

rstatsZH - Data Science mit R

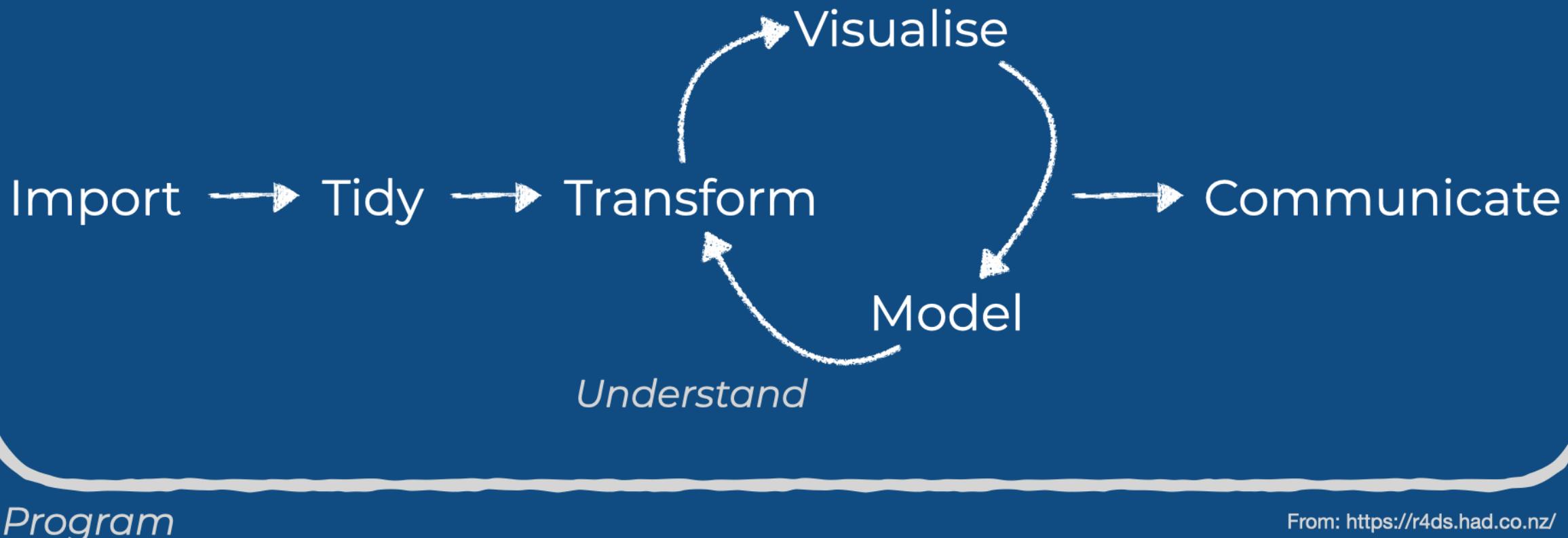
Daten Transformation mit dplyr

Lars Schöbitz

2021-11-08

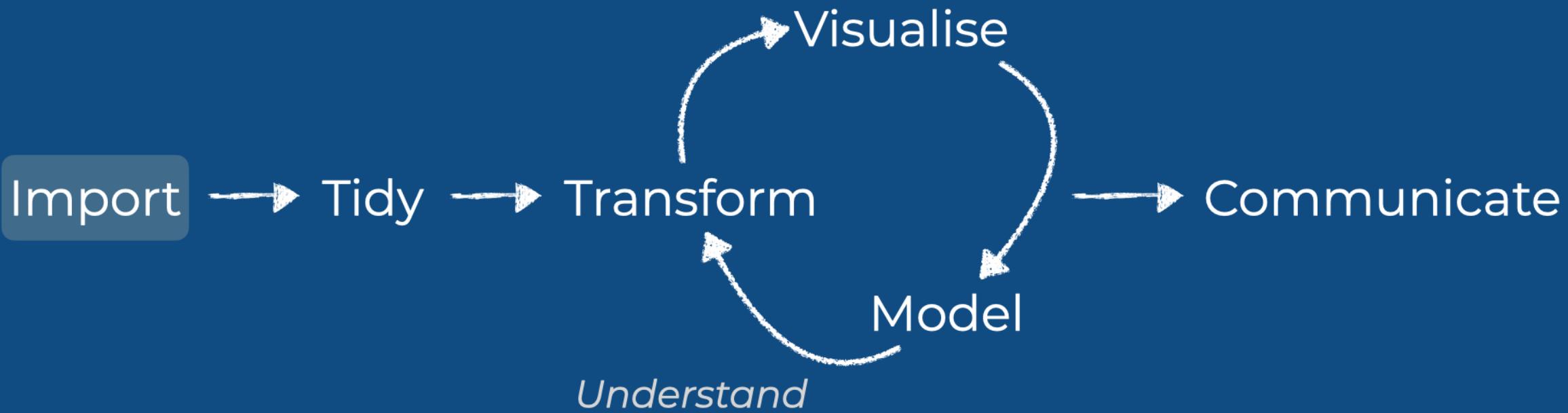
Data Science Lifecycle

Data Science Lifecycle



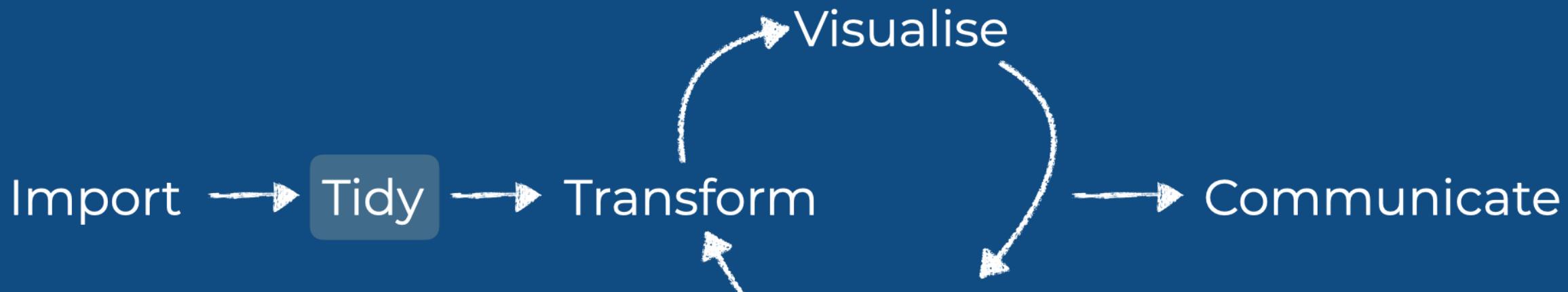
Data Science Lifecycle

Importiere deine Daten



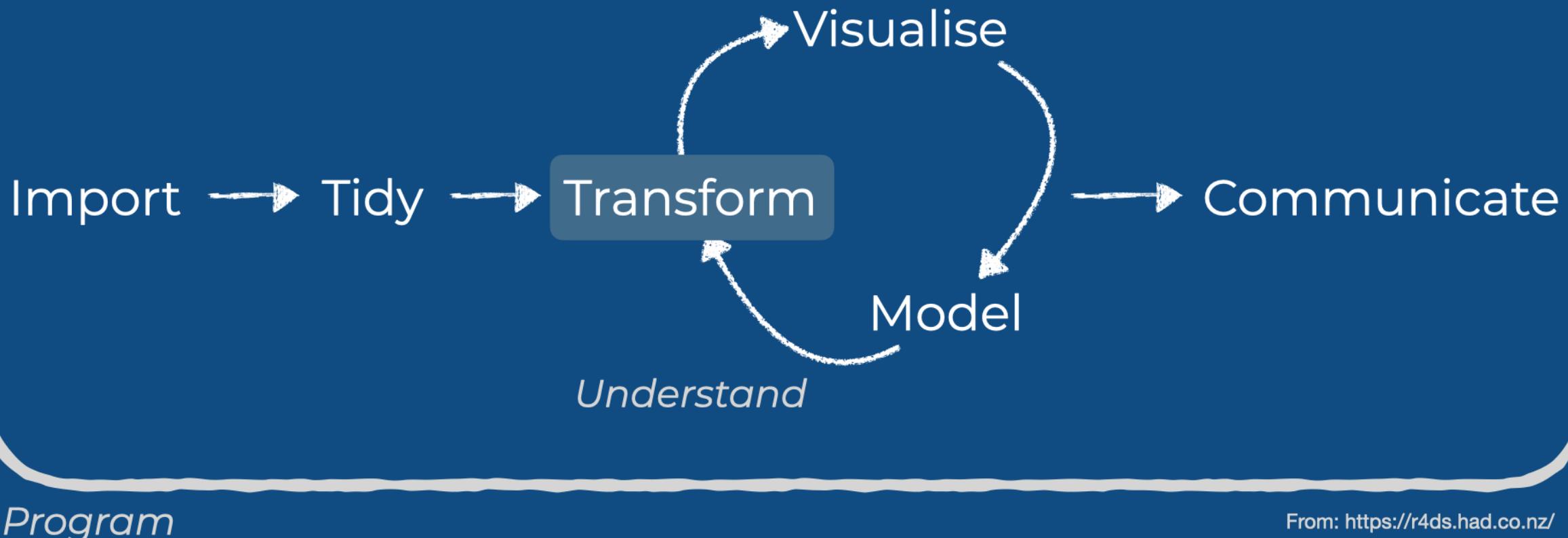
