Lecture 07

Testing / Logging / Parallelism

Announcements

- Group Formation due Wednesday
- Proposals: see specifics on course repository.
    - Check in with experts as you develop proposal!
        - Tip: "Start with a problem to solve/investigate", not a method or technique!
- Assignment 3: Due end of the quarter. Pipeline must run on test data, on a DSMLP server.

Testing

- Save a small amount of test data to Version Control
- Should test that pipeline runs end-to-end quickly
- Test data should:
  - be small and innocuous (no personal information!)
  - only be used to test pipelines 'fit together' -- they are not a substitute for analyses!
  - be transformed and loaded from disk, as if only the first step of the pipeline was removed.
- DEMO

Logging

- Logging/debugging with print statements doesn't work well for long running, remote jobs.

- Python's logging module allows you to log events to log files at different levels:
  - [CRITICAL], [ERROR], [WARNING], [INFO], [DEBUG]
  - Can log custom messages, along with the file the logging came from, the time of logging, etc...
  - Logging is does in a *thread safe* manner

- See: https://realpython.com/python-logging/

Parallelism

- Data Science tasks are often easily parallelized:
    - Download many files simultaneously; process many datasets; everything 'before the join' and modeling (BF on the processing DAG)
    - May instead want to process each file "all the way" and parallelize the processing of each observational unit.
    - Even better: more general concurrency (e.g. with Dask)

Parallelism with Bash

- **xargs** builds and executes commands from standard input. It converts input from standard input into arguments to a command.

- Passes input into a command as parameters for running the command!

- Has a flag -P, which runs the command through multiple processes!

- e.g. `seq 20 | xargs -P4 -n2 echo` passes the list of numbers 1..20, into echo, two numbers at a time. Note that the order changes when you (re)run it!

Parallelism with Python

- Intro to Threading (DEMO)