Essential Research Toolkit for the Humanities

Week 3: Looking at data

Anna Pryslopska

April 25, 2022

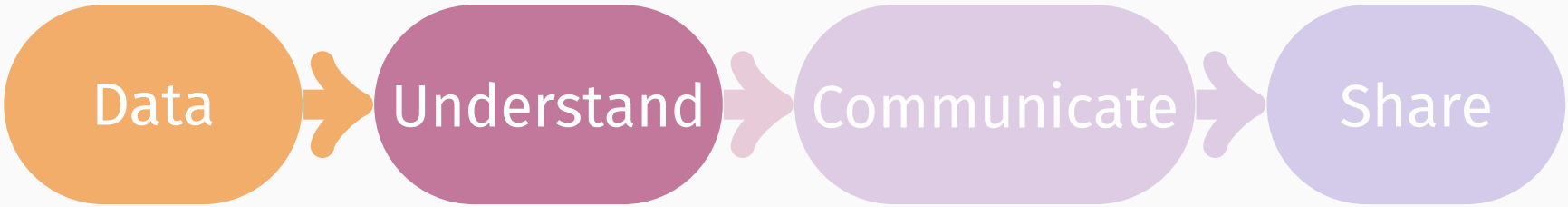
Psycholinguistics and Cognitive Modeling Lab

Questions?

Table of contents

Where are we this week?

Workflow



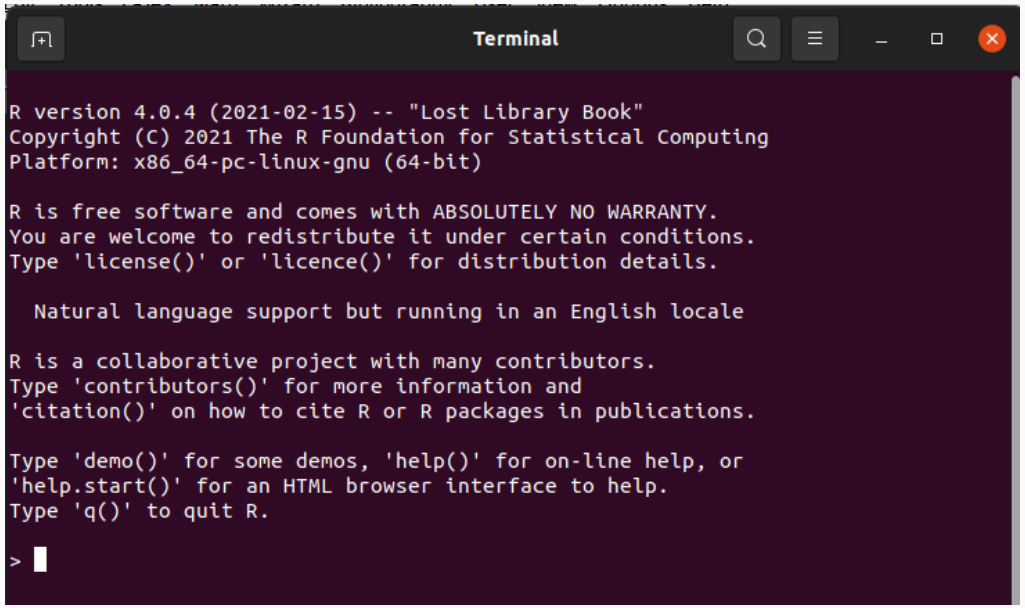
import, clean, transform, visualize, model

