Data Sci & Provenance

Ryan Peek

Fall 2021

E-mail: rapeek@ucdavis.edu
Office Hours: TBA
Office: CWS

Web: TBA
Class Hours: TBA
Class Room: *online*

Course Description

While there remain many obstacles to tackling the growing challenges we face, significant progress can be made through implementation of basic data science education including understanding and using best practices for data management; and fostering open science practices that support replication and synthesis (e.g., developing and teaching open science research workflows, making datasets and code freely available).

This seminar will be open to all CWS data denizens (so anyone who works with data) to train participants in the basic steps of using best practices in data science and project provenance using CWS specific projects. Each participant is expected to have a project in mind that they currently work on or expect to work on which we will actively use to learn each phase of the data life cycle. Furthermore, these steps are applicable regardless of the programming language or analytical tool being used, from Excel to R, to Python, and beyond.

Seminar Objectives

By the end of the seminar, each participant should understand and be able to implement:

1. A Data Plan & Process (what is data plan, where will it live, how will it be stored)
2. Metadata Collection and Processing
3. Data Assurance and Analysis (basic data wrangling skills)
4. How to Preserve and Store Data (version control and repositories)
5. Publish and Share (make data accessible)
6. Discover and Integrate (process is reproducible and can be built on)

Prerequisites

The class requires no prior knowledge of programming, nor is this an ecology or biology class. This seminar is about how to understand, organize, use, and share data using reproducible workflows.

Computing requirements

Data Science and associated workflows need to be learned by doing, and the seminar and associated class time will focus on working through problems and projects together.

Everything will be available online. All lectures will be recorded and made available.

Course Materials

All materials and assignments will be posted on a webpage, and seminars will be recorded.

Schedule

Week 01, 09/24 - 09/28: Data Plan and Process

Read associated documents on course website.

- Taking Good Notes

Week 02, 10/01 - 10/05: Metadata

- Understand and use metadata
- Identify your own metadata
- Create a metadata file to store with your datasets

Week 03, 10/08 - 10/12: Data Wrangling

Some basic data wrangling skills will be covered using R:

- Joining data
- Pivoting data (go from wide to long/long to wide)
- Filtering data
- Import/Export

Week 04, 10/15 - 10/19: Preserve and Store Data

How do we keep our data safe and accessible?

- Open source repositories
- Version control

Week 05, 10/22 - 10/26: Publish and Share Data

- Go over various repositories, pros/cons, and how to use (i.e., Dryad vs. OSF)

Week 06, 10/29 - 11/02: Integration and Updating

If we have interest and time, talk about reproducible workflows and toolsets

- `snakemake`
- `targets`