Setting up R projects

Theoretical and Empirical Research Methodology, Lab 3

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

Europa-Universität Flensburg

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

Goals for today

I. Learn how to set up an R project

II. Learn about the difference between absolute and relative paths

III. Learn how to use the here package

Our goal

- Learn about a **default directory structure** and a general way to **document everything** you do in your project

  → Facilitates the collaboration with future-you considerably

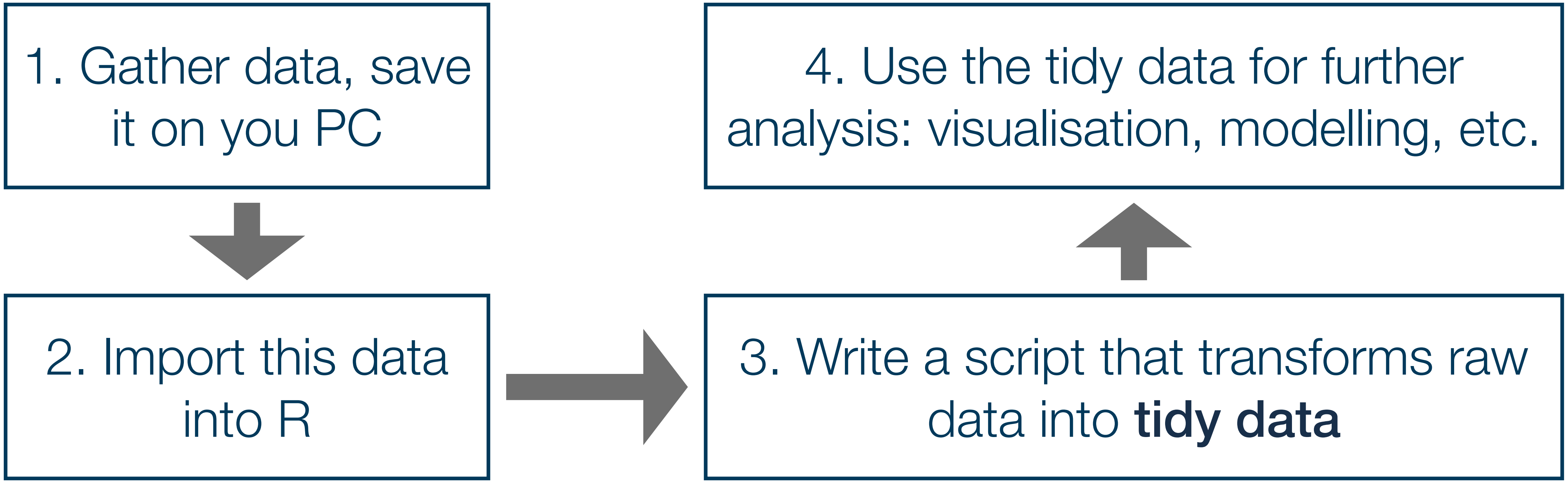
  ☝️Nothing is worse than hating your past-you for not documenting correctly where data came from, or how it has been prepared 🫣

- Introduce general workflow to avoid many problems in the context of project management

- Central idea: all results must always be **reproducible** from the raw data
  - You **must not manipulate your raw data** at any cost
  - Raw data = what you download from the internet, gather through an experiment, or code yourself

