

# 01 - Boite à outils

PRO1036 - Analyse de données scientifiques en R

Tim Bollé

September 9, 2024

# Boîte à outils

# Les outils

## Développement:

- R
- RStudio
- tidyverse
- R Markdown

## Gestion et collaboration:

- Git
- GitHub

# Objectif

# Objectif du cours

À la fin de ce cours, vous pourrez:

- Analyser des données
- Analyser des données de manière **répérable**
- Analyser des données de manière répérable, avec des **outils de programmation modernes**
- Analyser des données de manière répérable et **collaborative**, avec des outils de programmation modernes

# Répétabilité

Que signifie conduire une analyse de donnée de manière répétable ?

À court-terme:

- Pouvons nous reproduire les tableaux et les figures à partir des données
- Est-ce que le code fait ce que nous voulons ?
- Pouvons-nous reconstruire pourquoi et comment nous avons obtenus les résultats

À long-terme:

- Peut-on réutiliser le code pour d'autres données ?
- Peut-on réutiliser le code pour faire autre chose ?

# Les outils de la répérabilité

*Scriptability* → R

Documentation et communication → R Markdown

Gestion et collaboration → Git/GitHub

# R et RStudio



- R est un langage de programmation **open-source**
- R est un environnement pour faire des **calculs** et de la **visualisation statistiques**
- De nombreuses autres applications sont disponibles grâce à des **\*packages\***



- RStudio est un **IDE** (Environnement de Développement Intégré)
- C'est une interface pour R
- Pas nécessaire pour coder en R mais tellement pratique !



# R packages

Les packages sont les *building blocks* de la reproductibilité. Ils contiennent de nombreuses fonctions réutilisables, de la documentation et données de test ([Wickham and Bryan, 2023](#))

Nous allons en utiliser quelques une mais vous verrez que c'est tout une philosophie !

# RStudio tour

The screenshot displays the RStudio environment with the following components:

- Environment Pane:** Shows the 'Global Environment' with a variable 'x' containing the value 2.
- Console:** Contains the following R code and output:
 

```
R 4.4.1 - ~/Projets/PRO1036/
> 2 + 2
[1] 4
> x <- 2
> x * 3
[1] 6
> library(palmerpenguins)
> view(penguins)
> mean(penguins$flipper_length_mm)
[1] NA
> ?mean
> mean(penguins$flipper_length_mm, na.rm = TRUE)
[1] 200.9152
> penguins$flipper_length_mm
[1] 181 186 195 NA 193 190 181 195 193 190 186 180 182 191 198 185 195 197 184 194 174 180
[23] 189 185 180 187 183 187 172 180 178 178 188 184 195 196 190 180 181 184 182 195 186 196
```
- Data Frame (penguins):** A table with 15 rows and 8 columns: species, island, bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm, body\_mass\_g, sex, and year. The data is filtered to show rows 1 to 15 of 344 total entries.
 

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
10	Adelie	Torgersen	42.0	20.2	190	4250	NA	2007
11	Adelie	Torgersen	37.8	17.1	186	3300	NA	2007
12	Adelie	Torgersen	37.8	17.3	180	3700	NA	2007
13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007
14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007
15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007
- Documentation Pane:** Displays the 'Arithmetic Mean' documentation for the 'mean' function, including a description, usage, and arguments.
 

**Arithmetic Mean**

**Description**  
Generic function for the (trimmed) arithmetic mean.

**Usage**

```
mean(x, ...)
```

**Arguments**

  - x** an R object. Currently there are methods for numeric/logical vectors and [date](#), [date-time](#) and [time interval](#) objects. Complex vectors are allowed for `trim = 0`, only.
  - trim** the fraction (0 to 0.5) of observations to be trimmed from each end of `x` before the mean is computed.

