Data science lifecycle & Exploratory data analysis using visualization

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

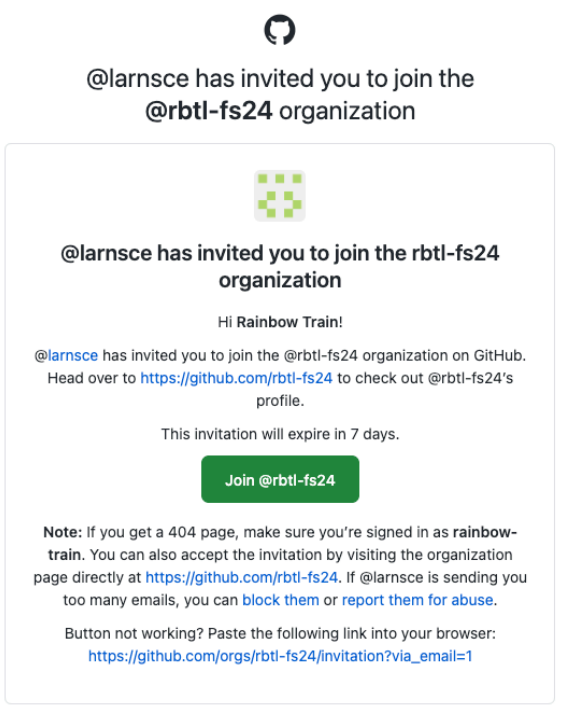
Lars Schöbitz

Feb 29, 2024

rbtl-fs24.github.io/website/

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).



@larnsce has invited you to join the **@rbtl-fs24** organization

@larnsce has invited you to join the rbtl-fs24 organization

Hi **Rainbow Train**!

@larnsce has invited you to join the @rbtl-fs24 organization on GitHub. Head over to https://github.com/rbtl-fs24 to check out @rbtl-fs24's profile.

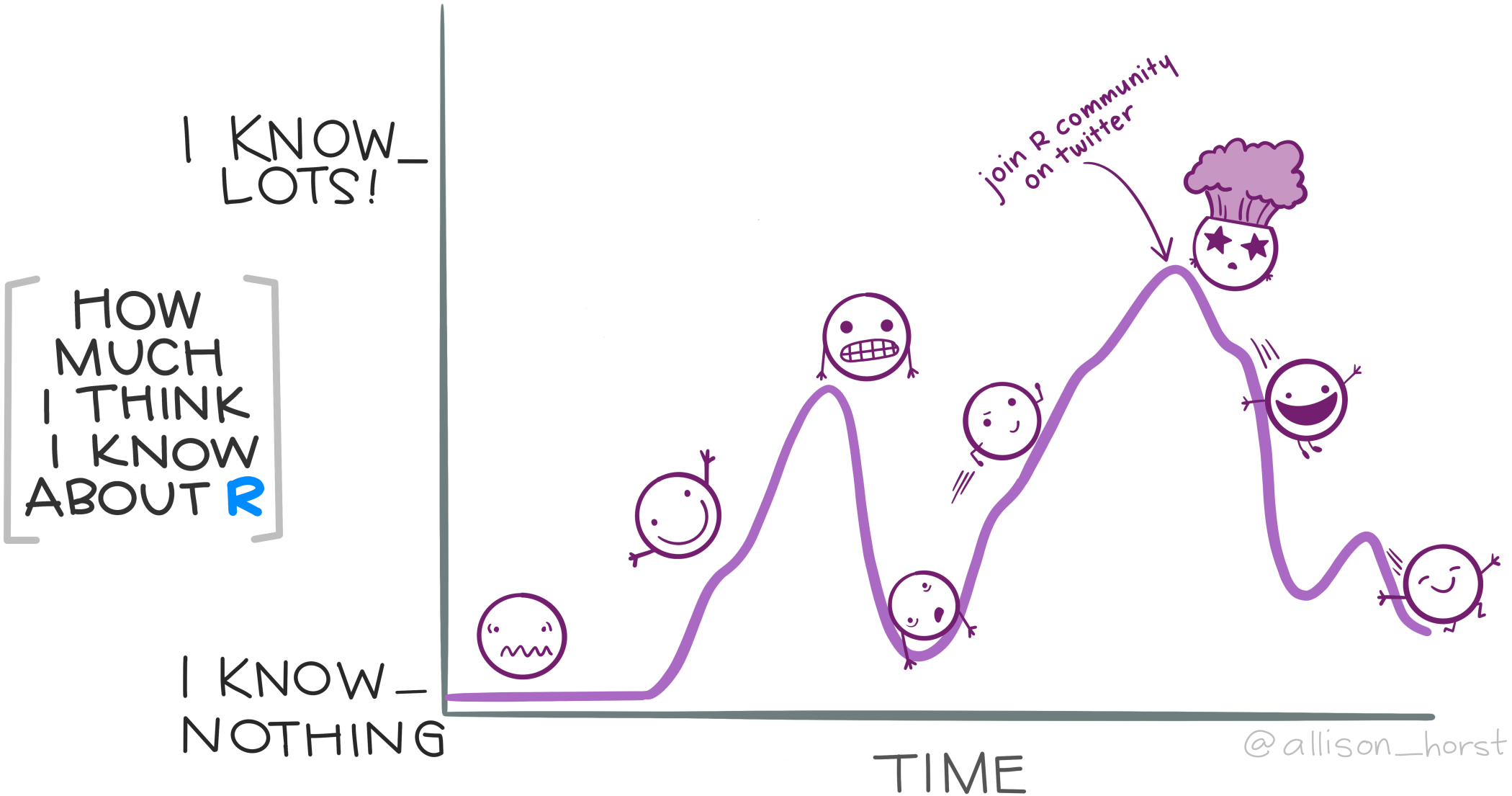
This invitation will expire in 7 days.

Join @rbtl-fs24

Note: If you get a 404 page, make sure you're signed in as **rainbow-train**. You can also accept the invitation by visiting the organization page directly at https://github.com/rbtl-fs24. If @larnsce is sending you too many emails, you can block them or report them for abuse.

Button not working? Paste the following link into your browser: https://github.com/orgs/rbtl-fs24/invitation?via_email=1

rbtl-fs24.github.io/website/



rbtl-fs24.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

rbtl-fs24.github.io/website/

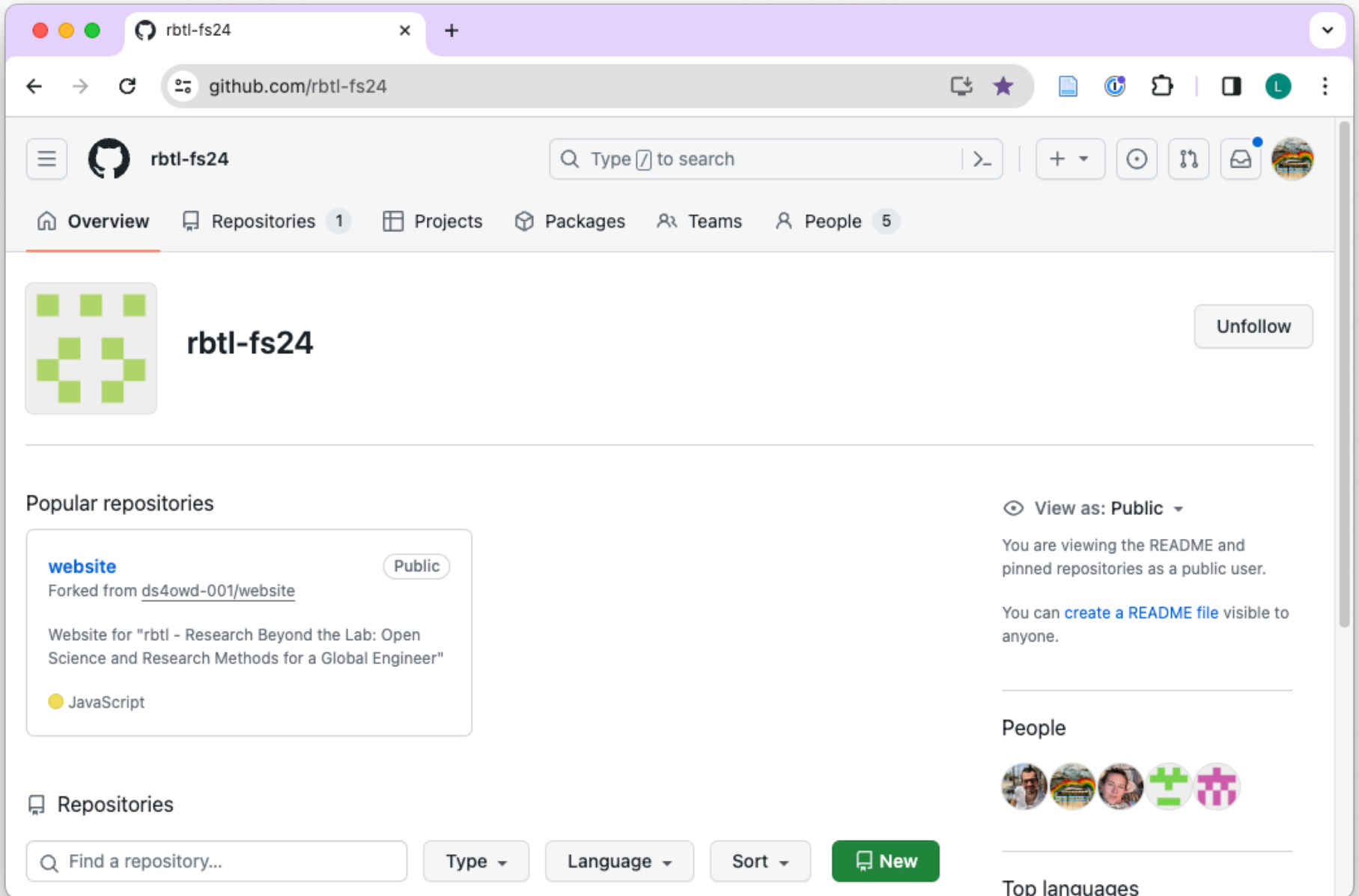
Interaction with GitHub

rbtl-fs24.github.io/website/

open GitHub organisation

Bookmark this link in your browser!

