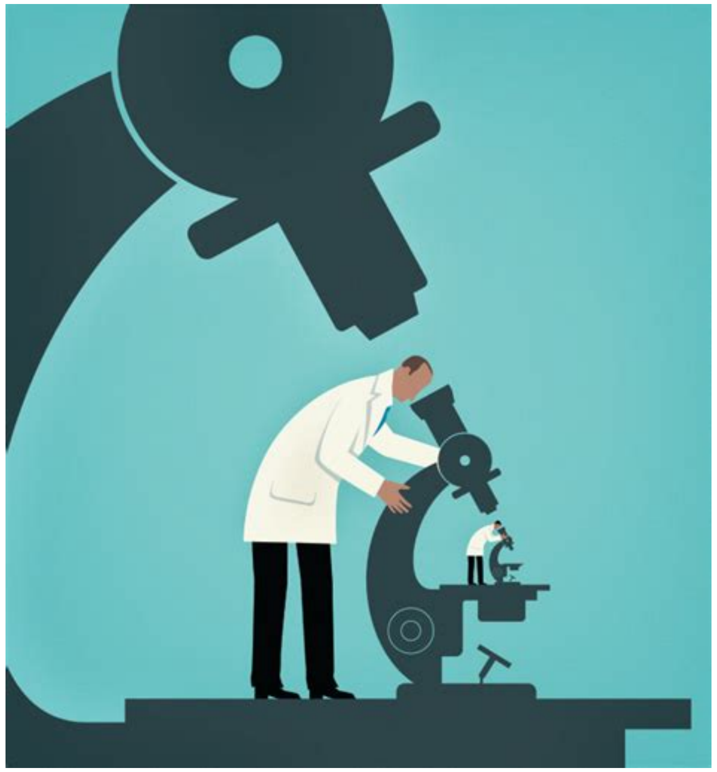


Intro to Reproducibility & Research Data Management

Hermina Ghenu
15 Oct 2024

What I’m going to tell you…

- There’s some unfamiliar words in the schedule and strange requests in the checklist

What I’m going to tell you…

- There’s some unfamiliar words in the schedule and strange requests in the checklist (e.g., “Data Management Plan”)
- What is reproducibility? 🧐

What I’m going to tell you…

- There’s some unfamiliar words in the schedule and strange requests in the checklist (e.g., “Data Management Plan”)
- What is reproducibility? 🧐
- Why do I have to annotate my code?
- Why do I have to hand in my code?
- Why shoud I document my file structure?

What I’m going to tell you…

- There’s some unfamiliar words in the schedule and strange requests in the checklist (e.g., “Data Management Plan”)
- What is reproducibility? 🧐
- Why do I have to annotate my code?
- Why do I have to hand in my code?
- Why should I document my file structure?

Plan for the afternoon: lecture interspersed with activities

Reserach practices:

Reproducibility *vs* Repeatability *vs* Generalizability

Reproducibility: Can other scientists *(or future you)* re-analyze your data & get the exact same result?

Repeatability: Can other scientists replicate your same experiment & achieve a consistent result?

Generalizability: Do other studies exploring the same research question come to the same conclusions?

(National Academies of Science, Engineering, Medicine 2019)

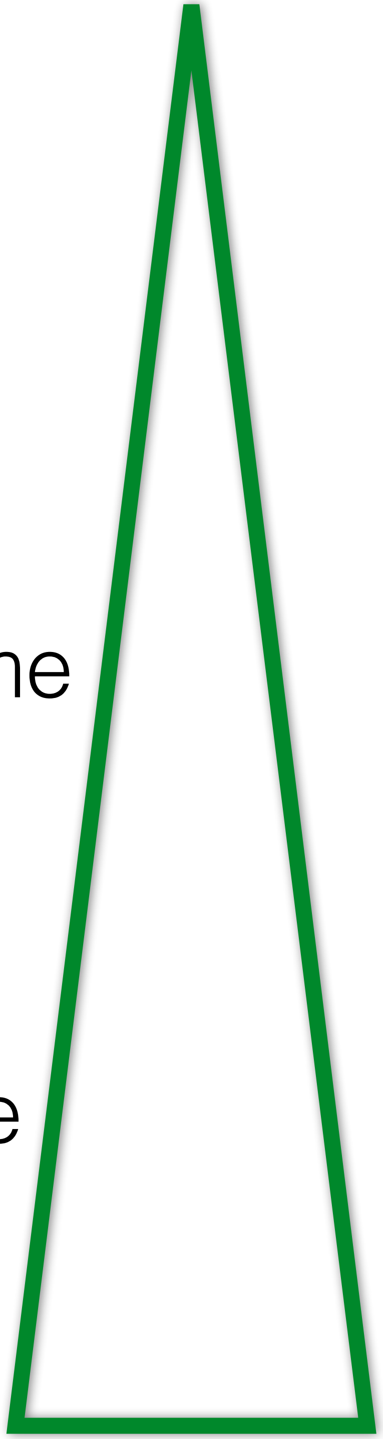
Reserach practices:

Reproducibility *vs* Repeatability *vs* Generalizability

Reproducibility: Can other scientists *(or future you)* re-analyze your data & get the exact same result?

Repeatability: Can other scientists replicate your same experiment & achieve a consistent result?

Generalizability: Do other studies exploring the same research question come to the same conclusions?



(National Academies of Science, Engineering, Medicine 2019)

Reserach practices:

Reproducibility *vs* Repeatability *vs* Generalizability

Reproducibility: Can other scientists *(or future you)* re-analyze your data & get the exact same result?

- first line of defense in creating repeatable research
- focused on computational or data analysis 👩🏾‍💻

Repeatability: Can other scientists replicate your same experiment & achieve a consistent result?

- from initial set-up to final results

Generalizability: Do other studies exploring the same research question come to the same conclusions?

- the ✨ideal✨ we strive for in our science

(National Academies of Science, Engineering, Medicine 2019)

Time for Action

- Why do you think we need *reproducibility*?

Why we need reproducibility

A reproducible research article is a trusted scientific contribution.

- >85% of ecology & evolution publications are **not** reproducible (e.g., no code).
- Papers that make data & code available are more highly cited.

(Kambouris et al., 2024; Maitner et al., 2024)

Nature is sometimes more complex than we imagined.

- e.g., mouse behavioural responses depend on how they are housed & handled

(Nigri et al., 2022)

Mistakes in research can have social / economic impacts.

- e.g., impoverished environment of mice during preclinical studies may explain why most new drug candidates don't work as expected in clinical trials.