# R 101

Les **fonction** sont souvent des verbes, suivi de parenthèses, contenant des arguments:

```
1 fait_ca(avec_ca)
2 fait_ca(avec_ca, et_ca, et_encore_ca)
```

Les packages peuvent être installés avec **install.package** et chargés avec **library**:

```
1 install.packages("package_name")
2 library(package_name)
```

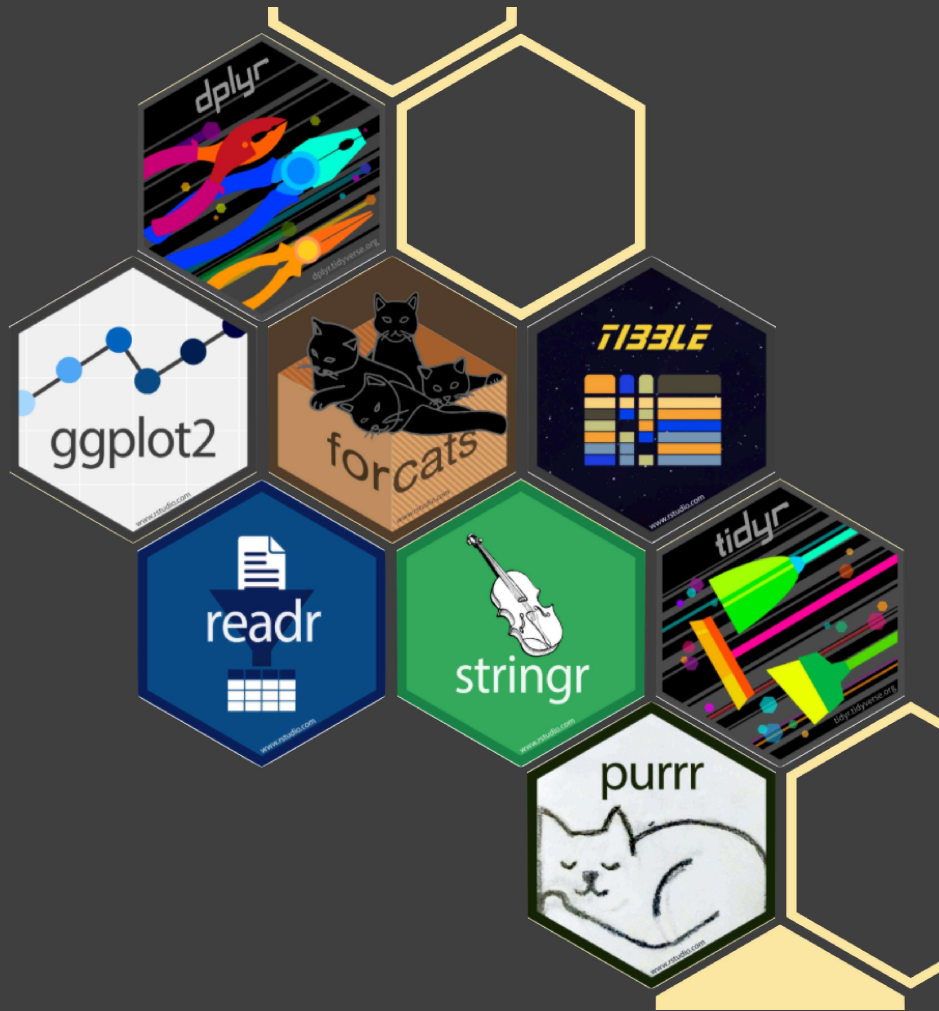
**\$** permet d'accéder aux colonnes des tableaux

```
1 dataframe$var_name
```

**?** permet d'accéder à l'aide sur les fonctions

```
1 ?mean
```

# Tidyverse



[tidyverse.org](https://tidyverse.org)

Le **Tidyverse** est une collection de packages développés pour faire de la data science

Il y a une philosophie et une grammaire commune à tous ces packages, que nous allons apprendre.

# R Markdown

[rmarkdown.rstudio.com](https://rmarkdown.rstudio.com)

**R Markdown** permet d'écrire des documents avec du code intégré (extension en **.Rmd**).

Va permettre de documenter et de communiquer directement nos analyses de données !

