



## PhD Program in Plant Sciences: Advanced Data Management and Manipulation using R

**Lecturers:** Dr. Jan Wunder  
**Location:** ETH center, CLA J1  
**Dates:** June 07 & 14, 2024  
**Credit Points:** 1 ECTS

### Course Objectives

The analysis of large data sets (“big data”) is becoming increasingly important in science and elsewhere. In this course you will learn how to use R to manage and manipulate large data sets, i.e. to sort, merge, subset, aggregate and reshape data. We will solve together several data quality issues (missing values, outliers, zeros, normalization etc.) and highlight the underlying theoretical foundations and computational frameworks to obtain high-quality data. For advanced data manipulation, we are going to use novel developments such as dplyr (“A Grammar of Data Manipulation”), the pipe operator (`%>%`) for simpler R-coding and `data.table` for the fast aggregation of large data sets. Furthermore, we will have a closer look at R-data base connections, SQL queries and the creation of new data bases from R. Depending on the course progress, there will be scope for individuals to work on small projects and / or their own data sets.

### Prior Knowledge:

Basic knowledge of R / attendance of the course “Introduction to R”

**Number of Participants:** Participation is limited to 16 people.

### Students are required to bring their own computers:

Please make sure that you install the most recent versions of R and RStudio.

### Individual Performance and Assessment:

In order to obtain the credit points, participants are required to attend *both* course days and hand in an assignment to be carried out at home. The details will be explained during the course. The assignment is due no later than one week after the course has ended.