Foundations 1

Christopher G. Wilson, Ph.D.

Loma Linda University
Dept. of Pediatrics and Center for Perinatal Biology

April 2, 2015

Outline

1. Overview of the class
2. Choosing a text editor
3. Version control with git

Section 1—Foundation skills

1. Using text editors for writing/reading code (emacs/vim/geany/gedit/Sublime)
2. Using the IPython notebook for data exploration
3. Using version control software to track changes (we will use git)
4. Using the class Discussions on Canvas

Section 2—Advanced Foundations

1. IPython/NumPy/SciPy for data analysis. Basic commands, magic commands, using the IPython notebook for reproducible data analysis and sharing
2. Loading and evaluating data with IPython (Pandas, PyEEG, PyHDF)
3. Data visualization tools accessible within Python (Matplotlib, Mayavi
4. Basic signal theory

Section 3—Analysis Methods

1. Basic signal theory overview (Time-series data analysis, Representing data)
2. Dynamical systems analyses of data variability (Poincaré maps, deterministic and non-deterministic complexity)
3. Information theory measures (entropy) of complexity
4. Spectral measures (frequency domain, Fast Fourier Transforms, time-varying spectrum, wavelets)

Section 4—Sequence Data

- Overview/foundations of sequence data
- Exploring sequence data—R/Bioconductor
- Understanding the differences between mRNA-Seq, gene-array, proteomics, and deep-sequencing
- Visualizing data from gene/RNA arrays (heat maps, volcano plots, etc.)

Section 5—Large data set storage and retrieval

- Relational databases
- SQL versus NOSQL
- Cloud storage, ocal NAS/servers, and computing clusters
- Using parallel computational tools to speed your work in IPython (interfacing with Hadoop/Pig/MapReduce)
- Meta-data and dynamic visualization

Data integrity and security

- The *Health Insurance Portability and Accountability Act* (HIPAA)
- Data security best practices
- De-identifying patient data
- Making data available to the public—Implications for data transparency and large-scale data mining

Section 7—Project Presentations

- There are three kinds of projects in this class
  1. Individual project (basics of using IPython, simple statistics computed via interaction with R?or using Pandas?and simple visualization of a dataset).
  2. Short projects (small group, designed to develop team-based distribution of workload).
  3. Large scale project using a Big Data dataset. This project will be the final exam for the class and each team will present their results.

Choosing your programming editor

There are many different programming editors and if you ALREADY have one that you like, feel free to use that. If you have not used editors before we suggest that you try *edit* https://wiki.gnome.org/Apps/Gedit

Version control with *git*

Abby will present a *git* tutorial now. . . .

For the rest of our session today

Please do the following:

- Use your text editor to write a paragraph about yourself, your career goals, and what you hope to get out of this class.
- Download *Canopy* [http://enthought.com]
- Work through the *CodeAcademy.com* tutorial on Python (and you'll probably need to continue this between now and next week).