Data Management and Visualization

BI 410/510 | Fall 2022 | Klamath 5 | MF 10-11:20 am | 4 credits

Overview

This course covers the non-statistical aspects of the data life cycle, including how to store, clean, visualize and communicate data (Figure 1). It is intended as a complement to statistics courses - we will cover how to get your data into shape for analysis, and how to communicate your findings visually. It is primarily a methods class and will be taught in R (but there is no expectation that students know R coming in).

This course satisfies the "Analytical Approaches" requirement for Environmental Science majors and the "MAPS" requirement for Biology majors.

Instructor:
Dr. Amy Webster (awebst10@uoregon.edu) is a postdoctoral scholar in the Institute of Ecology and Evolution with research interests in epigenetic regulation, gene expression, and evolutionary genetics. She received her PhD in Genetics and Genomics from Duke University and her BS in Genetics and Math from the University of Georgia. Amy frequently uses R for data analysis and visualization, and she is passionate about sharing the tools necessary to create reproducible and aesthetically pleasing visualizations of complex data.

Teaching assistant (GE):
Cal Penkauskas (cpenkaus@uoregon.edu) is a second-year master's student in the Biology Department and previously completed his undergraduate at the UO in Environmental Science. His research in the Hallett Lab seeks to improve ecosystem management of working lands and resolve conflict between habitat conservation and agricultural production. Currently, he is investigating how hazelnut orchards and wildland habitat can support bird diversity across the landscape.

Office hours:
Amy: 2 hours on Tuesday mornings, exact time TBD by class poll. Office hours will take place in Pacific Hall room 318 (the office right next to the kitchen) or, if needed, via Zoom.
Cal: 1 hour on Wednesdays, exact time TBD by class poll. Office hours will take place in a TBD location. Please fill out the **linked poll by 9/30** so we are ready to go by week 2!

Canvas site:
Our website is accessible via the UO Canvas server; use your UO email and password to access the site. Pre-recorded videos and scripts will be available in modules for each week. Problem sets will be distributed and submitted via Canvas. https://canvas.uoregon.edu/

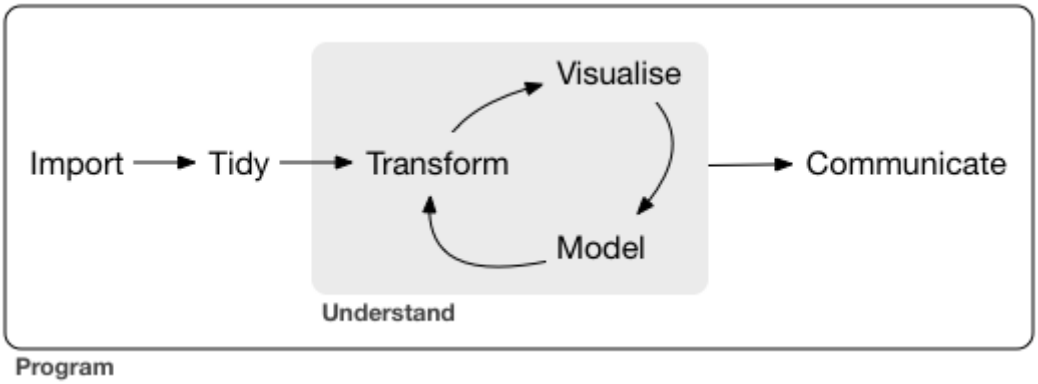

Figure 1: The data life-cycle (figure from Grolemund and Wickham)

Communication:
Our class will communicate primarily through our Canvas site. Announcements and emails are archived there and automatically forwarded to your UO email, and can even reach you by text. Check and adjust your settings under Account > Notifications. If you reach out to Amy or Cal, you can expect to receive a reply within approximately 1 business day; please do not expect quick replies on weekends or outside of business hours. If you send an email, please include 'BI 410/510' in the subject line of the email.

Course material

Our primary course material will be **R for Data Science** by Garrett Grolemund and Hadley Wickham, which is available for free online (http://r4ds.had.co.nz/); supplemental readings are uploaded to Canvas. Please make sure to read book sections and papers before the class in which they are assigned.

Objectives

By completing this course, students will be able to:
1) Interpret figures in scientific papers and popular media
2) Locate data relevant to biological and environmental questions
3) Understand the steps linking raw data to communicated findings
4) Create exploratory and publication-worthy graphs

Structure of the course

The course is broken into five approx. 2-week modules: Intro to R, Visualize, Transform, Wrangle, and Communicate. For each module the readings, scripts and assignments are posted in advance, along with short videos that overview key concepts and demonstrate their implementation in R.

Early in a module we will use class time to go over the concepts and scripts in depth, and later in a module we will use these times as "work sessions" for you to collaborate on the problem sets and other assignments. I strongly encourage watching the posted videos before class. This makes for essentially a "flipped" classroom setting, allowing us to spend less class time on lectures and more time completing the scripts and making substantial progress on the problem sets, aided by having your classmates and me available.

