

# Data Science for WASH

An opportunity for the sector

Lars Schöbitz

2021-04-28

Slides: [larnsce.github.io/co-wash-symposium-2021/](https://larnsce.github.io/co-wash-symposium-2021/)

Welcome! 🙌

# Lars Schöbitz

Environmental Engineer  
WASH Consultant  
Instructor for Data Science with R



# Georges Mikhael

Senior Urban Sanitation Specialist  
Consultant at Aguiconsult, UK  
Novice R user



# Housekeeping

- Keep your microphone muted
- Post question to the Zoom Chat
- Questions will be addressed at the end
- Technical difficulties cannot be addressed

**Note:** This Zoom Meeting will be recorded



## Do you sometimes wonder:

- Where people defecate in the open within a city? And if there are water bodies nearby?
- Who lives downstream of contaminated water bodies? And what the prevalence of diarrhea is in those communities?
- Whether access to safe drinking water decreases the rate of diarrheal disease?

If so, then you might also wonder:

- How to use data visualiation to answer these questions?
- How to combine your data with other open data to answer these questions?
- How to get data into the right structure to perform different types of analyses?



Artwork from @juliesquid for @openscapes (illustrated by @allison\_horst).

## Relevance of the topic

---

- Deriving actions from data plays a key role in every organisation
- Demand on WASH professionals to analyse data is increasing
- Little attention given to the competencies needed to satisfy this demand
- Investing into becoming a data-driven organisation will pay off in the long run

# Data Science - FAQ

---

## **Q: What is data science?**

A: In data science you turn raw data into understanding, insight and knowledge.

## **Q: What is R?**

A: It's a computing language used for data science.

## **Q: Is Data Science = Statistics?**

A: No, but they are closely related.

## **Q: Is Data Science = Computer Science?**

A: No, but many themes are shared.

# Data Science for WASH - FAQ

---

## **Q: Will I learn how to do machine learning?**

Nope, not at all. You will learn core concepts such as descriptive statistics, data visualisation, and (some) basic modeling.

## **Q: Do I need any prior experience?**

A: Nope, my courses are targeted at novice users.

## **Q: What do I need to learn data science?**

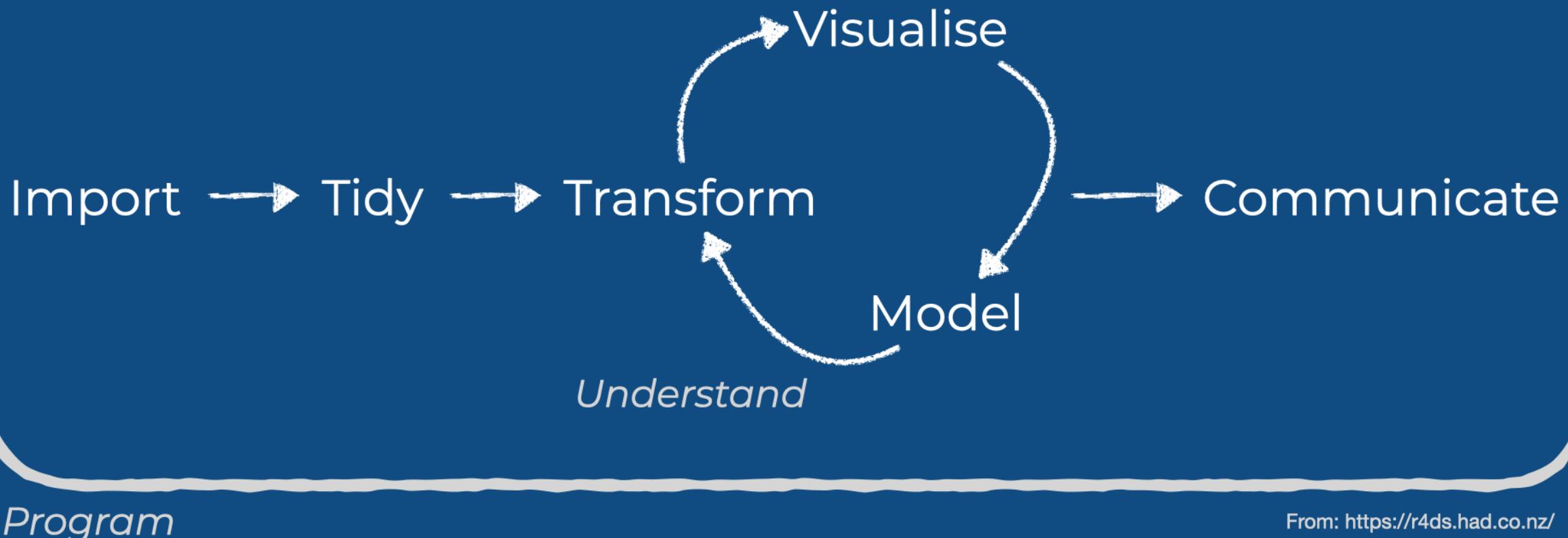
A: (1) A mindset for openness to change; (2) A good portion of vulnerability; (3) A friendly and open community.

## **Q: And how long does it take to learn data science?**

A: Three months with tutoring support and a time effort of 12 hours per week.

