Data Management and Visualization

BIO 410/510 | Winter 2020 | Klamath 5 | MF 12:00 - 1:20 pm | 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).

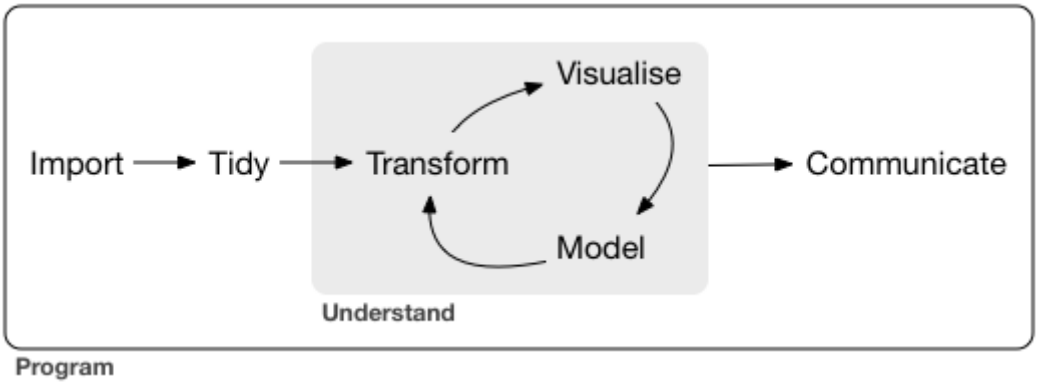

Figure 1: The data life-cycle (figure from Grolemund and Wickham)

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

Instructor: **Dr Lauren Hallett** is a plant community ecologist specializing in ecological restoration.
Email: hallett@uoregon.edu *Please include BIO 410/510 in email subject lines*
Office: Pacific Hall 220
Office hours: Tu 2-4 pm and by appointment

Canvas site: Our website is accessible via the UO Canvas server, use your UO email and password to access the site. Problem sets will be distributed and submitted via Canvas. https://canvas.uoregon.edu/

How we will contact you: Our communication to you outside of class will take place via Canvas 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/). I will post other readings 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

This class combines lectures on topics in data science and in-class exercises.

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 before class on 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, locate data relevant to answering the question, and interpret and communicate that data with a workflow in R. It is expected that students integrate data from multiple sources. In general, students will be using public data, 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 completion of 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	30

##Policies 1) Please note that assignments are due to Canvas before class. 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.

2) All missed classes need to be approved with the instructor prior to the start of class. Unexcused absences will result in the deduction of participation points.

3) This class includes frequent in-class exercises and workdays. I expect all students to actively participate in exercises and discussions.

4) 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).

5) Mac laptops will be provided for in-class use. The laptops wipe all personal data between classes and

whenever you are logged off. Please don't log off the computers, and make sure to back up your work at the end of class. If you wish to use your own laptop please make sure to have R, RStudio and TeX installed. Most UO computers in computing labs have these programs installed, but please talk to me if you have trouble accessing computers or software out of class.

##Deadlines Problem sets are due before class on:
F 1/17 PS 1
M 2/3 PS 2
F 2/17 PS 3
M 3/2 PS 4

Final project due dates are:
F 1/24 Topic proposed
F 2/7 Data sources identified
F 2/28 Literature review and workflow plan
F 3/6 Peer review
W 3/18 Final project

##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 will be posted to Canvas and emailed the week prior to when they should be read.

Day	Date	Topic	Lesson
1	M 1/6	Visualize	Plotting before analyzing
2	F 1/10	General	Overview of R, R Studio
3	M 1/13	General	`swirl` practice and importing data 11.1-11.6
4	F 1/17	Visualize	`ggplot2`: aesthetic mapping and facets 3.1-3.5
-	**M 1/20**	**MLK Day**	**No class**
5	F 1/24	Visualize	`ggplot2`: geometric objects, coordinate systems 3.6-3.10
6	M 1/27	Visualize	RMarkdown
7	F 1/31	Visualize	*Work day*
8	M 2/3	Transform	Rearranging data and `dplyr` 5.1-5.5
9	F 2/7	Transform	Grouping and summarizing with `dplyr` 5.6-5.7
10	M 2/10	Transform	Workflows and the pipeline 6.1-6.3, 18.1-18.3
11	F 2/14	Transform	*Work day*
12	M 2/17	Wrangle	Relational data and joins with `dplyr` 13.1-13.7
13	F 2/21	Wrangle	Workflows with relational data 12.3-12.6
14	M 2/24	Wrangle	*Work day*
15	F 2/28	Wrangle	Tidy data and `tidyr` 12.1-12.2, 12.7
16	M 3/2	Communicate	Good vs bad graphs
17	F 3/6	Communicate	Peer review
18	M 3/9	Communicate	Spatial visuals *with Joanna Merson*
19	F 3/13	Communicate	Odds and ends