(Shemesh & Chen 2023)

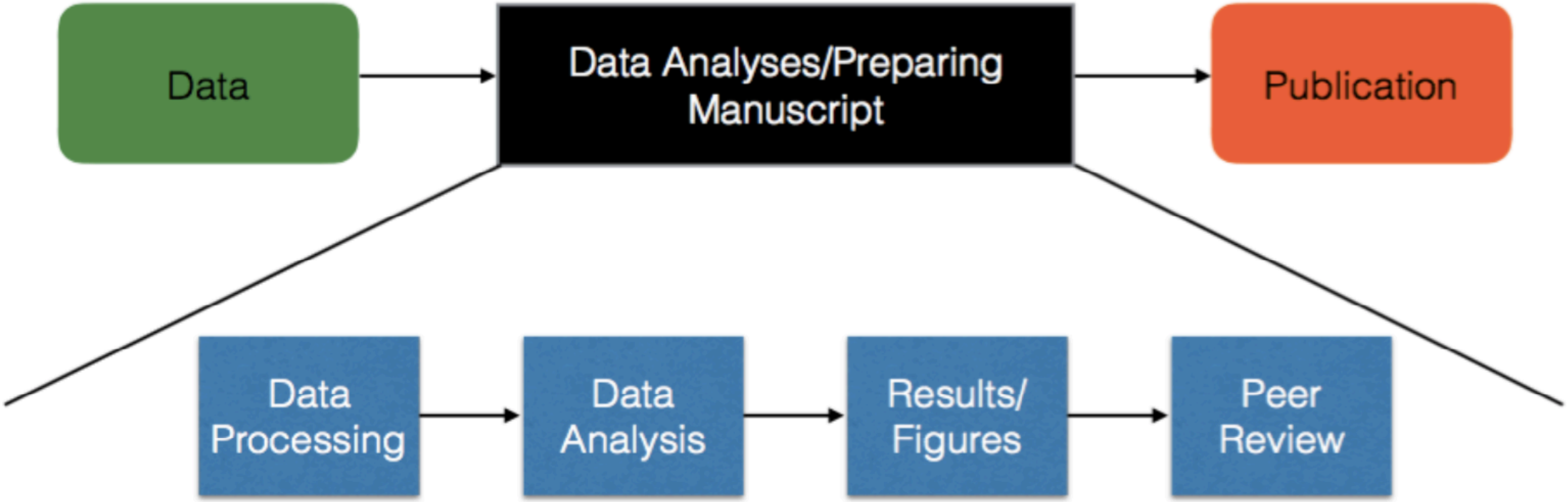
Bad faith actors can diminish our trust in science.

(Kozlov 2022; data forensics details)

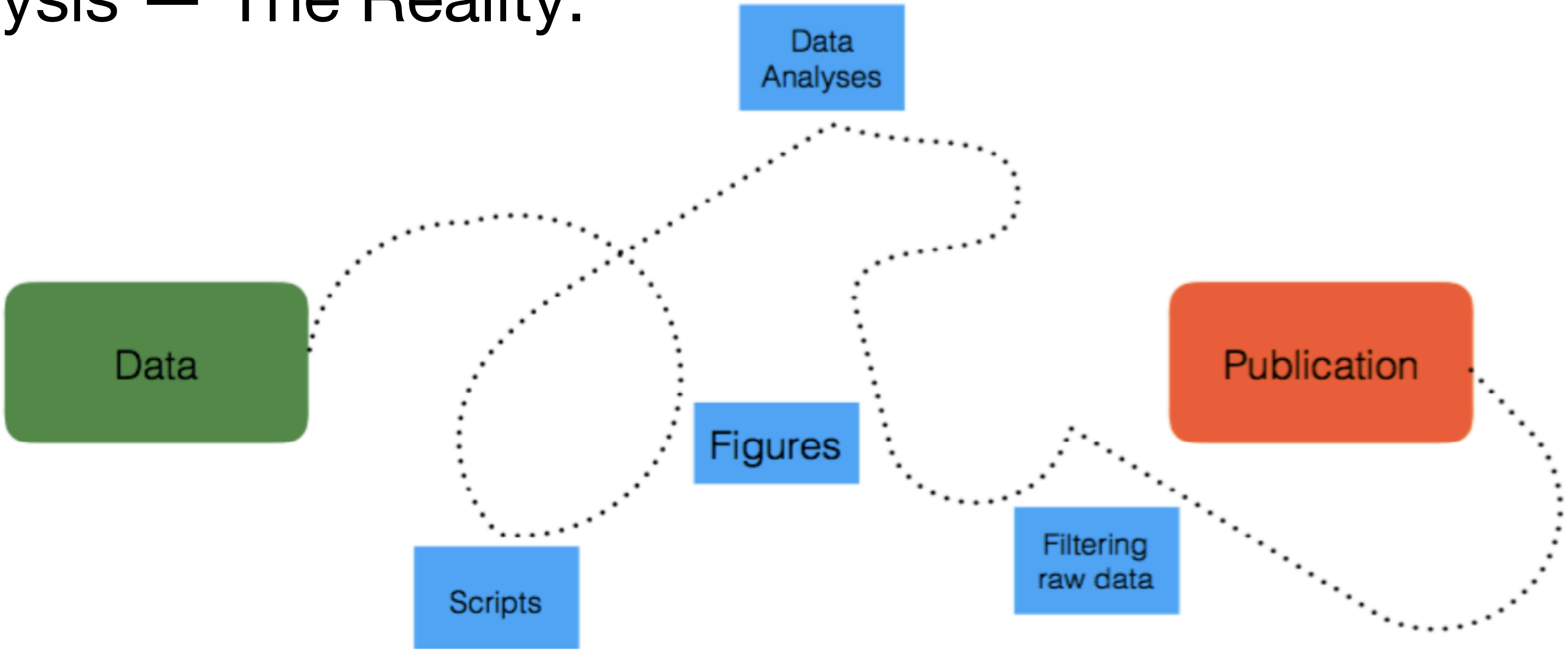
Consistent methods for globally coordinated research efforts.

e.g., in combating disease (Park et al., 2021) or climate change (Halbritter et al., 2019)

