Data science lifecycle & Exploratory data analysis using visualization

ds4owd - data science for openwashdata

Lars Schöbitz

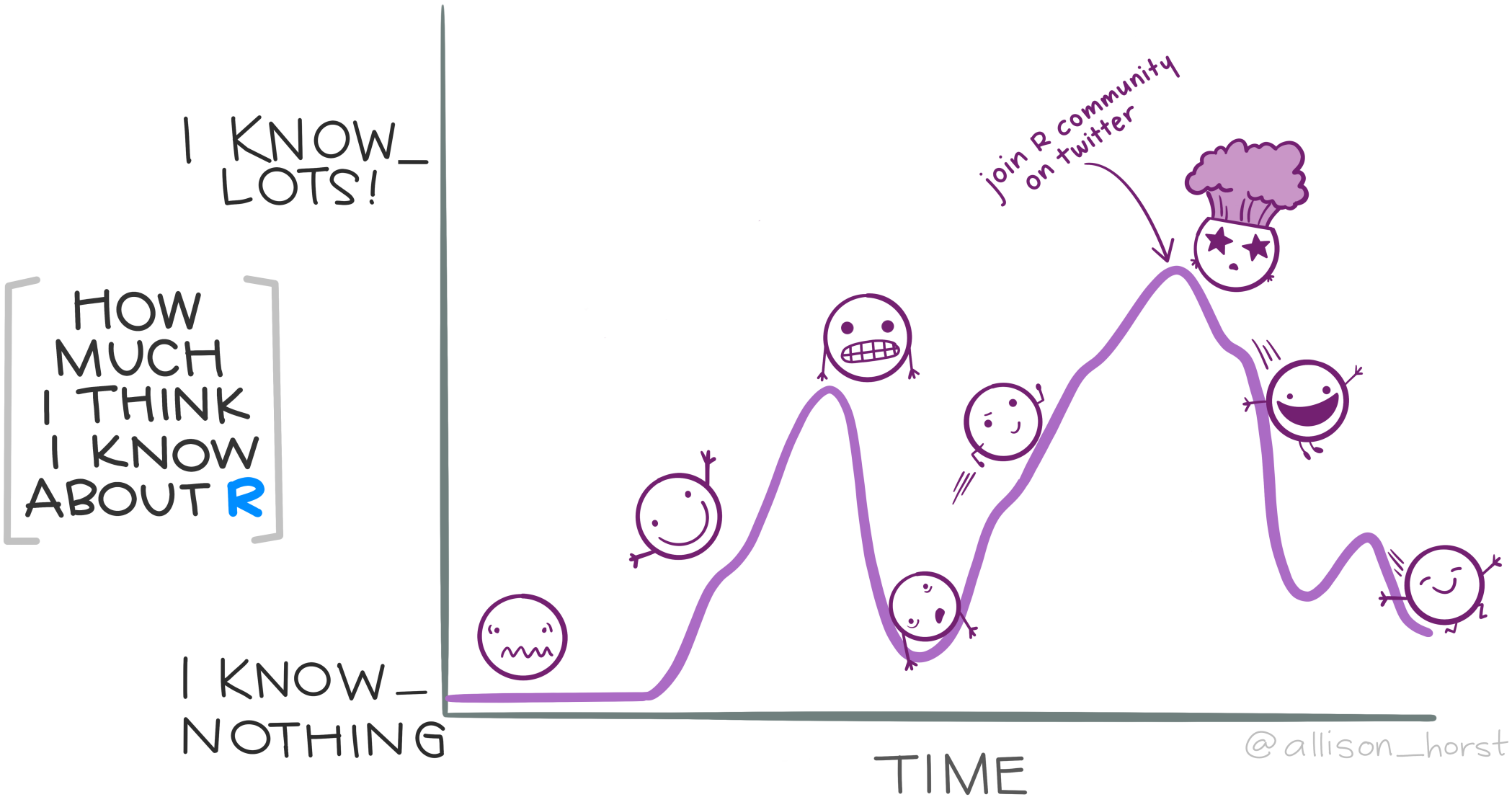
2023-11-07

Q: How do I successfully complete the course?

You successfully complete the course and you will receive a certificate of completion if you:

hand in a complete capstone project report that uses a dataset of your choice by 30 January 2024 (instructions will follow)

This is the only requirement to successfully complete the course, independent of how many classes you attended or how many homework assignments you completed.



Artwork by @allison_horst

ds4owd-001.github.io/website/

Solving coding problems

Tipps for search engines

- Use actionable verbs that describe what you want to do
- Be specific
- Add R to the search query
- Add the name of the R package name to the search query
- Scroll through the top 5 results (don’t just pick the first)

Example: “How to remove a legend from a plot in R ggplot2”

Stack Overflow

What is it?

- The biggest support network for (coding) problems
- Can be intimidating at first
- Up-vote system

Workflow

- First, briefly read the question that was posted
- Then, read the answer marked as “correct”
- Then, read one or two more answers with high votes
- Then, check out the “Linked” posts
- Always give credit for the solution

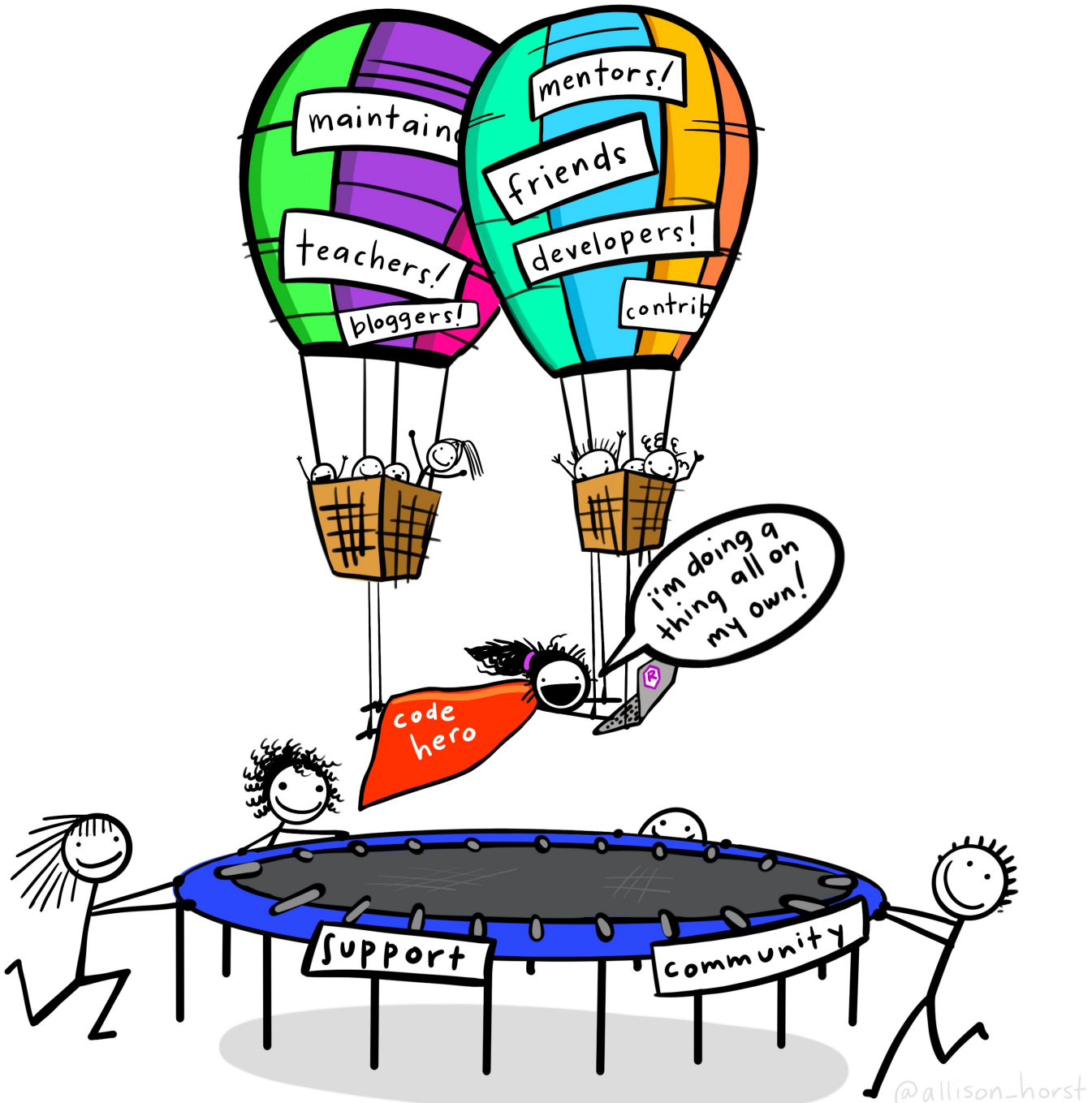
Tipps for AI tools

- Use actionable verbs that describe what you want to do
- Be specific
- Add R to the search query
- Add the name of the R package name to the search query

Example: “How to remove a legend from a plot in R ggplot2”

Other sources for help

- Posit Community Forum: https://community.rstudio.com/
- Documentation websites: https://ggplot2.tidyverse.org/
- Mastodon tag: #rstats
- Quarto GitHub Discussion: https://github.com/quarto-dev/quarto-cli/discussions



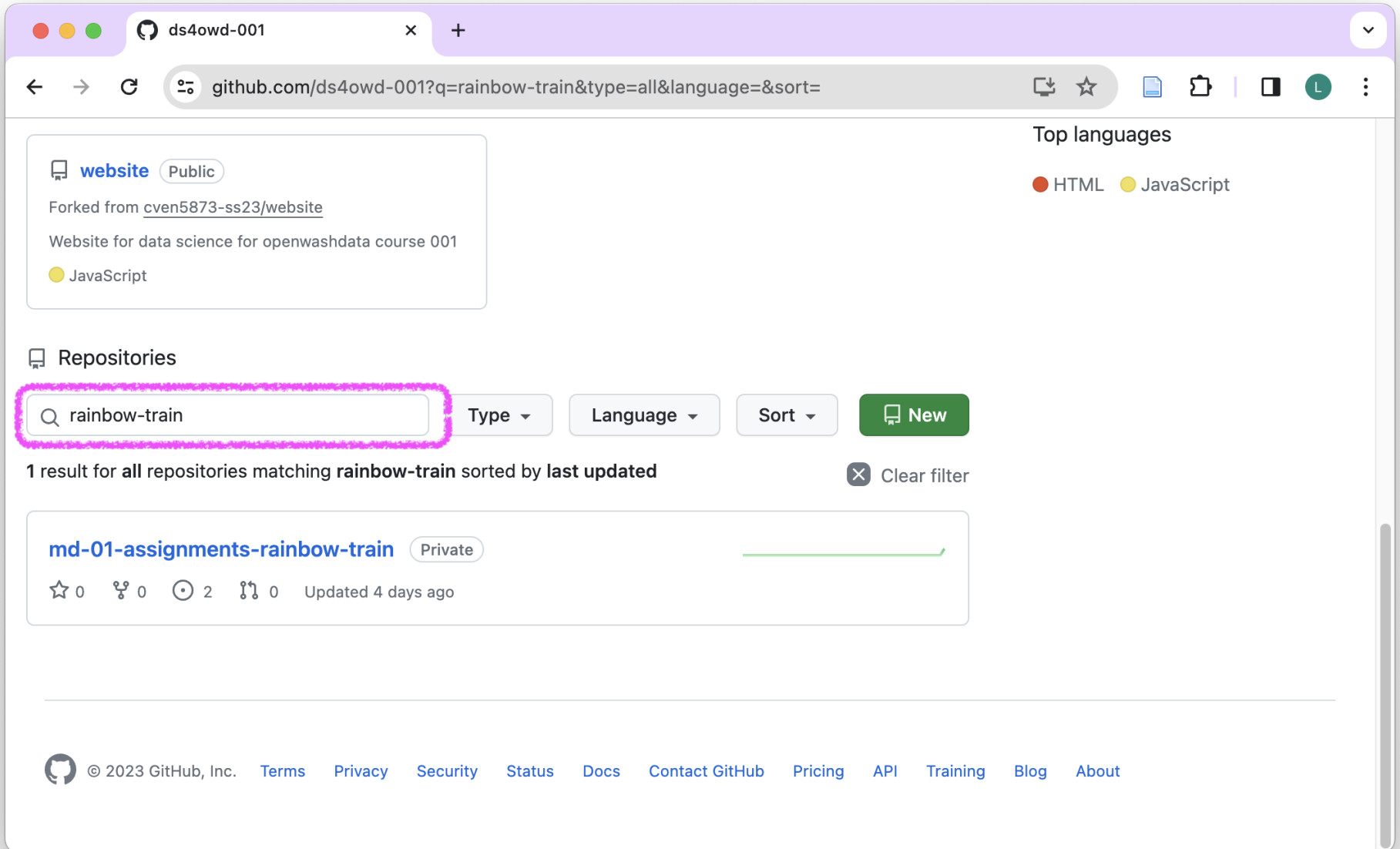
ds4owd-001.github.io/website/

Homework assignment module 1

on GitHub Organisation

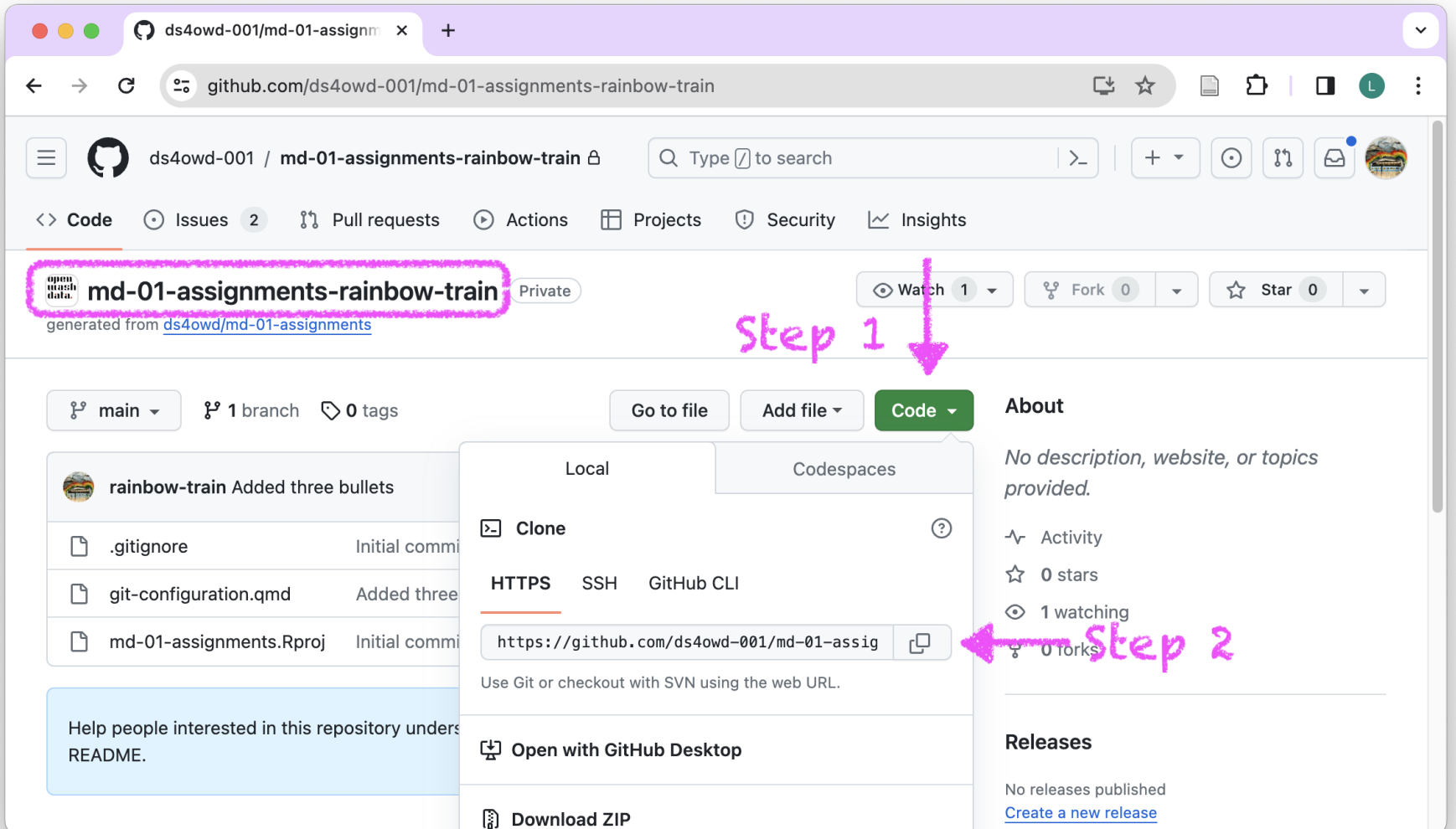
Bookmark this link in your browser!