document, create clean and beautiful reports

connect, collaborate, backup

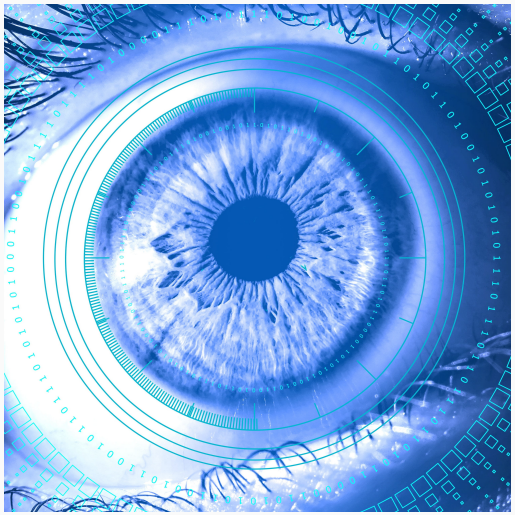
R and RStudio

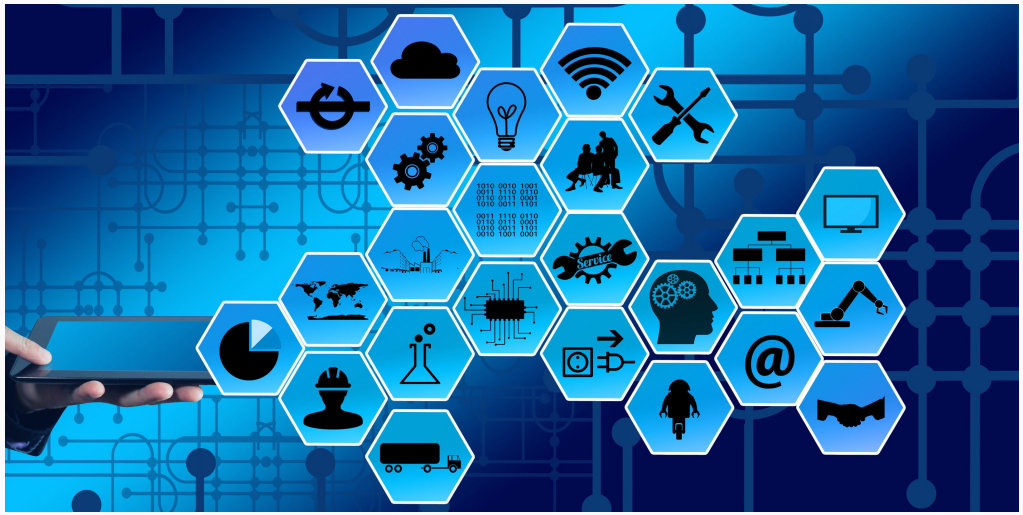
R



- language & environment for statistical computing & graphics
- used by data miners & statisticians for data analysis & developing statistical software
- developed by R Core Team & the R Foundation for Statistical Computing `https://www.r-project.org/`
- base + packages

R: Why bother?

