

Data Science for WASH

An opportunity for the sector

Lars Schöbitz

2021-03-10

Welcome! 🙌

Lars Schöbitz

Environmental Engineer
WASH Consultant
Instructor for Data Science with R

Georges Mikhael

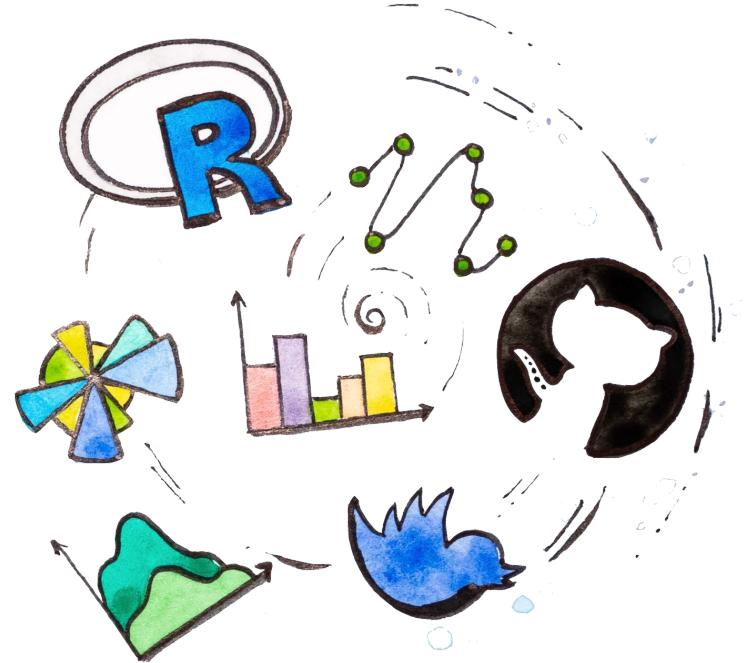
Senior Urban Sanitation Specialist
Consultant at Aguiconsult, UK
Novice R user

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 you to use your data, in combination with public open data, to answer these questions
- How you can read, transform and analyse data from different sources, and of different structures



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 and share knowledge is increasing
- Little attention given to the resources and competencies needed to satisfy this demand

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. We explore 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?

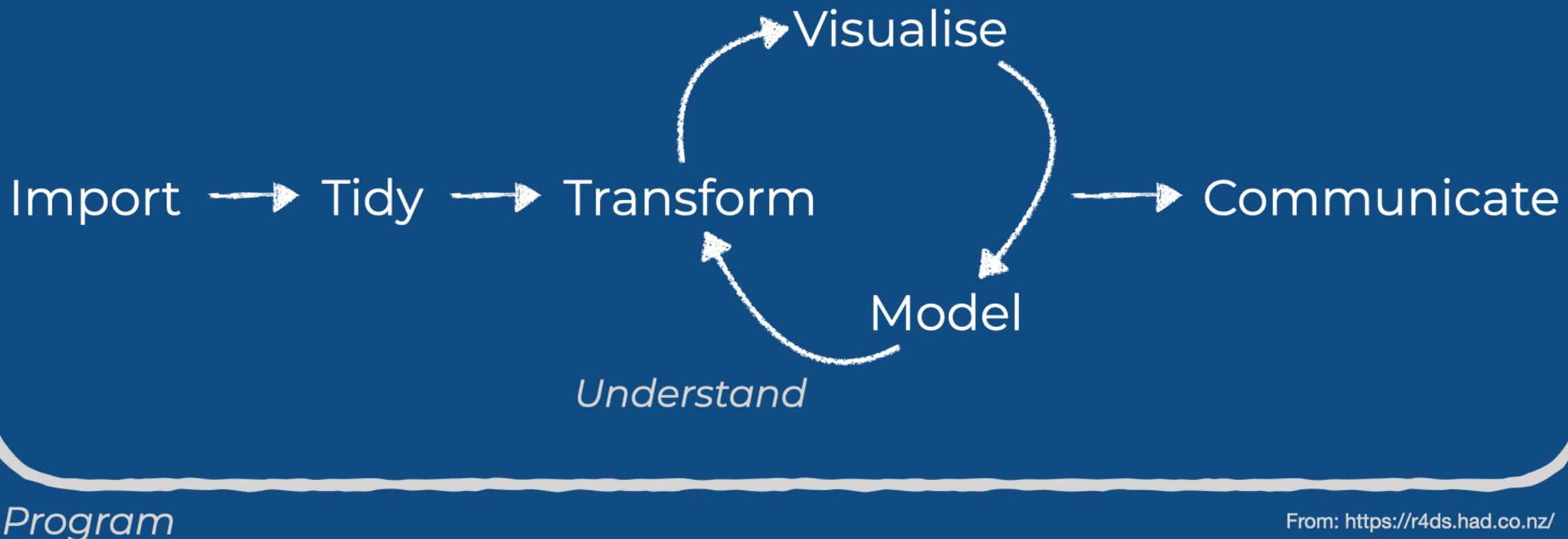
A: A mindset for openness to change; A good portion of vulnerability; A friendly and open community