Data Science Lifecycle

Bringe deine Daten in ein konsistentes Format



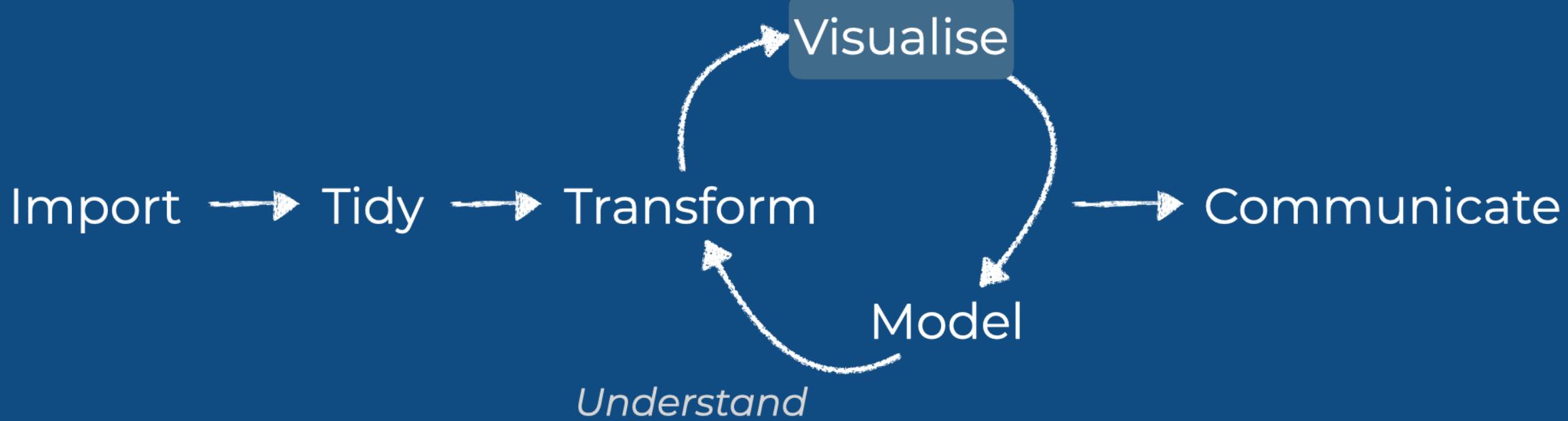
Data Science Lifecycle

Grenze ein + Erstelle neue Variablen + Fasse zusammen



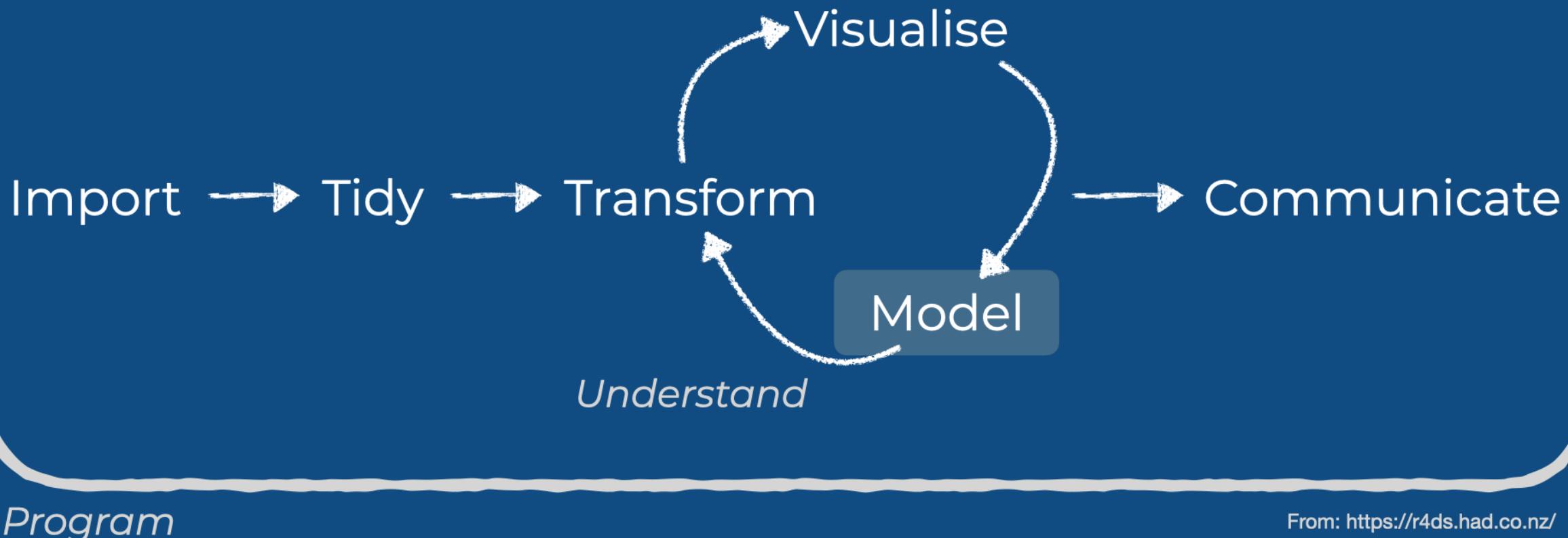
Data Science Lifecycle

Erkunde Daten mit Visualisierungen



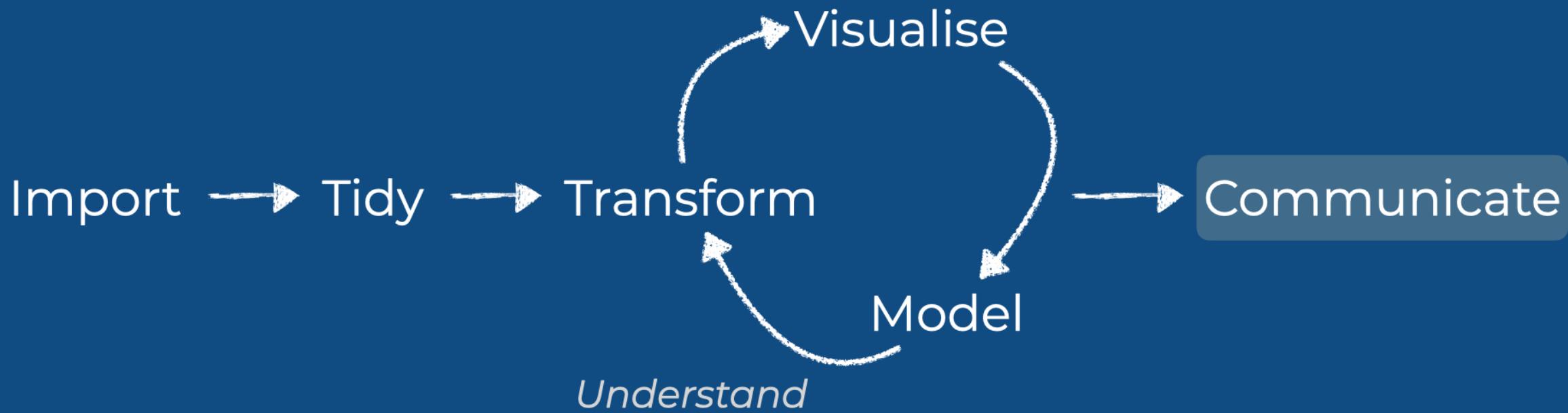