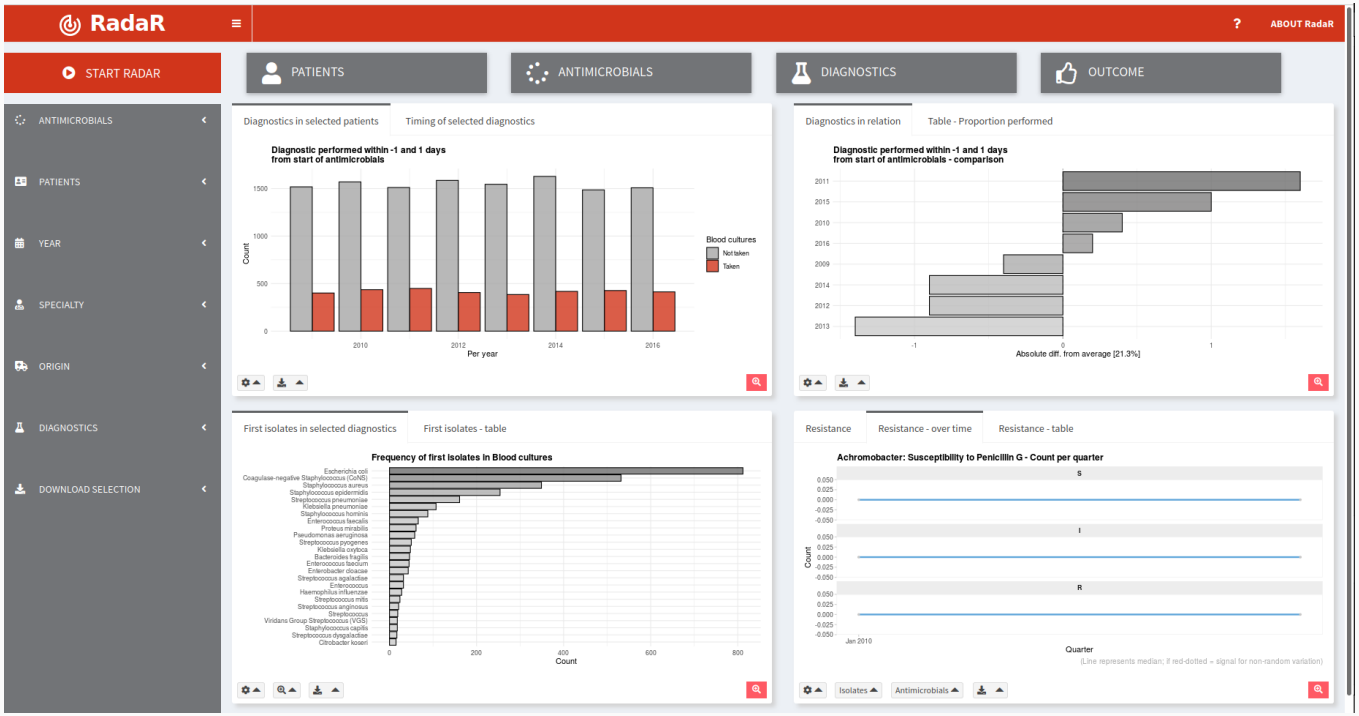




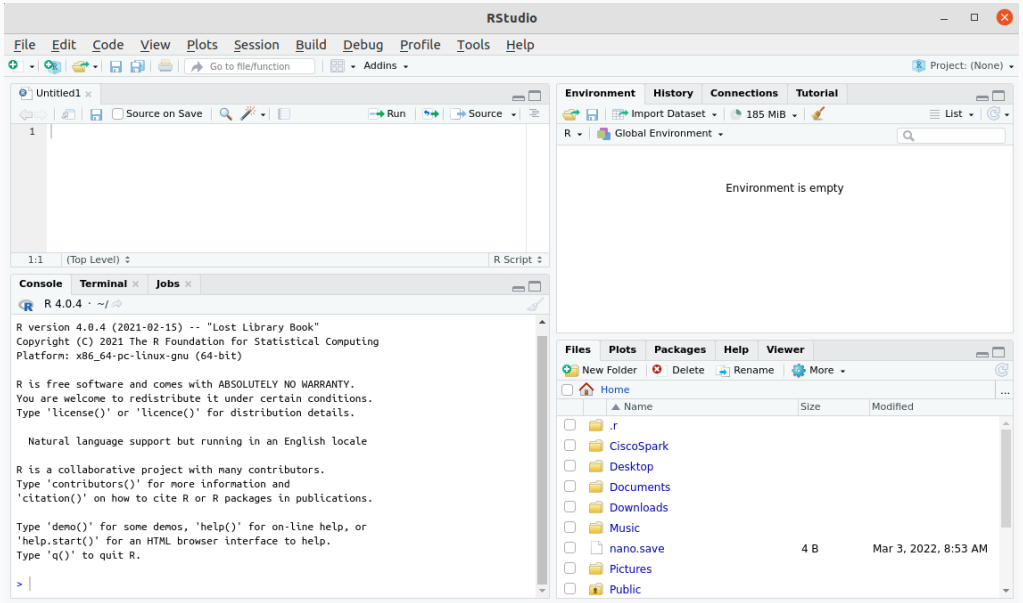
GPL Free Software

Free as in Freedom

R: Why bother?


RadaR
ABOUT RadaR
START RADAR
PATIENTS
ANTIMICROBIALS
DIAGNOSTICS
OUTCOME
ANTIMICROBIALS
PATIENTS
YEAR
SPECIALTY
ORIGIN
DIAGNOSTICS
DOWNLOAD SELECTION
Diagnostics in selected patients
Timing of selected diagnostics
Diagnostic performed within -1 and 1 days from start of antimicrobials
Count
Per year
Blood cultures
Not taken
Taken
Diagnostics in relation
Table - Proportion performed
Diagnostic performed within -1 and 1 days from start of antimicrobials - comparison
Absolute diff. from average [21.3%]
First isolates in selected diagnostics
First isolates - table
Frequency of first isolates in Blood cultures
Escherichia coli
Coagulase-negative Staphylococcus (CoNS)
Staphylococcus aureus
Staphylococcus epidermidis
Streptococcus pneumoniae
Klebsiella pneumoniae
Staphylococcus hominis
Enterococcus faecalis
Proteus mirabilis
Pseudomonas aeruginosa
Streptococcus pyogenes
Klebsiella oxytoca
Bacteroides fragilis
Enterococcus faecium
Enterobacter cloacae
Streptococcus agalactiae
Enterococcus
Haemophilus influenzae
Streptococcus mitis
Streptococcus anginosus
Streptococcus
Viridans Group Streptococcus (VGS)
Staphylococcus capitis
Streptococcus dysgalactiae
Citrobacter koseri
Count
Resistance
Resistance - over time
Resistance - table
Achromobacter: Susceptibility to Penicillin G - Count per quarter
S
I
R
Count
Quarter
Jan 2010
(Line represents median; if red-dotted = signal for non-random variation)
Isolates
Antimicrobials

RStudio



- Integrated Development Environment (=IDE) for R (& Python)
- free, open source
- “friendlier” than pure R
- developed by RStudio, who develop, maintain, and promote many awesome packages (e.g. tidyverse, RMarkdown, knitr)
- source, console, environment, plots, packages, help

