

Data science lifecycle & Exploratory data analysis using visualization

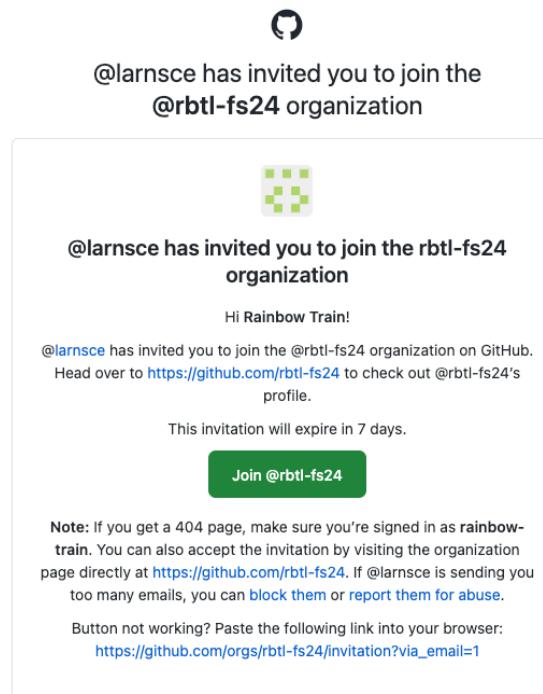
Research Beyond the Lab: Open Science and Research Methods
for a Global Engineer

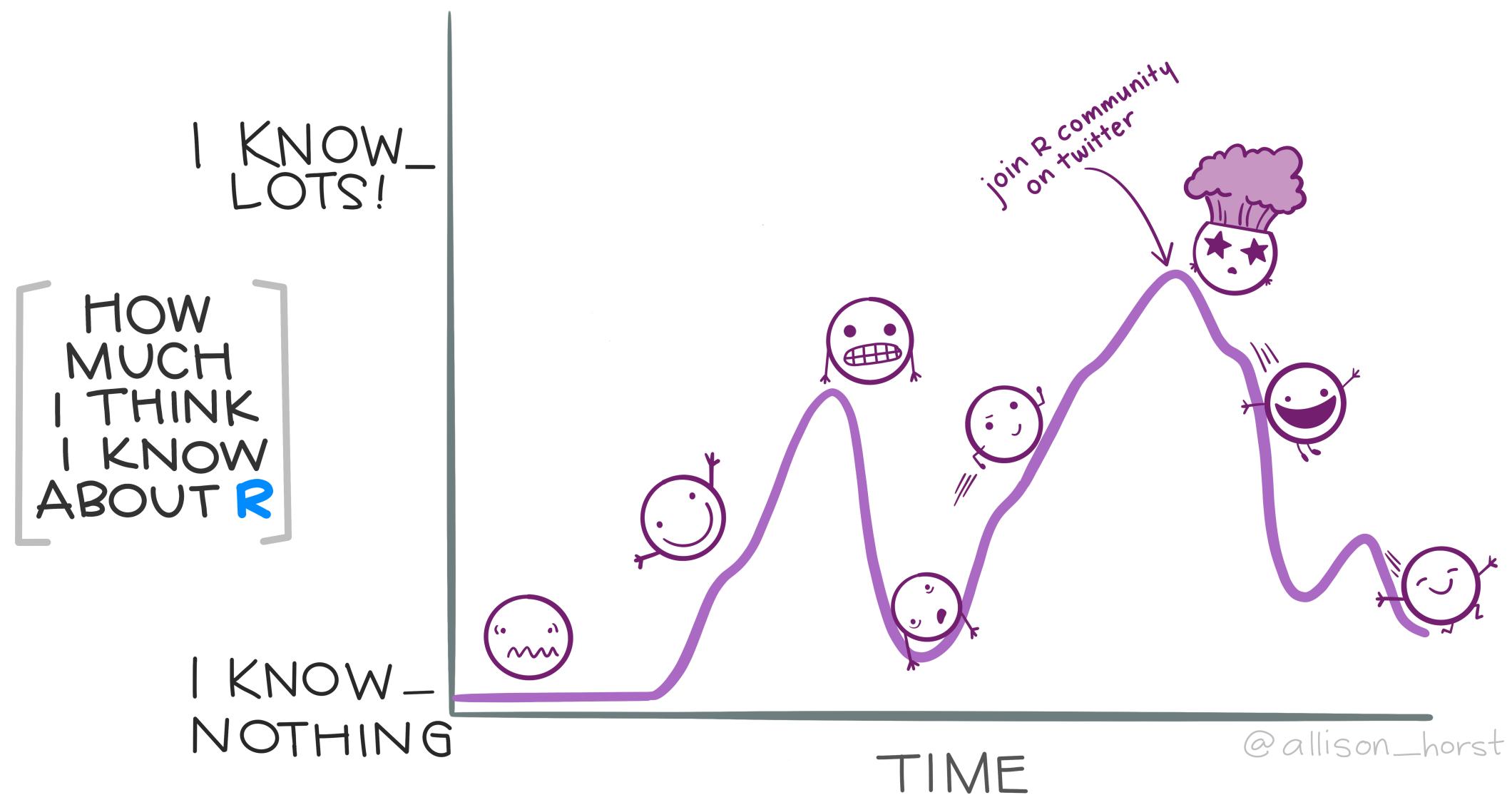
Lars Schöbitz

Feb 27, 2025

Email from GitHub?

While we are getting ready, please check for this email from GitHub and **accept the invitation** to join the GitHub organisation for the course. Used Gmail to sign up? Check the folders that aren't your primary inbox (e.g Updates).





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>





Interaction with GitHub

open GitHub organisation

Bookmark this link in your browser!

github.com/rbtl-fs25

rbtl-fs24

github.com/rbtl-fs24

rbtl-fs24

Type ⌘ to search

Overview Repositories 1 Projects Packages Teams People 5

Unfollow

Popular repositories

website Public

Forked from [ds4owd-001/website](#)

Website for "rbtl - Research Beyond the Lab: Open Science and Research Methods for a Global Engineer"

JavaScript

View as: Public

You are viewing the README and pinned repositories as a public user.

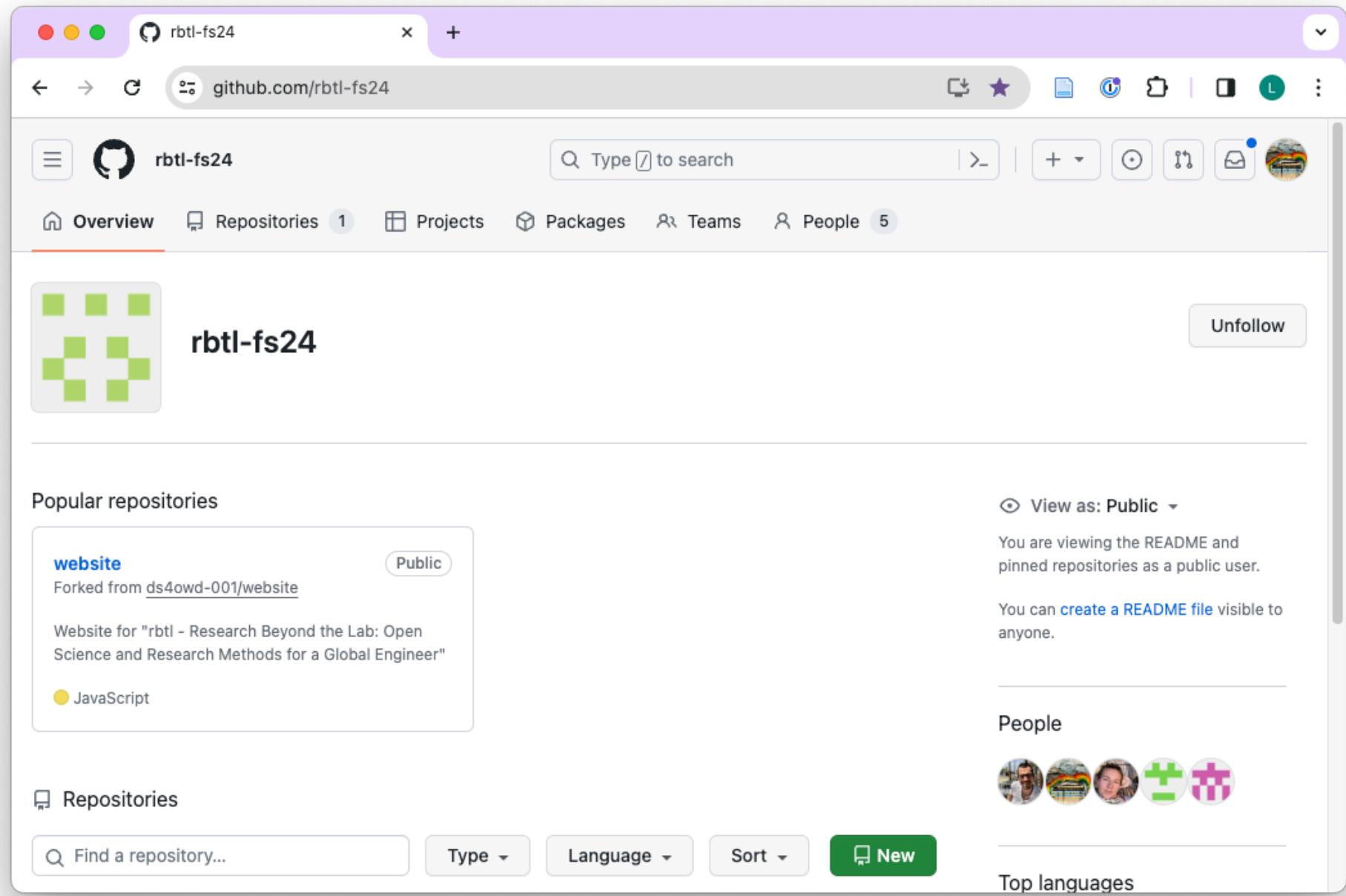
You can [create a README file](#) visible to anyone.

People

Repositories

Find a repository... Type Language Sort New

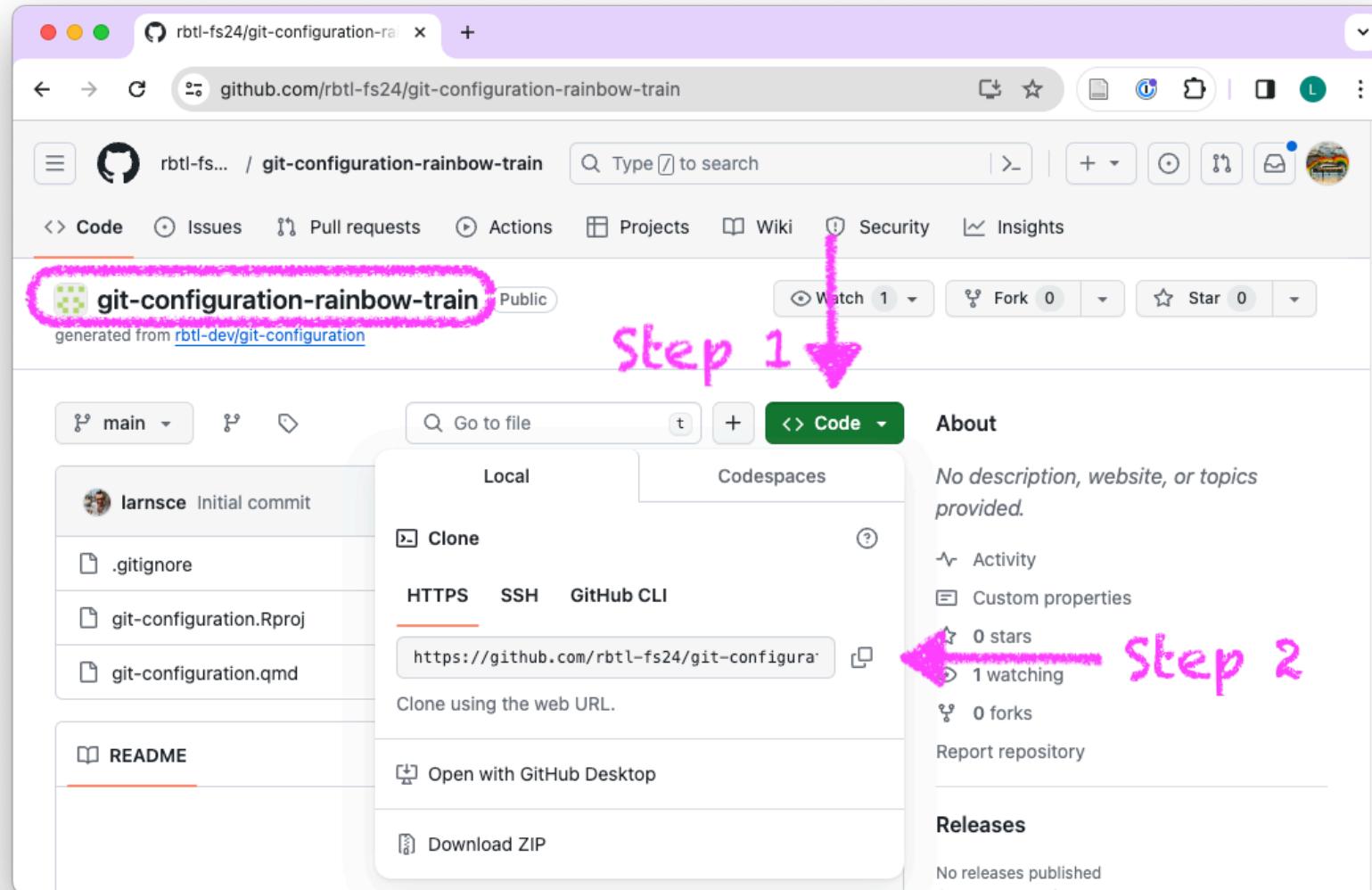
Top languages



on GitHub Organisation

- Search for your username in search bar under
Repositories

on your repository



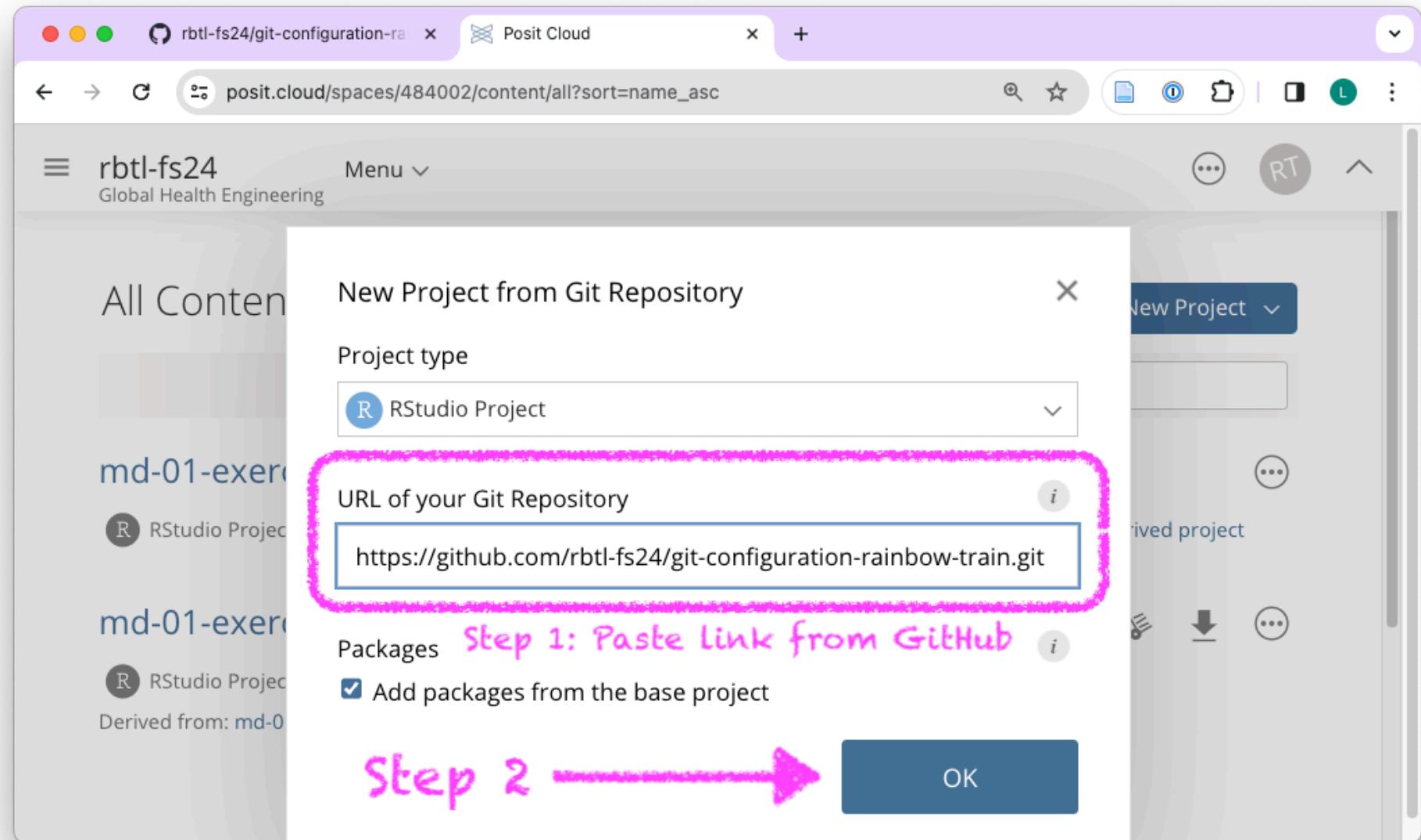
on Posit Cloud

Bookmark this link in your browser!

posit.cloud/spaces/611943/content/

The screenshot shows a web browser window for Posit Cloud. The address bar displays the URL posit.cloud/spaces/484002/content/all?sort=name_asc. The main content area shows a list of projects under the heading "All Content" (2). Two projects are listed: "md-01-exercises" (RStudio Project, Lars Schöbitz, Space members, Created Feb 22, 2024) and another "md-01-exercises" (RStudio Project, Rainbow Train, Private, Created Feb 22, 2024 1:51 PM). A dropdown menu titled "New Project" is open on the right, listing options: "New Project from Template" (selected), "New RStudio Project", "New Jupyter Project", and "New Project from Git Repository". A pink arrow labeled "Step 1" points to the "New Project from Template" option, and a pink arrow labeled "Step 2" points to the "New Project from Git Repository" option, which is highlighted with a pink rounded rectangle.

