

# rbtl - Data Science Lifecycle

Lars Schöbitz

Global Health Engineering - ETH Zurich

2022-04-28



Welcome back!

# You got your data!



via GIPHY

# What's happening next?

## Today

1. Classroom tools
2. Data Science Lifecycle
3. R basics: Functions, Arguments, Objects, Operators
4. Live Coding Exercise
5. Pair Programming Exercise
6. Homework and Project Report

# Learning Objectives

1. Learners can import their data from a CSV file to a team repository on GitHub
2. Learners can list the six elements of the data science lifecycle
3. Learners know three different ways of getting support in solving coding problems online

# Classroom tools

# Live Coding Exercises

- Instructor writes and narrates code out loud
- Instructor explains elements and principles that are relevant
- Code is displayed on projector screen
- Learners join by writing and executing the same code
- Learners “code-along” with the instructor



# Pair Programming Exercises

- Two learners work together on one computer
- One person (the driver) does the typing
- The other person (the navigator) offers comments and suggestions
- Roles get switched

# Taking Notes Together

- Questions in shared online document:

[https://docs.google.com/document/d/1B\\_fGhU2-p7GdMjDdRq73JXAhAM7VU2JDdu6t1foSQ0og/edit](https://docs.google.com/document/d/1B_fGhU2-p7GdMjDdRq73JXAhAM7VU2JDdu6t1foSQ0og/edit)

# Sticky Notes

- Use as status flags
- **Orange:** Exercise completed
- **Pink:** Problem, need support