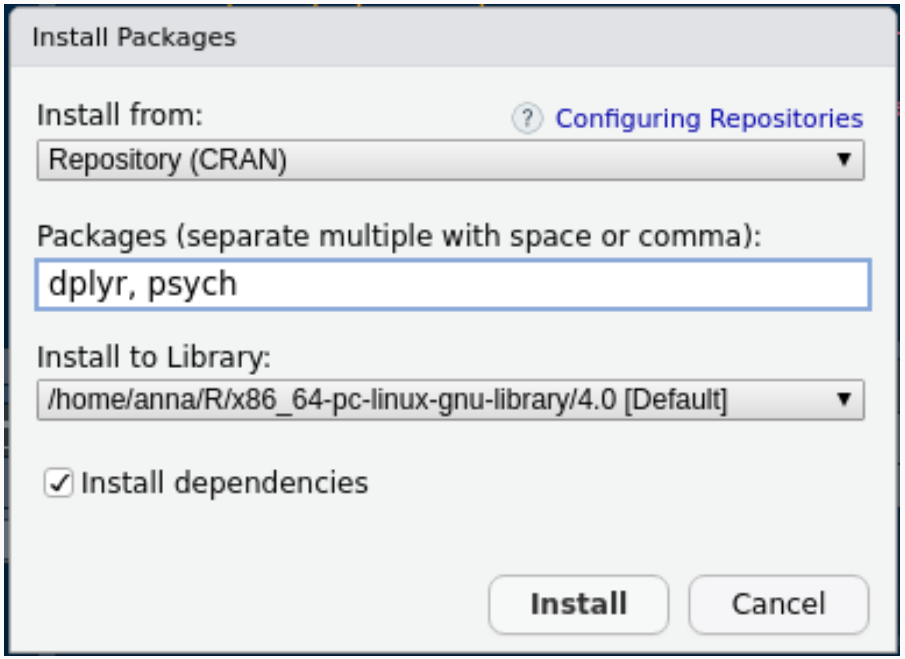
R packages

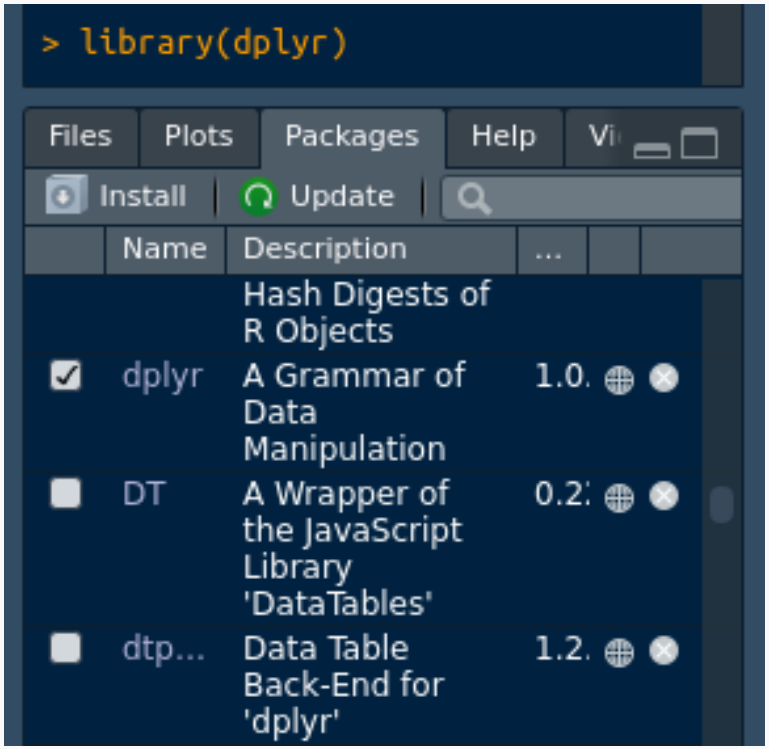
Installing, loading, checking

`install.packages("NAME","NAME2")` install from repository

`library(NAME)` load into workspace (=activate)