github.com/ds4owd-001



ds4owd-001.github.io/website/

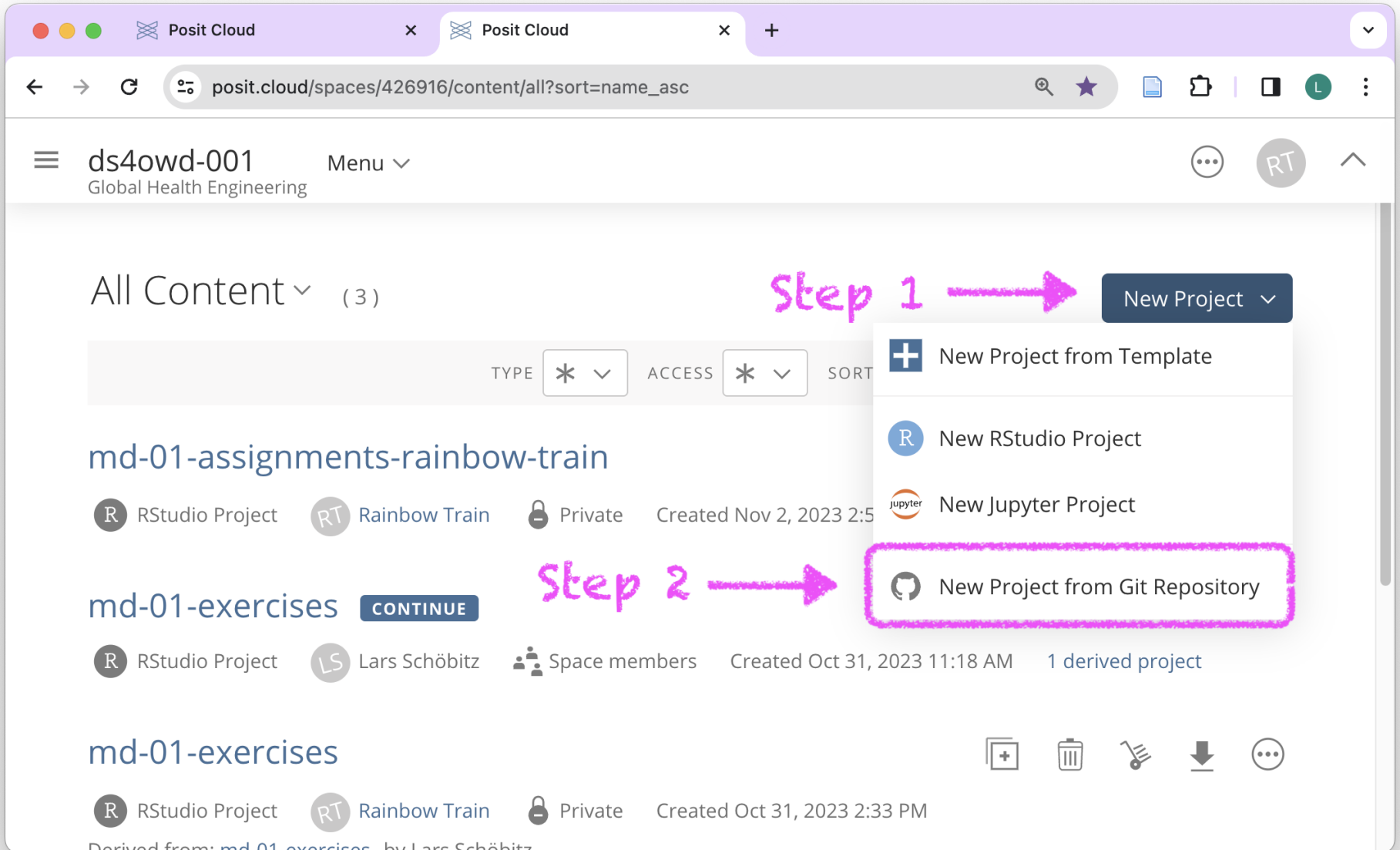
on your repository



on Posit Cloud

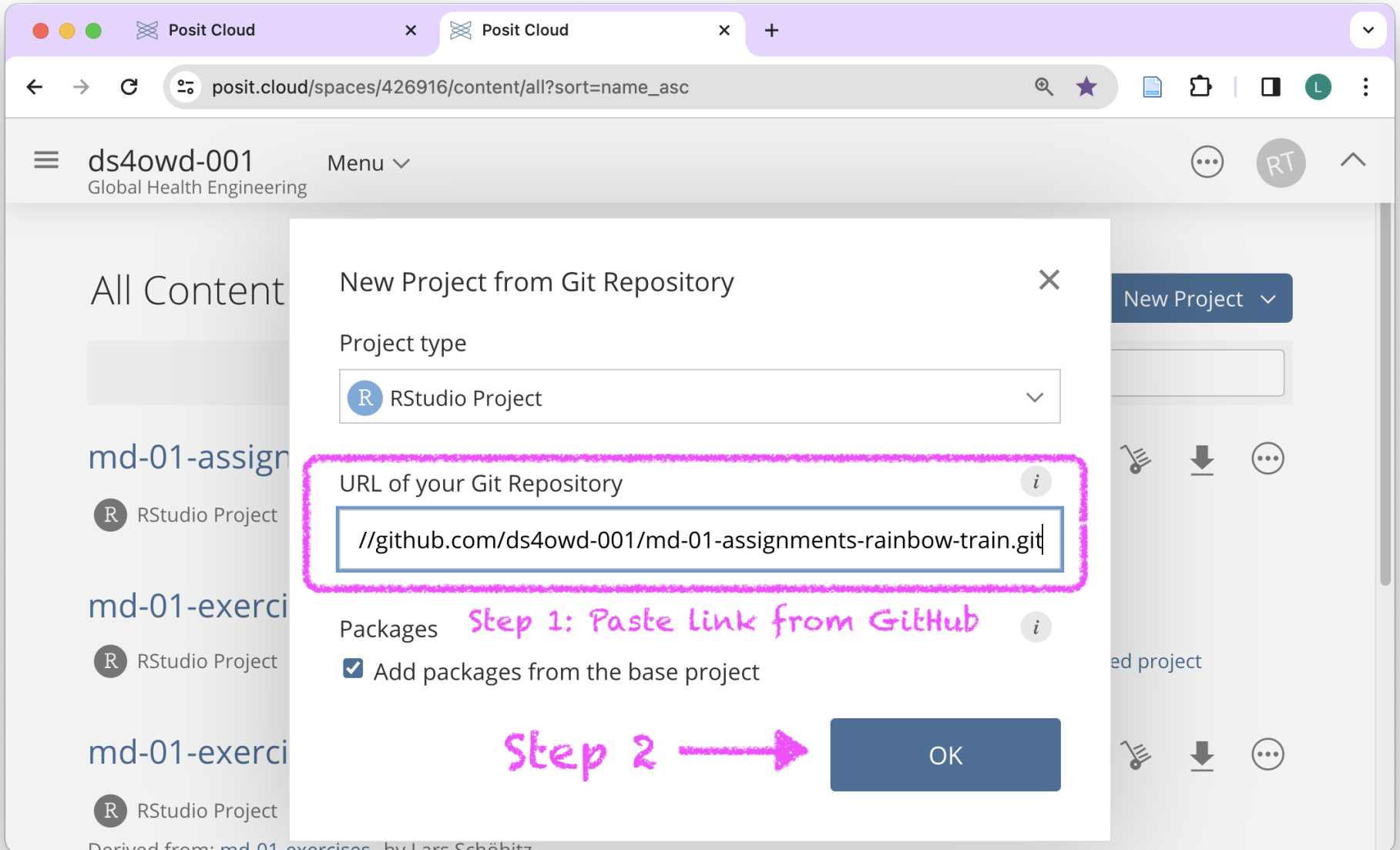
Bookmark this link in your browser!

posit.cloud/spaces/426916/content/

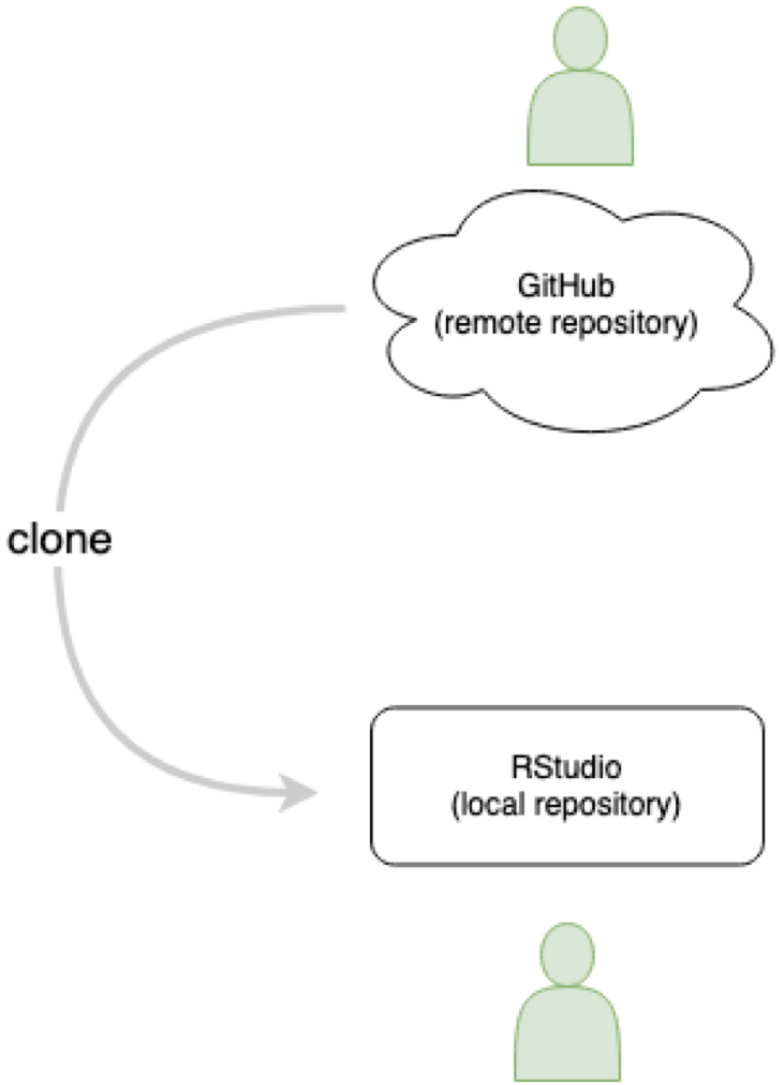

Posit Cloud
Posit Cloud
posit.cloud/spaces/426916/content/all?sort=name_asc
ds4owd-001
Global Health Engineering
Menu
RT
All Content
(3)
Step 1
New Project
New Project from Template
New RStudio Project
New Jupyter Project
New Project from Git Repository
Step 2
TYPE
ACCESS
SORT
md-01-assignments-rainbow-train
RStudio Project
Rainbow Train
Private
Created Nov 2, 2023 2:5
md-01-exercises
CONTINUE
RStudio Project
Lars Schöbitz
Space members
Created Oct 31, 2023 11:18 AM
1 derived project
md-01-exercises
RStudio Project
Rainbow Train
Private
Created Oct 31, 2023 2:33 PM

on Posit Cloud

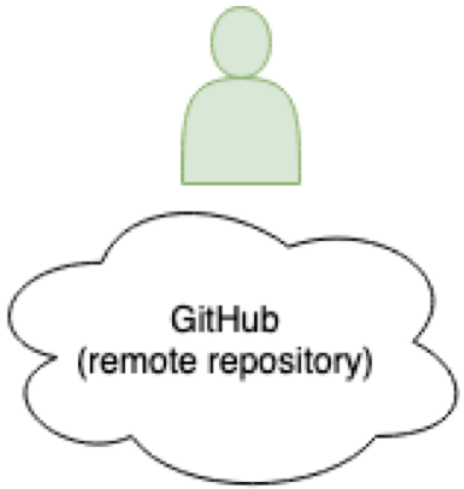
posit.cloud/spaces/426916/content/

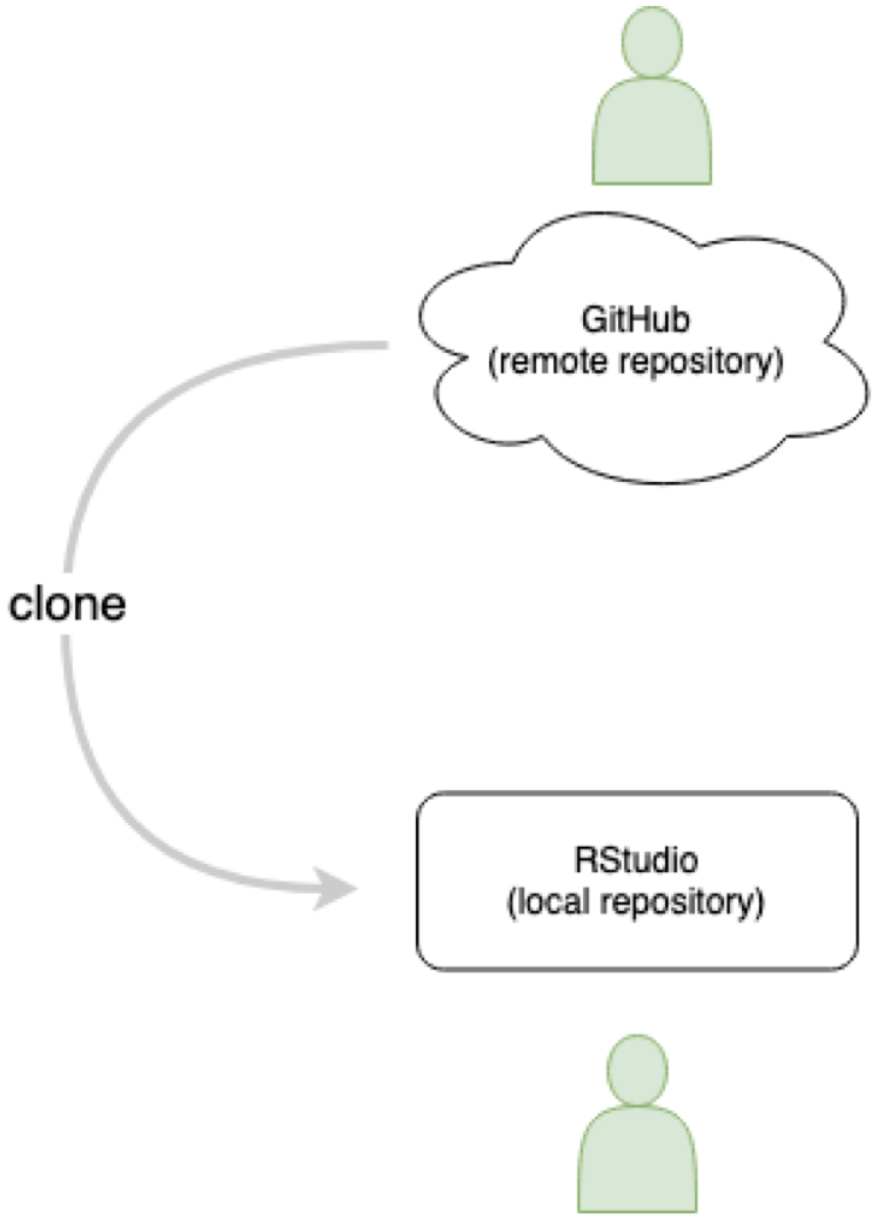

Posit Cloud
Posit Cloud
posit.cloud/spaces/426916/content/all?sort=name_asc
ds4owd-001
Global Health Engineering
Menu
RT
All Content
New Project
md-01-assign
RStudio Project
md-01-exerci
RStudio Project
md-01-exerci
RStudio Project
New Project from Git Repository
Project type
RStudio Project
URL of your Git Repository
//github.com/ds4owd-001/md-01-assignments-rainbow-train.git
Step 1: Paste link from GitHub
Packages
Add packages from the base project
Step 2
OK

command: `git clone`



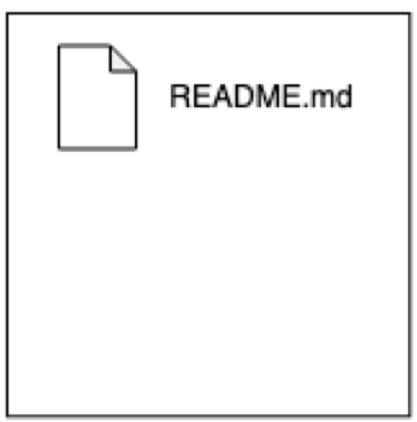
Version Control - Terminology

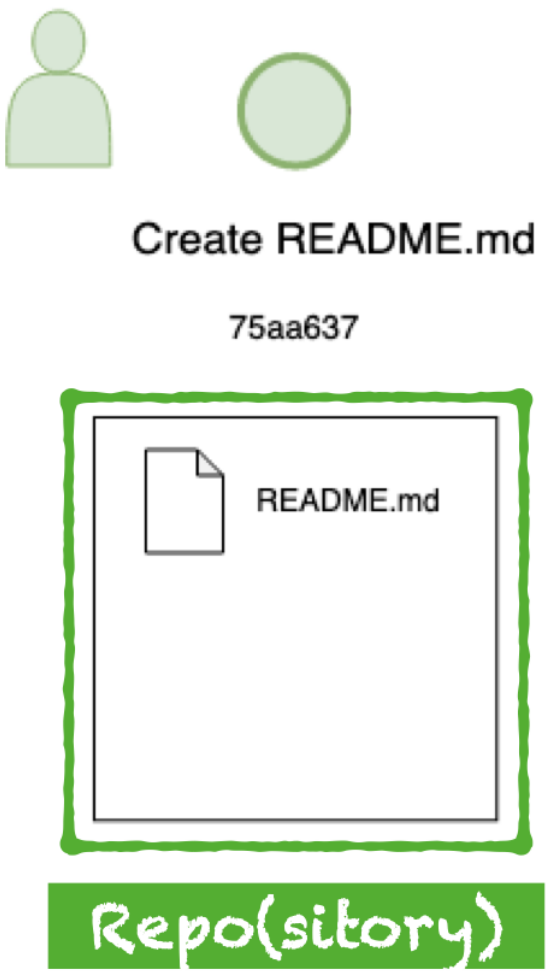

GitHub
(remote repository)

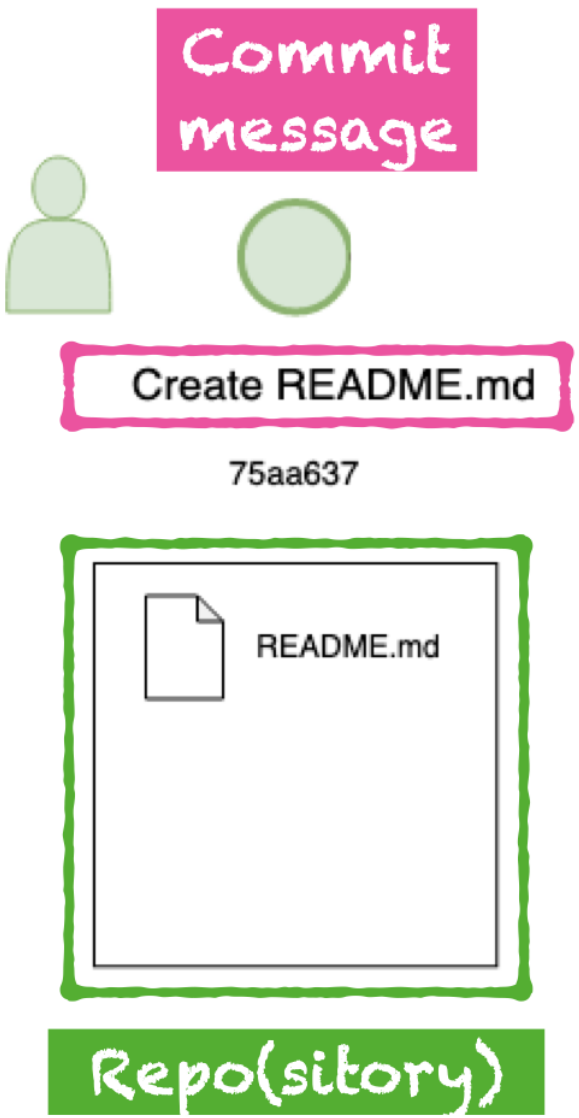

GitHub
(remote repository)
clone
RStudio
(local repository)

