BIOL 419/519: *Data Science for Biologists*

B. W. Brunton
Syllabus, Winter 2018

Summary

Modern biology is a quantitative science, and discoveries in biology are increasingly driven by quantitative understanding of data. The objective of this course is to provide you with hands-on knowledge in mathematics and basic tools in computation to practice data science, especially to answer biologically motivated questions. In short, we think of "Data Science" as the theory, practice, and art of "learning from data."

The course will focus on the basics of *data wrangling*, *data analytics*, *statistics* and *visualization*. The target audience is advanced undergrads and beginning graduate students, including students studying biology, neurobiology, microbiology, bioengineering, and others fields working with biologically relevant data.

Learning Objectives. By the end of the course, students will demonstrate the ability to take a new (potentially large and complex) dataset from an experiment, load it into relevant scientific computing software, then explore, visualize, and analyze the data.

Pre-requisites. Entering students are expected to have some basic familiarity with algebra, trigonometry, and calculus. Their last course in mathematics may have been some years ago. No prior experience with coding is required.

Course Logistics

Instructor. Prof. Bing Brunton; `bbrunton@uw.edu`

Teaching Assistant. John Huddleston; `jlhudd@uw.edu`

Class Meetings

- MW 2:30pm–3:50pm, MGH 231
- F computational labs, see time scheduler for exact time/room

Website and Communication

- The course website will be on Canvas.
- Make sure you check Canvas regularly for handouts, homework assignments, and communications. All course Announcements will also be made through Canvas.

Office Hours

- Both the instructor and TA will hold regular office hours, times to be announced.
- You may email to request an appointment if you cannot make office hours.

Weekly topics

I. Data science basics

Week 1 Basics of coding in Python using Jupyter Notebooks.

Week 2 Continued intro; managing data and visualizing data.

Week 3 Data visualization and matrix algebra.

II. Fitting data with models

Week 4 Systems of equations; data fitting and regression; model selection.

Week 5 Review and **Midterm on Wed Jan 31**

III. Advanced data techniques

Week 6 Dynamic models; intro to machine learning.

Week 7 Clustering and classification.

Week 8 Principal component analysis (PCA) and related techniques.

Week 9 Time series analyses and predictive analytics.

IV. Projects

Week 10 Data hygiene and provenance.

Finals Week No final exam; **Project presentations during scheduled exam time; final reports due**

Extras

I encourage you to keep your eyes and ears open for cool, recent examples of data-intensive discovery or education in Biology, broadly defined. This includes scientific publications, articles in mainstream media, and particularly well researched blogs. Email these to me and we will share with the class.

Extra credit may be granted at my discretion, particularly if the student supplies an independent analysis or a thoughtful critique.

Textbook and Software

Textbook: None

Software

- We will use Python with Jupyter Notebooks this term.
- Please follow the instructions for downloading/installing Anaconda and Jupyter Notebooks before this Friday's lab.
- If you own a laptop, feel bring it to lecture and to the Friday labs – we will have short in-class exercises on a regular basis.
- If you do not have access to a personal computer, contact the instructor.

Assessement

Undergraduate course: BIOL 419

Graduate course: BIOL 519

note: You must be a graduate student to enroll in BIOL 519.

10% In-class exercises, lab assignments, and participation

40% Homework

20% Midterm Exam

30% Course Project

Late Work

I expect all homework to be submitted on time. All students have a one-time, 2-day extension on a single homework assignment. This extension *cannot* be applied to any deadline associated with the course project. You must email John before the due date to indicate you are taking this extension. Otherwise, all late work will only be excused with a physician's note or other appropriate documentation. Any unexcused late homework will receive a 10% deduction per day.

Course Project

An important part of this course will focus on designing your own question and then answering it with data. The basic idea is that you will choose a data set, pose a hypothesis and propose an analytic approach, carry out the analysis, and then summarize your findings in a written report. The data set may be primary data from your own research, data from your research laboratory (with permission from the PI), or freely distributed data from the internet. You should interpret the field of Biology in a broad sense, to include topics such as genomics, paleontology, ecology, evolution, neuroscience, physiology, medicine, public health, among others. Talk to me if you need suggestions or want to talk about potential ideas!

Course projects will be carried out in teams of 2 or 3. You will be expected to submit a 1-page proposal of the project, carry out the project, present the project in class, and write a 5-page report. More details and deadlines about the project will be discussed in the second half of the class after the midterm.

Academic Honesty in Biology 419/519

I takes academic honesty very seriously and regard plagiarism as a form of fraud. Plagiarism is defined as representing someone else's work as though it were one's own. This definition includes cases of unintentional plagiarism, and even where there was not a conscious intention to deceive, the failure to make appropriate acknowledgement constitutes plagiarism.

Since this course aims in part to teach you how to code, we must acknowledge that a major part of the practice of coding is reusing code and building on others' code. It is often a waste of time to write everything from scratch; therefore, it is especially crucial that you respect others' work. If you duplicate or reimplement an algorithm or code from elsewhere, you must credit the original source (as an inline comment). At all times, what you turn in as solutions to your assignments must represent your own understanding and not copying-and-pasting others' solutions.

I encourage discussion and collaboration! In fact, I believe this is one of the best ways to rapidly improve your skills. If you discuss a solution of an assignment with a classmate, I ask that you clearly indicate this collaboration in the assignment.

What's not allowed is relying on people from others outside this class to solve your problems. Perhaps your roommate is a computer science major or your best friend has been doing research using Python for the past two years. Asking them to solve your problems will not help you learn; further, it is unethical to pass off their solutions as your own.

We will continue to discuss best collaboration practices and how to avoid plagiarism in our computational labs.

Please sign below to indicate that you have read, understood, and will comply with the above statement of academic honesty:

Student Signature	Date
Print Name	UW email