`sessionInfo()` collect information about current R session



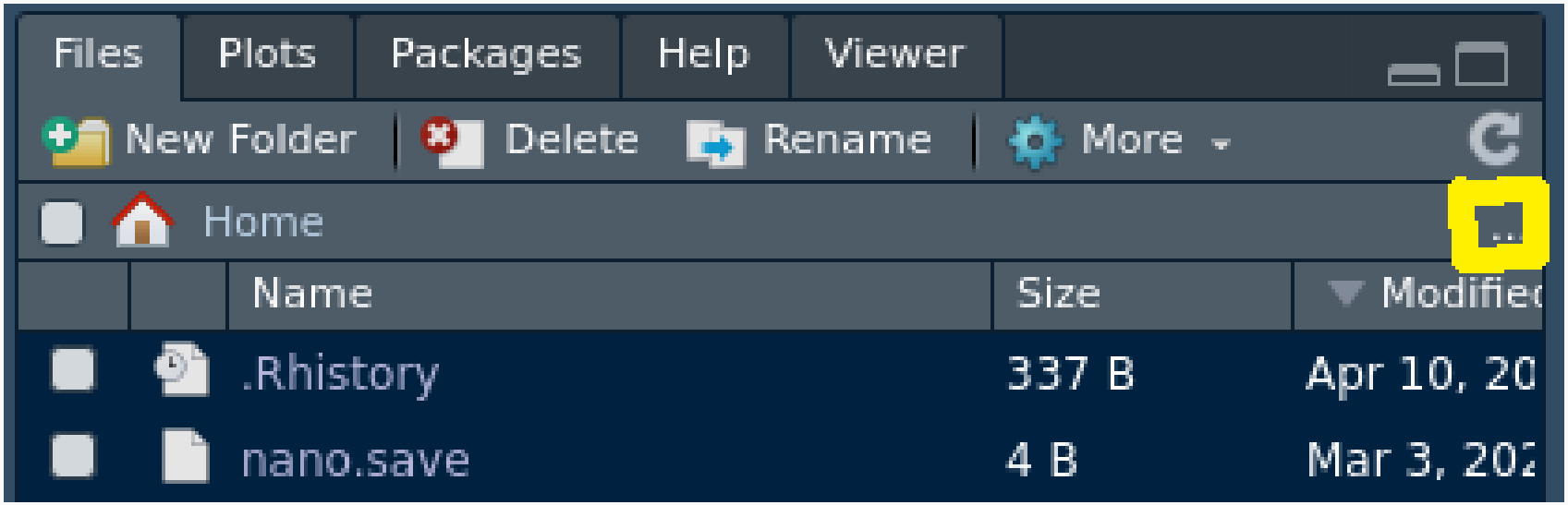


Working with RStudio

Make a home for your project

Create a new folder and set it as the working directory

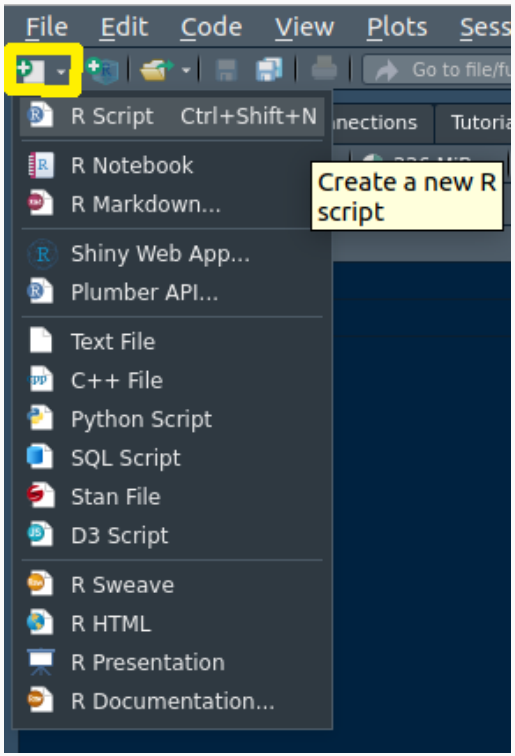
- where R will look for files
- where R will save visible and hidden files
- where R will automatically load files from



`setwd("PATH")` set working directory

`getwd()` check working directory

Don’t You Forget About Me



	Windows	Mac
work with scripts	`CTRL+SHIFT+N`	`CMD+SHIFT+N`
leave comments #	`CTRL+SHIFT+C`	`CMD+SHIFT+C`
make sections	`CTRL+SHIFT+R`	`CMD+SHIFT+R`
use autocompletion	`TAB`	`TAB`
run selected code	`CTRL+ENTER`	`CMD+RETURN`

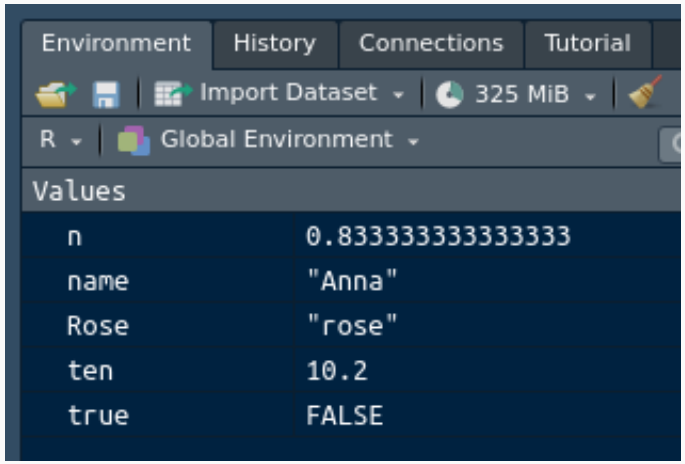
Assignment

Saves information for later, keeps it in the environment, but **overwrites existing values**.

```
x <- value          ten <- 10.2
value -> x      "rose" -> Rose

x = value          name = "Anna"

x <<- value     true <<- FALSE
value ->> x        13/12 ->> n
```



Questions?

Data types, formats, and encoding

What's your type?

From a statistical point of view:

- nominal — marital status, religion
- ordinal — grades, energy efficiency classes
- interval — IQ, temperature in C and F
- ratio — reaction times, population

From an empirical point of view:

- reading times
- brain waves
- transcription
- recordings

→ text files (`txt, csv, tsv`)

- acceptablity judgments
- texts
- annotation

The shape of data

From R's point of view:

- logical `TRUE`
- integer `1` or `1L`
- double `1.0`
- complex `1+0i`
- character `"one"`
- double "not a number" `NaN`
- double "infinity" `Inf`
- double "negative infinity" `-Inf`
- logical "missing" value `NA`
- special variable without a type `NULL`