Data Science Lifecycle

Ergänze Visualisierungen mit Modellen



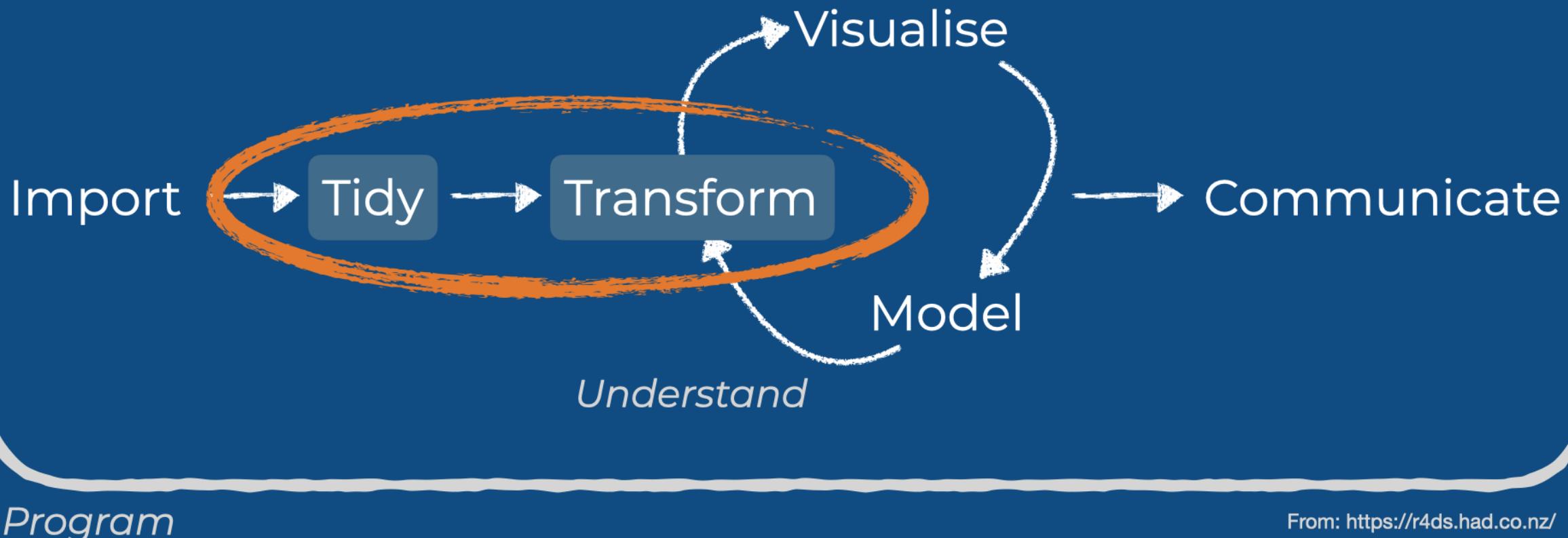
Data Science Lifecycle

Teile deine Erkenntnisse mit Anderen



Data Science Lifecycle

Data Wrangling - Kämpfe mit deinen Daten



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

Was bedeutet Tidy Data?