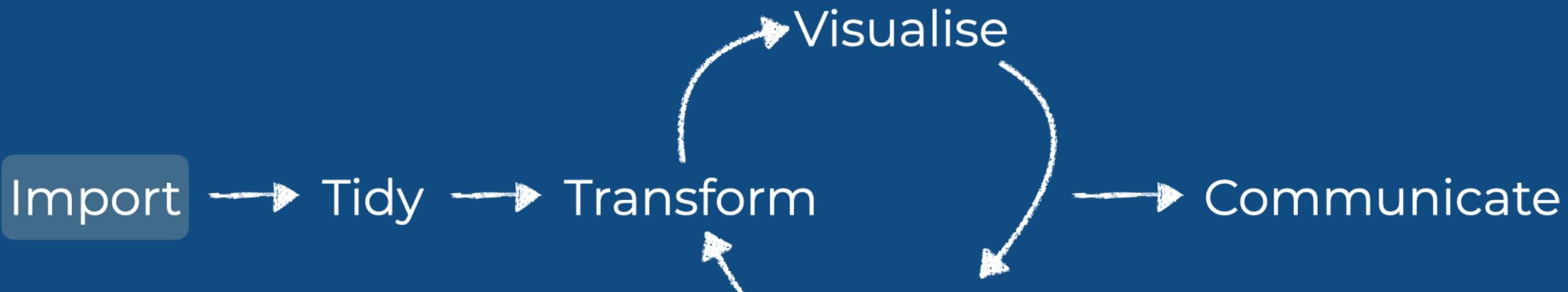
# Data Science Lifecycle

# Data Science Lifecycle



# Data Science Lifecycle

Get your data into R

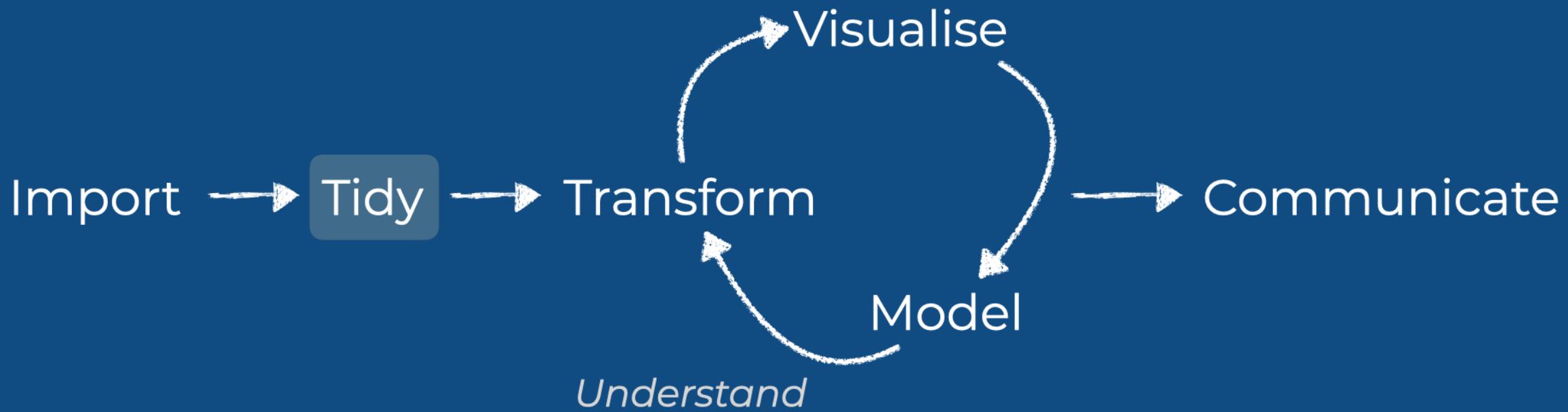


Program

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

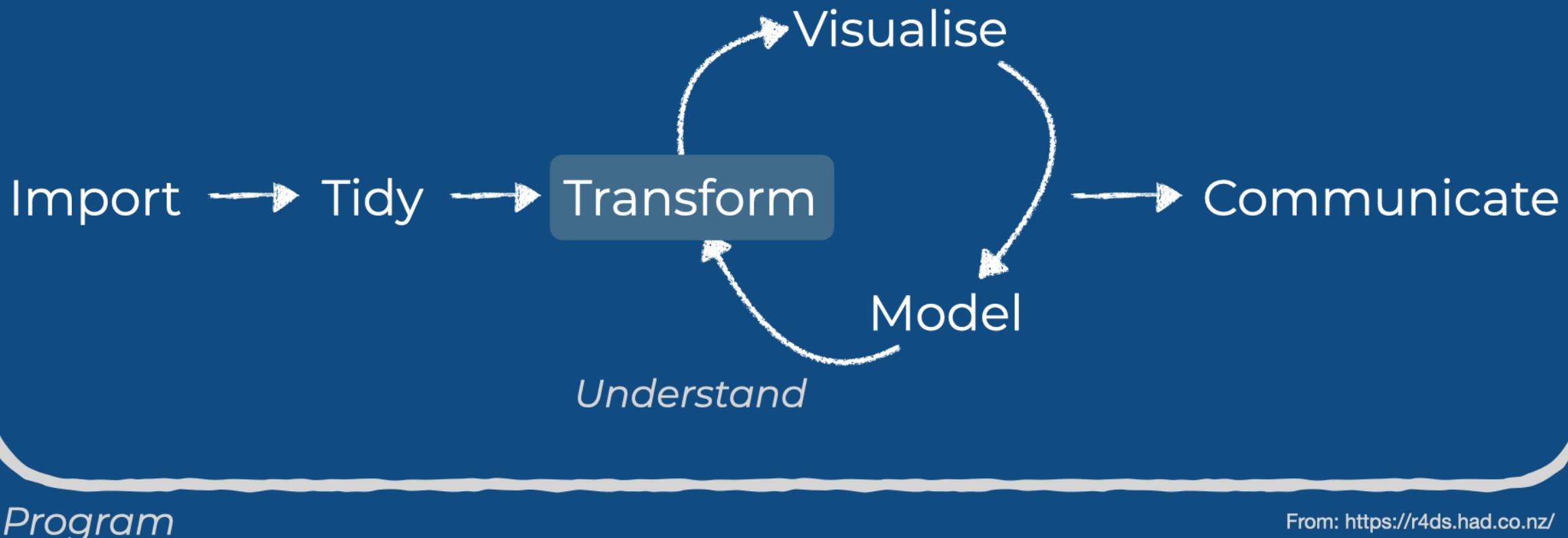
# Data Science Lifecycle

Store your data in a consistent form



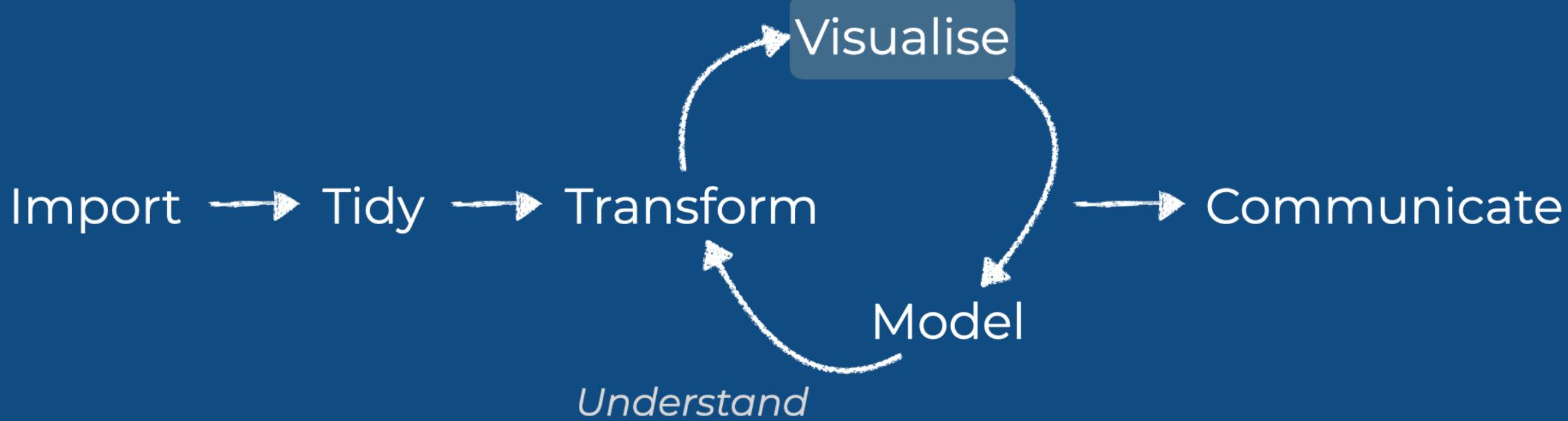
# Data Science Lifecycle

Narrow down + Create new variables + Summary stats



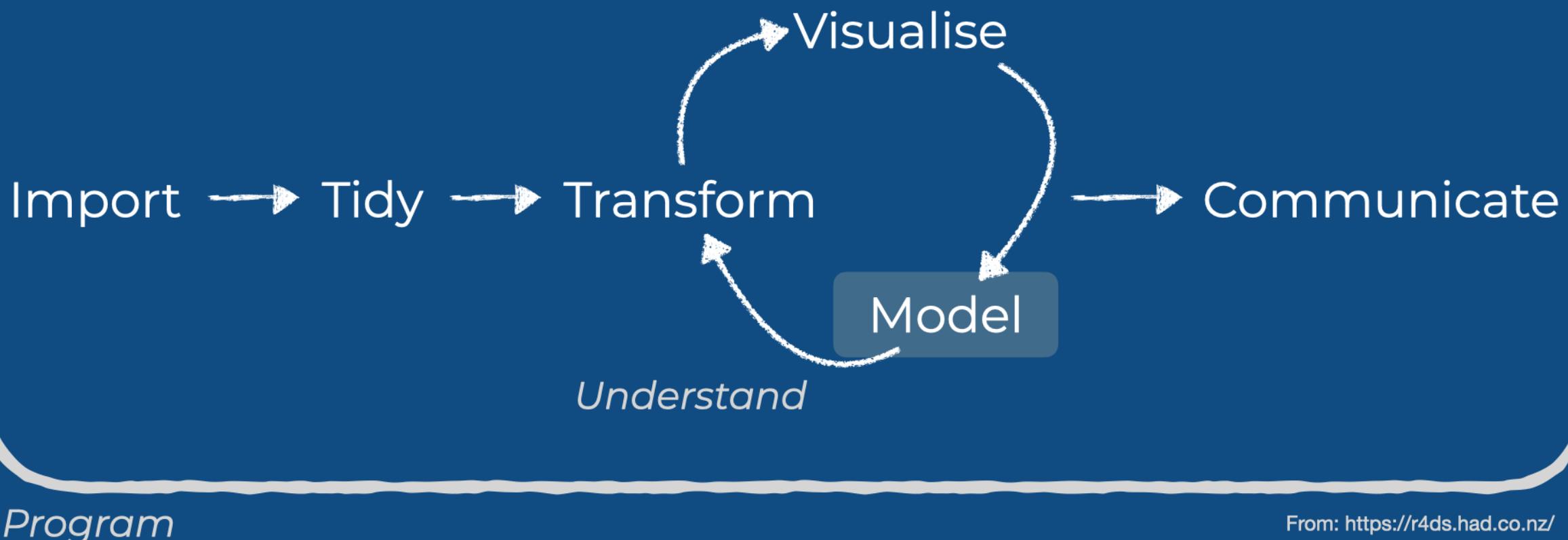
