Data Wrangling & Exploration with the Tidyverse in R

Introduction

Johannes Breuer
Thomas Ebel
Stefan Müller

2019-05-15

About this course

In this course you will learn how to **wrangle and explore data in R using packages from the tidyverse**. At the end of the course, you should be be familiar with the concepts of **tidy data** and the **pipe operator**, able to **import, transform, and explore** data with the tidyverse, and comfortable rewriting base R syntax in **tidyverse syntax**.

About us

Johannes Breuer

- senior researcher in the team Data Linking & Data Security at the GESIS Data Archive
    - main area: data linking
- Ph.D. in psychology, University of Cologne
- previously worked in several research projects investigating the use and effects of digital media (Cologne, Hohenheim, Münster, Tübingen)
- other research interests:
    - methods of media (effects) research
    - data management
    - open science

johannes.breuer@gesis.org, @MattEagle09

About us

Stefan Müller

- senior researcher in the team Data Linking & Data Security at the GESIS Data Archive
    - main area: geo data
- studied Sociology, Ethnology and Philosophy at the University of Cologne (M.A.)
- previously worked in the area of data ingest and curator of the data repository datorium and two DFG research projects on geo data

stefan.mueller@gesis.org

About you

- What's your name?
- Where do you work?
- What do you work on?
- What are your experiences with R and the tidyverse?
- What do you want to use the tidyverse for?

Prerequisites for this course

- Working version of R and RStudio
- Some basic knowledge of R
- The following packages: `tidyverse`, `sf`

Install the packages with

```
install.packages(c("tidyverse", "sf"))
```

Preliminaries

- Feel free to ask questions at any time
- We want to make this workshop as interactive as possible
- Slides and other materials are available at

  `https://github.com/jobreu/tidyverse-workshop-gesis-2019`

Course schedule

Wednesday, May 15th, 2019

When?	What?
10:00 - 11:00	Introduction
11:00 - 11:30	*Coffee break*
11:30 - 13:00	Tidy data
13:00 - 14:00	*Lunch break*
14:00 - 15:00	Importing data
15:00 - 15:30	*Coffee break*
15:30 - 16:30	Tidyverse terminology: tibbles & pipes
16:30 - 18:00	Data cleaning
18:00 - 20:00	*Wine & cheese*

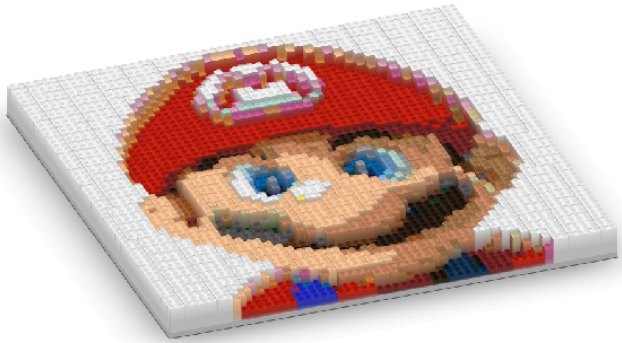
Course schedule

Thursday, May 16th, 2019

When?	What?
09:00 - 10:30	Data transformation
10:30 - 11:00	*Coffee break*
11:00 - 12:00	Relational data
12:00 - 13:00	*Lunch break*
13:00 - 14:00	Data exploration I: Summary statistics
14:00 - 15:30	Data exploration II: ggplot2 basics
15:30 - 16:00	Outlook

What to expect

- This is not a statistics workshop!
  - Plumbing (skills for wrangling & exploring) instead of engineering (knowledge about stats)