“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its **structure**. ”

—HADLEY WICKHAM

In tidy data:

- each **variable** forms a **column**
- each **observation** forms a **row**
- each **cell** is a **single measurement**

each column a variable

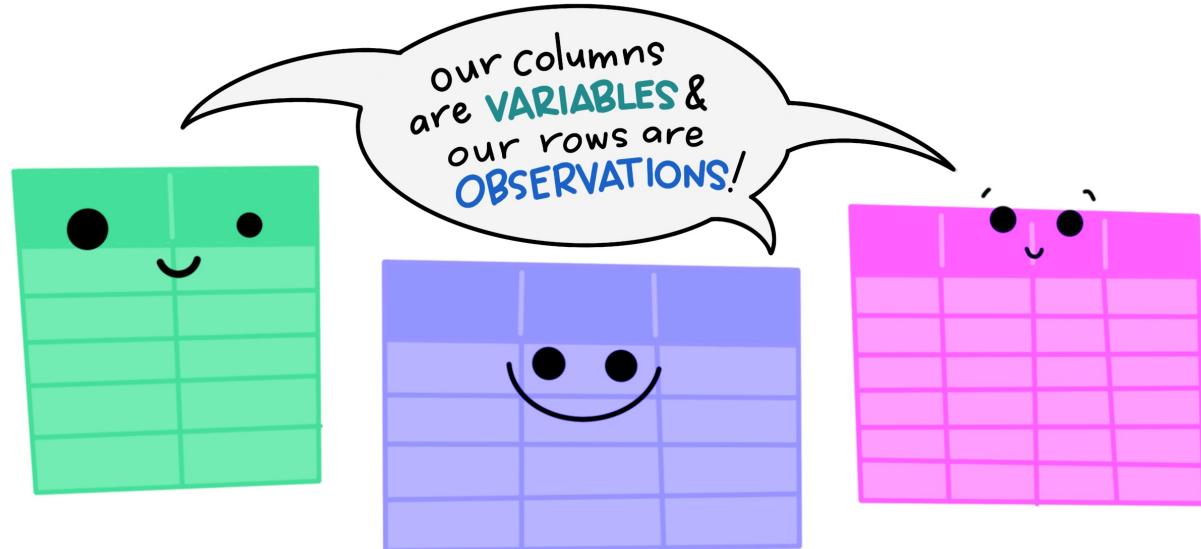
each row an observation

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

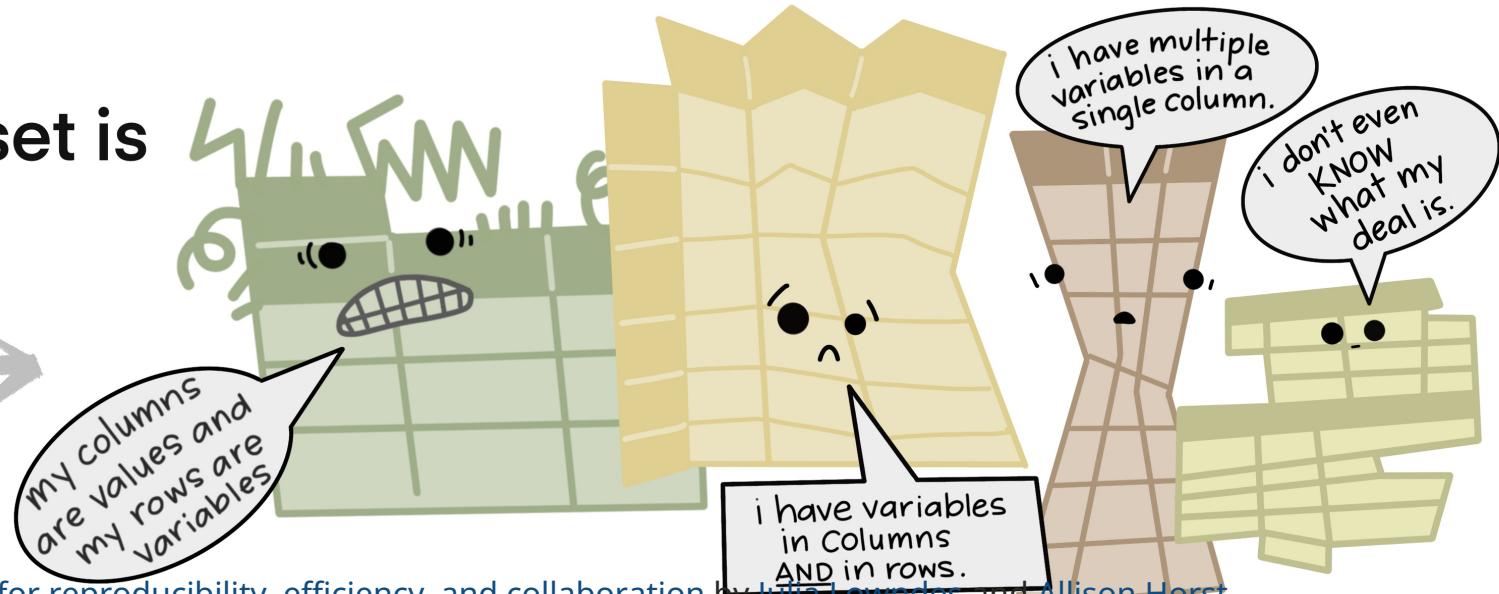
Illustrations from the [Openscapes](#) blog [Tidy Data for reproducibility, efficiency, and collaboration](#) by [Julia Lowndes](#) and [Allison Horst](#)

The standard structure of
tidy data means that
“tidy datasets are all alike...”