github.com/rbtl-fs24

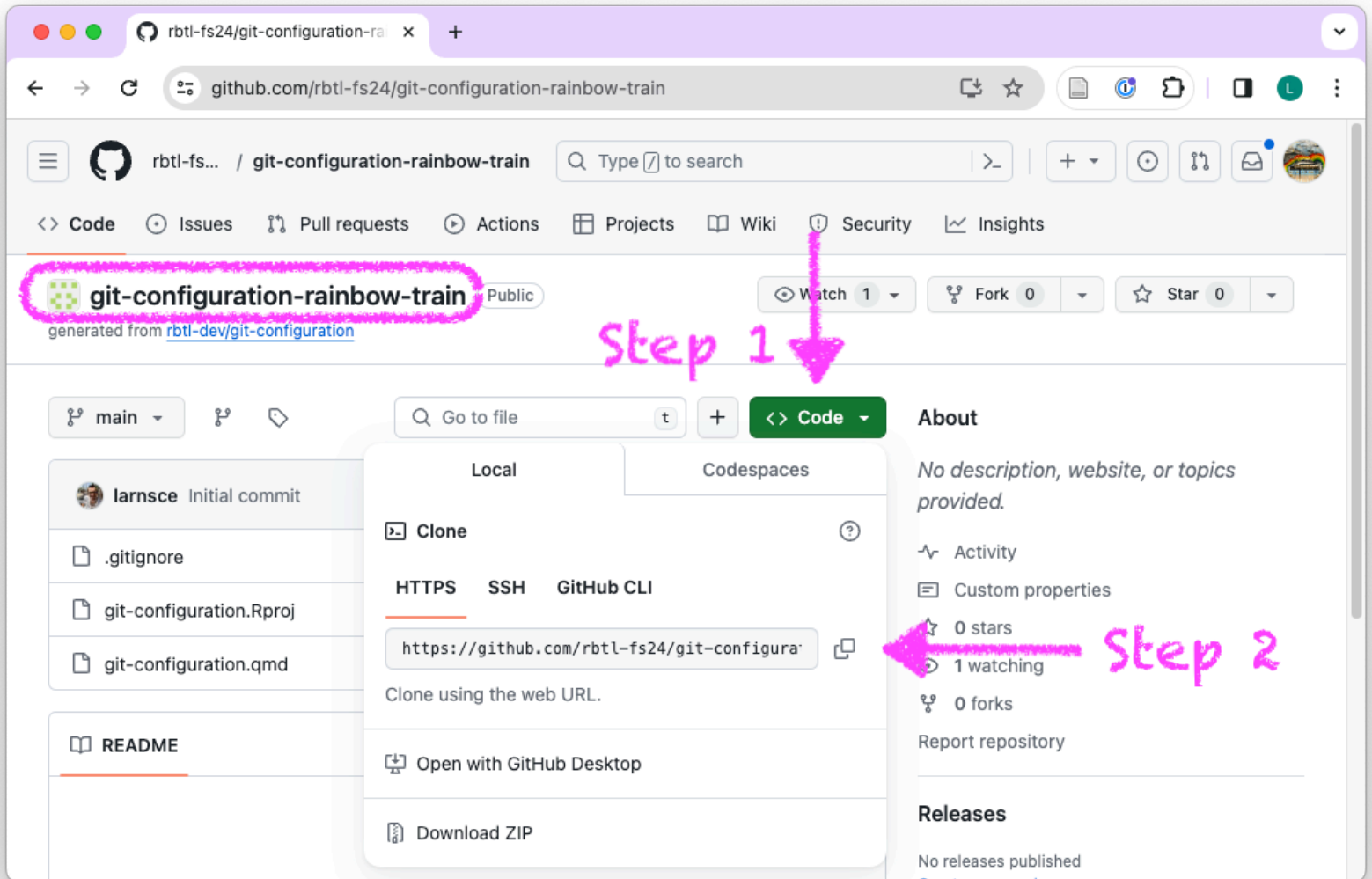


rbtl-fs24.github.io/website/

on GitHub Organisation

- Search for your username in search bar under `Repositories`

on your repository

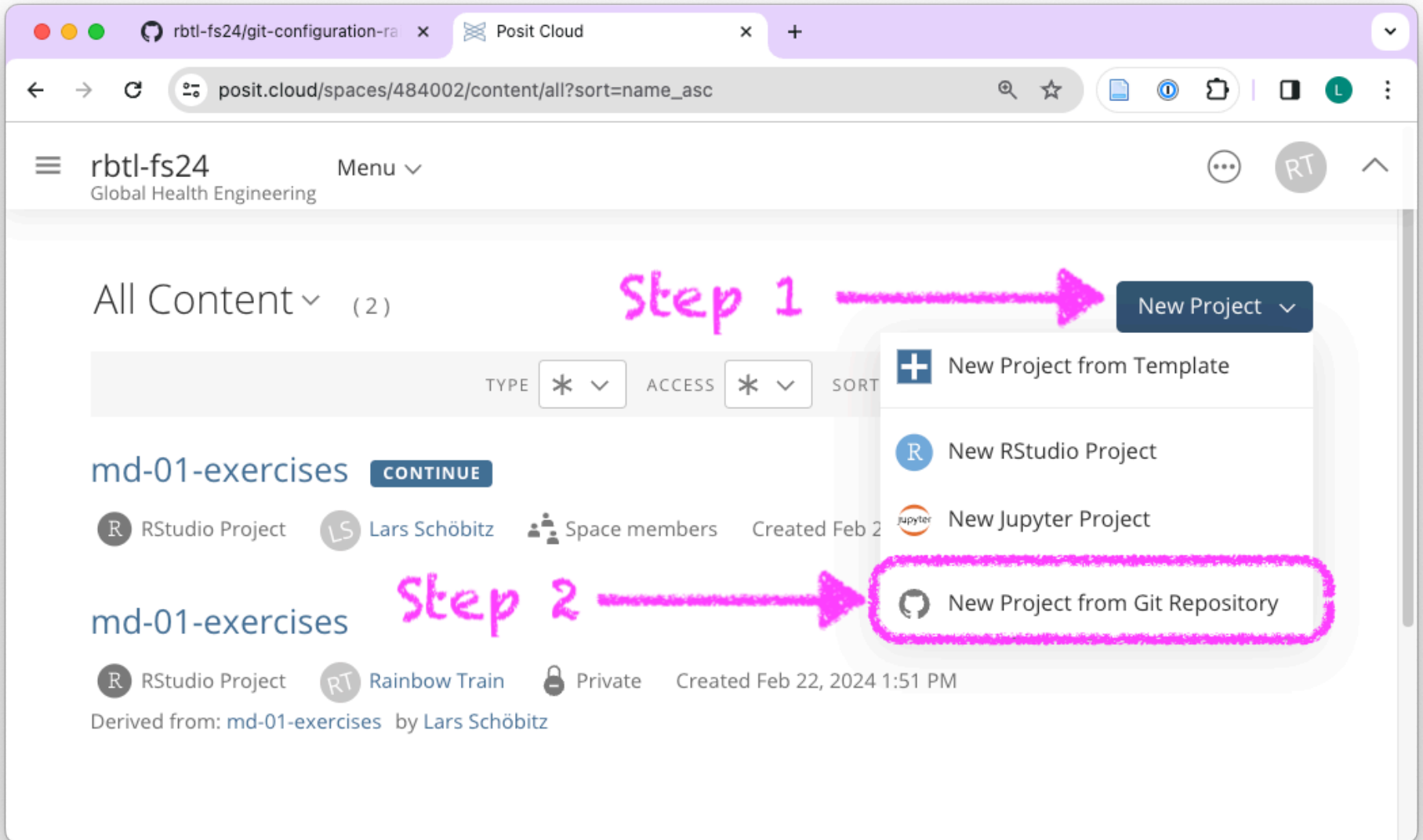


rbtl-fs24.github.io/website/

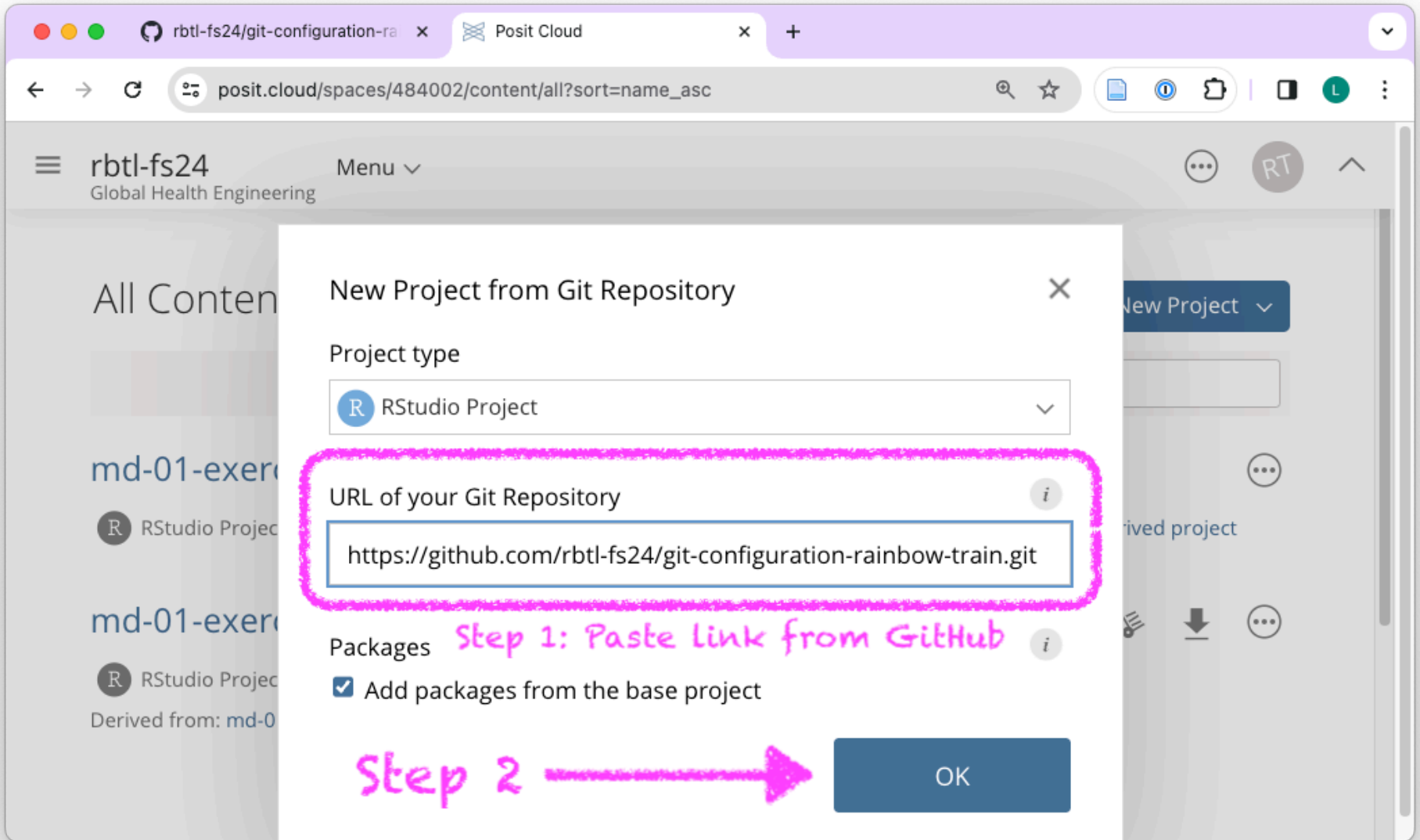
on Posit Cloud

Bookmark this link in your browser!

posit.cloud/spaces/484002/content/


rbtl-fs24/git-configuration-ra
Posit Cloud
posit.cloud/spaces/484002/content/all?sort=name_asc
rbtl-fs24
Global Health Engineering
Menu
RT
All Content (2)
Step 1
New Project
New Project from Template
New RStudio Project
New Jupyter Project
New Project from Git Repository
TYPE
ACCESS
SORT
md-01-exercises
CONTINUE
RStudio Project
Lars Schöbitz
Space members
Created Feb 2
Step 2
md-01-exercises
RStudio Project
Rainbow Train
Private
Created Feb 22, 2024 1:51 PM
Derived from: md-01-exercises by Lars Schöbitz

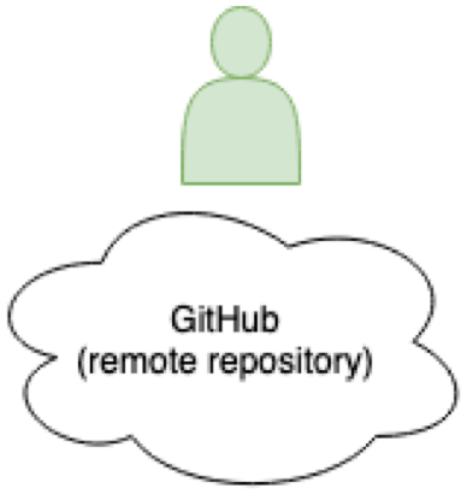
on Posit Cloud

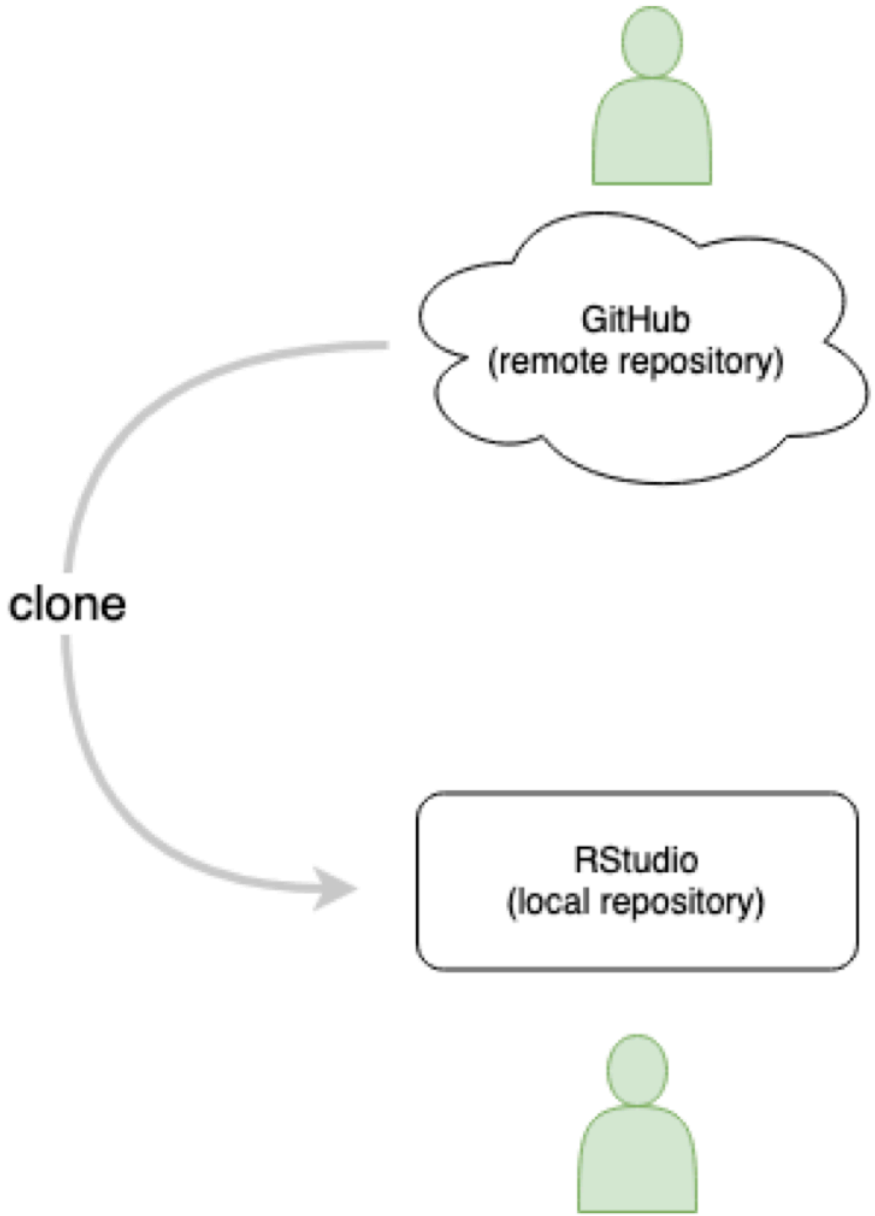

rbtl-fs24/git-configuration-ra
Posit Cloud
posit.cloud/spaces/484002/content/all?sort=name_asc
rbtl-fs24
Global Health Engineering
Menu
RT
All Conten
New Project
md-01-exer
RStudio Projec
ived project
md-01-exer
RStudio Projec
Derived from: md-0
New Project from Git Repository
Project type
RStudio Project
URL of your Git Repository
https://github.com/rbtl-fs24/git-configuration-rainbow-train.git
Step 1: Paste link from GitHub
Packages
Add packages from the base project
Step 2
OK

Your turn: Introduce yourself to git

1. Open a web browser on your laptop.
2. Navigate to the course website: rbtl-fs24.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

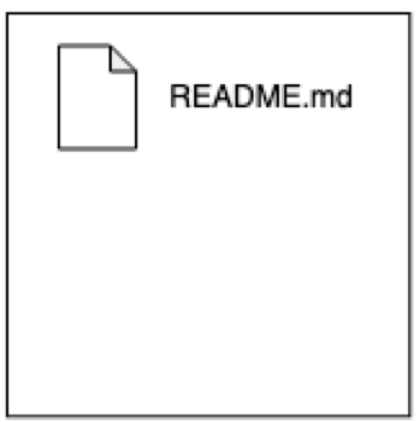
Version Control - Terminology

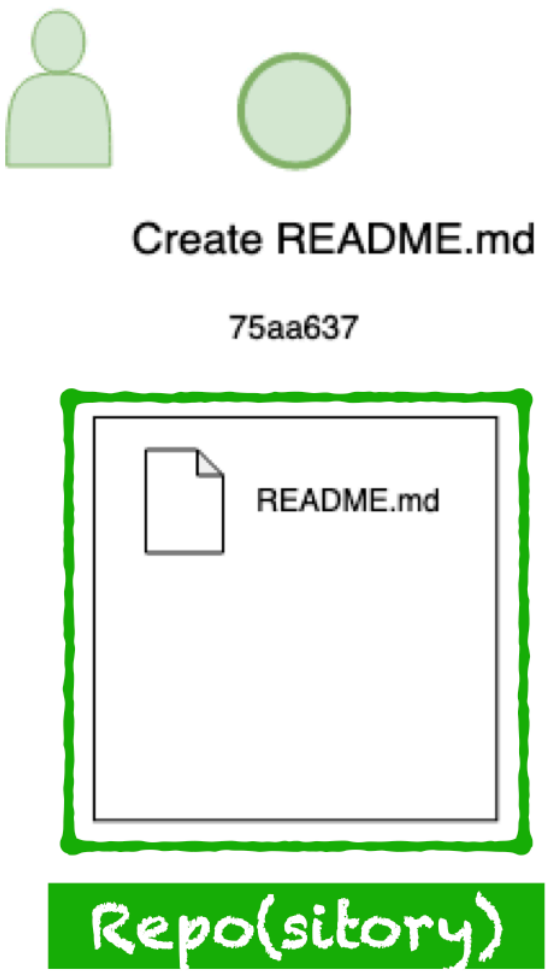

GitHub
(remote repository)

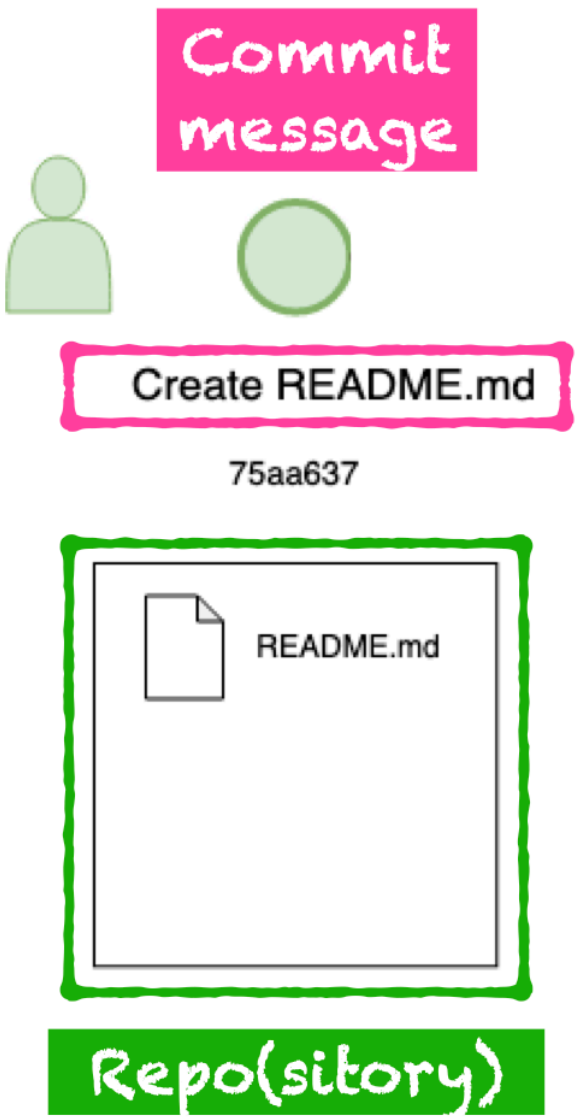

GitHub
(remote repository)
clone
RStudio
(local repository)