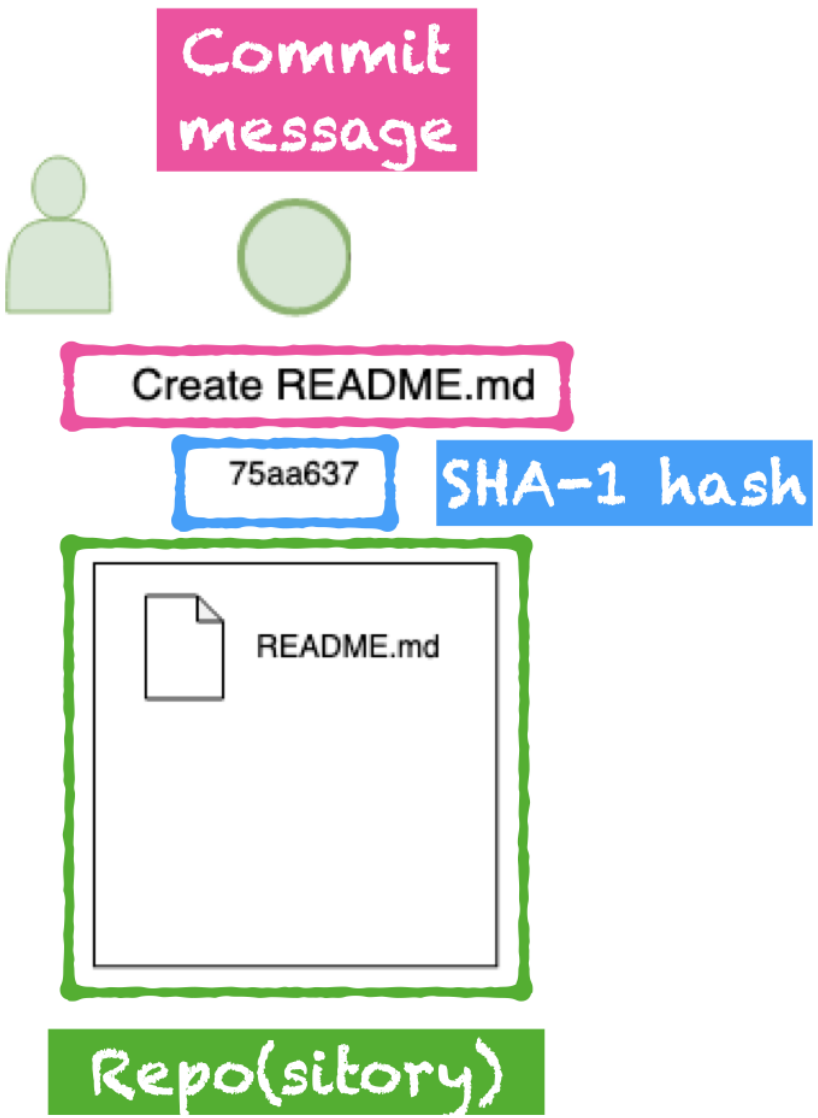
Create README.md

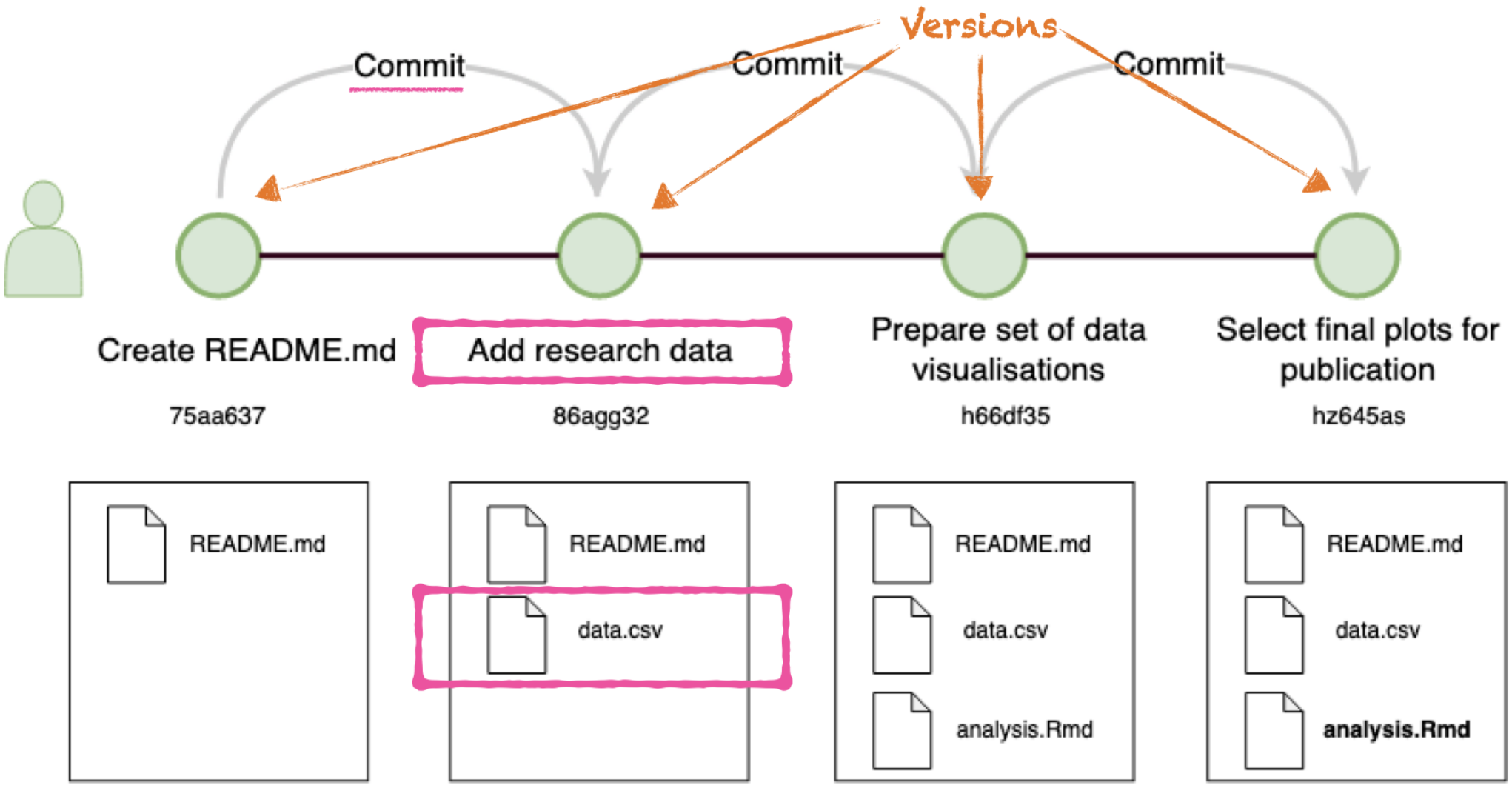
75aa637

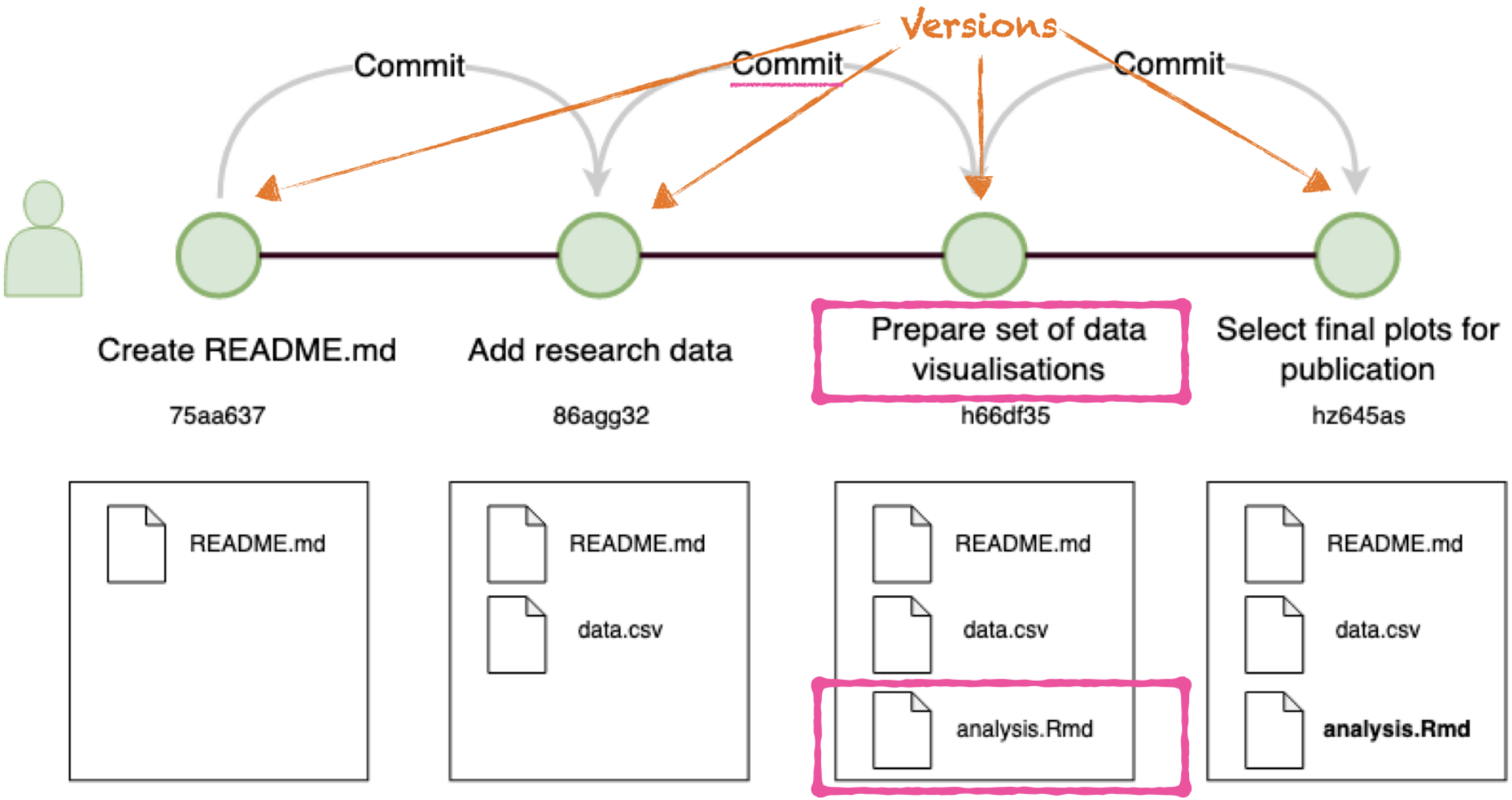


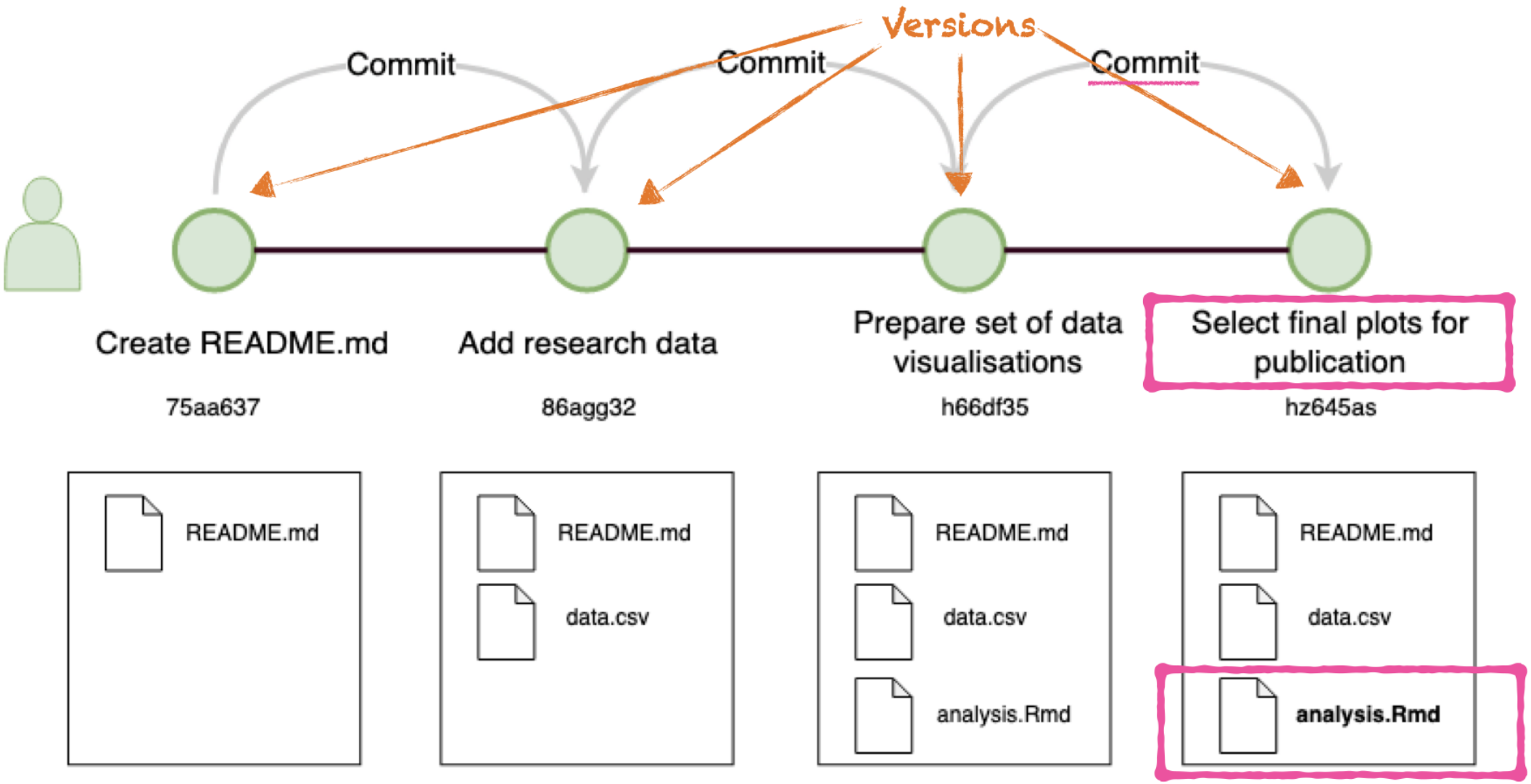

Create README.md
75aa637
README.md
Repo(sitory)

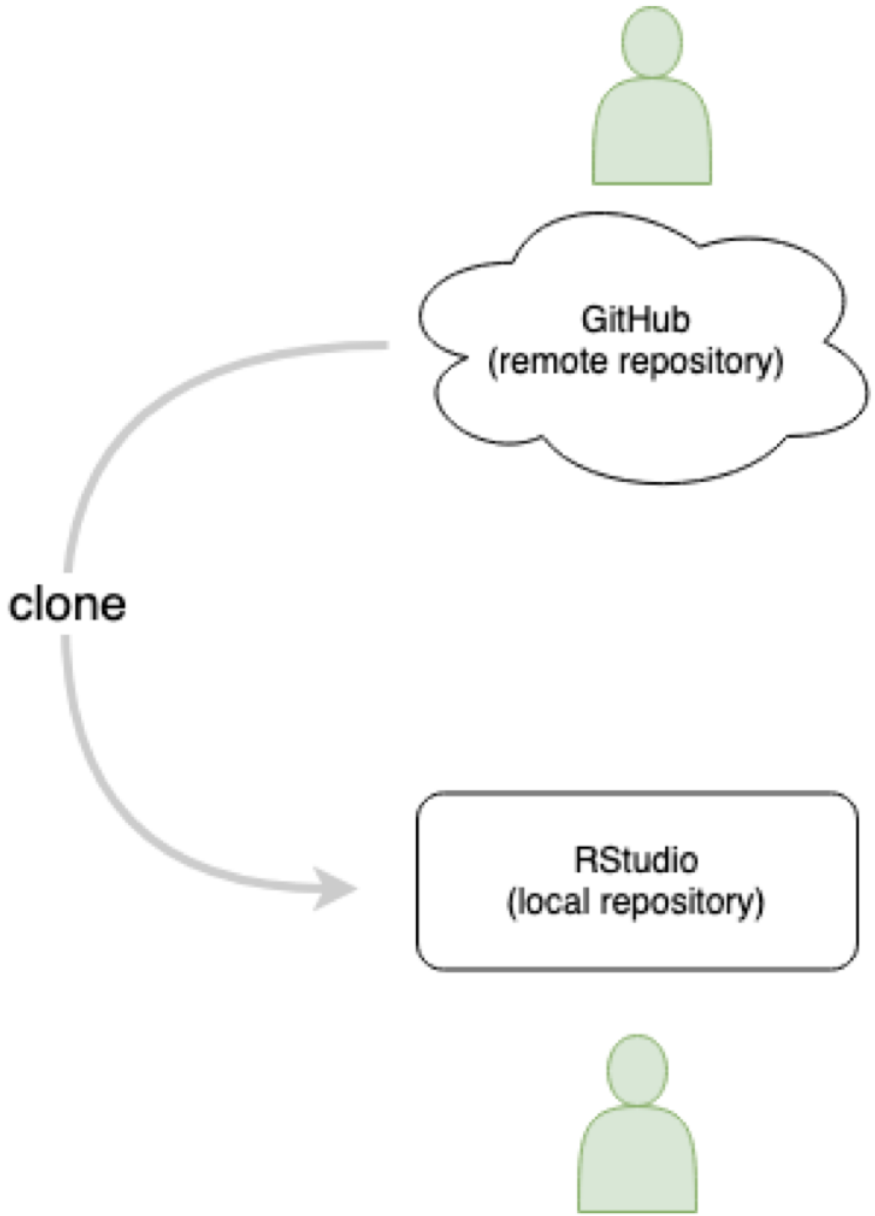

Commit message
Create README.md
75aa637
README.md
Repo(sitory)

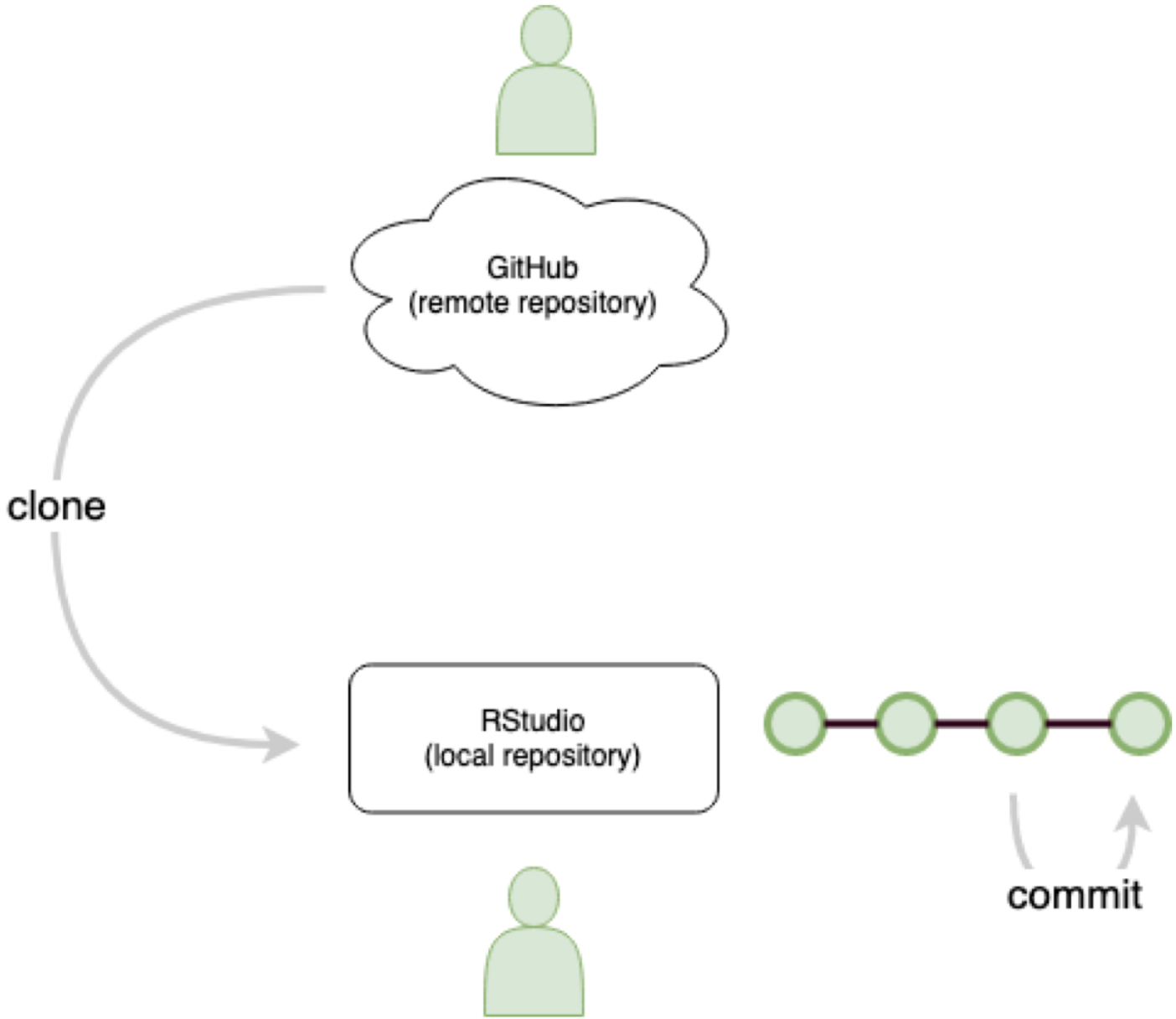

Commit message
Create README.md
75aa637
SHA-1 hash
README.md
Repo(sitory)