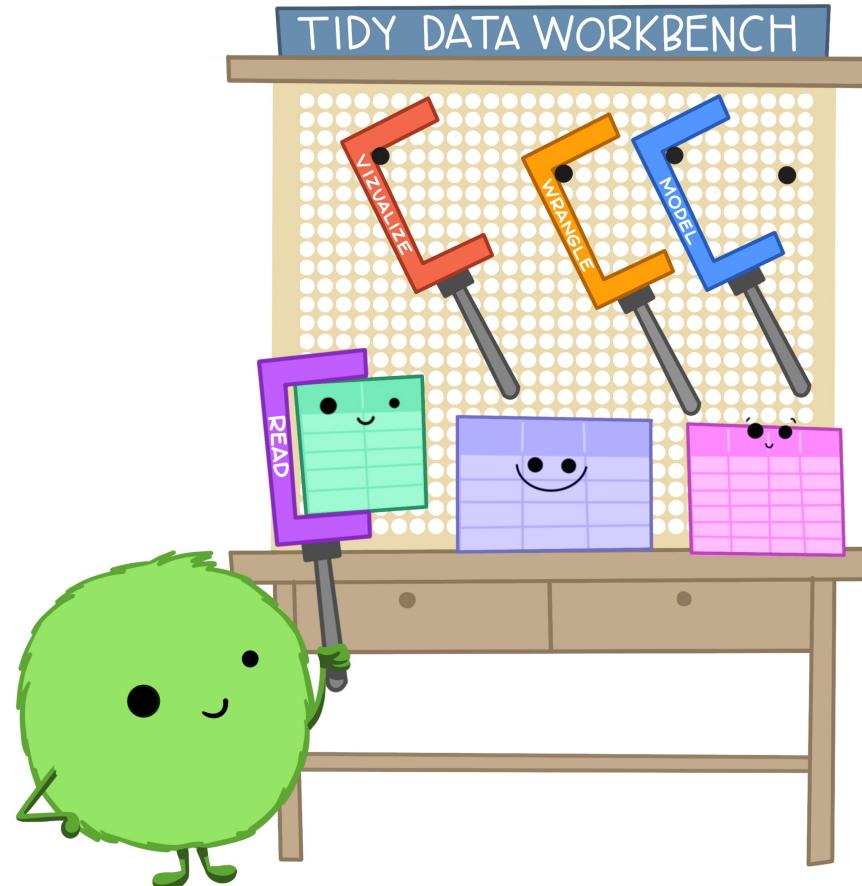


“...but every messy dataset is
messy in its own way.”