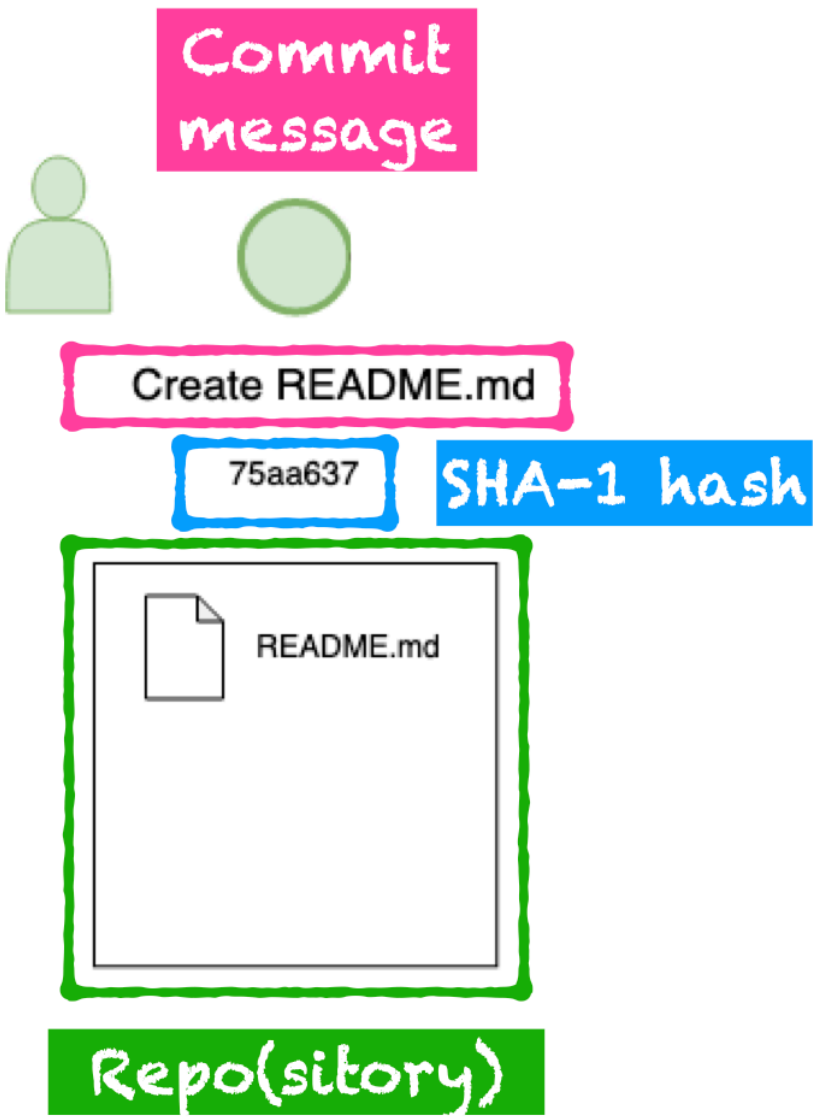
Create README.md

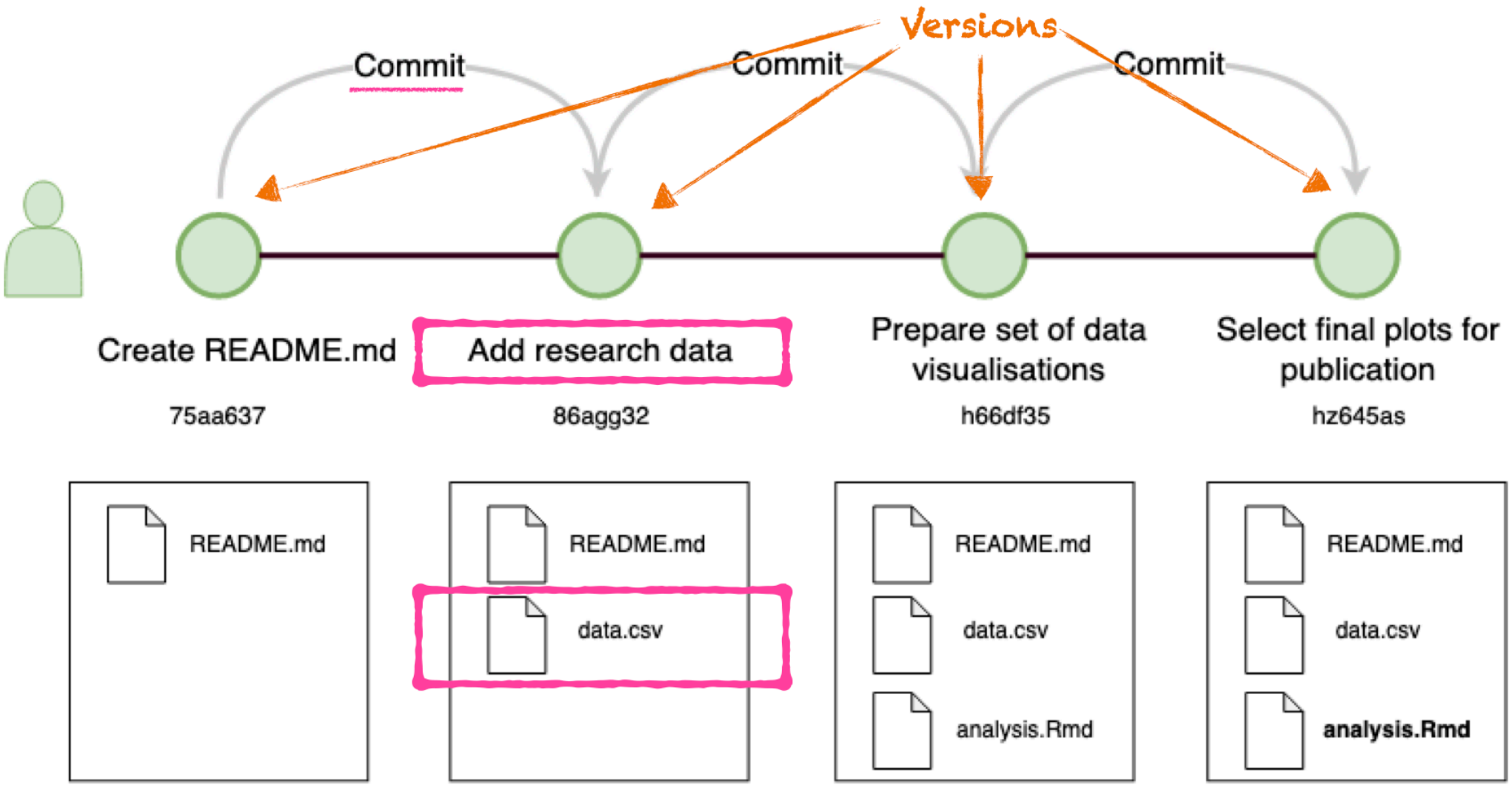
75aa637

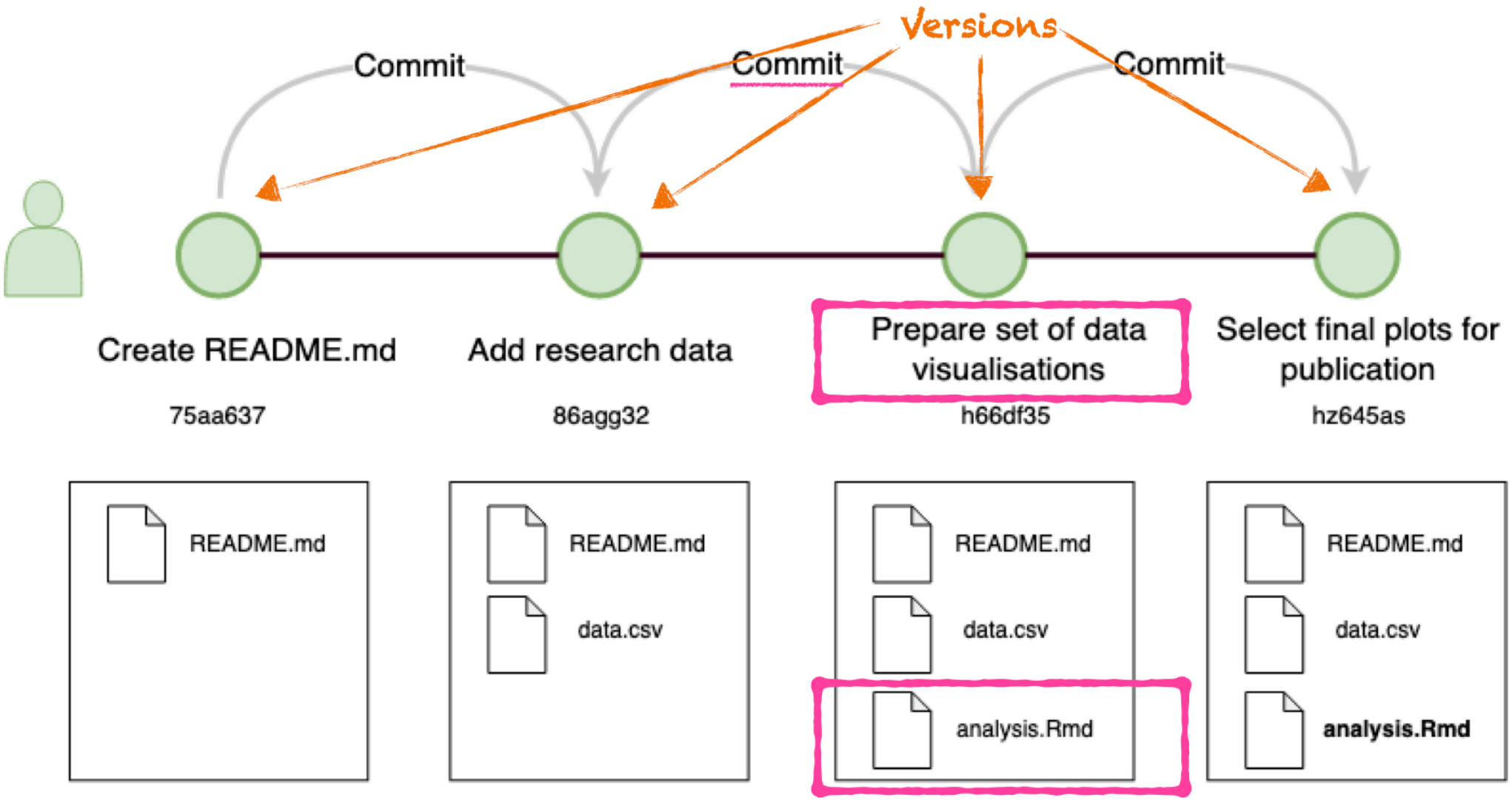



Create README.md
75aa637
README.md
Repo(sitory)

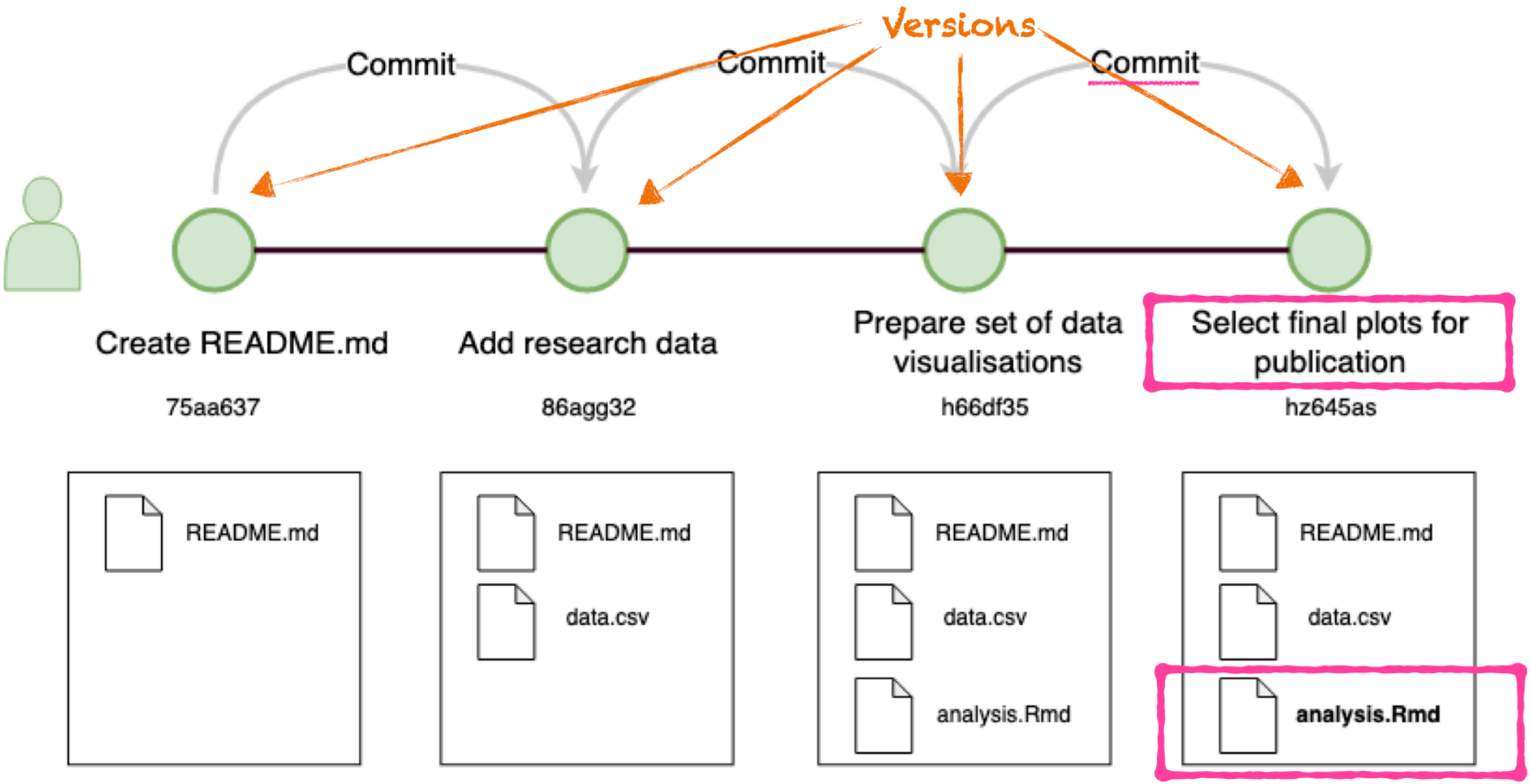

Commit message
Create README.md
75aa637
README.md
Repo(sitory)

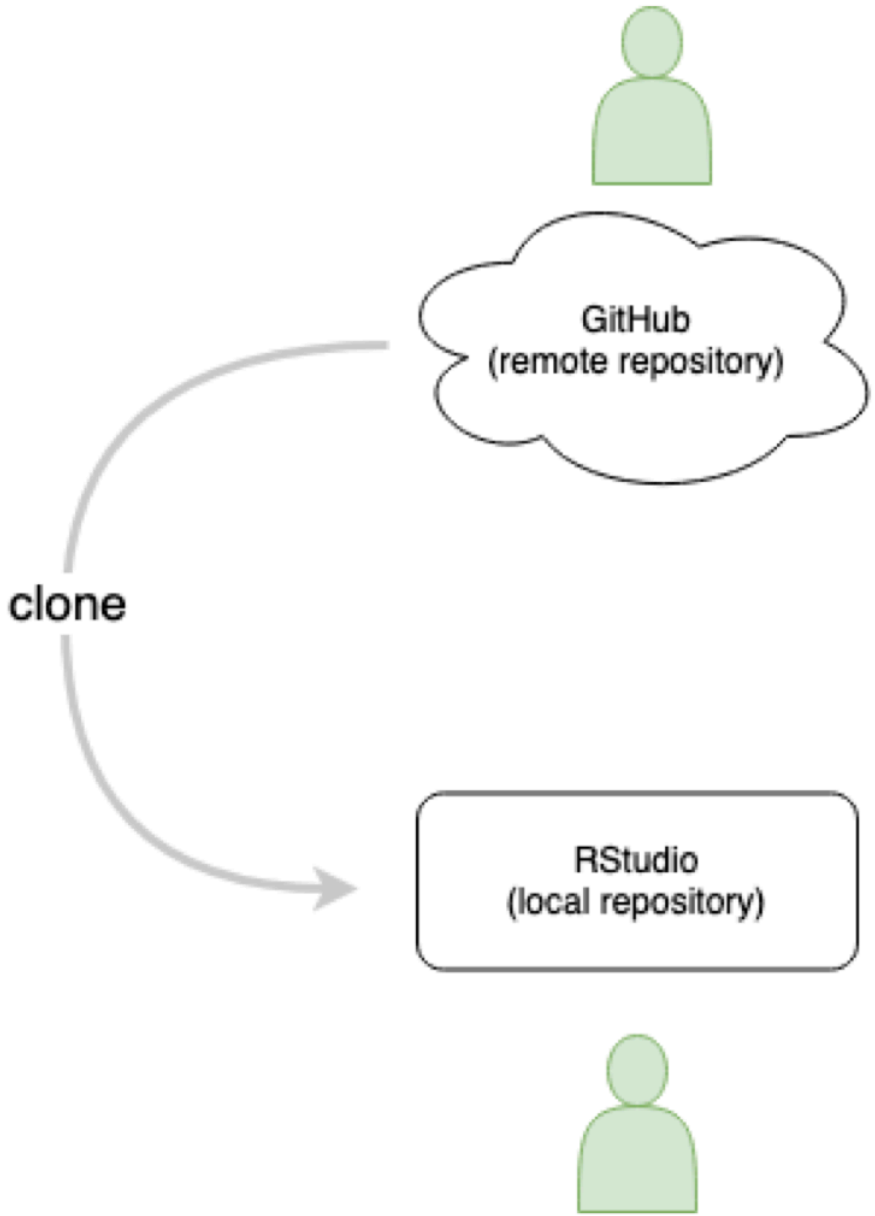

Commit message
Create README.md
75aa637
SHA-1 hash
README.md
Repo(sitory)

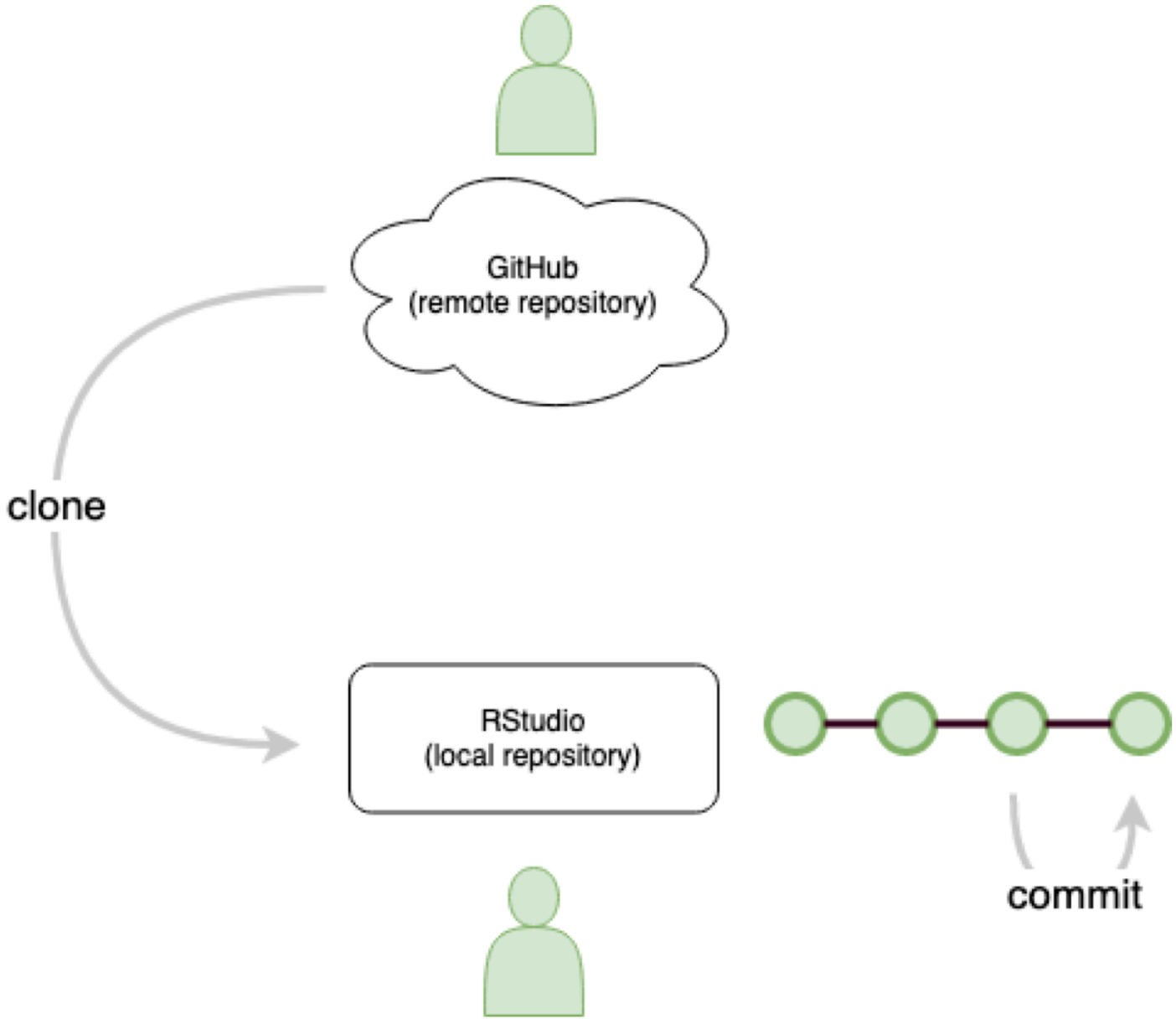

Versions
Commit
Commit
Commit
Create README.md
Add research data
Prepare set of data visualisations
Select final plots for publication
75aa637
86agg32
h66df35
hz645as
README.md
README.md
data.csv
README.md
data.csv
analysis.Rmd
README.md
data.csv
analysis.Rmd



