BIOS 338/538 Syllabus: Analysis and Visualization of Biological Data, Spring 2024

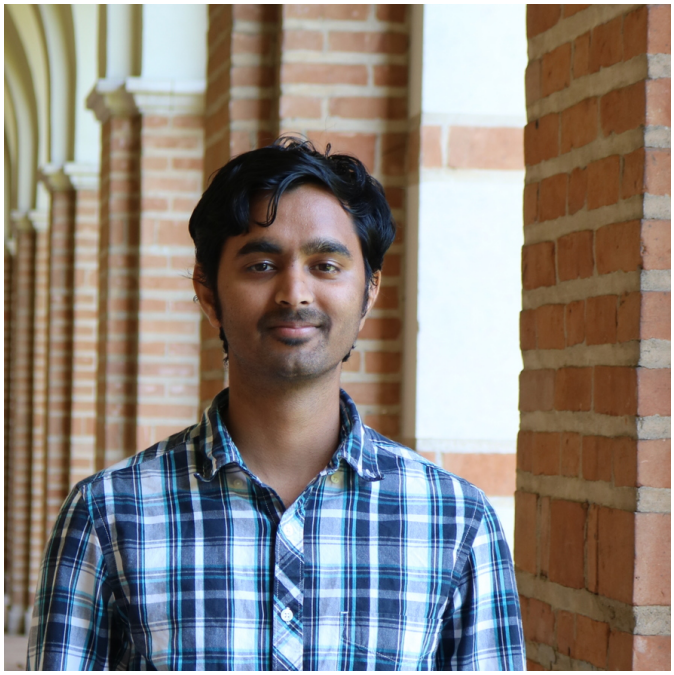
Course Description

This course addresses how to analyze, visualize and draw conclusions from biological data. It introduces basic concepts in statistics and the intuition behind them interwoven with training in data analysis using the `R` computing environment. Students will learn to identify underlying data structures and wrangle data. Students will also learn to effectively convey results using statistical graphics. Topics include basic `R` programming, data wrangling using modern `tidyverse` packages, t-tests, statistical modeling (*linear regression, dose response curves).* This class is targeted towards biologists, hence focuses on biological data for examples and omits rigorous statistical formulas. No prior experience with `R` required.

There will be a series of problem sets posted on the course website with instructions a week before they are due. You are encouraged to work with other students or use any outside resources when completing problem sets, but each student must submit their own written responses unless otherwise explicitly indicated in the problem set. Any group submissions or work should include attribution section detailing who did what. Students enrolled at the graduate level will produce a final project report detailing a specific biological research question, statistical methods and results and will present their final project to the class at the end of the semester.

Course Logistics

Instructor: Dr. Prashant Kalvapalle (he/him), Lecturer + Postdoctoral researcher, Biosciences, Rice University



Location: BRC 282 (*or via zoom with prior permission*)

Time: Tuesdays and Thursdays. 2:30-3:45 pm

Email: pbk1 [at] rice.edu

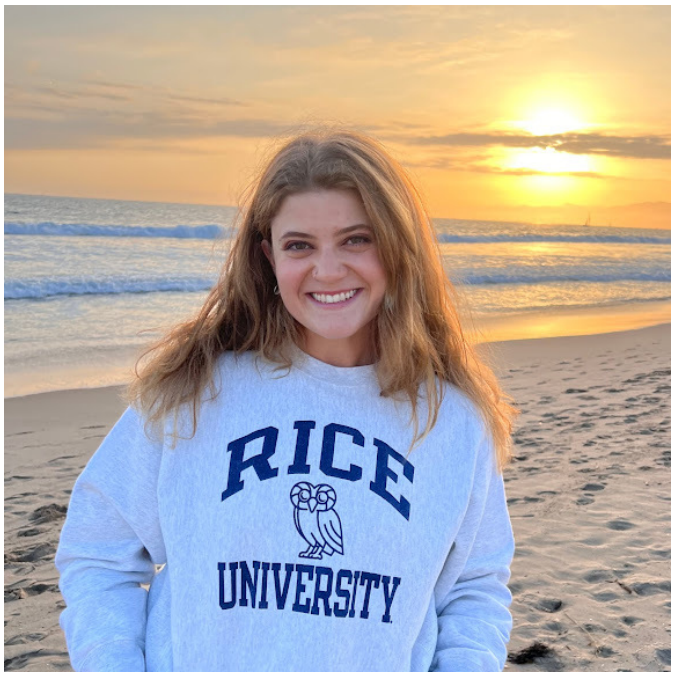
Office: Keck 205

Office hours: Friday, 3-4 pm, Keck 305 (*in person + zoom*)