# Data Science Lifecycle

Explore your with visual representations



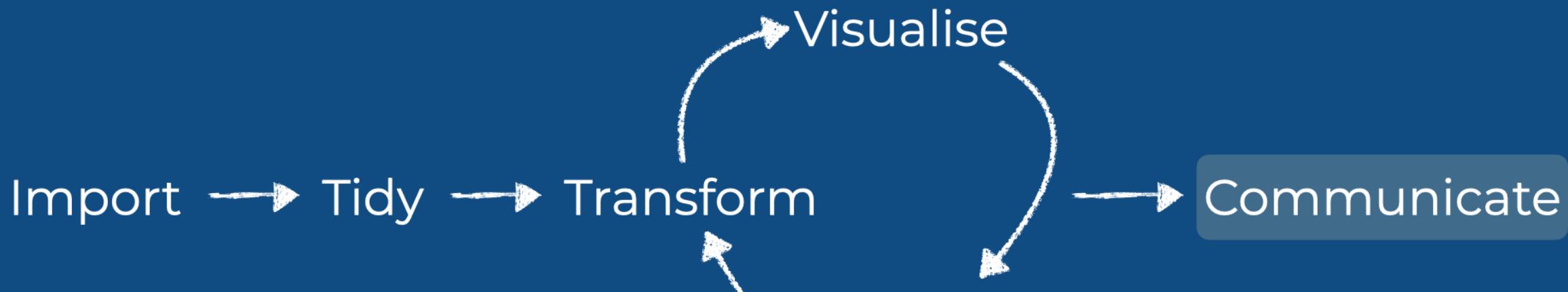
# Data Science Lifecycle

Explore your with visual representations



# Data Science Lifecycle

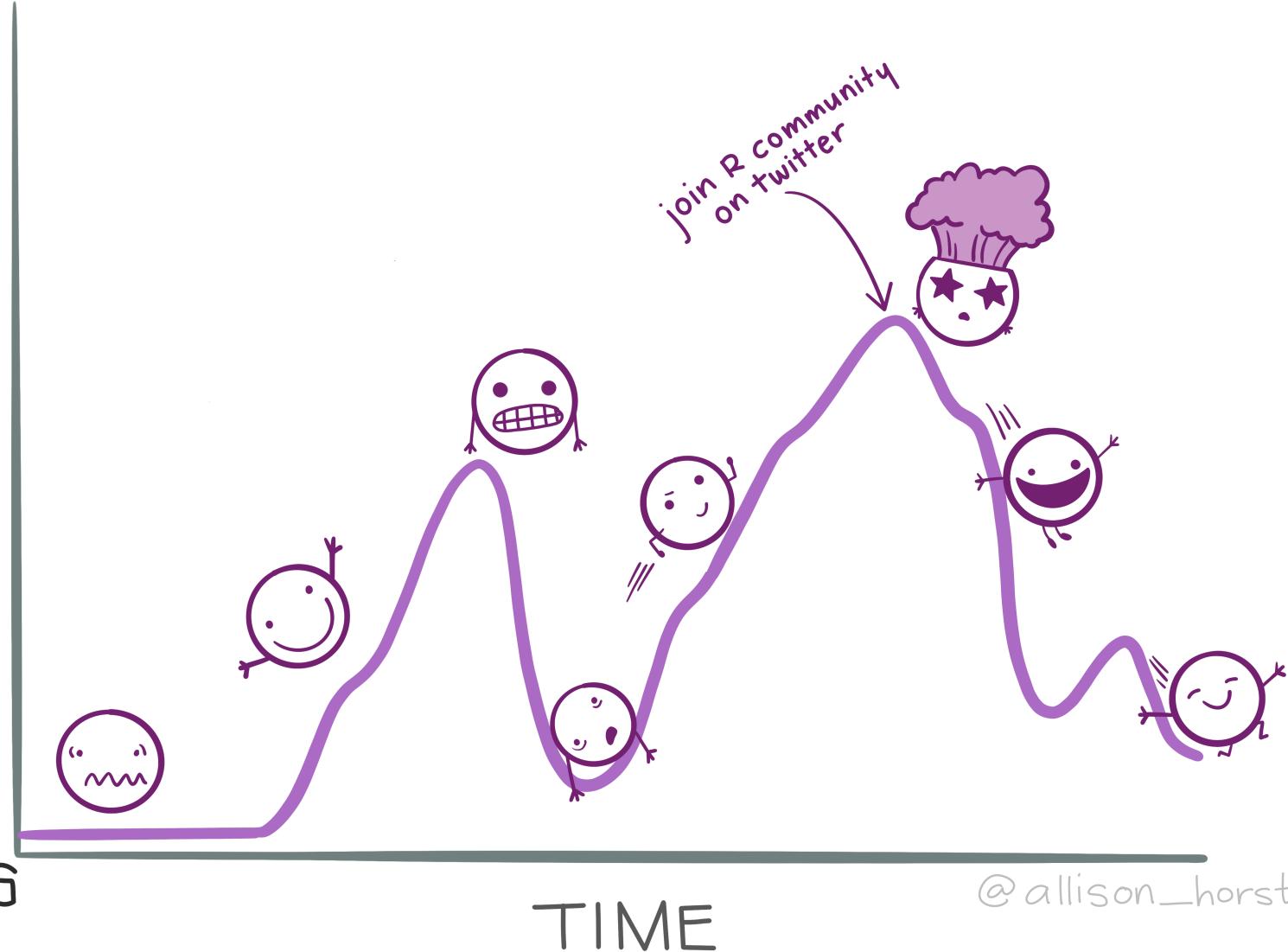
Share your findings with others



HOW  
MUCH  
I THINK  
I KNOW  
ABOUT R

I KNOW -  
NOTHING

I KNOW -  
LOTS!



@allison\_horst

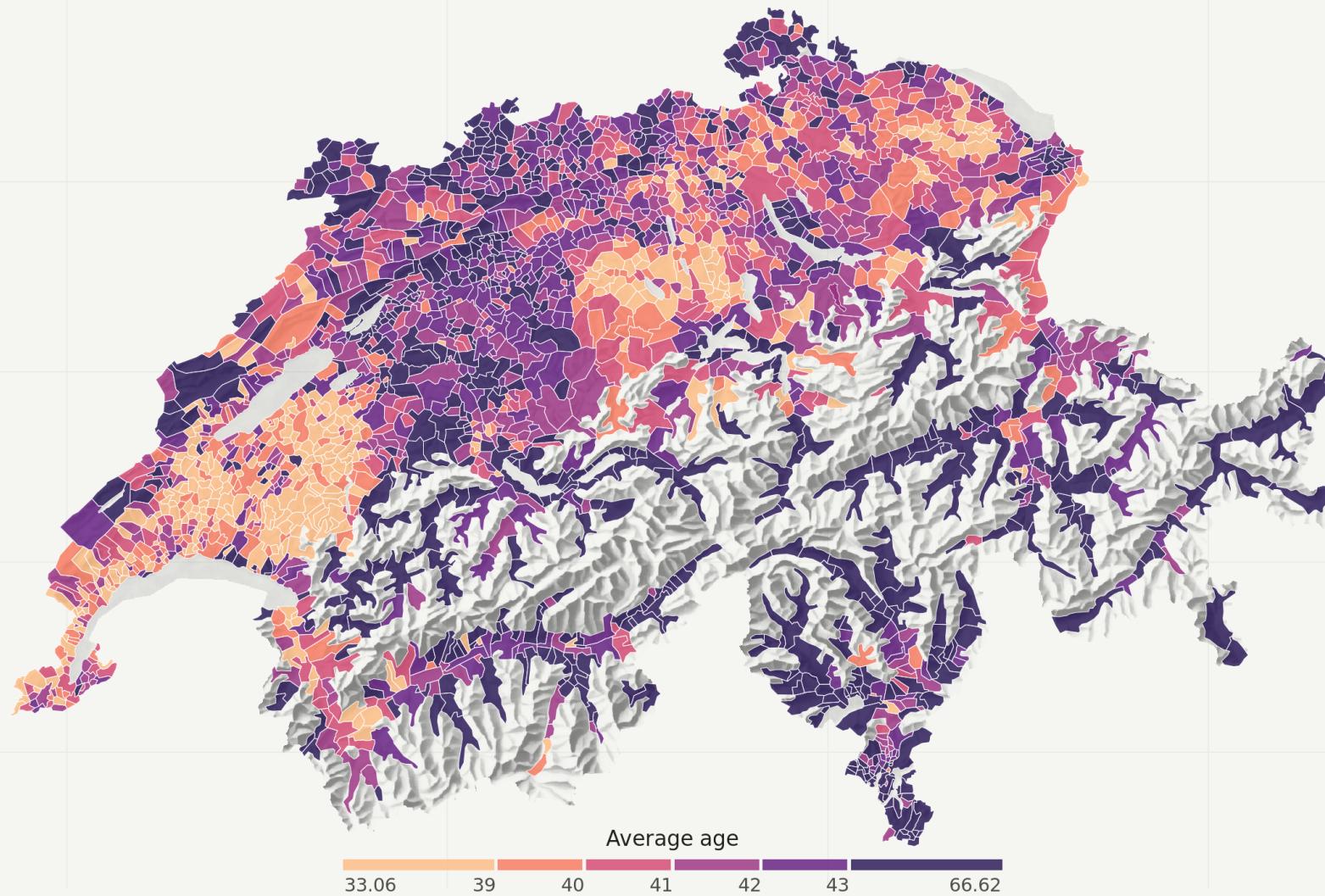
Awesome things you  
can do with R

These slides are made with R. They can include:

