Workshop: Simplifying data manipulation in R (90 min)

Workshop objectives

- Introduce the concept of "tidy data" and demonstrate how data can easily be reshaped using package `tidyr`
- Show how data manipulation in R can be made more efficient using package `dplyr`

NOTE that in 90 mins we have barely time to scratch these topics. The best outcome is therefore that the participants are aware of them and can start exploring the topics and packages in more detail.

Description

For already a good while, R has been *the* go-to computational environment for any one us in need of statistical tools. While R does offer a dizzying variety of analytic tools, much less effort has gone into simplifying the data wrangling parts that most of us actually spend the most of our time struggling with. Luckily, things have gotten a lot better lately specifically due to two packages: `tidyr` and `dplyr`.

tidyr is new package that makes it easy to "tidy" your data. Tidy data is data that's easy to work with: it's easy to munge (with `dplyr`), visualise (with `ggplot2` or `ggvis`) and model (with R's hundreds of modelling packages). The two most important properties of tidy data are:

- Each column is a variable.
- Each row is an observation.

`dplyr` package offers simple, clear and efficient way of working with your data. The package makes the most common data manipulation steps as fast and easy as possible by:

- Elucidating the most common data manipulation operations, so that your options are helpfully constrained when thinking about how to tackle a problem.
- Providing simple functions that correspond to the most common data manipulation verbs, so that you can easily translate your thoughts into code.
- Using efficient data storage backends, so that you spend as little time waiting for the computer as possible

Prerequisites

A working knowledge in R is assumed. If you've written more than 10 lines of code then you can probably follow what's going on. We'll be running some example code (to be distributed in before hand) in the workshop, so bring your own laptop. Make sure you have `R` (>=3.1.0), `tidyr` (>=0.2.0) and `dplyr` (>=0.4.1) installed. **We will not have time to install these in the workshop!** RStudio is *highly* recommended but not required.

Recommended reading

- Wickham, Hadley (2014): Tidy data. Journal of Statistical Software. 59(10). pp. 1-23.
- "Data Wrangling with dplyr and tidyr" cheat sheet