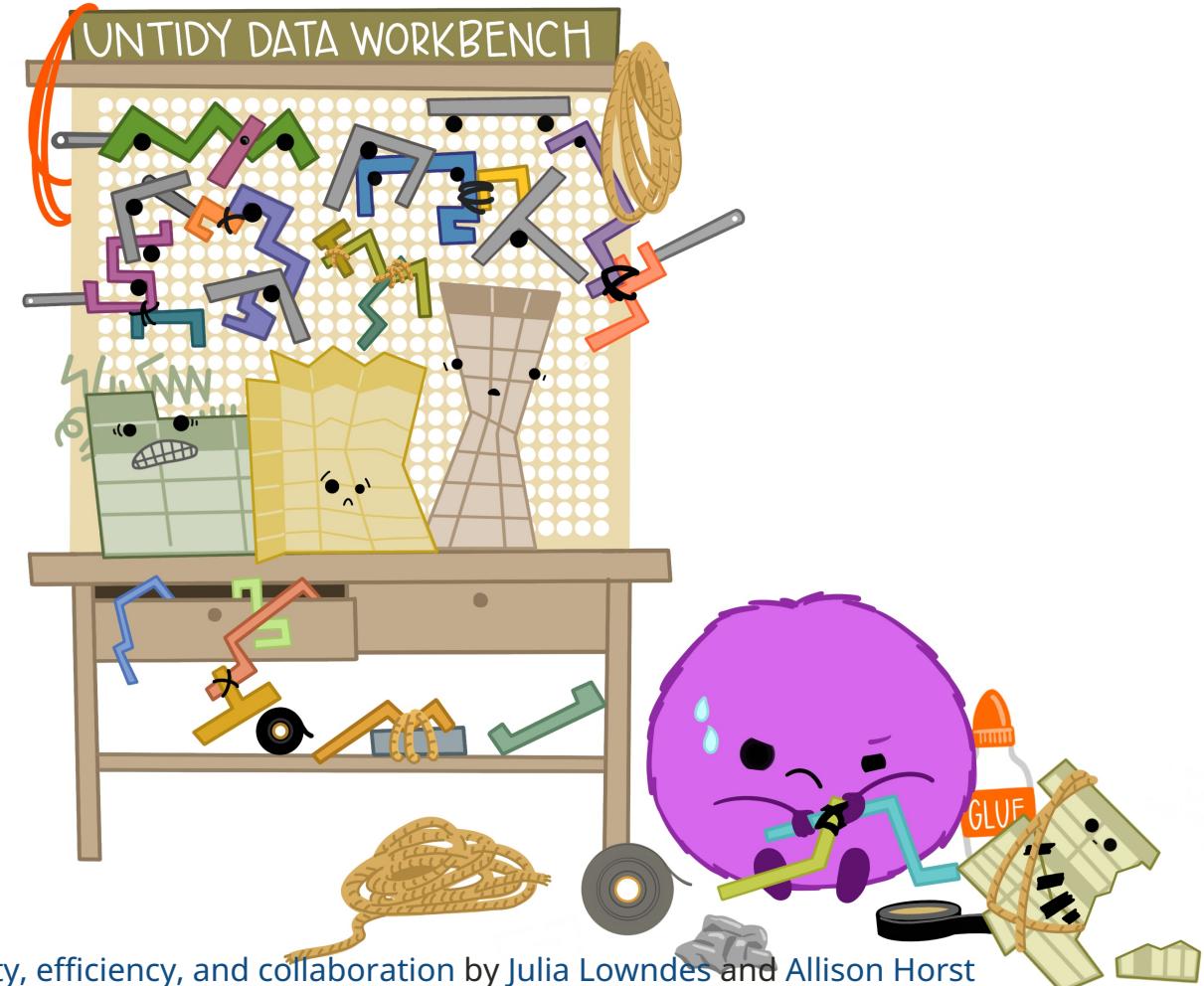
—HADLEY WICKHAM

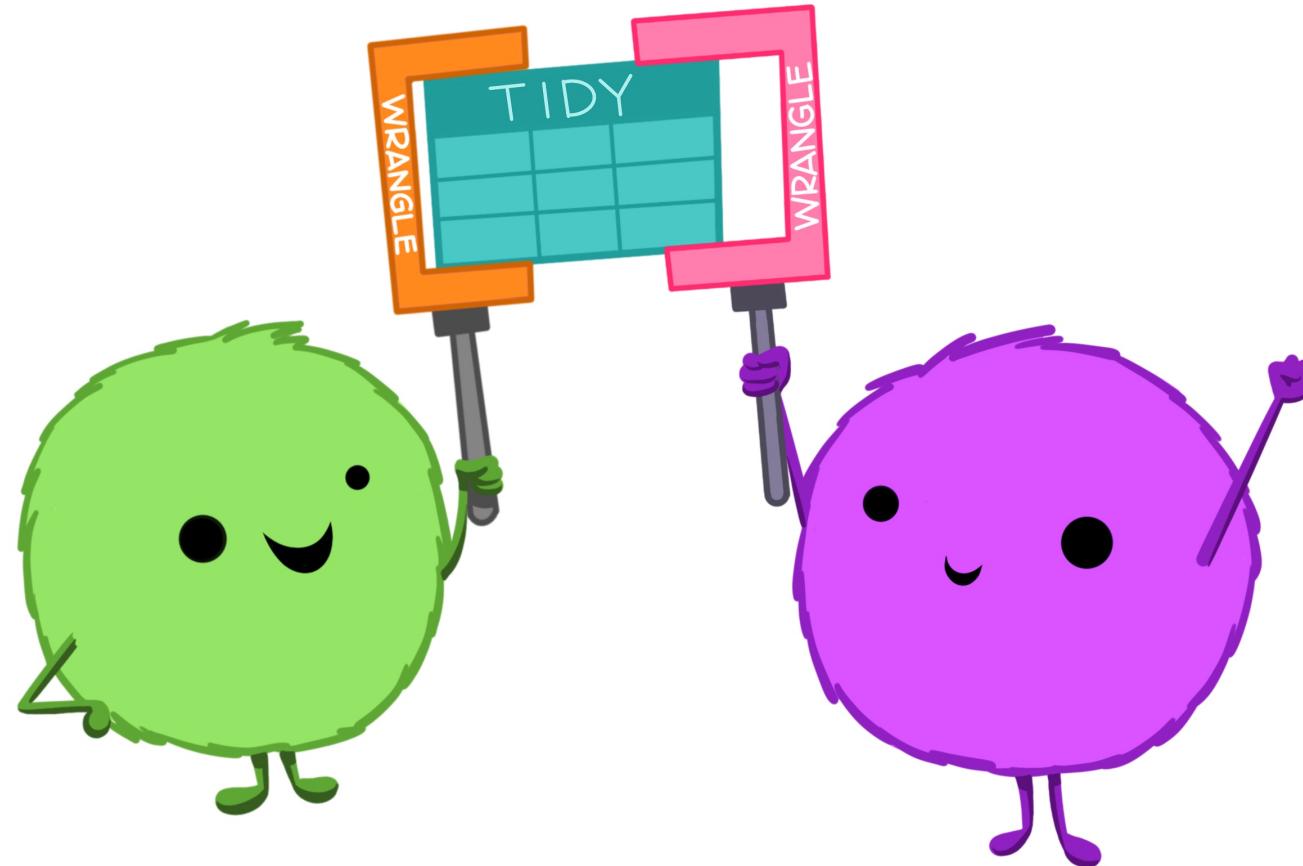


When working with tidy data,
we can use the **same tools** in
similar ways for different datasets...

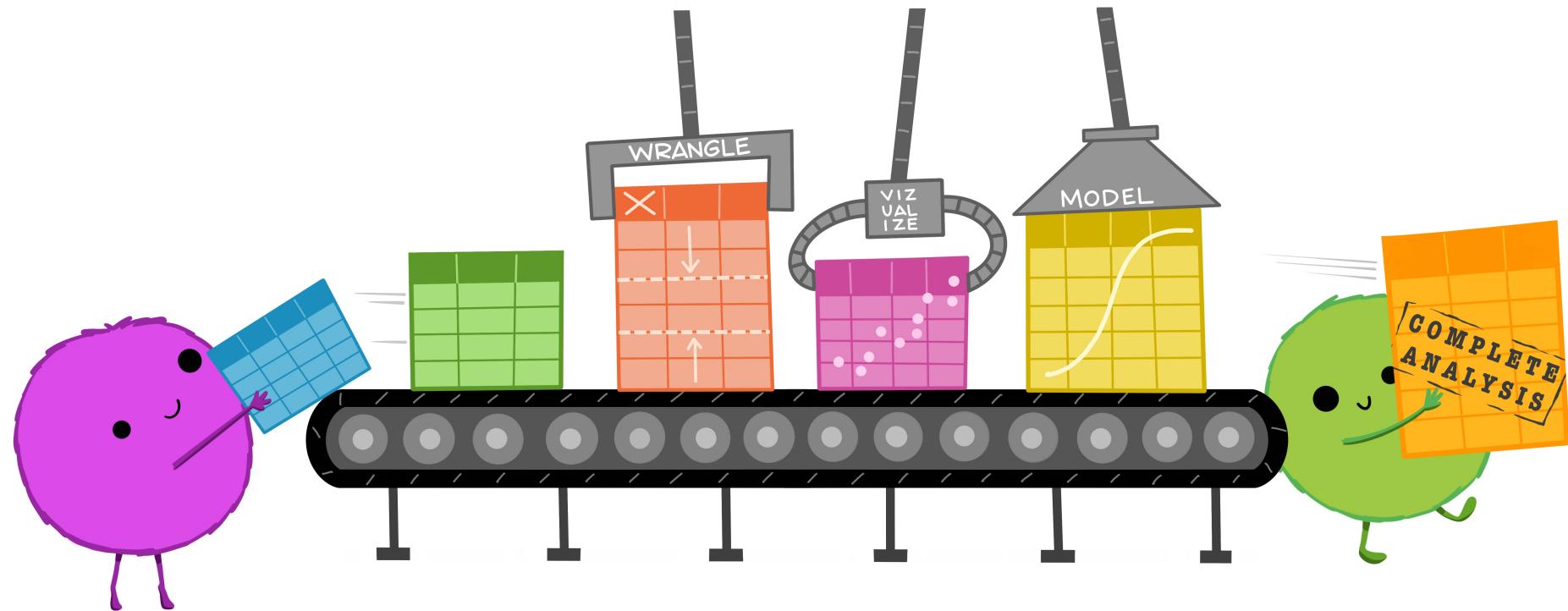


...but working with untidy data often means
reinventing the wheel with **one-time**
approaches that are hard to iterate or reuse.