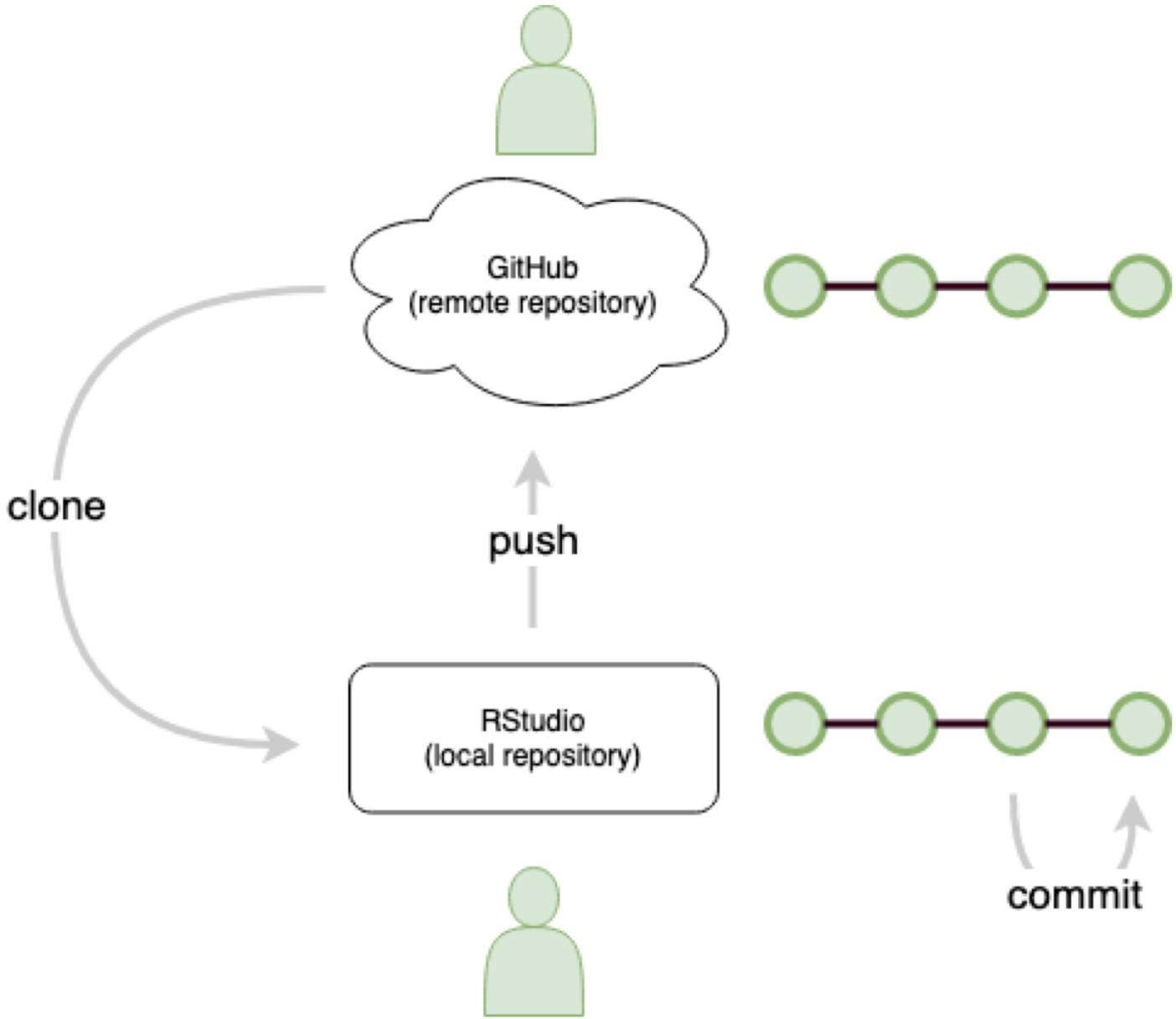

Versions
Commit
Commit
Commit
Create README.md
Add research data
Prepare set of data visualisations
Select final plots for publication
75aa637
86agg32
h66df35
hz645as
README.md
README.md
data.csv
README.md
data.csv
analysis.Rmd
README.md
data.csv
analysis.Rmd


Versions
Commit
Commit
Commit
Create README.md
75aa637
Add research data
86agg32
Prepare set of data visualisations
h66df35
Select final plots for publication
hz645as
README.md
README.md
data.csv
README.md
data.csv
analysis.Rmd
README.md
data.csv
analysis.Rmd

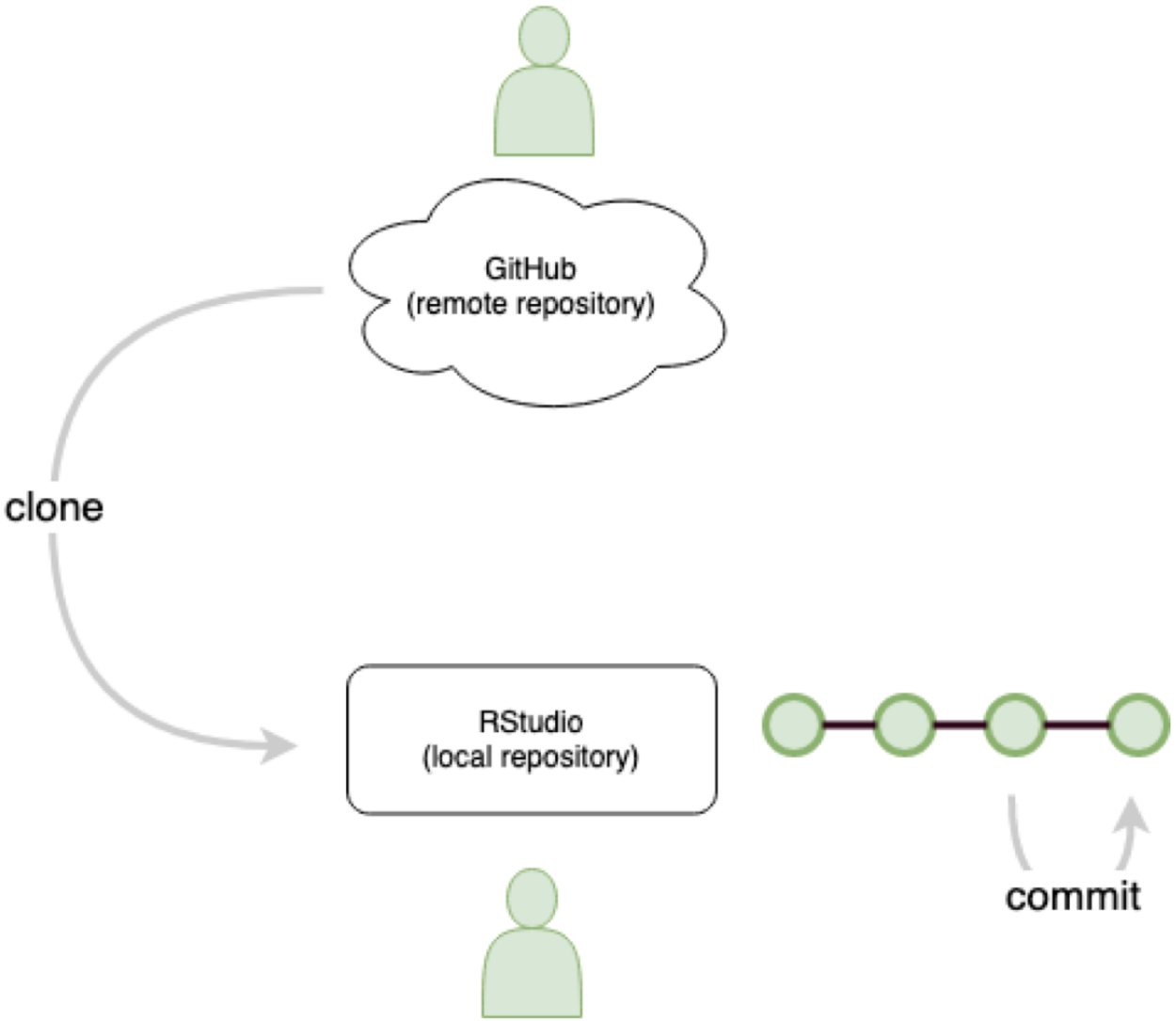

Versions
Commit
Commit
Commit
Create README.md
75aa637
Add research data
86agg32
Prepare set of data visualisations
h66df35
Select final plots for publication
hz645as
README.md
README.md
data.csv
README.md
data.csv
analysis.Rmd
README.md
data.csv
analysis.Rmd


GitHub
(remote repository)
clone
RStudio
(local repository)