Note: You will find the zoom links on the canvas website for the course

TAS

Annie Finneran

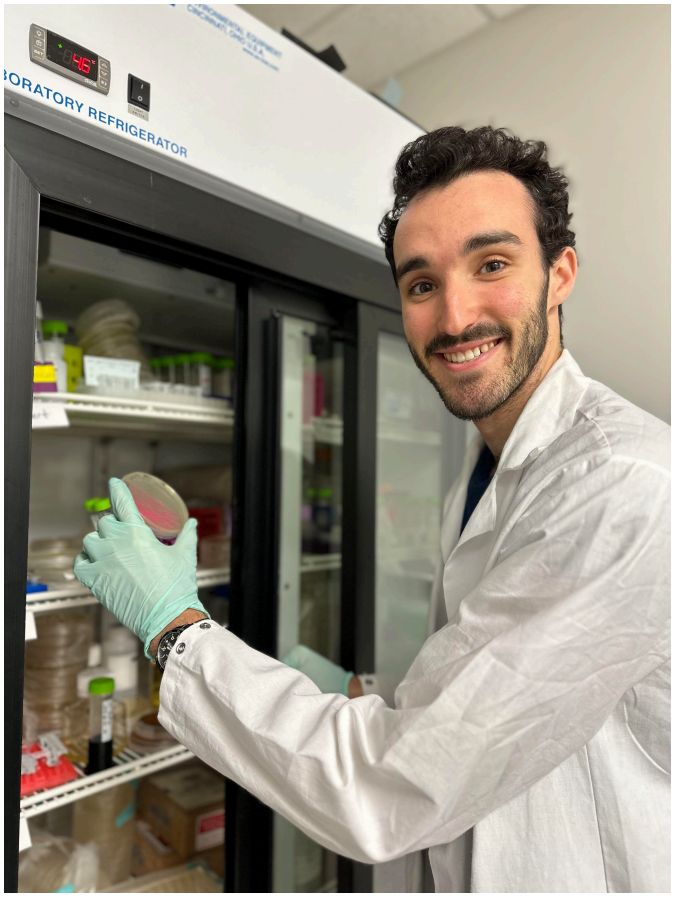


PhD student in Ecology & Evolutionary Biology (website)

Email: af58 [at] rice.edu.

Office hours: Tuesdays 11:30 am – 12:30 pm, ABL 100 (*in person + zoom*)

Sam Schwartz



PhD student in Systems, Synthetic and Physical Biology (website)

Email: Sam.Schwartz [at] rice.edu

Office hours: Mondays, 3 - 4 pm, Keck 305 (*in person + zoom*)

Assessment

Assignment	Total grade	Details
Attendance	10%	*on zoom with permission*
Participation	10%	online discussion board (*ask/answer*), office hours, in-class participation
Problem sets	40%	
Midterm	20%	

Assignment	Total grade	Details
Final / project	20%	For BIOS 538 : Final project presentation : `10%` and report `10%`

Grading Scale

- A 93-100
- A- 90-92
- B+ 87-89
- B 83-86
- B- 80-82
- C+ 77-79
- C 73-76
- C- 70-72
- D 60-69
- F <60

Assignment timelines

- *Thursday*: Assignment uploaded and introduced in class
- Get help in office hours during the week
- ***Friday 7 pm***: submission
- *Next Thursday*: Brief discussion on the previous week's assignment, common mistakes
- *Next Saturday*: receive your grades and feedback from TAs

Again, if you need any accommodations for particular submissions please email the instructor.

How challenging will it be if I am very new to `R`?

I believe that if you have one undergraduate level course where you **learnt** to code in **any programming language**, it should not be too challenging to pickup `R` syntax with 3 weeks of practice.

That said, the examples we walk through in class and working on the assignments (*with help from very enthusiastic TAs*) will be great resources to guide the learning process.

This classroom is a safe space, so no question is too simple or too "silly" to ask.

There are many assignments, seems like a lot of work?

The lectures are structured to be interspersed with **5-10 min** coding sessions where we will walk everyone through stuff. You essentially save all the work we did in class, and add to that by working on your own time for the assignments.

So 25% of the assignment will already be done in class!

Practicing by working through code at your own time is the only way to actually learn coding.