- Code
- Its output
- Interactive output
- Maps

# Switzerland's regional demographics

## Average age in Swiss municipalities, 2015



From: Timo Grossenbacher: Bivariate maps with `ggplot2` and `sf`

Map CC-BY-SA; Author: Timo Grossenbacher (@grssnbchr), Geometries: ThemaKart, BFS; Data: BFS, 2016; Relief: swisstopo, 2016

# **Data Science for WASH - An opportunity for the sector**

Workshop hosted at the Colorado WASH Symposium  
2021



*Before the Adventure Begins:*

# INTRO TO R



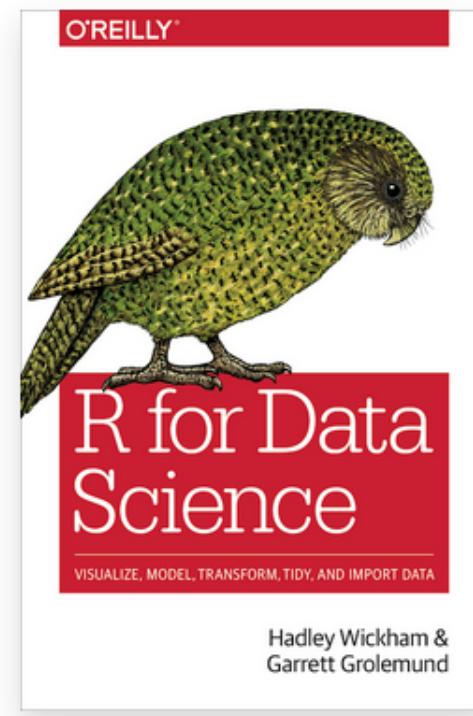
Ready to begin?

You're about to start an adventure to learn R and statistics. If this is your first time working with R, then you should begin on this page. If you're comfortable with R basics and you'd like to start with the statistical content, please proceed onto the islands with this [link](#).

From: <https://tinystats.github.io/teacups-giraffes-and-statistics/index.html>

# Welcome

This is the website for “**R for Data Science**”. This book will teach you how to do data science with R: You’ll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you’ll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You’ll learn how to use the grammar of graphics, literate



On this page

Welcome

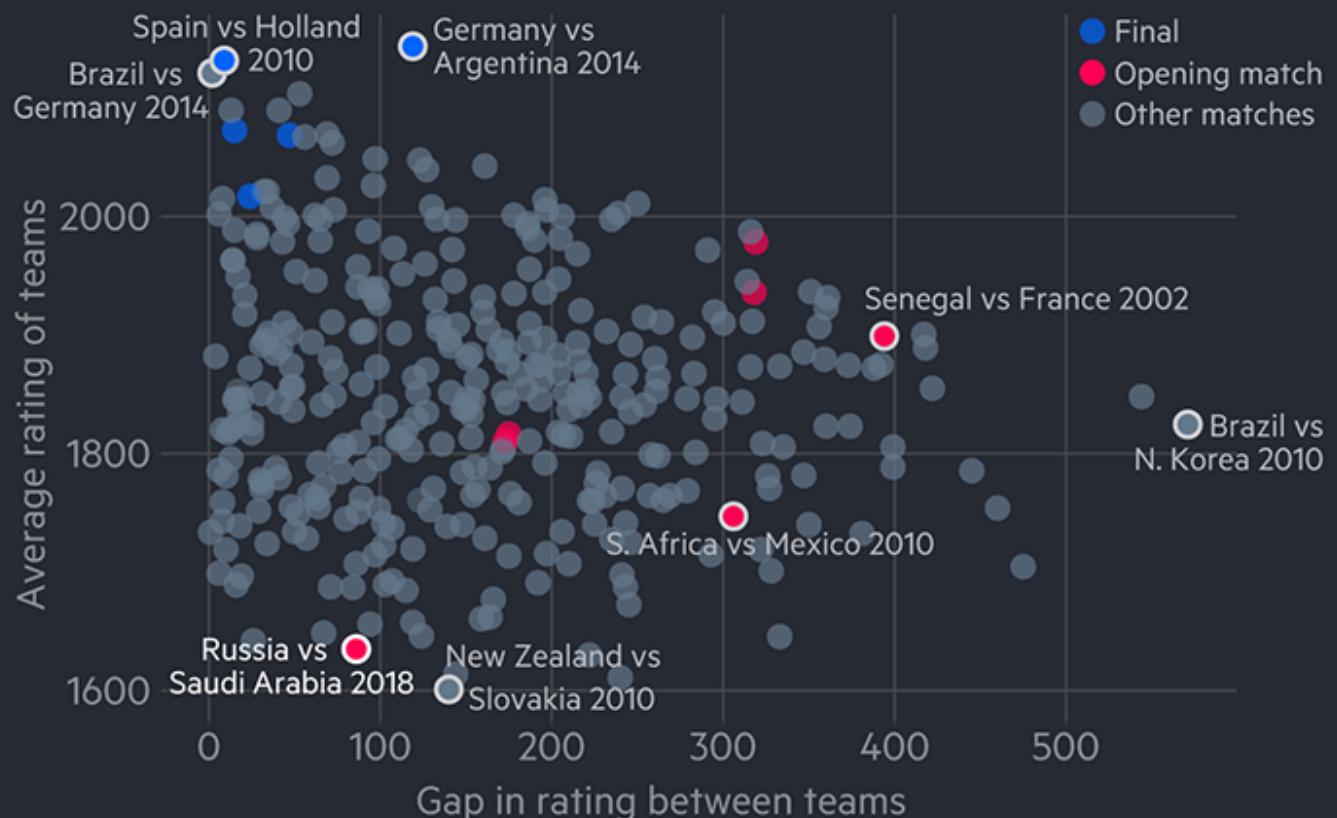
Acknowledgements

[View source](#)

[Edit this page](#)

