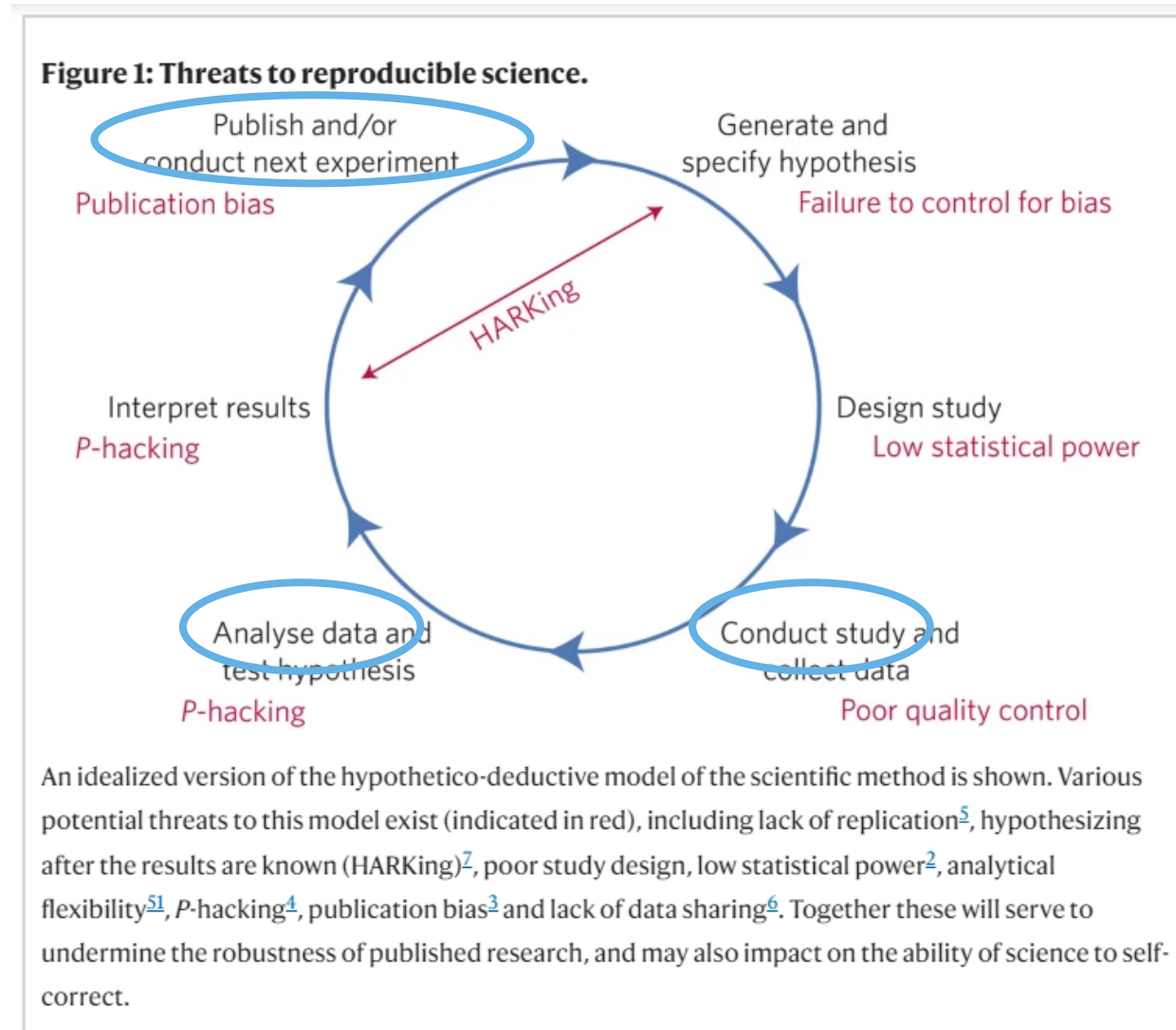# INTRODUCTION TO REPRODUCIBLE RESEARCH

LEARNING OBJECTIVES

- After this lecture, you should be able to:
  - Explain the value of performing reproducible research
  - Perform basic operations in R markdown

# WHY IS REPRODUCIBLE RESEARCH IMPORTANT?

Figure 1: Threats to reproducible science.