- Reproductible: À chaque fois qu'on génère le document, tout est exécuté depuis le début
- Syntaxe simple pour avoir des documents de qualité
- Le document se découpe en zones de texte et blocks de code



# R Markdown

The image shows a side-by-side comparison of an R Markdown document in its source and rendered states within the RStudio interface.

**Left Panel (Source):** Displays the R Markdown code for 'Bechdel.Rmd'. The code includes a YAML header, a title, author information, and a series of R code chunks for data loading and analysis. The console at the bottom is empty.

```

1 ---
2 title: "Bechdel"
3 author: "Mine Çetinkaya-Rundel"
4 format: html
5 editor: visual
6 ---
7
8 In this mini analysis we work with the data used in the FiveThirtyEight story
9 titled ["The Dollar-And-Cents Case Against Hollywood's Exclusion of
10 Women"](https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-
11 t-hollywoods-exclusion-of-women/). Your task is to fill in the blanks denoted
12 by _____.
13
14 ## Data and packages
15
16 We start with loading the packages we'll use.
17
18 ```{r}
19 #/ label: load-packages
20 #/ warning: false
21 #/ message: false
22
23 library(fivethirtyeight)
24 library(tidyverse)
25 ```
26
27 The dataset contains information on `r nrow(bechdel)` movies released between
28 `r min(bechdel$year)` and `r max(bechdel$year)`. However we'll focus our
29 analysis on movies released between 1990 and 2013.
30
31 ```{r}
32 bechdel190_13 <- bechdel %>%
33   filter(between(year, 1990, 2013))
34 ```
35
36 There are `r nrow(bechdel190_13)` such movies.
37
38 The financial variables we'll focus on are the following:

```

**Right Panel (Visual):** Shows the rendered HTML output of the document. It features a title, author, and paragraphs of text with embedded R code chunks. The output is styled with a light blue header and a search bar.

**Environment Panel (Top Right):** Shows the current environment with variables like 'PRO1036' and 'Bechdel'.

**Files Panel (Bottom Right):** Shows the file structure of the project, including 'Bechdel.Rmd' and 'Bechdel'.

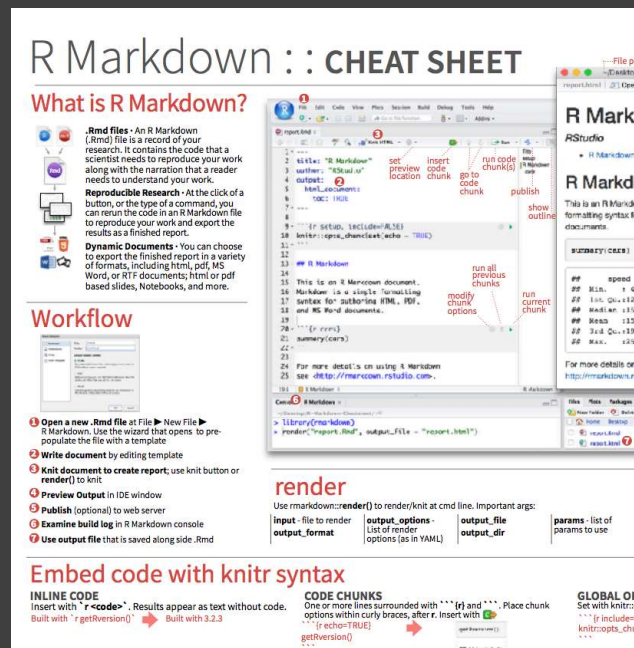
# R Markdown - Aide

Cheatsheet

Help > Cheatsheet

Markdown Quick Reference

Help > Markdown Quick Reference



**R Markdown :: CHEAT SHEET**

**What is R Markdown?**

**Rmd files** - An R Markdown (.Rmd) file is a record of your research. It contains the code that a scientist needs to reproduce your work along with the narration that a reader needs to understand your work.

**Reproducible Research** - At the click of a button, or the type of a command, you can rerun the code in an R Markdown file to reproduce your work and export the results as a finished report.

