JSM 2025 Short Course: Large Local Data

by: Kelly Bodwin, Tyson Barrett & Jon Keane

Course URL: kbodwin.github.io/jsm-big-data-2025

Welcome!

Welcome to this Short Course on **Storing, Importing, Managing, and Analyzing Large Data Locally with R**. The goal for the course is to provide training in modern tools for working with large data that fits on-disk on a local machine.

This workshop is geared towards researchers and practitioners who regularly use R for their data preparation and/or analysis, but who may not be familiar with strategies and tools for speeding up processes, especially on larger datasets.

Catalog Description

It is increasingly common in academic and professional settings to encounter datasets large enough to exceed the capabilities of standard data processing tools, yet small enough to be stored on local computers. Recent articles even claim that "the era of big data is over" and that data analysts and researchers should "think small, develop locally, ship joyfully". Such "medium" dataests are instrumental in measuring, tracking, and recording a wide array of phenomena across disciplines such as human behavior, animal studies, geology, economics, and astronomy.

In this workshop, we will present modern techniques for handling large local data in R using a tidy data pipeline, encompassing stages from data storage and importing to cleaning, analysis, and exporting data and analyses. Specifically, we will teach a combination of tools from the data.table, arrow, and duckDB packages, with a focus on parquet data files for storage and transfer. By the end of the workshop, participants will understand how to integrate these tools to establish a legible, reproducible, efficient, and high-performance workflow.

Intended Audience and Level

We expect attendees to have R fluency at the level of a typical introductory course, such as the textbook R for Data Science (Wickham, Çetinkaya-Rundel, & Grolemund 2023); as well as familiarity with some data application that may motivate tools beyond the introductory level.

Schedule

Time	Topic	Presenter
8:30 - 9:00	Introductions, Installs, Set-up	all
9:00 - 9:45	I. Identify slowdowns and solutions	Kelly
9:45 - 10:15	Activity 1	Kelly
10:15 - 10:30	Break	
10:30 - 11:30	II. data.table for fast wrangling	Tyson
11:30 - 12:00	Activity 2	Tyson
12:00 - 1:30	Lunch Break	
1:30 - 2:30	III. arrow, parquet, duckdb	Jon
2:30 - 3:00	Activity 3	Jon
3:00 - 3:30	Break	
3:30 - 4:30	IV. Workflow	all
4:30 - 5:00	Activity 4	all

Workshop Materials

Workshop materials can be found in the github repository github.com/kbodwin/jsm-big-data-2025/materials!