Data Management and Visualization

ENVS 410/510 | Fall 2020 | Online! | MW 4:15-5:45 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).

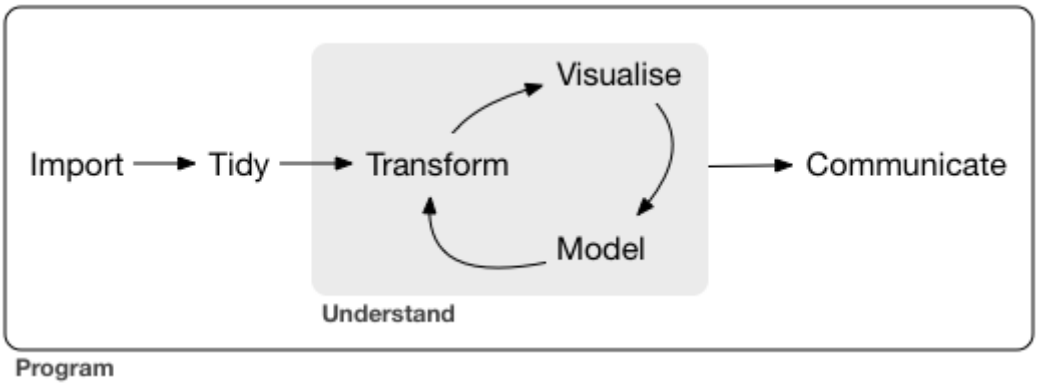

Figure 1: The data life-cycle (figure from Grolemund and Wickham)

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

Teaching team

Instructor: **Dr. Lauren Hallett** (she/her) is an assistant professor of plant community ecology specializing in ecological restoration. Data synthesis is fundamental to her work and she has a passion for developing and teaching open, reproducible science.

Co-instructor: **Dr. Andrew Muehleisen**. Andrew (he/him) is postdoc jointly affiliated with the Institute of Ecology and Evolution and the Data Science Initiative. His background is in tropical ecology, and his work with the DSI aims to improve data literacy and best practices for data science education.

GE: Lina Aoyama (Batas) (she/her) is a PhD candidate in ecology interested in producing and communicating science that improves restoration on rangelands.

Note: Lauren will lead the first half of the class, with Andrew and Lina available as resources. Lauren will be on maternity leave starting in early November, at which point Andrew will lead the rest of the term. Because the timing of this transition is not 100% predictable, we recommend directing emails about the class to both Lauren and Andrew.

Email: hallett@uoregon.edu, amuehlei@uoregon.edu, laoyama@uoregon.edu
Please include ENVS 410/510 in email subject lines

Office hours: TBD by a poll in first week of class

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/

How we will contact you: Our class will communicate 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.

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 2-week modules: Intro to R, Visualize, Transform, Wrangle, and Communicate. For each module we will post the readings, scripts and assignments in advance, along with short videos that overview key concepts and demonstrate their implementation in R. For those of you that said you liked the option for asynchronous learning and the ability to work at your own pace - this is for you!

In parallel, we will meet during class time over Zoom. Early in a module we will use this time to go over the 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. For those of you that said you wanted a synchronous time to ask questions and engage with others - this is for you!

For everyone, we **strongly encourage watching the videos and then attending the synchronous class sessions**. This is essentially a "flipped" classroom setting, and we anticipate that you will be able to complete the scripts and make substantial progress on the problem sets within class, aided by having your classmates and us available. Because this is an uncertain time, participation will be based primarily on your completion of the scripts and secondarily by interactive engagement (either in class or via discussion boards). This is to make sure you aren't penalized for in-the-moment tech issues or life events that conflict with class. However, there are three key days in which we ask you to prioritize synchronous attendance:

1) W 9/30 (the first day of class)
2) M 11/30 (peer review)
3) W 12/2 (guest lecture)

If you cannot attend one of these days please communicate with us in timely manner. Attendance during peer review is particularly important to give and receive feedback with your fellow classmates.

Please communicate with us if you have trouble with access and we will make a game plan to accommodate you. We will also host "live" office hours through Zoom each week.

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. In general, students will be using 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, we will deduct 10% of the total points allocated to that assignment, andwe will deduct 10% for each additional late day.
2) This class includes frequent exercises and workdays. We expect all students to actively participate in exercises and contribute to discussion (either in class, on the discussion boards, or both). Please communicate with us if you have difficulty attending class times.
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

Problem sets are due by midnight on:
F 10/9 PS 1
F 10/23 PS 2

F 11/6 PS 3
F 11/20 PS 4

Final project due dates are:
M 11/23 Literature review and workflow plan
Su 11/29 Peer review W 12/9 Final project

Please see the ENVS 410/510 calendar on Canvas (click on the Calendar icon in the dark green bar on the far left, then click on ENVS 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 will 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	W 9/30	Visualize	Plotting before analyzing
2	M 10/5	General	Overview of R, R Studio Chapter 1, 6.1-6.3
3	W 10/7	General	`swirl` practice
4	M 10/12	Visualize	`ggplot2`: aesthetic mapping and facets 3.1-3.5
5	W 10/14	Visualize	`ggplot2`: geometric objects, coordinate systems 3.6-3.10
6	M 10/19	Visualize	RMarkdown and importing data Chapter 11, 27.1-27.4
7	W 10/21	Visualize	*Work day*
8	M 10/26	Transform	Rearranging data and `dplyr` 5.1-5.5
9	W 10/28	Transform	Grouping, summarizing and piping with `dplyr` 5.6-5.7, 18.1-18.3
10	M 11/2	Transform	Workflows
11	W 11/4	Transform	*Work day*
12	M 11/9	Wrangle	Relational data and joins with `dplyr` 13.1-13.7
13	W 11/11	Wrangle	Tidy data and `tidyr` 12.1-12.7
14	M 11/16	Wrangle	*Work day*
15	W 11/18	Wrangle	*Work day*
16	M 11/23	Communicate	Beautiful graphs Chapter 28
17	W 11/25	Communicate	RMarkdown v2 27.5-27.6
18	M 11/30	Communicate	Peer review
19	W 12/2	Communicate	Color theory *with Joanna Merson*