**Dynamic Documents** - You can choose to export the finished report in a variety of formats, including html, pdf, MS Word, or RTF documents; html or pdf based slides, Notebooks, and more.

**Workflow**

1. Open a new .Rmd file at File > New File > New R Markdown. Use the wizard that opens to pre-populate the file with a template.
2. Write document by editing template.
3. Knit document to create report; use knit button or `render()` to knit.
4. Preview Output in IDE window.
5. Publish (optional) to web server.
6. Examine build log in R Markdown console.
7. Use output file that is saved along side .Rmd.

**render**

Use `markdown::render()` to render/knit at cmd line. Important args:

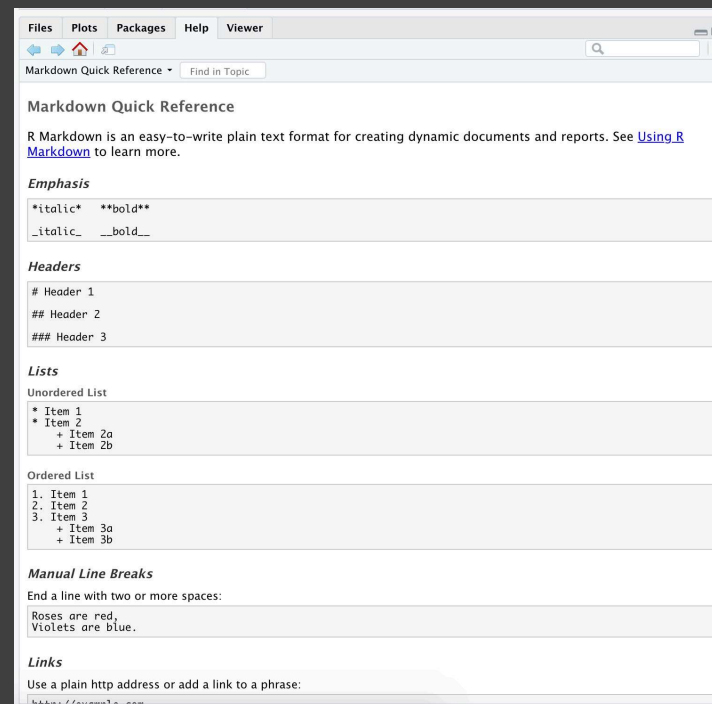
input	output_options	output_file	params
file to render	List of render options (as in YAML)	output_dir	list of params to use

**Embed code with knitr syntax**

**INLINE CODE**  
Insert with `<code>`. Results appear as text without code.  
Built with `r.getRversion()` Built with 3.2.3

**CODE CHUNKS**  
One or more lines surrounded with ````[r]` and `````. Place chunk options within curly braces, after `r`. Insert with `<code>`  
`{r chunk=TRUE}` `getRversion()`

**GLOBAL OPTIONS**  
Set with knitr options. ````{r} [include=FALSE] knitr.opts_chunk``



**Markdown Quick Reference**

R Markdown is an easy-to-write plain text format for creating dynamic documents and reports. See [Using R Markdown](#) to learn more.

**Emphasis**

`*italic*` `**bold**`  
`_italic_` `__bold__`

**Headers**

`# Header 1`  
`## Header 2`  
`### Header 3`

**Lists**

**Unordered List**

- \* Item 1
- \* Item 2
- \* Item 3
- \* Item 3a
- \* Item 3b

**Ordered List**

1. Item 1
2. Item 2
3. Item 3
- 3a. Item 3a
- 3b. Item 3b

**Manual Line Breaks**

End a line with two or more spaces:  
Roses are red,  
Violets are blue.