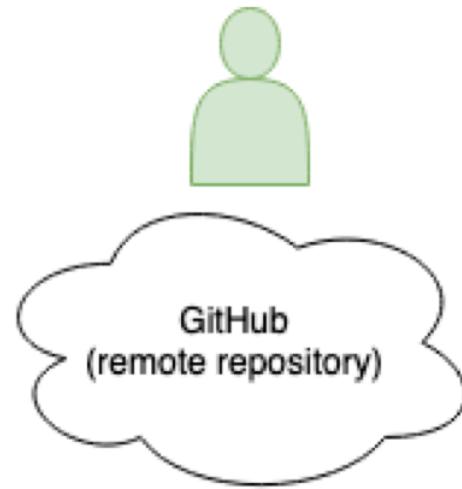
on Posit Cloud

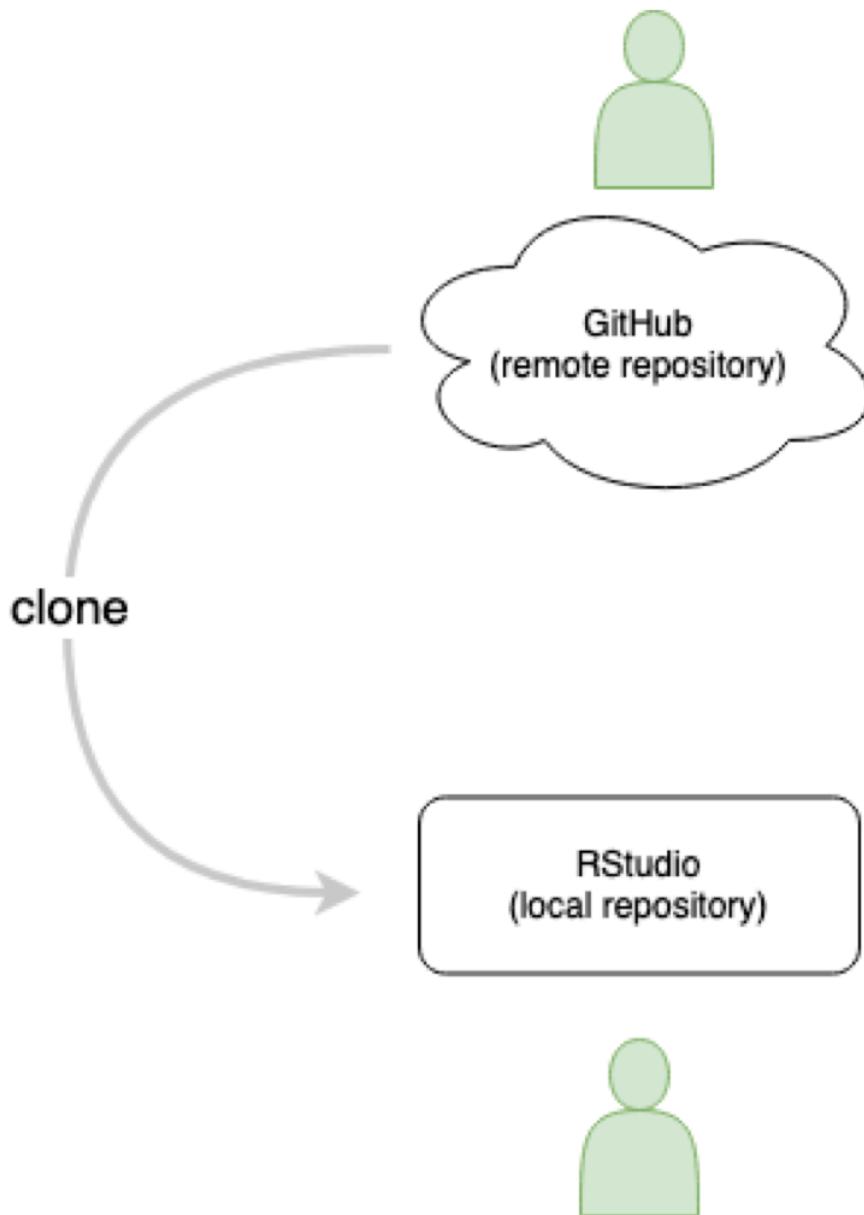


Your turn: Introduce yourself to git

1. Open a web browser on your laptop.
2. Navigate to the course website: rbtl-fs25.github.io/website/
3. If you haven't yet, bookmark the course website
4. In the left-hand menu, click on Module 2, then select am-01: Git configuration
5. Follow the instructions
6. Place a yellow sticky note on your laptop when you have completed the assignment

Version Control - Terminology

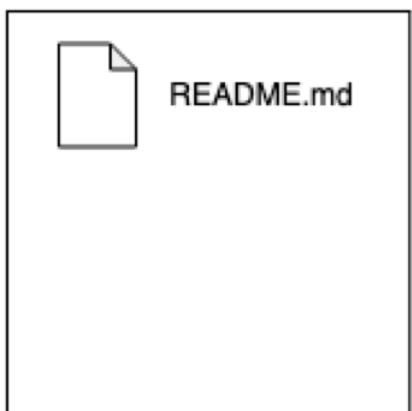






Create README.md

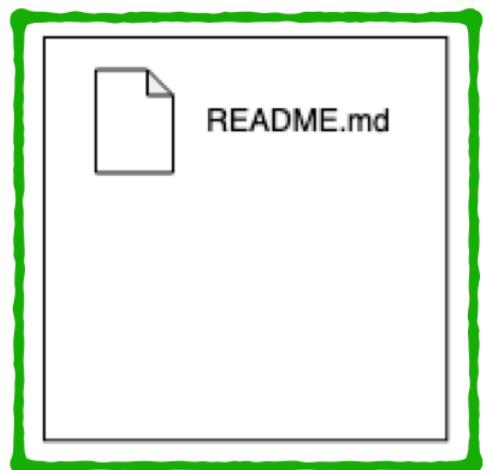
75aa637





Create README.md

75aa637



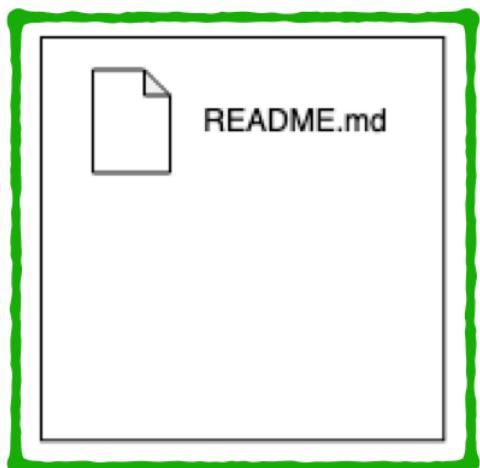
Repo(sitory)