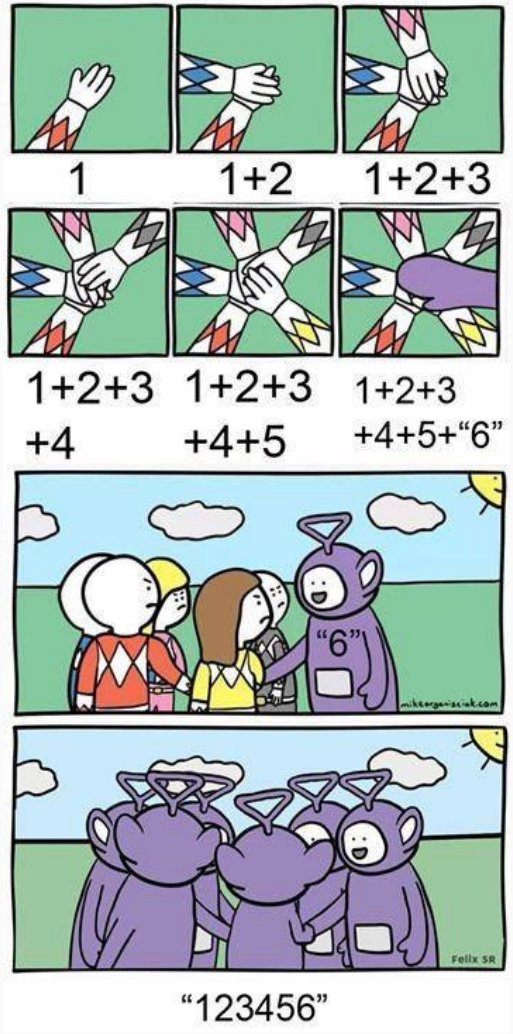
Reap what you sow

Unsure? `typeof(1L)` or `is.numeric(1)`

Certain? `as.character(1)`

`5L + 2 = 7`	integer
`3.7 * 3L = 11.1`	double
`99999.0e-1 - 3.3e+3 = 9999.9 - 3300 = 6699.9`	double
`10 / as.complex(2) = 5+0i`	complex
`as.character(5) / 5` = ⊘	non-numeric argument!

Lost in translation




```
悁獬档獗岘拔涋瘠湯堠泅灯滶浮獵步焠嵅沤⁴
数戢牧翟映牥涋娠敷杲ㄐ楄　畟⁸楦菶獗搠慦物
慬映　整岘ᶜᴱ(敨牵畤挠湩ⁱᵠ獚霾　窠ᵚ菡嫁ᵚ
菡β菡攜ᦈ♠銃菡嫁ᵚ菡攝邃♠ᦈ菡杖閃菡ₐ菡Ψ
鲃菡嫁隃菡ₔ菡攜邃菡ᵢ邃菡墜邃菡δ鼢菡ϒ靴
```

```
Falsches Äœben von Xylophonmusik
quÃ¤lt jeden grÃ¶ßeren Zwerg.
Dis aux filles de faire la fÃªte
Ã<0xa0> l'heure du cinq Ã<0xa0>
sept.
áƒ¥áƒ<0x90>áƒ<0xa0>áƒ–áƒ£áƒšáƒ¡
áƒ'áƒ<0x90>áƒ<0xa0>áƒ“áƒ<0x90>
áƒ¡áƒ®áƒ•áƒ<0x90>
áƒ”áƒœáƒ<0x90>áƒ–áƒ” áƒšáƒ<0x90>áƒ
žáƒ<0x90>áƒ<0xa0>áƒ<0x90>áƒ™áƒ
<0x9d>áƒ‘áƒ–?
```

```
Falsches √úben von Xylophonmusik
qu√§lt jeden gr√∂√üeren Zwerg.
Dis aux filles de faire la f√™te
√† l'heure du cinq √† sept.
ŠÉ•ŠÉźŠÉ†ŠÉóŠÉ£ŠÉöŠÉ°
ŠÉíŠÉźŠÉ†ŠÉďŠÉź ŠÉ°ŠÉģŠÉëŠÉź
ŠÉĒŠÉúŠÉźŠÉĖŠÉĒ
ŠÉöŠÉźŠÉěŠÉźŠÉ†ŠÉźŠÉôŠÉĔŠÉĎŠÉó?
```

