THAYER SCHOOL OF
ENGINEERING
AT DARTMOUTH

Winterim Course
Principles and Practices of Reproducible Research

Dates of course: December 9–12, from 9am to 12pm

Instructors: Library and Research ITC staff

Intended audience: All Dartmouth students, especially those with an interest in STEM research.

Prerequisites: None. This course is designed to be applicable to students engaging at every level of research and scholarly work.

Description: Computationally intensive and data-driven research and scholarly work has become ubiquitous across most disciplines, and students need to build on and expand their research skills. However, while researchers spend more time managing data and using code, we are rarely taught to do this efficiently, effectively, or in a computationally reproducible manner.

Our reproducible research short course integrates data management, computing resources, and open science principles. It will provide straightforward, hands-on, fundamental skills training that researchers at any level can apply to produce reliable and reproducible computationally intensive scholarship. Implementing reproducible research concepts and tools is a key to the scientific method. It allows others (including future you) to recreate, verify, and build upon your work by structuring, organizing, and analyzing your data and code using reproducible best practices.

By the end of the course, participants will be able to:
- Articulate the principles of reproducible research including terms and concepts related to data and data science
- Identify and utilize appropriate tools for different data needs and use cases
- Formulate specific strategies for incorporating data and data related reproducible computational tools into their own work

Daily schedule: Students attending this course will learn practical tools and tips to increase their research reproducibility, their computational reproducibility, and their research openness.

- Dec 9: In the first session, we will introduce the reproducible research workflow and best practices to more effectively manage data and code, including documentation, organization, and file management systems. We will also begin to use OpenRefine, an open-source tool for cleaning messy data.
- Dec 10: Focusing on computational reproducibility, we begin working with GIT as a version control tool for source code management allowing us to back-up, synchronize, and share our research with collaborators. We will also learn to clean and analyze research data using R, an open source programming language.
- Dec 11: We will continue with our scripting lesson, with a focus on dataframes and data visualization. Using scripting to analyze and present data cuts down on the time spent performing repetitive, time-consuming analyses, and makes research more transparent and reproducible.
- Dec 12: The last session will wrap-up the computational reproducibility content with the introduction of Docker containerization to create an executable package of software that includes everything needed to reproduce an application. The course will end with reproducible dissemination, including using open licensing options for sharing software, data, and results in trusted repositories and more.