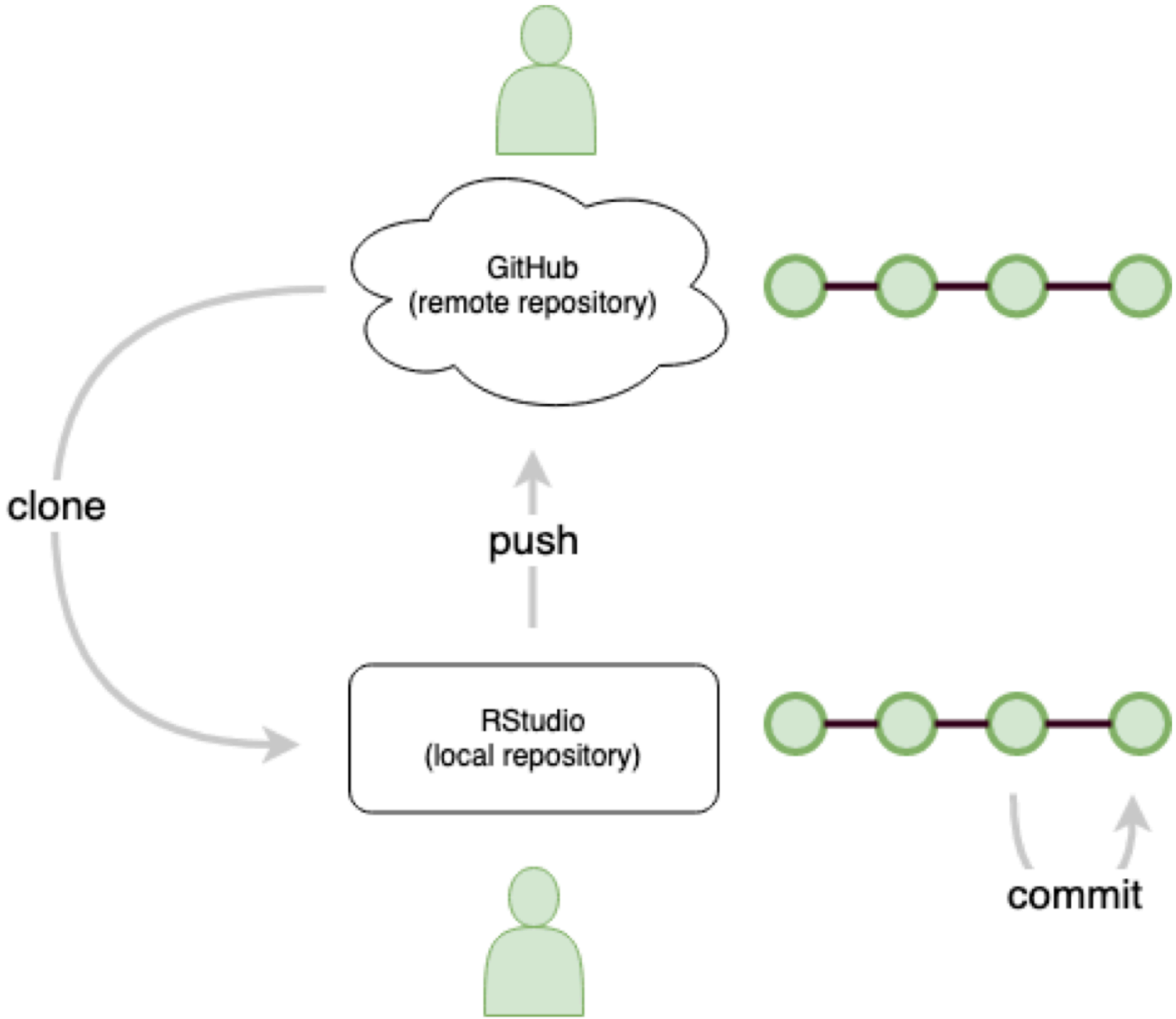
rbtl-fs24.github.io/website/


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)