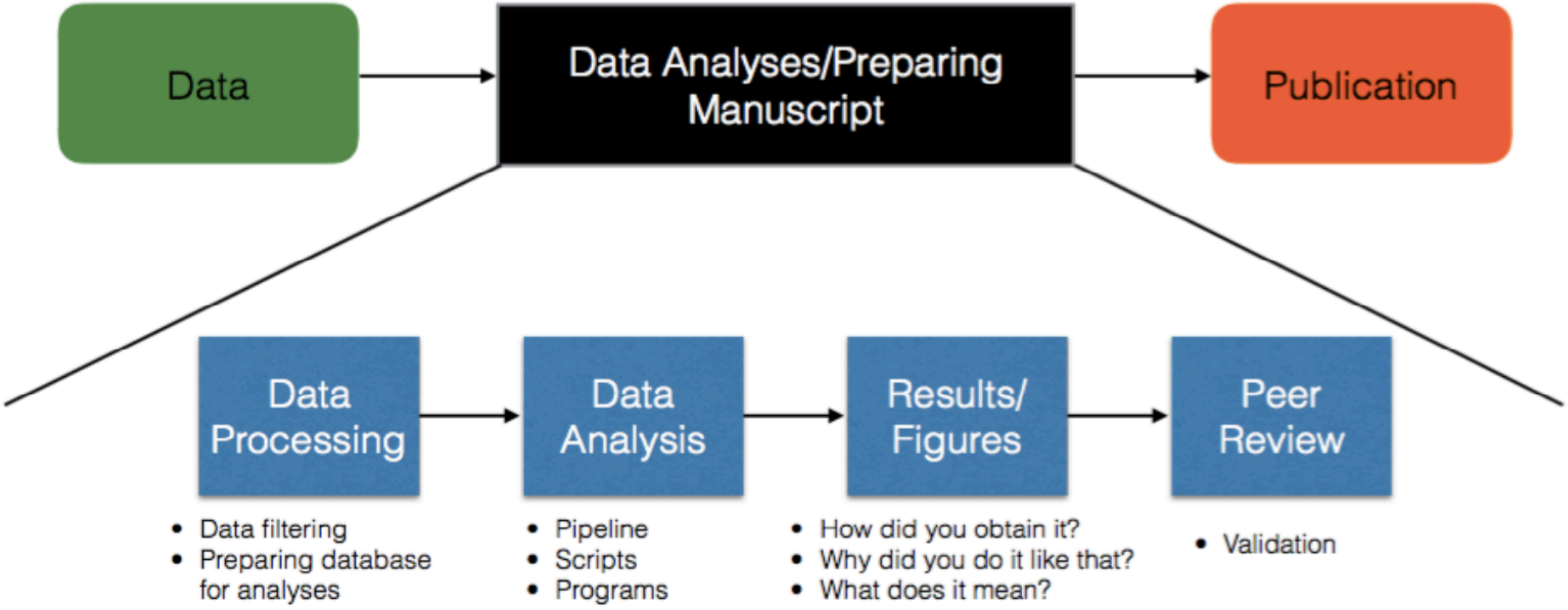
Data analysis — The Dream:



Data analysis — The Reality:



Data analysis — The Dream:



A Data Management Plan & Reproducibility Principles help us get closer to this dream

Doing reproducible research

- **Goal**: able to re-create data and analysis so that you and others can (ideally) arrive at the same interpretations of your results

Doing reproducible research

- **Goal**: able to re-create data and analysis so that you and others can (ideally) arrive at the same interpretations of your results

- Keep *everything!*

B662~
20
B881~

NOBODY WANTS TO DEAL WITH THIS!!!!!

Doing reproducible research

- **Goal**: able to re-create data and analysis so that you (ideally) arrive at the same interpretation/ conclusion from your results

- ~~Keep *everything!*~~

Doing reproducible research

- **Goal**: able to re-create data and analysis so that you (ideally) arrive at the same interpretation/ conclusion from your results

- ~~Keep *everything!*~~

- Keep everything in such a way that you, or people after you, can (happily?) go back to it

How do you achieve reproducibility in research?

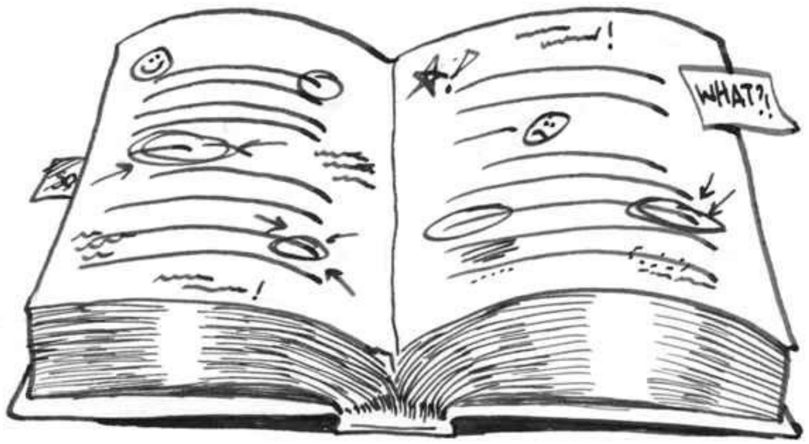
- **Annotate**
  explain what you’re doing and why

- **Automate**
  make your decisions explicit by using code

- **Share**
  provide access to your work

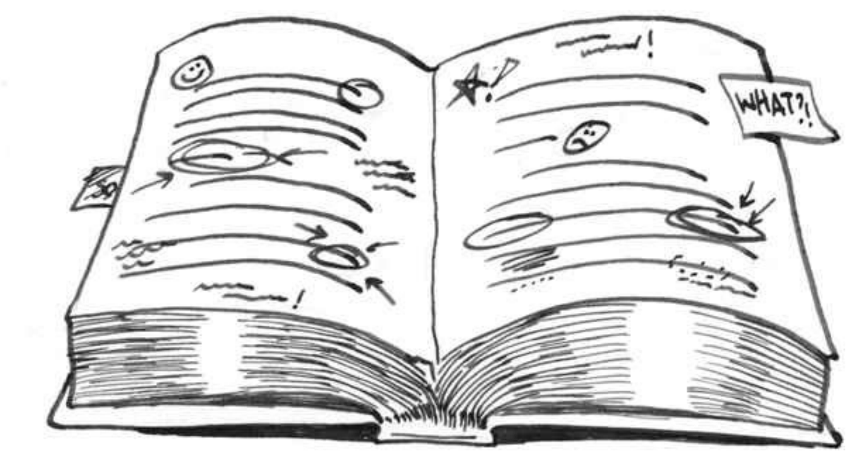
- **Hoard**
  keep (almost) everything

Annotate



- **Write explanations for your future collaborators**
  What? How? *Why?!*

Annotate



- **Write explanations for your future collaborators**
  What? How? *Why?!*

- **Habits:** use script headers, use meaningful & human-readable names, comment your code

- **Tools:** notebook documents

Time for Action

- Open RStudio and create a new R Notebook document. Save then knit the document to pdf. What does this do?

Time for Action

- Open RStudio and create a new R Notebook document. Save then knit the document to pdf. What does this do?

- **Recall annotate habits:** use meaningful & human-readable variable names, comment your code, use script headers.

  - How do you do that?

  - Type "# Annotate!" both *inside* and *outside* of the R code block. How are these displayed differently after you knit?

Time for Action

- Open RStudio and create a new R Notebook document. Save then knit the document to pdf. What does this do?

- **Recall annotate habits:** use meaningful & human-readable variable names, comment your code, use script headers.

  - How do you do that?

  - Type “# Annotate!” both *inside* and *outside* of the R code block. How are these displayed differently after you knit?

- Switch the markdown editing mode from Source to Visual. What does this do? What objects can you add to the text?

Time for Action

- Open RStudio and create a new R Notebook document. Save then knit the document to pdf. What does this do?

- **Recall annotate habits:** use meaningful & human-readable variable names, comment your code, use script headers.

  - How do you do that?