# Data Science Lifecycle

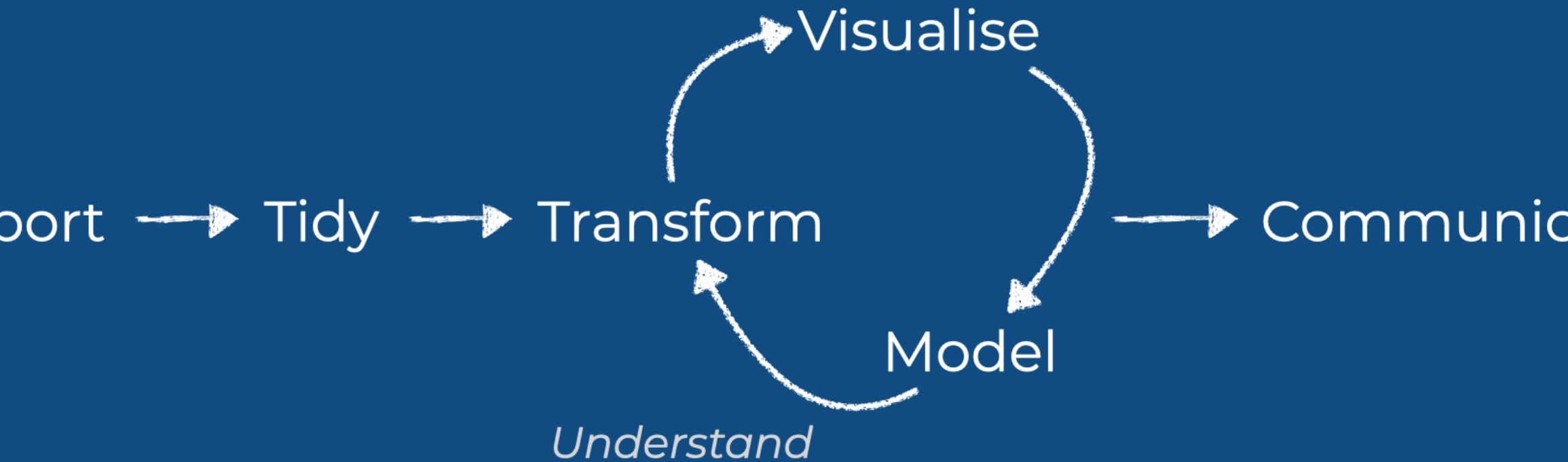
# Deep End



via GIPHY

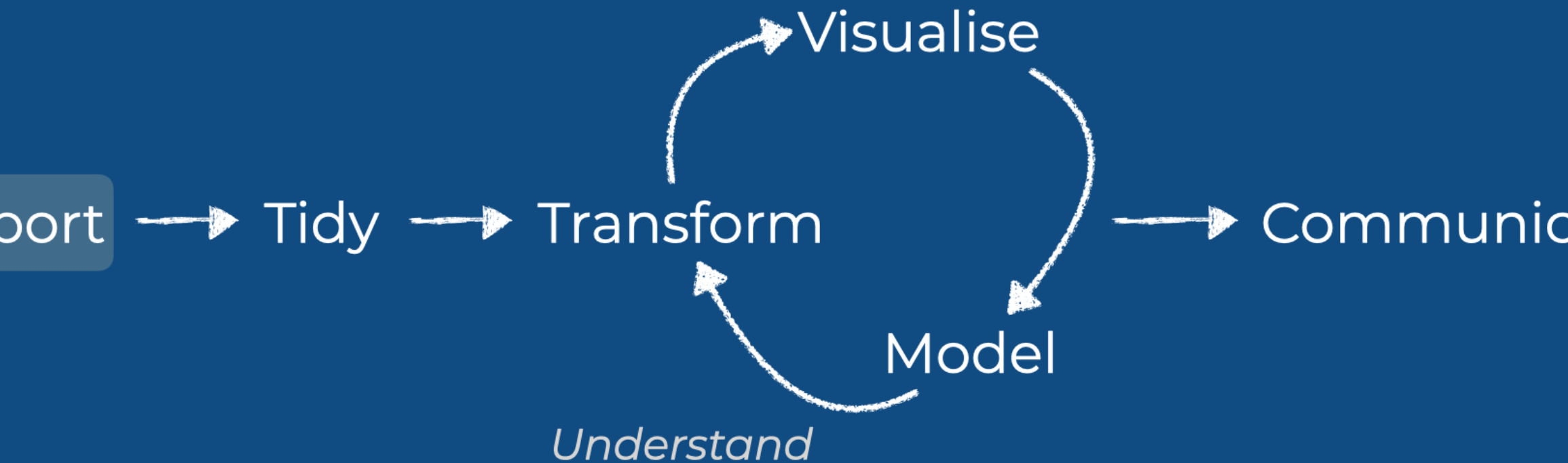


# Data Science Lifecycle



# Data Science Lifecycle

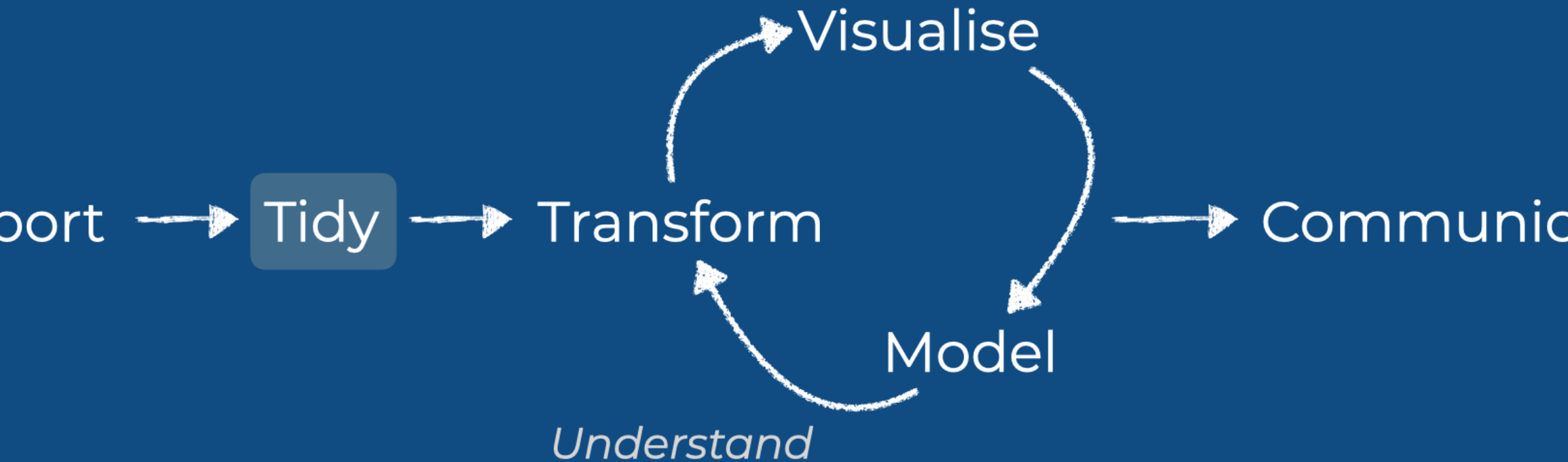
your data into R





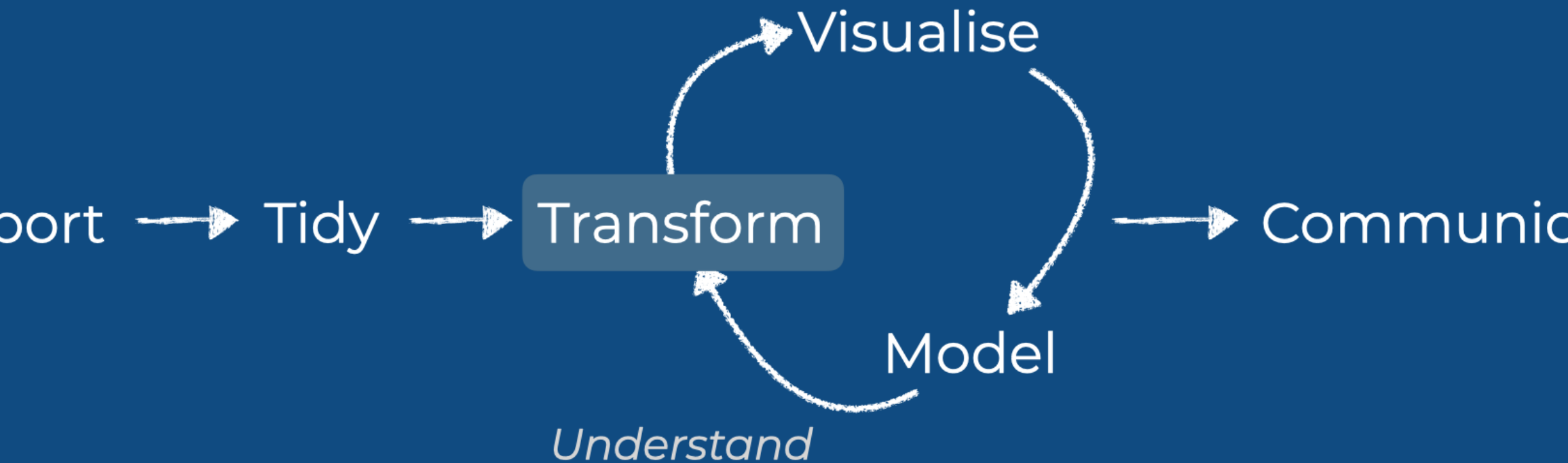
# Data Science Lifecycle

Prepare your data in a consistent form



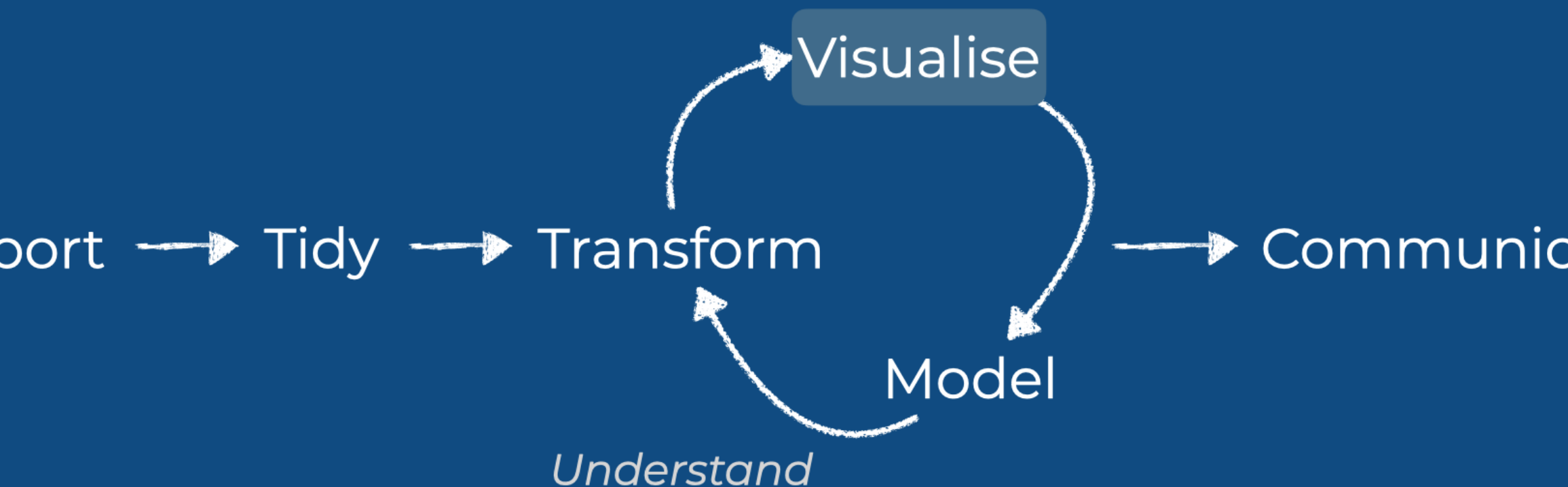
# Data Science Lifecycle

row down + Create new variables + Summary stat



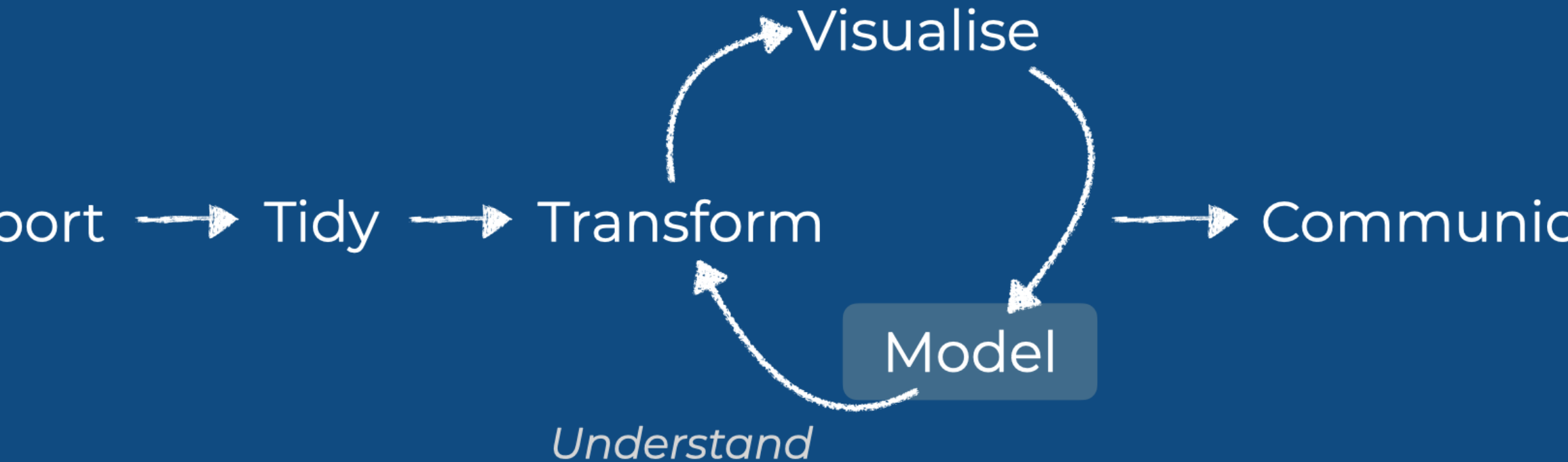
# Data Science Lifecycle

Explore your data with visual representations



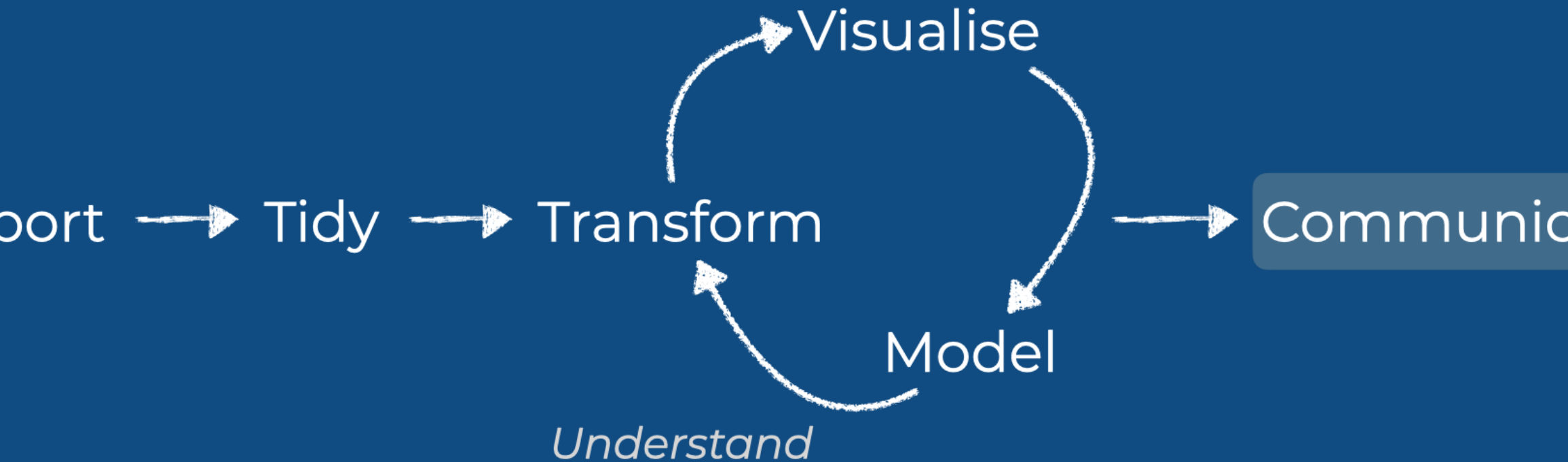
# Data Science Lifecycle

Explore your data with visual representations



# Data Science Lifecycle

Share your findings with others



R

# Packages

## base R

```