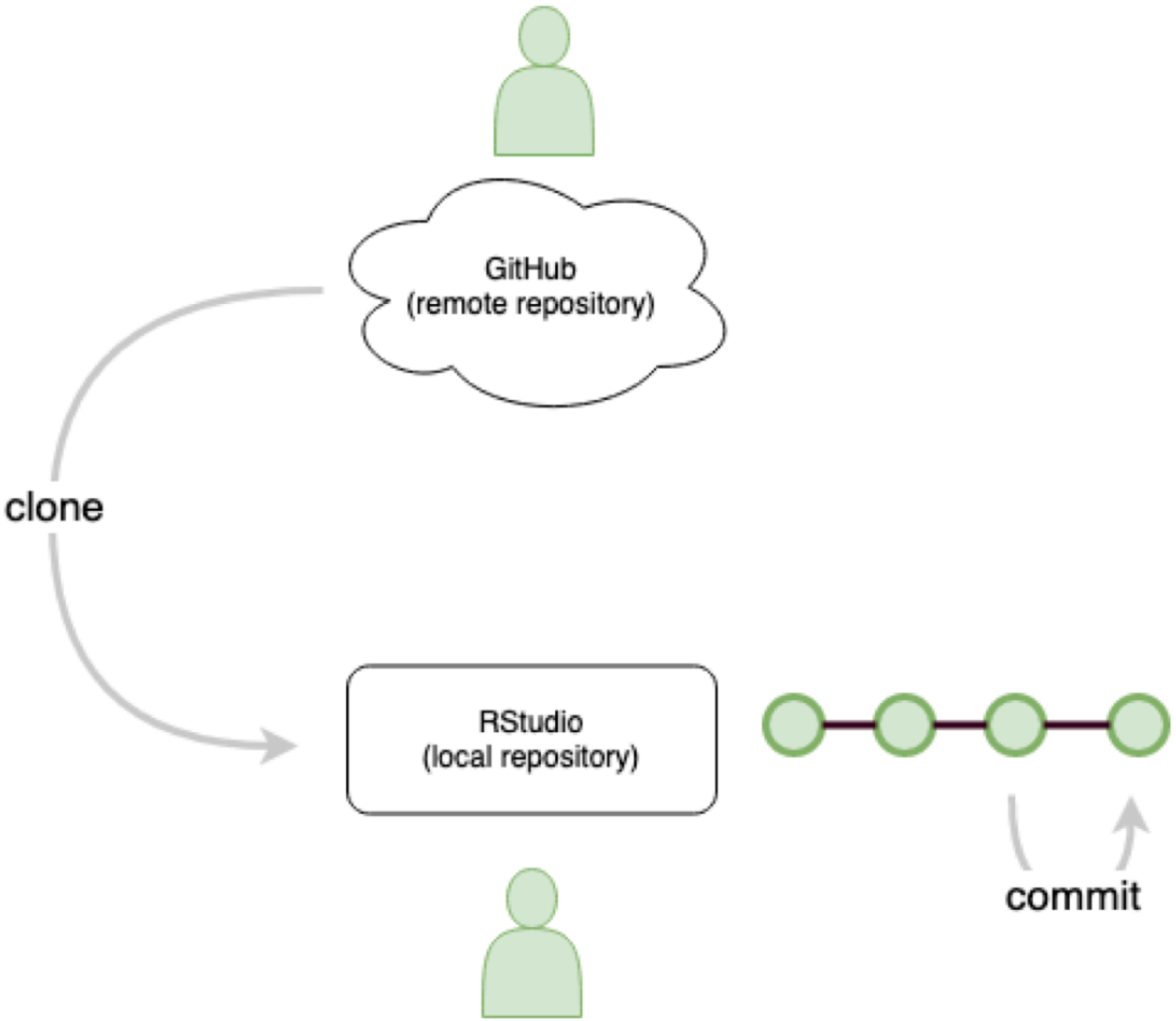


rbtl-fs24.github.io/website/



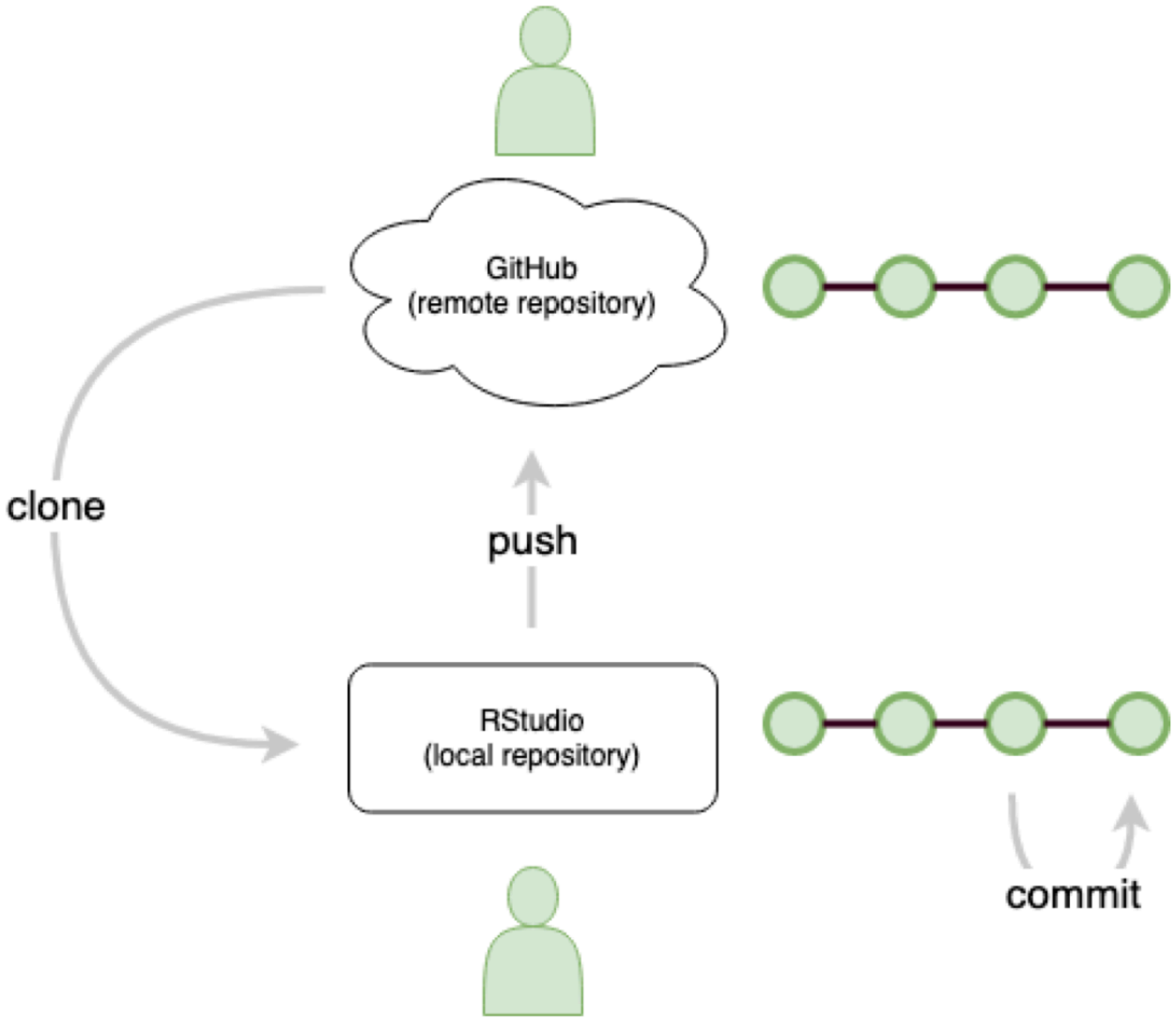
rbtl-fs24.github.io/website/

remember: `git commit`



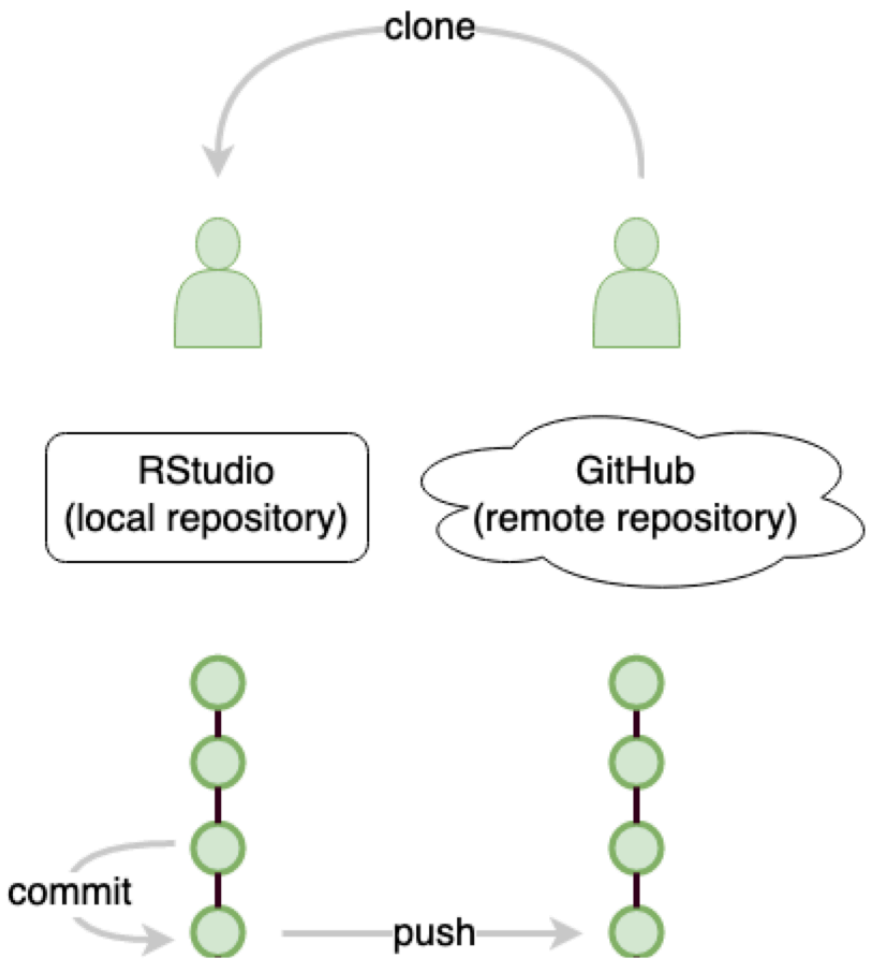
rbtl-fs24.github.io/website/

remember: `git push`

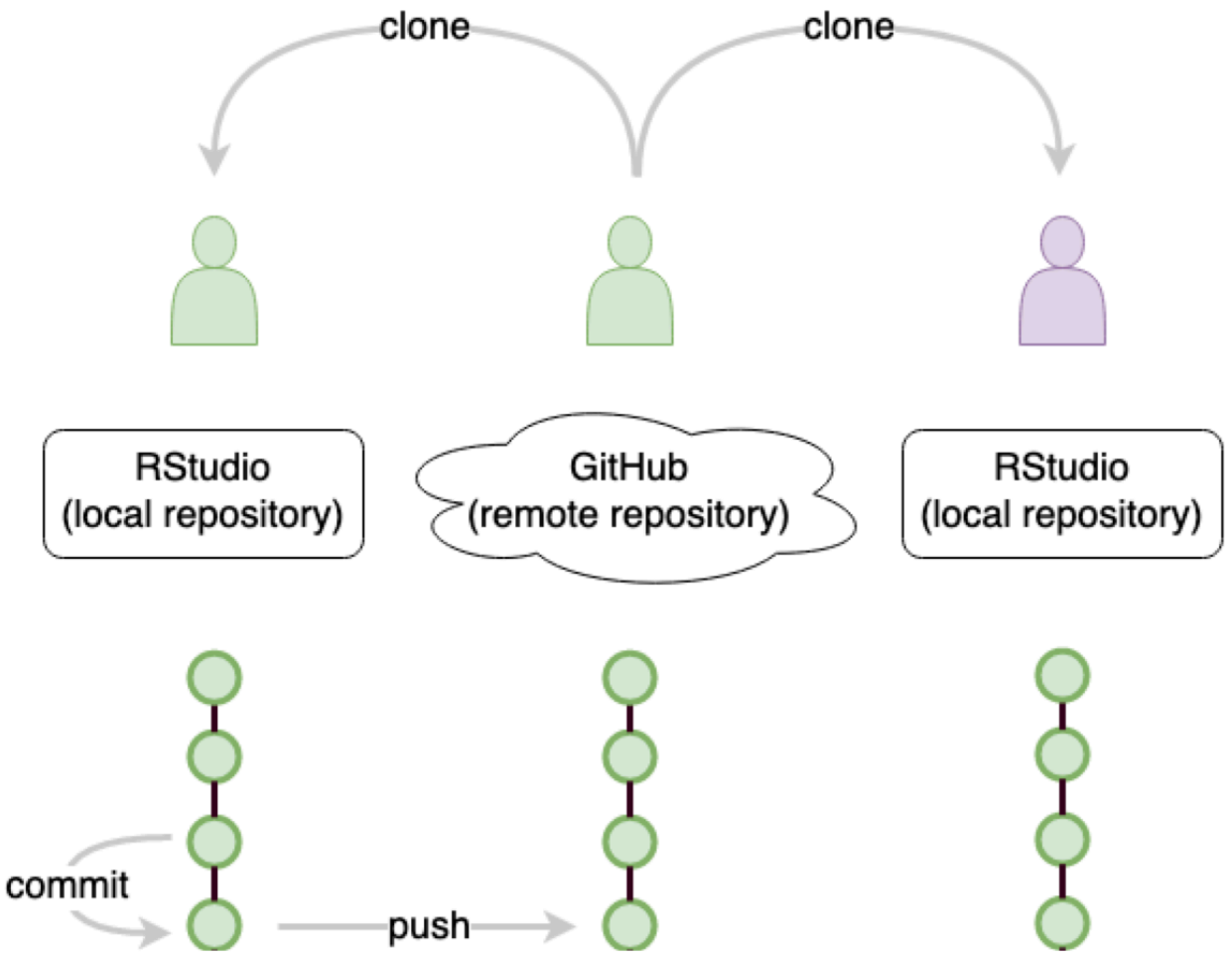


rbtl-fs24.github.io/website/

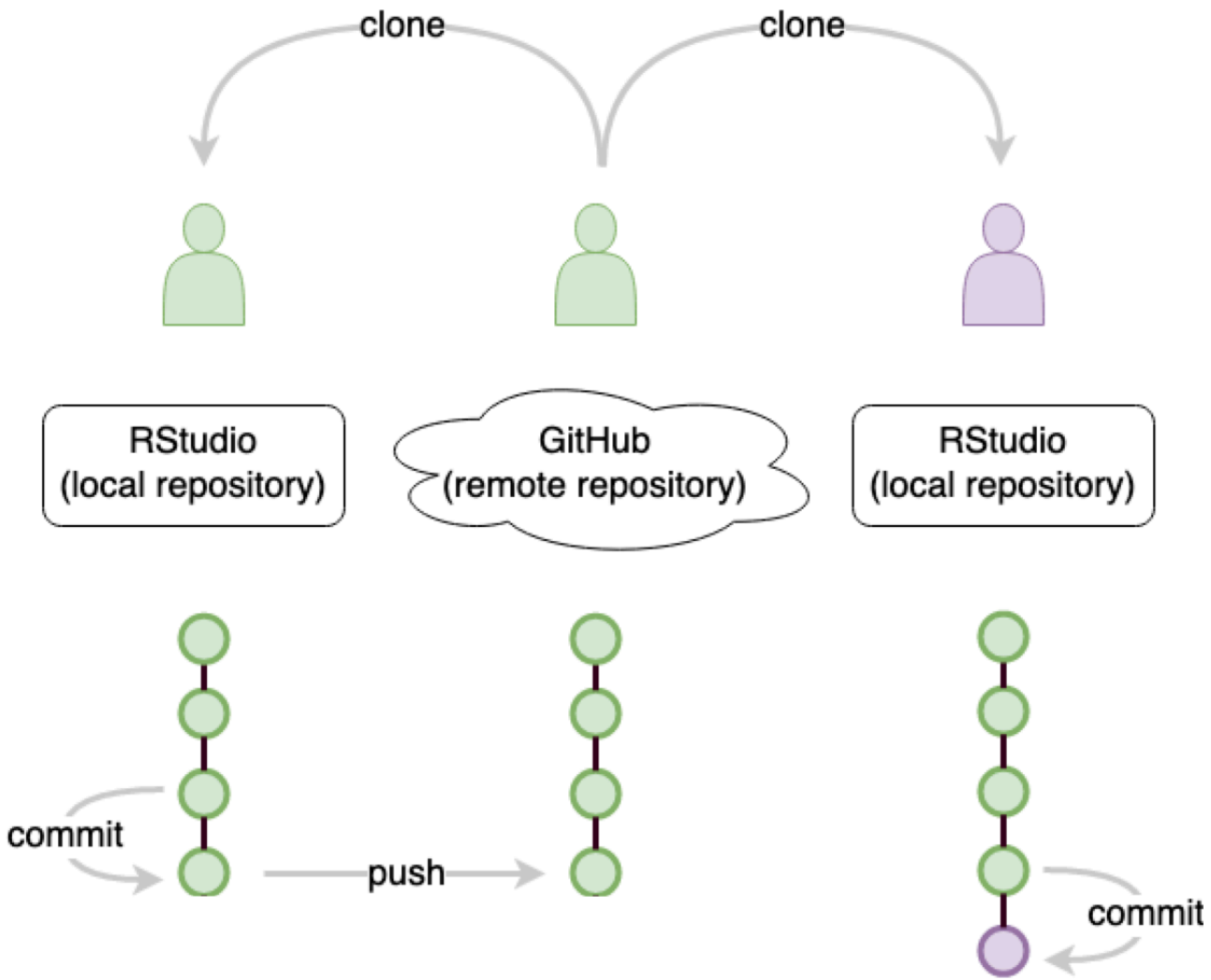
remember: `git push`



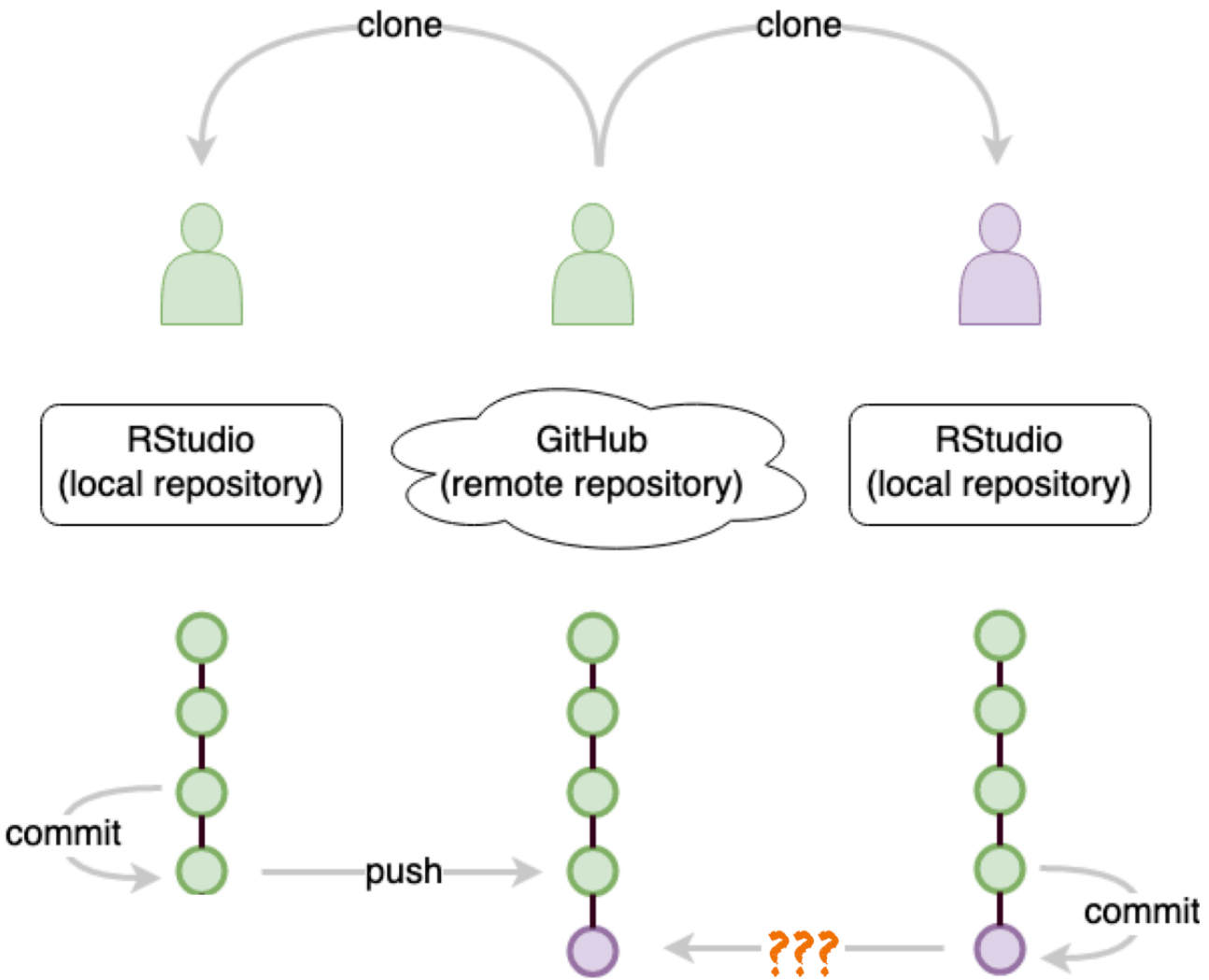
collaborate: `git clone`



track work: `git commit`

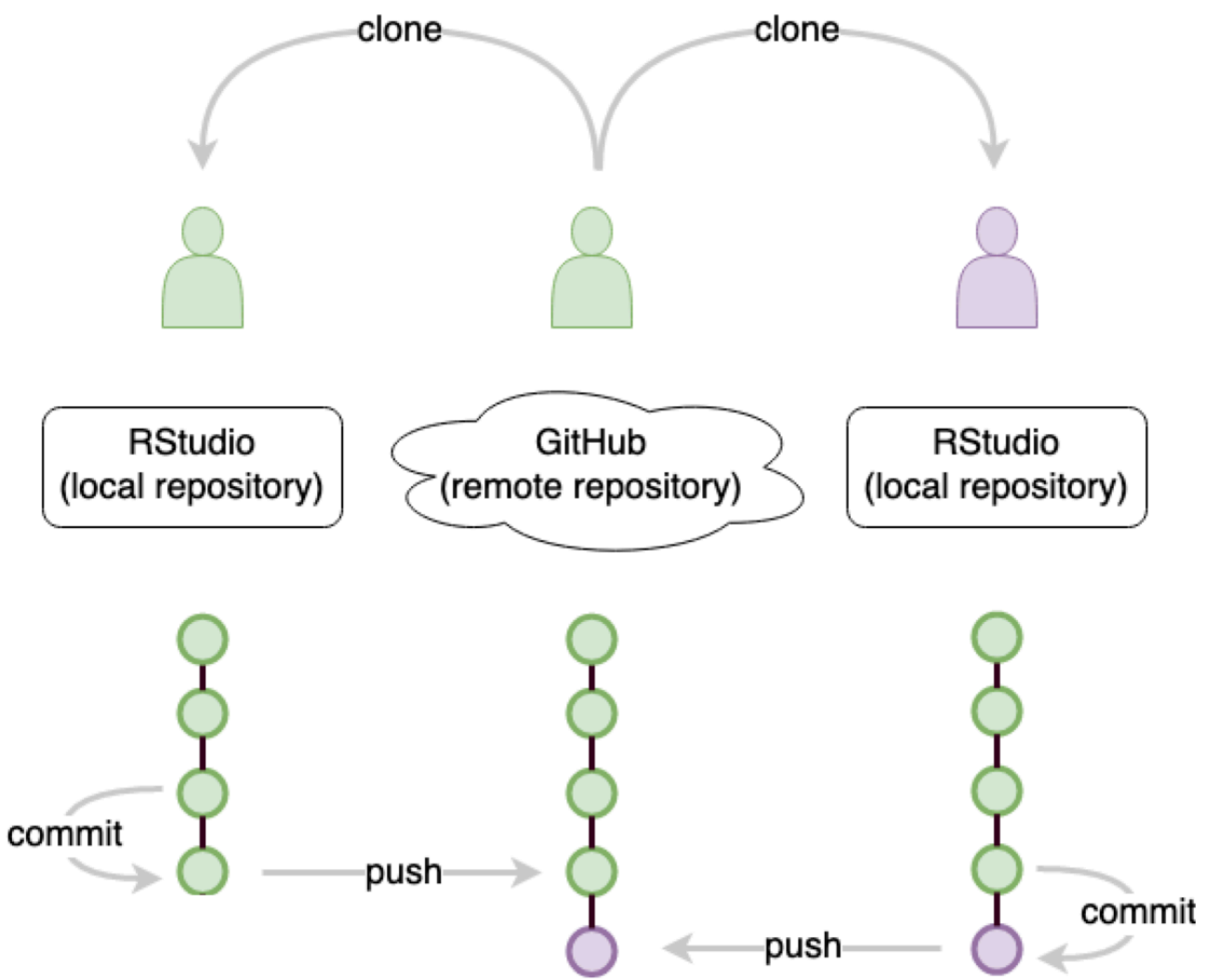


update: `git ???`



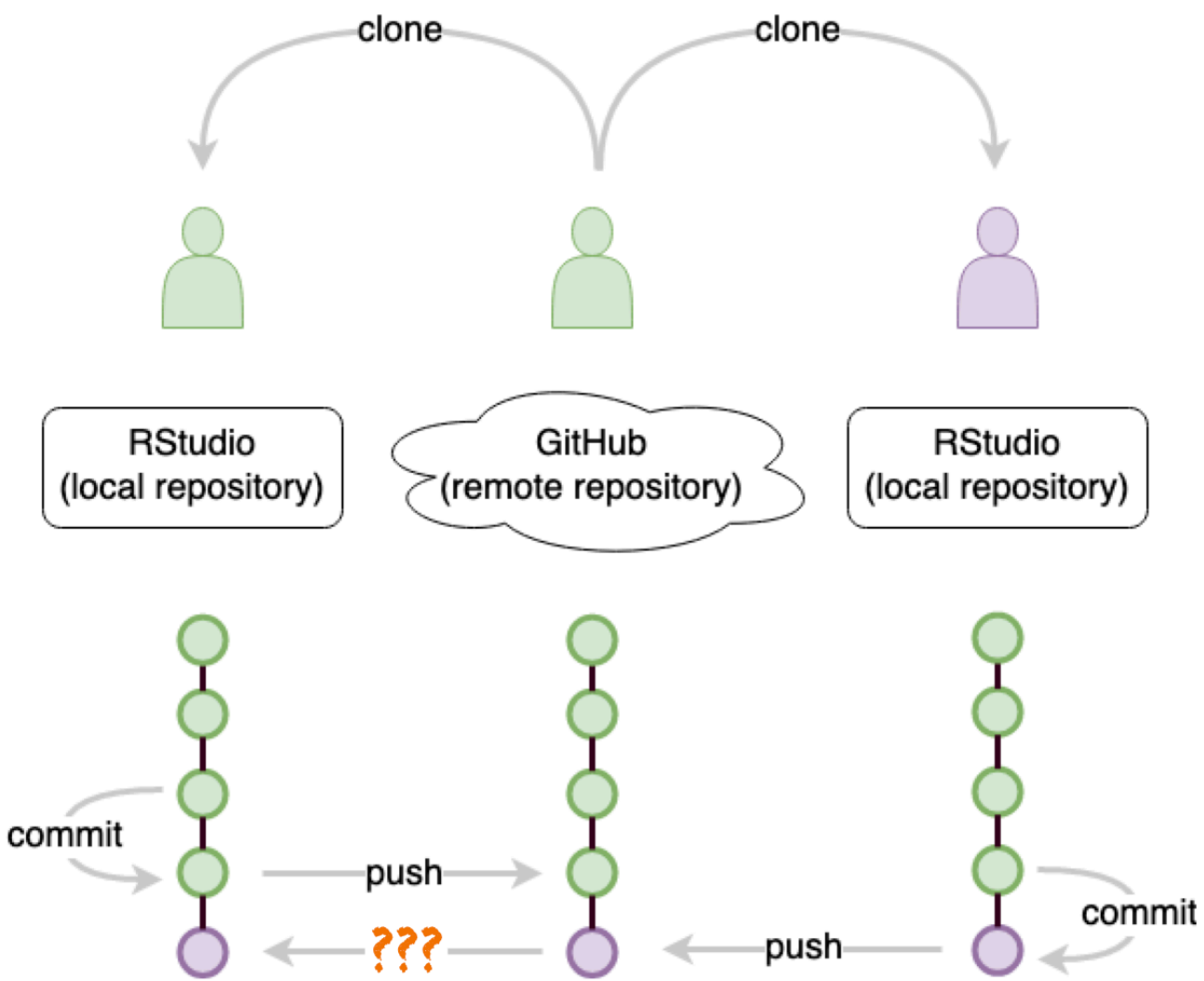
rbtl-fs24.github.io/website/

update: `git push`



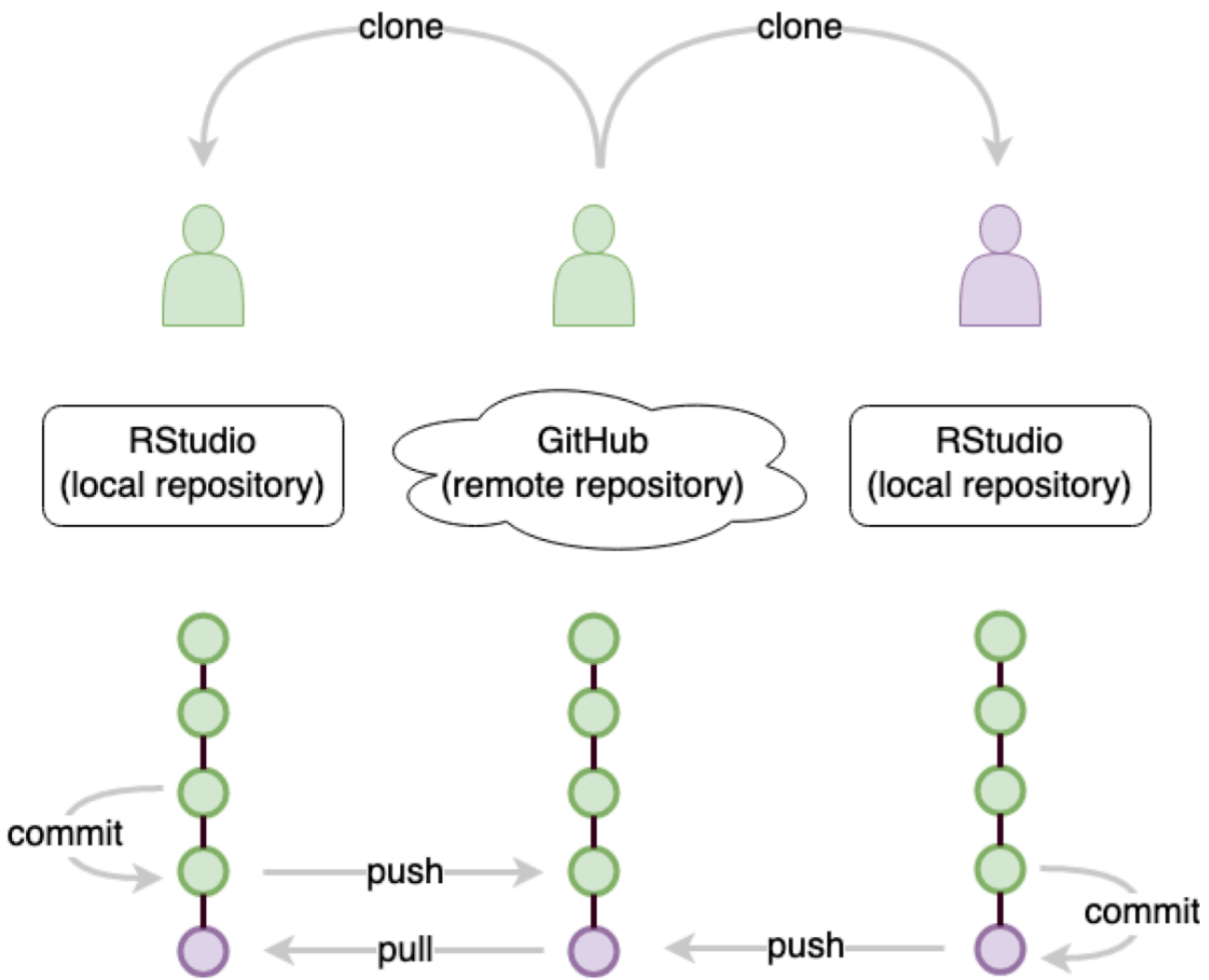
rbtl-fs24.github.io/website/

git ???



rbtl-fs24.github.io/website/

get updates: `git pull`



rbtl-fs24.github.io/website/

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.