  - Type “# Annotate!” both *inside* and *outside* of the R code block. How are these displayed differently after you knit?

- Switch the markdown editing mode from Source to Visual. What does this do?

- Modify the header to add new fields for “author:” and “date:”. What other authorship attribution information may be useful?

How do you achieve reproducibility in research?

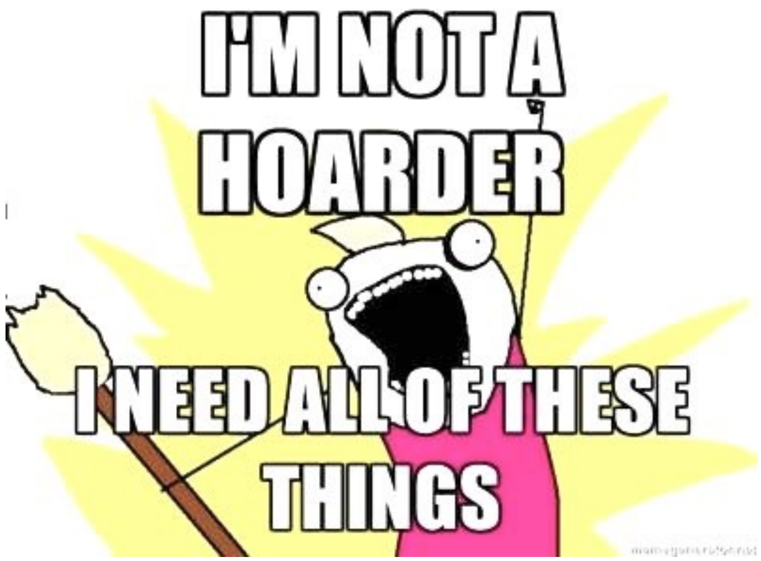
- **Annotate**
  explain what you’re doing and why

- **Automate**
  make your decisions explicit by using code

- **Share**
  provide access to your work

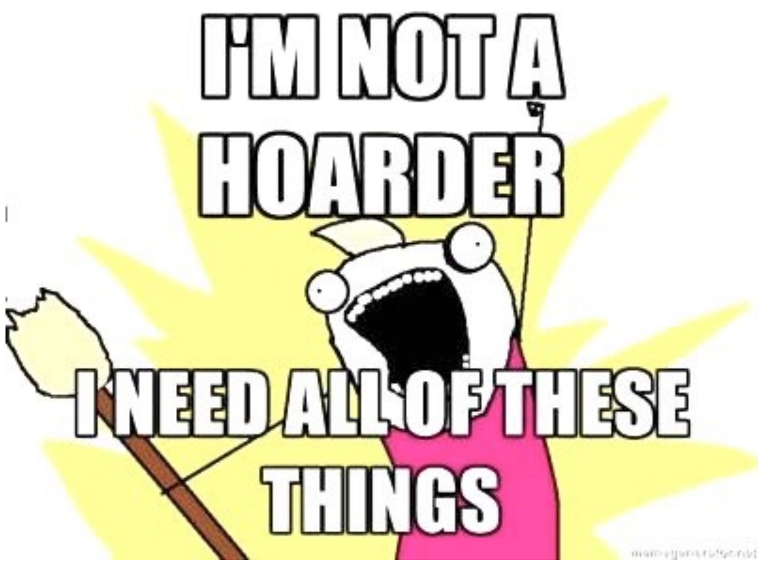
- **Hoard**
  keep (almost) everything

Hoard



- **Keep almost everything**

Hoard



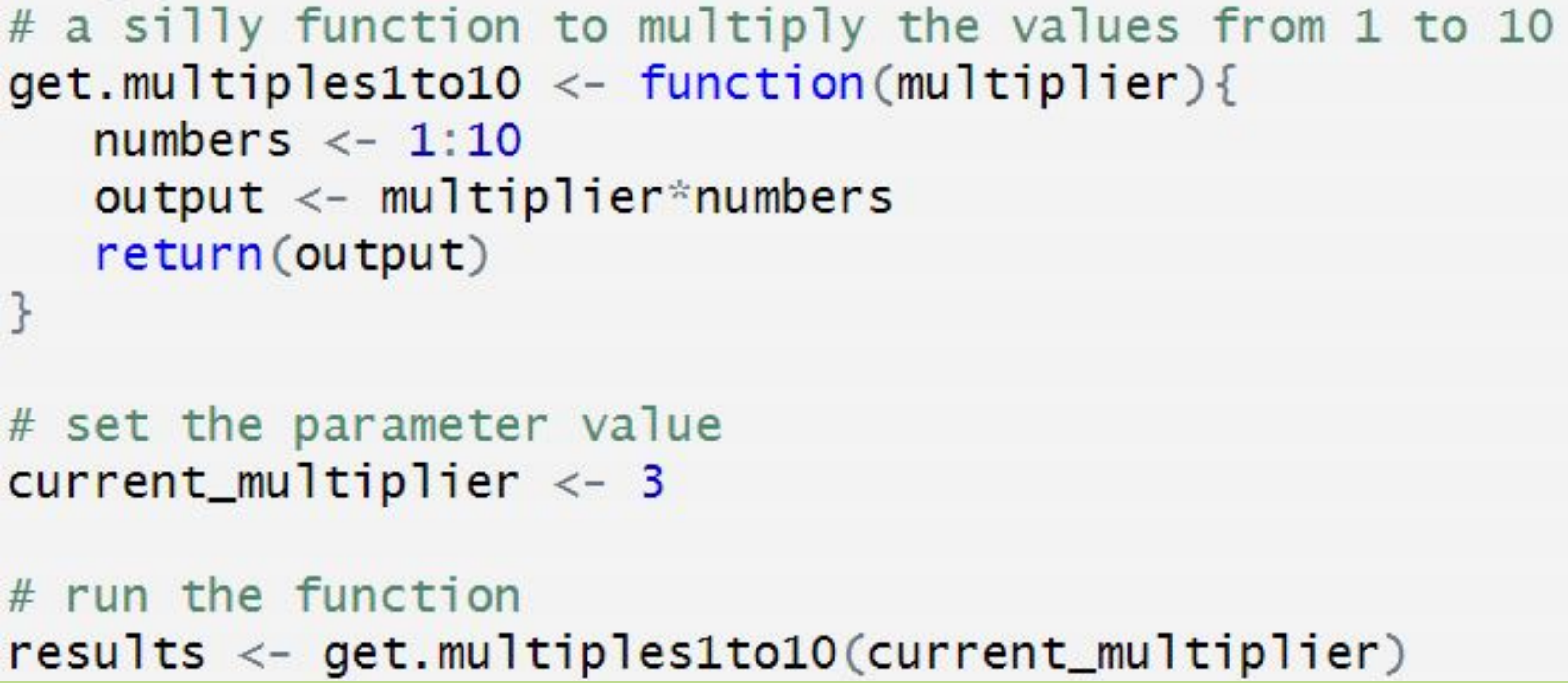
- **Keep almost everything**

- **Habits:** backup regularly (daily!), exact version of software, store raw data & intermediate steps in data processing, store code & progress on code