Commit
message

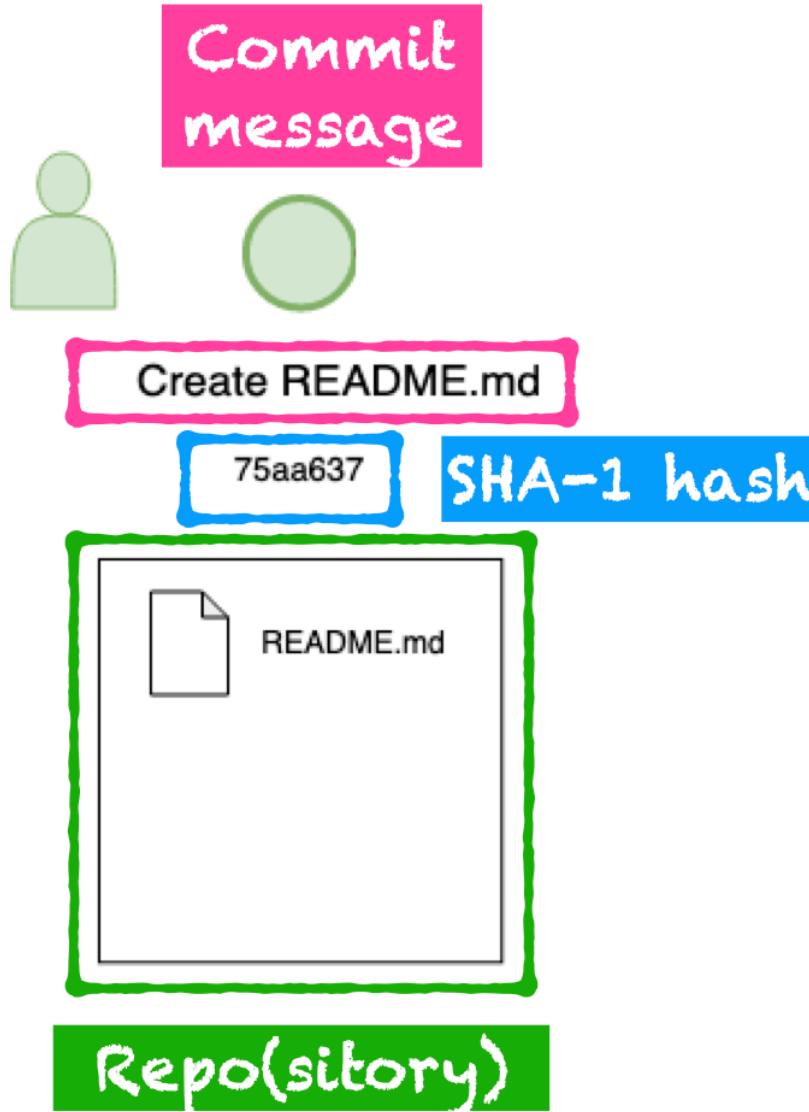


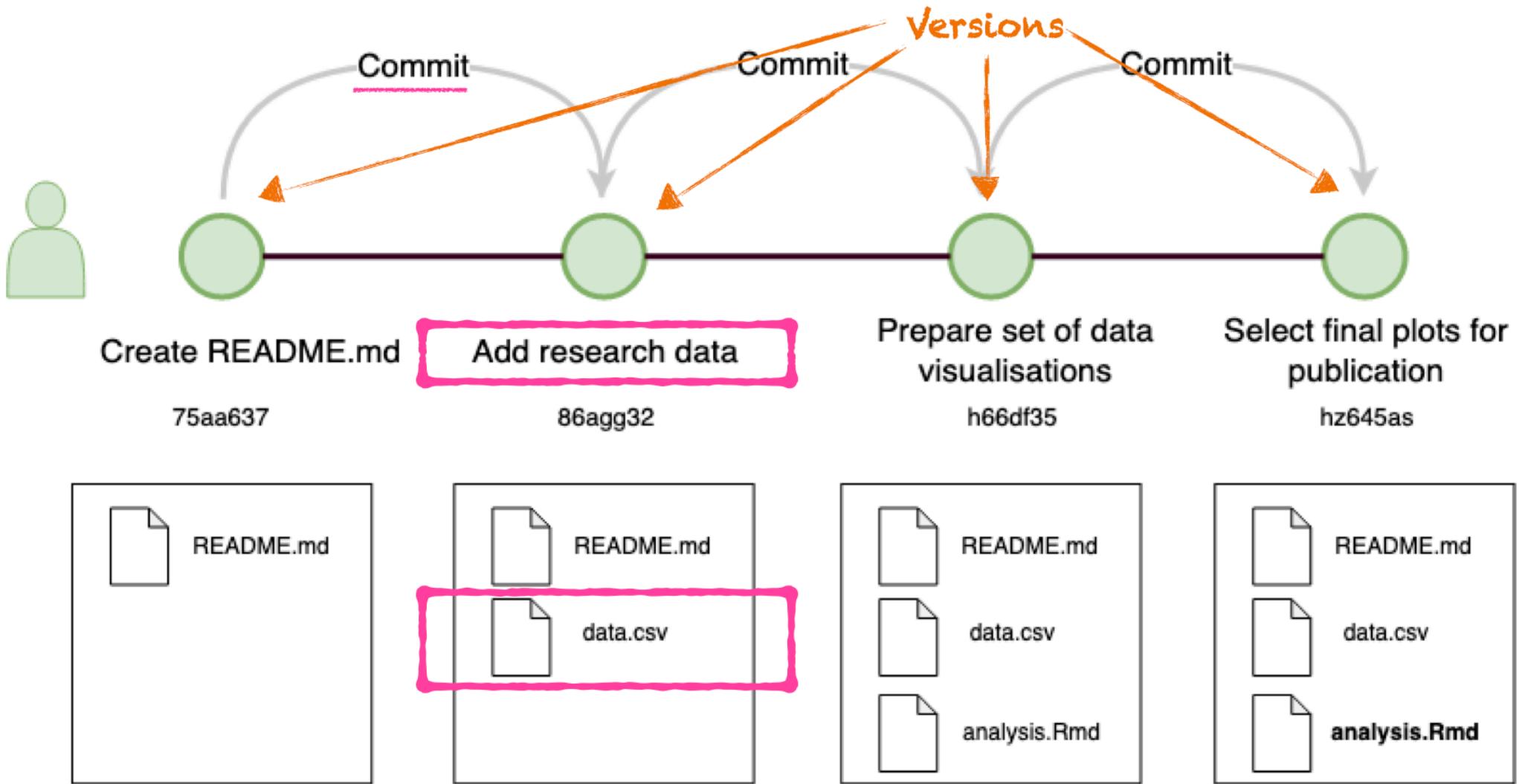
Create README.md

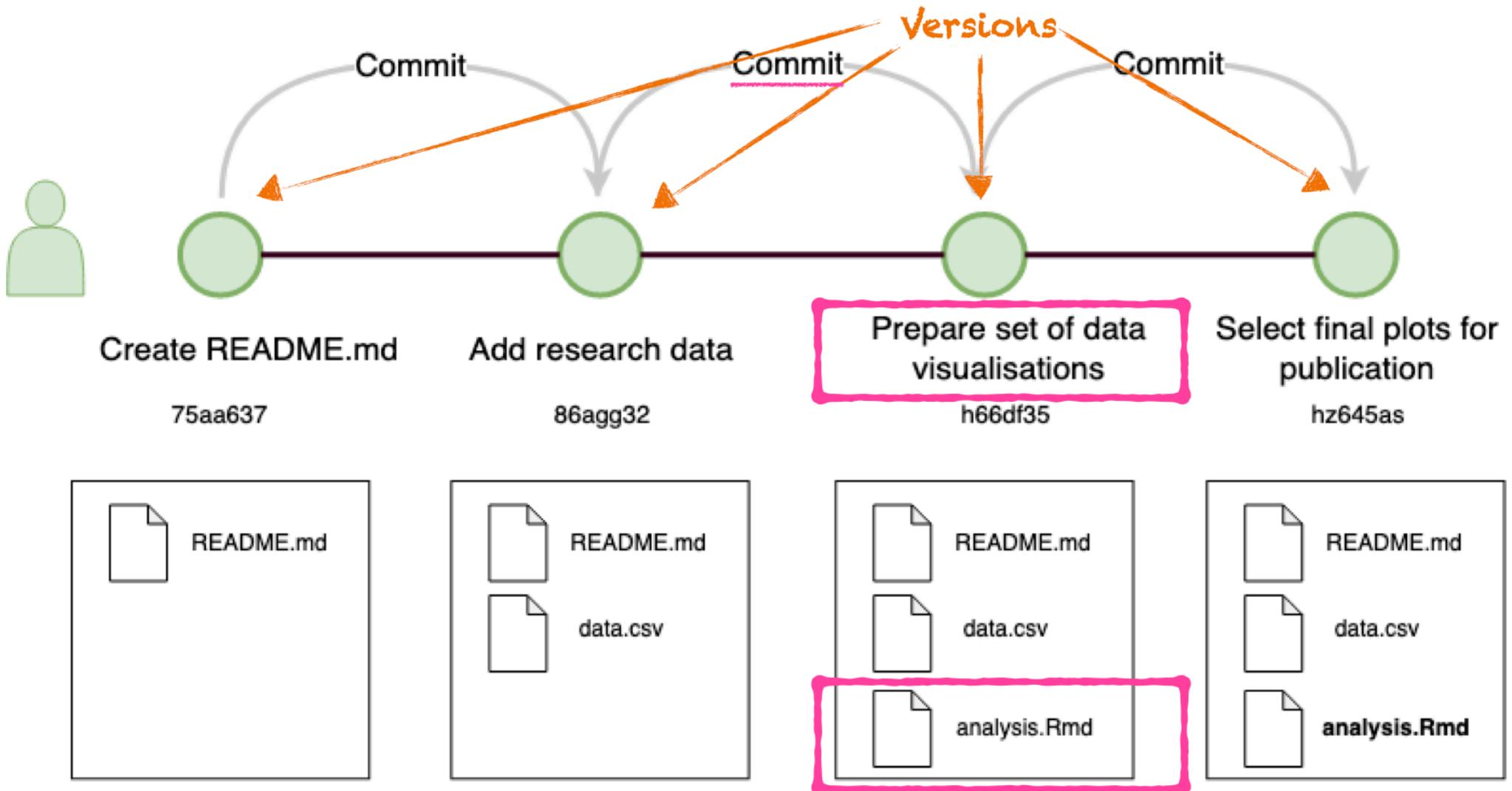
75aa637

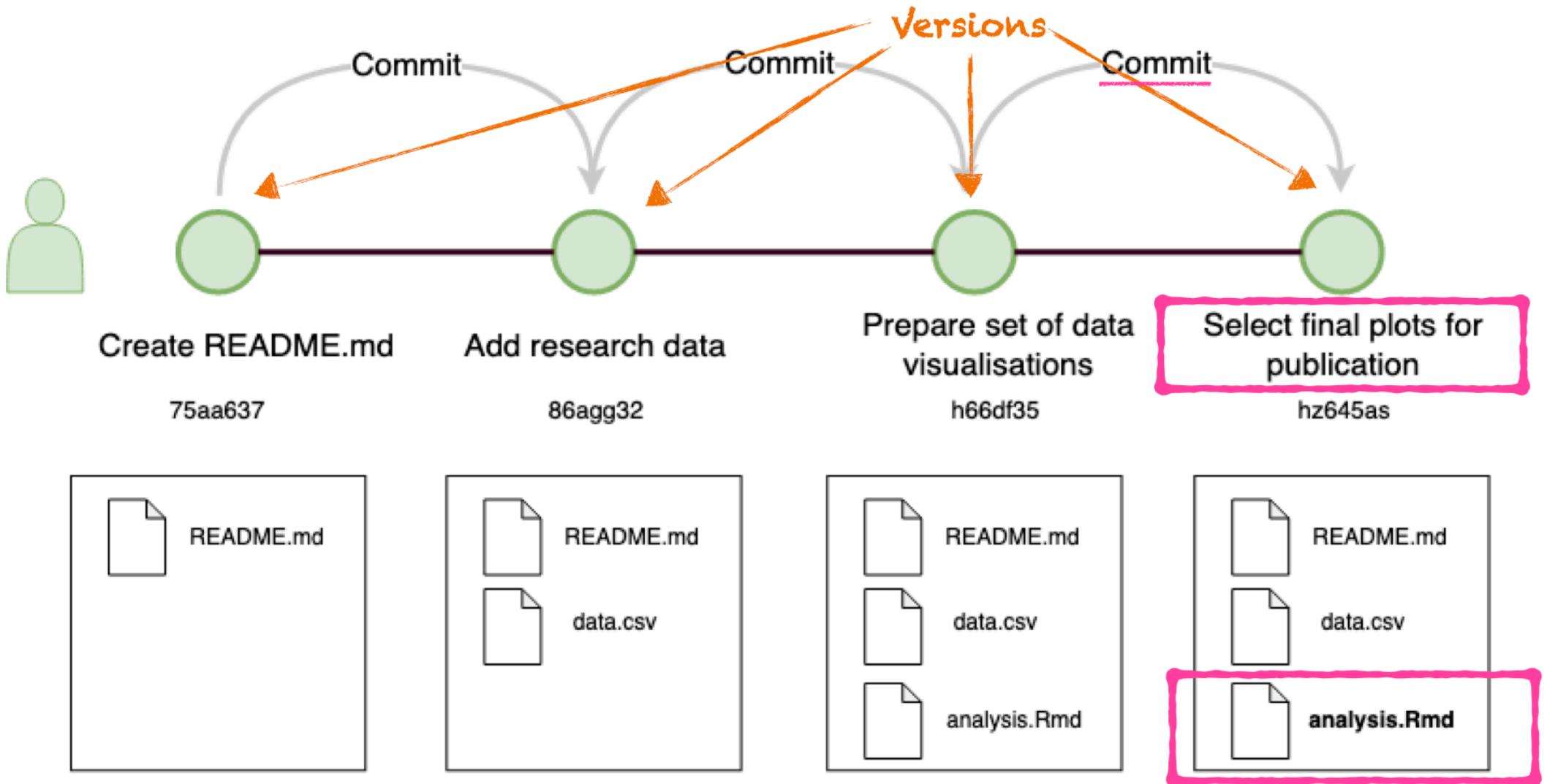


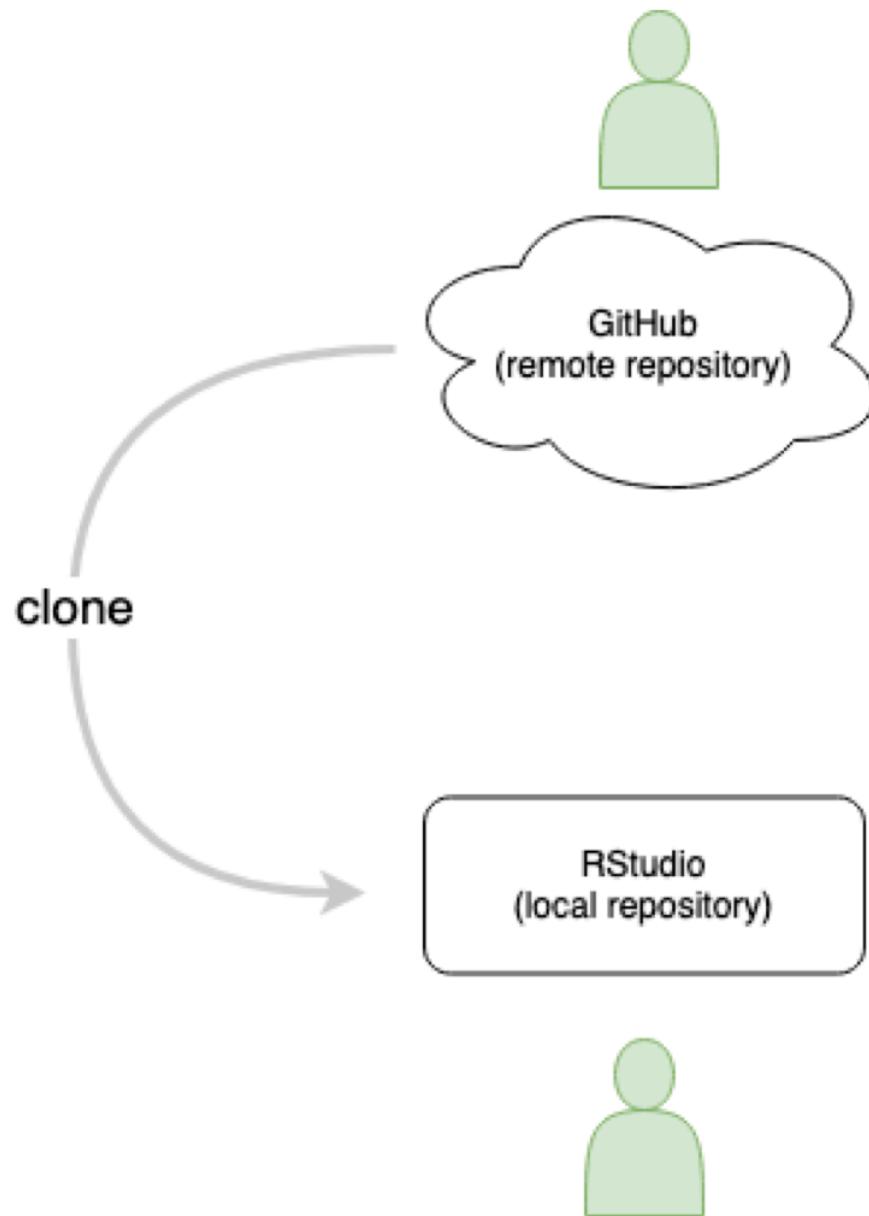
Repo(sitory)