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

A: A minimum of 3 months with tutoring support and time effort of 12 hours per week

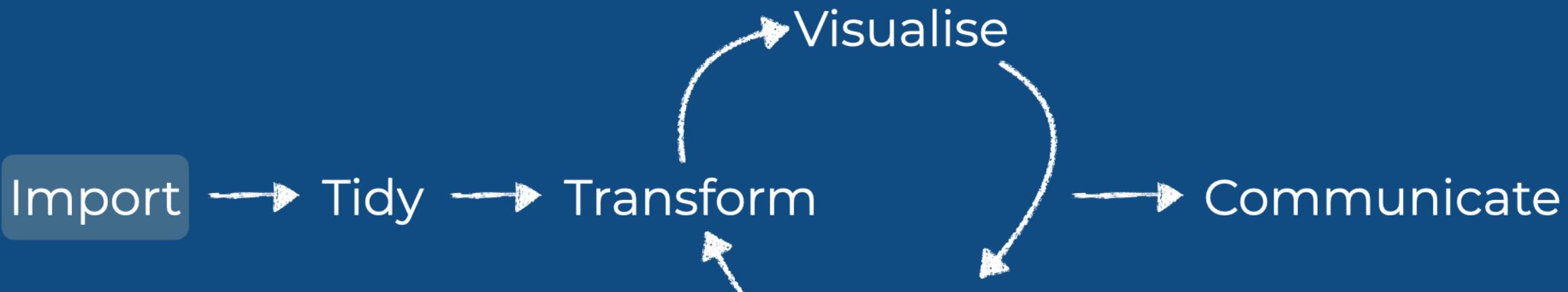
Data Science Lifecycle

Data Science Lifecycle



Data Science Lifecycle

Get your data into R

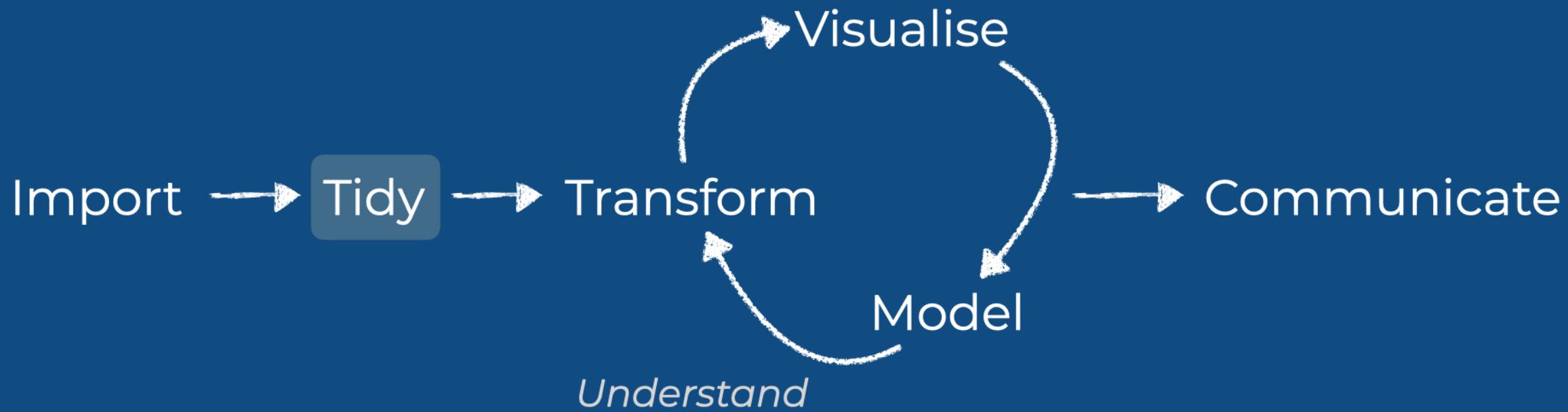


Program

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

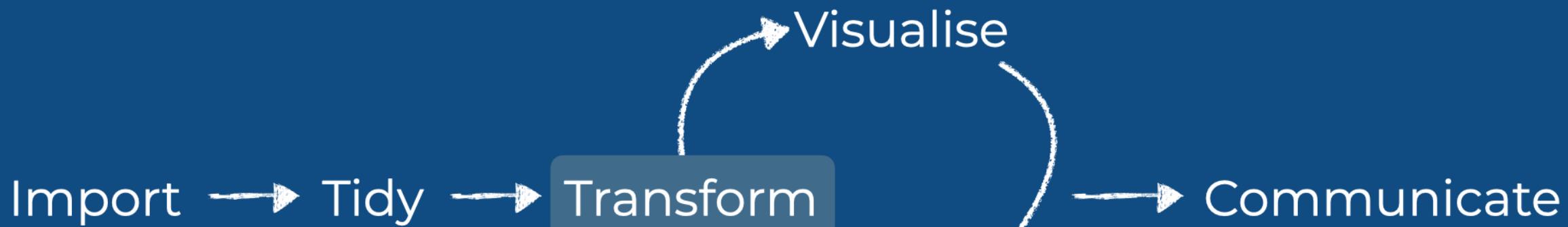
Data Science Lifecycle

Store your data in a consistent form



Data Science Lifecycle

Narrow down + Create new variables + Summary stats

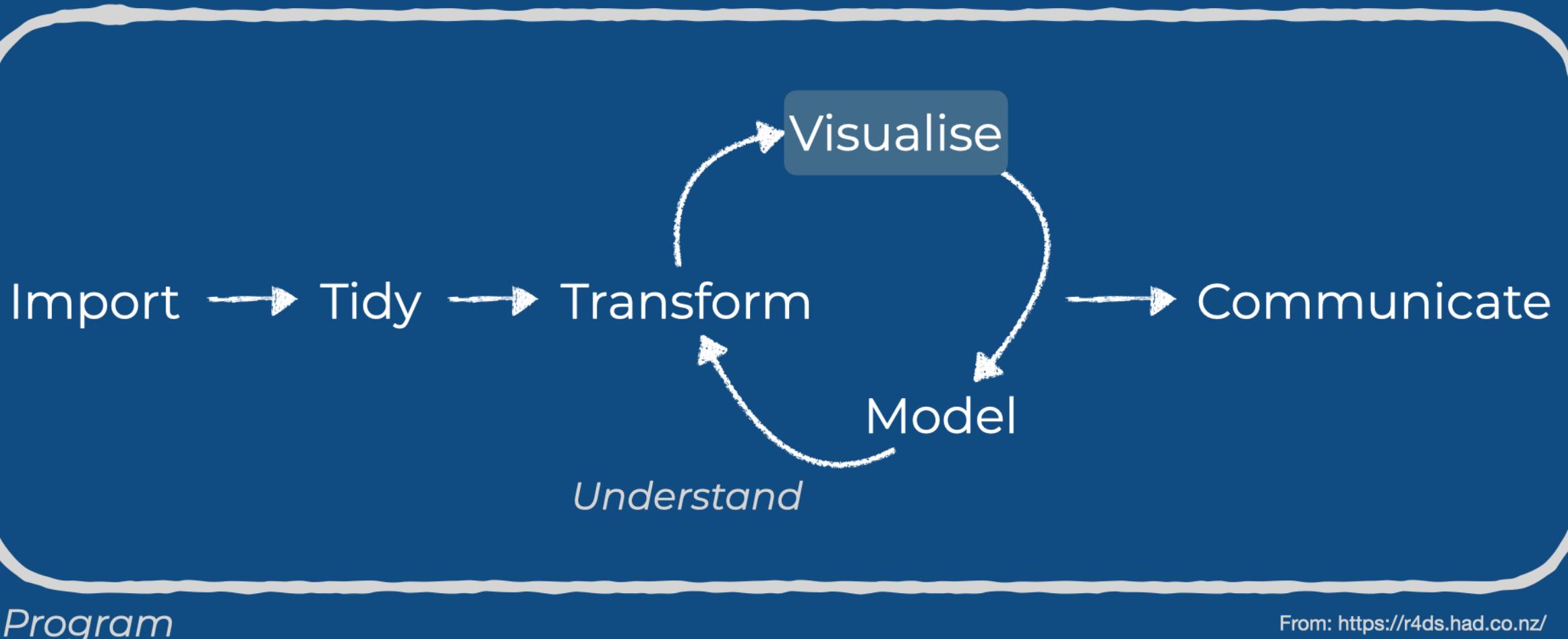


Program

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

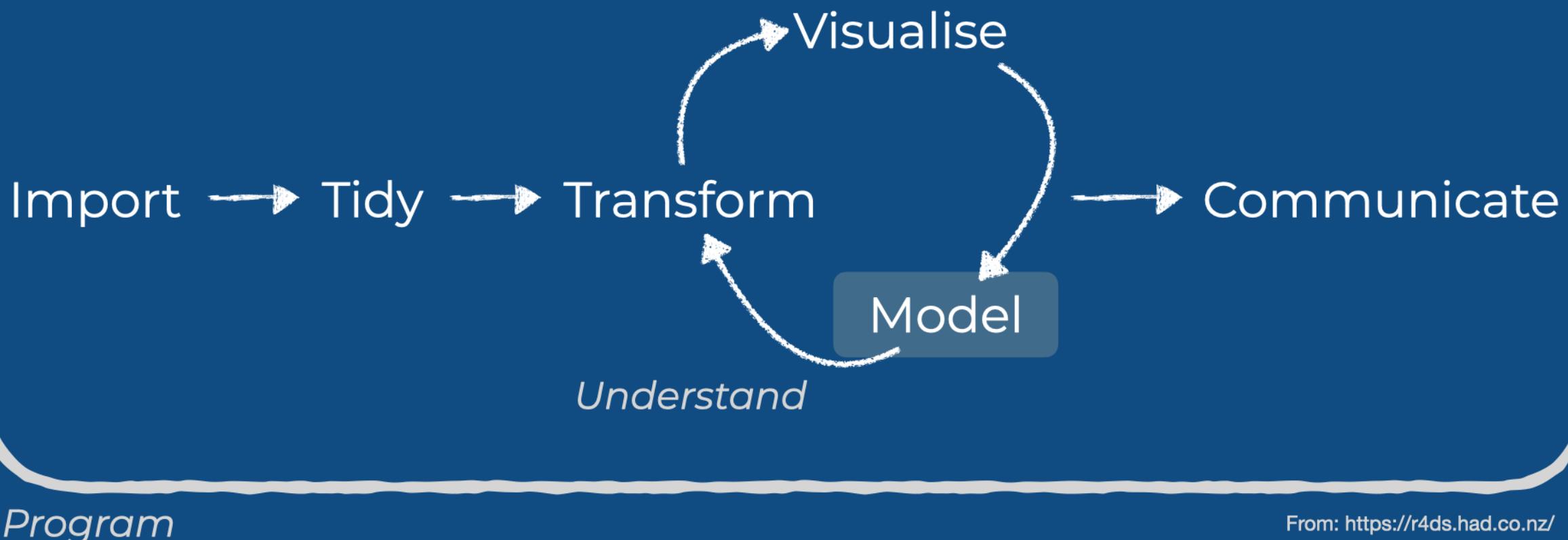
Data Science Lifecycle

Explore your with visual representations



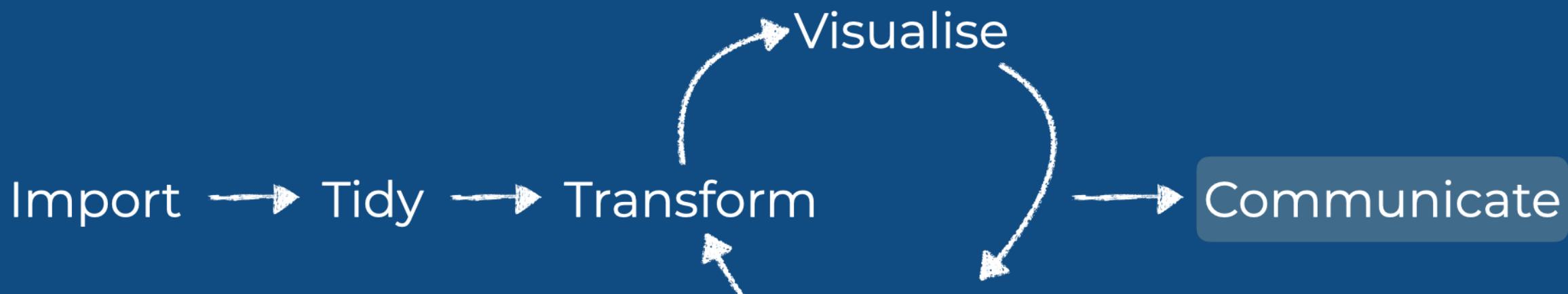
Data Science Lifecycle

Explore your with visual representations



Data Science Lifecycle

Share your findings with others



