

# Fundamentals of R

**Exchanging with the community**

Henrique Sposito and Livio Silva-Muller

**GENEVA  
GRADUATE  
INSTITUTE**

INSTITUT DE HAUTES  
ÉTUDES INTERNATIONALES  
ET DU DÉVELOPPEMENT  
GRADUATE INSTITUTE  
OF INTERNATIONAL AND  
DEVELOPMENT STUDIES

# Housekeeping

- All the homework answers are now available online on Moodle
- If you have any specific questions about the course, R code used, or would like to get some feedback on your homework assignments, please send us an email!



# Common Homework issues with R markdown

- Latex engines (for PDF rendering in Windows)
- The "yaml" header can sometimes be annoying about spaces and unrecognized symbols...
- The setup chunk (load data, packages, and options for the whole document)

```
```{r setup, include=FALSE}
library(dplyr)
library(ggplot2)
library(tidyr)
library(scales)
library(readr)
io_income_rs <- read_csv("io_income_rs.csv")
options(scipen = 999)
```
```

*Remember data and RMD file should ideally be in the same folder!*

- More on PDF knitting, figures, and tables [here](#)
- Other specific R Markdown issues you had?

# What did we learn last week?

Conceptually:

- Gestalt principles
- Tufte's law
- One plot, one story

Practically:

```
ggplot(data, mapping = aes()) # Do you  
remember mapping vs setting aesthetics?  
# geom_*( ) # What does the geom functions do?  
geom_line() # layering lines  
geom_points() # layering points  
geom_col() # layering columns  
geom_boxplot() # layering box plots  
geom_smooth() # layering line over patterns  
labs() # modifies plot labels  
theme() # customize plot components  
# theme_*( ) # family of preset themes
```

What about scales and normalization?

# Roadmap

## Lecture:

- R Markdown (a bit more)
- Minimal reproducible examples
- The ethics of data
- Where to go from here?

## Case Study:

*American Rappers*



# R Markdown, a one-stop shop (rmd 1/4)

- Write your text, execute code, and produce high quality (reproducible) reports in **one place!**
- Made for R, free, and with lots of support
- Do not underestimate R Markdown's potential or flexibility cause it is ugly

# R Markdown options (rmd 2/4)

- Code chunk options:
  - echo
  - eval
  - include
  - warning
  - message
  - fig.cap
- Global options: apply to every code chunk if used in setup

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
```

# From R Markdown with care (rmd 3/4)

- Articles
- Presentations
- Books
- Websites
- Interactive applications
- And more cool stuff ...

Write your IHEID thesis with [IHEIDDOWN](#)

# Is Quarto the future? (rmd 4/4)

- Quarto is handy for users of multiple programming languages, but is it really necessary for R users?
- Is Quarto just a fancier R Markdown?
- R Studio's, or shall I say "[Posit](#)", attempt to become a more flexible platform
- If you know a bit of Rmarkdown, getting started with Quarto is relatively easy, see this [video](#)

**Attention, Quarto requires the latest version of R Studio...**

# Github, worth more than this one mention...

- The worlds biggest code hosting platform for collaborations
- Make your research available online (just like [Livio](#) and [I](#))
- From hosting your websites and applications, to other types of online collaborative [work](#)

*If you are interested on how we work with GitHub to build, store, version, and release R packages, check out the [messydates GitHub site](#). If you are having issues with dates in R, see [how messydates can help you](#).*

# Exchanging with the community (reprex 1/8)

- Reproducibility of an error is the key
- Don't forget about the minimal part
- The better the question, the better the answer
- Writing a reprex often helps you debug!

# The question (reprex 2/8)

"Dear R friends,

I am having issues getting the mean for a variable in my dataset in R.

As you see, the code returns "NA", even though I am removing NAs to calculate the mean.

Below, you will find a sample of my data, code, and system info so that you can reproduce the issue.

Could you please help me?

Thank you in advance for your time and consideration!"

# The data (reprex 3/8)

A small sample of the data...