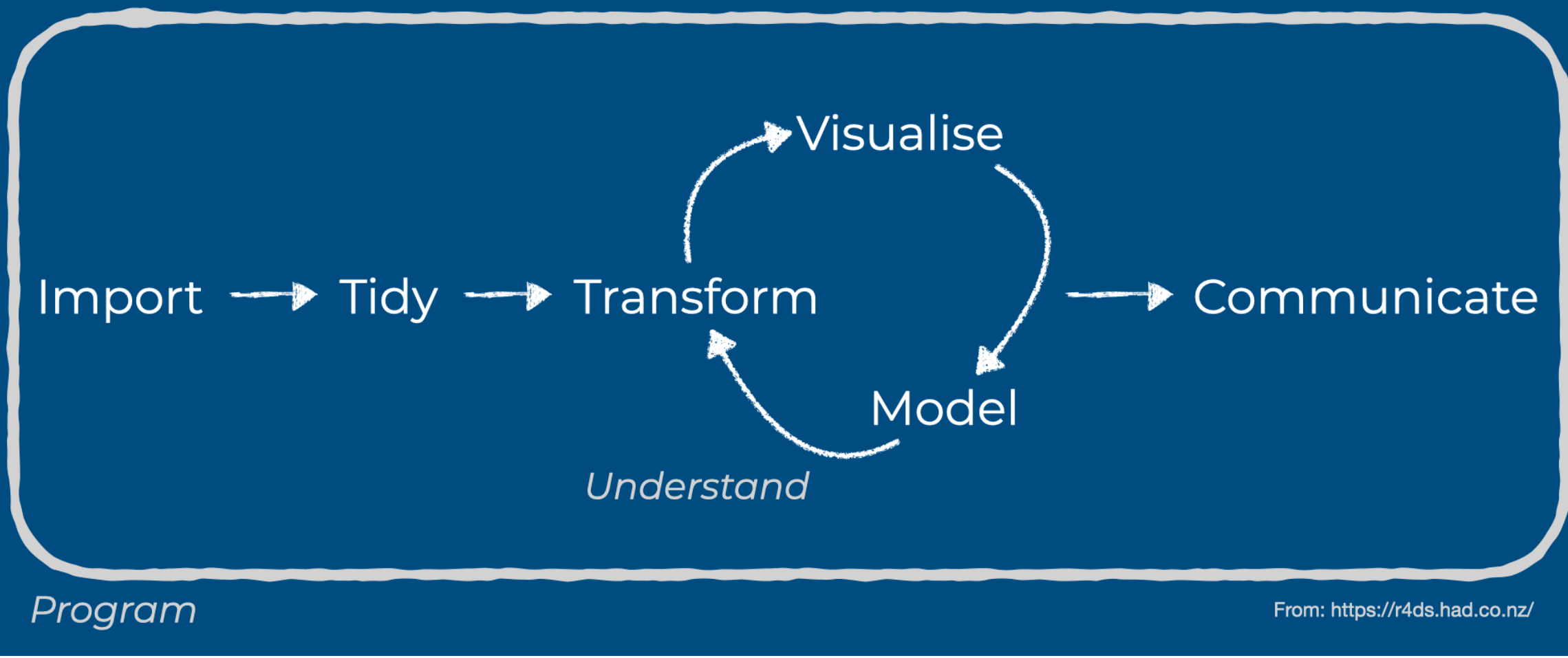
Data Science Lifecycle

Deep End

via GIPHY

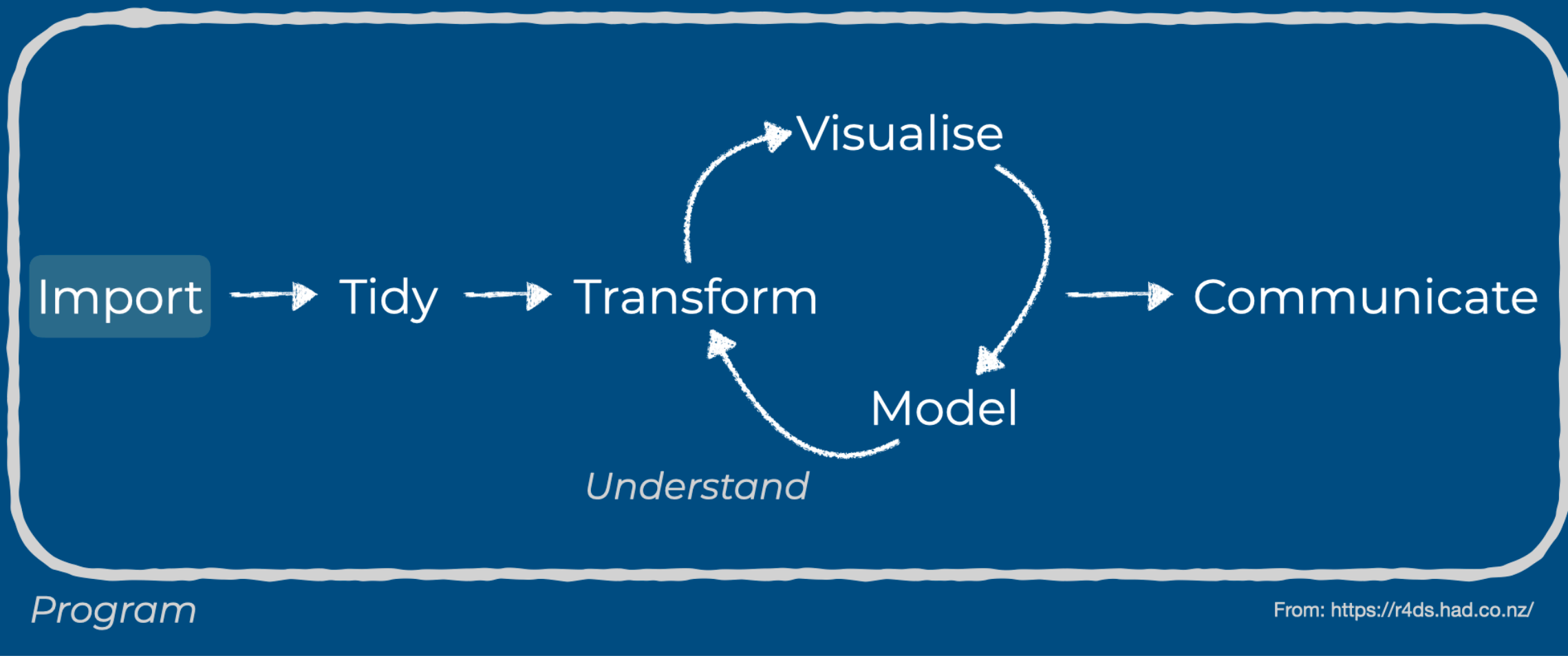
rbtl-fs24.github.io/website/

Data Science Lifecycle



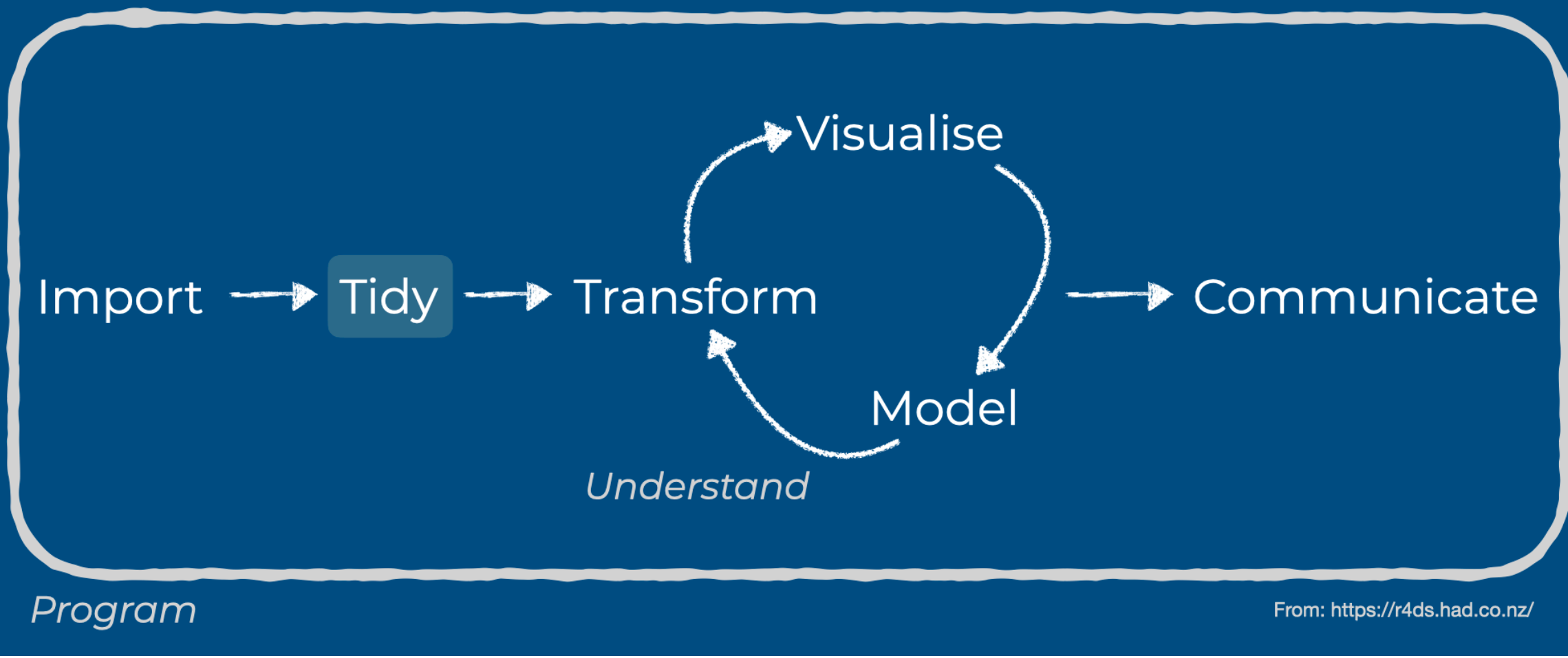
Data Science Lifecycle

Get your data into R



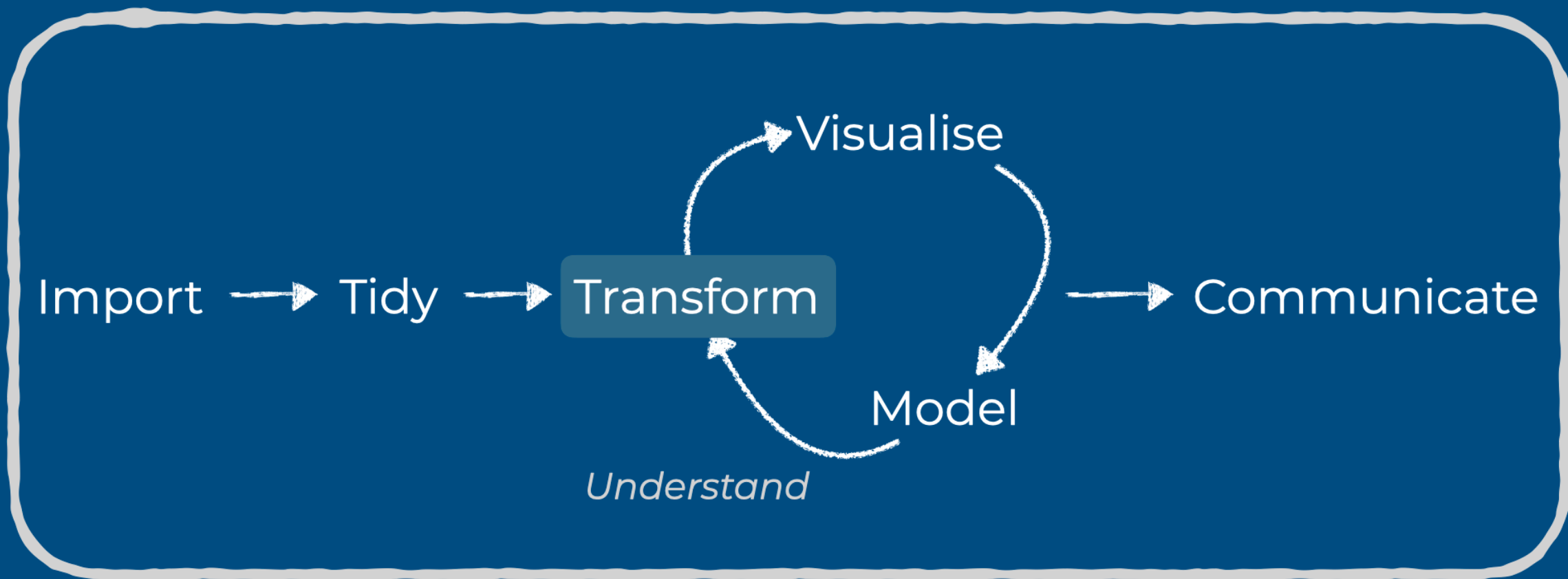
Data Science Lifecycle

Store your data in a consistent form



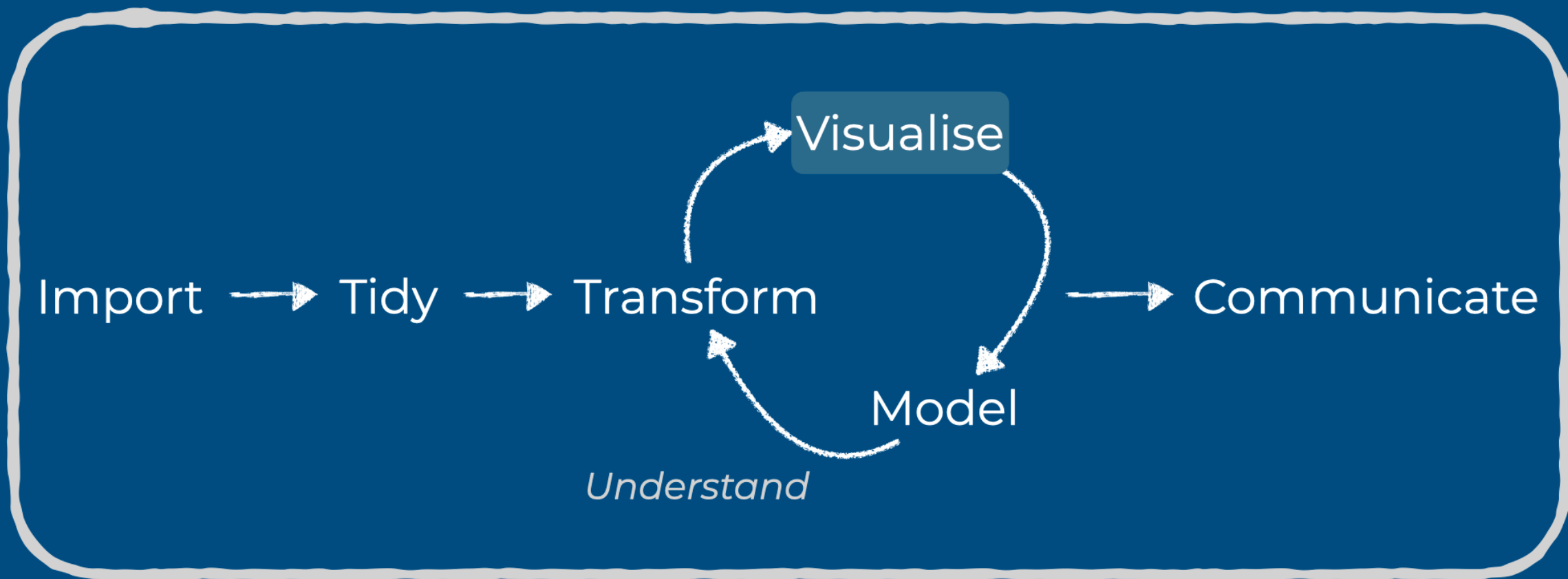
Data Science Lifecycle

Narrow down + Create new variables + Summary stats



Data Science Lifecycle

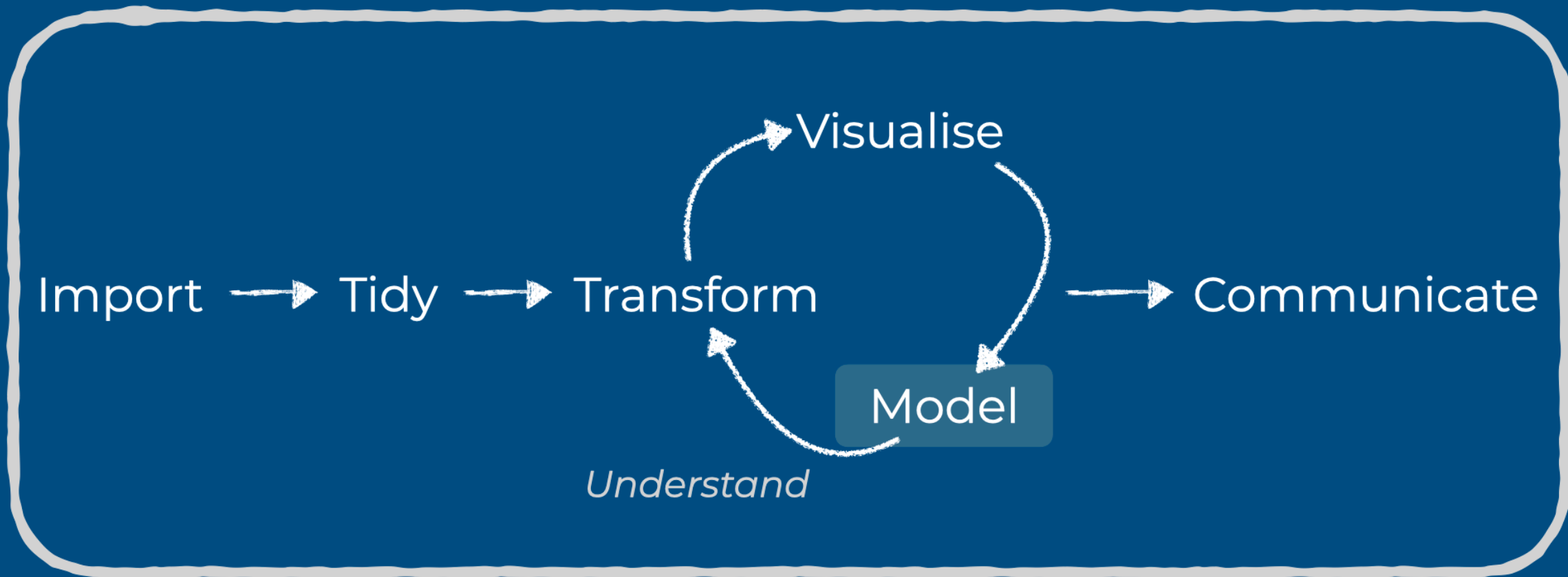
Explore your with visual representations