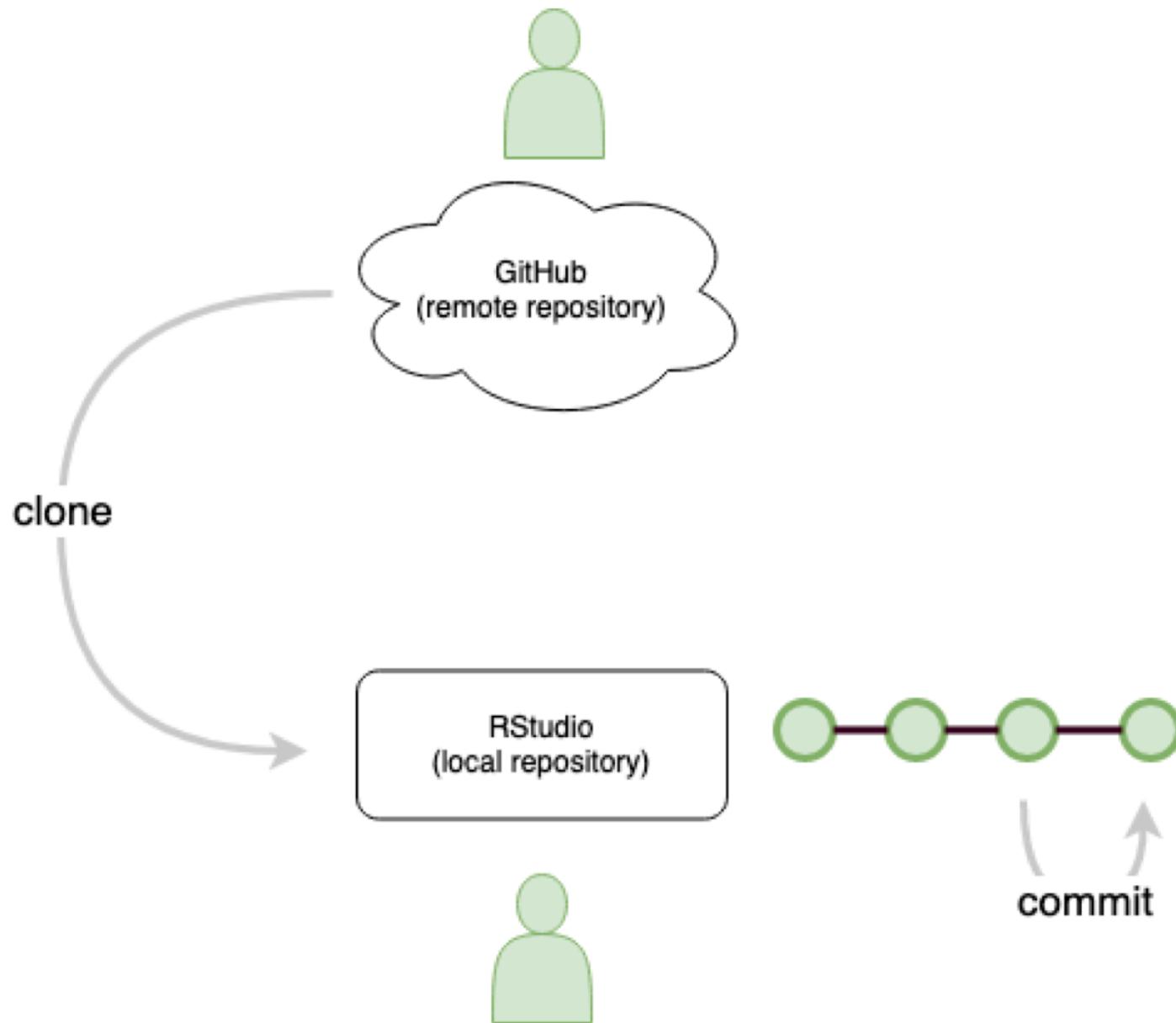


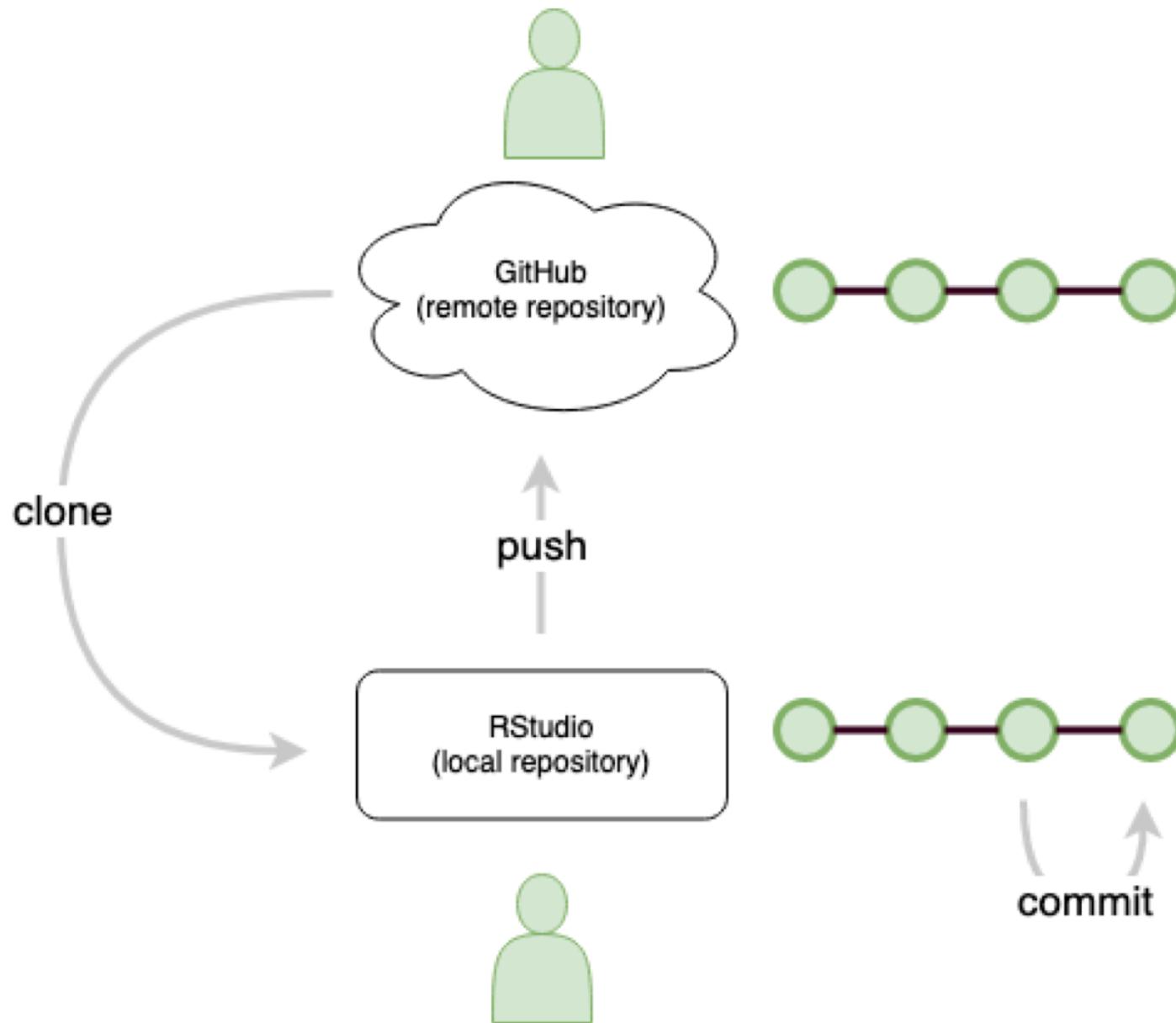




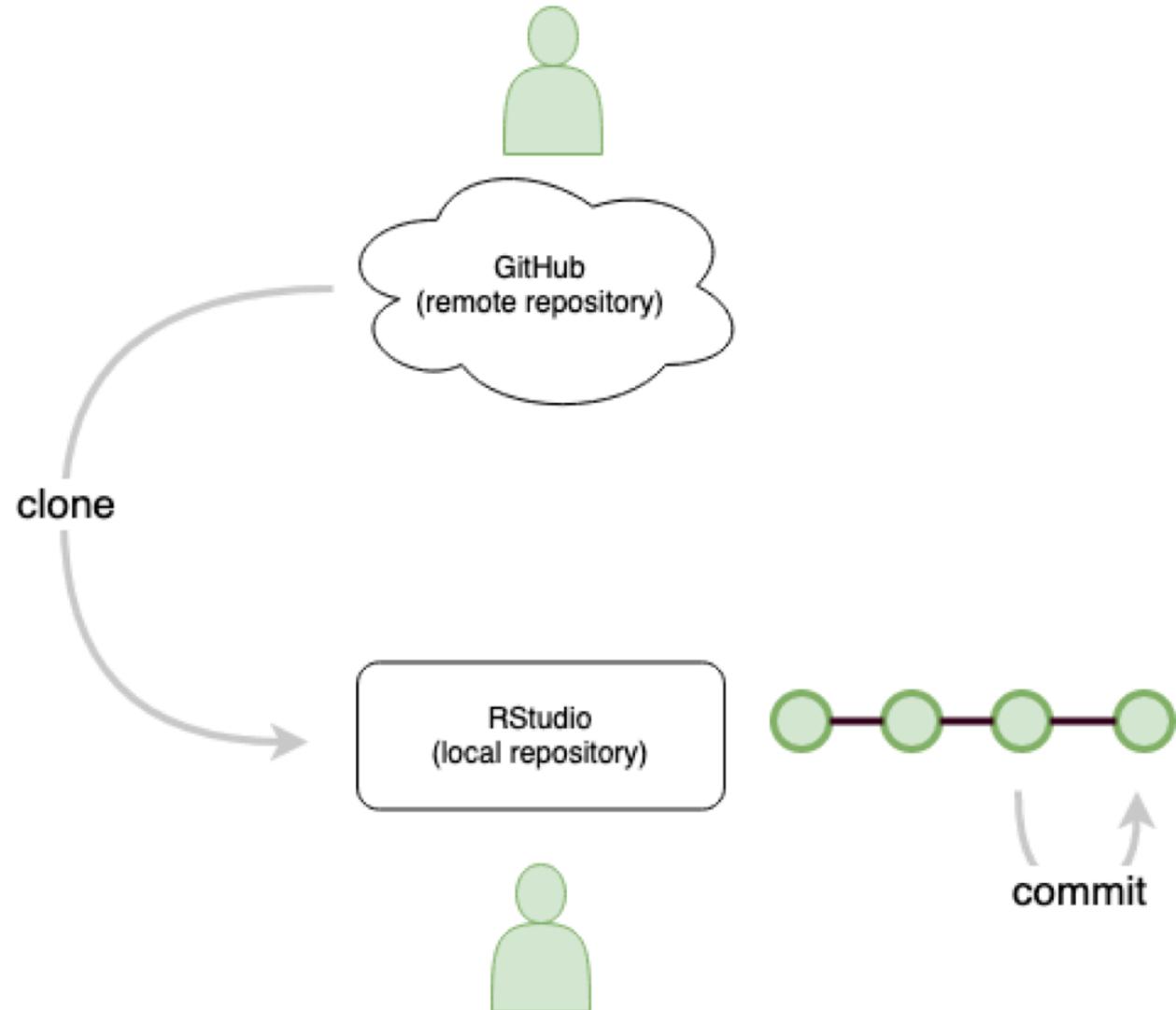




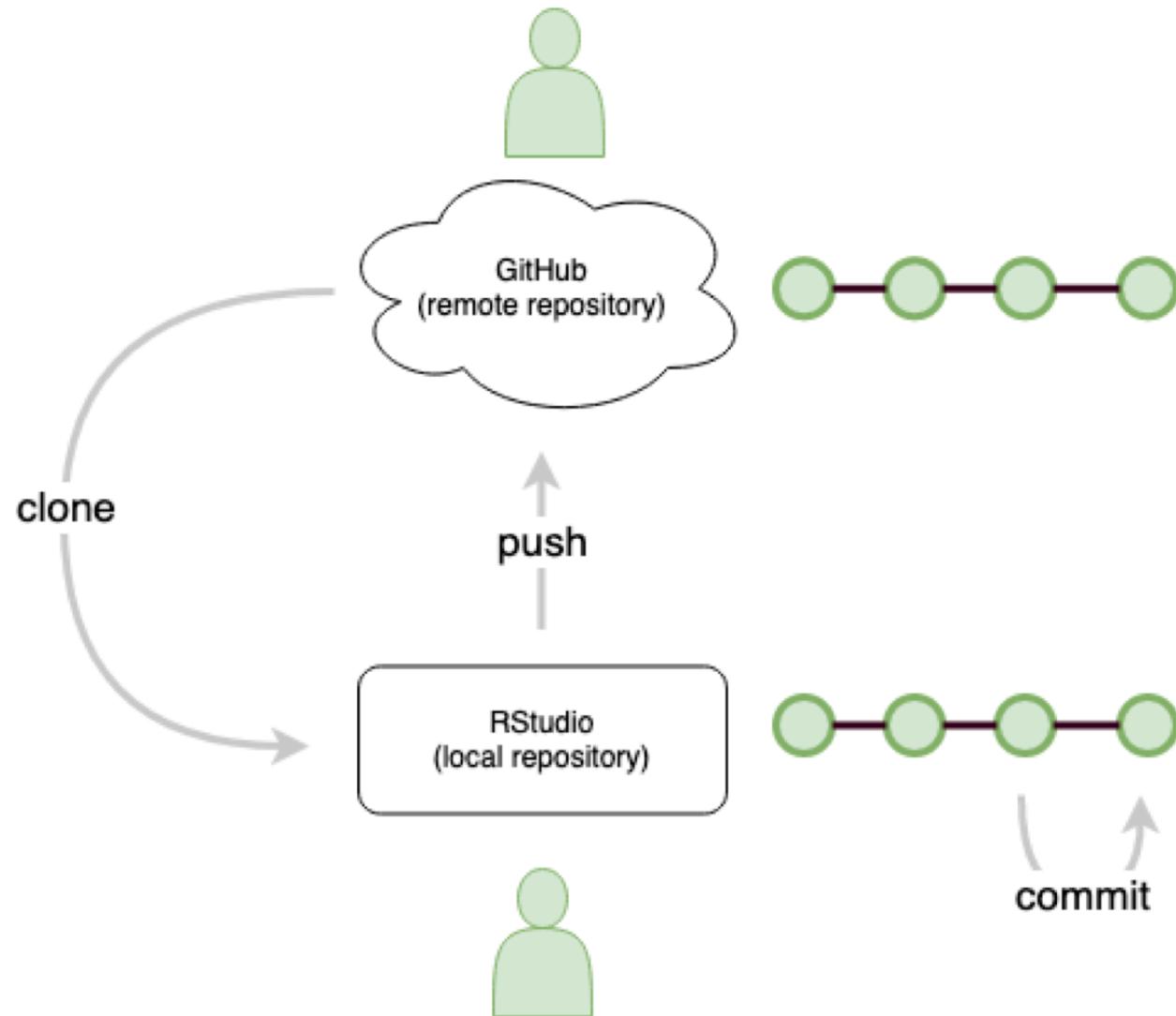




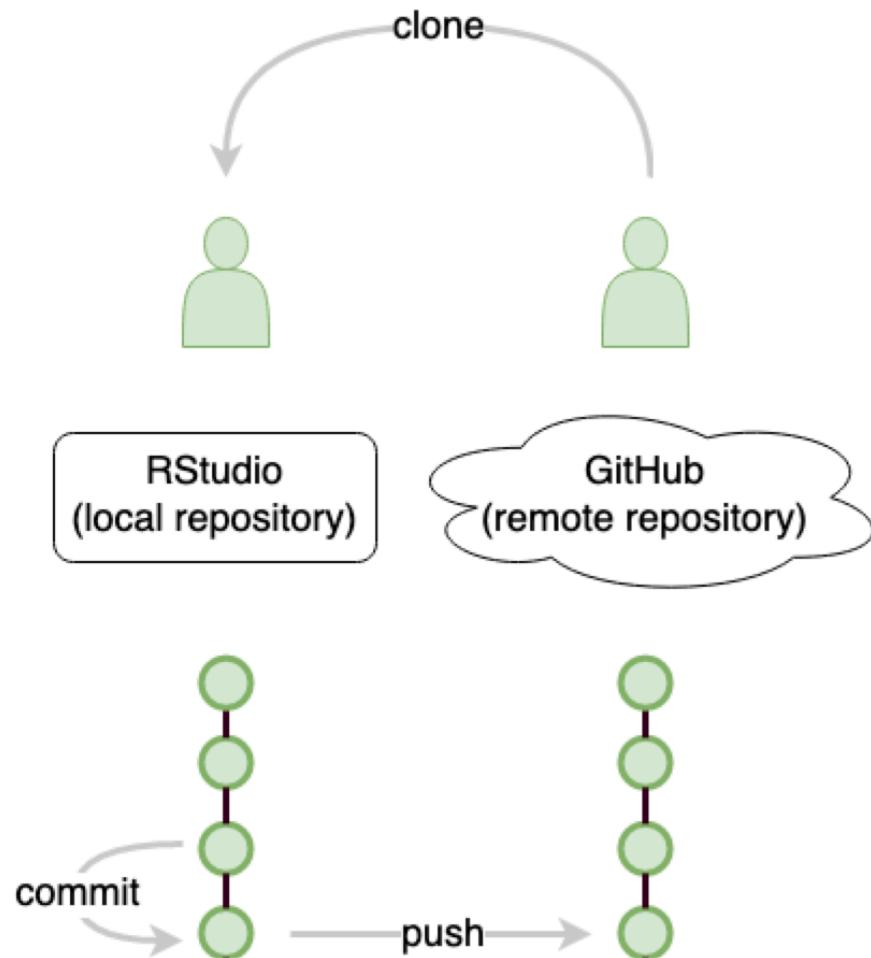
remember: git commit



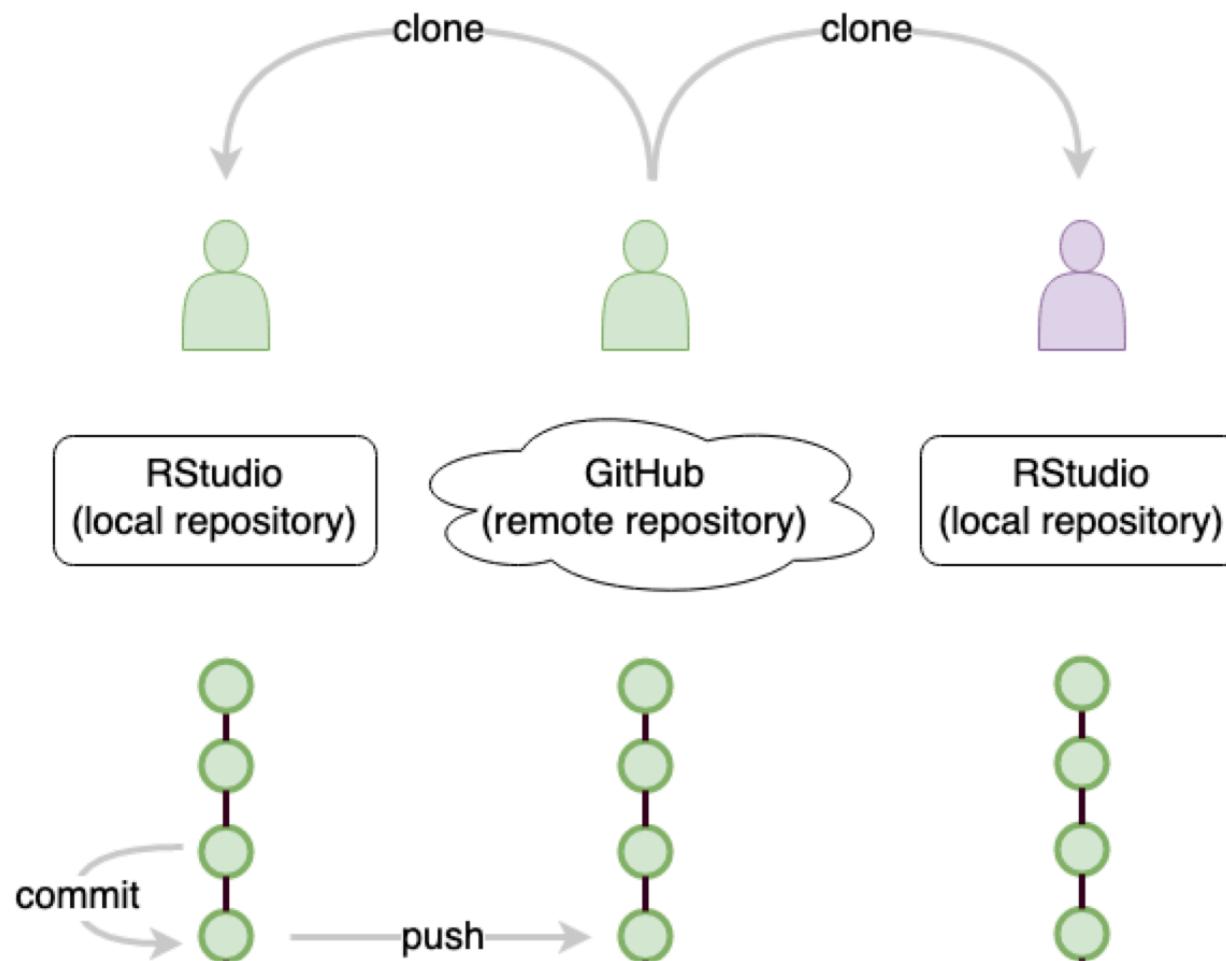
remember: git push



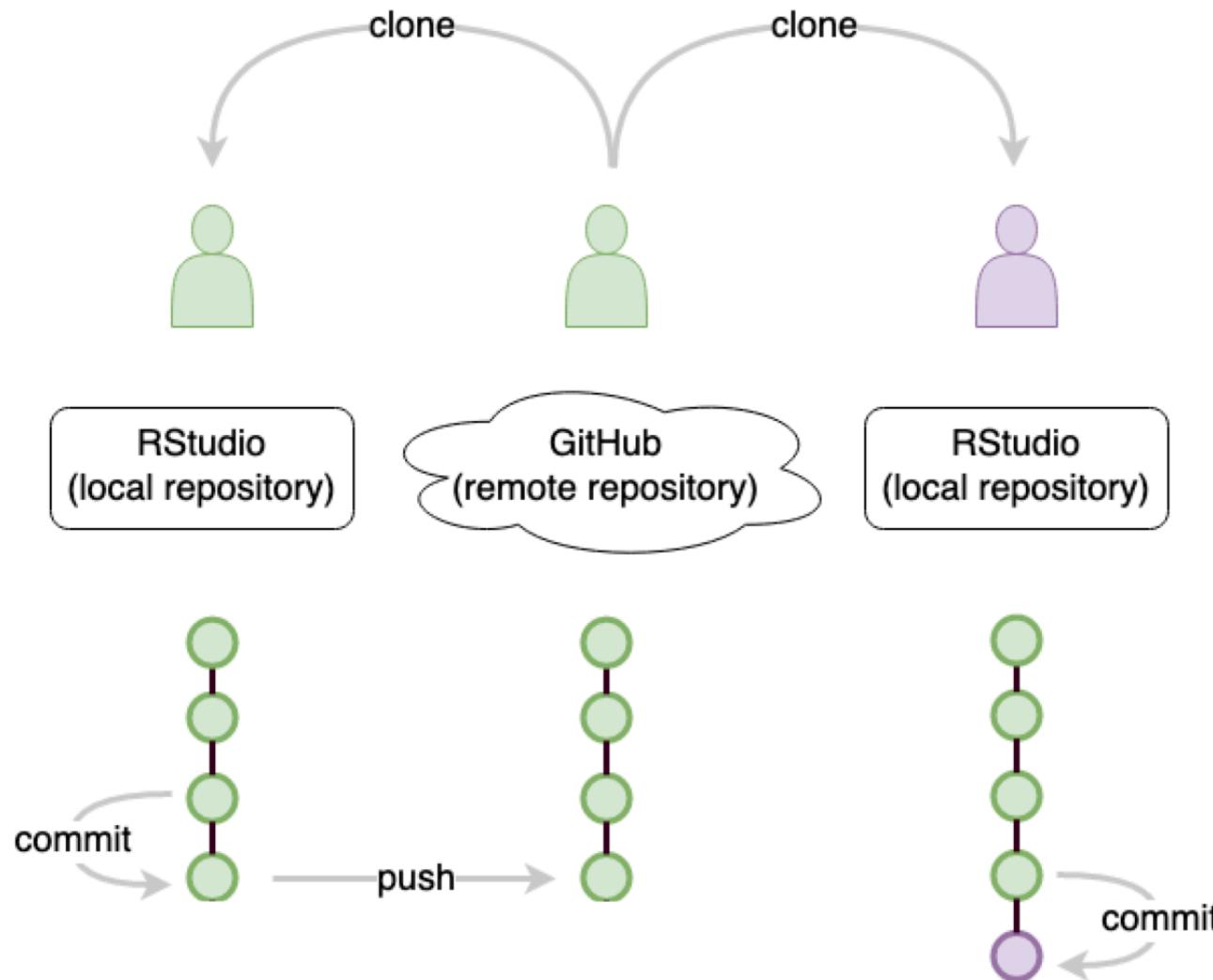