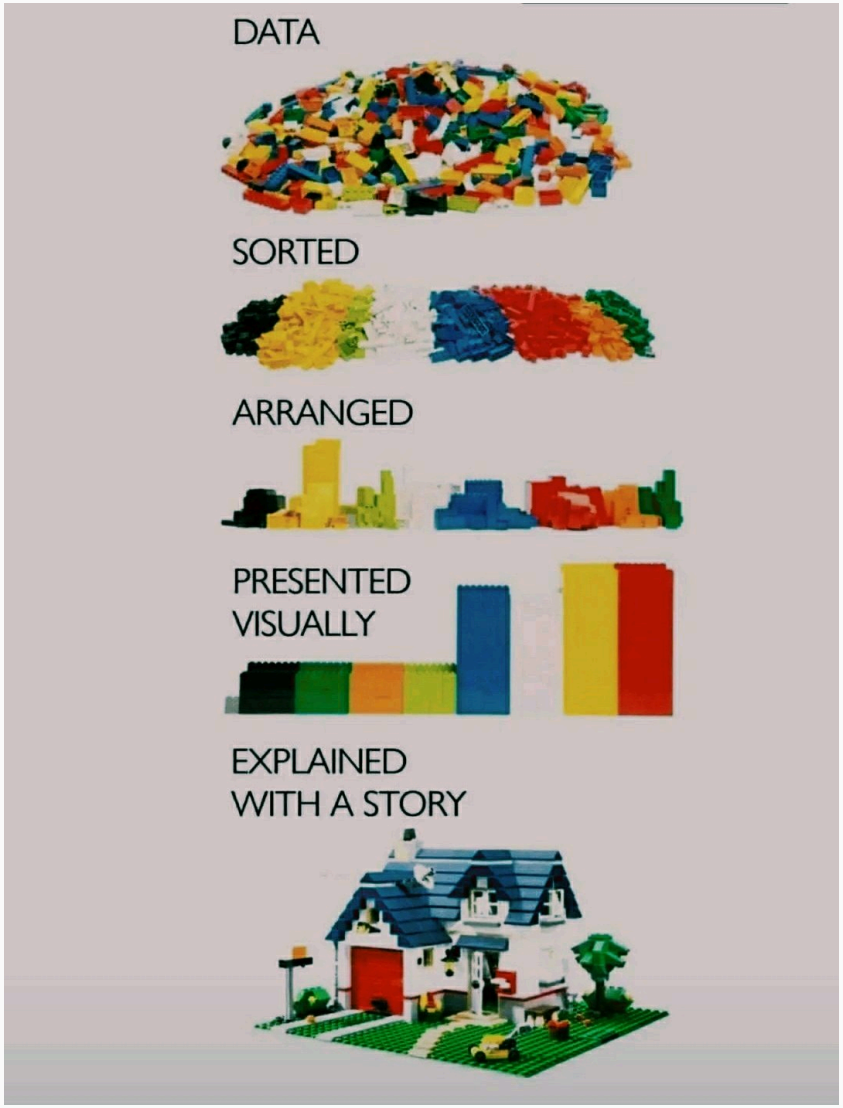
```
Falsches Üben von Xylophonmusik
quält jeden größeren Zwerg.
Dis aux filles de faire la fête à
l'heure du cinq à sept.
ქართულს გარდა სხვა ენაზე
ღაპარაკობთ?
```

→ Be careful and know what you put in

→ Be mindful of the character encoding (when in doubt, UTF-8)

