Foundational Statistics

Start-to-finish OLS power analysis (R and Talapas)

Rothamsted: An agriculture research station in England

Objectives for today:

1. Find and explore an agricultural dataset from the early 1900s

2. Set up a local working directory and R script

3. Transfer data and script to Talapas for a “resource-heavy” computational task (power analysis) using R

4. Transfer output back to local workspace for plots and analysis in RStudio

Relevant links:

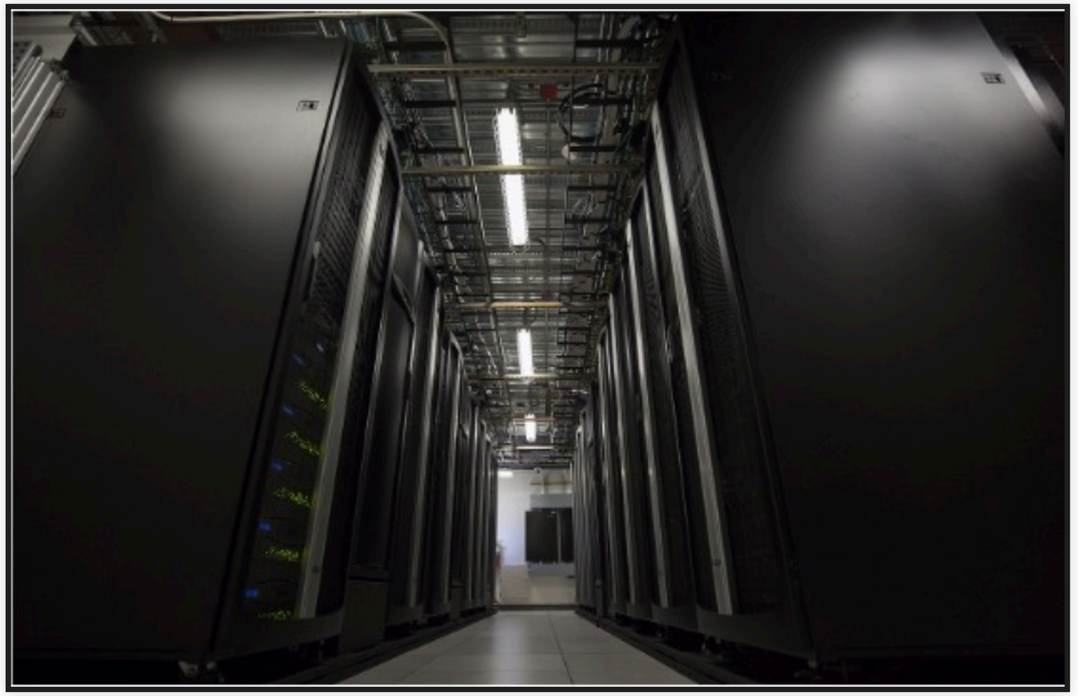
Tutorial from another course on the same dataset:
https://css.cornell.edu/faculty/dgr2/_static/files/R_PDF/mhw.pdf

The original paper:
https://zenodo.org/records/2220822/files/article.pdf

Mercer & Hall wheat yield dataset
(Scroll to the “General statistical methods” section):
https://www.css.cornell.edu/faculty/dgr2/tutorials/index.html

Main help site for UO’s HPC, Talapas 2:
https://uoracs.github.io/talapas2-knowledge-base/

Some key details about Talapas:



- A resource for researchers and students
- It is NOT a dependable long-term storage method
- Talapas takes its name from the Chinook word for coyote, who was an educator and keeper of knowledge

- It can run memory intensive jobs
- It can spread computationally expensive jobs across cpus
- Jobs won't die when you turn your computer off
- Makes it 'easy' to share files with labmates/classmates

- All people with a UO email are eligible for Talapas use!
- However, access is not automatically granted, you must fill out the necessary form for an account
- All accounts must be associated with a Primary Investigator Research Group (PIRG)

“Batch” or “job” scripts for Talapas:

- Starts with the header “#!/bin/bash”, just like scripts on your computer!
- SLURM requires additional headers to set up the parameters of the job
- Below the headers, write your job in normal unix/linux code

```
#!/bin/bash
#SBATCH --account=PIRG ### SLURM account which will be charged for the job
#SBATCH --job-name=MYjob-%j ### Job Name
#SBATCH --output= MYjob-%j.out ### File in which to store job output
#SBATCH --error= MYjob-%j.err ### File in which to store job error messages
#SBATCH --time=0-01:00:00 ### Wall clock time limit in Days-HH:MM:SS
#SBATCH --nodes=1 ### Node count required for the job
#SBATCH --ntasks-per-node=1 ### Number of tasks to be launched per Node
#SBATCH --cpus-per-task=1 ### Number of cpus (cores) per task
```

If you need to install packages to the default instance of R on Talapas:
https://uo-datasci-specialization.github.io/c4-ml-fall-2020/talapas-pkg-install.html