```
my.df <- data.frame(col1 = sample(c(1,2), 5, replace = TRUE),
                      col2 = as.factor(sample(5)),
                      col3 = letters[1:5],
                      col4 = sample(c(TRUE, FALSE), 5, replace = TRUE))
my.df
```

```
#>   col1 col2 col3  col4
#> 1    2    2     a  TRUE
#> 2    2    5     b FALSE
#> 3    1    1     c  TRUE
#> 4    2    3     d FALSE
#> 5    1    4     e FALSE
```

Using R datasets (e.g. `USArrests`, `mtcars`, `iris`) is also very handy!

*set.seed()* orders random processes, see `?set.seed()`

# The error (reprex 4/8)

If you are doing this manually, copy the exact error somewhere in the post!

```
mean(my.df$col1)
```

```
#> [1] 1.6
```

```
mean(my.df$col2, na.rm = TRUE)
```

```
#> [1] NA
```

```
.
```

# The system (reprex 5/8)

Sometimes system info can be important...

```
search()  
sessionInfo()  
R.Version()  
R.Version()$version.string
```

# The post (reprex 6/8)

*Dear R friends, I am having issues getting the mean for a variable in my dataset in R. As you see, the code returns "NA", even though I am removing NAs to calculate the mean. Below, you will find a sample of my data, code, and system info so that you can reproduce the issue. Could you please help me? Thank you in advance for your time and consideration!*

```
set.seed(1234)
```

```
my.df <- data.frame(col1 = sample(c(1,2), 5, replace = TRUE), col2 = as.factor(sample(5)), col3 =  
letters[1:5], col4 = sample(c(TRUE, FALSE), 5, replace = TRUE))
```

```
mean(my.df$col2, na.rm = TRUE)
```

```
'#> [1] NA'
```

```
'#> Warning in mean.default(my.df$col2, na.rm = TRUE): argument is not numeric or logical: returning  
NA'
```

```
'#> R version: 4.1.2 (2021-11-01) / Loaded packages: dplyr and tidyR'
```

# Reprex, the package (reprex 7/8)

You do not need any package to make a minimal reproducible example, but the [reprex](#) package can help!

- It may not look like much but, in practice, allows you to copy and paste the code directly to post so that the error can be reproduced by others!

---

Reprex code

Reprex output

