Best Practices in Environmental Data Science

FISH 549   University of Washington

Contents

Course overview

This course will provide an overview of some data science tools and best practices that can be used to create transparent and reproducible workflows when working with environmental data. Students will learn how to translate raw data from field and lab studies into databases and "tidy" digital formats, which can then be used for plotting, statistical analyses, etc. Students will learn how to track the history of file changes (version control), collaborate online with others, and generate "recipes" for re-creating one's work. Although failure and frustration in science are common, the open science community tries hard to be welcoming and helpful. Thus, students will also learn how and where to ask for help when attempting something new (*e.g.*, How do I create X from Y?), debugging or fixing code (*e.g.*, What does this error message mean?), etc.

Importantly, this course will also discuss the various types of data used in environmental science, which includes an exploration of what constitutes "raw" data versus derived products, and the implications of using the various types. The course will also focus on the ethical and social consequences of using various data types. For example, some recent studies have inferred gender identities from survey responses based on first names, and used that information to summarize hiring practices, demographics of successful grant applications, etc. However, doing so is widely regarded as inappropriate due to non-binary genders and inaccuracies in gender assignment.

Learning objectives

By the end of the quarter, students should be able to:

- Explain the 6 steps of a formal data analysis, which include

1. Defining the question of interest
2. Getting the data
3. Cleaning the data
4. Exploring the data
5. Fitting statistical models
6. Communicating the results

- Use a research compendium to organize data, code, and results in a unified framework

- Create a detailed workflow that will allow others to fully reproduce an analysis of data

Specifically, students should have the necessary hands-on skills to:

- Use **Git** to commit changes to files, recover old versions, push/pull changes to remote repositories, and manage merge conflicts

- Use **GitHub** to create repositories, manage projects, open/comment/close issues, and submit pull requests

- Create a reproducible example to use in a quest for help

- Import and clean a messy data file using a variety of `{packages}::functions` in **R**

- Properly describe and document data using **Ecological Metadata Language**

- Create and access a relational database with **PostgreSQL**

- Use unit tests in **R** to evaluate code functionality

- Create a package in **R** and document its contents

- Use **R Markdown** to combine text, equations, code, tables, and figures into reports, websites, and presentations

- Create dynamic html reports with **Shiny**

Instructor

Mark Scheuerell
Associate Professor, School of Aquatic & Fishery Sciences
scheuerl@uw.edu

Credits

This course is 3 credits.

Meeting times

M/W/F from 10:30-11:20.

Office hours

Fridays from 11:30-12:30.

Pre-requisites

Students should have a working knowledge of the **R** computing software, such as that provided in FISH 552/553.

Classroom conduct

I am dedicated to providing a welcoming and supportive learning environment for all students, regardless of their background, identity, physical appearance, or manner of communication. Any form of language or behavior used to exclude, intimidate, or cause discomfort will not be tolerated. This applies to all course participants (instructor, students, guests). In order to foster a positive and professional learning environment, I ask the following:

- Please let me know if you have a name or set of preferred pronouns that you would like me to use

- Please let me know if *anyone* in class says something that makes you feel uncomfortable[1]

In addition, I encourage the following kinds of behaviors:

- Use welcoming and inclusive language

- Show courtesy and respect towards others

- Acknowledge different viewpoints and experiences

- Gracefully accept constructive criticism

Although I strive to create and use inclusive materials in this course, there may be overt or covert biases in the course material due to the lens with which it was written. Your suggestions about how to improve the value of diversity in this course are encouraged and appreciated.

Please note: If you believe you have been a victim of an alleged violation of the Student Conduct Code or you are aware of an alleged violation of the Student Conduct Code, you have the right to report it to the University.

Access & accommodations

All students deserve access to the full range of learning experiences, and the University of Washington is committed to creating inclusive and accessible learning environments consistent with federal and state laws. If you feel like your performance in class is being impacted by your experiences outside of class, please talk with me.