- **Tools:** backup software (e.g. Time Machine), version control (e.g. git), online repositories (e.g. GitHub)

Time for Action

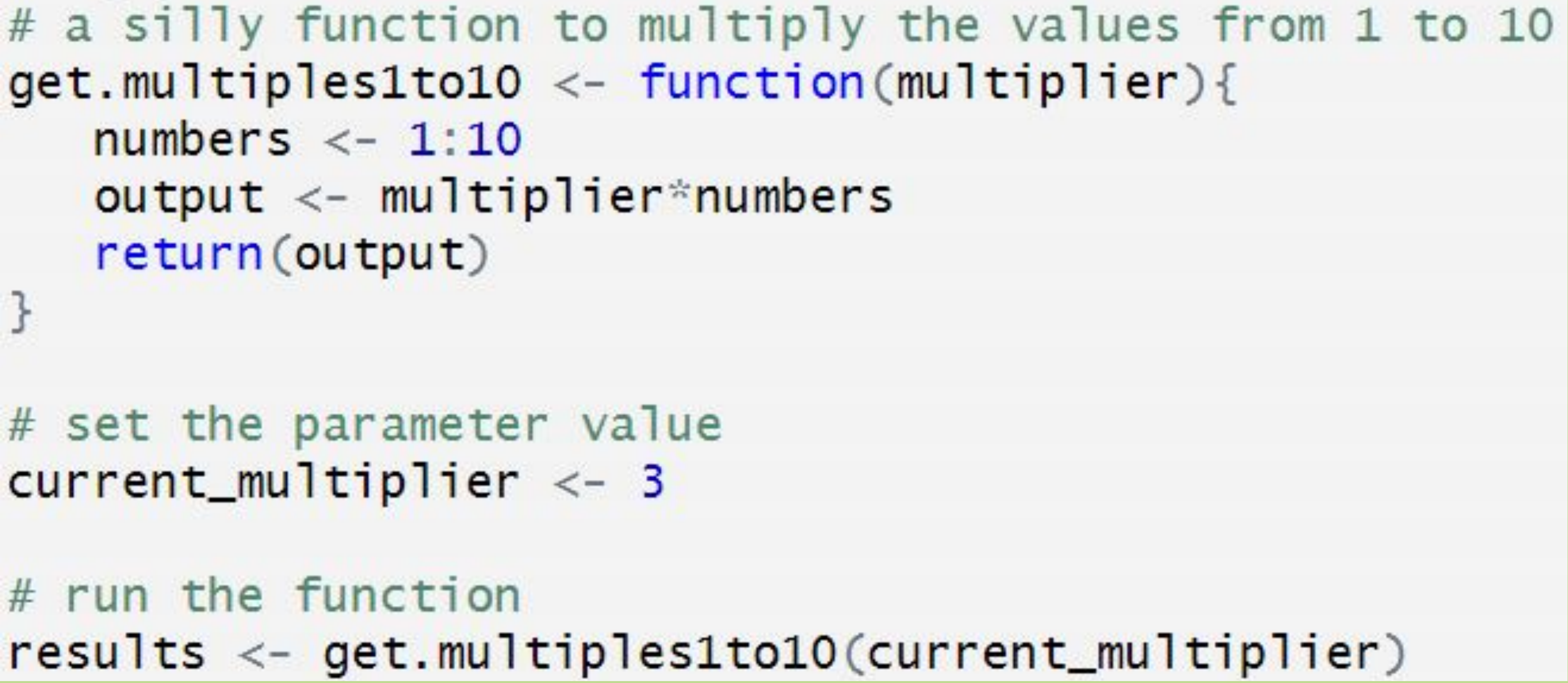
- Copy the following code into the code block in your Rnotebook file:



```
# a silly function to multiply the values from 1 to 10
get.multiples1to10 <- function(multiplier){
   numbers <- 1:10
   output <- multiplier*numbers
   return(output)
}

# set the parameter value
current_multiplier <- 3

# run the function
results <- get.multiples1to10(current_multiplier)
```

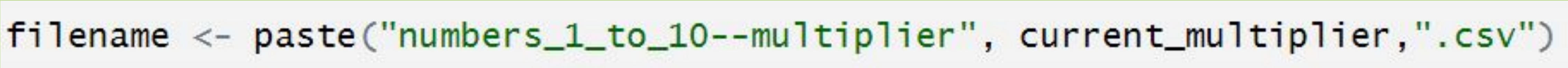
Time for Action

- Copy the following code into the code block in your Rnotebook file:



```
# a silly function to multiply the values from 1 to 10
get.multiples1to10 <- function(multiplier){
   numbers <- 1:10
   output <- multiplier*numbers
   return(output)
}

# set the parameter value
current_multiplier <- 3

# run the function
results <- get.multiples1to10(current_multiplier)
```

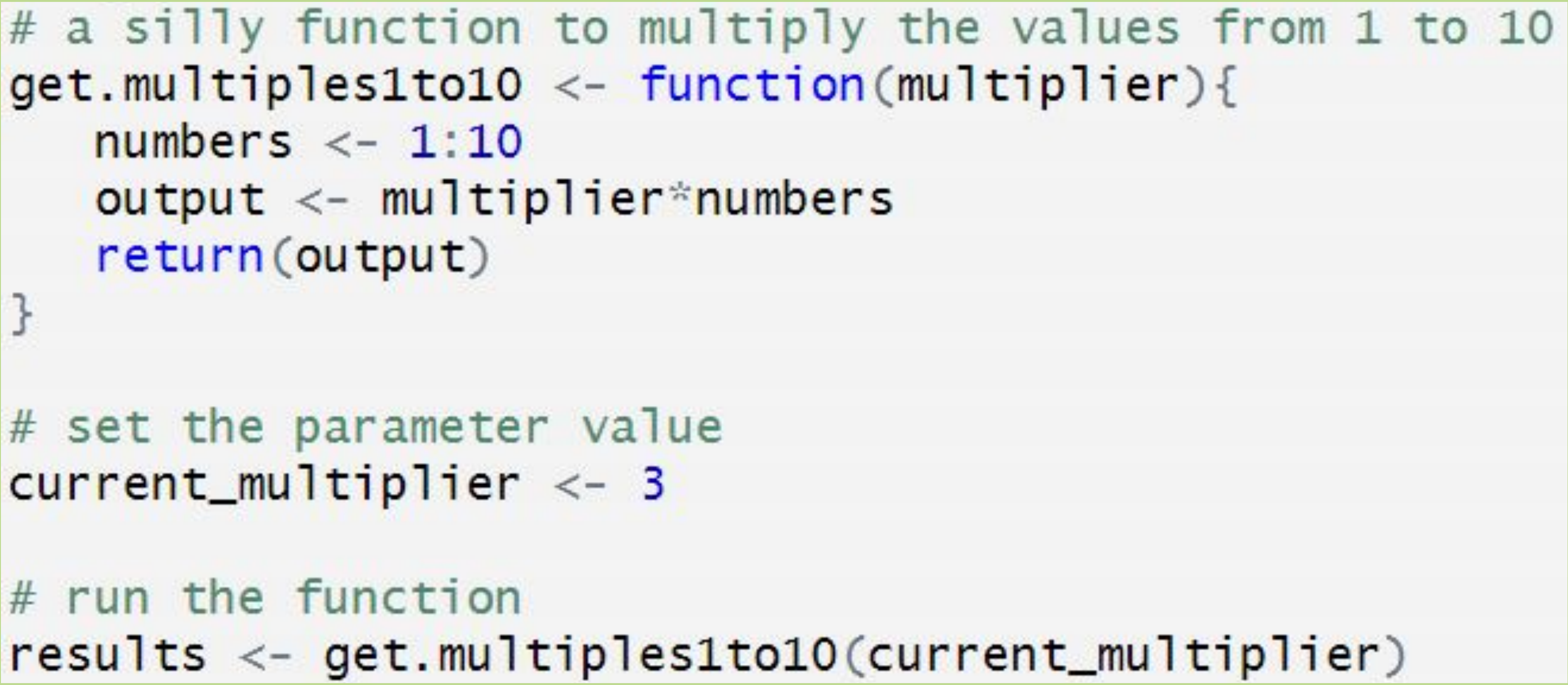
- Use paste(…) to store the parameter value in the filename:



```
filename <- paste("numbers_1_to_10--multiplier", current_multiplier,".csv")
```