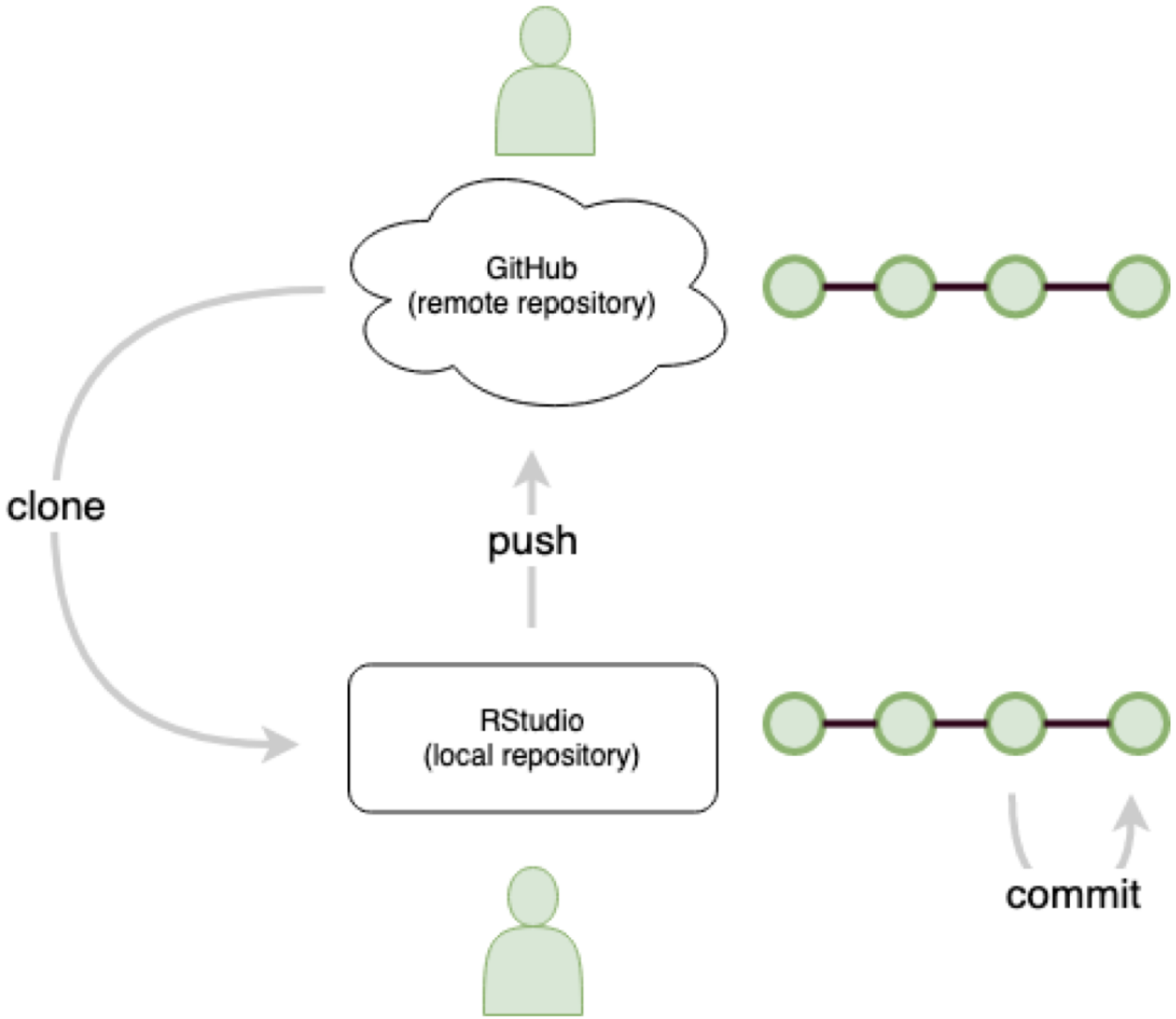

GitHub
(remote repository)
clone
RStudio
(local repository)
commit


GitHub
(remote repository)
clone
push
RStudio
(local repository)
commit

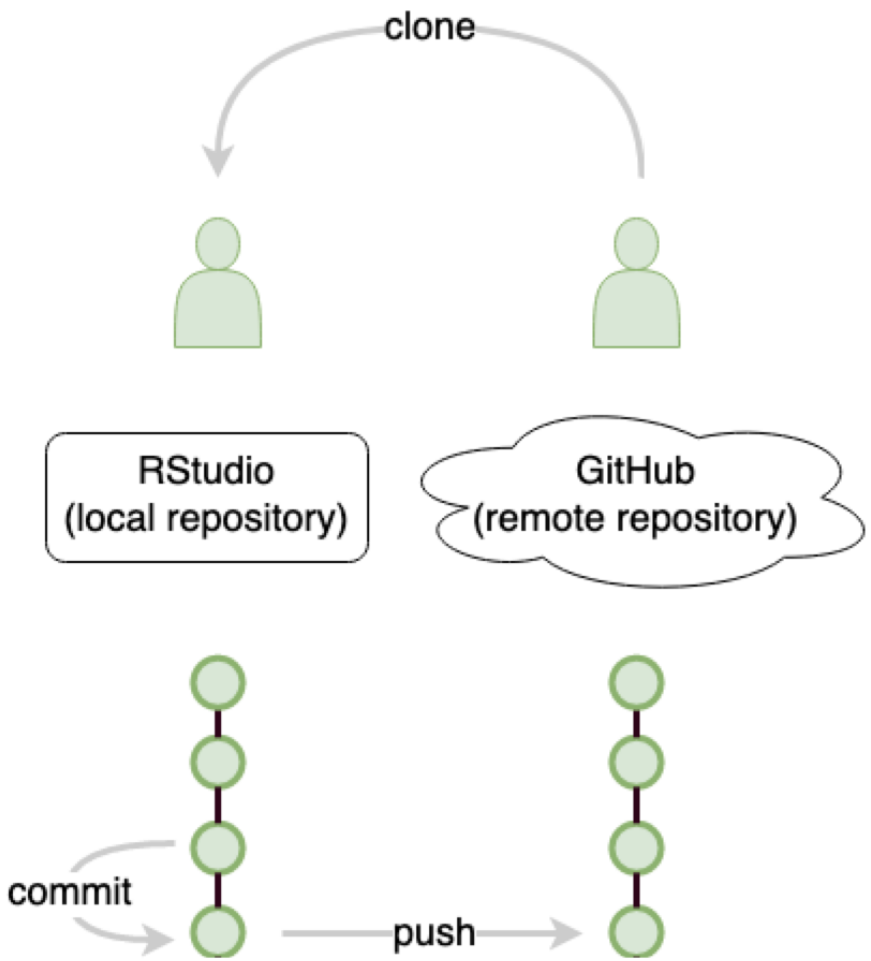
remember: `git commit`



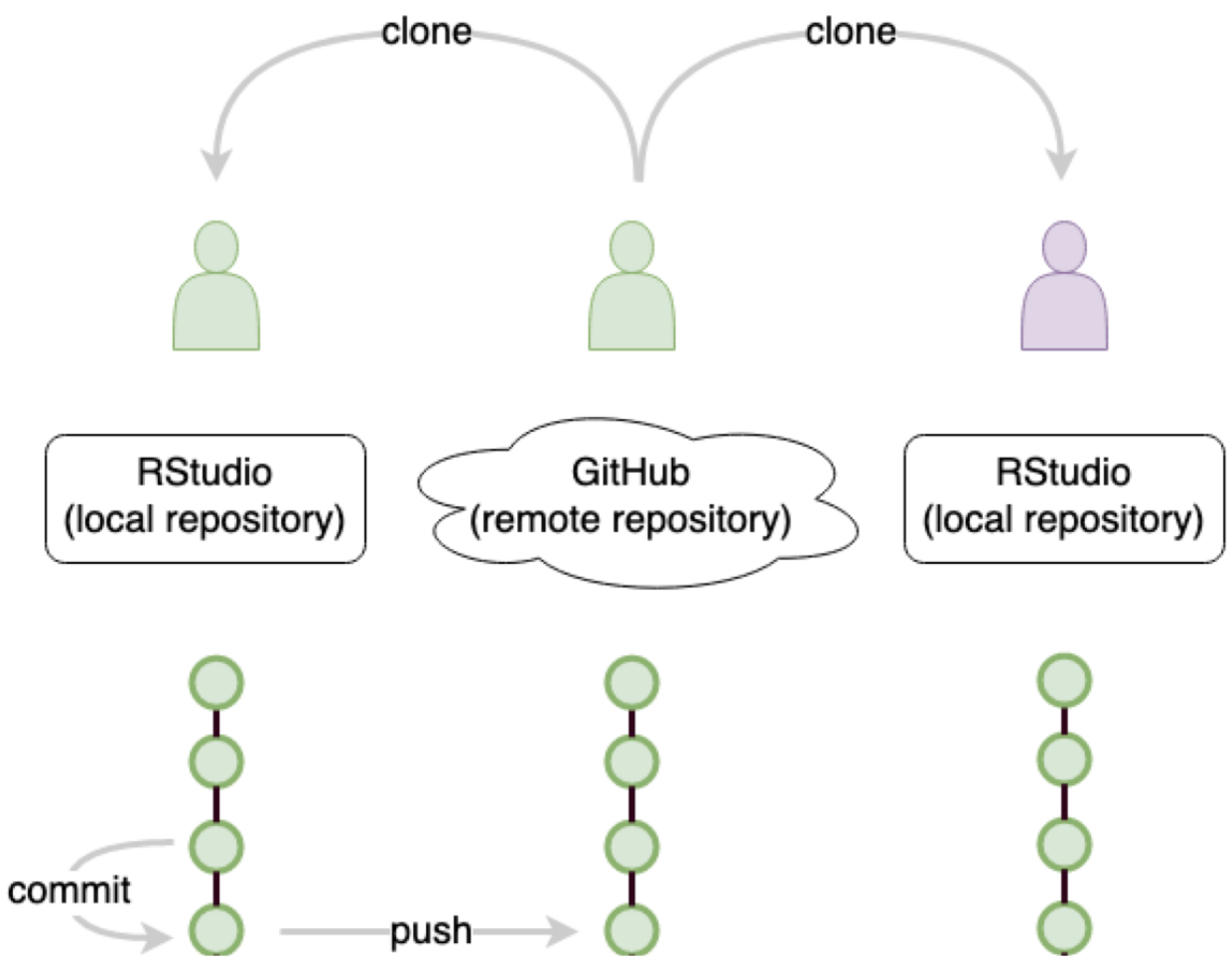
remember: `git push`



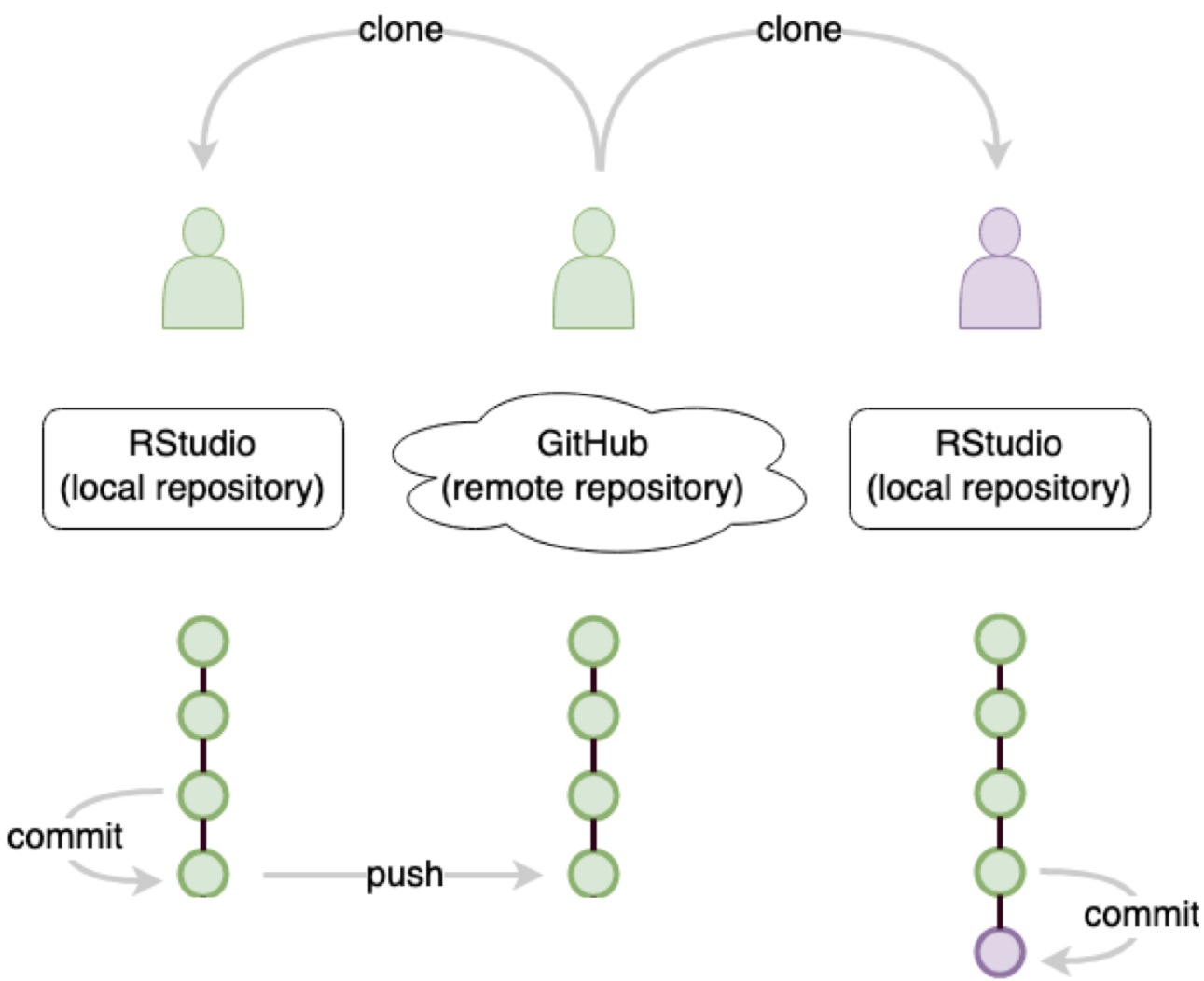
remember: `git push`



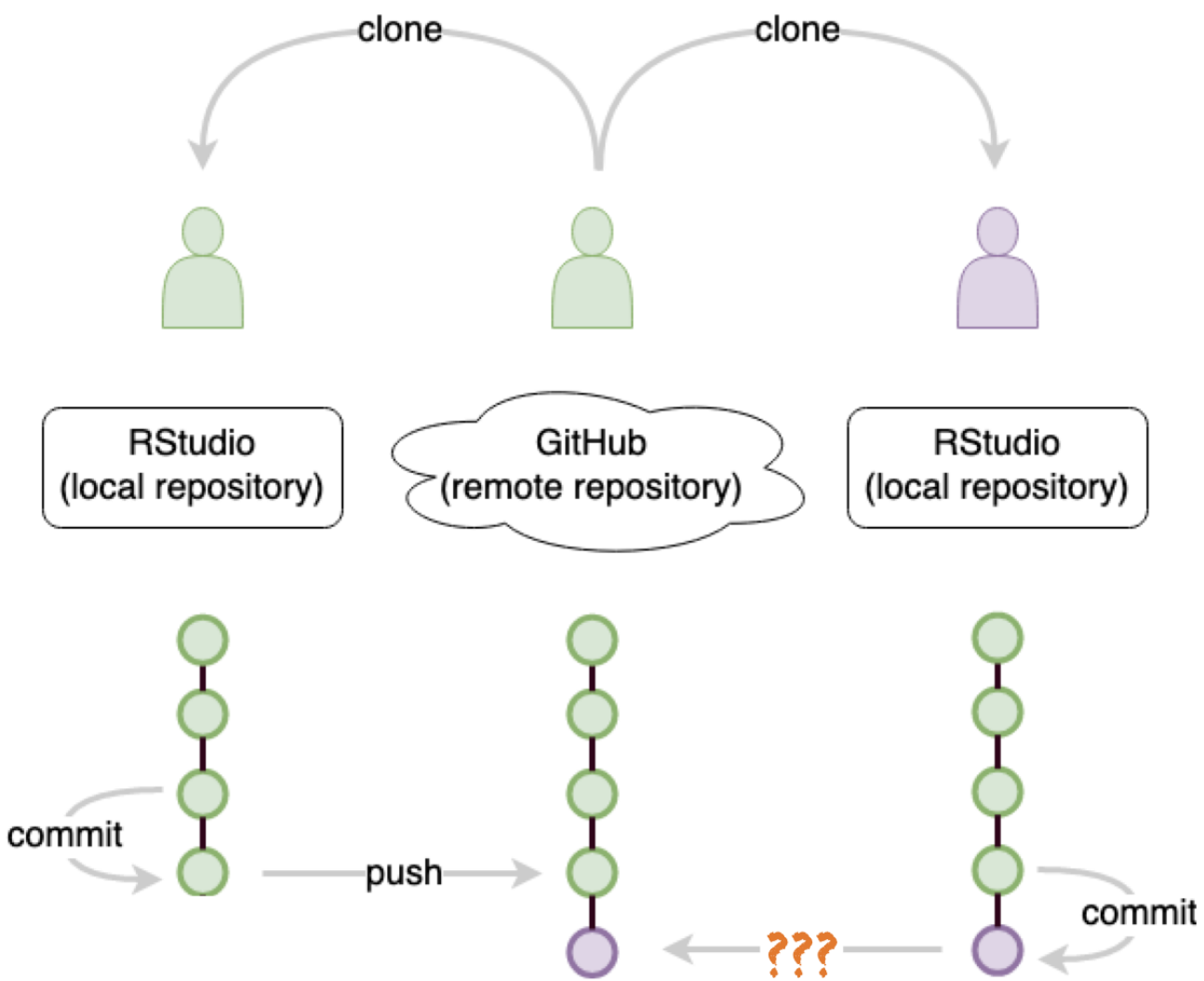
collaborate: `git clone`



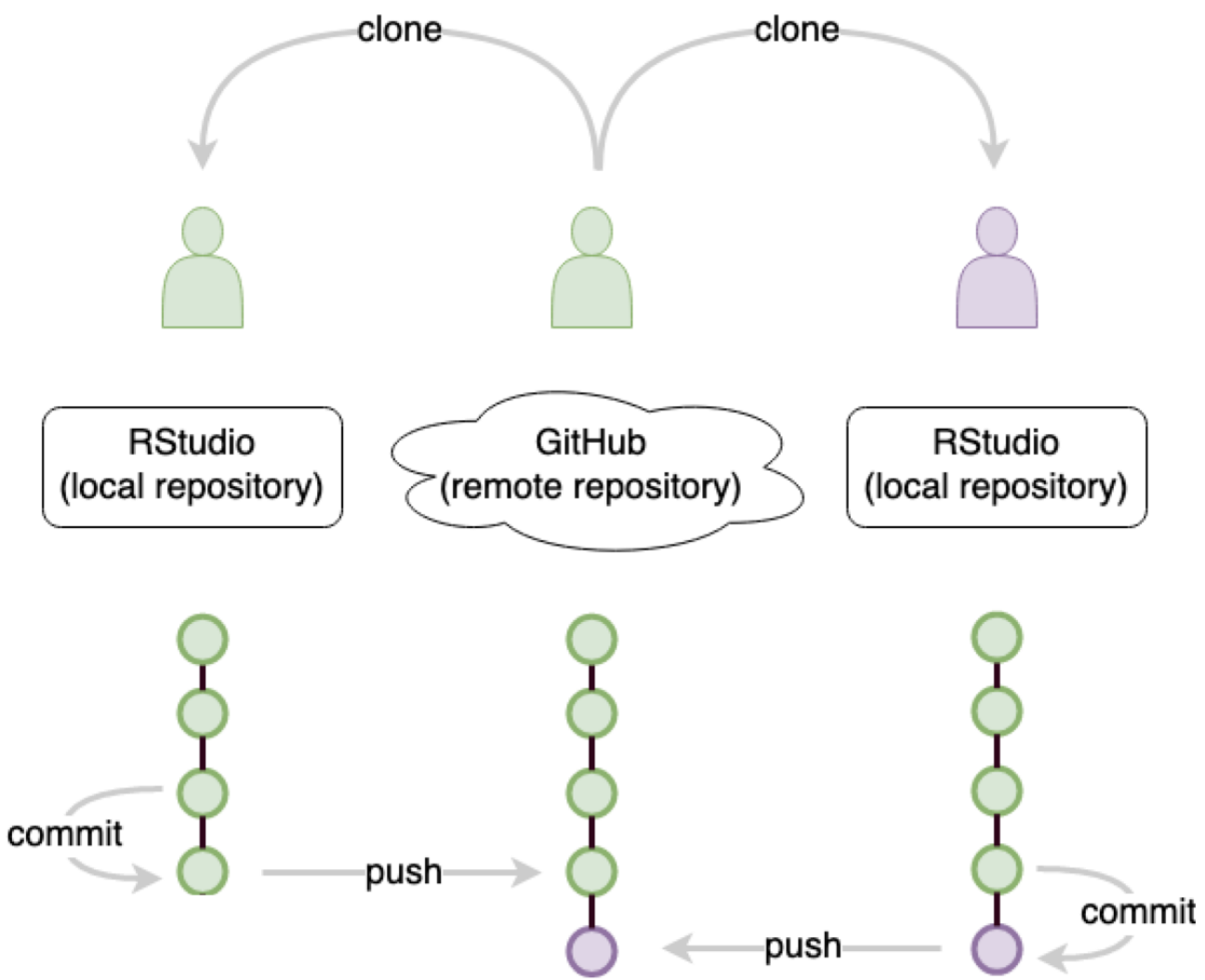
track work: `git commit`



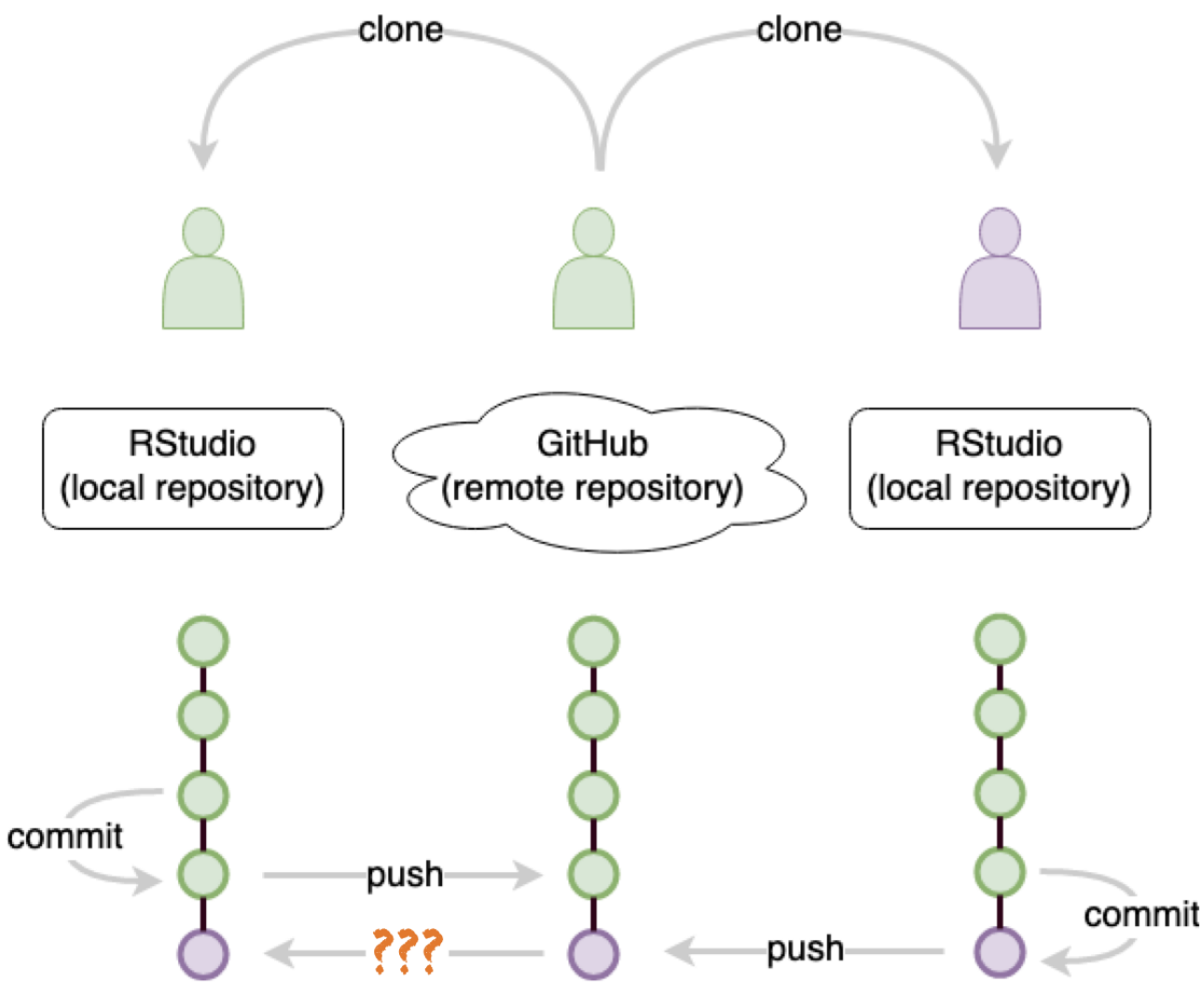
update: `git ???`


clone
clone
RStudio
(local repository)
GitHub
(remote repository)
RStudio
(local repository)
commit
push
???
commit

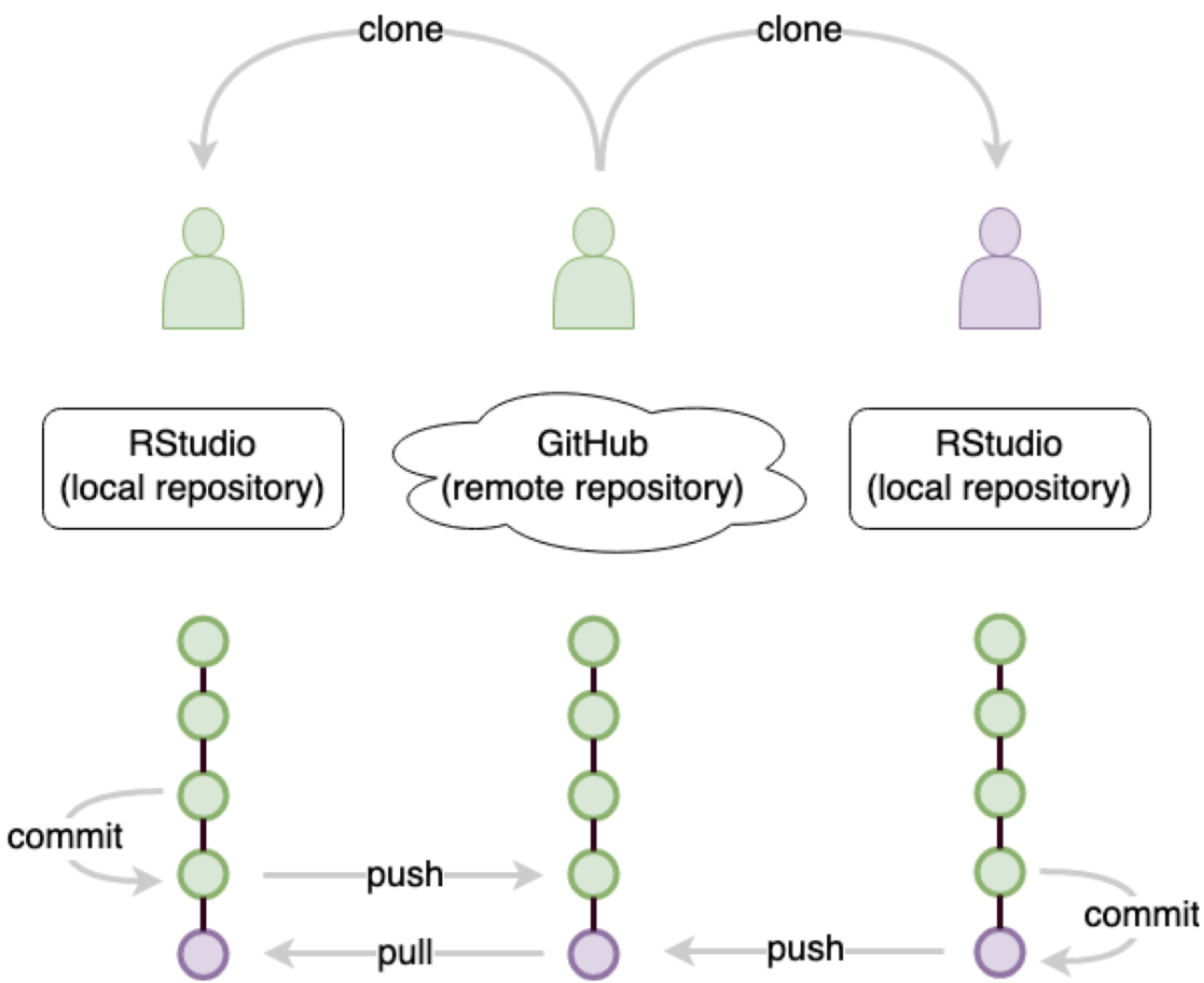
update: `git push`



git ???


clone
clone
RStudio
(local repository)
GitHub
(remote repository)
RStudio
(local repository)
commit
push
???
push
commit

new: `git pull`


clone
clone
RStudio
(local repository)
GitHub
(remote repository)
RStudio
(local repository)
commit
push
pull
push
commit

Learning Objectives (for this week)

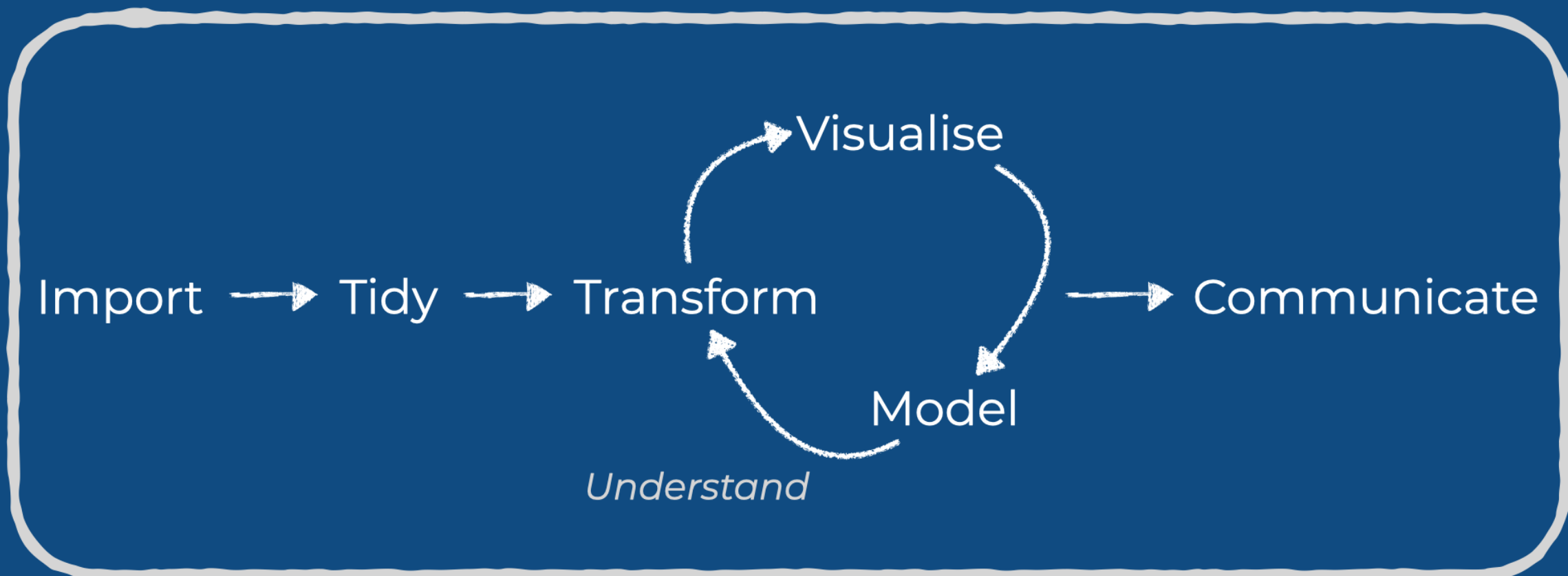
1. Learners can list the six elements of the data science lifecycle.
2. Learners can describe the four main aesthetic mappings that can be used to visualise data using the ggplot2 R Package.
3. Learners can control the colour scaling applied to a plot using colour as an aesthetic mapping.
4. Learners can compare three different geoms (bar/col, histogram, point) and their use case.