1 sqrt(49)
2 sum(1, 2)
```

- Functions come with R

## R Packages

```
1 library(dplyr)
```

- Installed once in the Console:  
`install.packages("dplyr")`
- Loaded per script

# Functions & Arguments

```
1 library(dplyr)
2
3 filter(.data = gapminder,
4        year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` **What do do with the data**



# Objects

```
1 library(dplyr)
2
3 gapminder_yr_2007 <- filter(.data = gapminder,
4                             year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` **What do do with the data**
- Object: `gapminder_yr_2007`

# Operators

```
1 library(dplyr)
2
3 gapminder_yr_2007 <- gapminder %>%
4   filter(year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` **What do do with the data**
- Object: `gapminder_yr_2007`
- Assignment operator: `<-`
- Pipe operator: `%<%`

# Rules

Rules of `dplyr` functions:

- First argument is always a data frame
- Subsequent arguments say what to do with that data frame
- Always return a data frame
- Don't modify in place

# Live Coding Exercise

# ae-10-data-science-lifecycle

1. Head over to the GitHub Organisation for the course.
2. Find the repo for week 10 that has your GitHub username.
3. Clone the repo with your username to the RStudio Cloud.
4. Open the file: `ae-10a-lifecycle.qmd`
5. Use your Sticky Notes to let me know when you are ready.

# Break



15:00

# Solving coding problems

# Tips for search engines

- Use describe 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
- Upvote 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

# Give credit



from [r cookbook](#), where bp is your ggplot:

528

Remove legend for a particular aesthetic (fill):



```
bp + guides(fill="none")
```



It can also be done when specifying the scale:

```
bp + scale_fill_discrete(guide="none")
```

This removes all legends:

```
bp + theme(legend.position="none")
```

[Share](#) [Edit](#) [Follow](#) [Flag](#)

edited Dec 2, 2021 at 7:07



Andrew Morris

408 ● 3 ● 8

answered Feb 25, 2016 at 8:48



user3490026

5,388 ● 1 ● 9 ● 4

# Give credit

Share Edit Follow Flag

edited Dec 2, 2021 at 7:07

Share a link to this answer (Includes your user id)

<https://stackoverflow.com/a/35622358/6816220>

Copy link

CC BY-SA 4.0



W Morris

3 ● 8

theme\_bw() can inte  
) , make sure to ad  
t 10:50

1



but when I do something like this `bp + theme(legend.position`

# Give credit