**Links**

Use a plain http address or add a link to a phrase:  
`<code>`



# Boîte à outils



# Les outils

## Développement:

- R
- RStudio
- tidyverse
- R Markdown

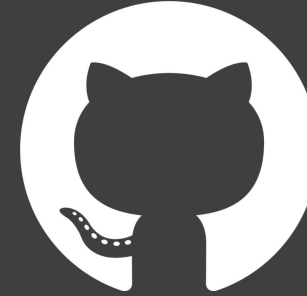
## Gestion et collaboration:

- Git
- GitHub

# Git et GitHub

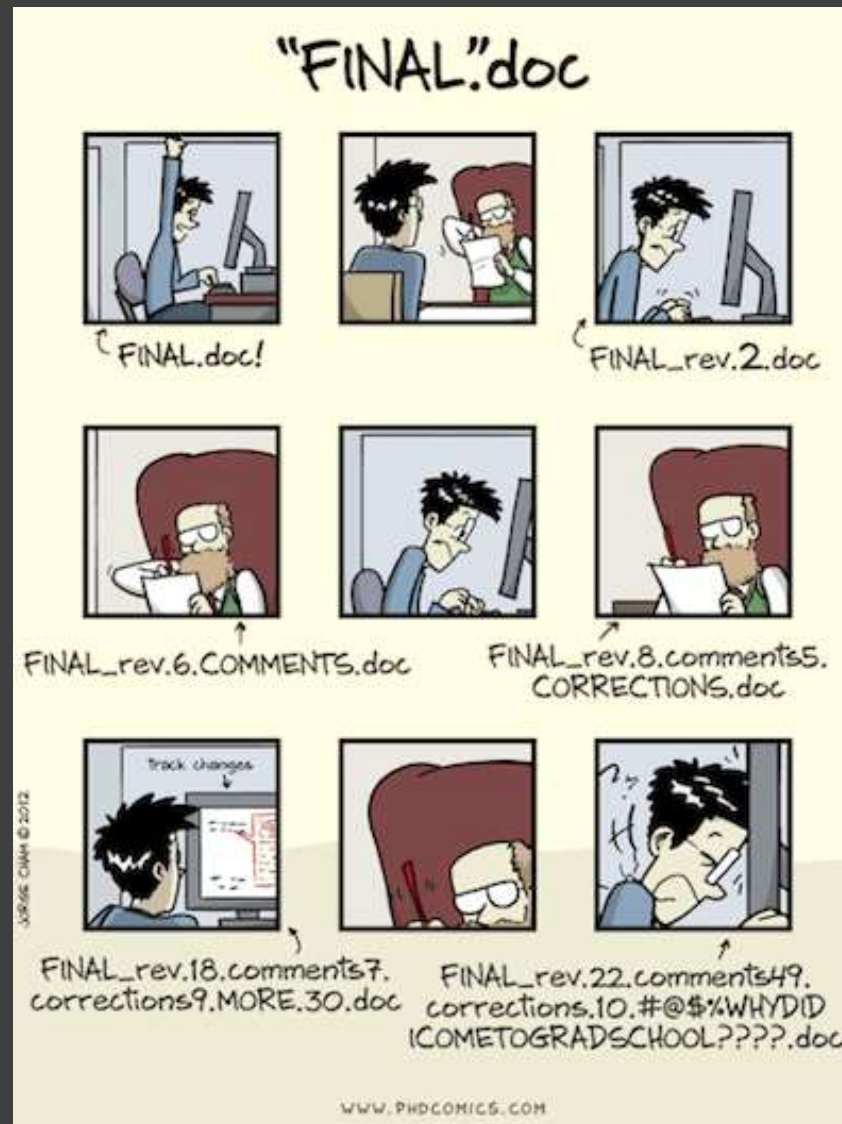


- Git est un outil de **gestion de version**
  - Comme le *track changes* sur Word
- Très populaire dans le monde de la programmation

