SDS 348: Computational Biology and Bioinformatics
Spring 2021

Lecture: MW 12-1:30pm CST
Lab: *Unique No 57785* T 4-5pm CST
Unique No 57790 T 5-6pm CST

Course Canvas Site	Sp21 - COMP BIOLOGY & BIOINFORMATICS https://utexas.instructure.com/
Instructor	Dr. Layla Guyot layla.guyot@austin.utexas.edu
Graduate Teaching Assistant	Derek Hanson dhanson@utexas.edu

Student hours

Instructor: Tuesday during lab 4-6pm & Thursday 4-5pm CST
TA: by appointment

Student hours are times outside of scheduled class time for students to ask in-depth questions, explore any topic of interest that was not fully addressed in class or discuss topics that need to be revisited. Before quizzes, these hours can be used as study sessions.

Meeting during student hours is also a way for me to get to know you, understand your study habits and learn about your career aspirations. For example, I would need such information in case you need a letter of recommendation for graduate school or employment applications.

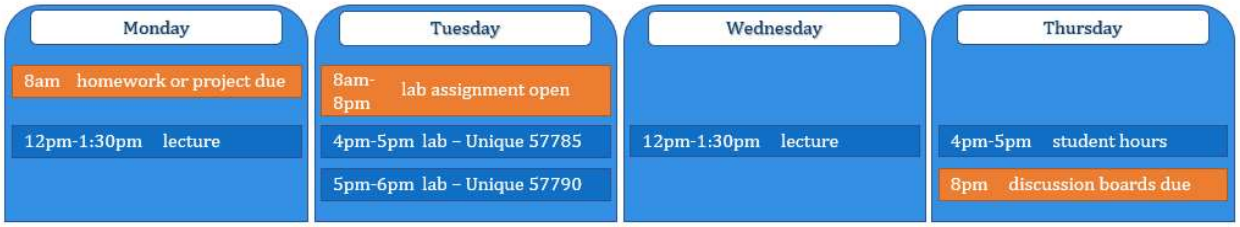
Course objectives and organization

This course is intended to be applied and hands-on. Through programming, discussions, and presentations, you will learn how to apply concepts, interpret analyses, and communicate results. Besides learning to program with R, we will use learn tools such as online repositories to collaborate on projects (e.g., GitHub or Jupyter Notebooks). You will develop crucial skills such as critical thinking, problem solving, and working collaboratively as well as independently.

This course is sequenced after SDS328M Biostatistics and before Bioinformatics. We will build upon your earlier statistical training (probability, tests, linear regression, ANOVA), getting you ready for more specific topics in statistics and bioinformatics (sequence

analysis, dynamic programming, genomic/proteomic techniques, networks/graph analysis). *First*, we will focus on tools for manipulating, visualizing, and interpreting patterns in data. We will primarily be working with the statistical computing language R through RStudio using advanced, industry-standard packages and workflows. *Second*, we will use these newly acquired tools to analyze data and will prioritize applications with common biological methods (e.g., binary/count data) and experimental designs (e.g., repeated measures or longitudinal studies). The goal is for you to gain familiarity with these techniques by learning how to code and interpret outputs. You will be able to judge when such techniques are appropriate and interpret research that uses these techniques. Though this course will be relatively light on theory, we will be extensively focus on computation, including simulations and Monte Carlo methods. *Finally*, in the third part of the course, we shift toward programming concepts and analyzing text and string-type data using regular expressions, leading us to sequence-alignment and other related topics. In the last few weeks, we also shift to programming in Python and see how it can be used to achieve many of the same analytic goals.

We will meet twice a week for lectures, on Monday and Wednesday, and as a lab section every Tuesday. To initiate interest in the key concepts before lectures, you will have reading assignments, then share your thoughts and ask questions through the discussion boards on Canvas. Activity in the discussion boards will count towards your participation grade. Worksheets will be used during class to practice concepts with examples and programming in breakout rooms. During labs, you will complete a short assignment to be completed individually which will be due by the end of the day within a certain time limit. How a typical week looks like:



During synchronous meetings, follow the *Live Meeting Etiquette:*

- ✓ Arrive on time – I'll start most lectures with important announcements.
- ✓ Eliminate or reduce sources of distraction – close all other windows and silence your phone.
- ✓ Have your video on and use your full name – whenever possible, make sure your video is on during live lectures. Feel free to use a virtual background if your computer can support one.
- ✓ Use a photo of yourself as a placeholder for when your video is off – this will help me, the TAs, and other students connect with you during activities.
- ✓ Mute yourself unless prompted – help reduce distractions for the other students in class.
- ✓ Ask questions by using the "raise hand" feature or by typing in the chat window - please refrain from using the chat for comments not related to the class.
- ✓ Participate to your breakout room to help your classmates report back to the main room.