# Russia vs Saudi Arabia: where does the oil state derby rank among the weakest World Cup matches?

Circles represent every World Cup match since 1998

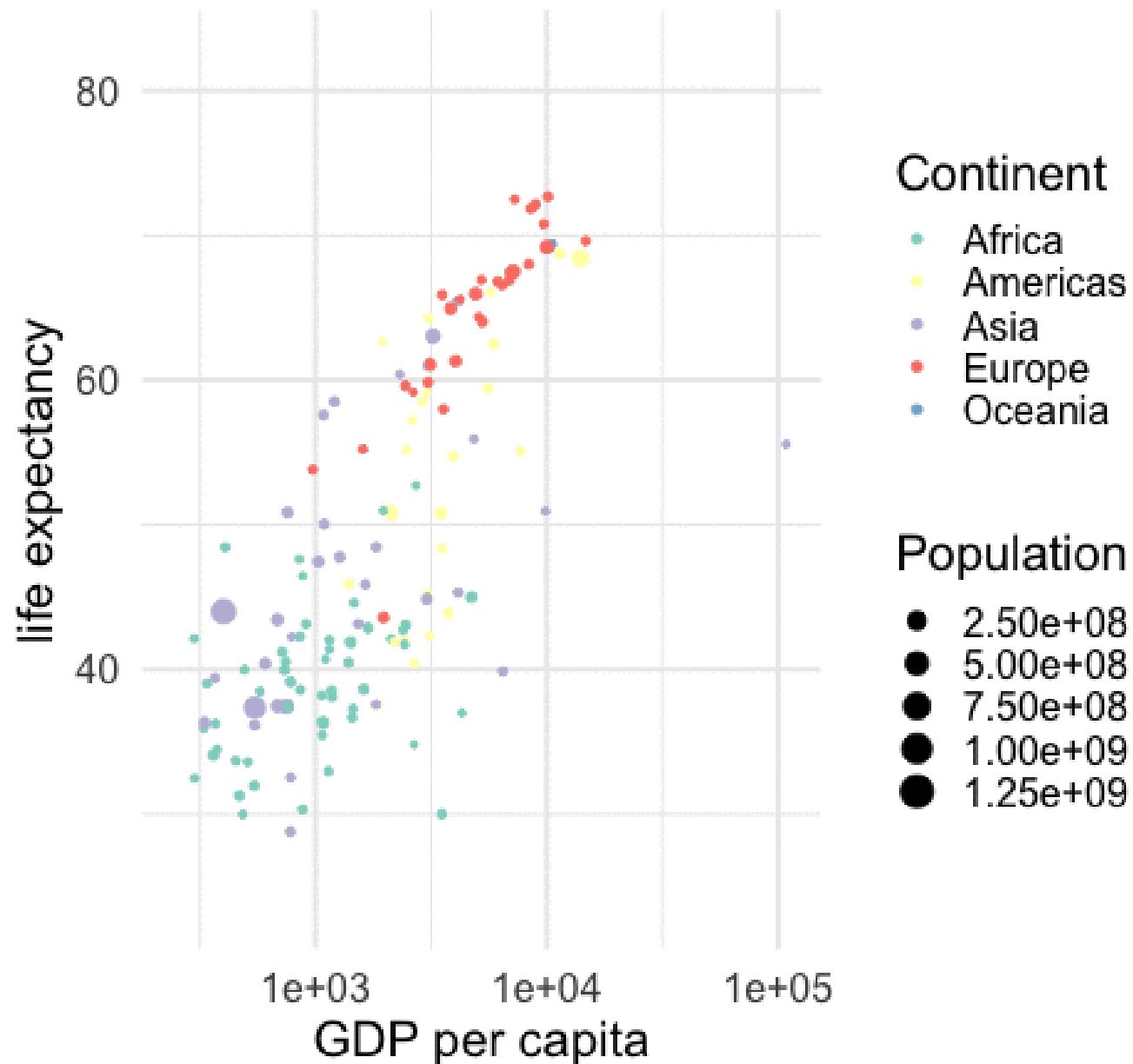


Source: eloratings.net

FINANCIAL TIMES

Year: 1952

# Hans Rosling's Gapminder



[Click here for the code](#)

# My data analysis projects

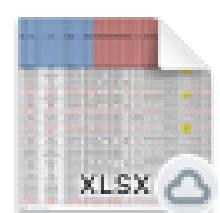
---

what it used to look like

< > raw data



FAQ data.csv



FAQ data.xlsx



FAQ\_Q\_test.csv



FAQ\_Q.csv



FAQ\_Quantificatio  
n\_1606...final.xlsx



quant-final.csv



**There are one or more circular  
references where a formula  
refers to its own cell either  
directly or indirectly. This might  
cause them to calculate  
incorrectly.**

Try removing or changing these  
references, or moving the formulae to  
different cells.

OK

Home Insert Draw Page Layout Formulas Data Review View Tell me

Share Comments

Default Page Break Preview Normal Custom Views

Zoom 150% Show 100% Zoom to 100% Zoom to Selection New Window Arrange All Freeze Panes Freeze Top Row Freeze First Column Split Hide Unhide Switch Windows View Macros Record Macro Use Relative References

	X	Y	Z	AA	AB	AC	AD
1	Solid waste production	Water usage	Excreta Production (1)	FS/WW production (2)	FS accumulation (3a) (mean)	FS accumulation (3a) (Q1)	FS accumulation (3a) (Median)
2	kg/cap*d	L/cap*d	L/d	L/d	L/d	L/d	L/d
3	0.60	150	0	0			
4	0.60	150	21'420	2'325'600	32'996	50'701	
5	0.60	150	64'260	6'976'800	98'988	152'103	
6	0.60	2	51'408	146'880	78'225	113'957	
7	0.60	2	205'632	587'520	312'898	455'827	
8	0.60	2	0	0	0		
9	0.60	2	85'680	244'800	130'374	189'928	
10	0.60	2	171'360	489'600	260'749	379'856	
11	0.60	2	171'360	489'600	79'793		
12	0.60	2	28'560	81'600	43'458	63'309	
13	0.60	2	57'120	163'200	86'916	126'619	
14	0.60	150	12'600	1'368'000	19'409	29'824	
15	0.60	150	37'800	4'104'000	58'228	121'683	
16	0.60	2	14'112	40'320	21'473	31'282	
17	0.60	2	56'448	161'280	85'894	125'129	
18	0.60	2	0	0	0		
19	0.60	2	22'680	64'800	34'511	50'275	
20	0.60	2	45'360	129'600	69'022	100'550	
21	0.60	2	45'360	129'600	21'122		
22	0.60	2	5'880	16'800	8'947	13'034	

# My data analysis projects

---

what it looks like now

RStudio Cloud

https://rstudio.cloud/project/2291449

RAM Lars Schöbitz

Your Workspace / data-science-for-wash-workshop

File Edit Code View Plots Session Build Debug Profile Tools Help

exercise-01.Rmd

```
1 ---  
2 title: "My first R Markdown report"  
3 author: "Add your name here"  
4 output: html_document  
5 editor_options:  
6   chunk_output_type: console  
7 ---  
8  
9 # R markdown file  
10  
11 This an R Markdown file. It combines text with code. This is text  
written in plain markdown and you can use markdown syntax to highlight  
text in bold, *italic* or underlined.
```

40:7 Chunk 2 R Markdown

Console Terminal R Markdown Jobs

```
/cloud/project/  
var_short = col_character(),  
percent = col_double(),  
var_long = col_character(),  
residence = col_character(),  
service = col_character(),  
indicator_type = col_character(),  
indicator = col_character()  
)  
  
>  
>  
> washdata_uga <- washdata %>%  
+ filter(iso3 == "UGA") %>%  
+ filter(year == 2017) %>%  
+ filter(residence == "national")  
>
```

Environment History Connections Git Tutorial

Import Dataset

Global Environment

Data

washdata	27171 obs. of 11 variables
washdata_uga	22 obs. of 11 variables

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Cloud > project

Name	Size	Modified
..	40 B	Mar 10, 2021, 10:15 AM
.gitignore	0 B	Mar 10, 2021, 10:15 AM
.Rhistory		
data		
exercise-01.Rmd	2 KB	Mar 10, 2021, 11:18 AM
LICENSE	1 KB	Mar 10, 2021, 10:15 AM
project.Rproj	205 B	Mar 10, 2021, 10:57 AM
README.md	0 B	Mar 10, 2021, 10:16 AM
setup		
exercise-02.Rmd	2 KB	Mar 10, 2021, 11:24 AM
exercise-01.html	724.8 KB	Mar 10, 2021, 11:34 AM

RStudio Cloud

https://rstudio.cloud/project/2291449

Your Workspace / data-science-for-wash-workshop

RAM Lars Schöbitz

File Edit Code View Plots Session Build Debug Profile Tools Help

Addins

exercise-01.Rmd

ABC Knit

---

title: "My first R Markdown report"

author: "Add your name here"

output: html\_document

---

---

# R markdown file

This an R Markdown file. It combines text with code. This is text written in plain markdown and you can use markdown syntax to highlight text in **bold**, \*italic\* or underlined.

40:7 Chunk 2 R Markdown

Console Terminal R Markdown Jobs

```
/cloud/project/
var_short = col_character(),
percent = col_double(),
var_long = col_character(),
residence = col_character(),
service = col_character(),
indicator_type = col_character(),
indicator = col_character()
)
>
>
> washdata_uga <- washdata %>%
+   filter(iso3 == "UGA") %>%
+   filter(year == 2017) %>%
+   filter(residence == "national")
>
```

Environment History Connections Git Tutorial

Import Dataset

R Global Environment

Data

washdata	27171 obs. of 11 variables
washdata_uga	22 obs. of 11 variables

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Cloud > project

Name	Size	Modified
..	40 B	Mar 10, 2021, 10:15 AM
.gitignore	0 B	Mar 10, 2021, 10:15 AM
.Rhistory		
data		
exercise-01.Rmd	2 KB	Mar 10, 2021, 11:18 AM
LICENSE	1 KB	Mar 10, 2021, 10:15 AM
project.Rproj	205 B	Mar 10, 2021, 10:57 AM
README.md	0 B	Mar 10, 2021, 10:16 AM
setup		
exercise-02.Rmd	2 KB	Mar 10, 2021, 11:24 AM
exercise-01.html	724.8 KB	Mar 10, 2021, 11:34 AM

RAM Lars Schöbitz

Code Editor

RStudio Cloud

https://rstudio.cloud/project/2291449

Your Workspace / data-science-for-wash-workshop

RAM Lars Schöbitz

File Edit Code View Plots Session Build Debug Profile Tools Help

Addins

exercise-01.Rmd x Knit Addins

---  
title: "My first R Markdown report"  
author: "Add your name here"  
output: html\_document  
edit: ch  
---

This is an R Markdown file. It combines text with code. This is text written in plain markdown and you can use markdown syntax to highlight text in **bold**, *italic* or underlined.