Disabilities

If you have already established accommodations with Disability Resources for Students (DRS), please communicate your approved accommodations to me at your earliest convenience so we can discuss your needs in this course. If you have not yet established services through DRS, but have a temporary health condition or permanent disability that requires accommodations (*e.g.*, mental health, learning, vision, hearing, physical impacts), you are welcome to contact DRS at 206-543-8924 or via email or their website. DRS offers resources and coordinates reasonable accommodations for students with disabilities and/or temporary health conditions. Reasonable accommodations are established through an interactive process between you, your instructor(s) and DRS.

Religious observances

Washington state law requires that UW develop a policy for accommodation of student absences or significant hardship due to reasons of faith or conscience, or for organized religious activities. The UW's policy, including more information about how to request an accommodation, is available at Religious Accommodations Policy. Accommodations must be requested within the first two weeks of this course using the Religious Accommodations Request Form.

Technology

This course will revolve around hands-on computing exercises that demonstrate the topics of interest. Therefore, students are strongly recommended to bring their own laptop to class, although students are certainly free to work with one another. **For students without access to a personal laptop**: it is possible to check out UW laptops for an entire quarter (see the Student Services office for details).

All of the software we will be using is free and platform independent, meaning students may use macOS, Linux, or Windows operating systems. In addition to a web browser, we will be using the free **R** software and the desktop version of the **R Studio** integrated development environment (IDE). We will also be using various packages not contained in the base installation of **R**, but we will wait and install them at the necessary time. The instructor will be available during the first week of class to help students troubleshoot any software installation problems.

We will be using **Git**, a free and open source version control system for tracking changes to our files.

We will be using **PostgreSQL** to create and access relational databases.

Students will also be required to have a user account on **GitHub**, which we will be using for file hosting and communications via "issues". If you do not already have an account, you can sign up for a free one here. The instructor will provide training on how to use the intended features in **GitHub**.

Teaching methodology

This course will introduce new material primarily through prepared slides and hands-on demonstrations. Students will be expected to work both individually and collaboratively (to the extent possible given the current conditions); course content and evaluation will emphasize the communication of ideas and the ability to think critically more so than a specific pathway or method. Other areas of this website provide an overview of the topics to be covered, including links to weekly reading assignments, lecture materials, computer labs, and homework assignments.

Communication

This course will involve a *lot* of communication between and among students and the instructor. Short questions should be addressed to me via email; I will try my best to respond to your message within 24 hours. Under more normal circumstances, detailed questions would be addressed to me in person–either after class or during a scheduled meeting. In this case, however, we will schedule one-on-one or group Zoom calls as needed.

In addition to email and Zoom, we will use the "Issues" feature in **GitHub** to ask questions and assist others. Specifically, questions and answers can be posted to the issues in the course's "assistance" repository here.

Evaluation

Students will be evaluated on their knowledge of course content and their ability to communicate their understanding of the material via individual homework assignments (90%) and their active participation in class (10%). There will be 9 homework assignments, each counting toward 10% of the final grade. Please note, **all assignments must be turned in to achieve a passing grade**.

Homework

Homework will be assigned each Friday of weeks 1-9 and will be due by 11:59 PM 9 days later on the following Sunday (weeks 2-10). The instructor will evaluate and return student homework assignments within one week of their due date. **If for some reason you cannot meet the homework deadline, contact the instructor as soon as possible to discuss other options**.

Participation

Data science is about communicating your ideas and findings to others. As such, this course will require students to engage with the instructor and one another in small groups. I expect all students to contribute to our discussions of concepts, results, bugs/errors, etc. In addition, there will be 2 in-class quizzes, each worth 5% of your final grade, which cannot be made up after the fact. Students should discuss any potential schedule conflicts with the instructor during the first week of class.

Academic integrity

Faculty and students at the University of Washington are expected to maintain the highest standards of academic conduct, professional honesty, and personal integrity. Plagiarism, cheating, and other academic misconduct are serious violations of the Student Conduct Code. I have no reason to believe that anyone will violate the Student Conduct Code, but *I will have no choice* but to refer any suspected violation(s) to the College of the Environment for a Student Conduct Process hearing. Students who have been guilty of a violation will receive zero points for the assignment in question.