There is a lot of value in working collaboratively, but there may be times when you are unable to make it to class. When needed, students have been successful in this class working through the videos and scripts on their own time. So if you have to miss a day - no sweat, just work through the scripts and videos. If you have to miss an extended period, *please* communicate with me about it so we can make a plan to keep you on track.

Participation will be based primarily on your completion of the scripts and secondarily by interactive engagement (either in class or via discussion boards). **The one day I expect you to prioritize attendance is peer review day.** Attendance during peer review is important to give and receive feedback with your fellow classmates. *If you cannot attend please communicate with me in advance.*

Class assignments and requirements

There are two main components required for successful completion of the course.

A. Problem sets Problem sets are designed to develop the skills you learn in class and to gain comfort in the R environment through practice. Problem sets will typically include designing and implementing code and interpreting code and figures. There will be four problem sets, to be submitted on Canvas by midnight the day they are due. Students are encouraged to collaborate on problem sets, and to come to "work day" classes with ideas and questions.

B. Final project The focal experience of the class will be to develop a research project that addresses a biological or environmental question with data. Students will be expected to identify a question, contextualize the question with a literature review, analyze data relevant to answering the question, and interpret and communicate that data with a workflow in R. Students can use one of our pre-identified datasets, but if you have a data project of your own (particularly graduate students) we can discuss its suitability for this requirement.

Grade allocation

Grading will be based on a total of 200 points, where 90% of the points will earn an A, 80% a B, etc. Participation will reflect attendance and involvement in discussion and in-class exercises. The breakdown by assignments is as follows:

Assignment	Points
4 Problem sets (20 pts each)	80
Final project	
Part I: Literature review and proposed workflow	30
Part II: Peer review	20
Part III: Final paper	40
Participation	
Completed scripts (2 pts each)	18
Interaction (in class or message board)	12

Policies

1) Please note that assignments are due on Canvas. If an assignment is late, I will deduct 10% of the total points allocated to that assignment, and I will deduct 10% for each additional late day, up to 5 days. After that, the assignment will be graded with a starting point of half off (i.e., you can always receive up to half credit). The final project must be turned in on time, as final grades are due.

2) Please do not come to class if you are sick! I expect all students to actively participate in exercises and contribute to discussion, but this can be either in class, on the discussion boards, or both. There are enough resources posted on Canvas that you can be successful even if you cannot come in person for a period of time. Please communicate with me so we can make a plan for keeping you on track.

3) We will follow school policy of plagiarism and academic dishonesty. All students need to be familiar with the Student Conduct Code (https://policies.uoregon.edu/vol-3-administration-student-affairs/ch-1-conduct/student-conduct-code).

Deadlines

Completed in-class scripts are due by midnight on:
M 10/3 Intro to R
M 10/17 ggplot2 and geoms
M 10/31 rearrange and pipelines
M 11/14 relational data and tidy data
M 11/21 better graphics

Problem sets are due by midnight on:
M 10/10 PS 1
M 10/24 PS 2

M 11/7 PS 3
T 11/22 PS 4 (note Tuesday due date instead of Monday!)

Final project milestones are due by midnight on:
F 10/14 Topic ID
F 10/28 Literature review and workflow plan
M 11/21 Peer review

M 12/5 Final project

Please see the BI 410/510 calendar on Canvas (click on the Calendar icon in the dark green bar on the far left, then click on BI 410/510 in the Calendars drop-down) for a calendar view of deadlines. You can also link this to your preferred calendar app to keep track of deadlines! To do so, click on "Calendar Feed" and copy the link to your calendar app.

Course topics and tentative schedule

The topics on the tentative outline are subject to change. This is a guess, but we will take as long as needed on each lesson. Topics and lessons generally correspond to the noted chapter numbers in the book, additional readings may be posted to Canvas and emailed the week prior to when they should be read. Please note that the reading information is also summarized within each module on Canvas.

Day	Date	Module	Lesson
1	F 9/30	Visualize	Plotting before analyzing
2	M 10/3	General	Overview of R, R Studio Chapter 1, 6.1-6.3
3	F 10/7	General	`swirl` practice
4	M 10/10	Visualize	`ggplot2`: aesthetic mapping and facets 3.1-3.5
5	F 10/14	Visualize	`ggplot2`: geometric objects, coordinate systems 3.6-3.10
6	M 10/17	Visualize	RMarkdown and importing data Chapter 11, 27.1-27.4
7	F 10/21	Visualize	*Work day*
8	M 10/24	Transform	Rearranging data and `dplyr` 5.1-5.5
9	F 10/28	Transform	Grouping, summarizing and piping with `dplyr` 5.6-5.7, 18.1-18.3
10	M 10/31	Transform	Workflows
11	F 11/4	Transform	*Work day*
12	M 11/7	Wrangle	Relational data and joins with `dplyr` 13.1-13.7
13	M 11/14	Wrangle	Tidy data and `tidyr` 12.1-12.7
14	F 11/18	Communicate	Beautiful graphs Chapter 28
15	M 11/21	Wrangle	*Work day*
16	M 11/28	Communicate	Peer review
17	F 12/2	Communicate	RMarkdown v2 27.5-27.6, Odds and ends