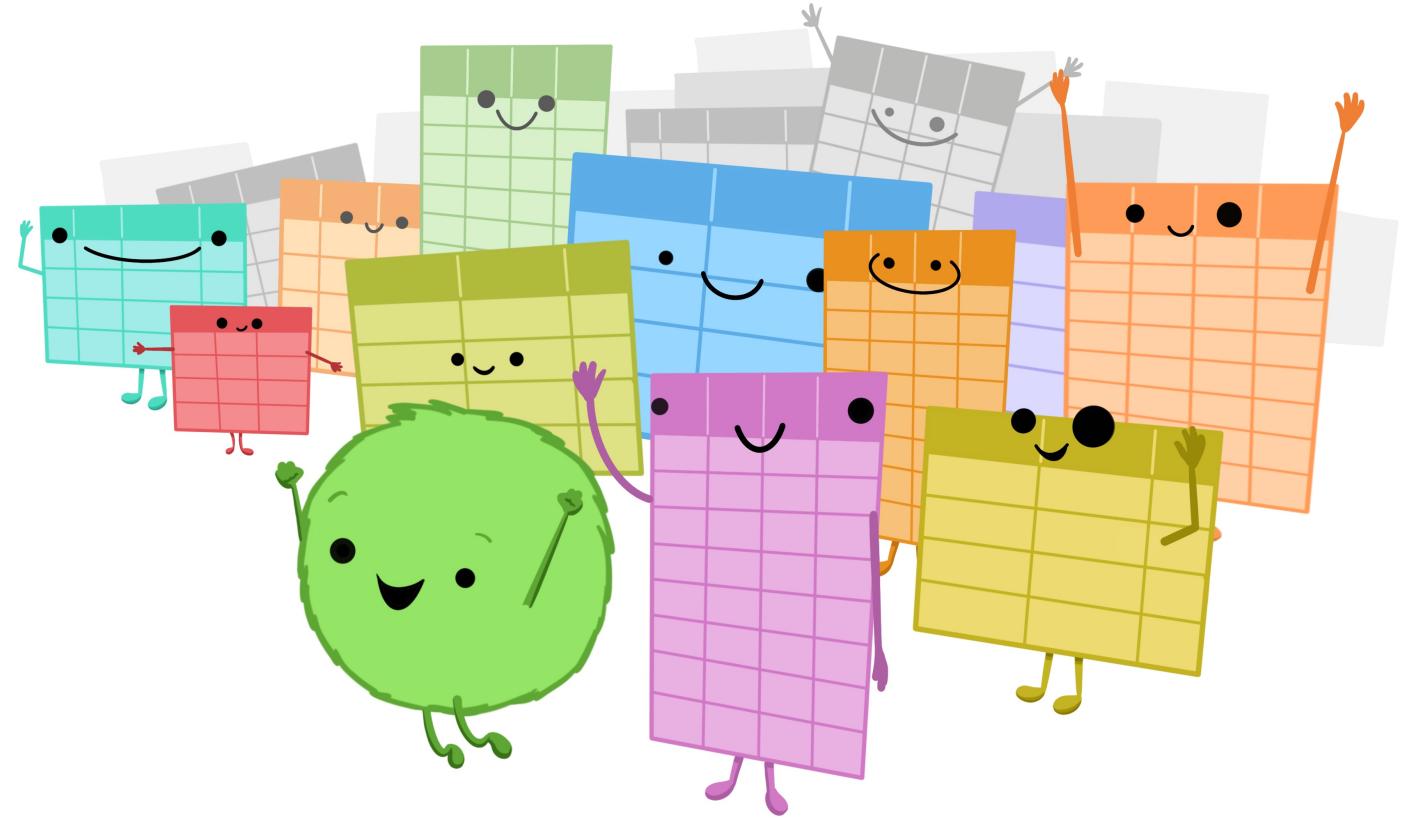
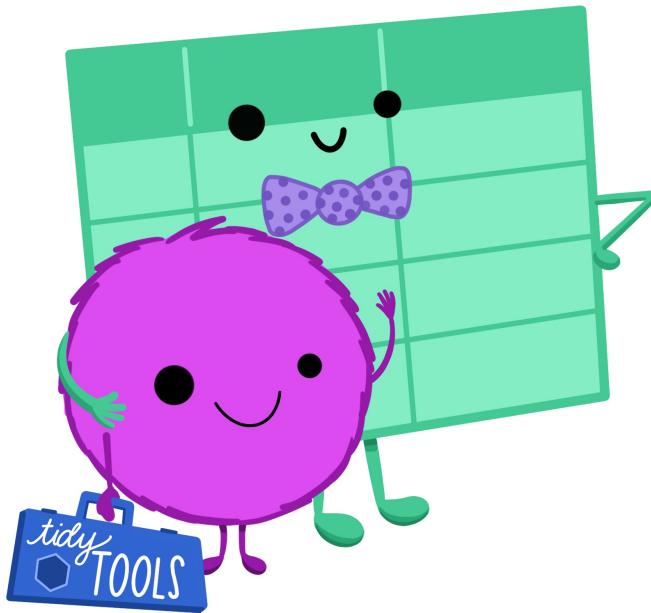




Illustrations from the Openscapes blog [Tidy Data for reproducibility, efficiency, and collaboration](#) by [Julia Lowndes](#) and [Allison Horst](#)

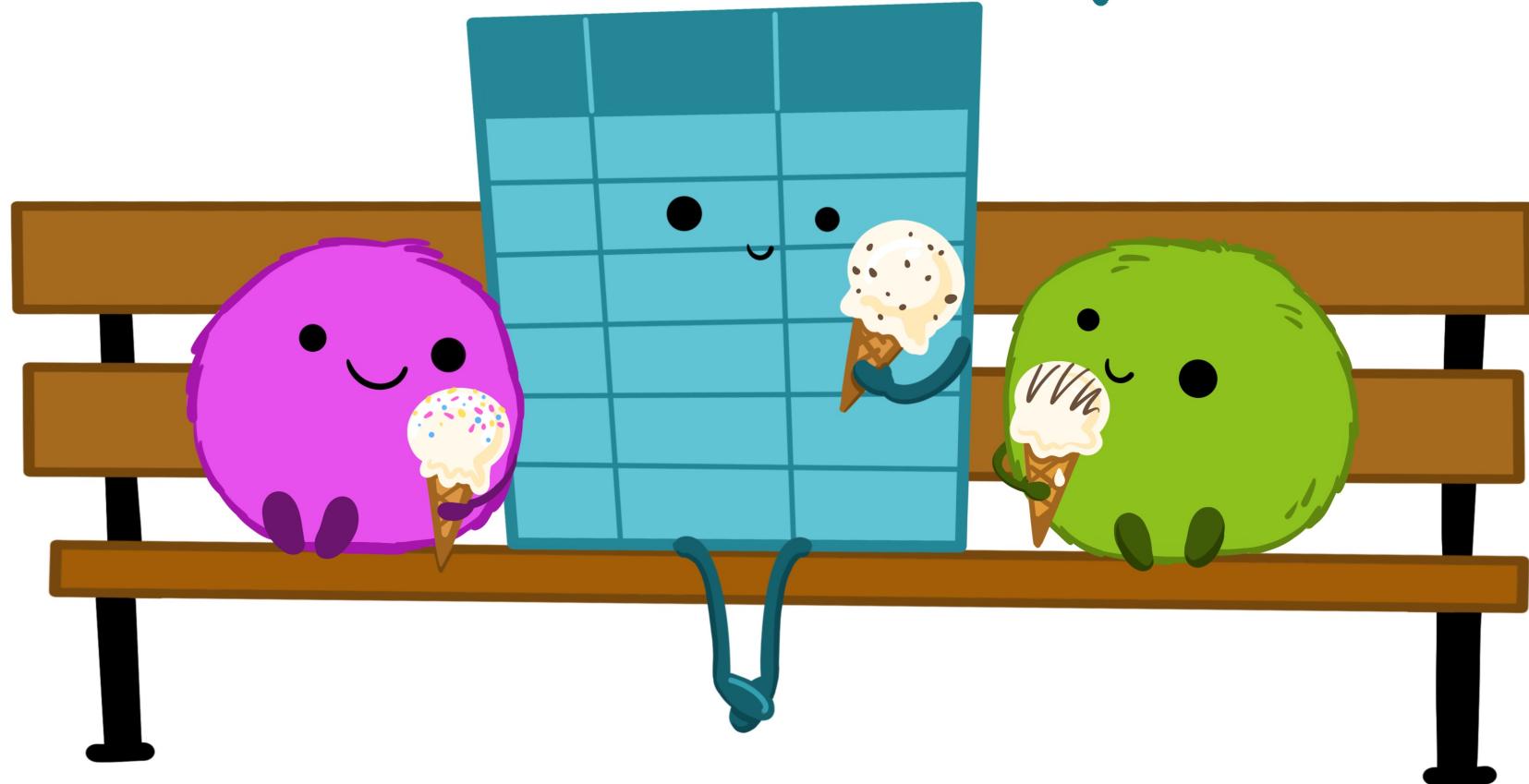


Illustrations from the [Openscapes](#) blog [Tidy Data for reproducibility, efficiency, and collaboration](#) by [Julia Lowndes](#) and [Allison Horst](#)