```
1 ggplot(data = global_waste_data_kg_year,  
2         mapping = aes(x = income_id,  
3                       y = capita_kg_year,  
4                       color = income_id)) +  
5   ## Remove legend ref: https://stackoverflow.com/a/35622358/6816220  
6   theme(legend.position = "none")
```

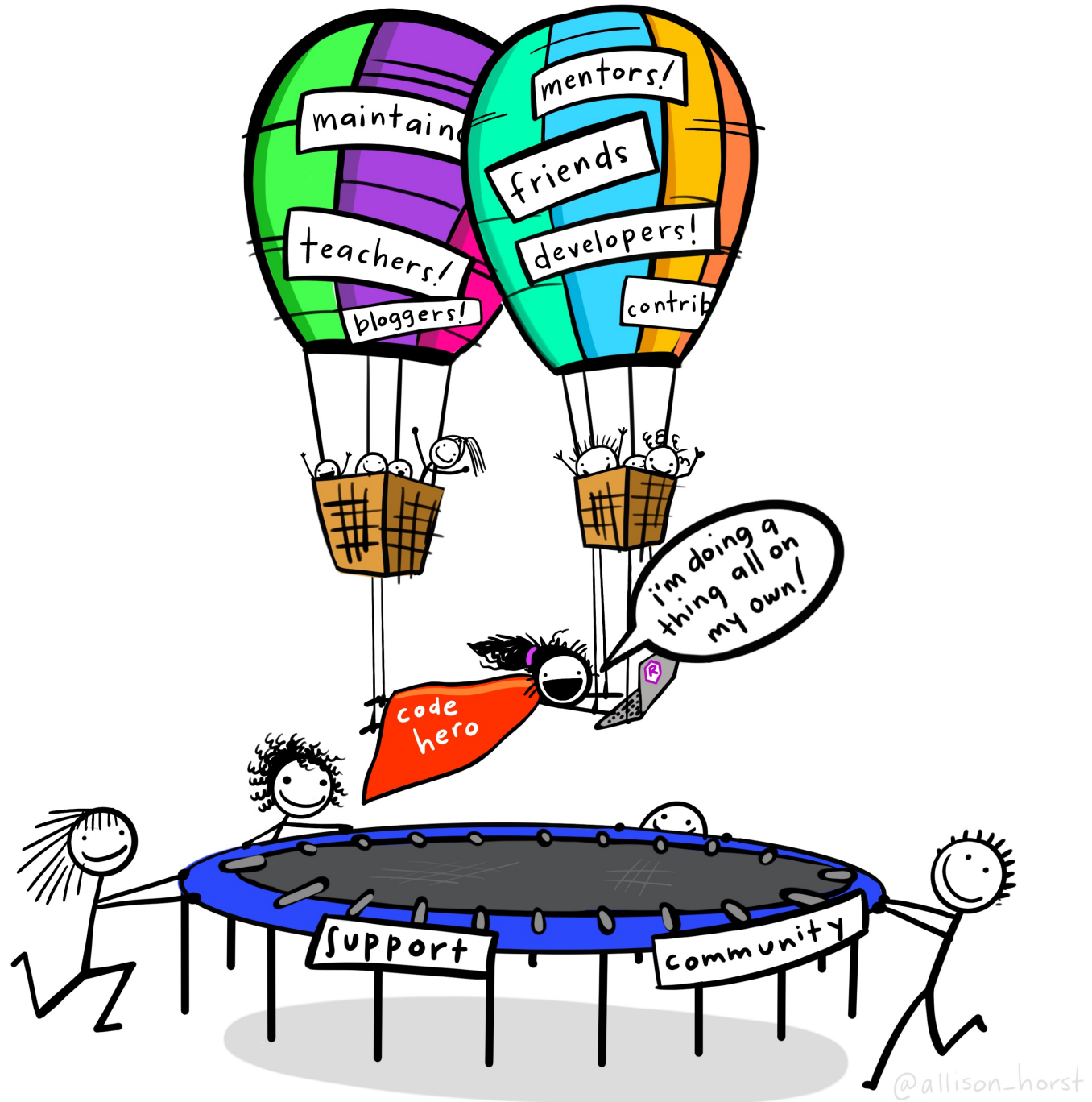
# Other sources for help

- RStudio Community Forum:  
<https://community.rstudio.com/>
- Our rbt1 Slack channel
- Documentation websites:  
<https://dplyr.tidyverse.org/>
- Twitter community: [#rstats](#)



# Minimal reproducible example (reprex)

- Needed when asking questions online
- We will practice this in another class
- Good support information: <https://www.tidyverse.org/help/#reprex>



# Break



10:00



# Pair Programming

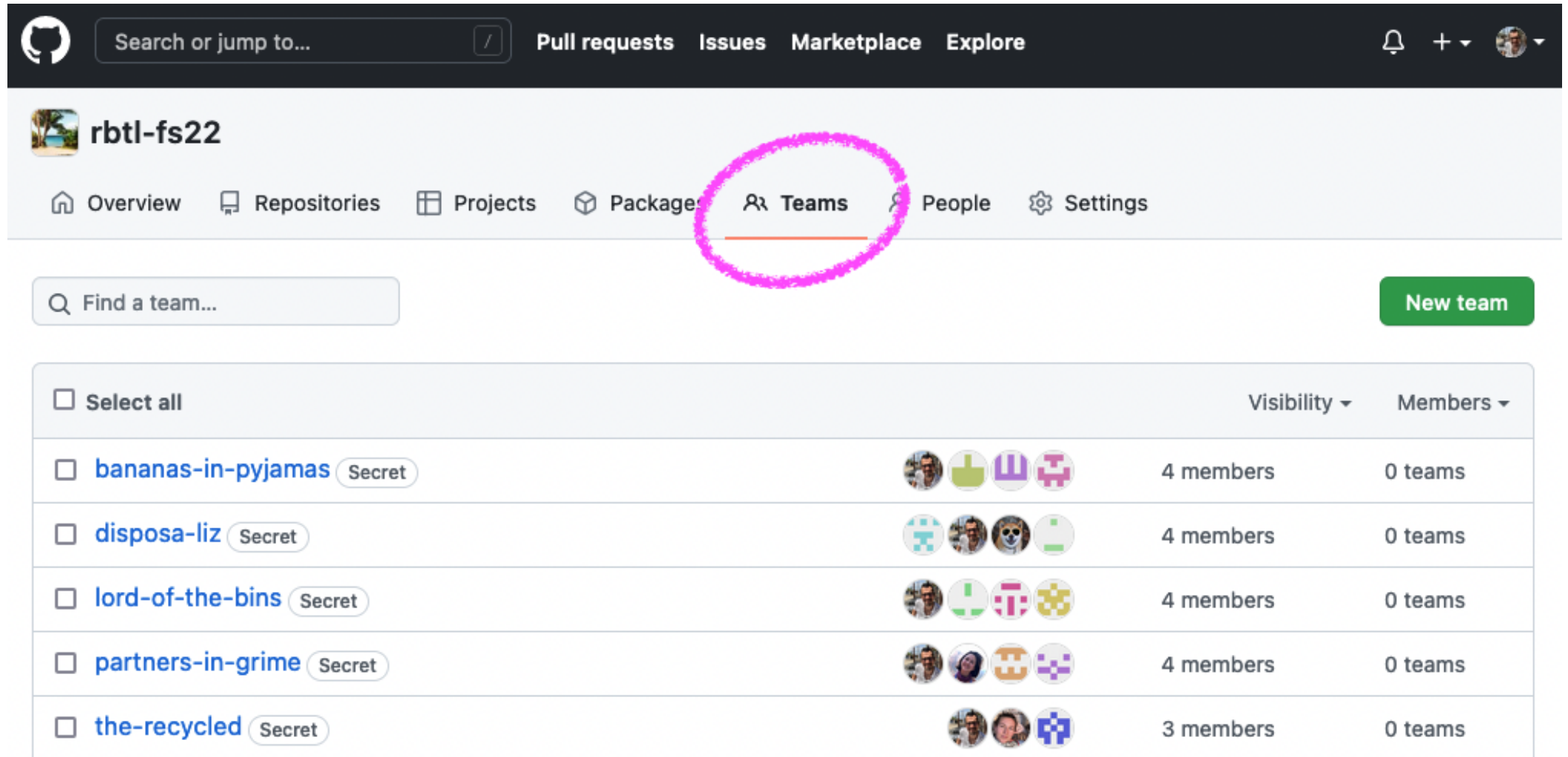
# ae-10-data-science-lifecycle

1. Team up in pairs
2. Decide who writes code, and who supports without writing code
3. Open the file: `ae-10b-lifecycle.qmd`
4. Work through the exercises
5. Use your sticky notes to indicate if you need support

30:00

# Homework Assignment

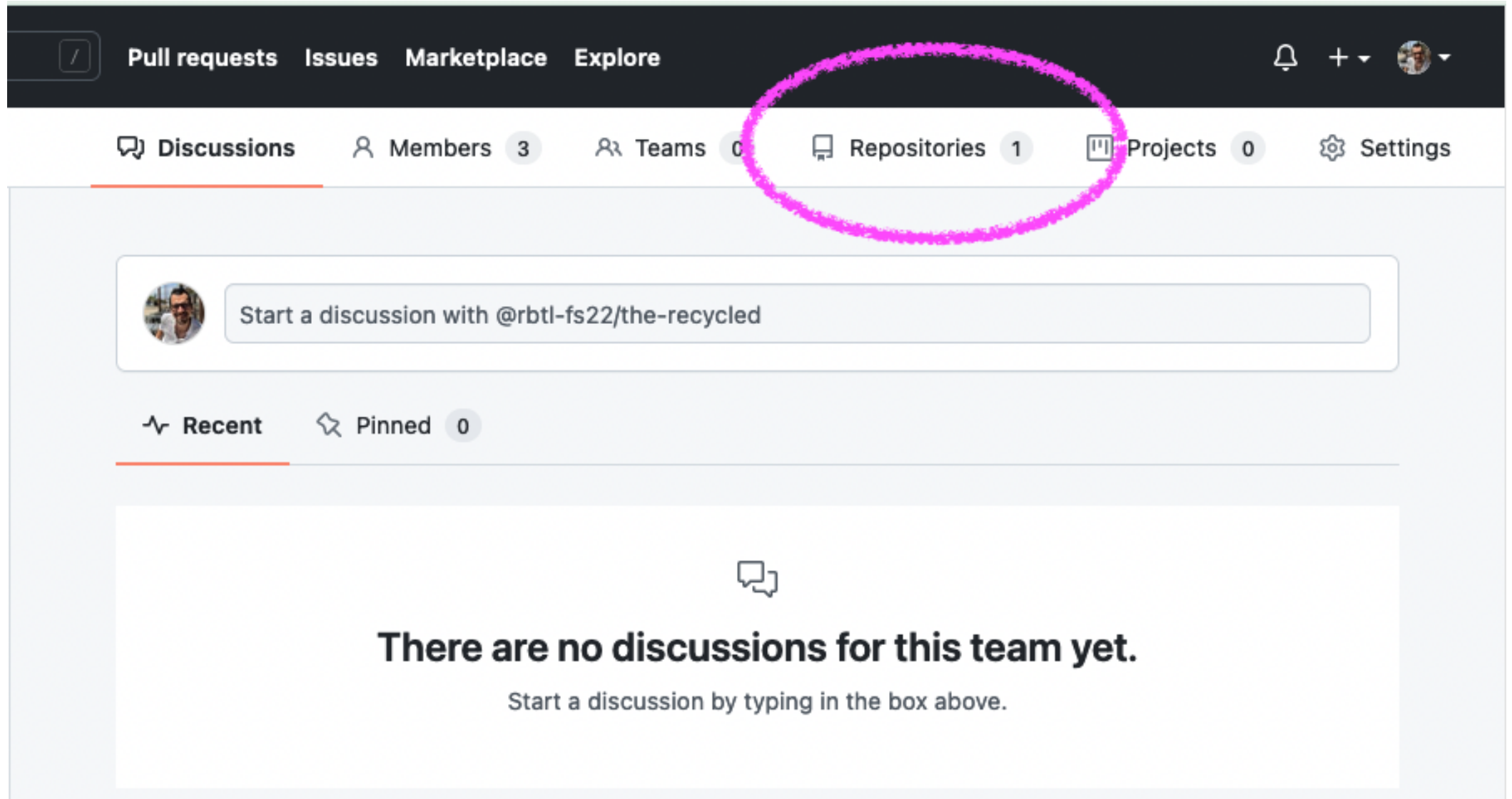