remember: git push



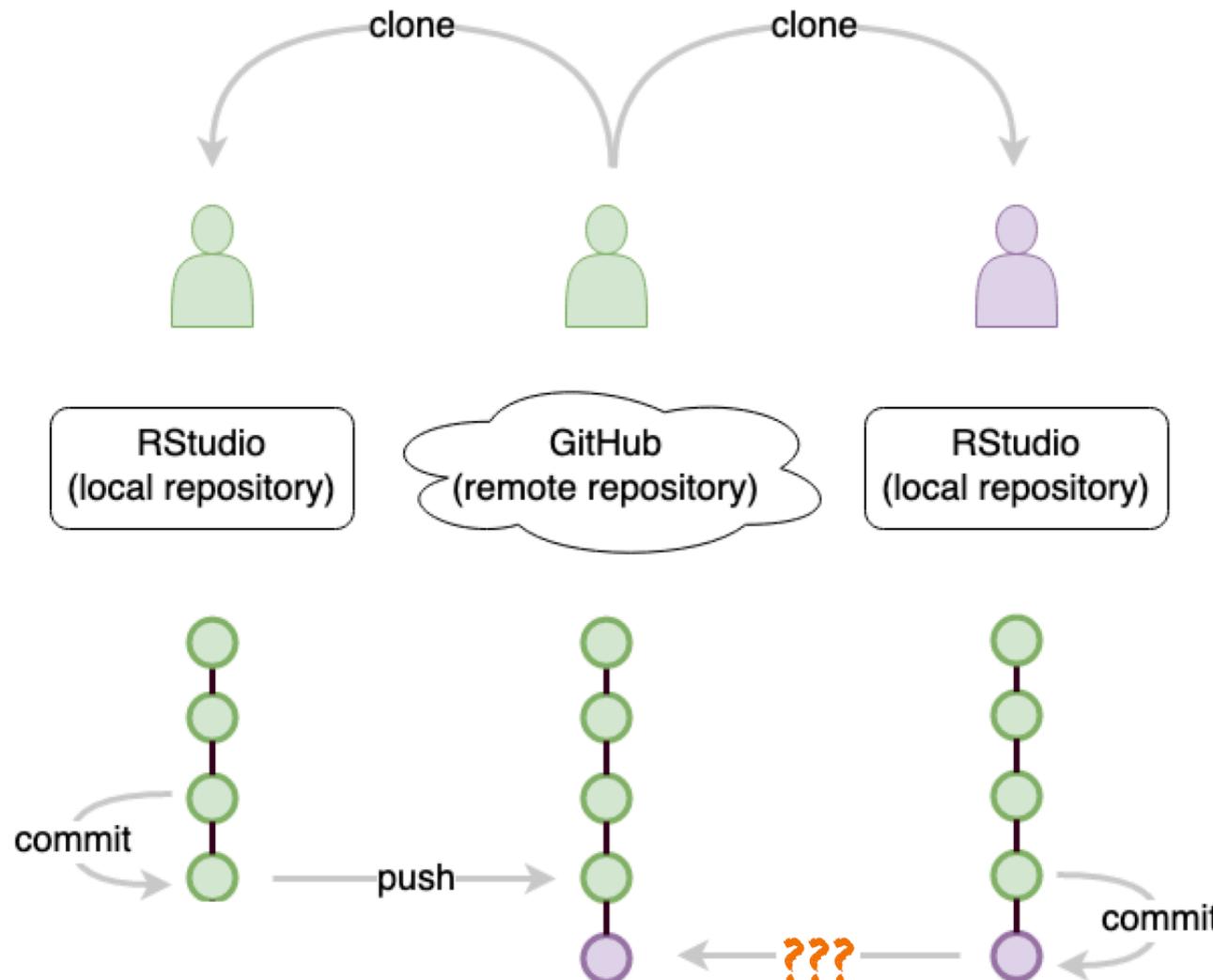
collaborate: git clone



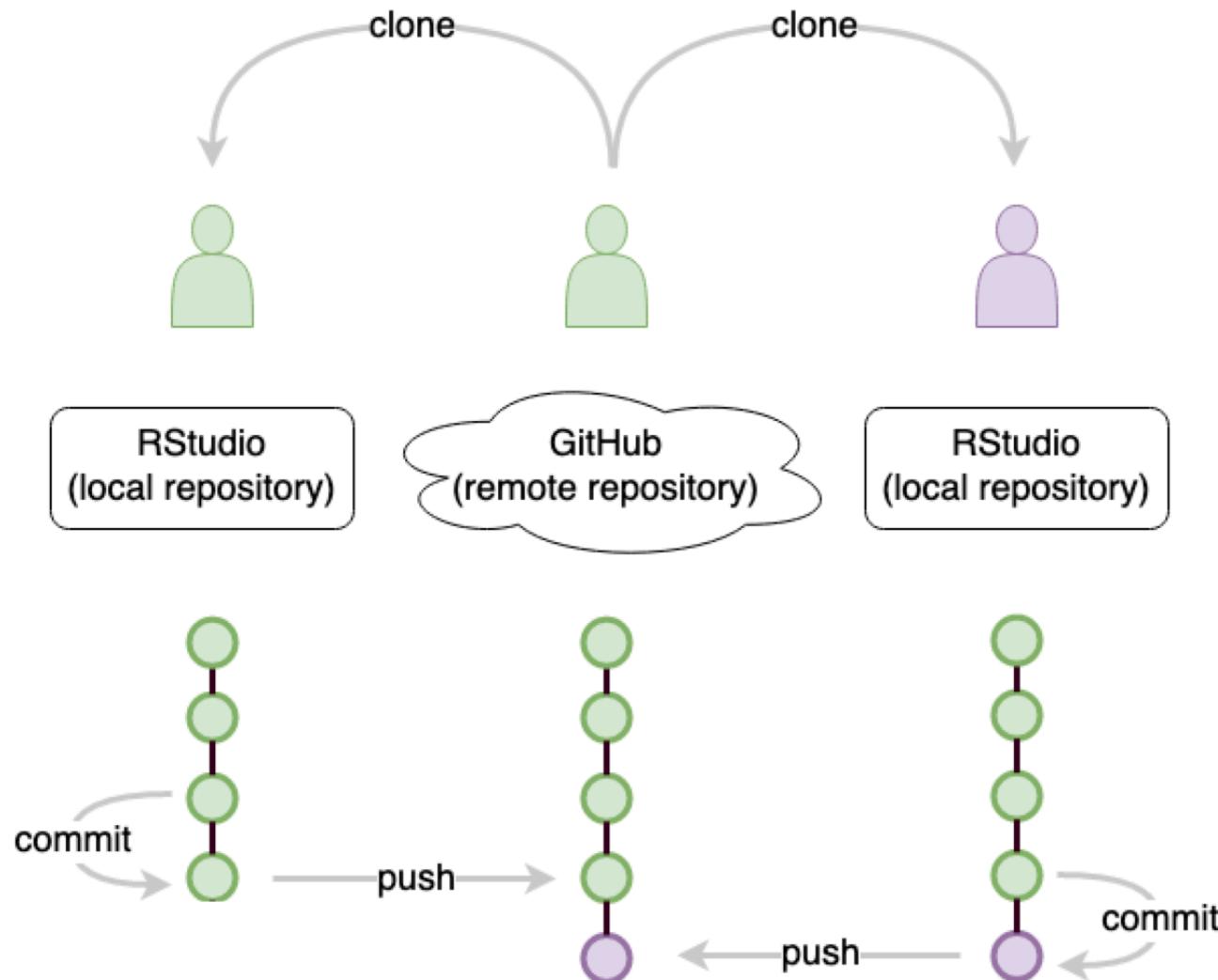
track work: git commit



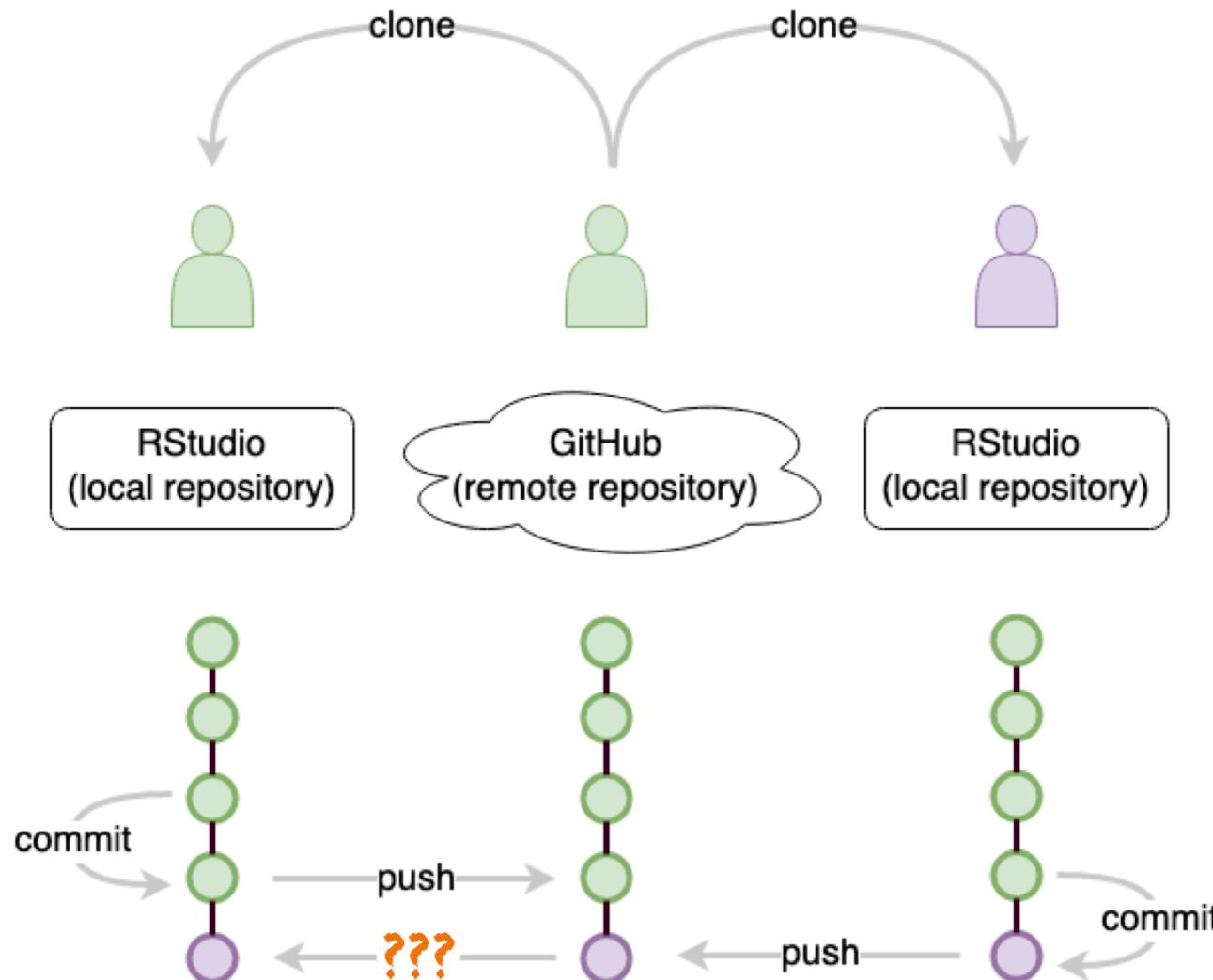
update: git ???



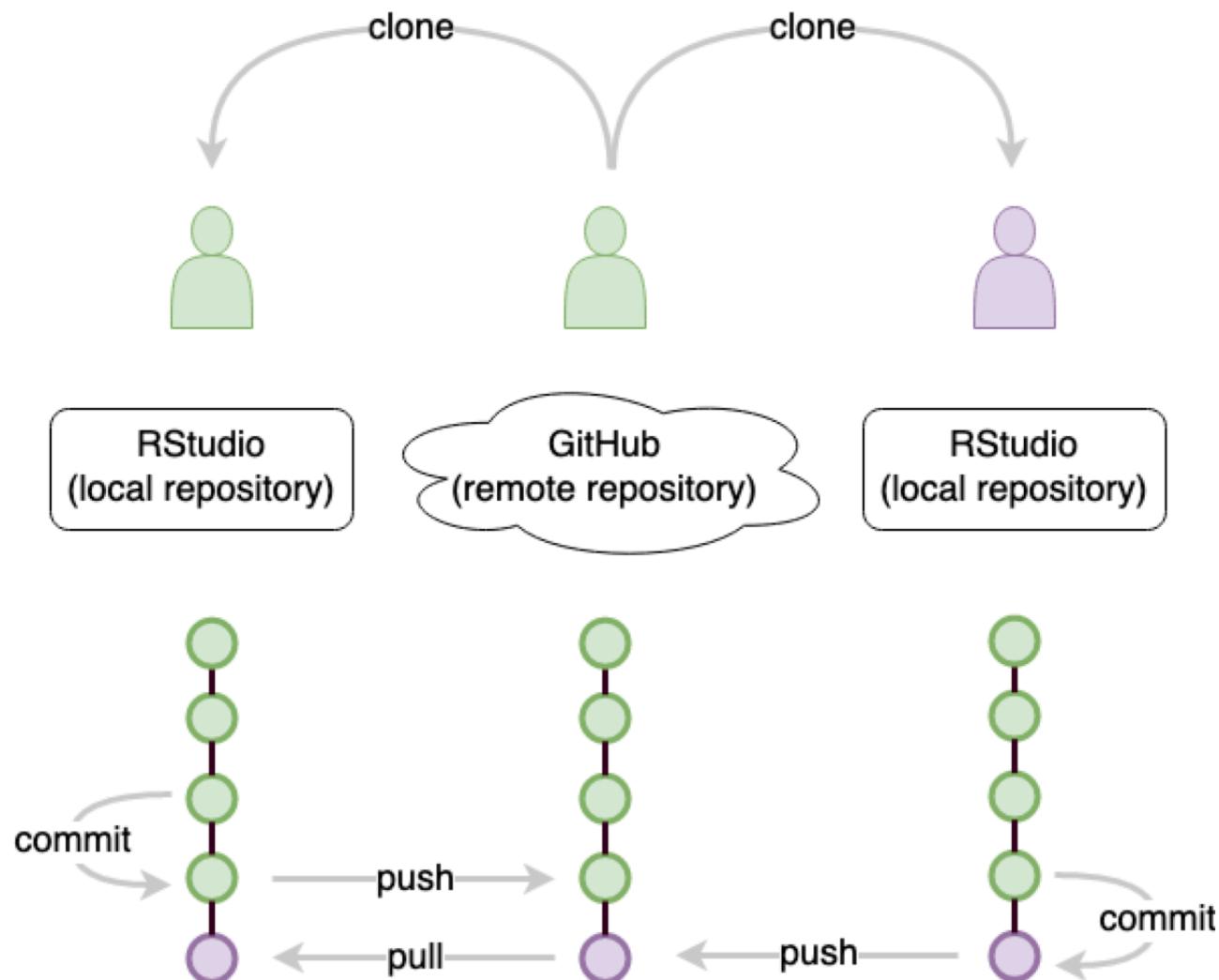
update: git push



git ???



get updates: git pull



Learning Objectives (for this week)