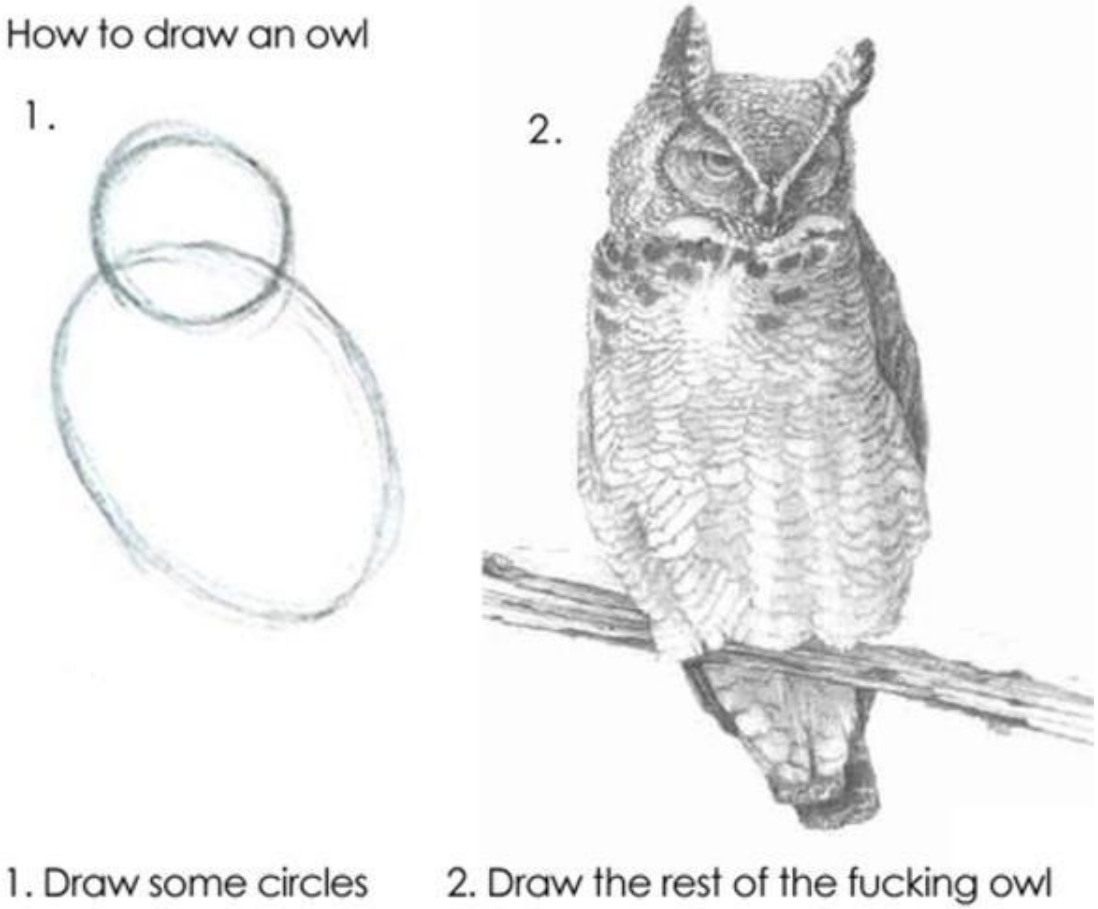
- You will become a super (data) plumber who knows how to use pipes

Why do we focus on wrangling & exploring?

The (in)famous **80/20-rule**: 80% wrangling, 20% analysis

> Given almost every data task, you'll almost certainly need to clean your data, visualize it, and do some exploratory data analysis Moreover, they are also important as you move into more advanced topics. Do you want to start doing machine learning, artificial intelligence, and deep learning? You had better know how to clean and explore a dataset. If you can't, you'll basically be lost (Sharp Sight Labs, 2017).

What we want to avoid



Source: https://bit.ly/2Xhz81a

What is the tidyverse?

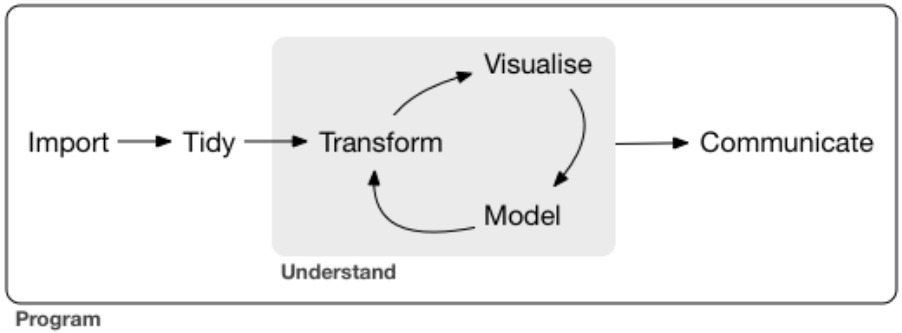
> The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures (Tidyverse website).
>
> The tidyverse is a coherent system of packages for data manipulation, exploration and visualization that share a common design philosophy (Rickert, 2017).

Benefits of the tidyverse

Most of the things we are going to show you can also be achieved with base R. However, the syntax for this is typically (more) verbose and not intuitive and, hence, difficult to learn, remember, and read (plus many tidyverse operations are faster than their base R equivalents).

Tidyverse syntax is designed to increase **human-readability**. This makes it especially **attractive for R novices** as it can facilitate the experience of **self-efficacy** (see Robinson, 2017). The tidyverse also aims for **consistency** (e.g., data frame as first argument and output) and uses **smarter defaults** (e.g., `stringsAsFactors = FALSE` & no partial matching of data frame and column names).