An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], P-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Source: Munafò et al. 2017

REMEMBER
HE MODEL REPLICATION

- Berjemo et al. 2017
  - 5 health economic models using publicly available information
  - 1 replication stopped, 2 replicated, 2 not replicated

- McMannus et al. 2019
  - 5 health economic models using publicly available information
  - Comparison with original outcomes: -4.54% to 108.00% for costs; -3.81% to 0.40% for outcomes

HE MODELLING WORKFLOW
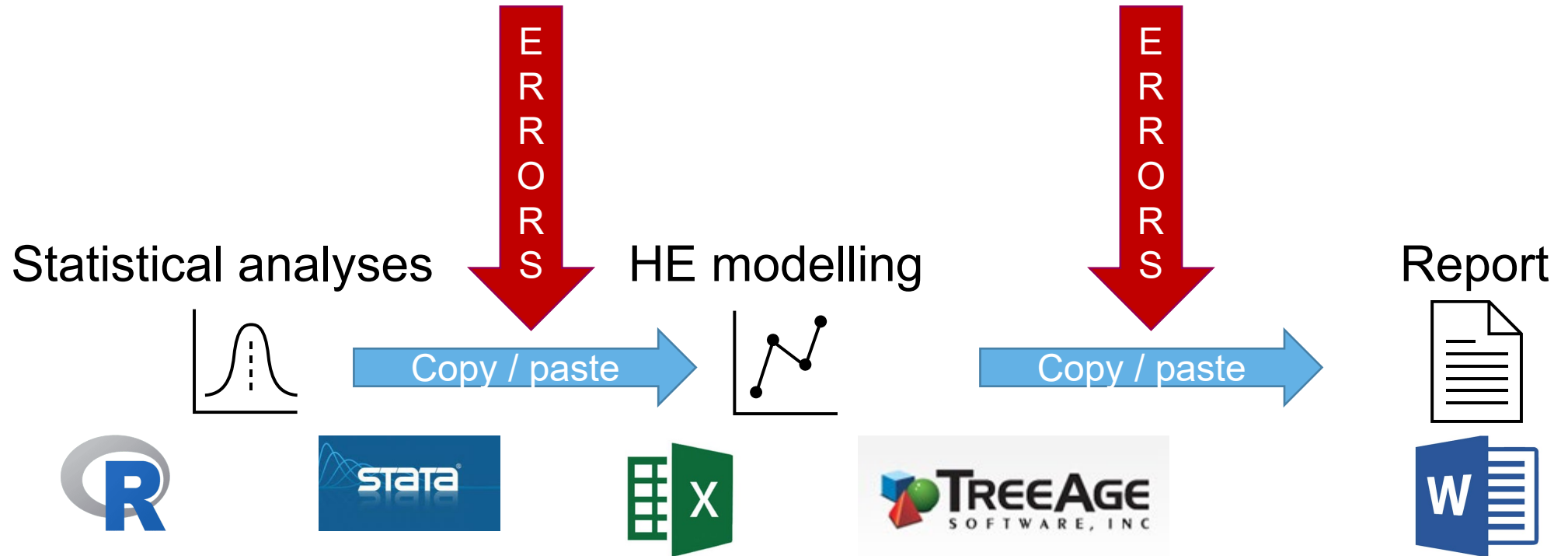
HE MODELLING WORKFLOW IDEALLY
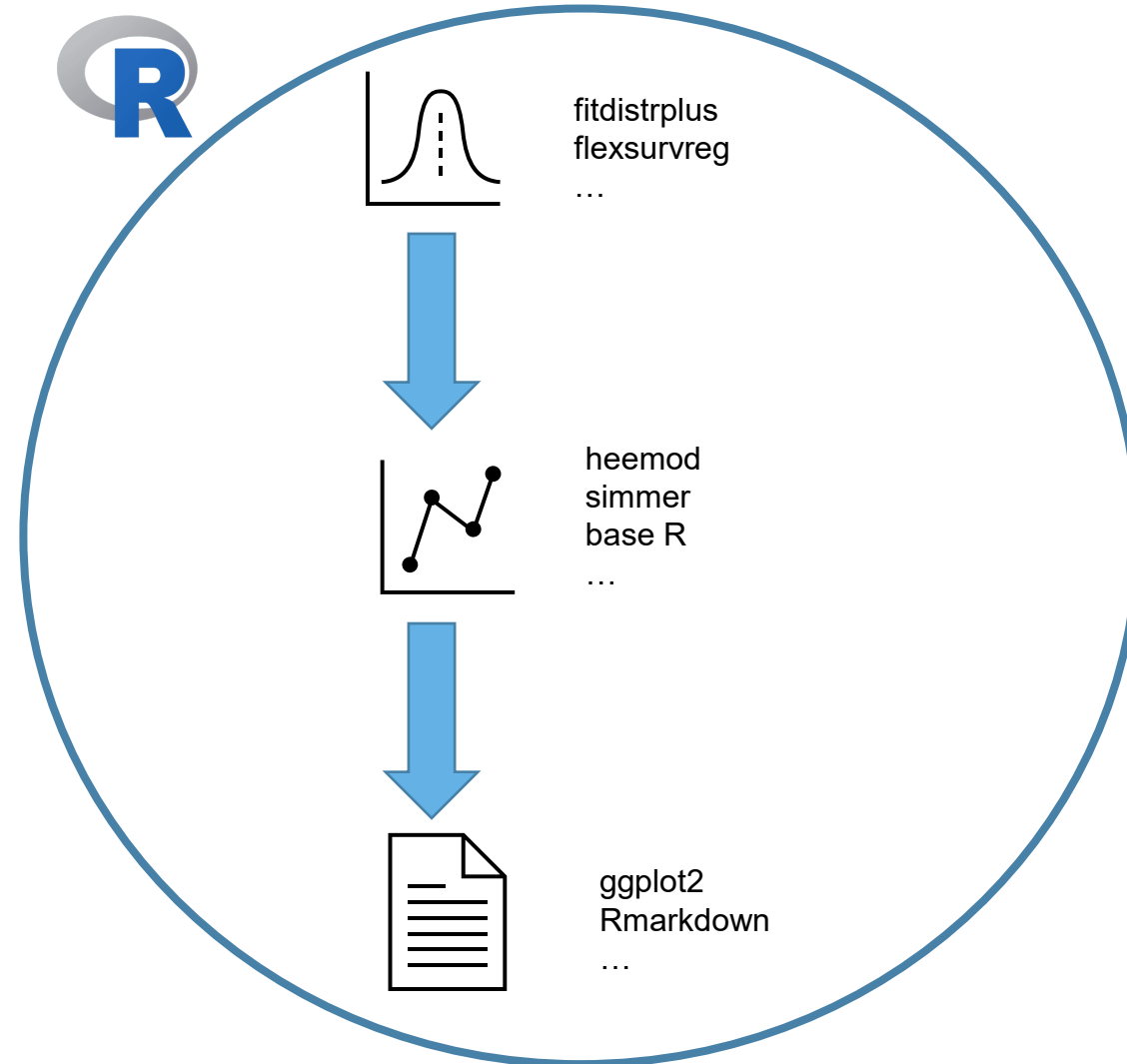
SOME BARRIERS TO PERFORMING REPRODUCIBLE RESEARCH