How to keep your work transparent

- Raw data must not be changed, but is usually not in a state we can work with 🤨



1. Gather data, save it on you PC

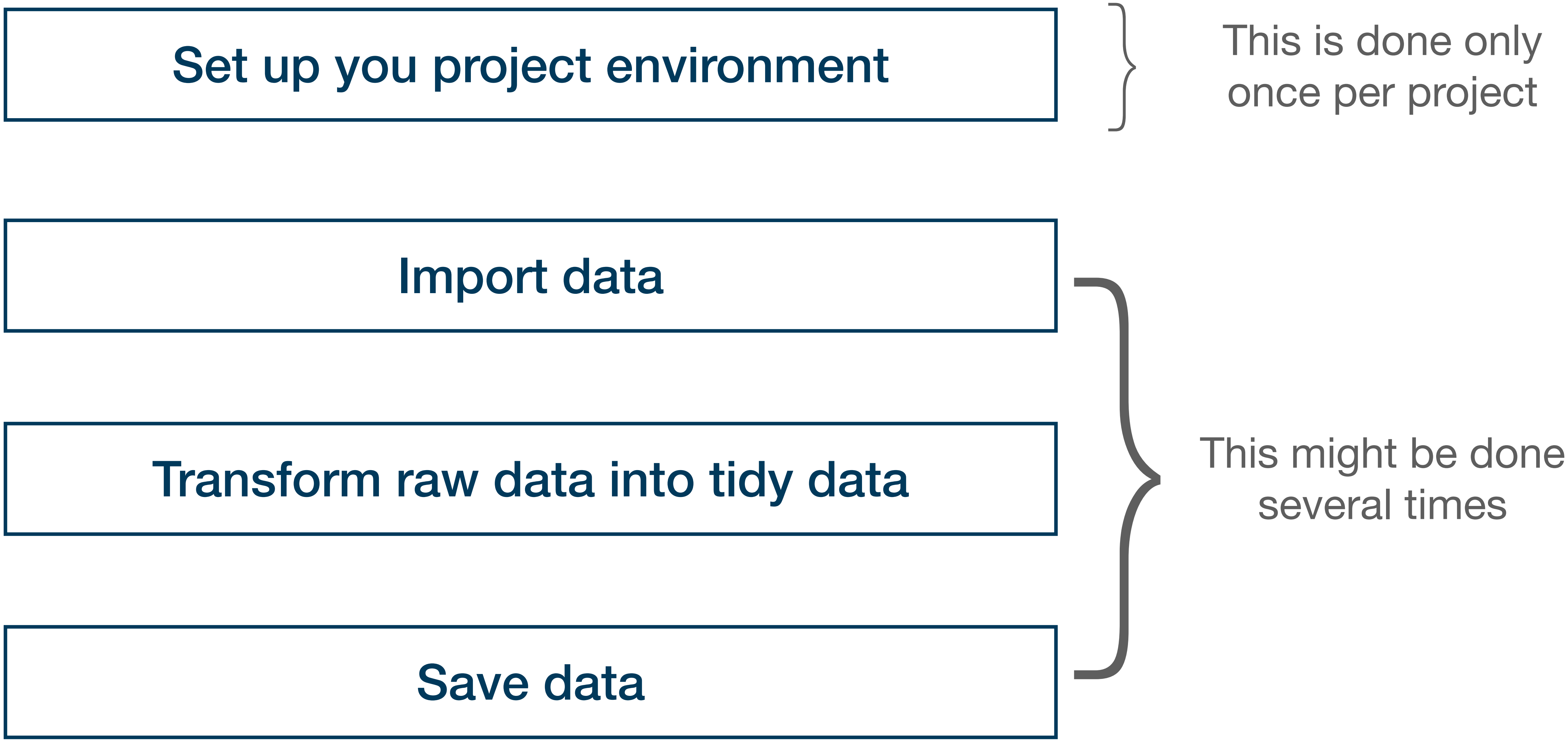
2. Import this data into R

3. Write a script that transforms raw data into **tidy data**

4. Use the tidy data for further analysis: visualisation, modelling, etc.

- Saving the scripts in steps 2 & 3 makes your work **fully reproducible**
- The scripts always tell you what you did to your raw data → you can heal basically every mistake you made!

Outlook



Set up you project environment

This is done only once per project

Import data

Transform raw data into tidy data

Save data

This might be done several times

Outlook

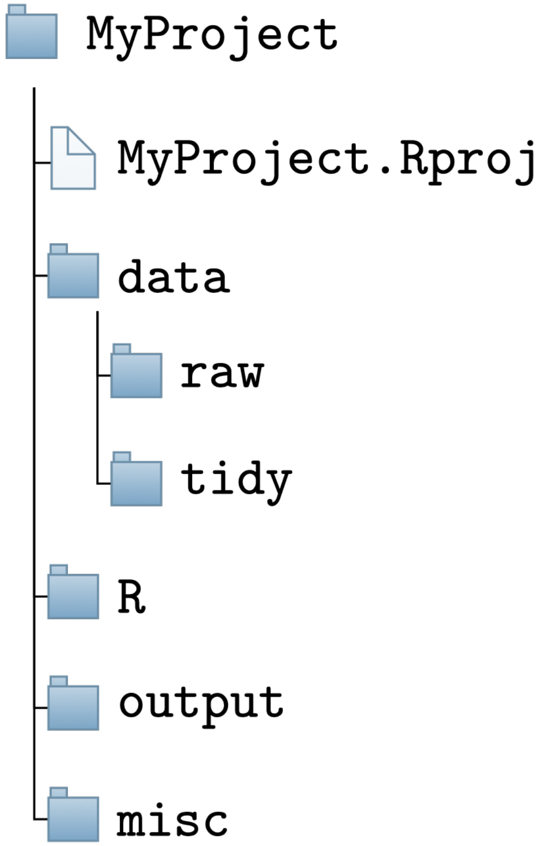
Set up you project environment

This is done only once per project

Set up your R project

Setting up your working environment

- Where you save your data (and files) is important!
- The right **directory structure** as a prerequisite for transparent, reproducible, and easy-to-work-with project
- For every task in R you should set up your project like this:
- All relevant steps to set this up, and the rationality for this structure are described in the respective **tutorial**