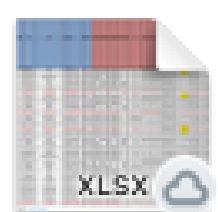
My data analysis projects

what it used to look like

< > raw data



FAQ data.csv



FAQ data.xlsx



FAQ_Q_test.csv



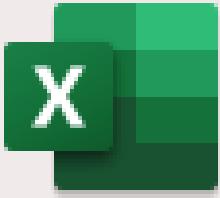
FAQ_Q.csv



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	FS accu
1	Solid waste production	Water usage	Excreta Production (1)	FS/WW production (2)	FS accumulation (3a) (mean)	FS accumulation (3a) (Q1)	FS accu
2	kg/cap*d	L/cap*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 133% ... RAM Settings Help Lars Schöbitz

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

File Edit Code View Plots Session Build Debug Profile Tools Help

Addins

exercise-01.Rmd

ABC Knit

1 title: "My first R Markdown report"
2 author: "Add your name here"
3 output: html_document
4
5 edit
6 ch
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

Code Editor

Environment History Connections Git Tutorial

Import Dataset

R Global Environment

Data

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

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")
>
```

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 133% RAM Lars Schöbitz

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

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

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 Settings Help Logout

Code Editor

The screenshot shows the RStudio Cloud interface with several highlighted sections:

- Code Editor (Pink Box):** The 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 "Run". A large pink box covers the entire Code Editor area.
- Environment (Blue Box):** The top-right pane shows the RStudio environment. It includes tabs for Environment, History, Connections, Git, and Tutorial. Under the Data tab, two datasets are listed: "washdata" (27171 obs. of 11 variables) and "washdata_uga". A large blue box covers the Environment and Data sections.
- Files (Grey Box):** The bottom-right pane shows the project's file structure. It includes tabs for Files, Plots, Packages, Help, and Viewer. The "Files" tab is active, showing a list of files and their details. A grey box covers the Files section.
- Console (Bottom Left):** The bottom-left pane shows the RStudio console output, which includes R code and its execution results.

Code Editor Content (exercise-01.Rmd):

```
1 ---  
2 title: "My first R Markdown report"  
3 author: "Add your name here"  
4 output: html_document  
5 edit  
6 ch  
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 C Chunk 2
```

Environment Data:

- washdata: 27171 obs. of 11 variables
- washdata_uga

Files Content:

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 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 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", "exercise-01.html", and "exercise-02.Rmd".