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

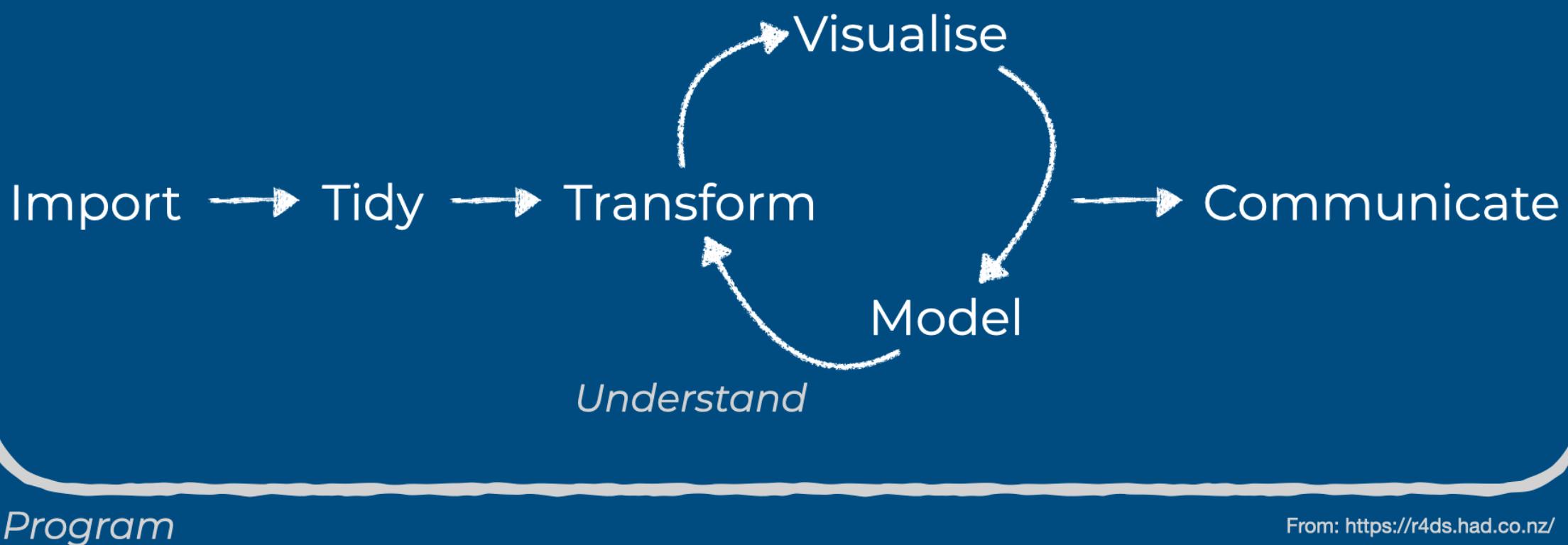
Data Science Lifecycle

Deep End

[via GIPHY](#)

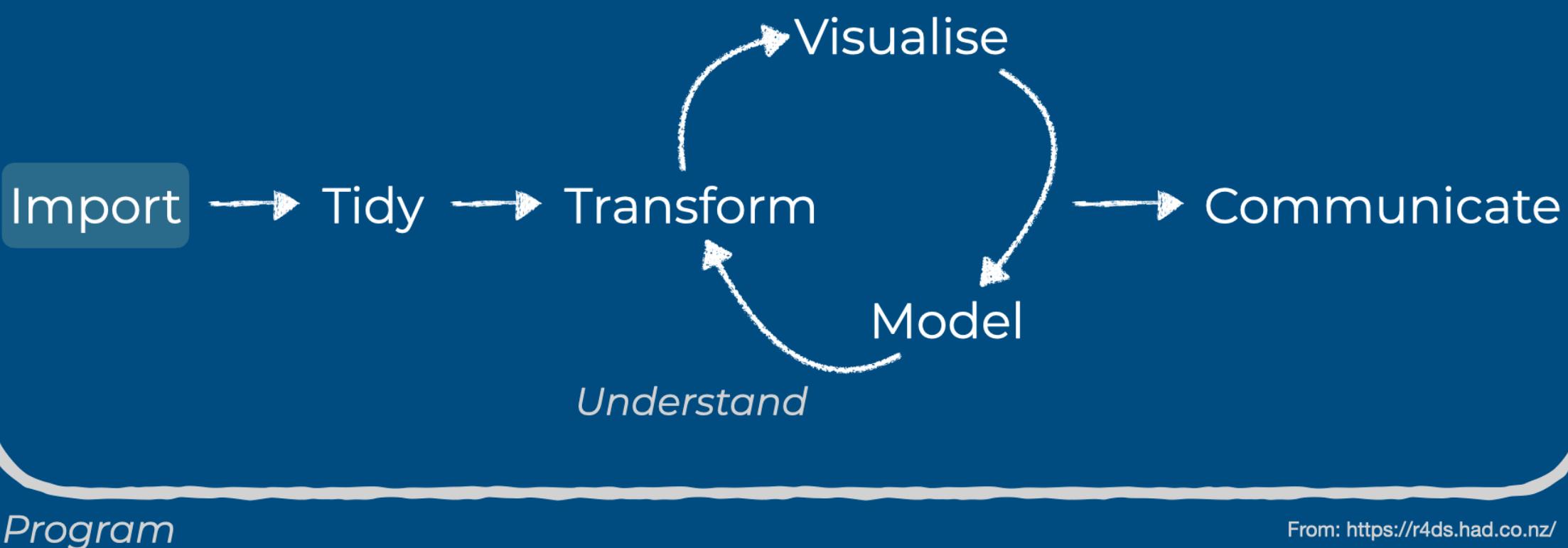
 rbtl-fs25.github.io/website/

Data Science Lifecycle



Data Science Lifecycle

Get your data into R

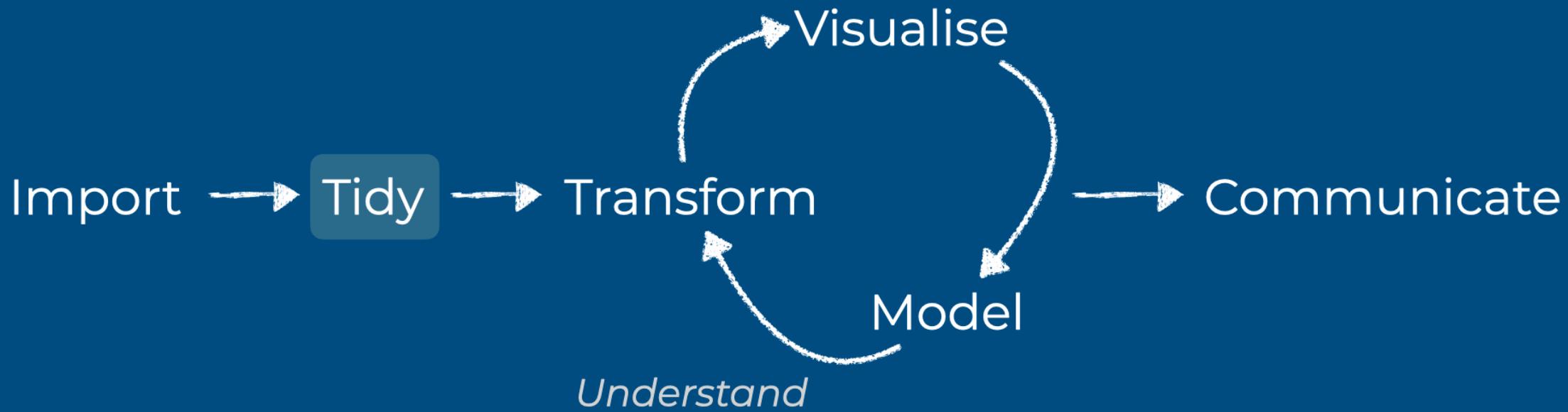


Program

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

Data Science Lifecycle

Store your data in a consistent form

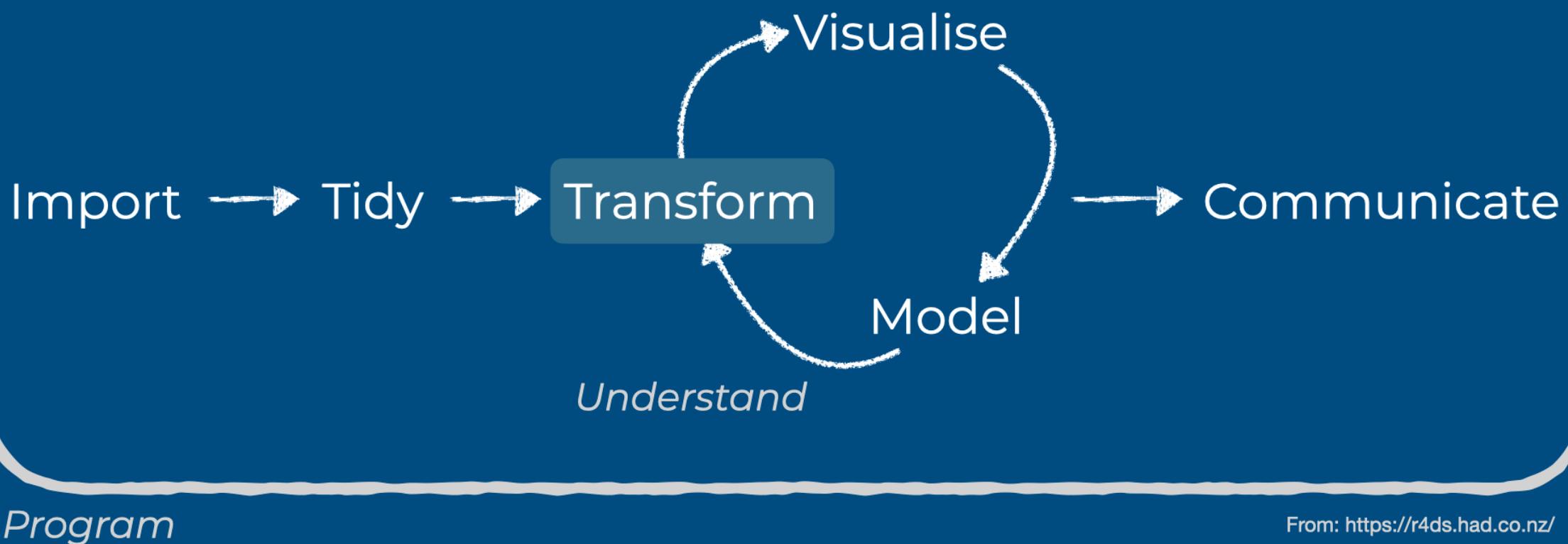


Program

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

Data Science Lifecycle

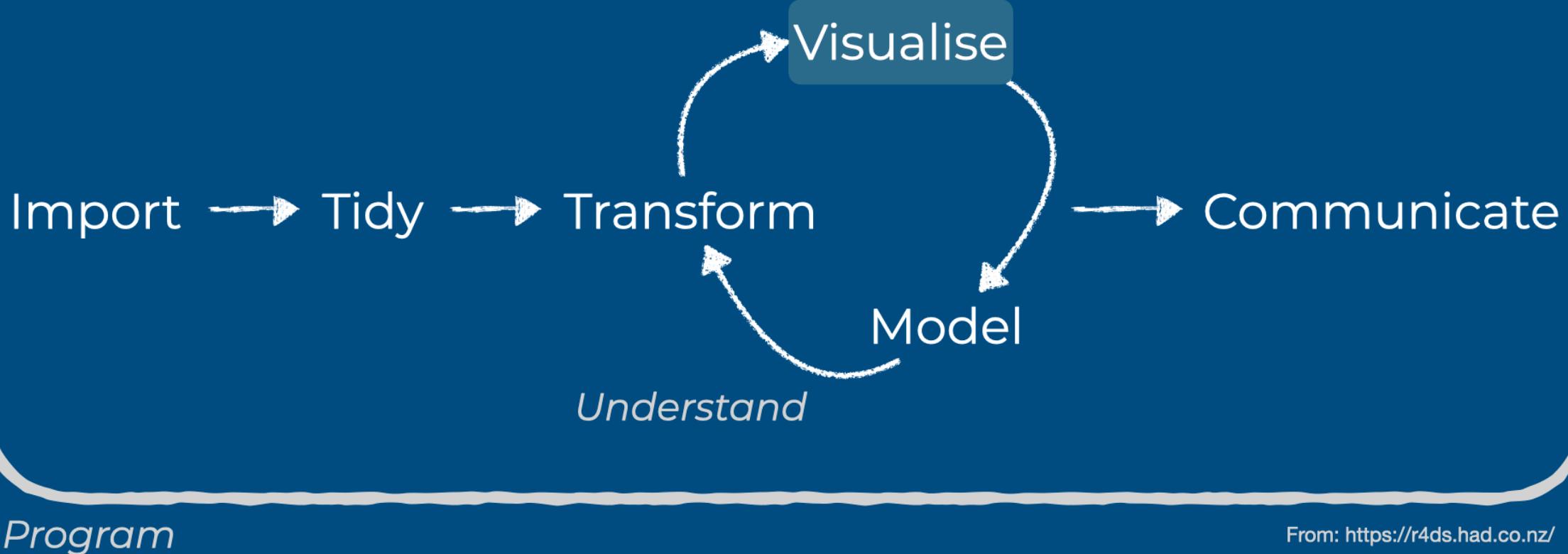
Narrow down + Create new variables + Summary stats



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

Data Science Lifecycle

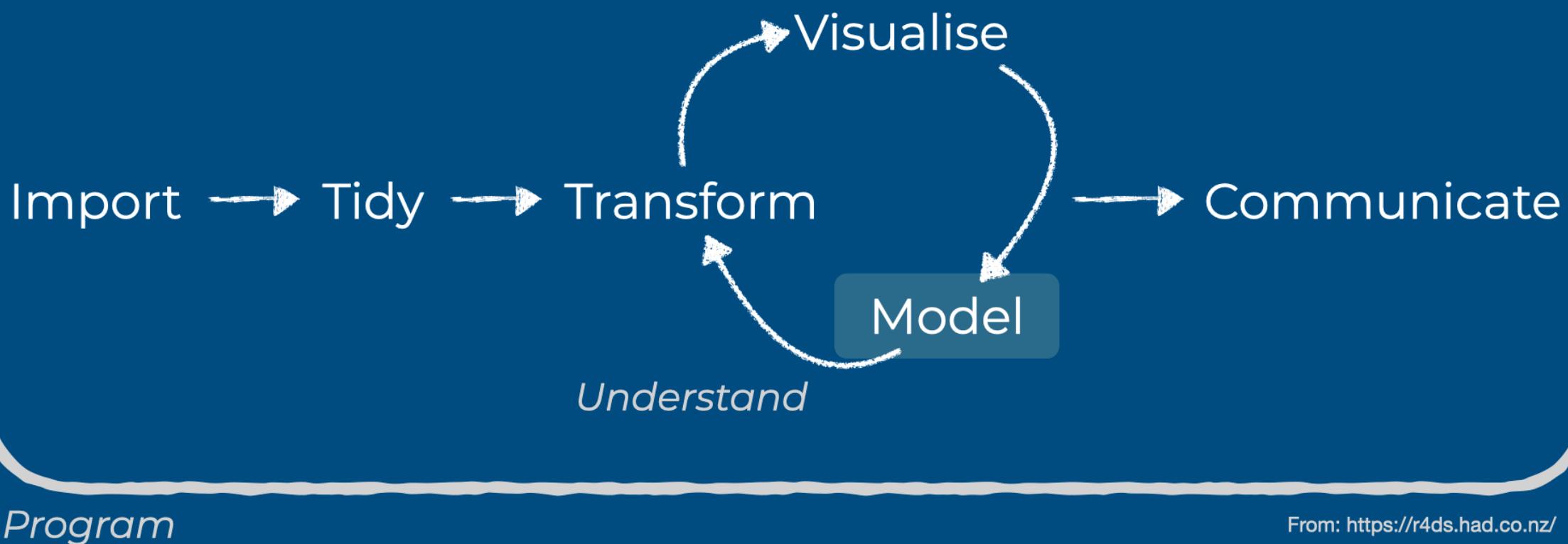
Explore your with visual representations