The live lectures will be recorded and posted on Canvas. ***If you have barriers for attending class regularly, contact me ASAP to discuss your options.***

Communication

We will communicate via announcements, discussion boards, and email:

- I will frequently post information to the class using Canvas announcements. Please check your notifications settings to ensure that you receive immediate email versions of announcements.
- We will have discussion boards about current topics on Canvas that will be monitored by the instructor and TA everyday Monday through Friday. Other students can also see, comment, and answer your questions on the discussion boards.
- For more personal question/feedback, feel free to email me at layla.guyot@austin.utexas.edu or through Canvas messaging. I will get back to you **within 2 business days** (don't expect answers during the weekend, we all need to stay sane!).

To ensure fast response, follow the *Email Etiquette:*

- ✓ In the email subject, include the code of the course (SDS 348).
- ✓ In the email subject, include the name of the assignment you have questions about (e.g., HW 1, Lab 2, Project 1, …) or reason for emailing (time conflict, sickness, …).
- ✓ Questions about assignments that are due within 48 hours might not be answered on time so plan ahead!

Course materials

Recommended textbook: There will be assigned reading for this course from textbook that available online for free.

- **R For Data Science**: http://r4ds.had.co.nz
  Written in part by the infamous Hadley Wickham, this book is the foundational text of tidyverse. It is well written, concise, and easy to follow.
- **Modern Statistics for Modern Biology**: http://web.stanford.edu/class/bios221/book/
  This book, by Susan Holmes and Wolfgang Huber, both of Stanford, is terrific in terms of coverage but tends to be a bit advanced and does not give many practice opportunities. Still, a very good book!
- **Broadening Your Statistical Horizons**: http://bookdown.org/roback/bookdown-bysh/
  This book was made using R via bookdown. It is very thorough and very helpful: it goes slow through complicated topics and gives lots of examples.

Software: Download and install the statistical software package R and the user-friendly interface RStudio (both are free):

www.r-project.org www.rstudio.com

If you have used R before, make sure to update your version to the last available at the beginning of the course.

Later in the semester, I will give instructions on how to download and access Python.

Slides/Notes: You will be provided lecture slides and worksheets on Canvas and will collaborate on activities with Google Docs.

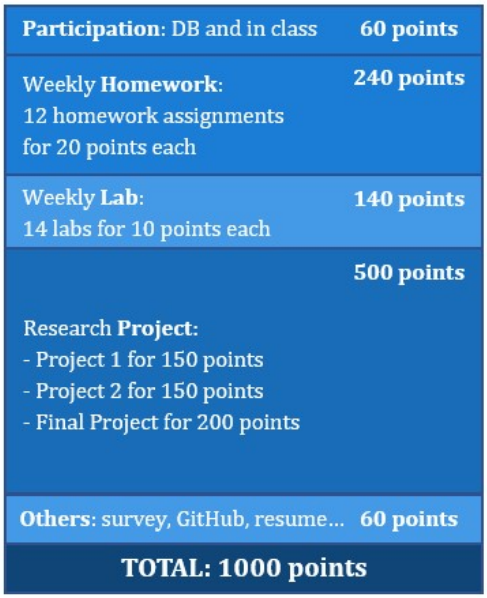
Learning assessment

Participation: Activity in weekly discussion boards on Canvas will count towards the participation grade. Students will post questions and/or comments about the readings or about the key concepts that we will further discuss during lecture (comment, ask or answer questions on at least 10 discussion boards out of 15 over the semester to earn full credit). You need to contribute to class activities in breakout rooms and make sure your group reports back to the main room after activities. ***If you have barriers for attending class meetings regularly, contact me ASAP to discuss your options.***

Weekly Homework Assignments: There will be 12 homework assignments for this course, due on Gradescope each Monday morning at 8am. You will submit a knitted *.Rmd* file as a *.pdf* (we will cover the details in class). Each HW will be worth 20 points, for a possible total of 240 points. Because homework solutions may be discussed during the subsequent lab session, no late submissions will be accepted.

Weekly Lab Assignments: There will be 14 lab sessions to complete via a short assignment (usually a quiz) and submit to Canvas within a time limit (usually 60 minutes). The assignment will be open between 8am and 8pm. The labs are relatively low-stakes and designed to give you additional practice applying concepts from class. Some will be more challenging so it is always a good idea to organize your notes beforehand. Each lab will be worth 10 points, for a possible total of 140 points. You may work within a group of 2-3 students in breakout rooms (unless otherwise noted: sometimes there will be special labs that individual) but each student must submit separately.

Research Project: There will be three projects throughout the course (an exploratory data analysis project, a statistical modeling project, and a final project), due near the end of each unit. For each project, you will produce a report and a presentation that will be peer-reviewed. You may collect data yourself or use data sources available online. Links to possible data sources will be posted on Canvas as the semester progresses. The final project will also include an online portfolio to showcase your work in this course. The first two projects are worth 150 points and the final project is worth 200 points, for a possible total of 500 points. I highly encourage to submit the final project to the Undergraduate Statistics Project Competition under the USCLAP category. This competition encourages the development of skills in statistics and data science and is sponsored by the Consortium for the Advancement of Undergraduate Statistics Education and the American Statistical Association.