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

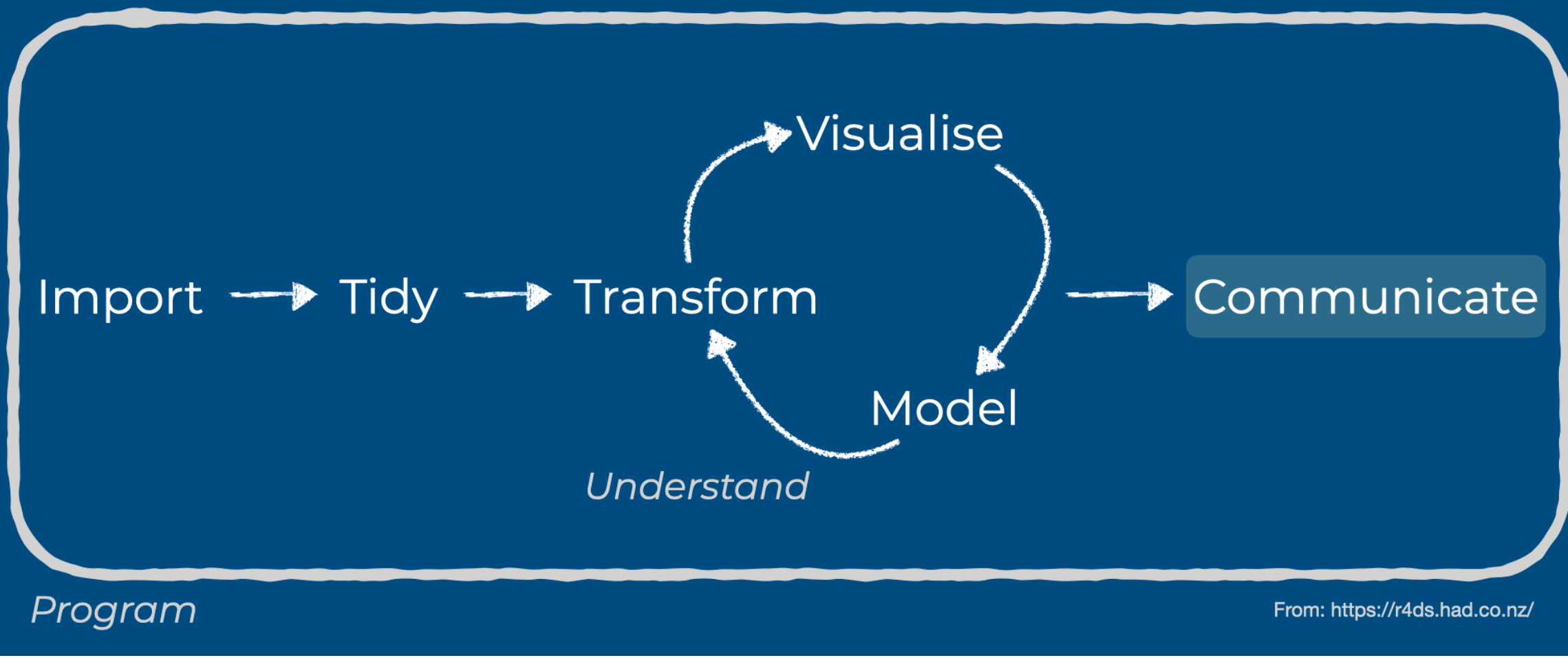
Data Science Lifecycle

Explore your with visual representations



Data Science Lifecycle

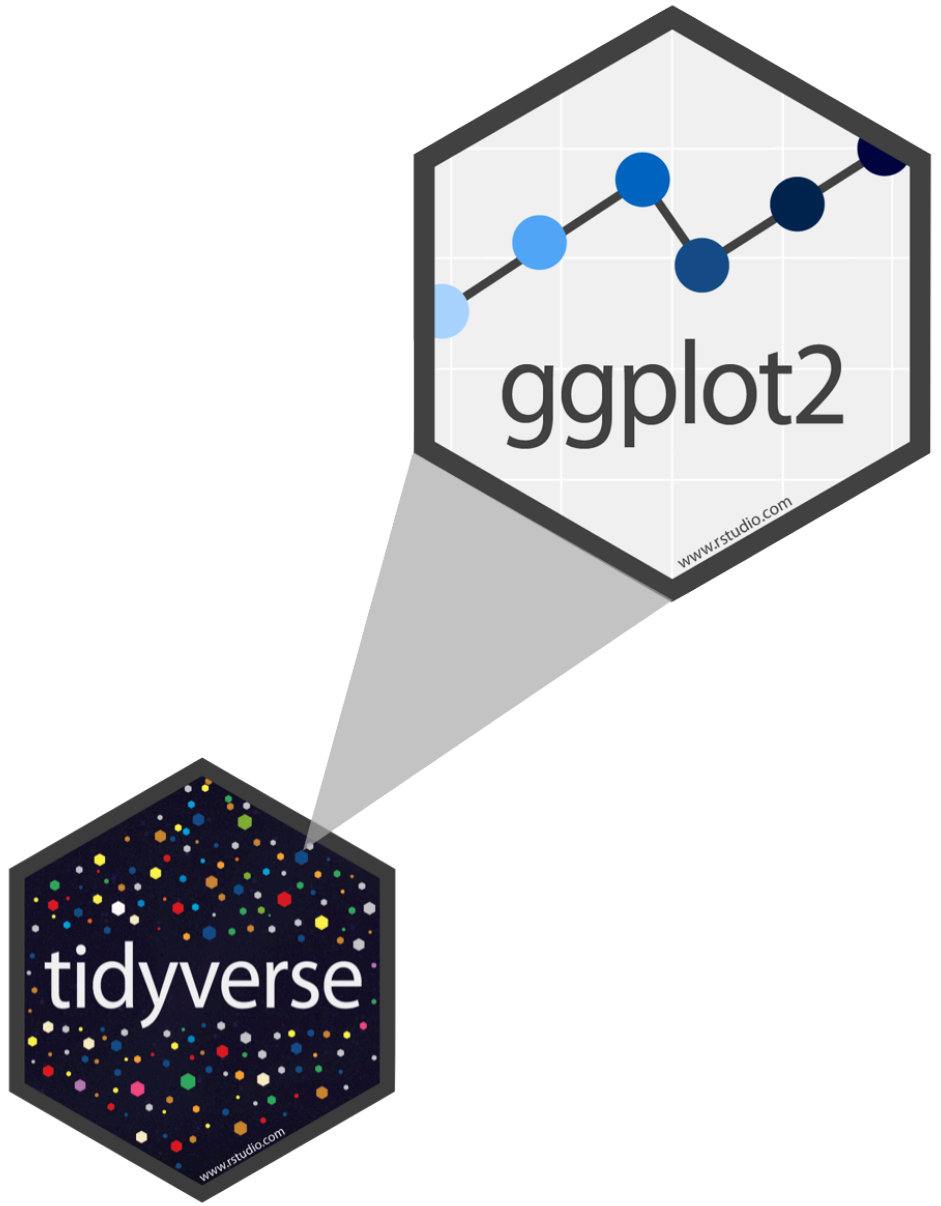
Share your findings with others



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.