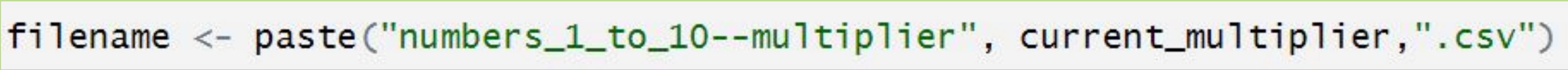
Time for Action

- Copy the following code into the code block in your Rnotebook file:



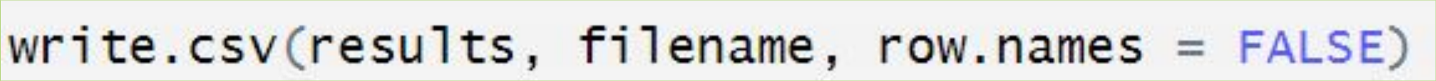
```
# a silly function to multiply the values from 1 to 10
get.multiples1to10 <- function(multiplier){
   numbers <- 1:10
   output <- multiplier*numbers
   return(output)
}

# set the parameter value
current_multiplier <- 3

# run the function
results <- get.multiples1to10(current_multiplier)
```

- Use paste(…) to store the parameter value in the filename:



```
filename <- paste("numbers_1_to_10--multiplier", current_multiplier,".csv")
```

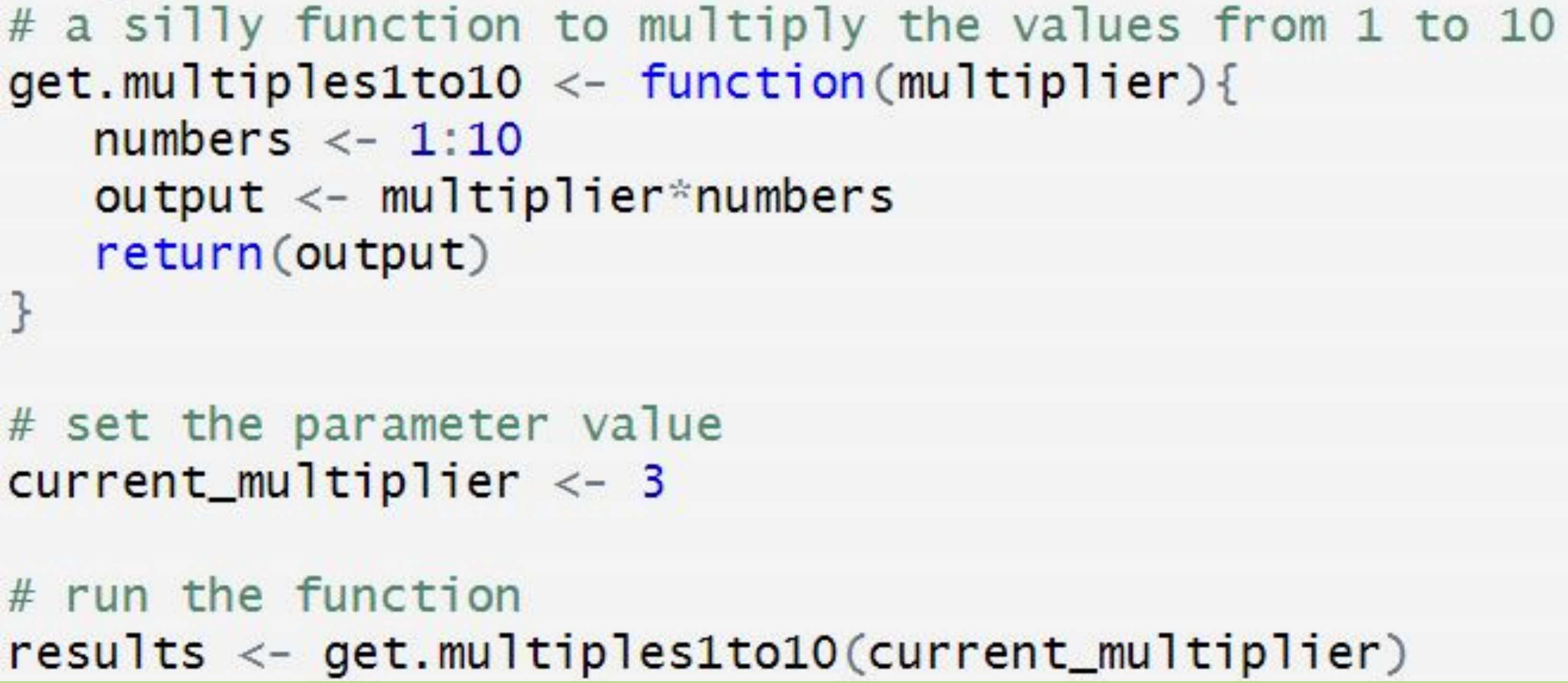
- Save your result to an output file with appropriate filename:



```
write.csv(results, filename, row.names = FALSE)
```

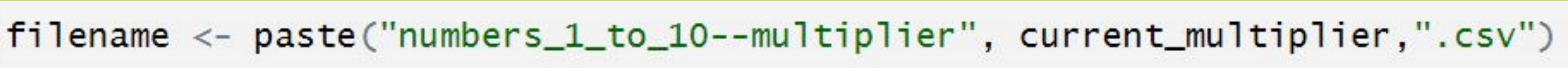
Time for Action

- Copy the following code into the code block in your Rnotebook file:



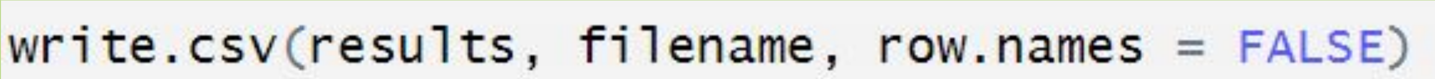
```
# a silly function to multiply the values from 1 to 10
get.multiples1to10 <- function(multiplier){
   numbers <- 1:10
   output <- multiplier*numbers
   return(output)
}

# set the parameter value
current_multiplier <- 3

# run the function
results <- get.multiples1to10(current_multiplier)
```

- Use paste(…) to store the parameter value in the filename:



```
filename <- paste("numbers_1_to_10--multiplier", current_multiplier,".csv")
```

- Save your result to an output file with appropriate filename:



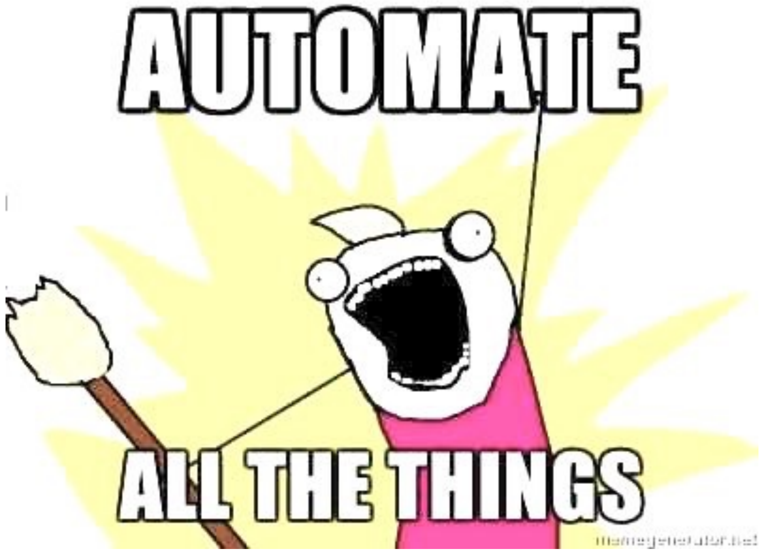
```
write.csv(results, filename, row.names = FALSE)
```

- Are there any annotate habits that we are using here?
  (annotate habits: variable names, annotate code, script headers)

How do you achieve reproducibility in research?

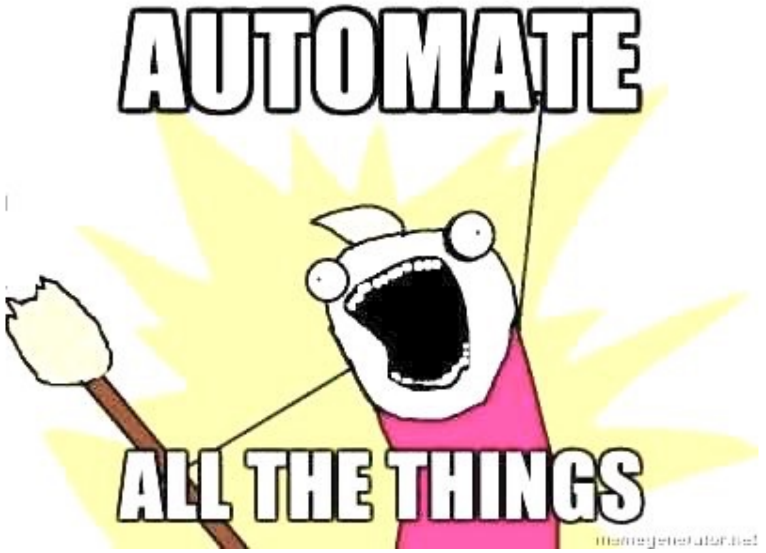
- **Annotate**
  explain what you’re doing and why
- **Automate**
  make your decisions explicit by using code
- **Share**
  provide access to your work
- **Hoard**
  keep (almost) everything

Automate



- **Avoid manual manipulation**
  waste of time, error-prone,
  decisions are <u>not</u> explicit and can be inconsistent

Automate



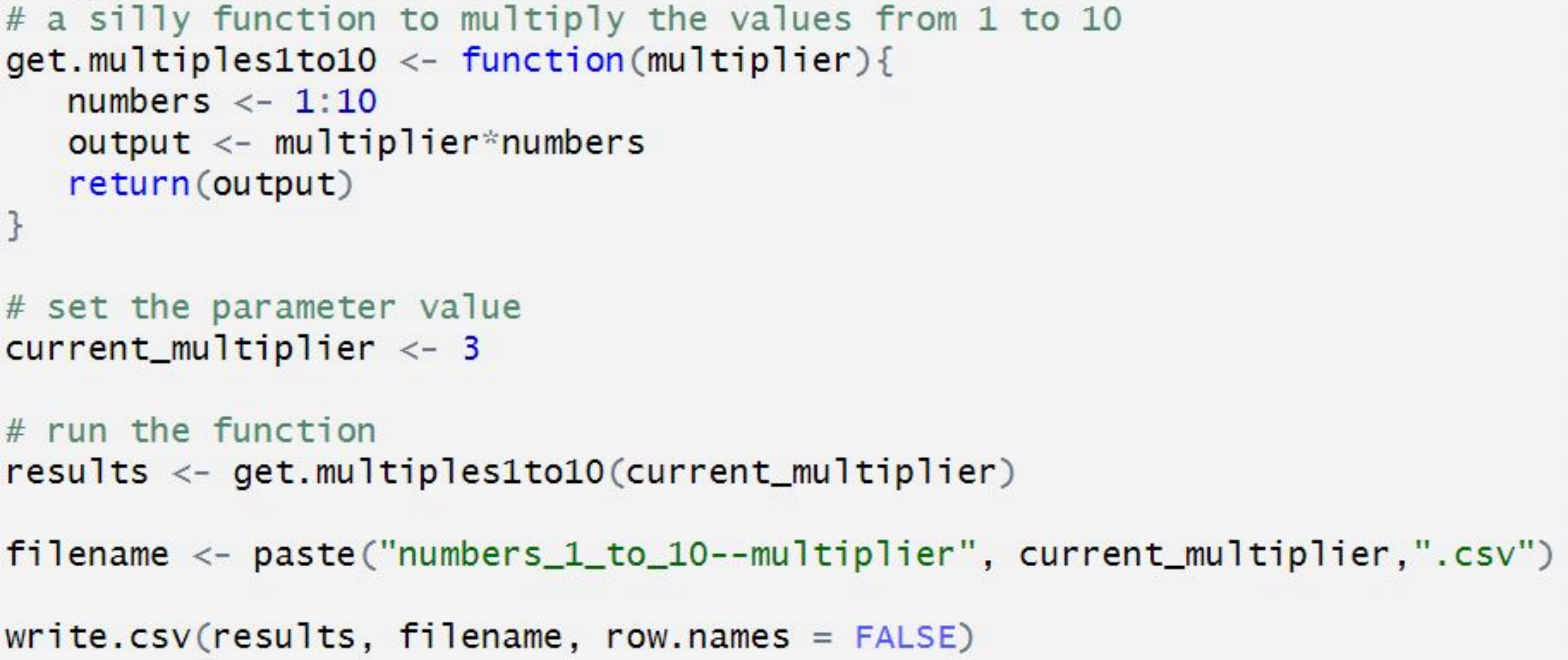
- **Avoid manual manipulation**
  waste of time, error-prone,
  decisions are <u>not</u> explicit and can be inconsistent

