Data Science for PhD student in BioSci

KRT

March 2, 2015

Objective

Students doing PhDs in the Biological Sciences are required to deal with large data sets growing larger at an unprecedented rate. Current PhD programs in BioSci at UCI are not adequately training students in how to handle such tasks. This lack of training results in at least the following issues (from most common to, hopefully, the least common):

1. Disorganization
2. Irreproducible research
3. Publication of erroneous results
4. Publication of fradulent results

This course covers several concepts and tools that cover a suite of "best practices" that, when implemented as a regular part of day-to-day research, solve these issues.

Course structure

Two meetings per week. One lecture/overview per topic (Tuesday) and one practical session led by a team of graduate students (Thursday). During the practical session, students will actively use tools and apply concepts to relevant data sets.

Tentative topics (10 wks)

- Week 1: Set up UCI HPC accounts for students and GitHub accounts. Students will join a GitHub "organization" for the course. Discuss challenges of data science in general.
- Week 2: Using git/GitHub (first week of practicals, too)
- Week 3: Rstudio
- Week 4: Data manipulation in R – data.table, dplyr. Learn how to read data from remote servers.
- Week 5: Plotting in R – base graphics and ggplot2
- Week 6: Reproducible research reports – Rmd, Rnw (the latter is not for the faint of heart). Emphasize that Rmd documents are one form of a "lab notebook" for computational work. For writing papers, Rnw lets you put *numbers in your paper based on calculations done when the manuscript is processed, meaning it is always up to date.*
- Week 7: Data sharing: iPlant collaborative, Dryad, sharing via UCI HPC. The pros and cons of each approach. Emphasize how it is relatively easy to share large data sets.
- Week 8: iPython Notebook–an alternative to Rstudio
- Week 9: Advanced topics: file formats (binary, hdf5), managing large workflows on clusters like HPC. How not to get yelled at by our IT guys.
- Week 10: Where to go from here. Additional training available at UCI. When "big data" is too big for "regular" R: Rcpp, etc.