Workflow

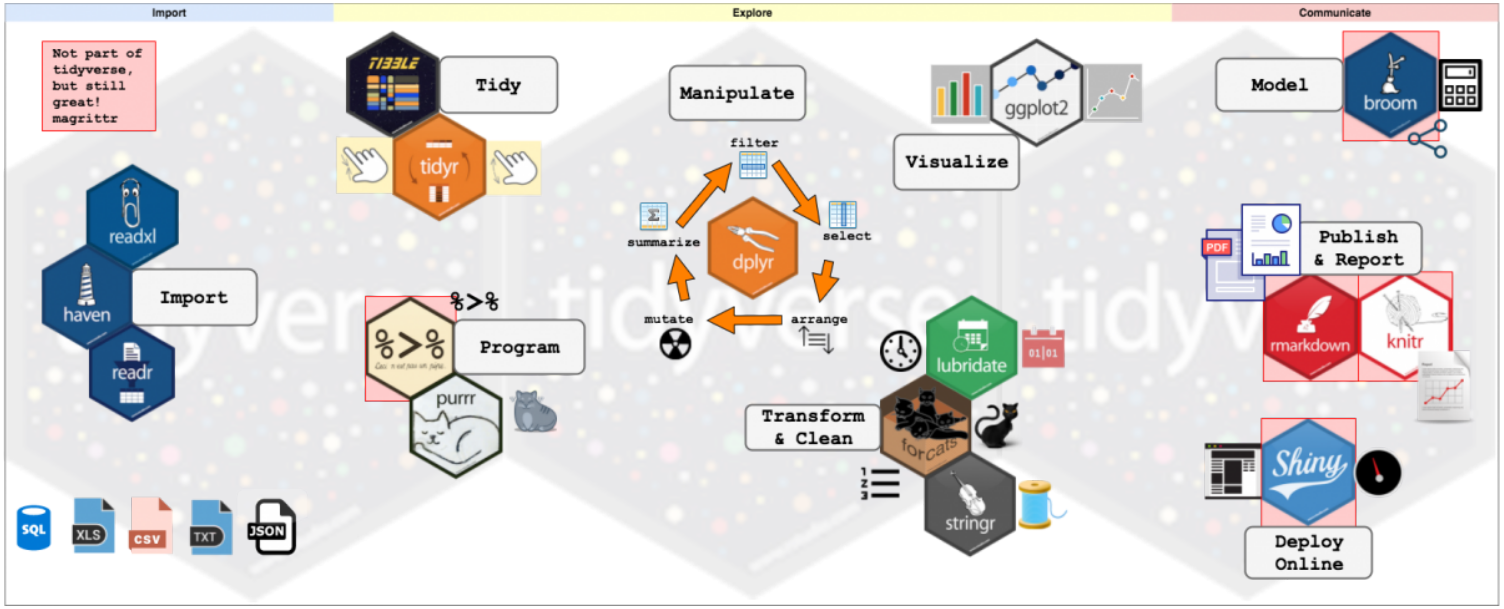


Source: http://r4ds.had.co.nz/

- **Import**: read in data in different formats (e.g., .csv, .xls, .sav, .dta)
- **Tidy**: clean data (1 row = 1 case, 1 column = 1 variable), rename & recode variables, etc.
- **Transform**: prepare data for analysis (e.g., by aggregating and/or filtering)
- **Visualize**: explore/analyze data through informative plots

- **Model**: analyze the data by creating models (e.g, linear regression model)
- **Communicate**: present the results (to others)

Tidyverse workflow



Source: http://www.storybench.org/getting-started-with-tidyverse-in-r/

Lift-off into the tidyverse 🚀

install all tidyverse packages[1]

```
install.packages("tidyverse")
```

load core tidyverse packages[2]

```
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------------
## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
## -- Conflicts ------------------------------------------------------ t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

[1] For the full list of tidyverse packages see https://www.tidyverse.org/packages/.
[2] To save time and reduce namespace conflicts it often makes sense to install and load the tidyverse packages individually.

Coding in style 😎

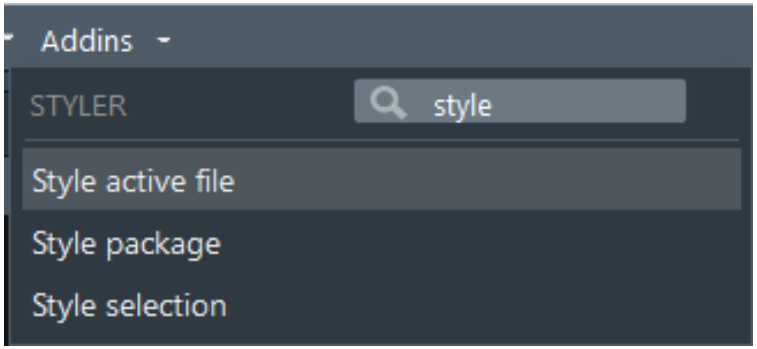
The tidyverse style guide by Hadley Wickham

`styler` package (incl. RStudio add-in)

```
install.packages("styler")
library(styler)
```

From the package documentation:

- `style_file()` styles .R and/or .Rmd files.
- `style_dir()` styles all .R and/or .Rmd files in a directory.



Time for