# Teams on GitHub



The screenshot shows the GitHub profile page for user 'rbtl-fs22'. The navigation bar at the top includes 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. Below the profile name, the 'Teams' tab is highlighted with a pink circle. A search bar for teams and a 'New team' button are also visible. The main content area displays a list of teams with columns for selection, team name, visibility, members, and teams.


<input type="checkbox"/> Select all		Visibility ▾	Members ▾
<input type="checkbox"/>	<a href="#">bananas-in-pyjamas</a> <span>Secret</span>		4 members 0 teams
<input type="checkbox"/>	<a href="#">disposa-liz</a> <span>Secret</span>		4 members 0 teams
<input type="checkbox"/>	<a href="#">lord-of-the-bins</a> <span>Secret</span>		4 members 0 teams
<input type="checkbox"/>	<a href="#">partners-in-grime</a> <span>Secret</span>		4 members 0 teams
<input type="checkbox"/>	<a href="#">the-recycled</a> <span>Secret</span>		3 members 0 teams


# Teams on GitHub



The screenshot displays the GitHub interface for a team. At the top, a dark navigation bar contains links for 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. On the right side of this bar are a notification bell, a plus sign, and a user profile icon. Below this, a secondary navigation bar features 'Discussions', 'Members 3', 'Teams 0', 'Repositories 1', 'Projects 0', and 'Settings'. The 'Teams 0' link is circled in pink. The main content area shows a profile picture and a text input field with the placeholder 'Start a discussion with @rbtl-fs22/the-recycled'. Below the input field are tabs for 'Recent' and 'Pinned 0'. A large white box at the bottom contains a speech bubble icon and the text: 'There are no discussions for this team yet. Start a discussion by typing in the box above.'

# Teams on GitHub

 Search or jump to... / Pull requests Issues Marketplace Explore

 [rbtl-fs22](#) / [the-recycled](#) Discussions Members 3 Teams 0 Repositories 1

Q Find a repository...

Select all

[rbtl-fs22/research-project-template-the-](#)  
[recycled](#) Private  
Updated 3 minutes ago

# Research Project Template

- Report template
- Complete list of items for grading (also as [open Google Sheet](#))
- Information on intentions for publishing

# Discuss and decide

- **Data:** Who will upload which data as part of the homework?
- **Results & Discussion:**
  - Who writes in file:
    - `03-01-results.qmd`?
    - `03-02-results.qmd`?
    - `03-03-results.qmd`?

03:00



# Submission

- All details in assignment week 10
- Due: Tuesday, 3rd May at 23:59 (2 points)

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](#).