```
#install.packages("reprex")
library(reprex)
reprex::reprex({
  R.Version()$version.string
  search()
  set.seed(1234)
  my.df <- data.frame(col1 = sample(c(1,2), 5, replace = TRUE),
                        col2 = as.factor(sample(5)), col3 = letters[1:5],
                        col4 = sample(c(TRUE, FALSE), 5, replace = TRUE))
  mean(my.df$col2, na.rm = TRUE)
})
```

# Reprex for R Markdown, more or less the same (reprex 8/8)

*Rules-of-thumb:*

- If issue is within a code chunk (code issue), make a normal reprex
- If issue is with knitting, copy the document (yaml) header and the code chunk(s) that fails to knit or displays the issue
- The rest is the same but remember to try and keep it minimal...
- If you want to improve the knitting itself, post a (minimal) version of the RMD file with what you want to improve and the knitted PDF/HTML document

# Stack overflow, one last time

- You can find virtually anything about an [R issue](#)
- R community can sometimes be annoying, so ask well
- If anything, remember the golden rule of copy and paste (i.e. others need to be able to copy and paste your code and reproduce the error)

[R bloggers](#), [Community R](#), [Twitter](#), and even [Reddit](#), if that is your thing

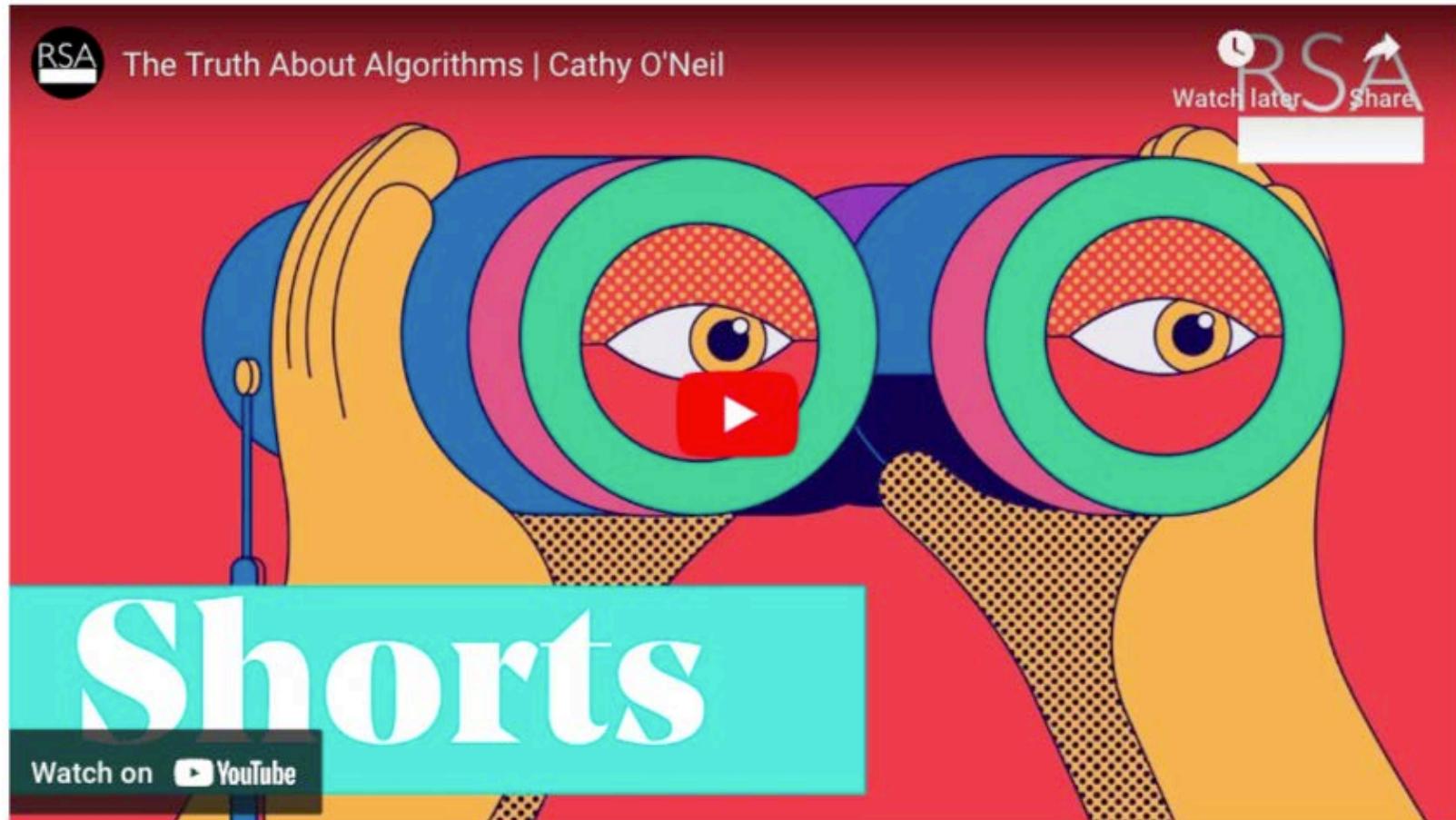
# Individual data ethics (Ethics 1/2)

This course, hopefully, has provided you with the tools to keep on learning R on your own but...

- Consent (also for data)
- Privacy vs. transparency (a thin-line)
- Fairness (data is interpreted but...)
- Reproducibility



# The ethics of big data (Ethics 2/2)



# Where to go from here (future 1/3)

- As everything else in R, there are lots of free material available online to keep on learning!
- The steepest part of the R learning curve should be behind you (if you took this course serious).
- Keep on taking advantage of the slides and the syllabus from now onwards
- Finally, the big secret to learning R is:

**There is no secret... you will have to try, try, and try some more...**



Hopefully, this course has made it easier for you to understand R code and errors, think through data, and find help online for your specific needs!

# Courses (future 2/3)

- Datacamp is not free, but it is good!
- For more on data wrangling check out Kelsey Gonzalez' course; for more on data science check Dave Langer's course; for more on modelling and regression see this Stanford course
- Several free courses available from various universities as Harvard, for example, and Coursera which are free to audit
- Learn R on your own within R with the "swirl" package

```
install.packages('swirl')
library(swirl)
swirl()
```

# Summer methods schools (future 3/3)

- Can be expensive and often very specific, but worth it if you can afford (or get a scholarship)
- [GSERM](#), we cannot recommend enough!
- A lot more out there in person and online (see [ECPR](#) and the [USI](#) for instance)
- Last but not least, the [DS3](#) is a summer online with several R courses pre-recorded online and available for free!

# You made it!