rbtl-fs24.github.io/website/

10:00

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

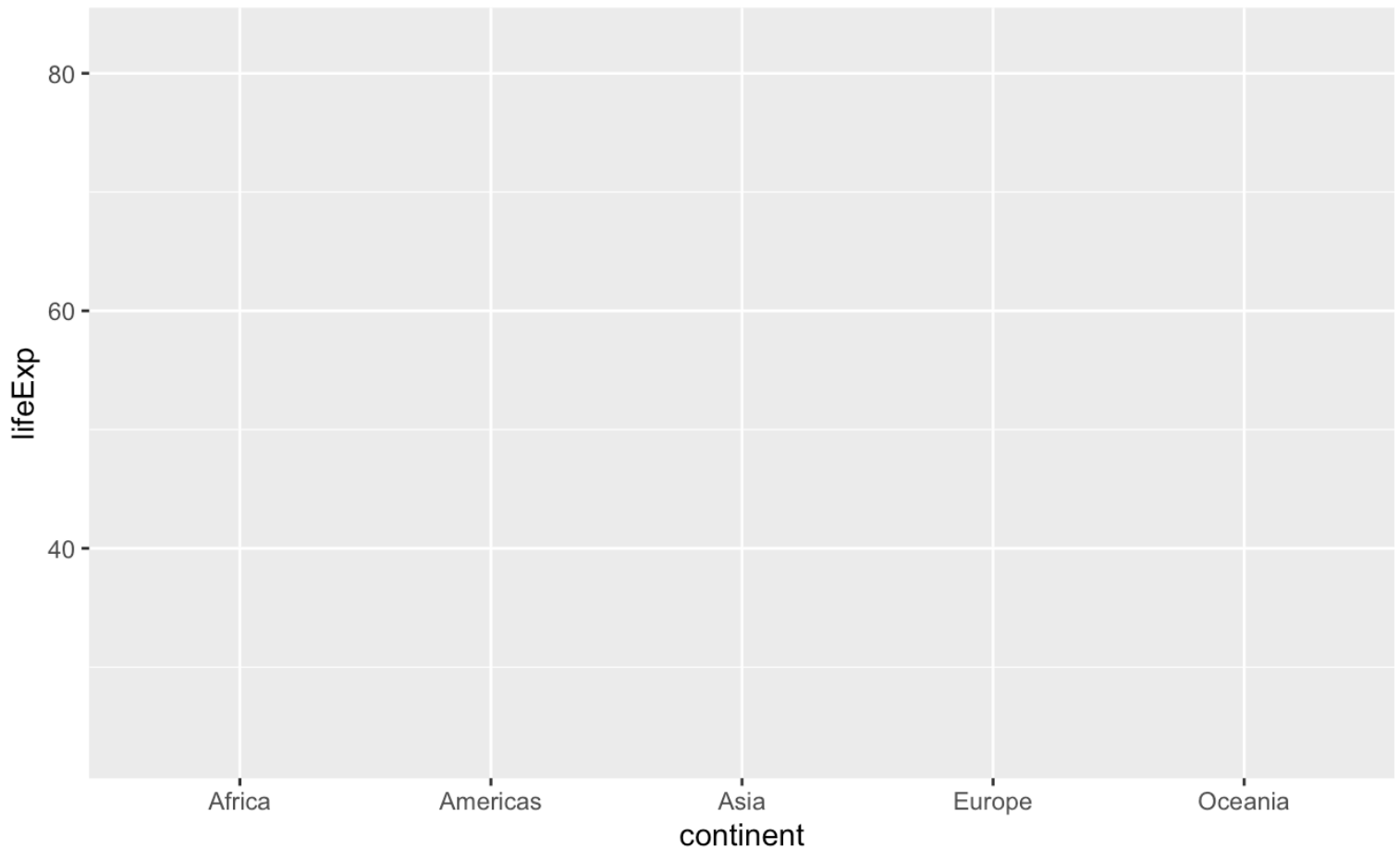
```
ggplot(data = gapminder)
```

Code structure

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

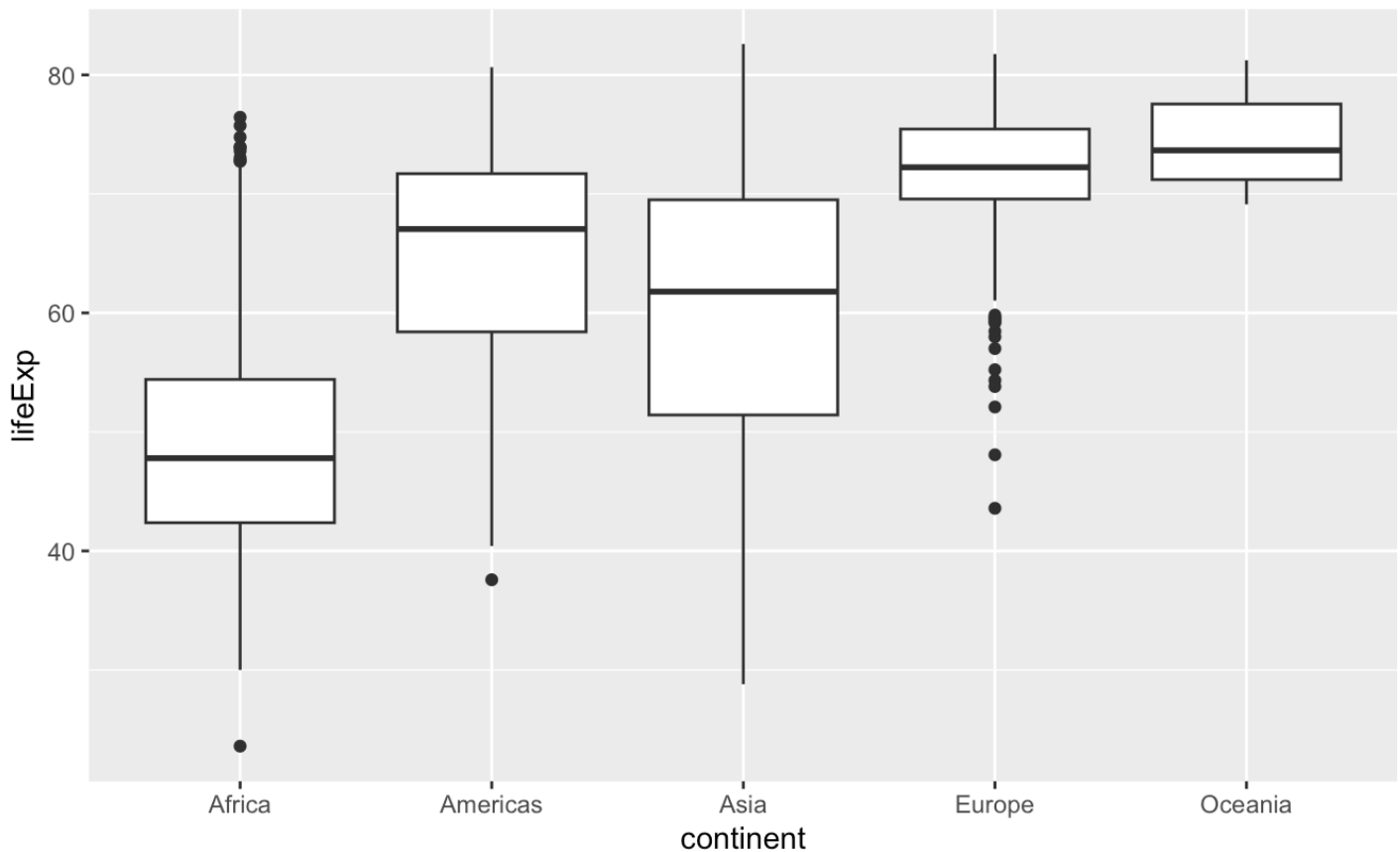
Code structure

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



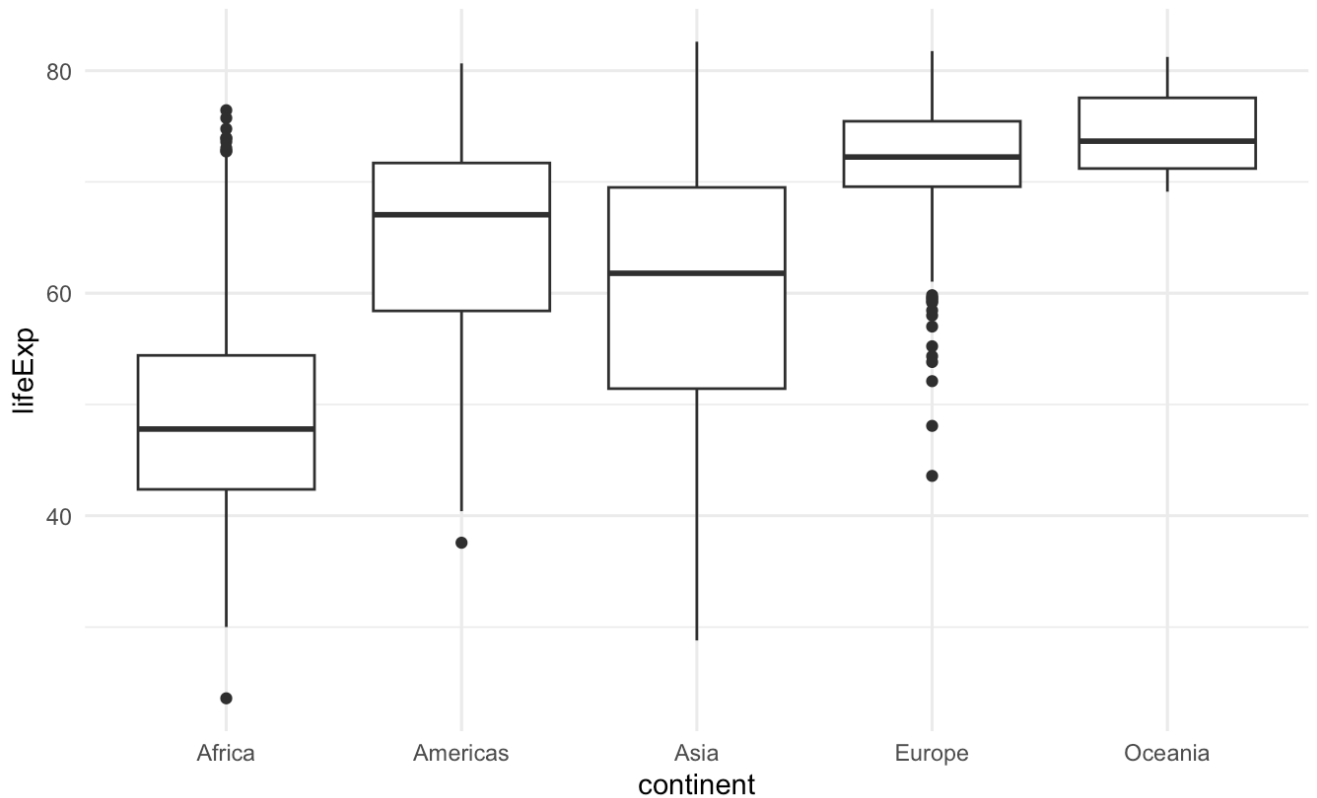
Code structure

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



Code structure

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

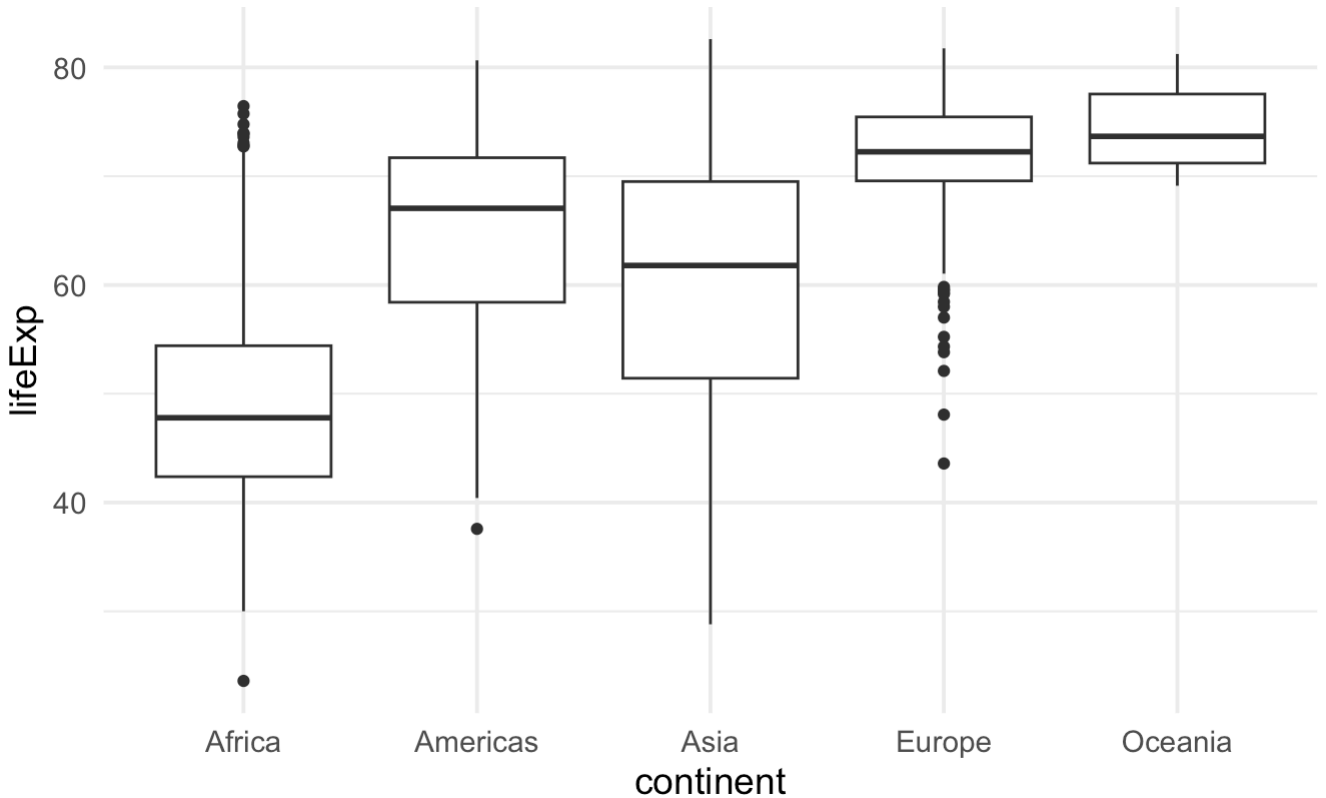


rbtl-fs24.github.io/website/

Polls

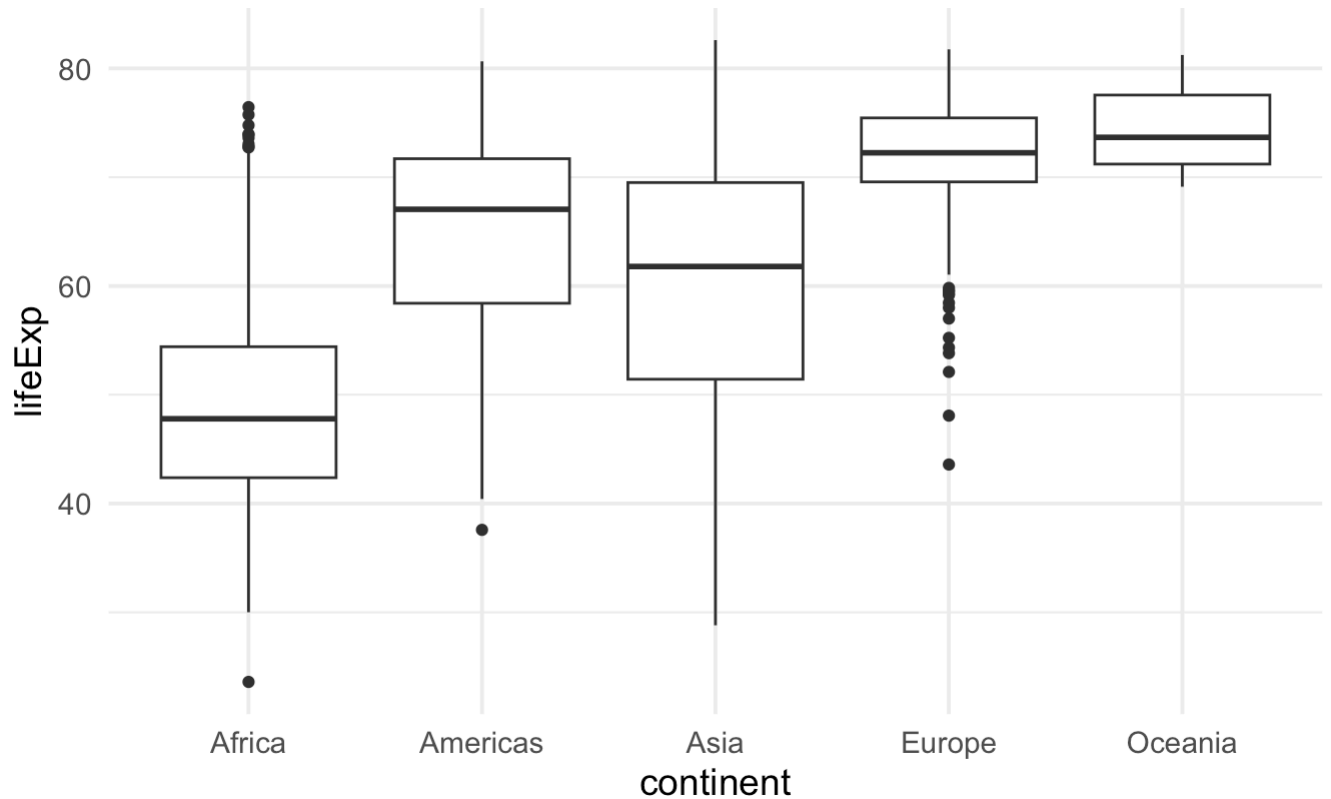
Poll 1: What does the thick line inside the box of a boxplot represent?

1. the mean of the observations
2. the middle of the box
3. the median of the observations
4. none of the above
5. I don't know



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
5. I don't know



rbtl-fs24.github.io/website/

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.
5. I don’t know

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
5. I don’t know

Boxplot, explained

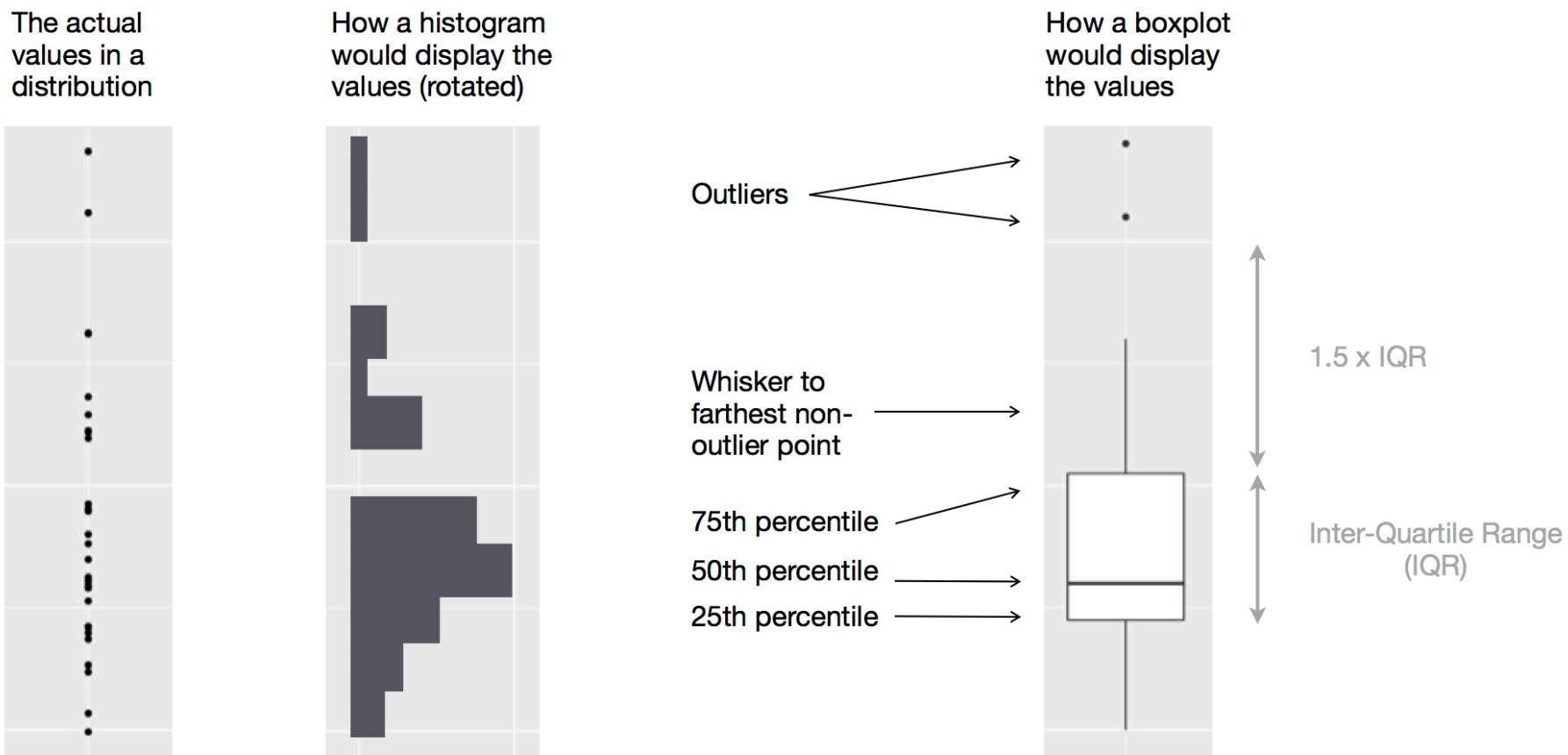


Figure 1: Diagram depicting how a boxplot is created.

rbtl-fs24.github.io/website/

Our turn: md-02-exercises

1. Open posit.cloud in your browser (use your bookmark).
2. Open the `rbtl-fs24 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.

rbtl-fs24.github.io/website/

20:00

Take a break

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

rbtl-fs24.github.io/website/

10:00

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

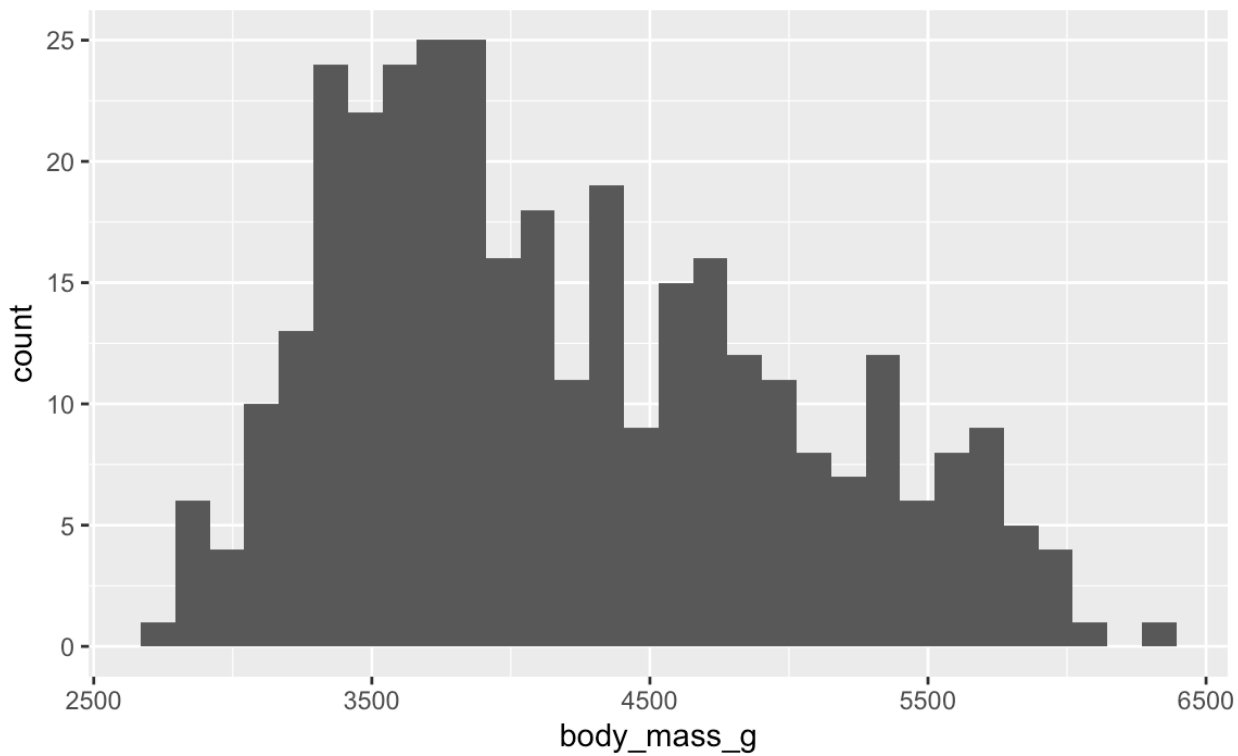
categorical variables

- 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

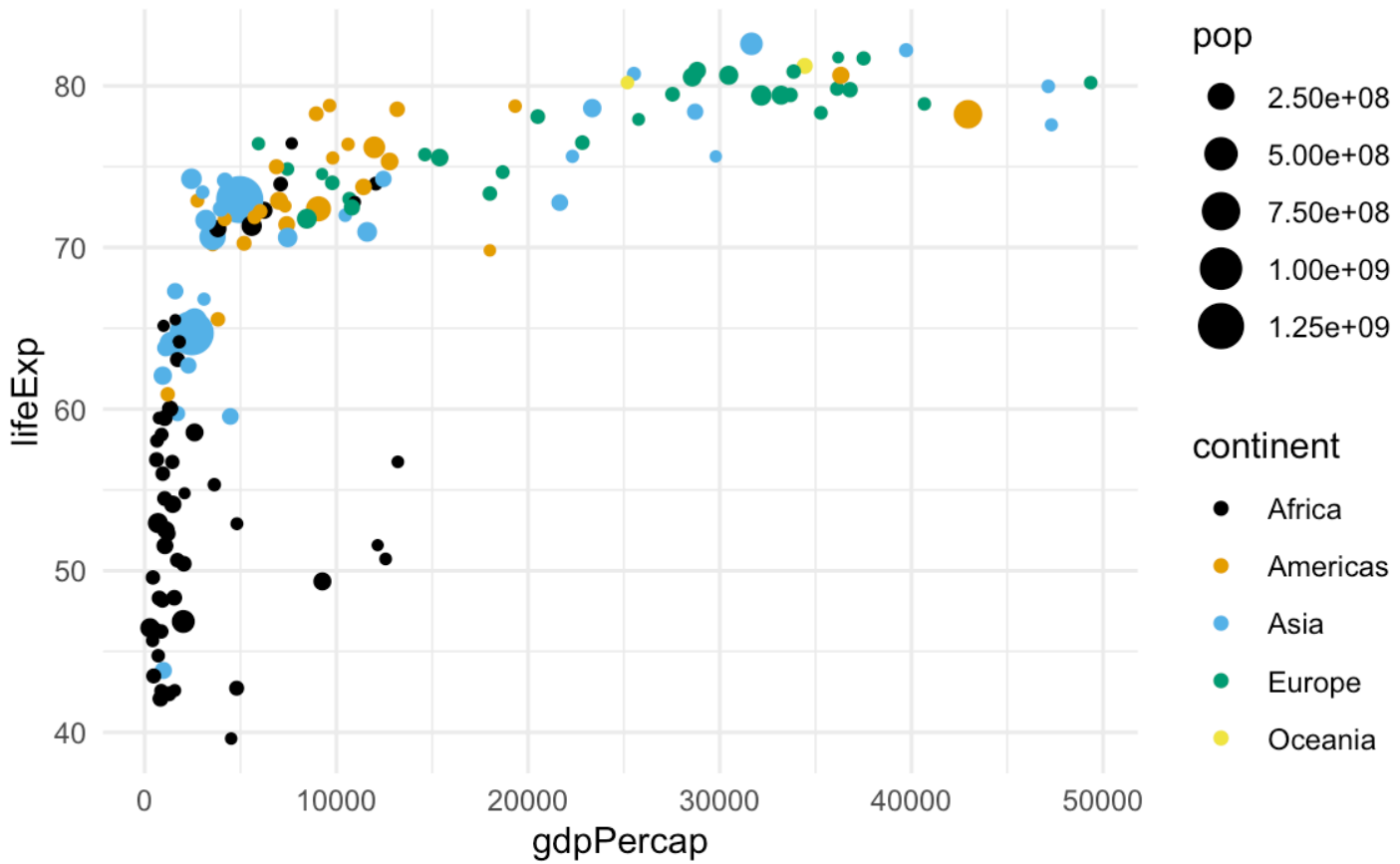
- for visualizing distribution of categorical (non-numerical) variables

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

Scatterplot

- for visualizing 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()
```



rbtl-fs24.github.io/website/

Your turn: md-02-exercises

1. Open posit.cloud in your browser (use your bookmark).
2. Open the rbtl-fs24 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

rbtl-fs24.github.io/website/

Module 2 documentation

rbtl-fs24.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 6th
- Correction & feedback phase up to: Tuesday, March 12th

Wrap-up

rbtl-fs24.github.io/website/

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.