Course Topics / Schedule

Below is a general description of the material to be covered. *Note that topics and dates are subject to change*. See course website for most up to date version.

Date	Topic	Details
9/Jan	Introductions and installations	
11/Jan	Refresher in statistics, *hands-on activity*	working with distributions - *without a computer*
16/Jan	R basics, working in Rstudio	navigating Rstudio, syntax, data types, functions, data frames/tibbles, debugging and looking for help.
18/Jan	Data wrangling in R with `tidyverse`, `dplyr`	Load data from `.csv` or `excel` file, commands to arrange rows and columns
23/Jan	Data wrangling **workshop**	re-hash concepts from the last 2 classes, get students to wrangle data to answer a question
25/Jan	`ggplot()` plotting	`ggplot`, `geom_point()` and `geom_line()`, interactive plots with `plotly::ggplotly()`. What should you show in a plot?
30/Jan	Version control using `git`	why version control? Setup git with Rstudio. Learn some command line basics : `cd`, `ls`, `git add, git commit`, `grep "function-name"`
6/Feb	Rmarkdown/quarto to produce reports	motivation for reproducible data analysis, *benefits of keeping thoughts, code and outputs in one document*

Date	Topic	Details
8/Feb	Data -> figure pipeline from a research article, **workshop**	using `Rmd/quarto` to walk through the steps required to reproduce plots from a published paper
13/Feb	Normal-distributions	_understand central limit theorem intuitively + using R simulations; why ~normal is a useful default. Learn cases when it doesn't apply. Students distribution is got by sampling from a normal and calculating the mean.
15/Feb	Intro to hypothesis testing (*t-tests*)	experimental science usually involves comparisons => hence hypothesis testing to compare distributions / means. *Concept: Students t-distribution is got by sampling from a normal and calculating the mean.*
20/Feb	Problems with p-values	motivate misunderstandings people have, how p-values contribute to non-reproducible science. *Touch upon Surprise value (S-value) ; alternatives : bayesian t.test - when to use them*
22/Feb	Linear regression (*2 dimensional data*)	2D data with one independent variable. *Rehash straight line equation:* `y = mx + c`. _show an interactive tutorial on how fitting works. Show `geom_smooth()` in R.
27/Feb	Linear regression fitting	key concept is the % of variance explained by the fit. *Do `lm` and show how to interpret results and p.values*
5/Mar	Re-capitulate t-tests using linear regression fitting	*hypothesis testing is akin to fitting 1D data + 2nd D of categorical variable*
7/Mar	Data -> statistics pipeline from a research article, **workshop**	using `Rmd/quarto` to walk through the steps required to reproduce statistics from a published paper
9-17/Mar	*Spring break*	no class
19/Mar	Non-linear regressions	examples: *fitting dose-response curves and bacterial growth curves*. Explain why initial conditions matter for nls, using self starting functions. `map; safely()` `workflow` to avoid breaking code due to convergence issues_

Date	Topic	Details
21/Mar	Bootstrapping	benefits: *a better non-parametric test without making any assumptions about the data distribution*
26/Mar	Bootstrapping in R	*using a vectorized workflow to achieve bootstrapping and simple hypothesis testing*
28/Mar	Working with higher dimensional data	explain dimensionality reduction concept. Techniques: `PCA`, weighted techniques, clustering. *likely using sequencing microbiome datasets examples*
2/Apr	critique statistics from a research article, **workshop**	using `Rmd/quarto` to recreate statistics from a published paper, finding faults, redoing analysis with better methods
4-9/Apr	TBD based on feedback (2 classes)	*Will push further to expand initial few topics if people are struggling..*
11/Apr	Student presentations 1	
16/Apr	Student presentations 2	
18/Apr	Student presentations 3	
24-30/Apr	Final exam	

Textbooks/Reference material

There is no single textbook for the course. There will be multiple references, mostly from free textbooks or materials available online. All required readings and supplementary materials will be posted on the course website. All code and data files will also be posted on the course canvas site. Code and data files will typically be made available before class meetings; slides will be posted afterwards.

Open textbooks on R and statistics

- Modern statistics with R
- Statistical Thinking for the 21st Century (Poldrack) : Libre text - Simulation, bayesian?
- An introduction to data analysis
- Introduction to Statistics with R : Libre text – lot of basics, hypothesis testing?