Grievance procedures

The College of the Environment Student Academic Grievance Procedures provide mechanisms for enrolled students to address academic problems or grievances in an equitable, respectful and timely manner. Academic grievances are defined as those involving conflicts between a student or students and their course instructors (including faculty and teaching assistants) or research mentor(s) with respect to differences arising within credit-bearing work and while the student is registered at the University of Washington. If you have or are experiencing such a conflict in this class, and have not, cannot, or do not wish to attempt resolution with me, I encourage you to explore additional options open to you by accessing the website above.

Mental health

We are in the midst of an historic pandemic that is creating a variety of challenges for everyone. If you should feel like you need some help, please consider the following resources available to students.

If you are experiencing a life-threatening emergency, please dial 911.

Crisis Clinic
Phone: 206-461-3222 or toll-free at 1-866-427-4747

UW Counseling Center
Phone: 206-543-1240
Immediate assistance

Let's Talk

Hall Health Mental Health

Safety

If you feel unsafe or at-risk in any way while taking any course, contact SafeCampus (206-685-7233) anytime–no matter where you work or study–to anonymously discuss safety and well-being concerns for yourself or others. SafeCampus can provide individualized support, discuss short- and long-term solutions, and connect you with additional resources when requested. For a broader range of resources and assistance see the Husky Health & Well-Being website.

Food Pantry

No student should ever have to choose between buying food or textbooks. The UW Food Pantry helps mitigate the social and academic effects of campus food insecurity. They aim to lessen the financial burden of purchasing food by providing students access to shelf-stable groceries, seasonal fresh produce, and hygiene products at **no cost**. Students can expect to receive 4 to 5 days' worth of supplemental food support when they visit the Pantry, located on the north side of Poplar Hall at the corner of NE 41st St and Brooklyn Ave NE. Visit the Any Hungry Husky website for additional information, including operating hours and additional food support resources.

Schedule

Note: The following specific dates are based upon the Winter 2021 Quarter and will change accordingly in the future.

Date	Topics
4 Jan	Course overview; What is data science?; Types of questions; Storytelling
6 Jan	Steps to an analysis; Informal vs formal analysis
8 Jan	Organization & styles; Research compendia; Coding styles
11 Jan	Introduction to GitHub: Repositories, Markdown, Issues, Projects
13 Jan	Version control with Git: Adding, Committing, Reverting
15 Jan	Remotes in GitHub; RStudio projects; Pushing/pulling
18 Jan	** Martin Luther King, Jr. Day - No class **
20 Jan	Understanding workflows: Centralized, Feature branch, Forking
22 Jan	Working with data: Chain of custody, Tidy data, Manual entry
25 Jan	Getting data: Importing text files, Reading from Excel & Google
27 Jan	Getting help; Creating a Minimal Reproducible Example (MRE)
29 Jan	Guest presentation: Kara Woo - Data science for fun and nonprofit
1 Feb	Introduction to relational databases
3 Feb	Creating a database; New users & databases; Adding data
5 Feb	Working with a database; Cleaning data; Joining tables
8 Feb	Review session; Workflows, Reprexes, Databases
10 Feb	Metadata, Ecological Metadata Language
12 Feb	Data analysis & visualization: Data summaries, Design principles, Exploratory & Expository figures
15 Feb	** President's Day - No class **
17 Feb	R Markdown: Introduction, Publishing papers
19 Feb	R Markdown: Publishing papers, Reporting
22 Feb	Guest presentation: Dr. Christine Stawitz - Data science skills for a career in fisheries management
24 Feb	Building R packages: Functions, Documentation
26 Feb	Building R packages: Unit tests, Vignettes, Websites
1 Mar	Introduction to Shiny apps
3 Mar	Deploying Shiny apps
5 Mar	R Markdown: Designing presentations
8 Mar	R Markdown: Building simple websites
10 Mar	Code reviews: Checklists, Do's and Don'ts
12 Mar	Ethics in data science

Endnotes

[1] If the instructor should be the one to say something that makes a student uncomfortable, the student should feel free to contact the Director of the School of Aquatic and Fishery Sciences.