Others: other assignments will provide opportunities for personal and professional development. Some example of assignments: present yourself, build your resume, set up a GitHub profile and portfolio, …

Grading

Deadlines for all labs, homework, and project assignments are strict. Please keep track of deadlines with the due dates and road map on Canvas. I understand this semester will be challenging, adapting to online learning and juggling between academic and non-academic priorities. Assignments are expected to be turned in by their stated deadlines, but I will be flexible as long as you communicate with me prior to the deadline or within a reasonable time of the missed deadline and that there is not a consistent pattern of requesting to turn in assignments late.

Make-ups for any graded assignments are guaranteed only under one of the following circumstances:

- ✓ you are away from UT as part of a UT-sponsored activity including athletics,
- ✓ the deadline is in conflict with a religious observance,
- ✓ you provide documentation for an illness or serious emergency that resulted in missing a deadline.

Final grades will be assigned according to the grade cutoffs listed below with the maximum points possible being 1000. Grades are based on point values and cutoffs are firm. Keep track of your progress on Canvas (consider the total points as the % can be misleading).

Grade:	A	A-	B+	B	B-	C+	C	C-	D+	D	D-
Minimum Points:	930	900	870	830	800	770	730	700	670	630	600

Statements

DIVERSITY AND INCLUSION

In accordance with federal and state law, UT Austin prohibits unlawful discrimination, including harassment, on the basis of race, skin color, religion, national origin, gender, gender identity, gender expression, sexual orientation, age, disability, citizenship, and veteran status.

In a perfect world, statistics would be objective. However, some studies are subjective and were historically built on a small subset of privileged voices. I acknowledge that some examples for this course may have biases due to the lens with which it was written. Integrating a diverse set of studies is important for a more comprehensive understanding of statistics. Furthermore, I would like to create a learning environment for my students that supports a diversity of thoughts, perspectives and experiences, and honors your identities. If something was said in class that made you feel uncomfortable, please talk to me about it.

Learn about The Division of Diversity and Community Engagement at UT Austin:
https://diversity.utexas.edu/about-ddce/

UT Austin provides upon request appropriate academic accommodations for qualified students with disabilities. Regardless of whether or not you plan to use your accommodations, bring me your letter within 2 weeks of receiving it. For more information, contact the Office of the Dean of Students at 471-6259.

ACADEMIC HONESTY

This course is built upon the idea that team-based learning is an important and powerful way to learn. I encourage you to study and work on homework assignments together. However, **any work you turn in must be your own.** Everything turned in under your name must be from your brain. Simply copying another student's answers is always unacceptable. **Students who violate the academic honesty expectations for this class will be penalized, up to receiving a failing grade for the class and being reported to the Office of the Dean of Students for academic dishonesty.**

Students should be aware that project assignments will be submitted to the plagiarism-detection tool Turnitin, which is intended to address plagiarism and improper citation. The software works by cross-referencing submitted materials with an archived database of journals, essay, newspaper articles, books, and other published work. Other methods may be used to determine the originality of the paper, and this software is not intended to replace or substitute for my judgment regarding detection of plagiarism.

SHARING OF COURSE MATERIALS IS PROHIBITED

No materials used in this class, including, but not limited to, lecture slides, videos, assessments (quizzes, labs, projects), and in-class materials, may be shared online or with anyone outside of the class unless you have my explicit, written permission. Unauthorized sharing of materials promotes cheating. It is a violation of the University's Student Honor Code and an act of academic dishonesty. I am well aware of the sites used for sharing materials, and any materials found online that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in sanctions, including failure in the course.

LIVE LECTURE RECORDINGS NOTICE

Please be aware that your video and audio might be included in the Zoom recordings of lectures that I post to Canvas for students who are unable to join us live. **Class recordings are reserved only for students in this class for educational purposes and are protected under FERPA.** The recordings should not be shared outside the class in any form. **Violation of this restriction by a student could lead to Student Misconduct proceedings.**

IMPORTANT SAFETY INFORMATION

If you have concerns about the safety or behavior of fellow students, TAs, or instructors, call the Behavior Concerns Advice Line (BCAL) at 512-232-5050. Your call can be anonymous. If something doesn't feel right, it probably isn't. Trust your instincts and share your concerns.

Tentative Course Schedule

See Canvas for full schedule, materials, and assignments.

Unit I: Data Science Skills, Exploratory Analysis

- Introduction and Course Overview
- R Review, R Markdown
- Data visualization
- Data generating processes, Probability functions
- Manipulation/Wrangling
- Exploratory data analysis
- Distances, Associations, Matrix operations
- Clustering
- Dimensionality reduction (PCA)

Unit II: Statistical Methods

- Hypothesis Testing revisited
- Simulation; Randomization tests
- ANOVA, MANOVA
- Regression/GLM revisited
- Binary prediction/classification, Logistic regression
- ROC/AUC, Cross-validation, Regularization
- Count Data: PERMANOVA, Poisson/Multinomial regression

Unit III: Analyzing Strings/Text, Doing Everything in Python

- Strings, Regular expressions
- Working with sequences
- Python programming
- Python data science tools
- Biopython, Python bioinformatics