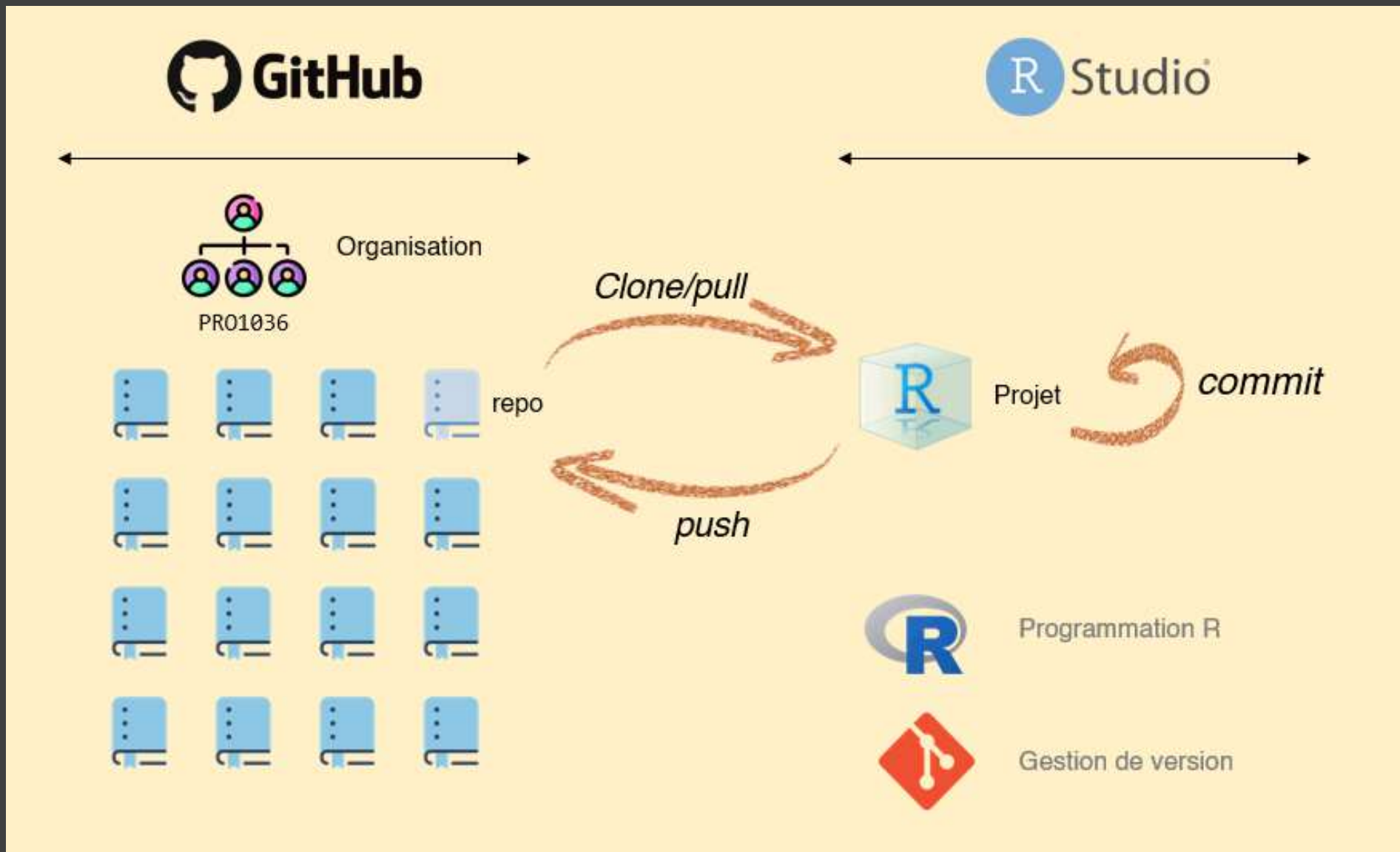


- GitHub est une plateforme de stockage de **repo** Git
  - Comme un Onedrive/Dropbox pour Git
- Nous allons essayer de l'utiliser pour... tout !

# Pourquoi la gestion de version ?



# Fonctionnement



# Mise en place

Git peut être utilisé depuis le terminal de commande

- Utilisation plus avancée
- Nous pouvons normalement tout faire depuis R Studio

Github:

- Créez un compte avec votre adresse UQTR
- Vérifiez votre adresse courriel

# Références

Wickham, H. and Bryan, J. (2023). *R Packages: Organize, Test, Document, and Share Your Code* (2nd edition). O'Reilly Media.