40:7 Chunk 2 R Markdown

Console Terminal R Markdown Jobs

```
/cloud/project/ ↗  
var_short = col_character(),  
percent = col_double(),  
var_long = col_character(),  
residence = col_character(),  
service = col_character(),  
indicator_type = col_character(),  
indicator = col_character()  
)  
  
>  
>  
> washdata_uga <- washdata %>%  
+ filter(iso3 == "UGA") %>%  
+ filter(year == 2017) %>%  
+ filter(residence == "national")  
>
```

Environment History Connections Git Tutorial

Import Dataset

R Global Environment

Data

washdata	27171 obs. of 11 variables
washdata_uga	22 obs. of 11 variables

Code Editor

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Cloud > project

Name	Size	Modified
..	40 B	Mar 10, 2021, 10:15 AM
.gitignore	0 B	Mar 10, 2021, 10:15 AM
.Rhistory		
data		
exercise-01.Rmd	2 KB	Mar 10, 2021, 11:18 AM
LICENSE	1 KB	Mar 10, 2021, 10:15 AM
project.Rproj	205 B	Mar 10, 2021, 10:57 AM
README.md	0 B	Mar 10, 2021, 10:16 AM
setup		
exercise-02.Rmd	2 KB	Mar 10, 2021, 11:24 AM
exercise-01.html	724.8 KB	Mar 10, 2021, 11:34 AM

The screenshot shows the RStudio Cloud interface with several panes:

- Code Editor (Left Pane):** Displays an R Markdown file named "exercise-01.Rmd". The code includes a YAML header and a chunk of text. Two specific buttons in the toolbar above the editor are circled in pink: "Knit" and "Build". A large pink box covers the entire Code Editor area.
- Environment (Top Right Pane):** Shows the Global Environment and Data sections. The Data section lists "washdata" and "washdata\_uga" datasets. A large blue box covers the Environment and Data sections.
- Files (Bottom Right Pane):** Shows the project's directory structure with files like ".gitignore", ".Rhistory", "data", "exercise-01.Rmd", "LICENSE", "project.Rproj", "README.md", "setup", "exercise-02.Rmd", and "exercise-01.html". A large blue box covers the Files section.
- Console (Bottom Left Pane):** Displays R code related to data manipulation, including the creation of variables like "var\_short", "percent", "var\_long", etc., and filtering data from "washdata" to "washdata\_uga".

The screenshot shows the RStudio Cloud interface with four main panels:

- Code Editor** (Top Left, pink border): An R Markdown file titled "exercise-01.Rmd". The code includes a YAML header and a chunk that outputs text. Two specific buttons in the toolbar are circled in red: "Knit" and "Run".

```
1 ---  
2 title: "My first R Markdown report"  
3 author: "Add your name here"  
4 output: html_document  
5  
6 edit  
7 ch  
8 ---  
9 # R markdown file  
10  
11 This an R Markdown file. It combines text with code. This is text  
written in plain markdown and you can use markdown syntax to highlight  
text in bold, italic or underlined.  
40:7 C Chunk 2
```
- Environment** (Top Right, blue border): Shows the Global Environment and Data pane. The Data pane lists two datasets: "washdata" (27171 obs. of 11 variables) and "washdata\_uga".

```
Environment History Connections Git Tutorial  
Import Dataset  
R Global Environment  
Data  
washdata 27171 obs. of 11 variables  
washdata_uga
```
- File Manager** (Bottom Left, green border): Shows the project directory structure. The "data" folder contains "washdata.csv" and "washdata\_uga.csv". Other files include ".gitignore", ".Rhistory", "LICENSE", "project.Rproj", "README.md", "setup", "exercise-01.Rmd", and "exercise-01.html".

```
Files Plots Packages Help Viewer  
New Folder Upload Delete Rename More  
Cloud > project  
Name Size Modified  
.. 1, 10:15 AM  
.gitignore 1, 10:15 AM  
.Rhistory 1, 11:18 AM  
data 1, 10:15 AM  
washdata.csv 1, 10:57 AM  
washdata_uga.csv 0 B Mar 10, 2021, 10:16 AM  
LICENSE 1, 10:15 AM  
project.Rproj 1, 10:57 AM  
README.md 0 B Mar 10, 2021, 11:24 AM  
setup 2 KB Mar 10, 2021, 11:34 AM  
exercise-01.Rmd 724.8 KB Mar 10, 2021, 11:34 AM  
exercise-01.html 0 B Mar 10, 2021, 11:34 AM
```
- Viewer** (Bottom Right, green border): A placeholder panel for viewing results.

The screenshot shows the RStudio Cloud interface with four main panels:

- Code Editor** (Top Left, pink border): An R Markdown file titled "exercise-01.Rmd". The code includes a YAML header and a chunk that outputs text. Two specific buttons in the toolbar are circled in pink: "Knit" and "Run".

```
1 ---  
2 title: "My first R Markdown report"  
3 author: "Add your name here"  
4 output: html_document  
5  
6 edit  
7 ch  
8 ---  
9 # R markdown file  
10  
11 This an R Markdown file. It combines text with code. This is text  
written in plain markdown and you can use markdown syntax to highlight  
text in bold, italic or underlined.  
40:7 C Chunk 2
```
- Environment** (Top Right, blue border): Shows the Global Environment and Data pane. The Data pane lists two datasets: "washdata" (27171 obs. of 11 variables) and "washdata\_uga".

```
Environment History Connections Git Tutorial  
Import Dataset  
R Global Environment  
Data  
washdata 27171 obs. of 11 variables  
washdata_uga
```
- File Manager** (Bottom Right, green border): Shows the project directory structure. A file named "exercise-02.Rmd" is highlighted with a green oval.

Name	Size	Modified
..		1, 10:15 AM
.gitignore		1, 10:15 AM
.Rhistory		
data		
exercise-01.Rmd		1, 11:18 AM
LICENSE		1, 10:15 AM
project.Rproj		1, 10:57 AM
README.md		
setup		
exercise-02.Rmd	2 KB	Mar 10, 2021, 11:24 AM
- Viewer** (Bottom Left, white background): Shows the R Markdown preview of the "exercise-01.Rmd" file.

The screenshot shows the RStudio Cloud interface with four main panels:

- Code Editor** (Top Left, pink border): An R Markdown file titled "exercise-01.Rmd". The code includes a YAML header and a body with a note about R Markdown syntax. Two buttons in the toolbar are circled in red: "Knit" and "Run".
- Environment** (Top Right, blue border): Shows the Global Environment and Data pane. The Data pane lists "washdata" and "washdata\_uga" datasets.
- Console** (Bottom Left, green border): Displays R code for creating variables and a chain of filter operations on the "washdata" dataset.
- File Manager** (Bottom Right, green border): Shows the project directory structure with files like ".gitignore", ".Rhistory", "data", "LICENSE", "project.Rproj", "README.md", "setup", "exercise-01.Rmd", and "exercise-01.html".

Let's put *work* into this workshop

# Requirements

---

1. A free account on RStudio Cloud
  - <https://rstudio.cloud/plans/free>
2. One of Mozilla Firefox, Google Chrome, Microsoft Edge, Safari, Opera
  - just **not** the Internet Explorer
3. A laptop or desktop computer
  - it will be hard to do on a phone or tablet
4. And if you haven't seen it, please read the [Code of Conduct](#) after this workshop

**Window 1: RStudio Cloud**

File Edit Code Help  
R 4.0.3

exercise-02.Rmd

```
1 ---  
2 title: "My first R Markdown report"  
3 author: "Lars Schöbitz"  
4 output: html_document  
5 editor_options:  
  chunk_output_type: console  
7 ---  
8  
9 # R markdown file  
10  
11 This an R Markdown file. It combines text  
with code. This is text written in plain  
markdown and you can use markdown syntax to  
highlight text in **bold**, *italic* or  
_underlined_.
```

Console Terminal Jobs

where you write code yourself

**Window 2: Zoom**

Your Workspace / data-science-for-wash-workshop

File Edit Code View Plots Session Build Debug Profile Tools Help

exercise-02.Rmd

```
1 ---  
2 title: "My First R Markdown report"  
3 author: "Lars Schöbitz"  
4 output: html_document  
5 editor_options:  
6 chunk_output_type: console  
7 ---  
8  
9 # R markdown file  
10  
11 This an R Markdown file. It combines text with code. This is text written in plain markdown and you can use markdown syntax to highlight text in **bold**, *italic* or _underlined_.  
12  
13 # Tasks  
9:18
```

Console Terminal Jobs

Cloud project

where you watch me write code

Environment History Connections Git Tutorial

Global Environment

Data

- washdata 45936 obs. of 11 variables
- washdat... 22 obs. of 11 variables

Files Plots Packages Help Viewer

New Folder Upload Delete

- Name
- .Rhistory
- data
- exercise-01.html
- exercise-01.Rmd
- exercise-02.html
- exercise-02.Rmd
- exercise-03.html
- exercise-03.Rmd
- LICENSE
- project.Rproj
- README.md

R Markdown

Cloud project

Name	iso3	year	pop_n	var_short	percent	var_long	residence	service
Afghanistan	AFG	2015	33736.	san_bas	40.7	At least..	national	sanita..
Afghanistan	AFG	2018	34656.	san_bas	42.1	At least..	national	sanita..
Afghanistan	AFG	2017	35530.	san_bas	43.4	At least..	national	sanita..
Albania	ALB	2015	2923.	san_bas	97.7	At least..	national	sanita..
Albania	ALB	2016	2926.	san_bas	97.7	At least..	national	sanita..
Albania	ALB	2017	2930.	san_bas	97.7	At least..	national	sanita..