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

Data Science Lifecycle

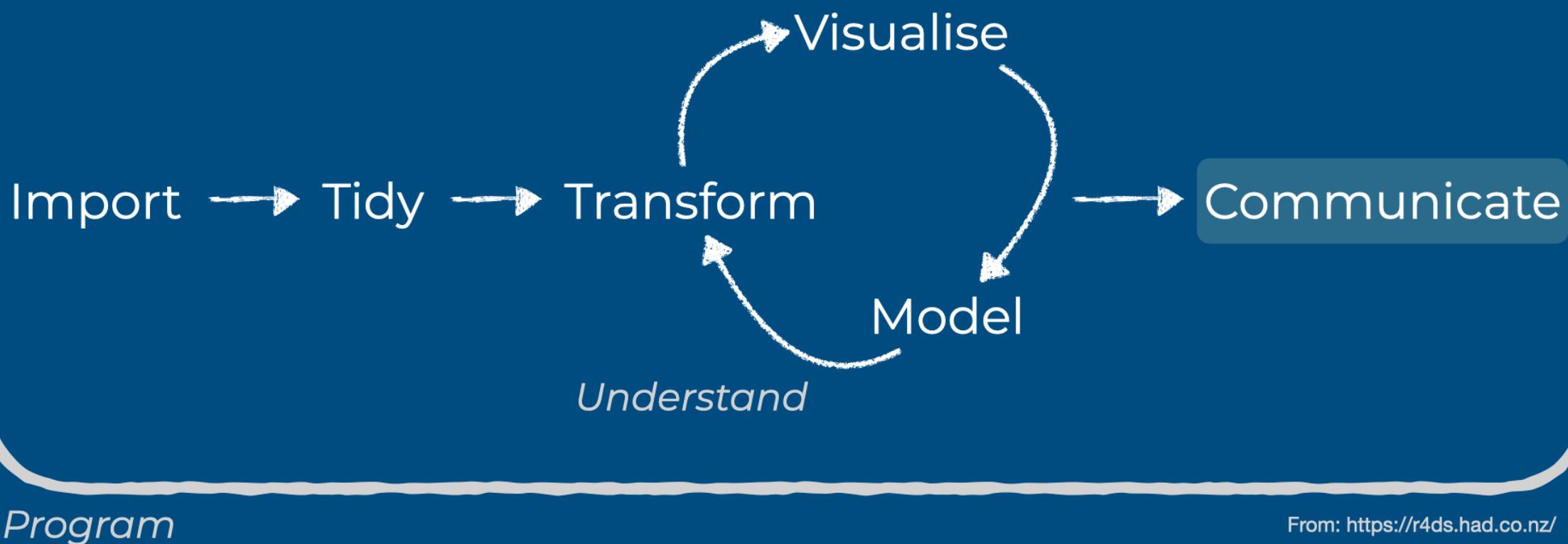
Explore your with visual representations



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

Data Science Lifecycle

Share your findings with others

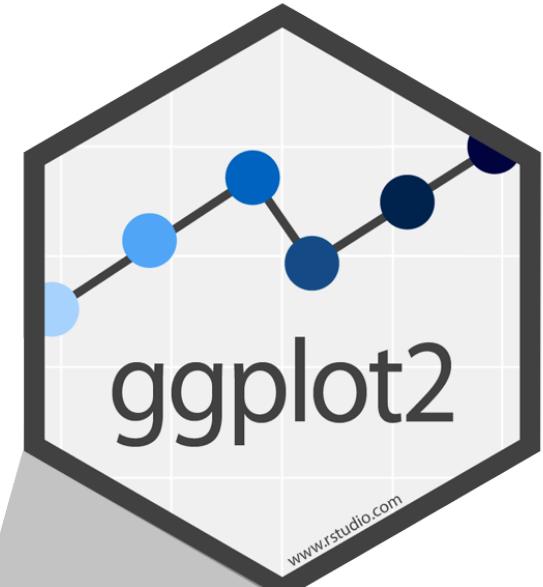


From: <https://r4ds.had.co.nz/>

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!

Take a break

Please get up and move! Let your emails rest in peace.



Code structure

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

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

Code structure

```
1 ggplot()
```

Code structure

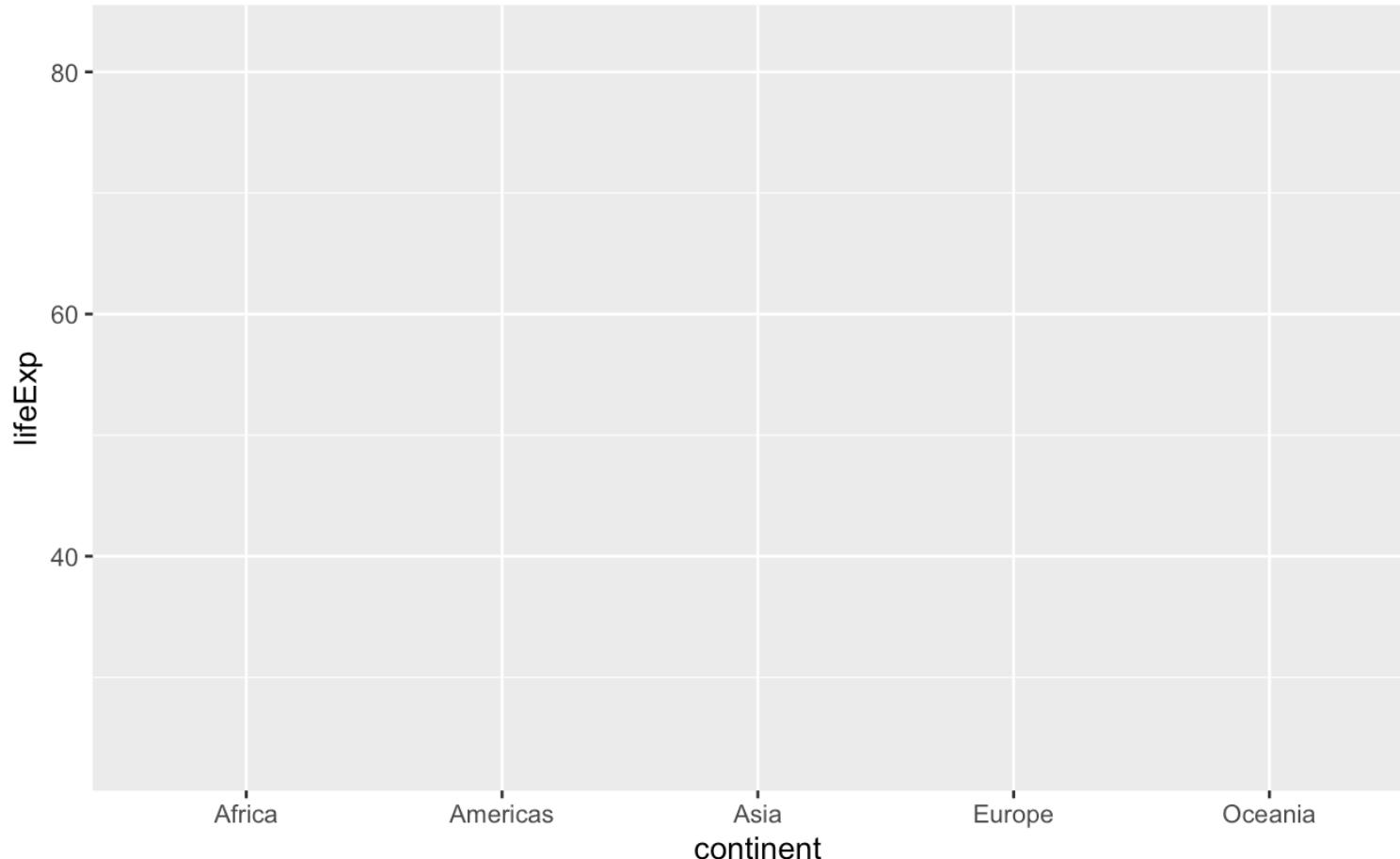
```
1 ggplot(data = gapminder)
```

Code structure

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

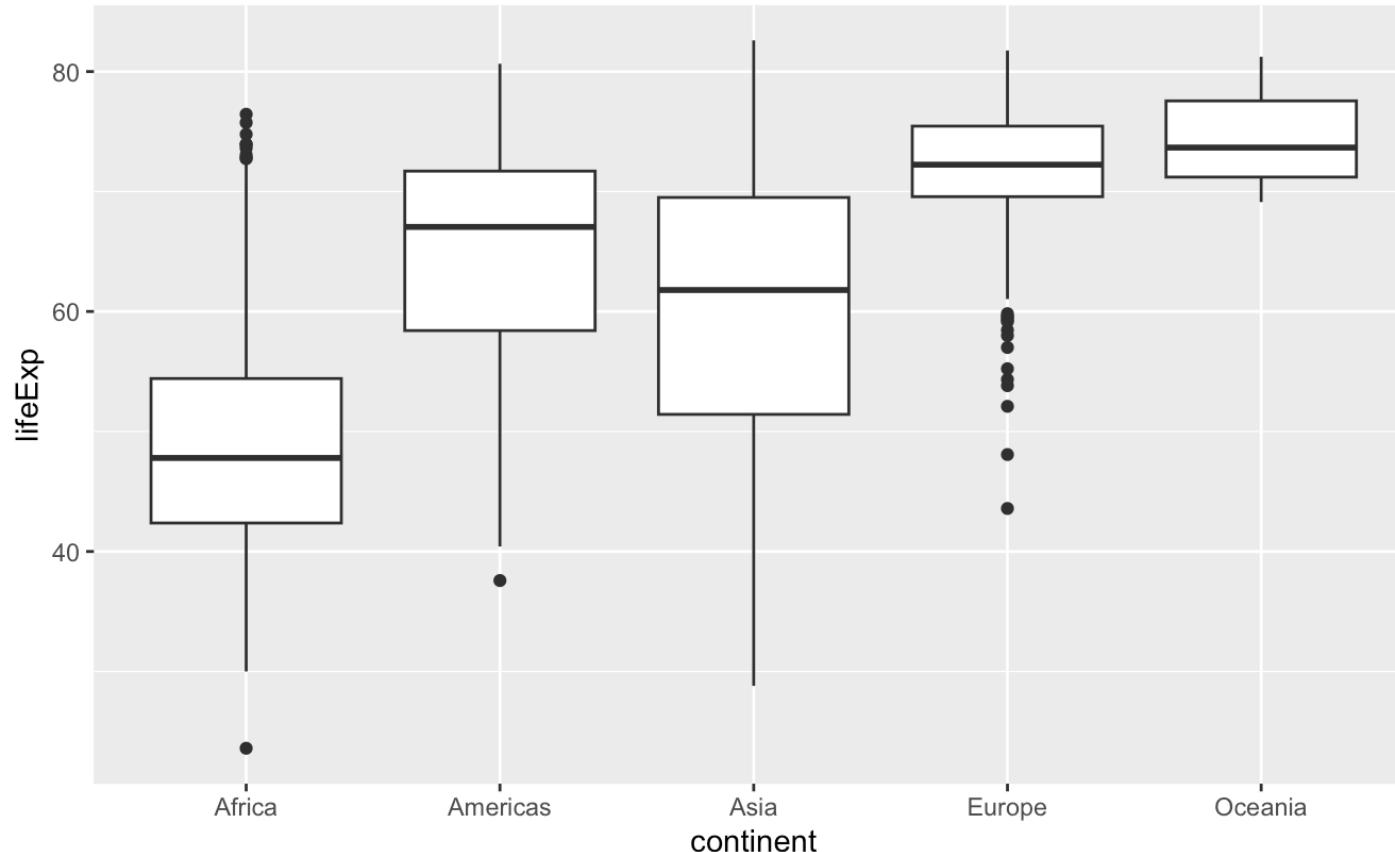
Code structure

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



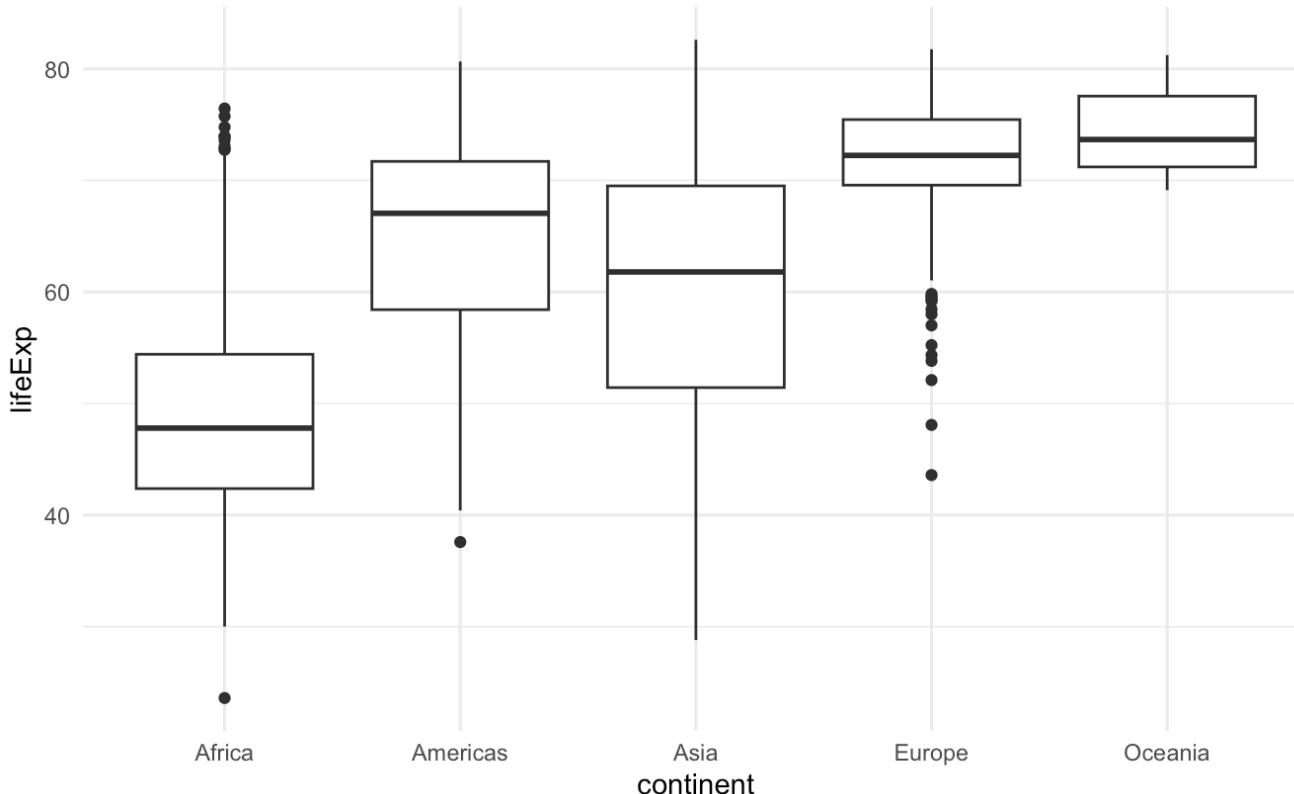
Code structure

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



Code structure

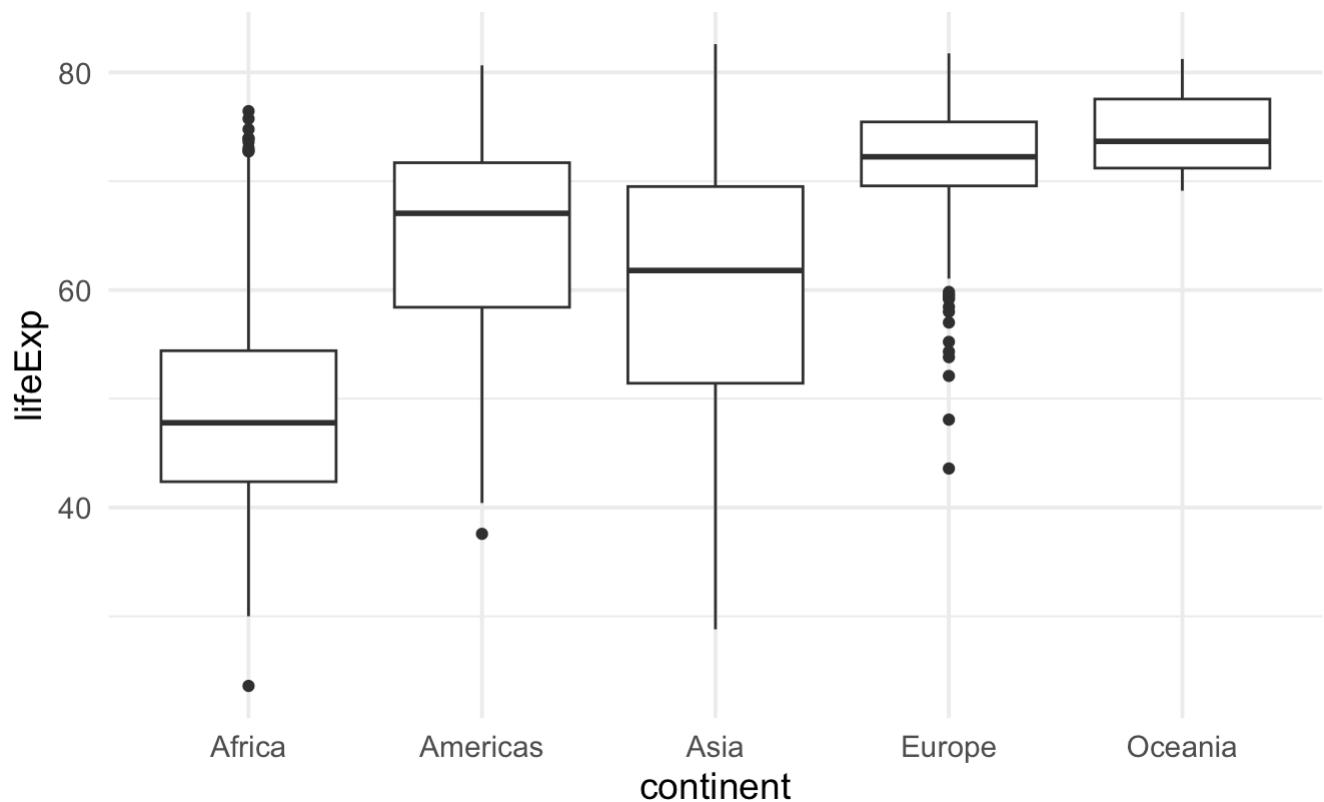
```
1 ggplot(data = gapminder,  
2         mapping = aes(x = continent,  
3                           y = lifeExp)) +  
4   geom_boxplot() +  
5   theme_minimal()
```