- Knowledge & Skills: how to do it? What do I need to report?
- Time: being transparent requires additional efforts
- Lack of incentives: you do not get rewarded for producing reproducible research

BUT…

- ~~Knowledge & skills~~
  - ➢ You will learn about it today!
  - ➢ There are a lot of resources available online! (e.g. [FORRT](#))
- ~~Time~~
  - ➢ You have to document your work anyways for your (future) self and colleagues
  - ➢ Does not require extra time if you do it during your research
- ~~Lack of incentives~~
  - ➢ Archiving and citations of software code is possible
  - ➢ All the time and efforts you've put into programming…
    - ➢ Waste to keep it on your computer, someone may benefit!
    - ➢ Leads to more visibility (and potential collaborations)
  - ➢ Recognition and rewards movement

FURTHER ADVANTAGES OF PERFORMING REPRODUCIBLE RESEARCH

- More transparent & allows external review

  → Increases the credibility of Science

- May increase the efficiency of performing Science (re-use)

- Sharing newly acquired knowledge with society

REPRODUCIBLE RESEARCH

1. Share data and software code
2. Document your code & workflow
   - See 'introduction to R' lecture
3. Document version number of R and packages
4. Containerise your code

- See Konkol et al. 2020 for an extensive review of existing infrastructure for transparent and reproducible research

1. SHARING DATA

1. Archive data on an open repository
   - Dataverse (Dutch)
   - DANS-EASY (Dutch)
   - ZENODO (European)
   - Open Science Framework repository (US)
   - …
   - Read the policies!