## Your Turn

---

Step 1: Open this link in your browser

[rstudio.cloud/project/2301653](https://rstudio.cloud/project/2301653)

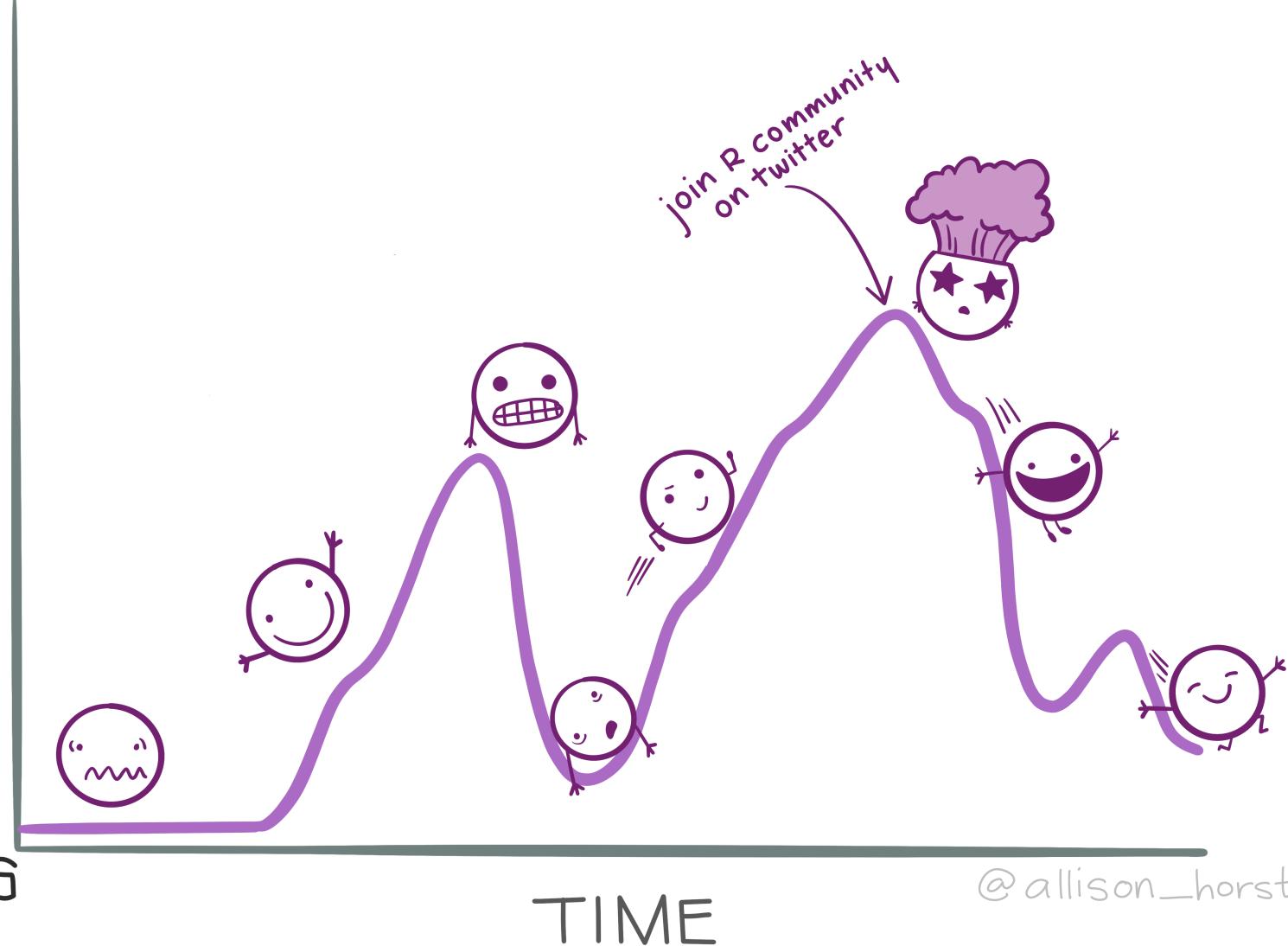
Step 2: Create your own permanent project copy

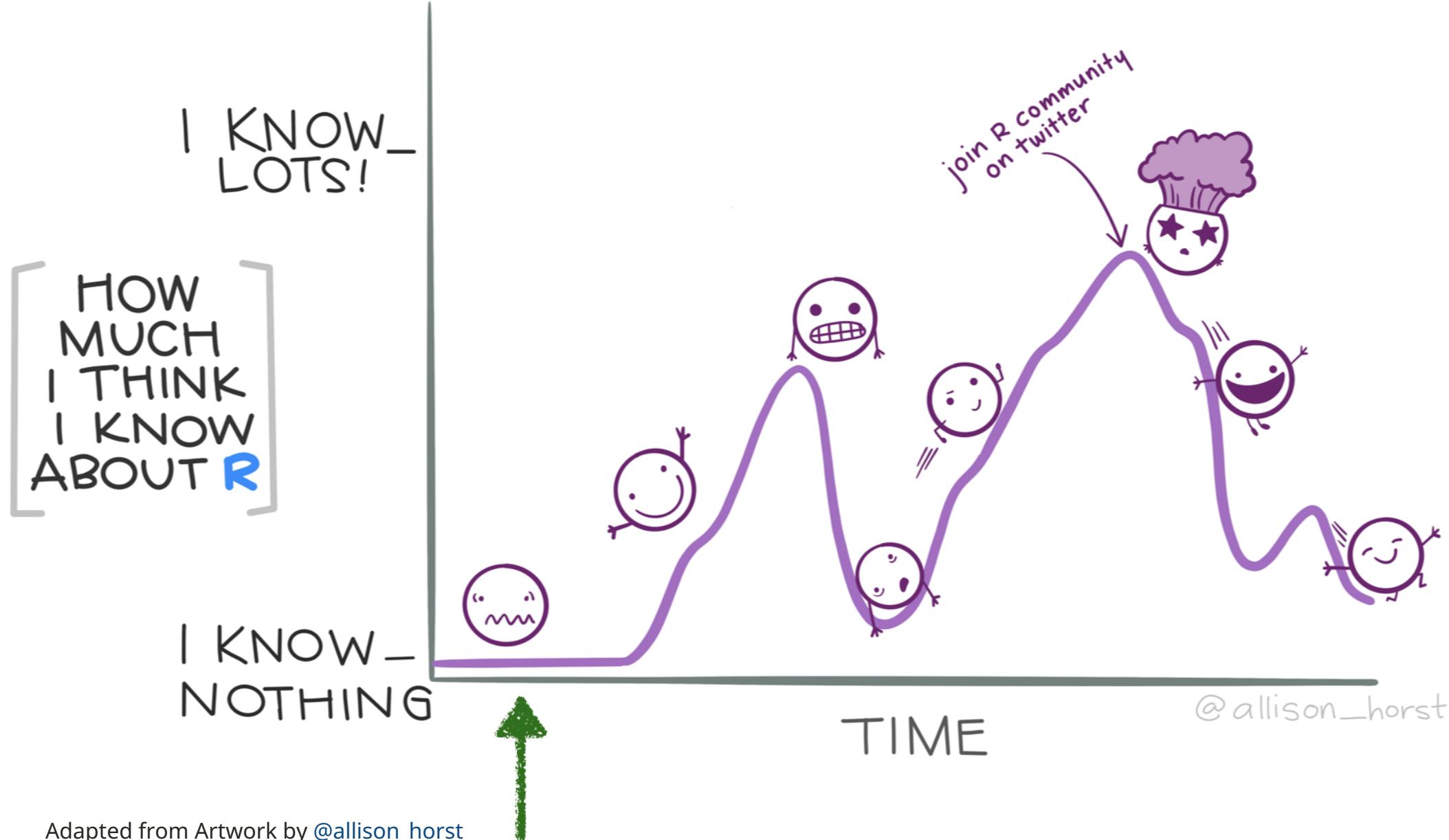
# What's next?