```
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 724.8 KB Mar 10, 2021, 11:34 AM  
exercise-01.Rmd 2 KB Mar 10, 2021, 11:24 AM  
exercise-01.html 724.8 KB Mar 10, 2021, 11:34 AM
```
- Viewer** (Bottom Right, green border): A placeholder panel with a large green background and white text.

File Manager
Viewer

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

- Code Editor** (Top Left): 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 edit  
6 ch  
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 C Chunk 2
```
- Environment** (Top Right): 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): A file browser showing the project 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 Center): A preview pane showing the rendered content of the R Markdown 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 and its output, including variable definitions and a chain of operations involving "washdata" and "washdata_uga".
- File Manager/Viewer** (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. ~~Registration for Colorado WASH Symposium 2021~~
2. A free account on RStudio Cloud
 - <https://rstudio.cloud/plans/free>
3. One of Mozilla Firefox, Google Chrome, Microsoft Edge, Safari
 - just **not** the Internet Explorer
4. A laptop or desktop computer
 - it will be hard to do on a phone or tablet

Screen setup

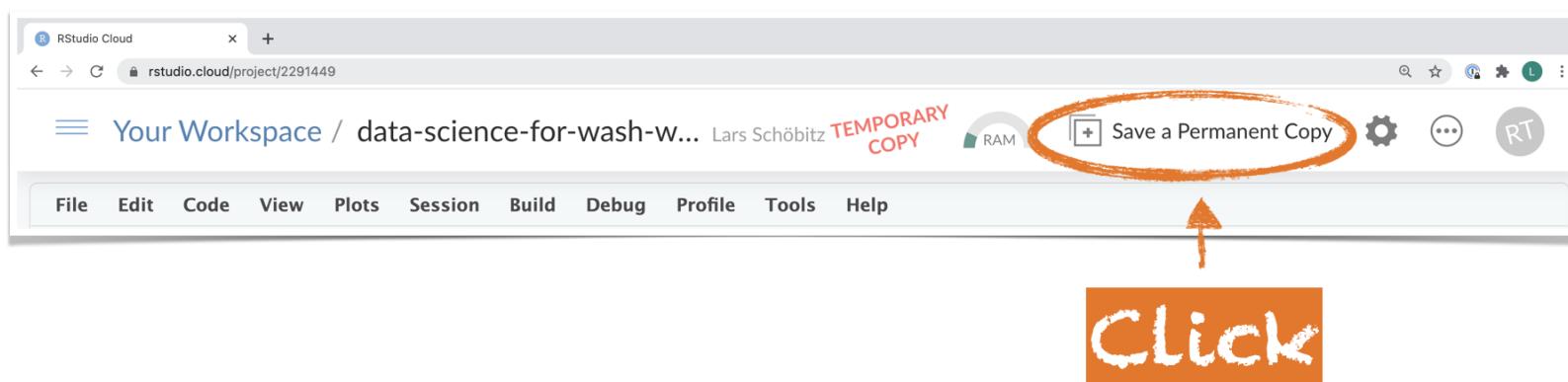
Add screenshot that shows setup on one small screen