Data

Tell a story


DATA
SORTED
ARRANGED
PRESENTED
VISUALLY
EXPLAINED
WITH A STORY

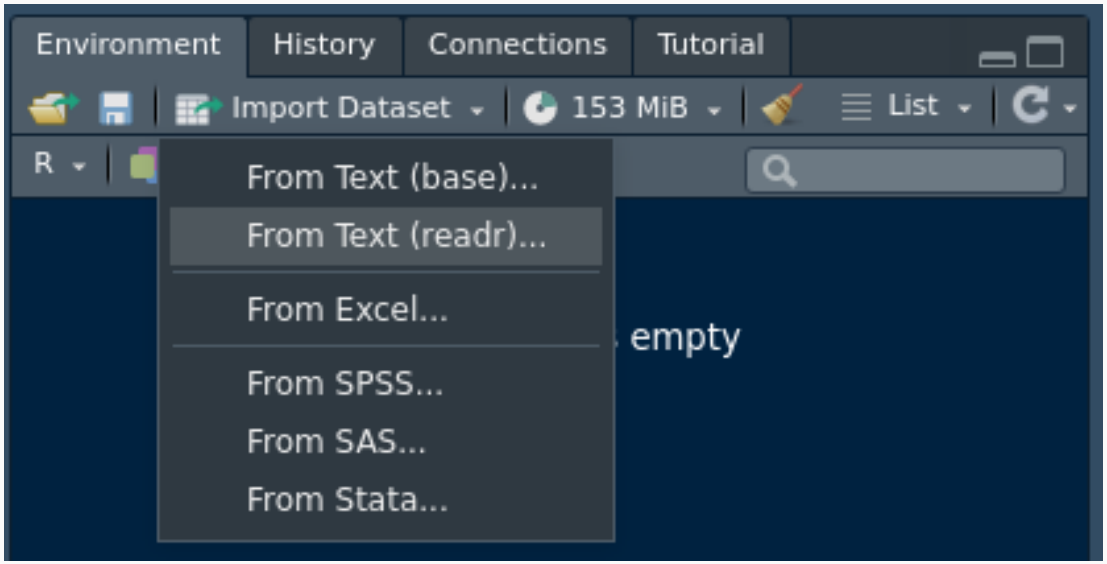
Reading in data

Install and load the package `tidyverse`

Download `moses.csv` from ILIAS and save it to your working directory

Read in the file `moses.csv` and assign it to the name `moses`

Select `Environment > Import Dataset > readr`



or `read_csv("moses.csv")`

Experiment week 1

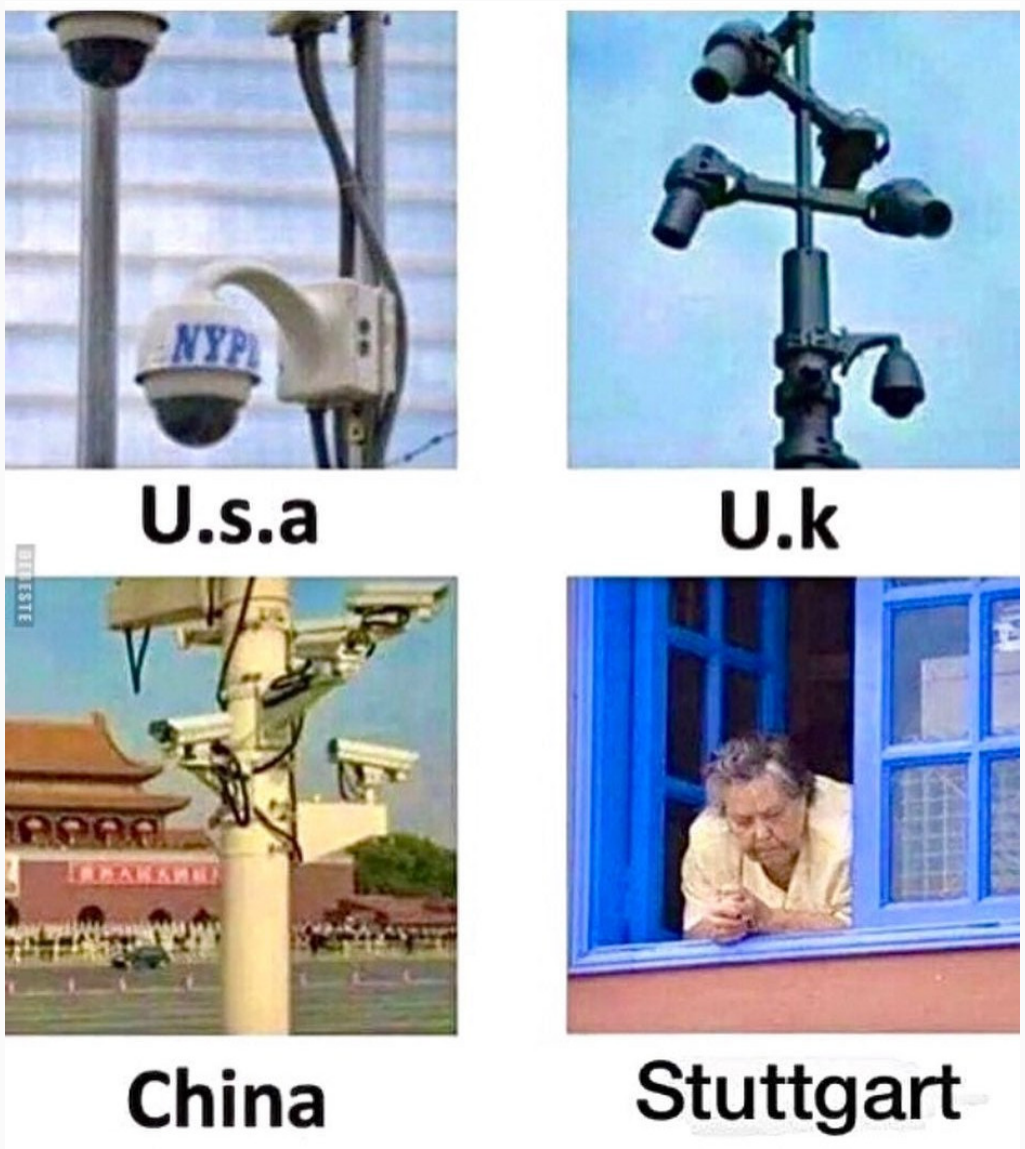
Q: Can a man marry his widow's sister? No 💀💀🪦

Q: According to the Bible, how many animals of each kind did Moses take on the ark? can't answer

Q: How many numbers are in the English alphabet? can't answer

Moses Illusions or *semantic illusions* occur when readers fail to recognize inconsistency in a text even if they were warned and know the correct word (Erickson and Mattson 1981).

Inspecting data



Look at what you did

- in the RStudio window
- `View(moses)`

Wrap-up

Summary

- R and RStudio IDE
- packages (installing and loading)
- working directory
- scripts
- assignment
- data types and encoding
- reading in and inspecting data

Homework assignment due May 2

- Read chapters 4–6 of *R for Data Science*
  `https://r4ds.had.co.nz/`
- Complete assignment 1 ($\rightarrow$ ILIAS)

Next week: R programming basics