Data Science Lifecycle

Deep End

via GIPHY

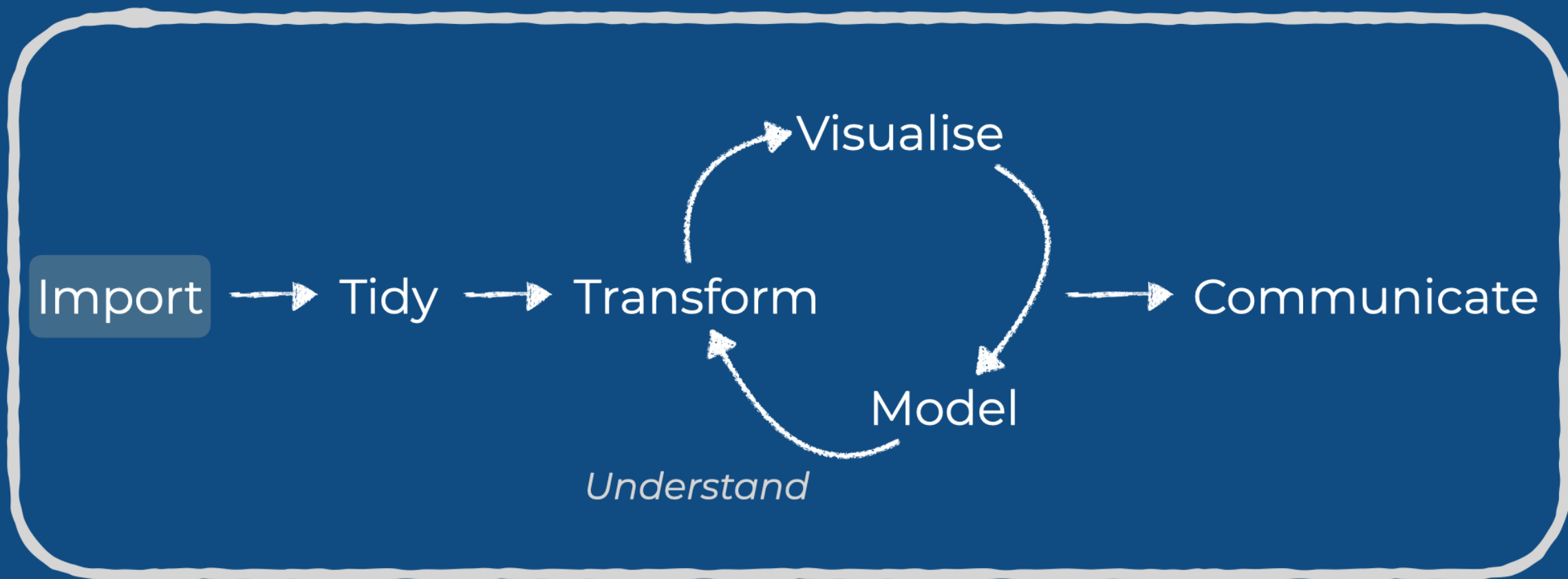
Data Science Lifecycle



From: https://r4ds.had.co.nz/

Data Science Lifecycle

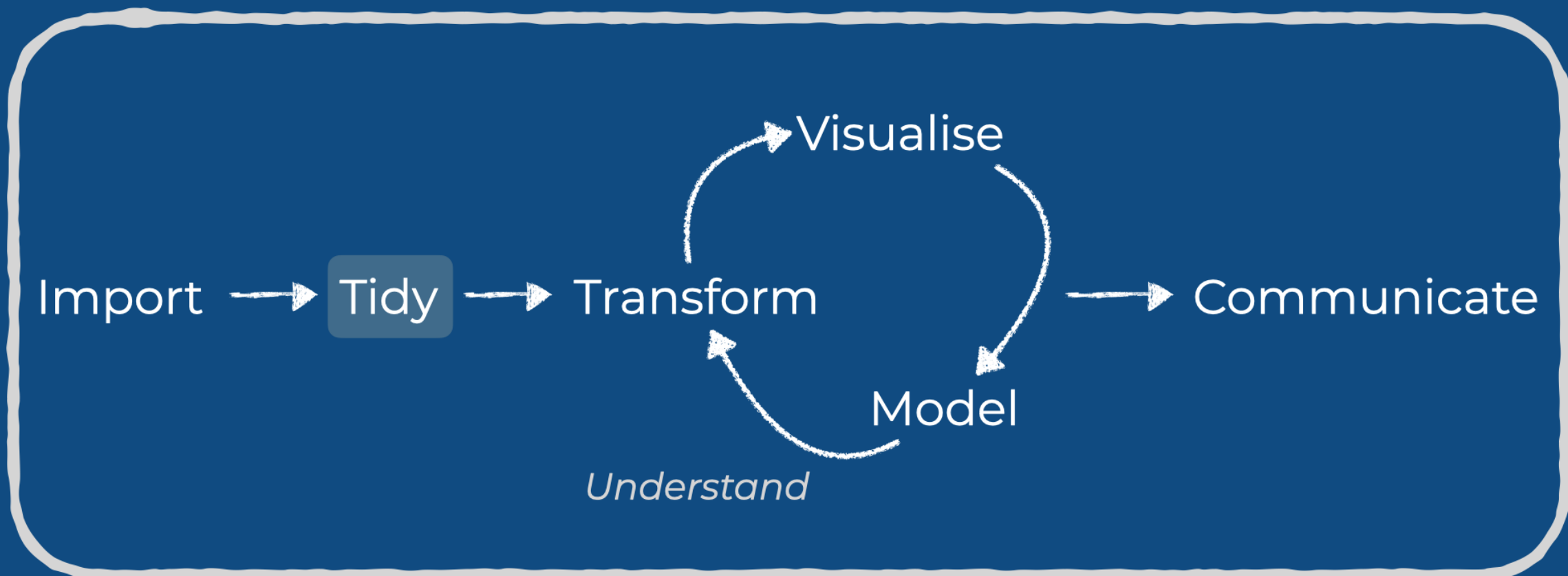
Get your data into R



From: https://r4ds.had.co.nz/

Data Science Lifecycle

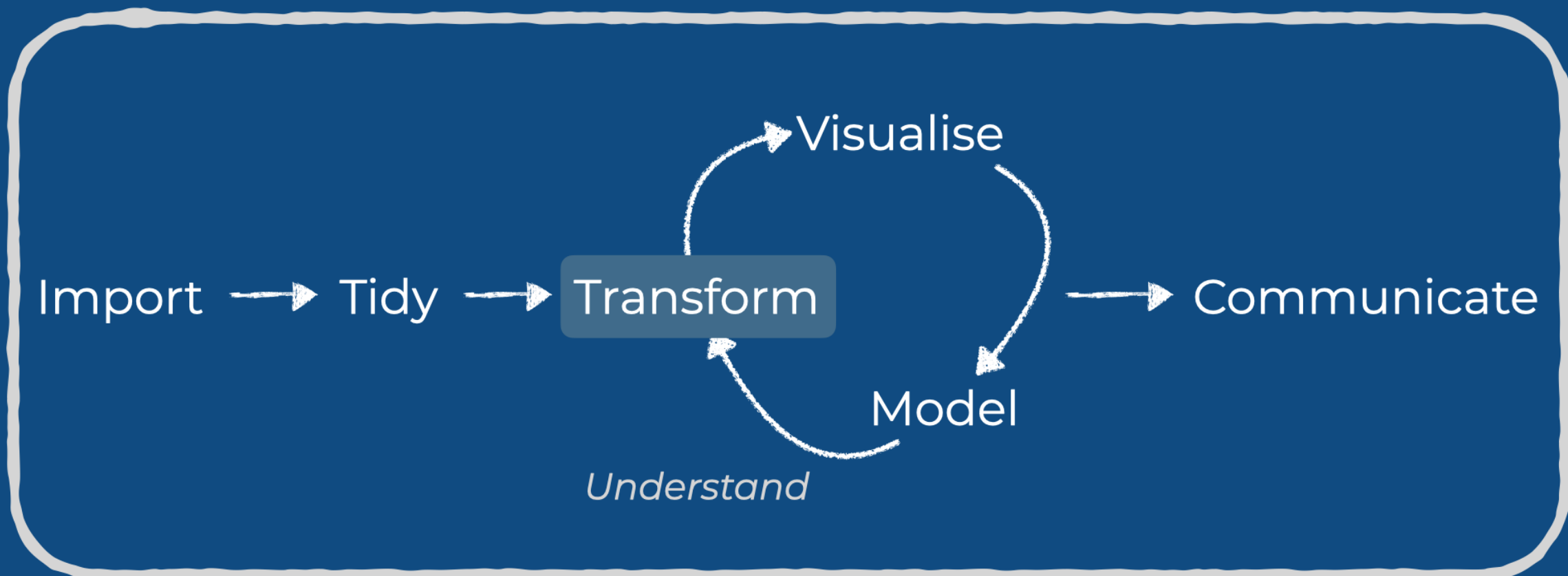
Store your data in a consistent form



From: https://r4ds.had.co.nz/

Data Science Lifecycle

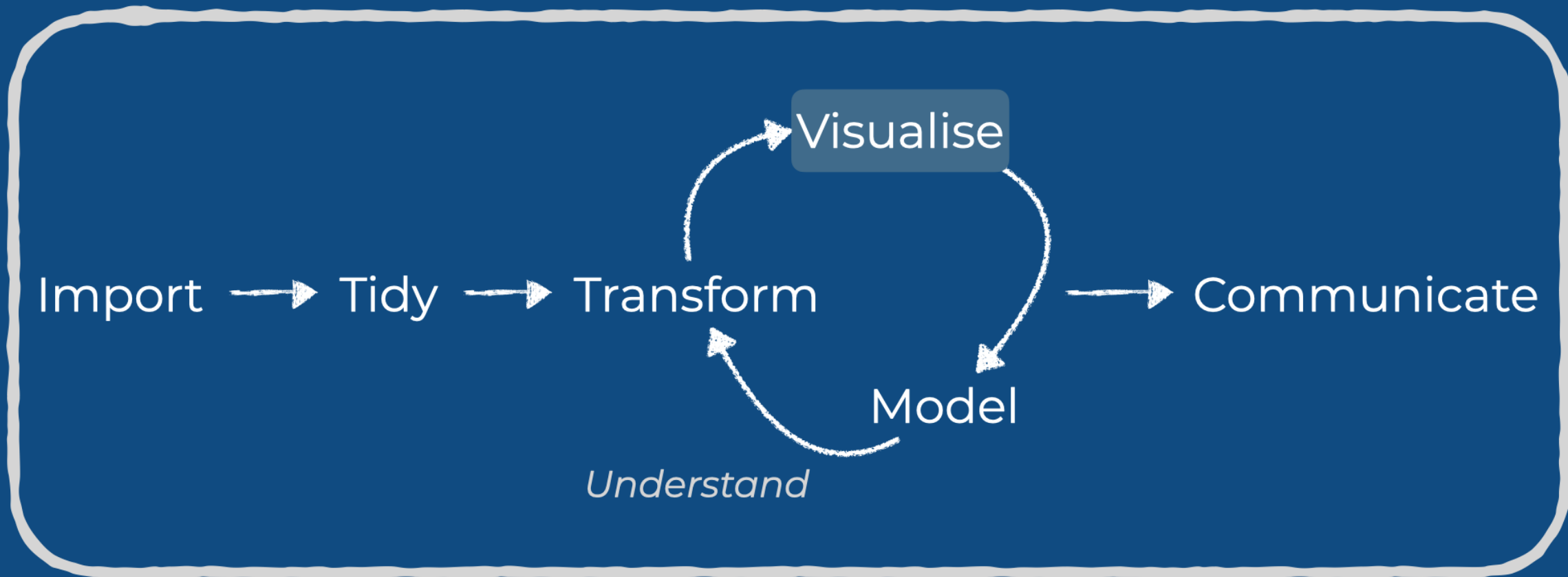
Narrow down + Create new variables + Summary stats



From: https://r4ds.had.co.nz/

Data Science Lifecycle

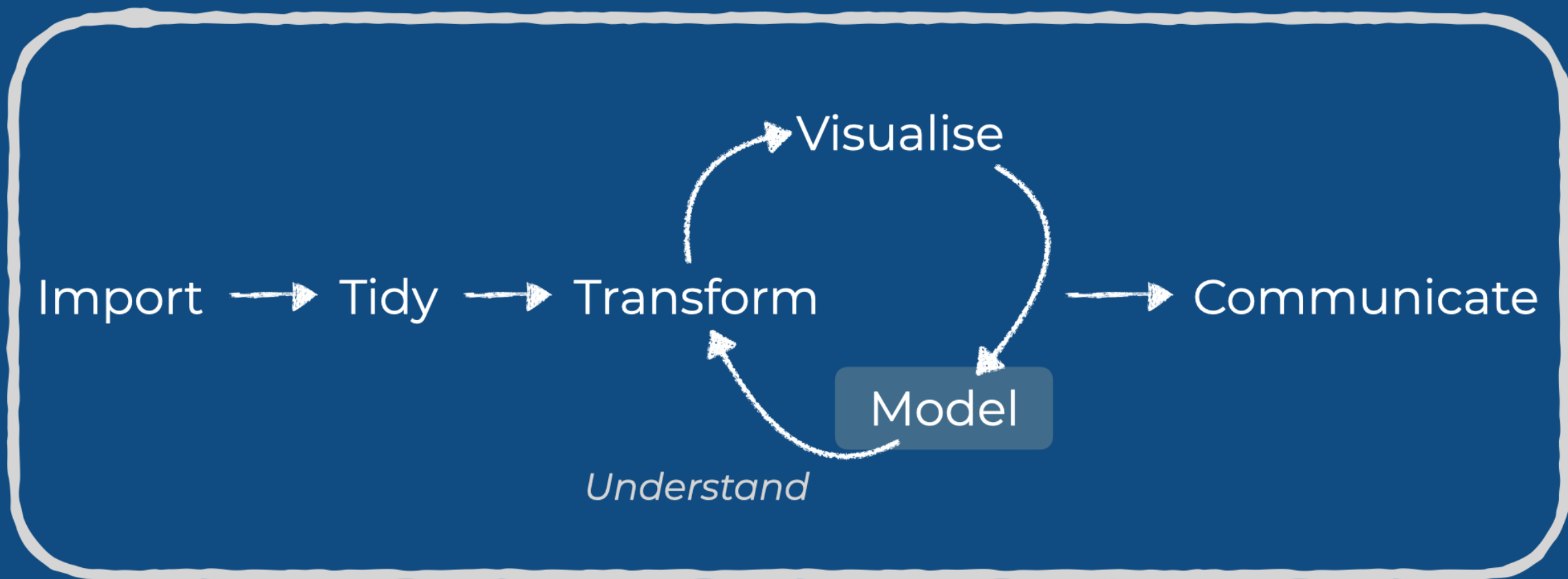
Explore your with visual representations



From: https://r4ds.had.co.nz/

Data Science Lifecycle

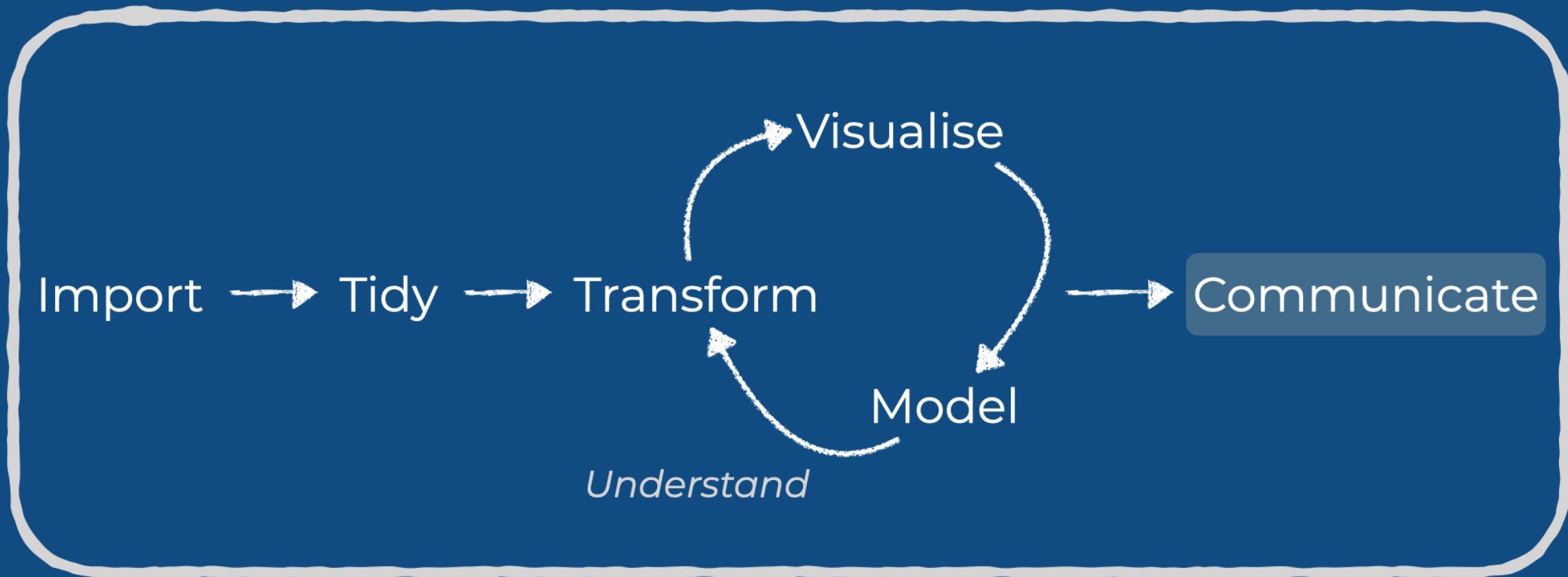
Explore your with visual representations