2. Include (raw) data in your software code if you share it
   - On Github for instance

→ More on this in a following lecture!

2. DOCUMENT YOUR CODE & WORKFLOW

- R Markdown
  - https://rmarkdown.rstudio.com/
- Quarto
  - R Markdown extension
  - https://quarto.org/about.html
- Workflow for Open Reproducible Code in Science (WORCS)
  - https://cjvanlissa.github.io/worcs/
- Alternatives: JupiterLab & Jupiter Notebook
- ...

R MARKDOWN
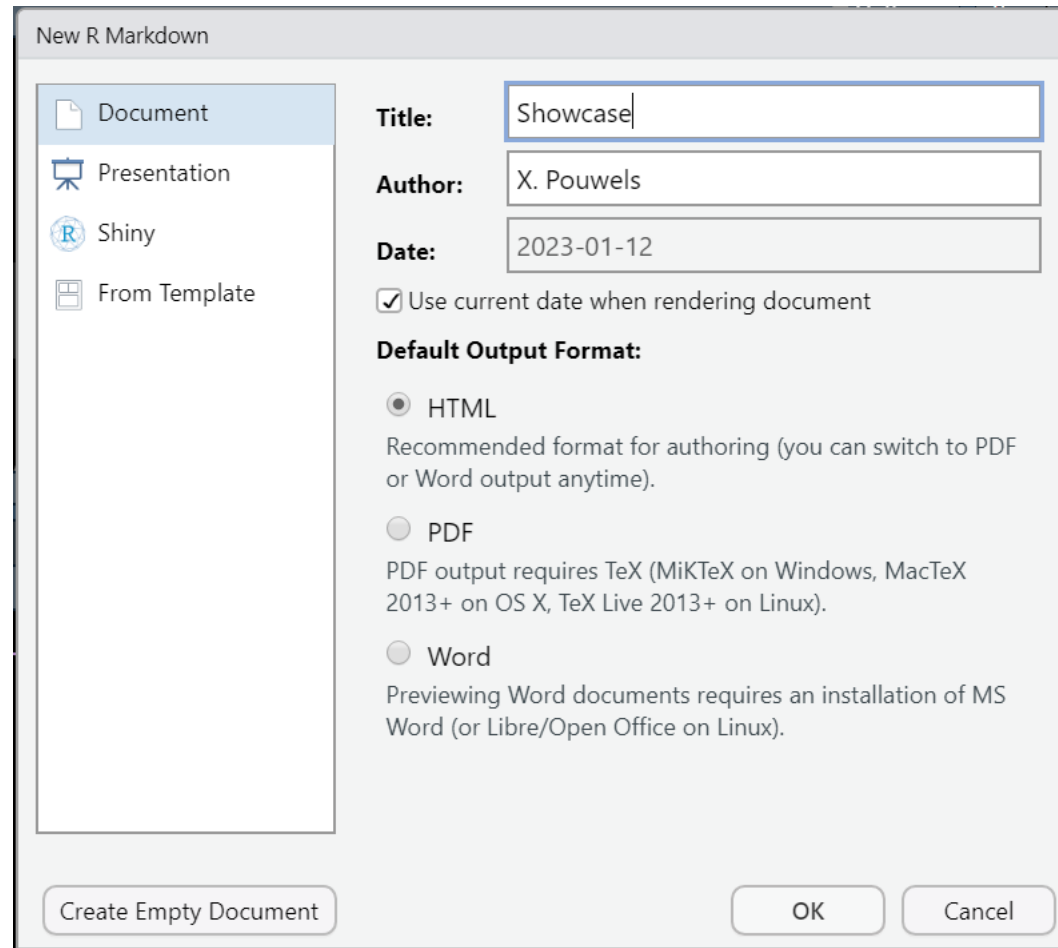
- R package / tool allowing to combine text editing and code execution
    - Perfect to document what you do and explain your results!
    - Easy to share insights with non-R users