MyProject

MyProject.Rproj

data

raw

tidy

R

output

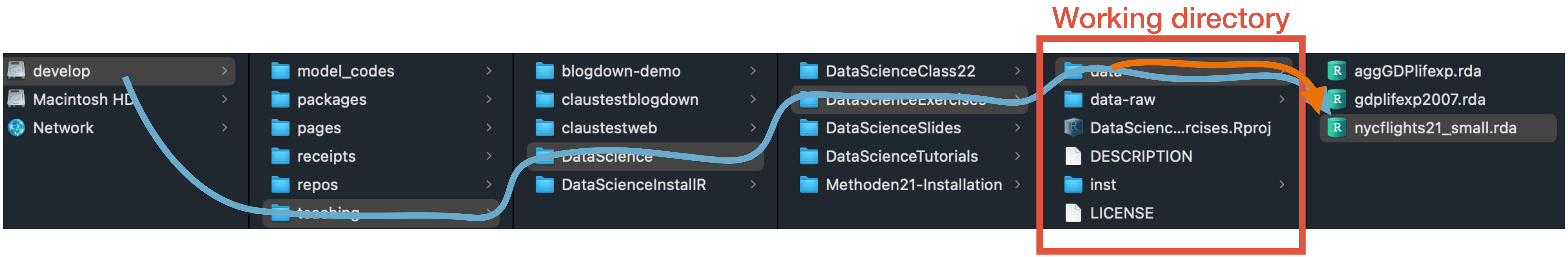
misc

Creating an R project

See the associated tutorial and video for the documentation of the relevant steps (slides focus on selected background concepts only)

Paths and the here-package

- Two ways to you tell your computer where a certain file is located:
  - Via an **absolute path**: description starts at the root directory 🌲
  - Via a **relative path**: description starts at your current position in the file system



- Assuming we are 'located' in the folder `DataScienceExercises` and want to point to the file `nycflights21_small.rda`:
  - `"/Volumes/develop/teaching/DataScience/DataScienceExercises/data/nycflights21_small.rda"`
  - `"data/nycflights21_small.rda"`

Relative paths and setwd()

- The relative path seems nicer…
  - Its shorter 😇 and you can share code without forcing others to adjust the path
- Problem: how to set our location to the directory `DataScienceExercises`?
- One option: use `setwd()` to provide the absolute path to `DataScienceExercises` as an argument:

  -
    ```
    setwd("/Volumes/develop/teaching/
            DataScience/DataScienceExercises")
    ```
  - Then we can use `"data/nycflights21_small.rda"`
- Many people put `setwd()` at the top of their scripts
  - **BUT YOU MUST NEVER EVER DO THIS!!!!!!!!!!!!!!!!!!!!!!!!!!!!**

Why setwd() is evil and not to be used

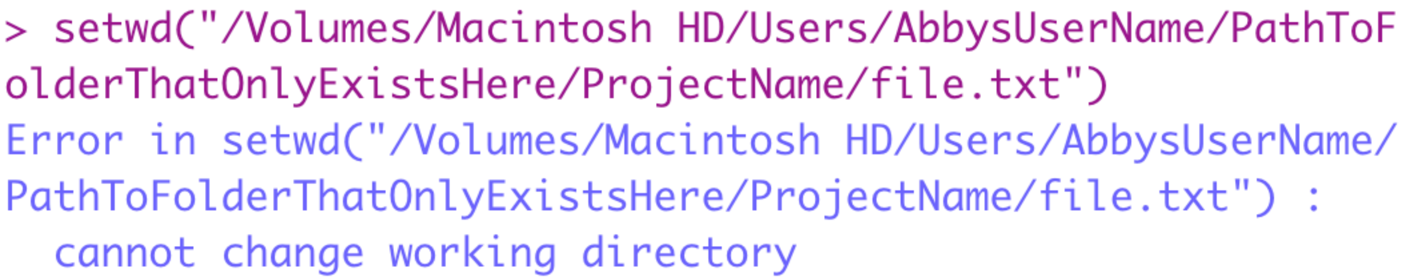
- You should **never ever** use `setwd()` in your scripts
- First, does not help avoiding absolute paths → you have to provide an absolute path to `setwd()` 🤯
- Second, it makes others hate you:

Abby writes amazing_script.R 👩🏼‍💻