From: https://r4ds.had.co.nz/

Data Science Lifecycle

Share your findings with others

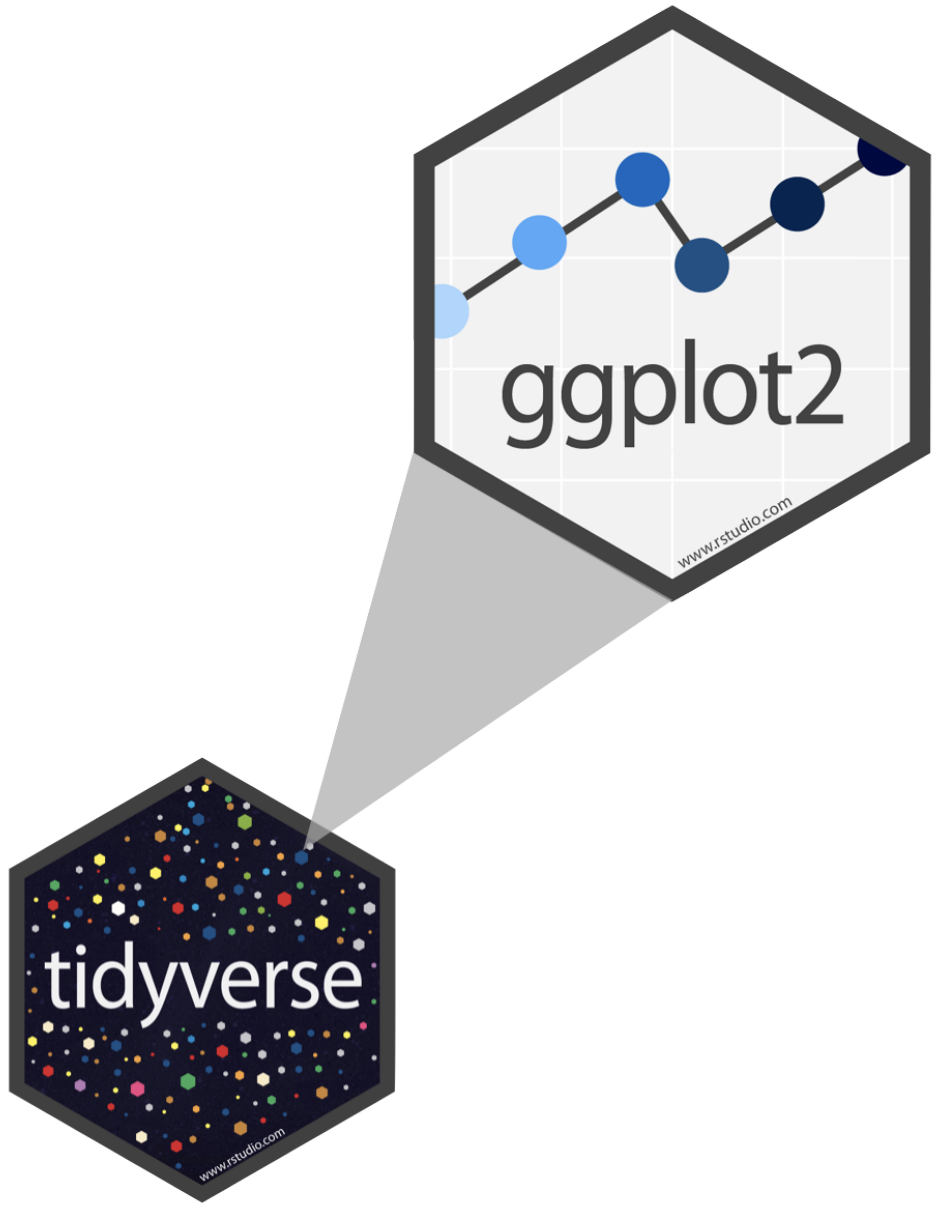


From: https://r4ds.had.co.nz/

Exploratory Data Analysis with `ggplot2`

R Package `ggplot2`

- **ggplot2** is tidyverse's data visualization package
- `gg` in `ggplot2` stands for Grammar of Graphics
- Inspired by the book **Grammar of Graphics** by Leland Wilkinson
- **Documentation:** https://ggplot2.tidyverse.org/
- **Book**: https://ggplot2-book.org



My turn: Working with R

Sit back and enjoy!

Code structure

- `ggplot()` is the main function in ggplot2
- Plots are constructed in layers
- Structure of the code for plots can be summarized as

```
ggplot(data = [dataset],
       mapping = aes(x = [x-variable],
                     y = [y-variable])) +
  geom_xxx() +
  other options
```

Code structure

```
1  ggplot()
```

Code structure

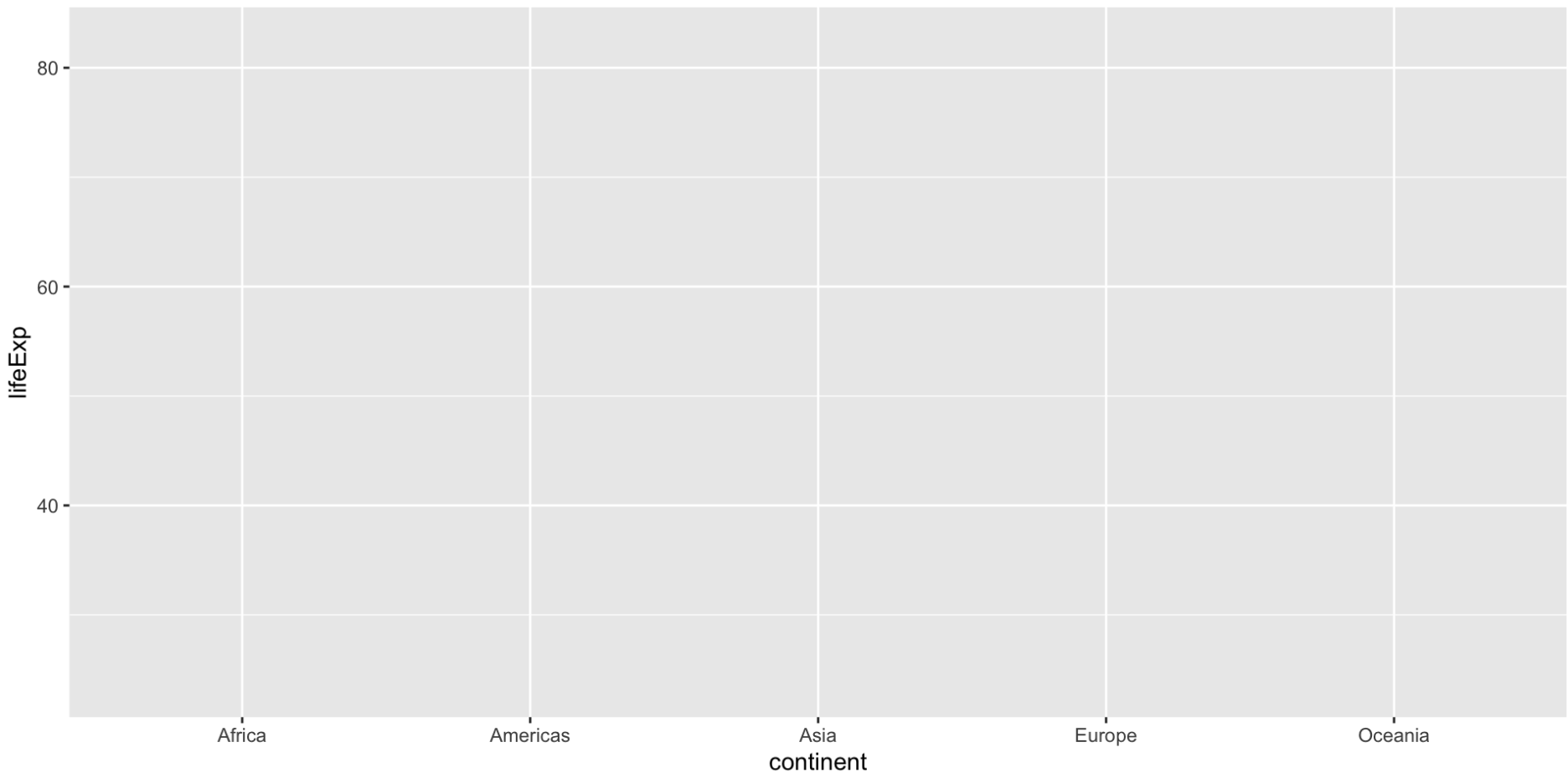
```
1  ggplot(data = gapminder)
```

Code structure

```
ggplot(data = gapminder,
       mapping = aes())
```

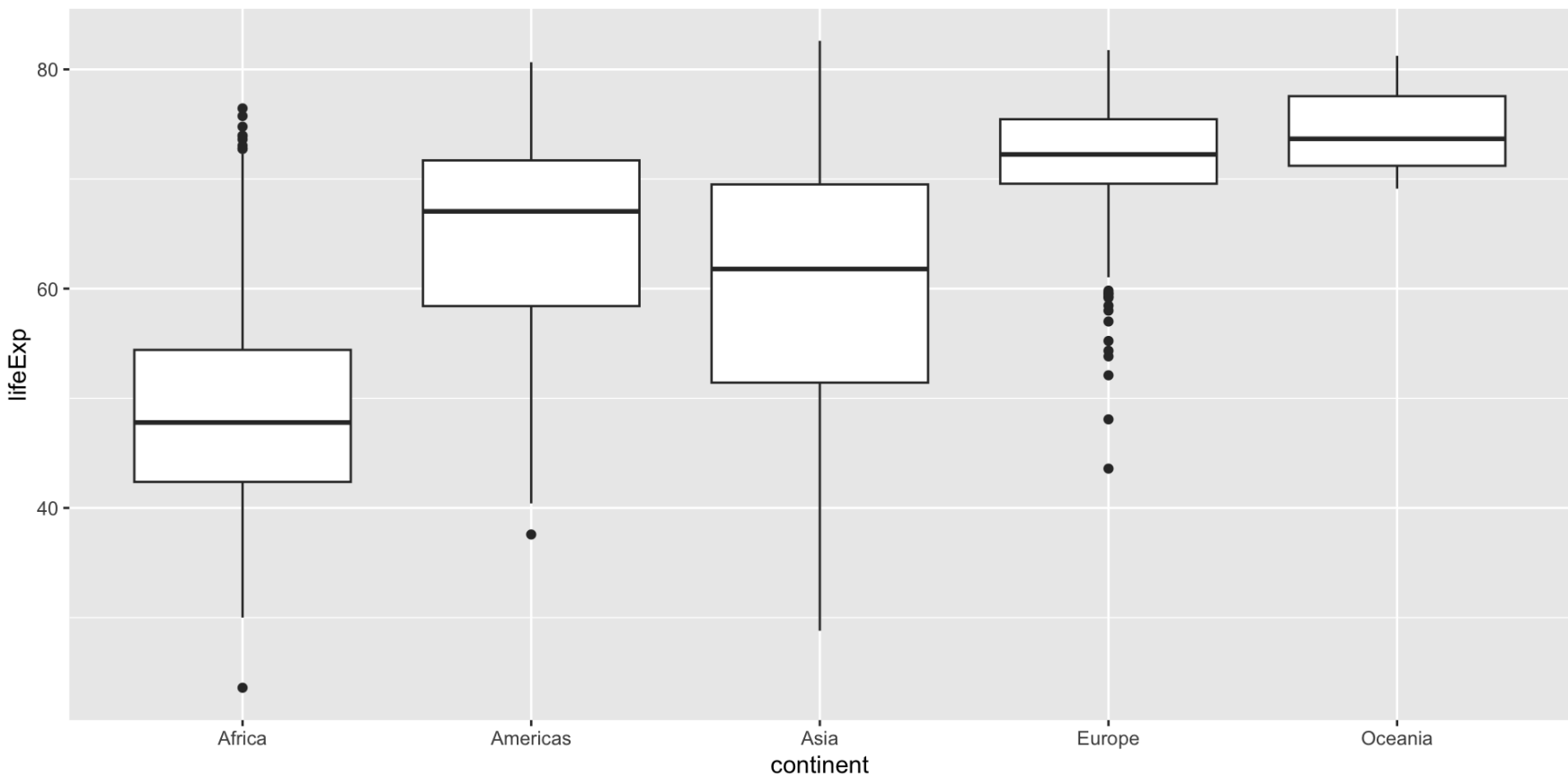
Code structure

```
1  ggplot(data = gapminder,
2         mapping = aes(x = continent,
3                       y = lifeExp))
```



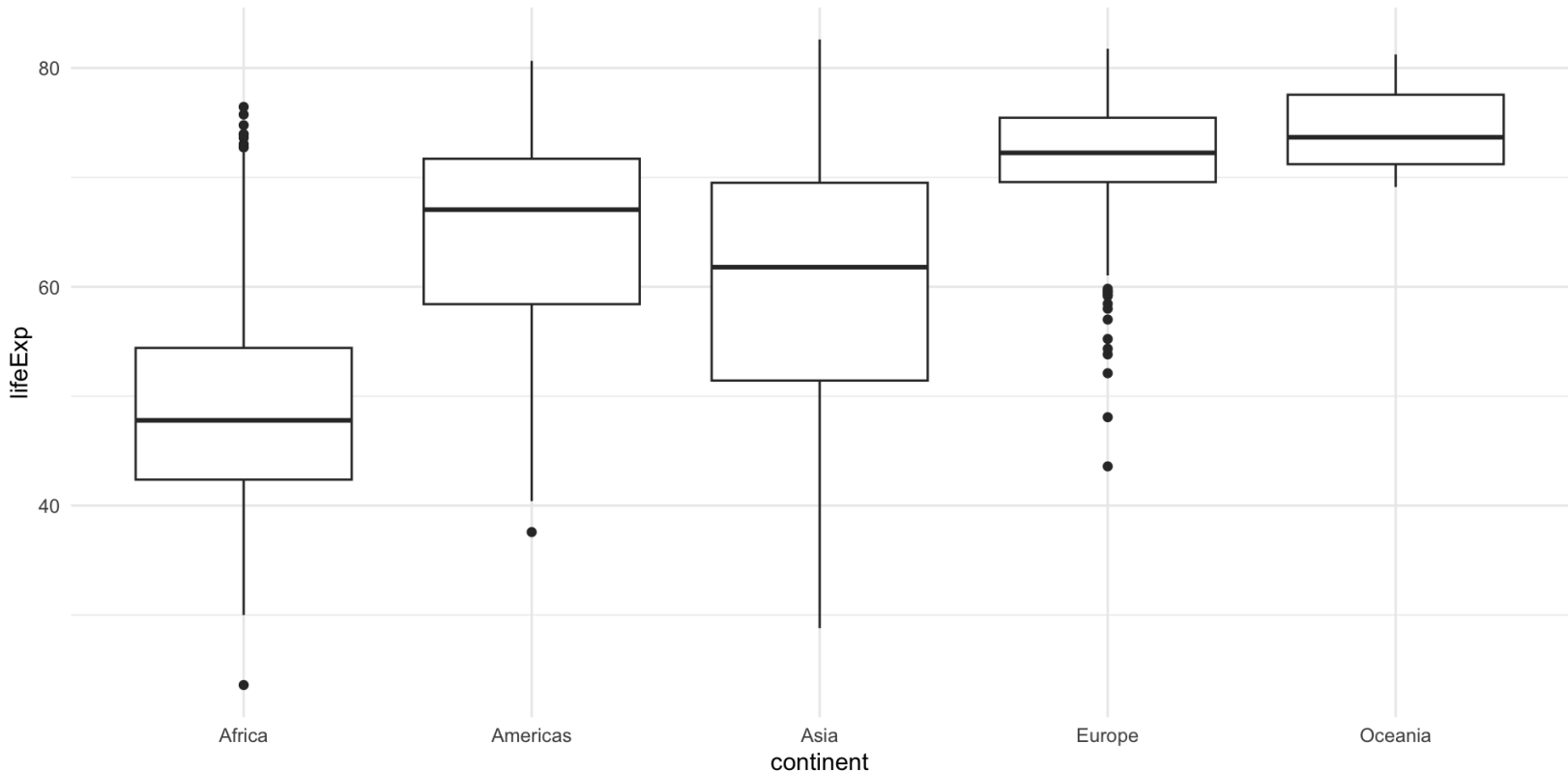
Code structure

```
1  ggplot(data = gapminder,
2         mapping = aes(x = continent,
3                       y = lifeExp)) +
4    geom_boxplot()
```



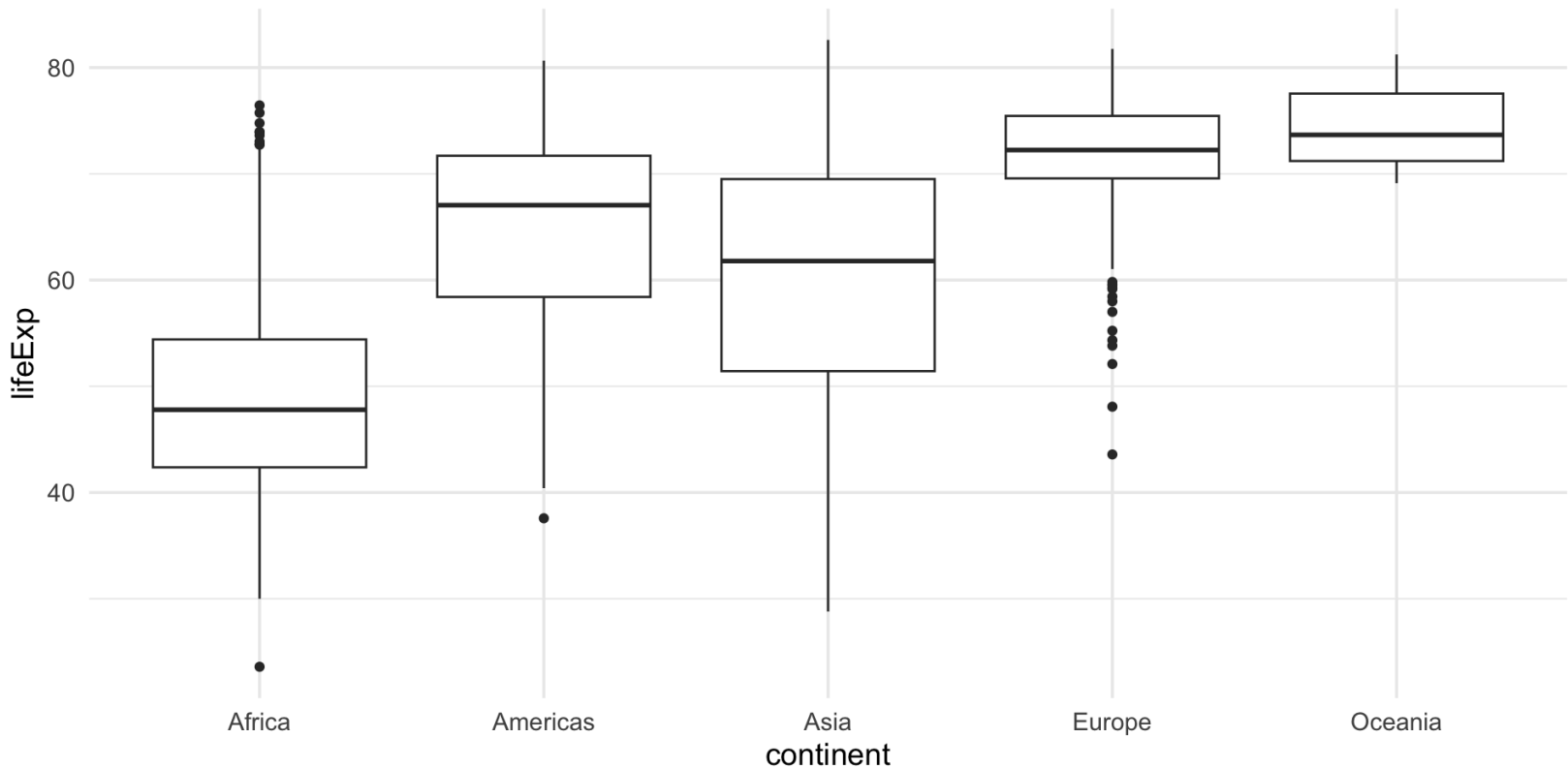
Code structure

```
ggplot(data = gapminder,
       mapping = aes(x = continent,
                     y = lifeExp)) +
  geom_boxplot() +
  theme_minimal()
```