HOW  
MUCH  
I THINK  
I KNOW  
ABOUT R

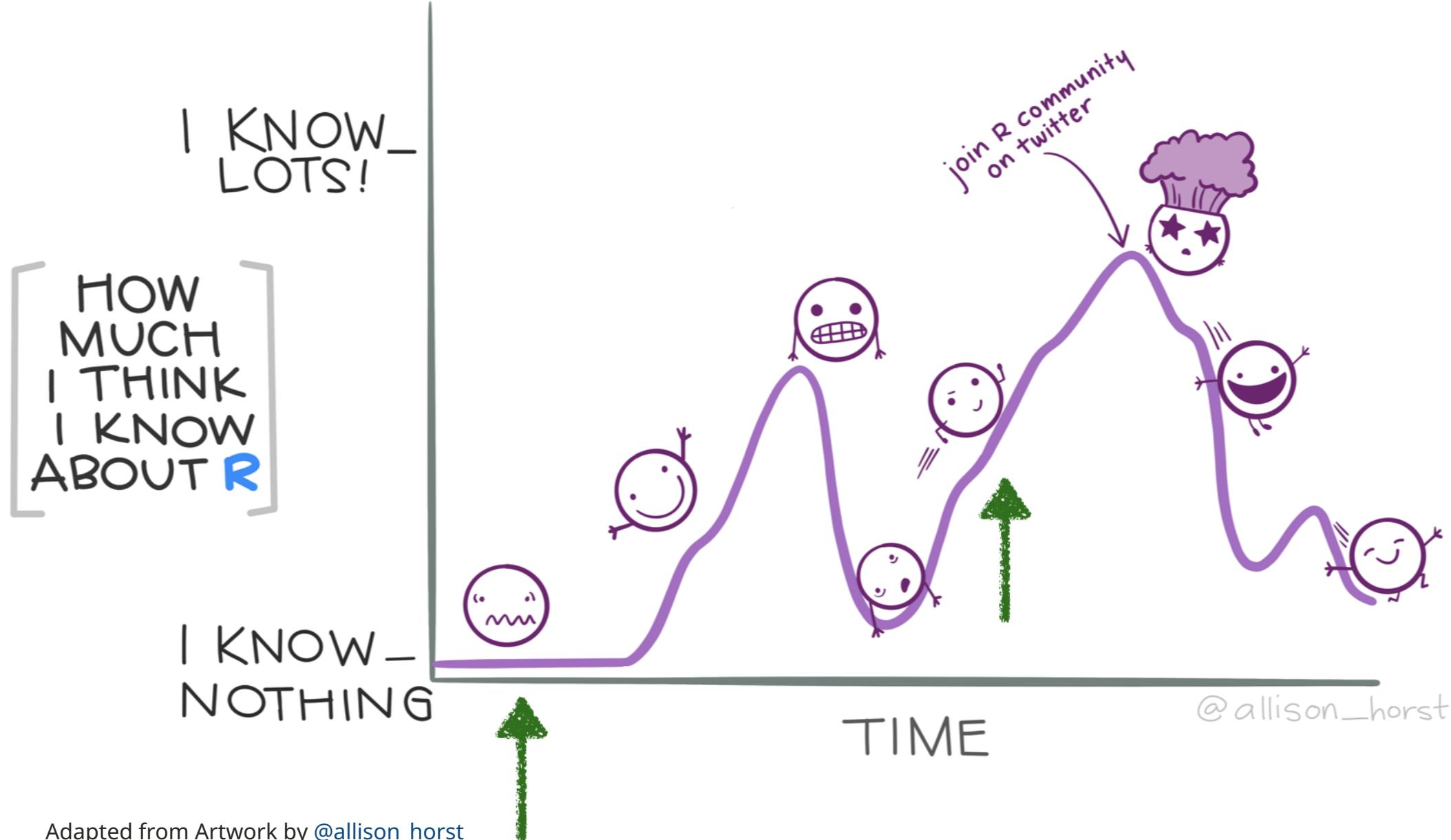
I KNOW\_  
LOTS!

I KNOW\_  
NOTHING





Adapted from Artwork by [@allison\\_horst](#)



Adapted from Artwork by [@allison\\_horst](#)

If you could not follow through the exercises for any reason

---

Contact me: [Lars@Lse.de](mailto:Lars@Lse.de)

If you are interested in a course on Data Science for WASH

---

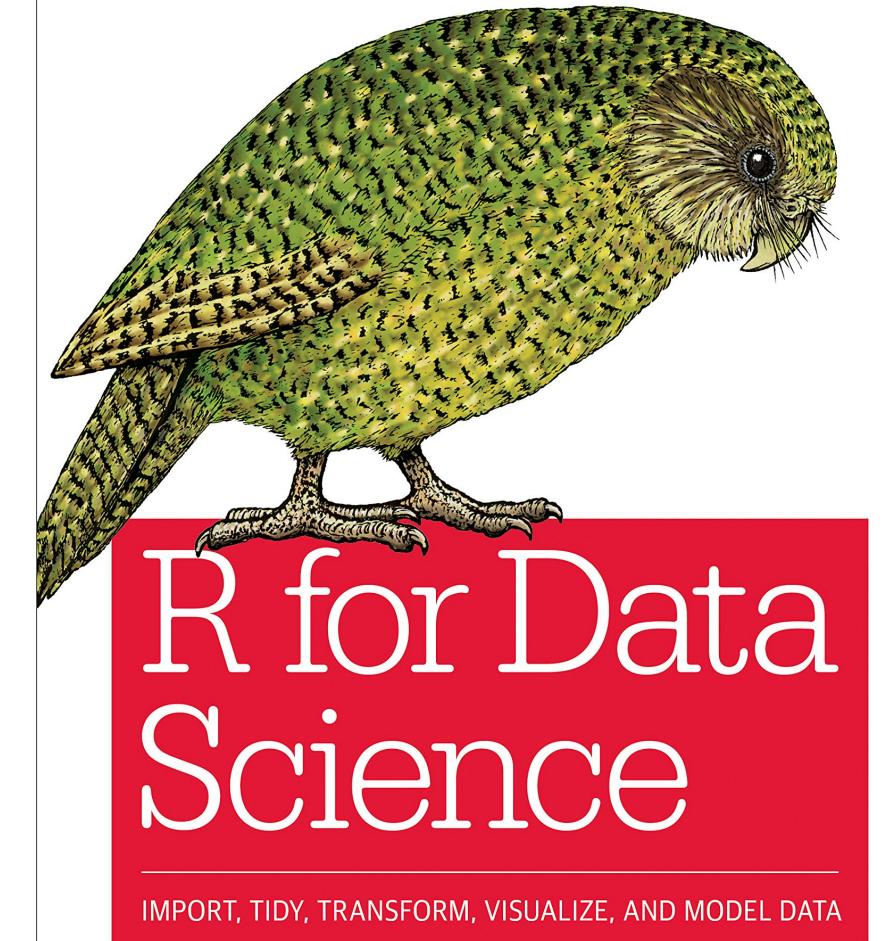
Fill out this form:

[forms.gle/XaDNE3Z6SJrxo2BM8](https://forms.gle/XaDNE3Z6SJrxo2BM8)

- 10 questions
- done in 5 minutes

## If you want to continue learning now

- Book: <https://r4ds.had.co.nz/>
- Community: <https://www.rfordatasci.com/>



Hadley Wickham &  
Garrett Grolemund

If you are interested in following along the development of  
Data Science for WASH

---

- Join Slack Workspace: [\*\*Click here\*\*](#)
- Follow along on Twitter: [\*\*@washdata\*\*](#)
- Get in touch via E-Mail: [\*\*Lars@Lse.de\*\*](mailto:Lars@Lse.de)

## If you know of someone that could be interested

---

Share info about next round of this workshop

- **Date:** 29th July 2021
- **Time:** 3 PM (CET)
- **Sign-up Link:** <https://us02web.zoom.us/meeting/register/tZcucO-qpjItGdRYUuhfU120KpUCnnTejkqm>



# Thank you!

---

For joining!

For R packages `{xaringan}` and `{xaringanthemer}`, which where used to create these slides.

All material is licensed under Creative Commons Attribution Share Alike 4.0 International.

---

## Slides

PDF version: [Download here](#)

Web version: <https://larnsce.github.io/co-wash-symposium-2021/>