```
setwd("/Volumes/Macintosh HD/Users/AbbysUserName/
          PathToFolderThatOnlyExistsHere/ProjectName”)
data_file <- data.table::fread(“data/file.csv”)
```

Sends file to Ellie 📧

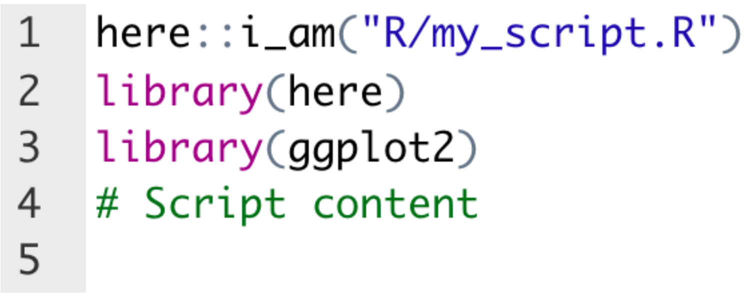
Ellie opens file and executes it 😀



```
> setwd("/Volumes/Macintosh HD/Users/AbbysUserName/PathToF
olderThatOnlyExistsHere/ProjectName/file.txt")
Error in setwd("/Volumes/Macintosh HD/Users/AbbysUserName/
PathToFolderThatOnlyExistsHere/ProjectName/file.txt") :
  cannot change working directory
```

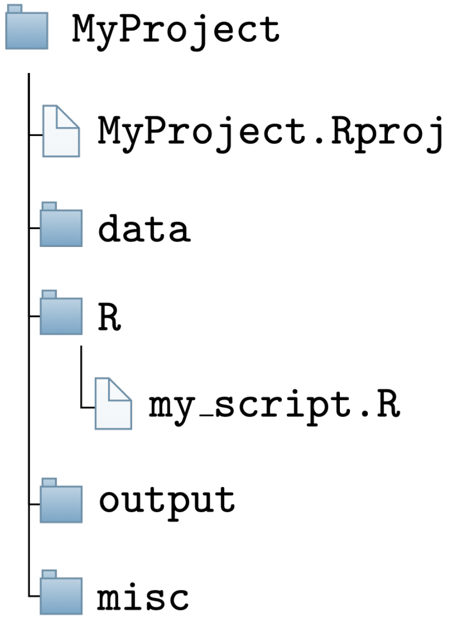
The better alternative to setwd() is `here`

- A very simple solution: the package **here**
- Sets an anchor ⚓ in your project directory
- Then create paths relative to this anchor using **`here::here()`**
  - Will always work on every machine
- Always put **`here::i_am()`** into the first line of your scripts
  - Provide the location of the script relative to the project root
- Then only provide paths relative to this root using **`here::here()`**



