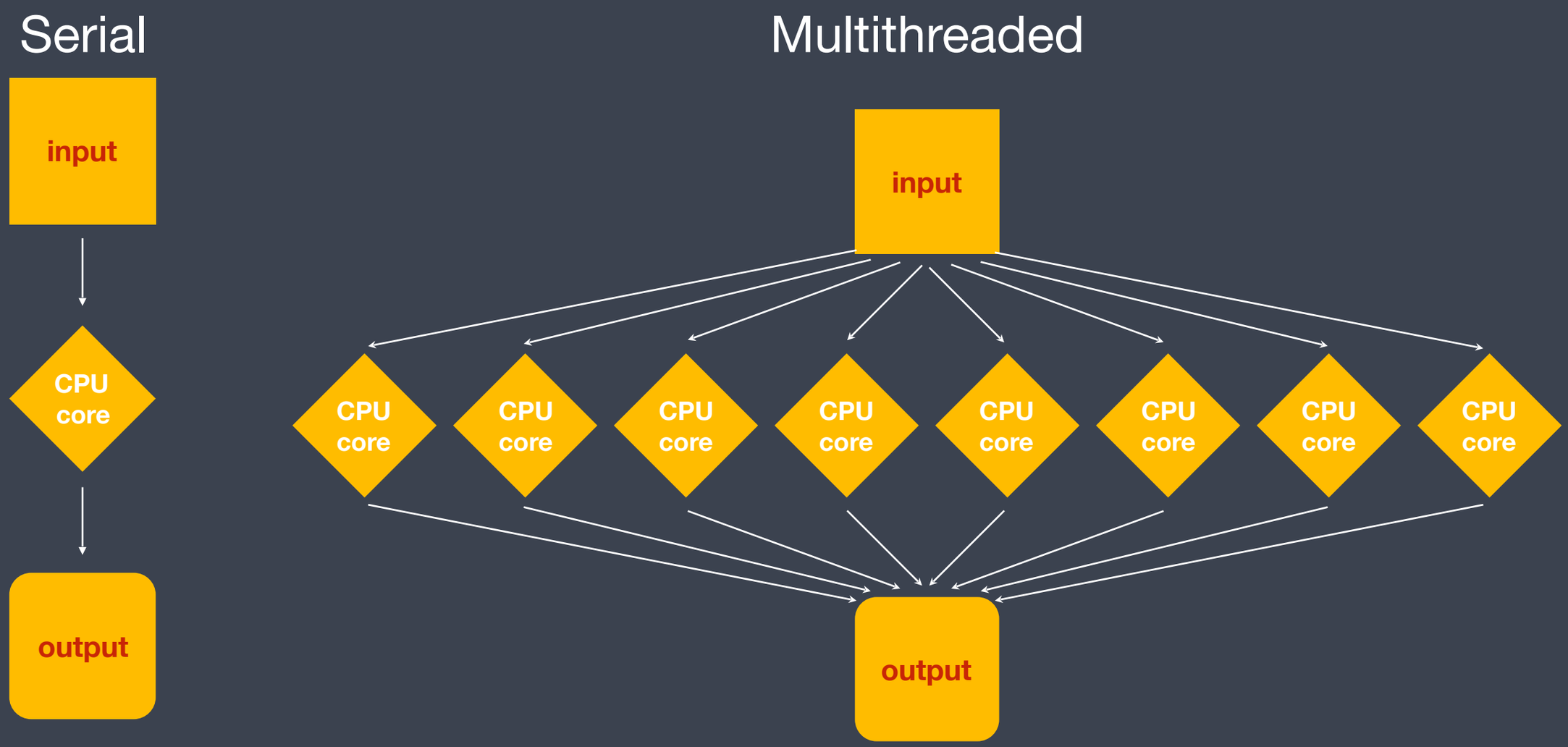
100s of Gigabytes of memory!

High-Performance Computing



HPC == efficiency

For 1 sample



Faster and more efficient…

NGS data analysis is very amenable to this strategy

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 ≠ 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

Genome Biology

HOME ABOUT ARTICLES SUBMISSION GUIDELINES

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 | © 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!

Credits

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 Countway library