- Fundamentals of data visualization
- R for Data Science
  - Happy Git and GitHub for the useR
- Statistical Inference via Data Science : A ModernDive into R and the Tidyverse - moderndive

  > *We have intentionally minimized the number of mathematical formulas used. Instead, you'll develop a conceptual understanding of statistics using data visualization and computer simulations. We hope this is a more intuitive experience than the way statistics has traditionally been taught in the past and how it is commonly perceived.*
- Tidy Modeling with R -Tmwr : *how to use tidymodels packages ; develop good statistical practice*
- Foundations in Statistical Reasoning (Kaslik) - libretexts
- Other uses of R for graphics, reports, general automation in R without Statistics *by David Keyes*

Open online courses

- Allison horst's Intro data analysis and stats : Google slides and exercise material available
- Intro to R for biologists : Check out syllabus, _how to get course material?
- Computational biology foundations - syllabus
  - Check out any courses from this nice faculty teaching R for 10 years, *interested in incorporating LLMs into teaching from this question on posit.* Faculty : boris steipe
- Rfun/duke: Topics covered
  - Getting started: import data, data wrangling
  - Data wrangling with dplyr
  - Visualization with ggplot2
  - Coding with ChatGPT
  - Tidy data, pivot, join, and iteration (part 1)
  - Functions & {purrr}; iteration part2
  - Regression and tidymodels
- Nice slides for basics of R - tinystats/teacups-giraffes-and-statistics course/module 1

Classroom policies

Attendance

Since this course requires hands-on R use in class, and values participation, in-person attendance is highly encouraged. However, if you are unable to attend class in-person, please contact me in advance to avail the zoom option. There is no penalty for mussing upto 3 classes, additional absences will proportionally reduce the attendance part of the final grade.

Aside from excused absences, you are permitted to absences for *religious holidays* that are not included in the list of Federal and Academic holidays. Please let me know if you will be absent due to a religious holiday.

Late submissions

You are automatically entitled to two late submissions throughout the course. Any additional late submissions have a penalty of 10% of the grade for that assignment. If you need any accommodations please email the instructor.

In addition, the policy for religious holidays under **Attendance** extends into late submissions - if you have a religious holiday that occurs on the day of or the day before an assignment is due, please let me know and I can provide an additional extension.

Communication policy

You are welcome to email me at any time with your questions, concerns, and appointment requests. To ensure that I do not miss your email, please include *[`course ID`]-[short summary of the query]* in the email subject. For example, if you are requesting an additional appointment, the email subject should be: *[course ID]-scheduling appointment.*

Names and pronouns

If you prefer a different name or gender pronoun than the one displayed on Canvas, please let me know. All students are expected to refer to each other by their correct names and pronouns during class. When addressing groups of people, use gender-neutral language. Here is a resource to assist you with inclusive, respectful language: Pronouns.

Etiquette.

In addition, no racist vocabulary will be permitted in class in *any* capacity. Please refer to each other with respect.

Needs

If you face any challenges in securing personal or educational resources, such as a safe space and uninterrupted access to computing or the internet, and believe that these challenges may impact your performance in this class, please notify me. I will work with you to ensure that I can assist in any way possible.

Inclement weather policy

In the event of extreme weather conditions that impact your ability to get to class in person, please inform me and I will announce opening the zoom to everyone for that day/class

Academic honesty

Rice Honor Code

In this course, all students will be held to the standards of the *Rice Honor Code*, a code that you pledged to honor when you matriculated at this institution. If you are unfamiliar with the details of this code and how it is administered, you should consult the Honor System Handbook. This handbook outlines the University's expectations for the integrity of your academic work, the procedures for resolving alleged violations of those expectations, and the rights and responsibilities of students and faculty members throughout the process.

Attribution policy

Any collaborative work or help received should be duly acknowledge in the report under attributions section. For group submissions, please detail who did what for the assignment/report.

It is ideal to pre-determine how the work will be split by writing down the attribution section initially to prevent any confusion among the group.

Accessibility

Disability resource center

If you have a documented disability or other condition that may affect academic performance you should: 1) make sure this documentation is on file with the Disability Resource Center (Allen Center, Room 111, adarice@rice.edu, x5841) to determine the accommodations you need; and 2) talk with me to discuss your accommodation needs.

Furthermore, if you require course materials to be delivered in an alternate way to online .pdfs, please let me know.

Attributions

This course was taught by Prof. Lydia Beadrout for 6 years and materials from the spring 2023 course were heavily relied upon to create the current course. Classroom policies for this page were verbatim borrowed from my wife, Swetha Sridhar.