How it works

Source: https://rmarkdown.rstudio.com/lesson-2.html

R MARKDOWN
NEW FILE

- *File > New File > R Markdown*

R MARKDOWN ELEMENTS

YAML metadata: title, type of output, and other options (e.g. table of content,…)

Set up chunk

Write text and use # to make headers

Embed executable R code: code chunks

R MARKDOWN
R CODE CHUNKS

```
40 ## R object
41 1. Use `<-` to assign a value to an object
42 2. Use `+, -, *, /` for basic transformation
43 ```{r obj, echo =  TRUE}
44 a <- 2 # single object
45 b <- a + 1
46 b
47 a / b
48 ```

   [1] 3
   [1] 0.6666667
```

Chunk options: e.g. should it be printed, and executed?
Some options (TRUE | FALSE) :
➢ echo : display code chunk in output document ?
➢ eval : run code chunk?
➢ message : display code messages in output document?
➢ warning : display code warnings in output document?

QUARTO

- *File > New File > Quarto document*

- Use '/' or 'CTRL+/' to insert different types of elements in markdown file

- R Markdown 'made easy' (read: with more buttons)
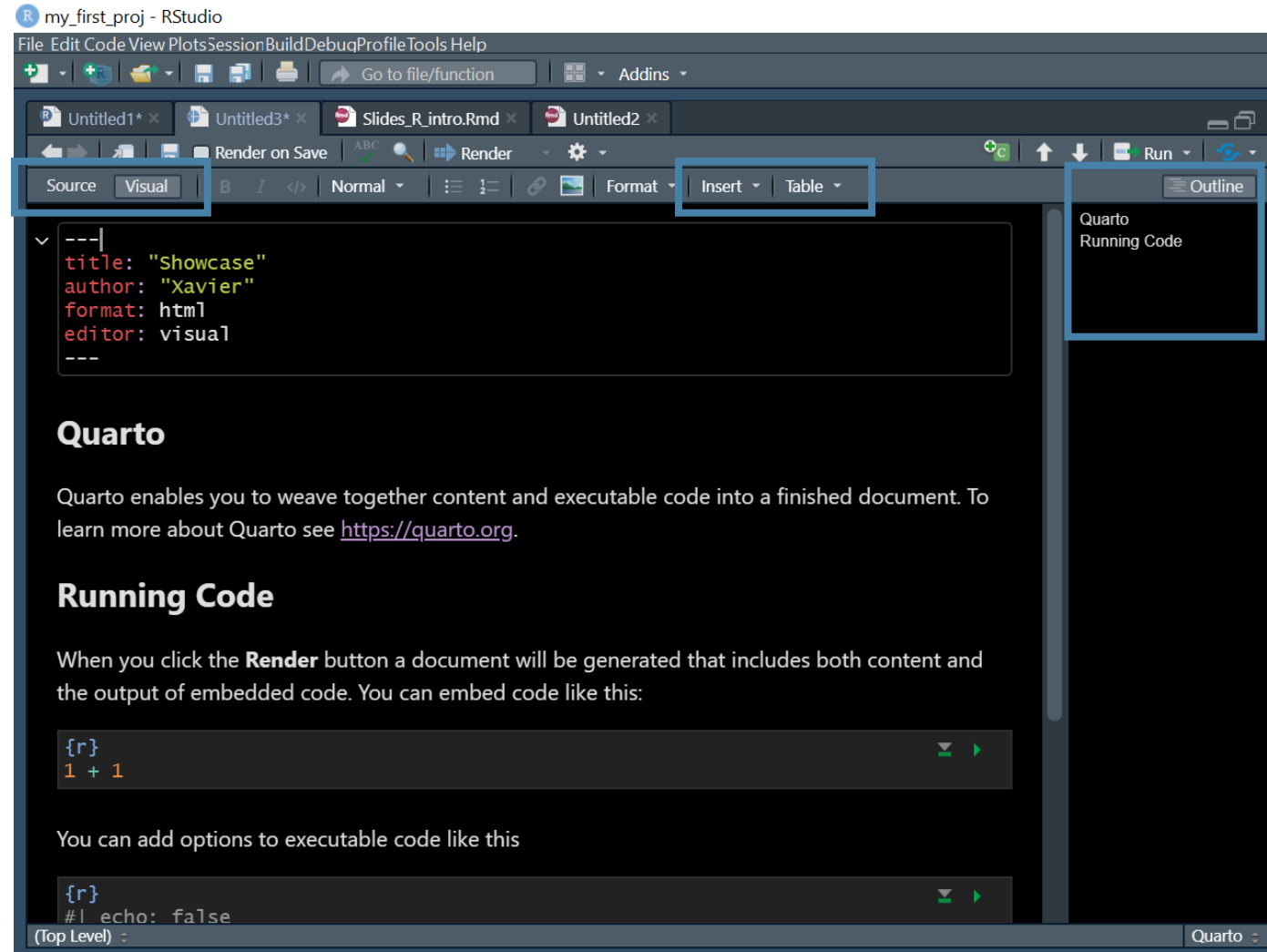
ONE SHORT NOTE ON R MARKDOWN
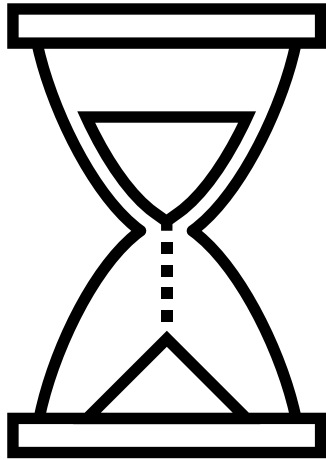
Avoid extensive computations within R markdown…!