Illustrations from the Openscapes blog [Tidy Data for reproducibility, efficiency, and collaboration](#) by [Julia Lowndes](#) and [Allison Horst](#)

make friends with tidy data.



Eigenschaften von Tidy Data

Eigenschaften von Tidy Data:

- Eigenschaft 1: Jede Spalte ist eine Variable
- Eigenschaft 2: Jede Reihe ist eine Beobachtung
- Eigenschaft 3: Jede Zelle enthält eine Messung

Penguins

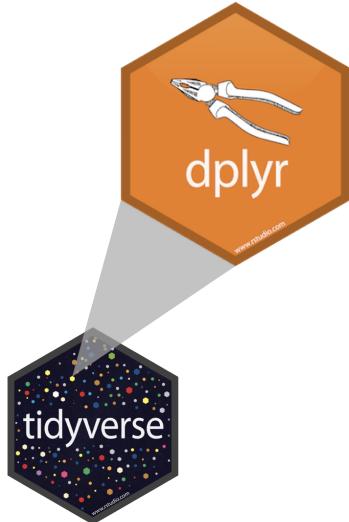
Erfüllen die Daten die Eingeschafften für Tidy data?

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
Adelie	Torgersen	42.0	20.2	190	4250	NA	2007

Daten Transformation mit {dplyr}

Eine Grammatik für Daten Transformation...

...basierend auf dem Konzept von Funktionen als Verben, die Daten manipulieren



- `select`: Spalten beim Namen auswählen
- `arrange`: Zeilen neu ordnen
- `slice`: Zeilen über einen Index auswählen
- `filter`: Zeilen mit bestimmten Kriterien auswählen
- `mutate`: Neue Variablen hinzufügen
- `summarise`: Variablen auf Werte reduzieren
- `group_by`: Gruppierte Arbeitsgänge
- ... (und viele viele mehr)

Regeln für `{dplyr}` Funktionen

1. Das erste Argument ist immer ein Dataframe (genau wie bei ggplot)
2. Die folgenden Argumente geben an was mit dem Dataframe gemacht werden soll
3. Geben immer einen Dataframe zurück
4. Ändern nichts an Ort und Stelle

```
filter(.data = penguins, species == "Adelie")
```

	# A tibble: 152 × 8	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
		<fct>	<fct>	<dbl>	<dbl>	<int>	
1	Adelie	Torgersen		39.1	18.7	181	
2	Adelie	Torgersen		39.5	17.4	186	
3	Adelie	Torgersen		40.3	18	195	
4	Adelie	Torgersen		NA	NA	NA	
5	Adelie	Torgersen		36.7	19.3	193	
6	Adelie	Torgersen		39.3	20.6	190	

Praktikum 5 - dplyr

Praktikum 5 - dplyr

Live Code

1. **E-Mail:** Öffne deine Email und klicke auf den Link zu deinem persönlichen GitHub repo für **prak-05**
2. **GitHub:** Klicke auf den grünen Button "Code" und kopiere den Link für das Repo in deine Zwischenablage
3. **RStudio Cloud:** Öffne deinen Arbeitsbereich für den Kurs in der RStudio Cloud
4. **RStudio Cloud / Projects:** Klicke auf "New Project from GitHub Repository"
5. **RStudio:** Finde den Datei Manager und den Git Reiter
6. **Zoom Chat:** Schreibt wenn ihr soweit seid

Praktikum 5 - Lösungen

- **GitHub Organisation:** rstatsZH
 - github.com/rstatsZH
- **Repo:** prak-05-wrangle-dplyr
 - <https://github.com/rstatsZH/prak-05-wrangle-dplyr>
- **R Markdown Datei:** prak-05-solutions.Rmd
 - <https://github.com/rstatsZH/prak-05-wrangle-dplyr/blob/main/prak-05-solutions.Rmd>

Praktikum 6 - dplyr

Praktikum 6 - dplyr

In 2er Teams

1. **E-Mail:** Öffne deine Email und klicke auf den Link zu deinem persönlichen GitHub repo für **prak-06**
2. **GitHub:** Klicke auf den grünen Button "Code" und kopiere den Link für das Repo in deine Zwischenablage
3. **RStudio Cloud:** Öffne deinen Arbeitsbereich für den Kurs in der RStudio Cloud
4. **RStudio Cloud / Projects:** Klicke auf "New Project from GitHub Repository"
5. **RStudio:** Finde den Datei Manager und Git
6. **Zoom Chat:** Schreibt wenn ihr soweit seid

tidyverse.org

- Lesezeichen
 - ggplot2.tidyverse.org
 - dplyr.tidyverse.org
 - readr.tidyverse.org
 - ...



Feedback

Ziele erreicht?

Bitte ausfüllen: kutt.it/rstatszh-eval



Hausaufgabe

Hausaufgabe

- Öffnet jetzt eure E-Mail Inbox
- Bestätigt mir im Zoom Chat, dass ihr eine Email mit Betreff "**rstatsZH - Lars hat das Repo ha02-GitHubName**" erhalten habt
- Die Anweisungen für die Heausaufgabe 02 findet ihr oben rechts auf unserer Kurswebseite
- Kontaktiert mich unter der Woche jederzeit auf Slack



Danke

Für die Aufmerksamkeit!

Für die R packages `{xaringan}` und `{xaringanthemer}` mit welchen die Folien geschrieben wurden.

Eine PDF Version der Folien kann hier heruntergeladen werden:

https://github.com/rstatsZH/website/raw/master/slides/e1_d04-data-transform/e1_d04-data-transform.pdf

Für Data Science in a Box und Remaster the Tidyverse, von welchen ich Materialien für diesen Kurs nutze und welche genau wie diese Folien mit Creative Commons Attribution Share Alike 4.0 International lizenziert sind.