Code structure

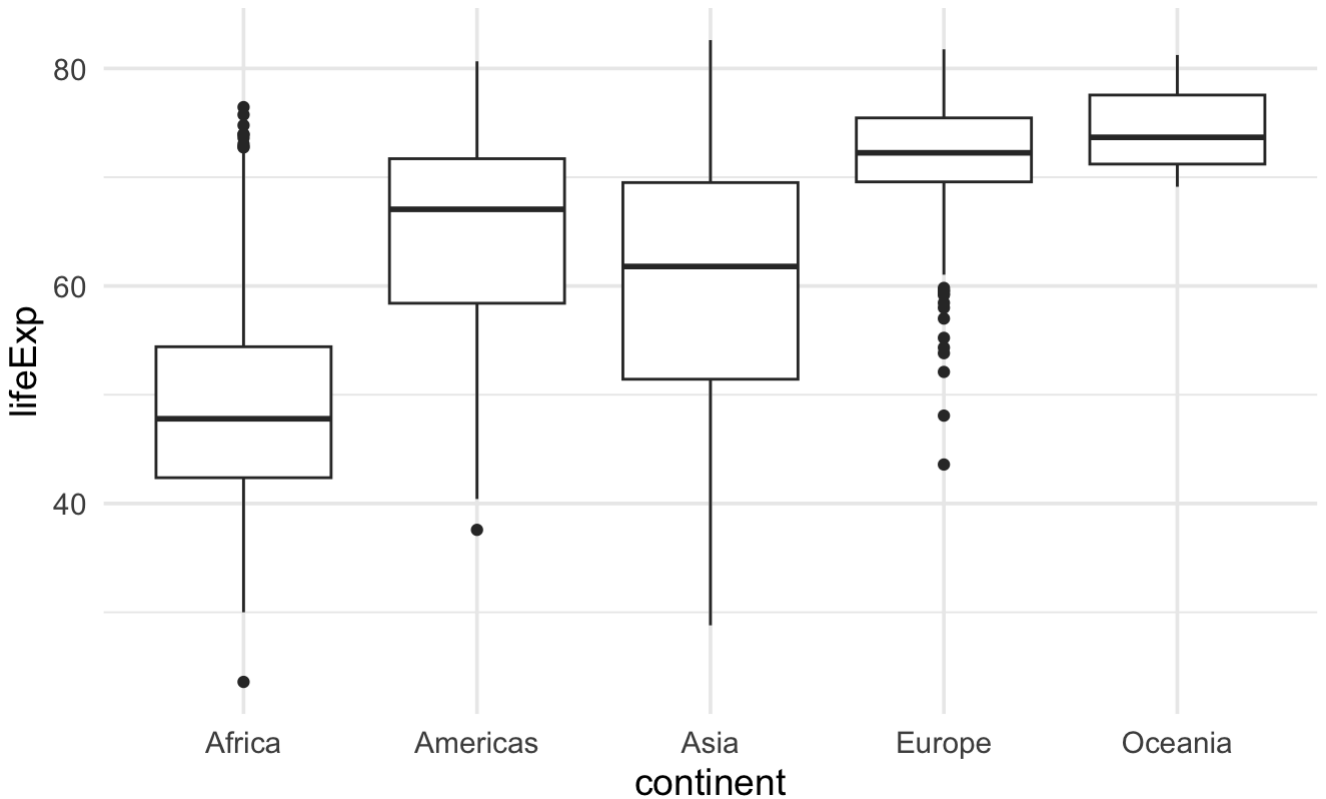
```
ggplot(data = gapminder,
       mapping = aes(x = continent,
                     y = lifeExp)) +
  geom_boxplot() +
  theme_minimal(base_size = 14)
```



Polls

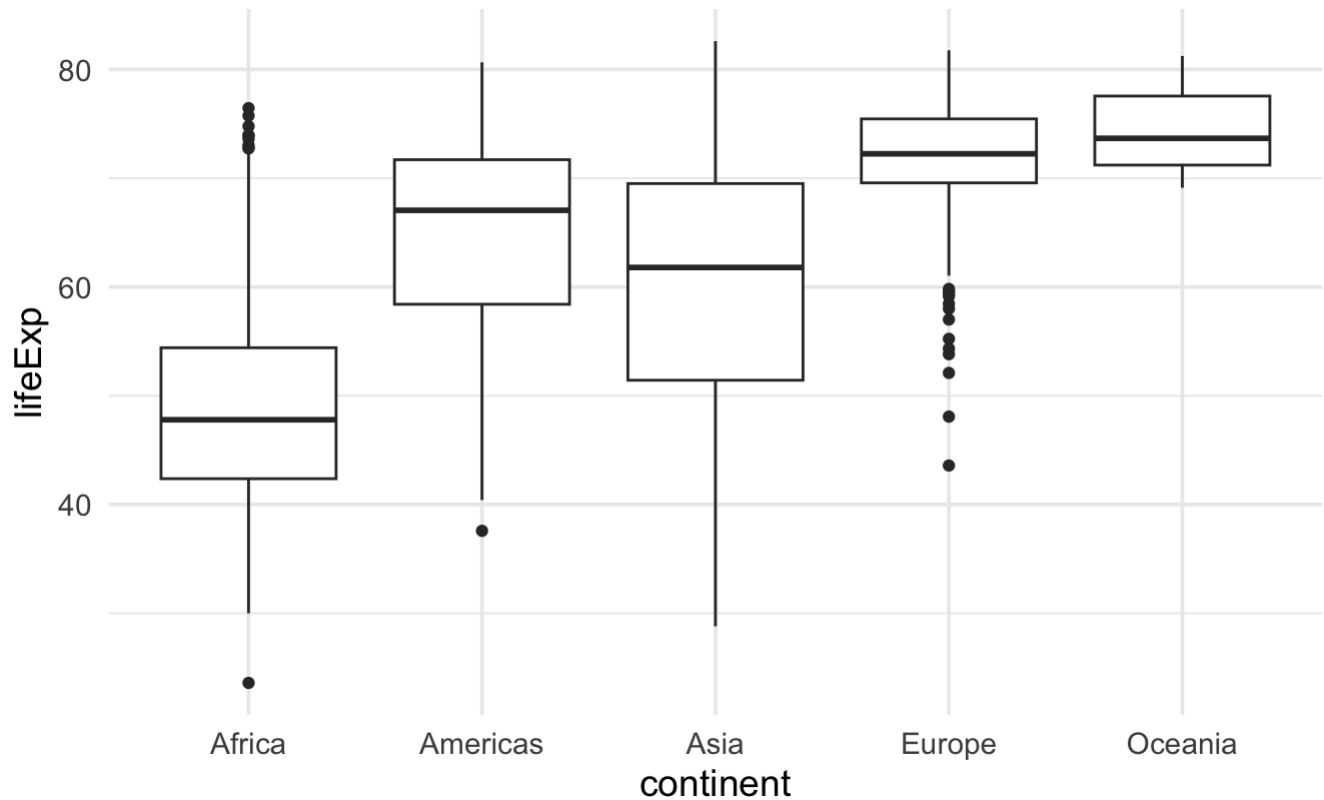
Poll 1: What does the thick line inside the box of a boxplot represent?

1. the mean of the observations
2. the middle of the box
3. the median of the observations
4. none of the above



Poll 2: What percentage of observations are contained inside the box of a boxplot (interquartile range)?

1. 25%
2. depends on the median
3. 50%
4. none of the above



Poll 3: What is the median of a set of observations?

1. The median is the most frequently occurring value in a dataset.
2. The median is the sum of all values in a dataset divided by the number of observations.
3. The median is the point above and below which half (50%) of the observations falls.
4. The median is the square root of the sum of the squares of each value in a dataset.

Poll 4: If you have the values: 1, 2, 3, and 10: which statistical measure best represents the “true” value?

1. The mean
2. The standard deviation
3. The median
4. The interquartile range

Boxplot, explained

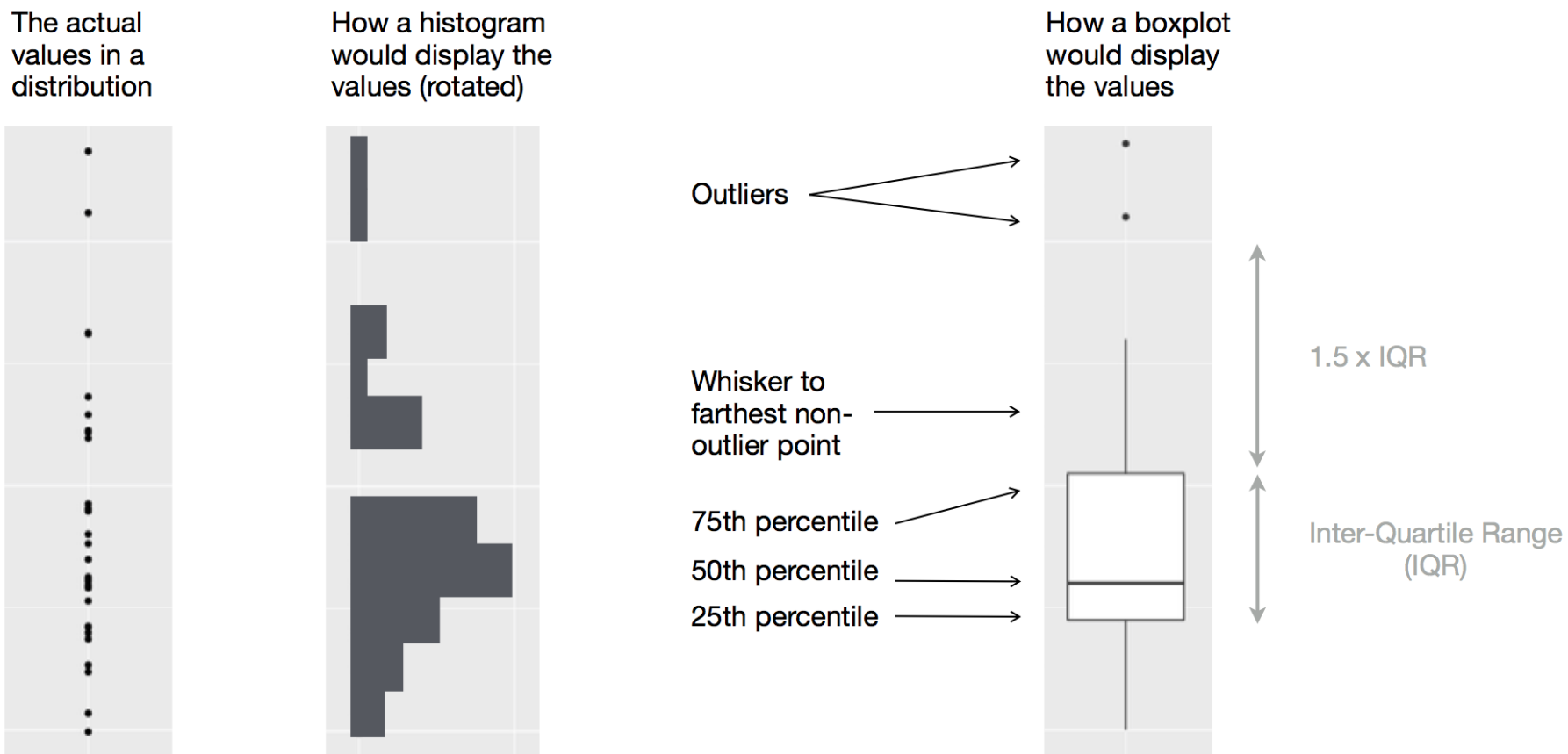


Figure 1: Diagram depicting how a boxplot is created.

ds4owd-001.github.io/website/

Our turn: md-02-exercises

1. Open posit.cloud in your browser (use your bookmark).
2. Open the `ds4owd workspace` for the course.
3. Click `Start` next to `md-02-exercises`.
4. In the File Manager in the bottom right window, locate the `md-02b-data-visualization.qmd` file and click on it to open it in the top left window.

30:00

Visualising data

Types of variables

numerical

discrete variables

- non-negative
- whole numbers
- e.g. number of students, roll of a dice

continuous variables

- infinite number of values
- also dates and times
- e.g. length, weight, size

non-numerical

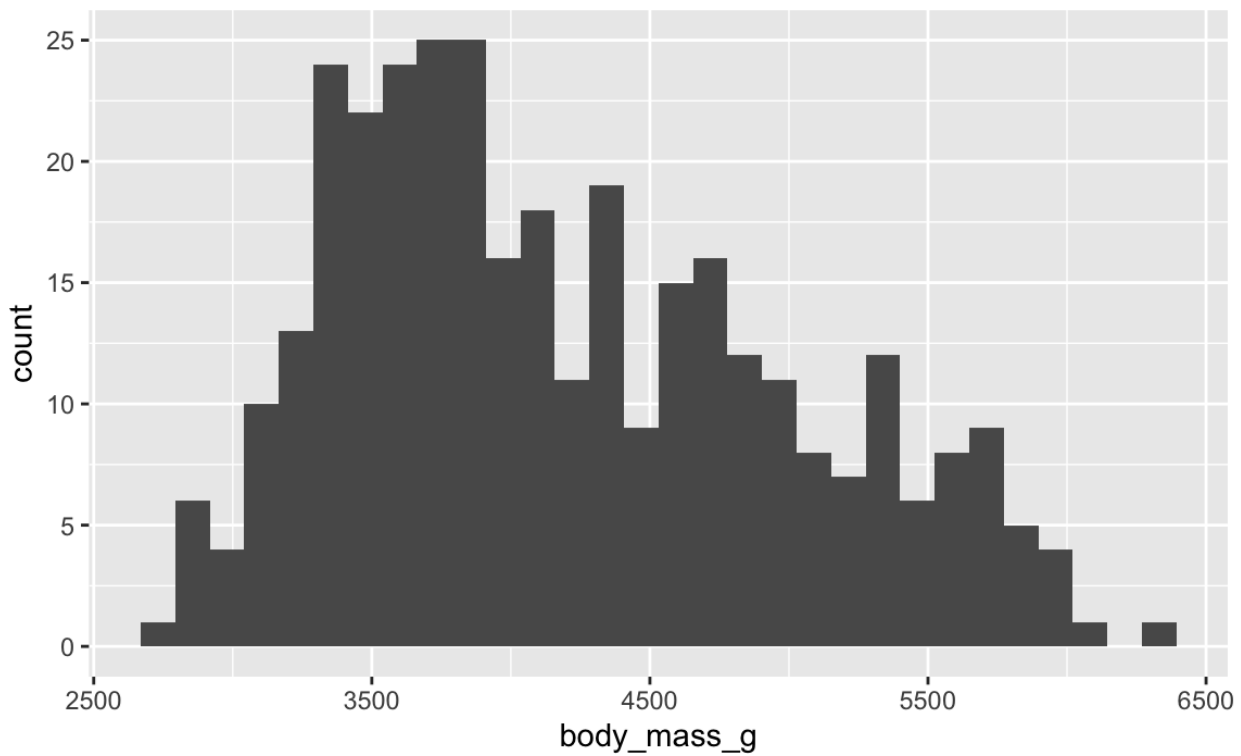
categorical variables

- finite number of values
- distinct groups (e.g. EU countries, continents)
- **ordinal** if levels have natural ordering (e.g. week days, school grades)

Histogram

- for visualizing distribution of continuous (numerical) variables

```
ggplot(data = penguins,
       mapping = aes(x = body_mass_g)) +
  geom_histogram()
```



Barplot

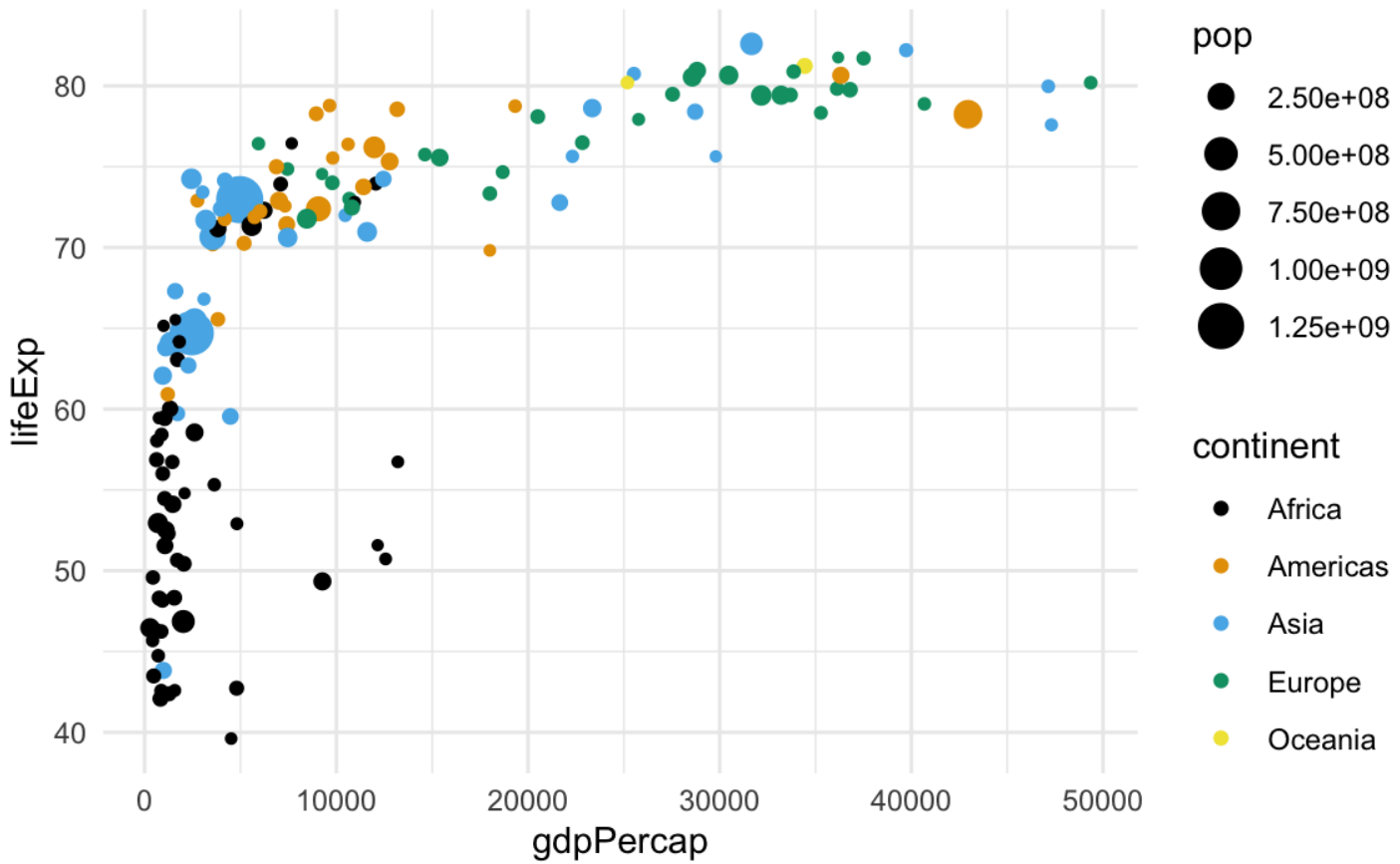
- for visualizing distribution of categorical (non-numerical) variables

```
ggplot(data = penguins,
       mapping = aes(x = species)) +
  geom_bar()
```

Scatterplot

- for visualising relationships between two continuous (numerical) variables

```
ggplot(data = gapminder_2007,
       mapping = aes(x = gdpPercap,
                     y = lifeExp,
                     size = pop,
                     color = continent)) +
  geom_point() +
  scale_color_colorblind() +
  theme_minimal()
```



ds4owd-001.github.io/website/

Your turn: md-02-exercises

1. Open posit.cloud in your browser (use your bookmark).
2. Open the ds4owd workspace for the course.
3. In the File Manager in the bottom right window, locate the `md-02c-make-a-plot.qmd` file and click on it to open it in the top left window.
4. Follow instructions in the file

15:00

Homework assignments module 2

ds4owd-001.github.io/website/

Module 2 documentation

ds4owd-001.github.io/website/modules/md-02.html

Module 2

Data science lifecycle & Quarto

Learning Objectives

1. Learners can render a Quarto file to an output file in HTML, PDF and DOCX format.
2. Learners can list the six elements of the data science lifecycle.
3. Learners can identify four components of a Quarto file (YAML, code chunk, R code, markdown).

Slides

- In preparation

Readings

Homework due date

- Homework assignment due: Monday, November 13th
- Correction & feedback phase up to: Thursday, November 16th

Wrap-up

ds4owd-001.github.io/website/

Thanks! 🌻

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/ Access slides as PDF on GitHub