Your Turn

Step 1: Open this link in your browser

kutt.it/cowash-ws

Step 2: Create your own permanent project copy

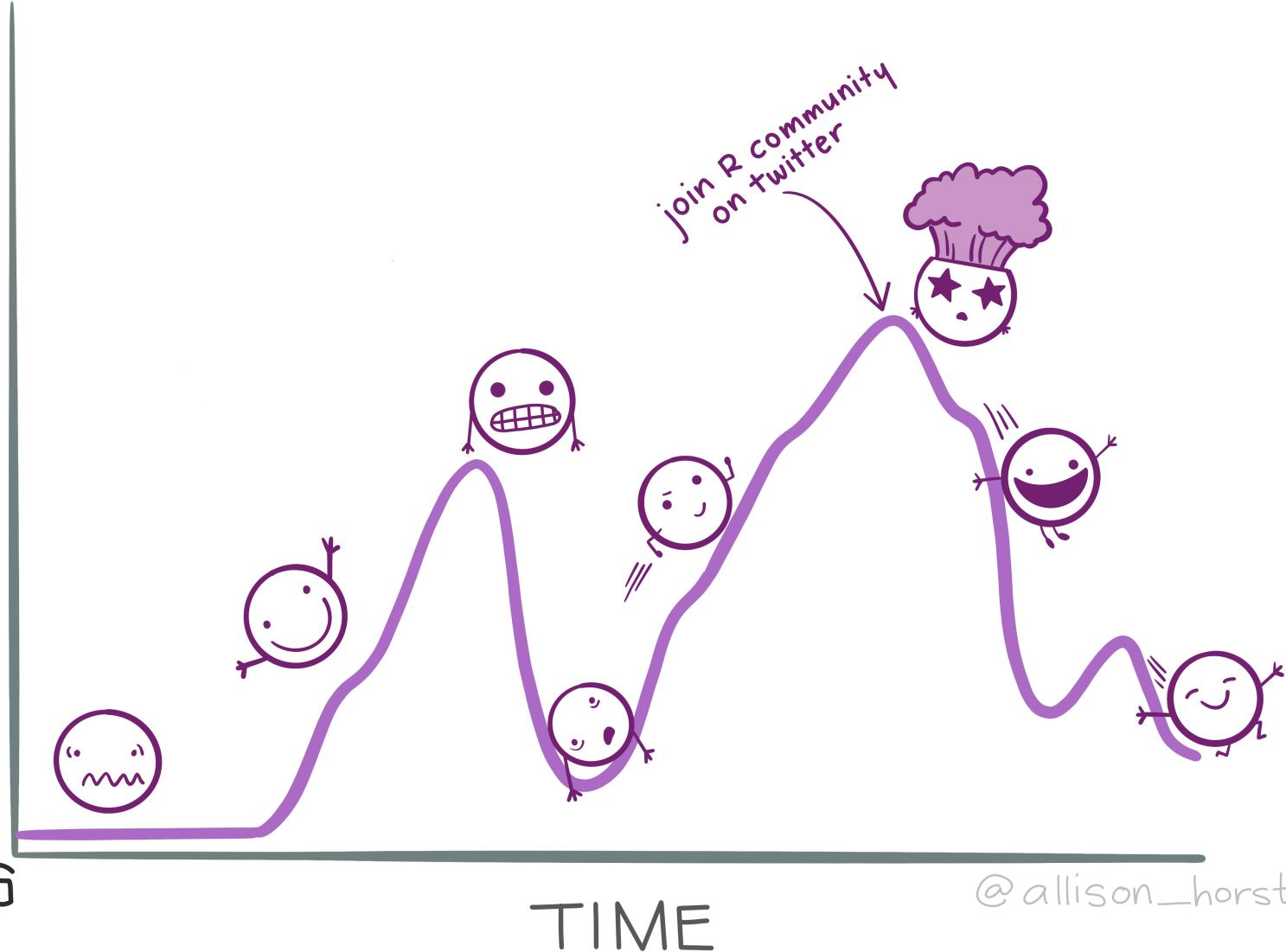


What's next?

HOW
MUCH
I THINK
I KNOW
ABOUT R

I KNOW -
NOTHING

I KNOW -
LOTS!



@allison_horst

If you could not follow through the exercises for any reason

- Sign up for another round of this workshop

TODO: Schedule a Zoom Meeting that people can sign up for

If you want to continue learning online

- Twitter: #rstats
- Online learning community @R4DScommunity
- Minority R Users @miR_community
- We are @R-Ladies
- RStudio Community



If you want a complete introduction to R Markdown

Tutorial by Danielle Navarro

- Slides: <https://slides.djnavarro.net/startling-rmarkdown>
- YouTube videos: <https://youtube.djnavarro.net/startling-rmarkdown>



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

- Slack: kutt.it/washdata-slack
- Twitter: [@washdata](https://twitter.com/washdata)
- E-Mail: Fill out this form [Add link to form](#)

If you have unanswered questions or want to leave a comment (anonymously)

kutt.it/cowash-gd

Survey form

We want to learn a bit more about you

- Add Google Form Link



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](#). To download a PDF version of these slides [click here](#).