- **Habits:** find+replace, use a scripting language for your analyses, automatically save parameters in the filename.

- **Tools:** notebook documents

Time for Action

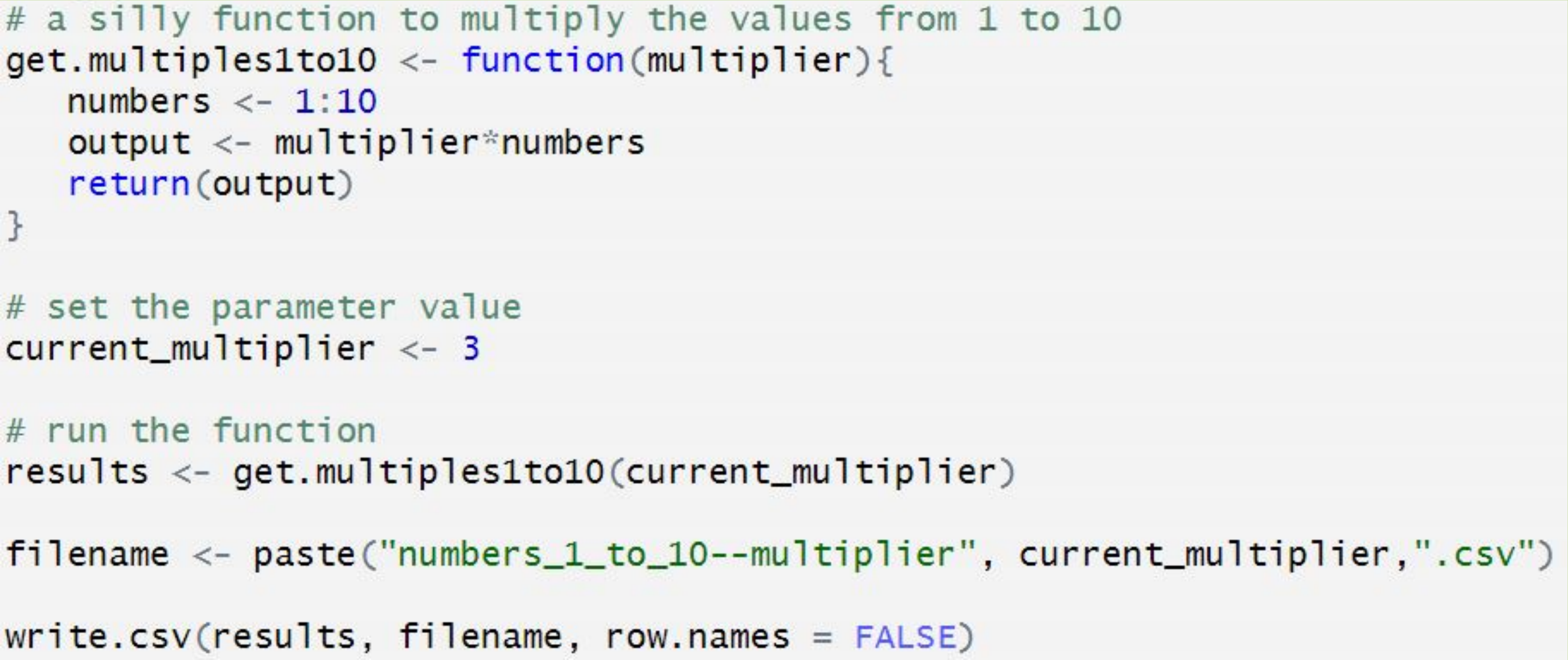
- Are there any automate habits that we used in this code? (automate habits: find+replace, scripting language, filename)



```
# a silly function to multiply the values from 1 to 10
get.multiples1to10 <- function(multiplier){
   numbers <- 1:10
   output <- multiplier*numbers
   return(output)
}

# set the parameter value
current_multiplier <- 3

# run the function
results <- get.multiples1to10(current_multiplier)

filename <- paste("numbers_1_to_10--multiplier", current_multiplier,".csv")

write.csv(results, filename, row.names = FALSE)
```

Time for Action

- Are there any automate habits that we used in this code? (automate habits: find+replace, scripting language, filename)



```
# a silly function to multiply the values from 1 to 10
get.multiples1to10 <- function(multiplier){
   numbers <- 1:10
   output <- multiplier*numbers
   return(output)
}

# set the parameter value
current_multiplier <- 3

# run the function
results <- get.multiples1to10(current_multiplier)

filename <- paste("numbers_1_to_10--multiplier", current_multiplier,".csv")

write.csv(results, filename, row.names = FALSE)
```

- Use find+replace to change the name of the function from `get.multiples1to10` to a new name that makes sense to you.

How do you achieve reproducibility in research?

- **Annotate**
  explain what you're doing and why
- **Automate**
  make your decisions explicit by using code
- **Share**
  provide access to your work
- **Hoard**
  keep (almost) everything

Share

- **Fundamentally, research is about sharing**
  with collaborators, with other scientists

Share

- **Fundamentally, research is about sharing** with collaborators, with other scientists
- **Habits:** think about your audience when analysing (see annotate), share early and often
- **Tools:** online repositories for data (e.g. Dryad), code (e.g. GitHub), and papers (e.g. bioRxiv)

How do you achieve reproducibility in research?

- **Annotate**
  explain what you’re doing and why

- **Automate**
  make your decisions explicit by using code

- **Share**
  provide access to your work

- **Hoard**
  keep (almost) everything

Hopefully, by implementing reproducibility principles, our workspace can be more like this:

Review of what I told you…

- What is reproducibility?

Review of what I told you…

- What is reproducibility?
- Why do you have to annotate your code?

Review of what I told you…

- What is reproducibility?
- Why do you have to annotate your code?
- Why do you have to hand in your code?

Review of what I told you…

- What is reproducibility?
- Why do you have to annotate your code?
- Why do you have to hand in your code?
- Why should you document your file structure?