Polls

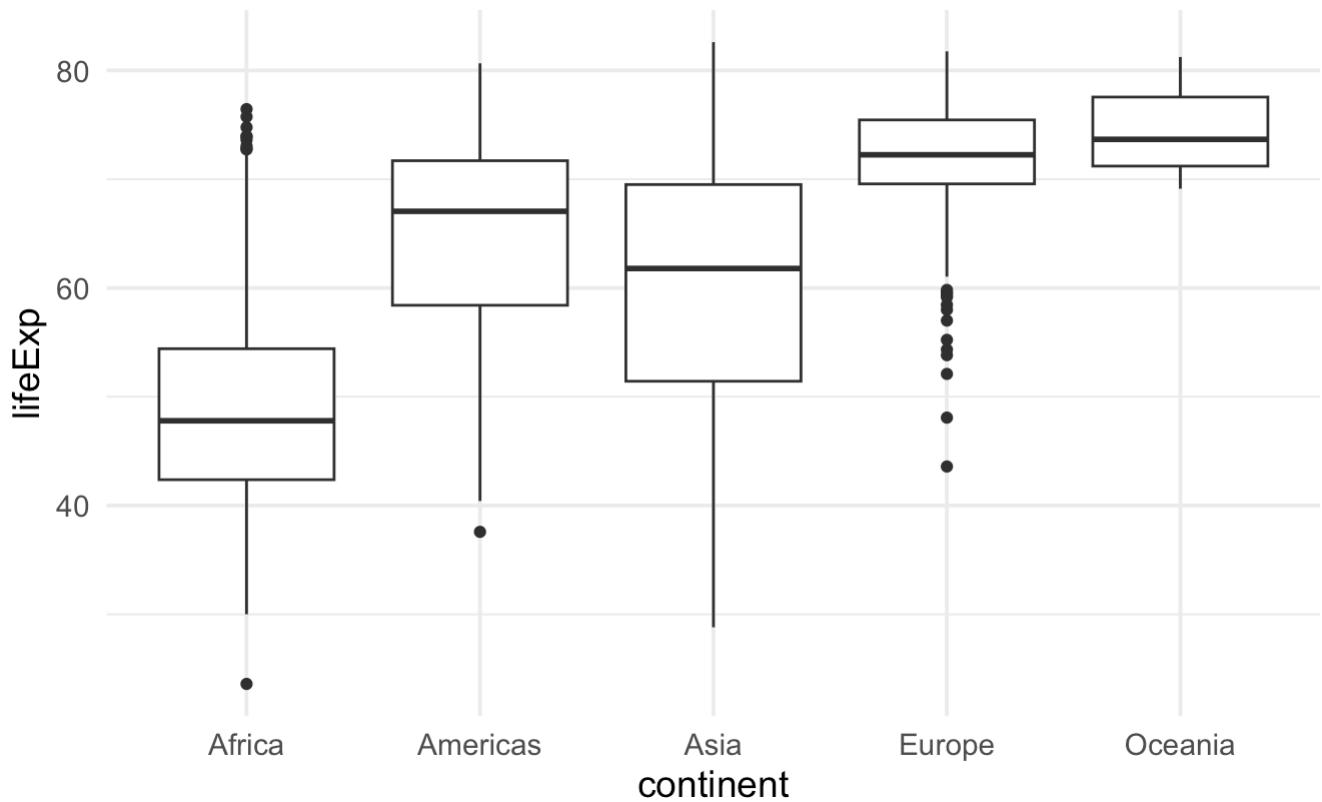
Poll 1: What does the thick line inside the box of a boxplot represent?

1. I don't know
2. the mean of the observations
3. the middle of the box
4. the median of the observations



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

1. I don't know
2. 25%
3. depends on the median
4. 50%



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

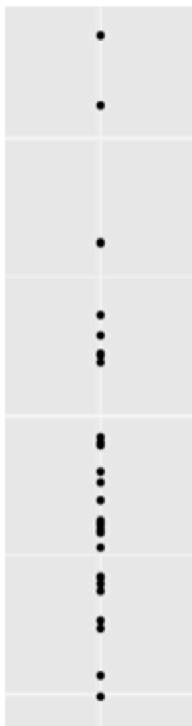
1. I don't know
2. The median is the most frequently occurring value in a dataset.
3. The median is the sum of all values in a dataset divided by the number of observations.
4. The median is the point above and below which half (50%) of the observations falls.

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

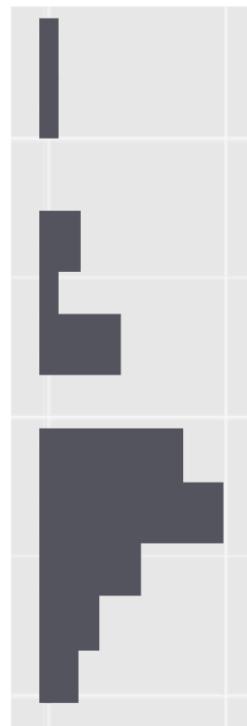
1. I don't know
2. The maximum
3. The standard deviation
4. The median

Boxplot, explained

The actual values in a distribution



How a histogram would display the values (rotated)



How a boxplot would display the values

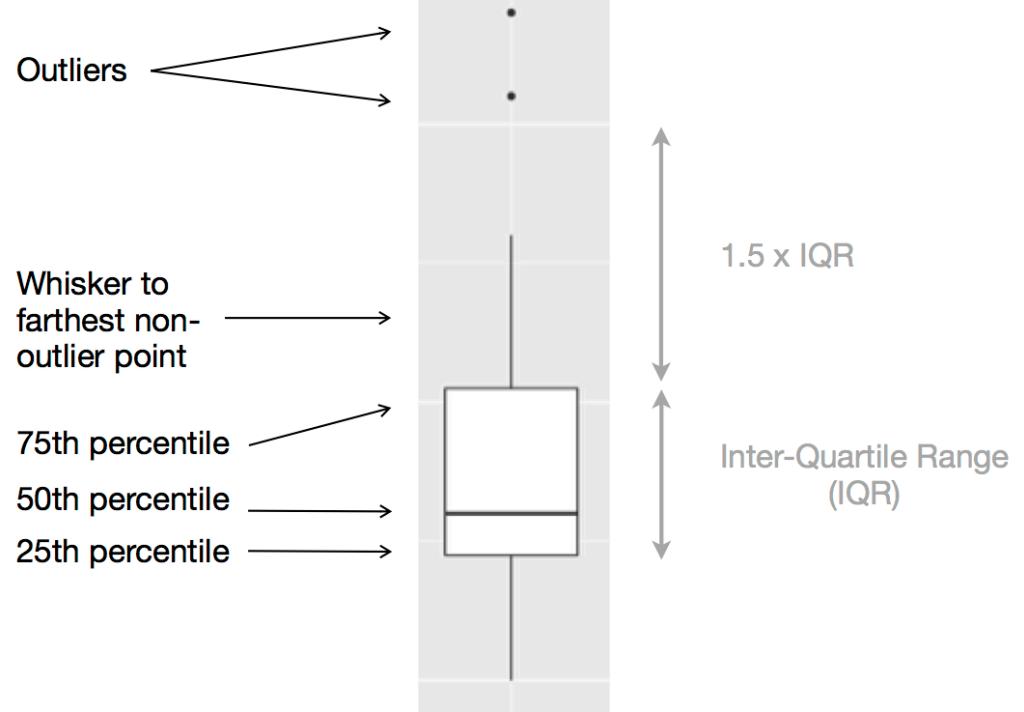


Figure 1: Diagram depicting how a boxplot is created.

Our turn: md-02-exercises

1. Open [posit.cloud](#) in your browser (use your bookmark).
2. Open the [rbtl-fs25 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.

Take a break

Please get up and move! Let your emails rest in peace.



Visualizing 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

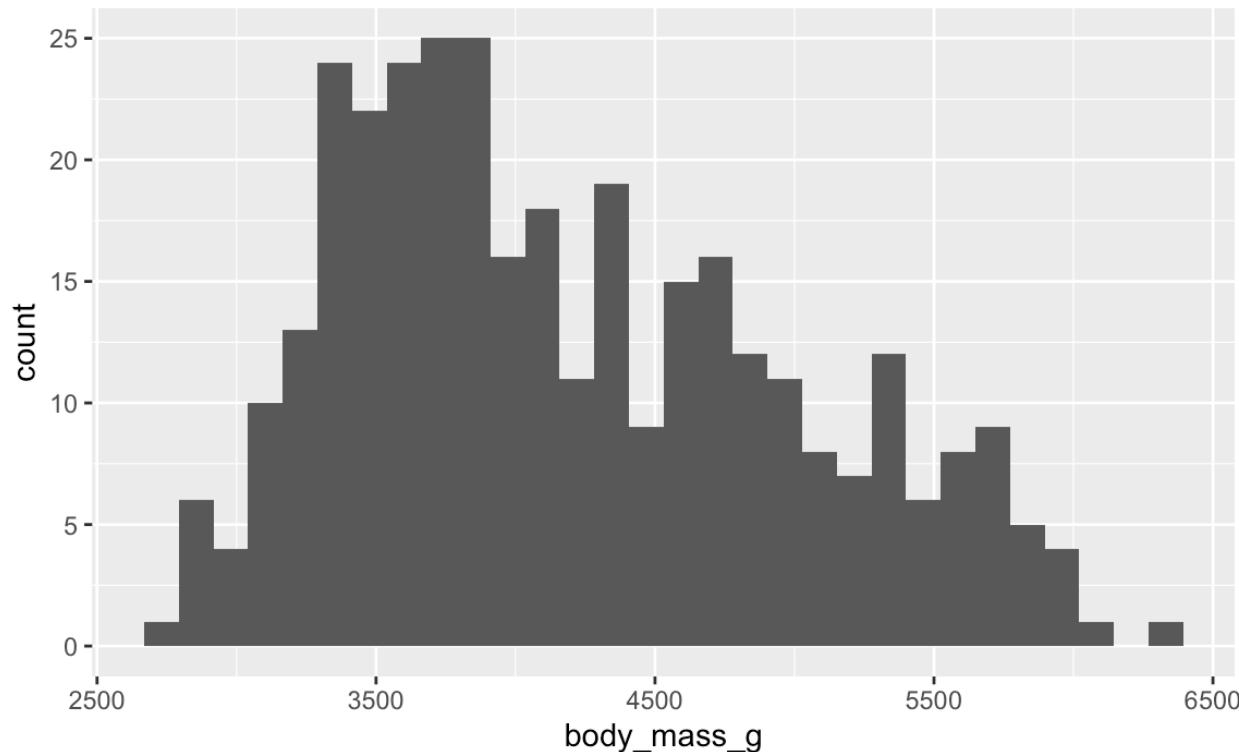
categorical variables

- 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

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



Barplot

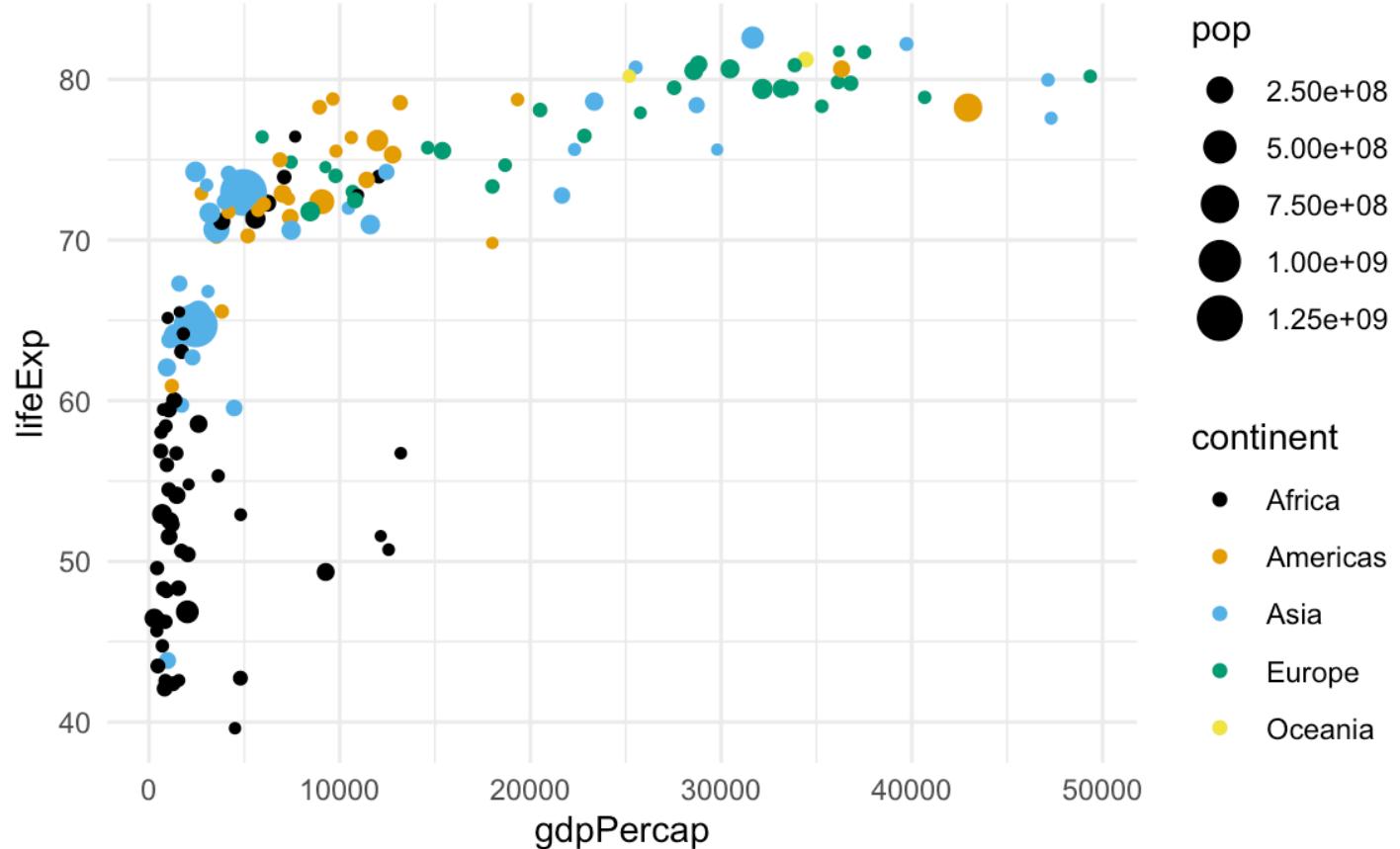
- for visualizing distribution of categorical (non-numerical) variables

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

Scatterplot

- for visualizing relationships between two continuous (numerical) variables

```
1 ggplot(data = gapminder_2007,
2         mapping = aes(x = gdpPercap,
3                          y = lifeExp,
4                          size = pop,
5                          color = continent)) +
6   geom_point() +
7   scale_color_colorblind() +
8   theme_minimal()
```



Your turn: md-02-exercises

1. Open [posit.cloud](#) in your browser (use your bookmark).
2. Open the [rbtl-fs25 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

Homework assignments

module 2

Module 2 documentation

rbtl-fs25.github.io/website/modules/md-02.html

Module 2

Data science lifecycle & Exploratory data analysis using visualization

Are you ready for some data visualisations? This week is all about exploring data with `ggplot2` R package. We will also learn about the data science lifecycle.

◎ Learning Objectives

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

Homework due date

- Homework assignment due: Wednesday, March 5th
- Correction & feedback phase up to: Tuesday, March 11th

Wrap-up

Thanks! 🌻

Slides created via revealjs and Quarto:

<https://quarto.org/docs/presentations/revealjs/> Access slides as
[PDF on GitHub](#)

All material is licensed under [Creative Commons Attribution Share Alike 4.0 International.](#)