WORCS

Workflow for Open Reproducible Code in Science (WORCS)
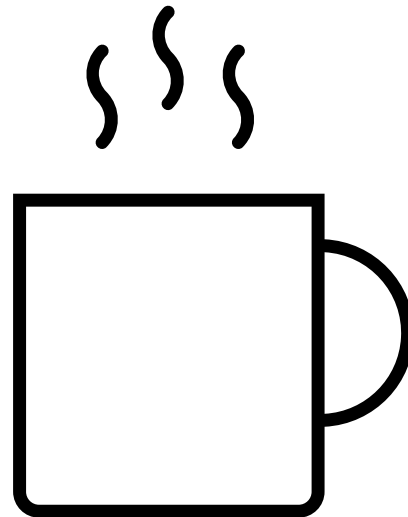
Phase 1: Study design — Create Git remote repository; Dependency management; New RStudio project from WORCS template; Add open materials; Add preregistration; Upload to dedicated repository

Phase 2: Writing/analysis — open_data() closed_data(); Regularly commit changes and push to remote repository; Add manuscript; Add analysis script; Cite @essential or @@non-essential

Phase3: Submission publication — Add badge WORCS open; Snapshot dependencies; Finished project; Knit to PDF; Submit to journal; OSF; Publish preprint

Source: van Lissa et al. 2021

- Workflow based on R integrating R Markdown and GitHub (version control)
- Creates the files necessary to perform reproducible research within your R project

3. DOCUMENT VERSION NUMBER OF R AND PACKAGES

- Why? To ensure your research results are reproducible across machine and time

- R package 'renv' (https://rstudio.github.io/renv/)
  - Creates a library of package for a project
    - Isolate your project
    - Makes your project portable & reproducible

- Note for Windows user: it requires Rtools to work properly
  - e.g. for building packages from source

4. "CONTAINERISE" OF YOUR CODE

- E.g. Docker, Singularity, or Shifter, …
- Code, data, and dependencies are self-contained
  - Makes it 'portable'
- Facilitates sharing and running the software code
  - No need to reproduce the environment!