```r
here::i_am("R/my_script.R")
library(here)
library(ggplot2)
# Script content

```



MyProject

MyProject.Rproj

data

R

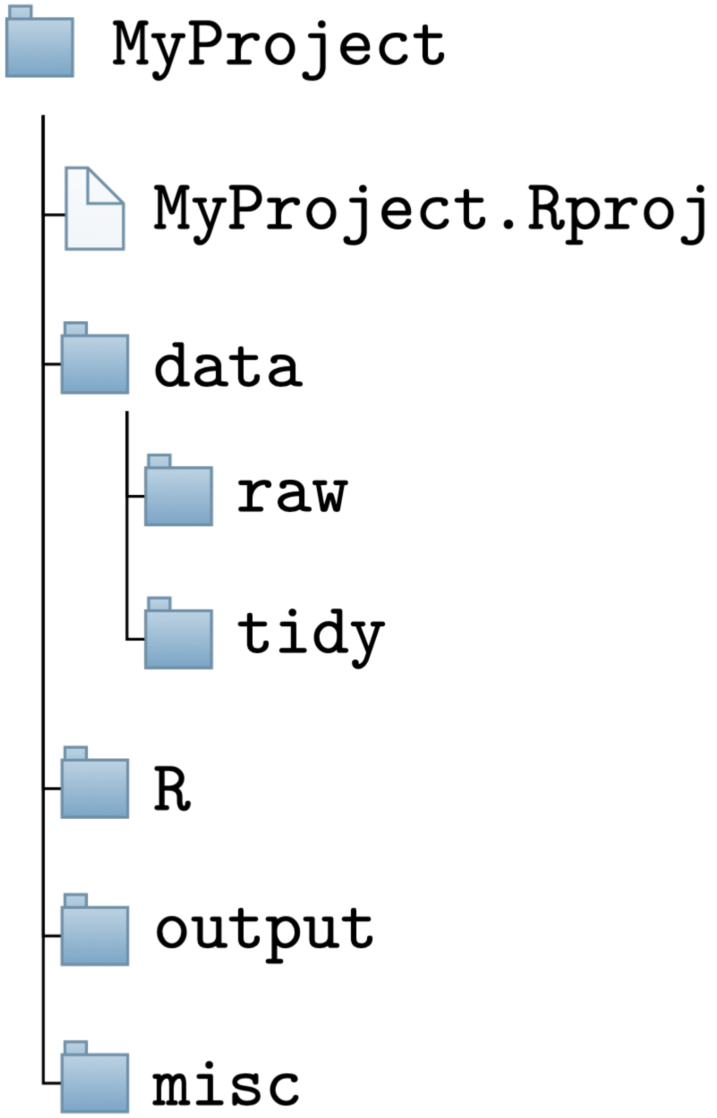
my_script.R

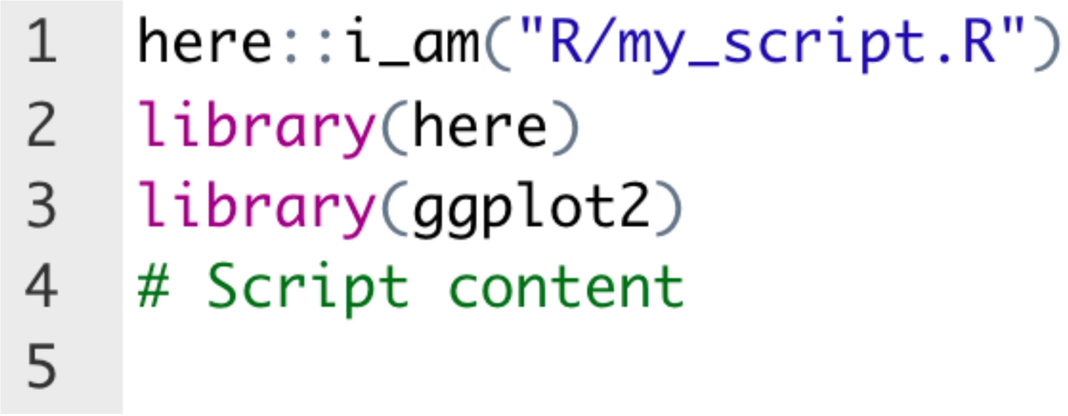
output

misc

Your turn

- Create a new R-Project on your computer
- Create all the required folders
- Write an R script, put it into the right directory, and make it usable for the `here`-package
- Check out what the function `here::here()` returns and experiment with its use

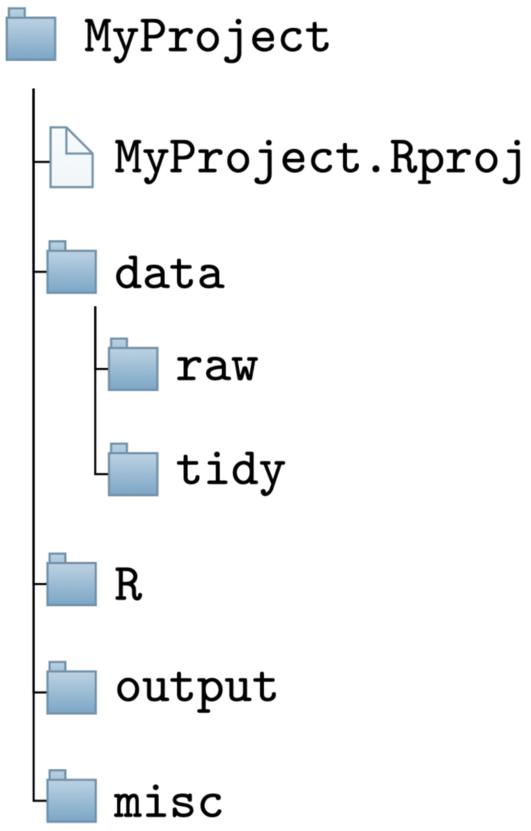




```r
here::i_am("R/my_script.R")
library(here)
library(ggplot2)
# Script content

```

Summary and outlook



MyProject
MyProject.Rproj
data
raw
tidy
R
output
misc

- We now know how to organise our **working directory**
- Important difference between **absolute and relative paths**
- Challenge of using code on different machines can be addressed using the **here package**
  - Better alternative than using `setwd()`
- Project management essential but often under-appreciated!

- Further topics:
  - Using a version control system (such as Git)
  - Using virtual programming environments (e.g. via the package `renv`)