REPRODUCIBLE RESEARCH IN HEALTH ECONOMIC MODELLING

- R package 'darthpack': framework to improve the transparency of health economic models (Alarid-Escudero et al. 2019)
  - Coding style, file and variable naming conventions, project structure, workflow for building and describing health economic models
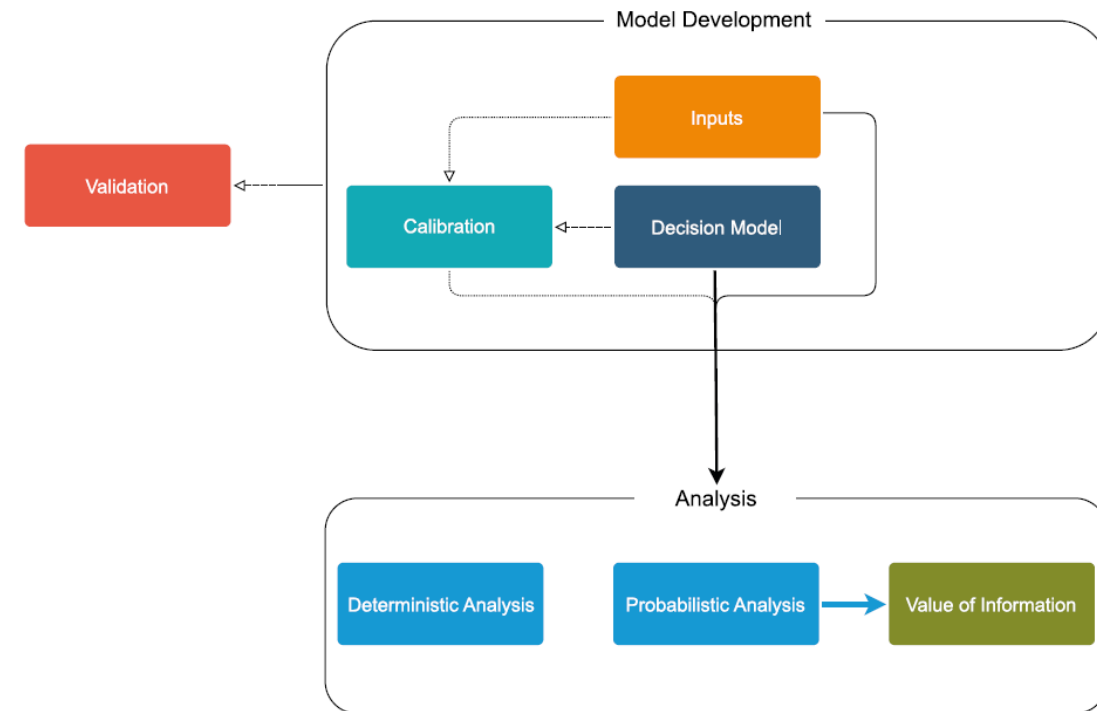 - Paper: http://dx.doi.org/10.1007/s40273-019-00837-x
  - Github: https://github.com/DARTH-git/darthpack

Fig. 1 Schematic representation of the connectivity between the different components of the proposed DARTH framework

ANY QUESTION?

DO-IT-YOURSELF!
WRITE YOUR FIRST REPRODUCIBLE REPORT!

1. Create a new R markdown (or Quarto) document
    - Install the 'rmarkdown' package if necessary

2. Reproduce this document:
    https://doi.org/10.5281/zenodo.7625710

3. Consult these resources if needed:
    - R Markdown reference guide
        - https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf?_ga=2.64091686.867202621.1673517167-2144887827.1580912391
    - R Markdown cheatsheet
        - https://github.com/rstudio/cheatsheets/raw/main/rmarkdown-2.0.pdf

4. The solution is here: https://doi.org/10.5281/zenodo.7625725

RESOURCES

- Alarid-Escudero F, Krijkamp E, Pechlivanoglou P, Jalal H, Kao SY, Yang A, Enns EA. "A need for change! A coding framework for improving transparency in decision modeling". PharmacoEconomics 2019;37(11):1329–1339. http://dx.doi.org/10.1007/s40273-019-00837-x

- Konkol M, Nüst D, Goulier L. Publishing computational research - a review of infrastructures for reproducible and transparent scholarly communication. Res Integr Peer Rev. 2020 Jul 14;5:10. doi: 10.1186/s41073-020-00095-y.

- Munafò, M., Nosek, B., Bishop, D. et al. A manifesto for reproducible science. Nat Hum Behav 1, 0021 (2017). https://doi.org/10.1038/s41562-016-0021

- Van Lissa, C. J., Brandmaier, A. M., Brinkman, L., Lamprecht, A., Peikert, A., , Struiksma, M. E., & Vreede, B. (2021). WORCS: A Workflow for Open Reproducible Code in Science. Data Science. Data Science, vol. 4, no. 1, pp. 29-49. DOI: 10.3233/DS-210031.