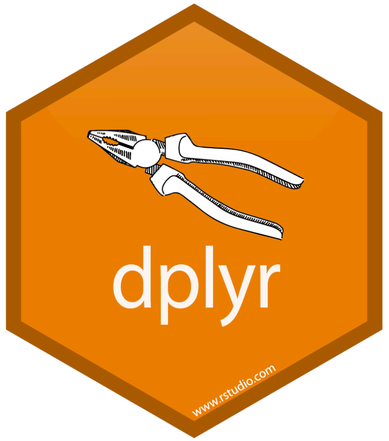
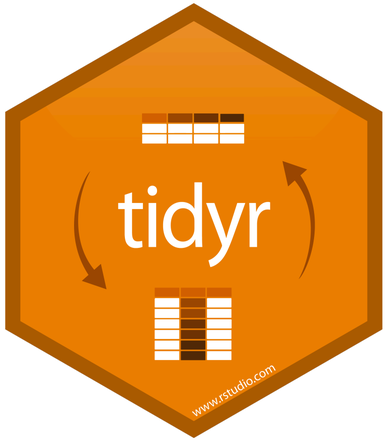
Programme for today

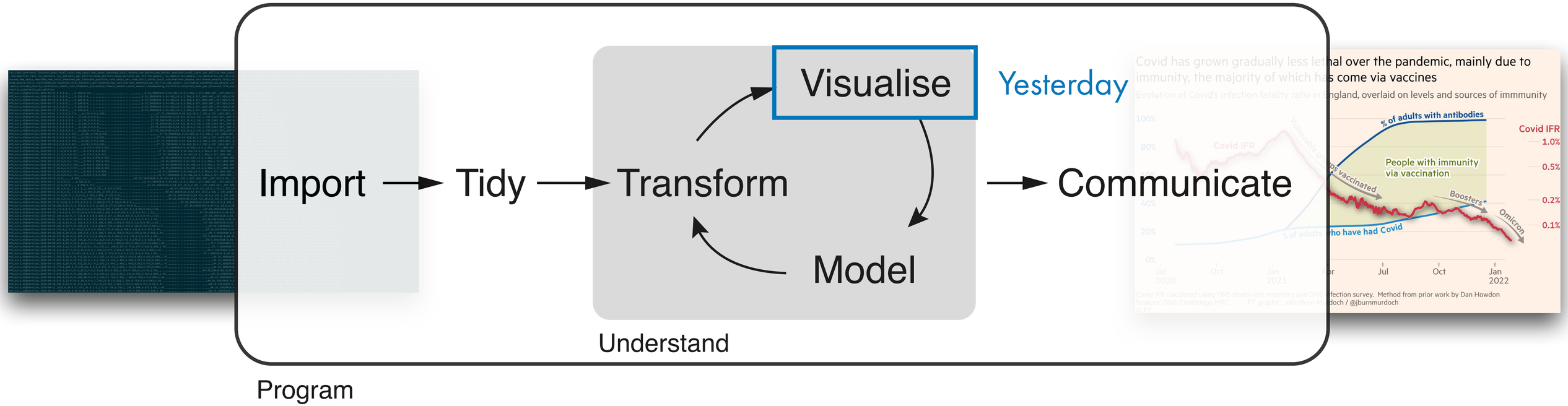
- Introductory lecture
- Hands-on tutorial on data wrangling
- Individual exercises





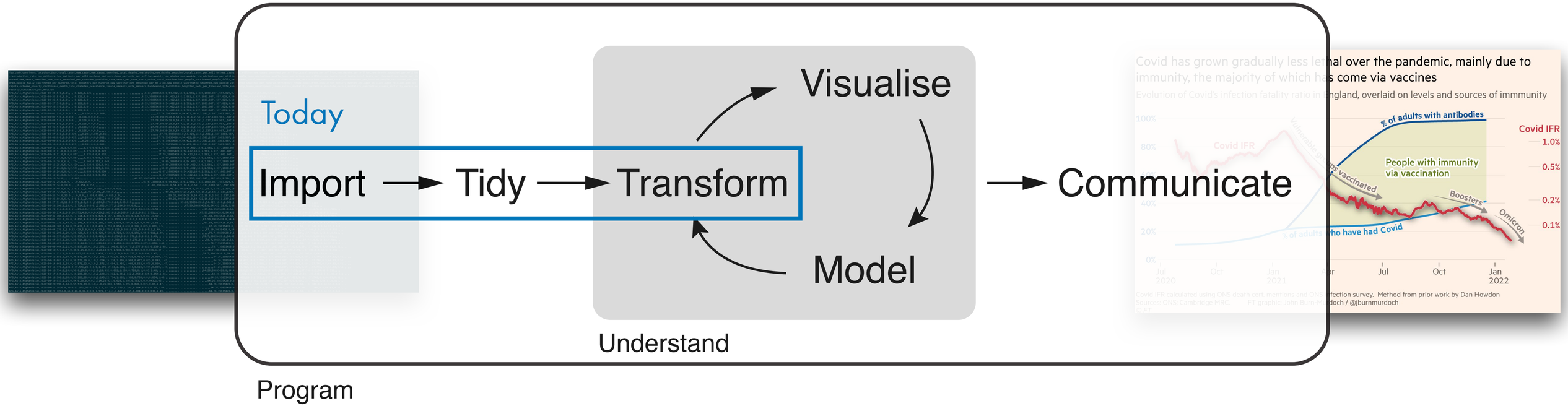
At the end of the day you should be familiar with transforming and tidying data in R using the dplyr and tidy packages

A data science project workflow



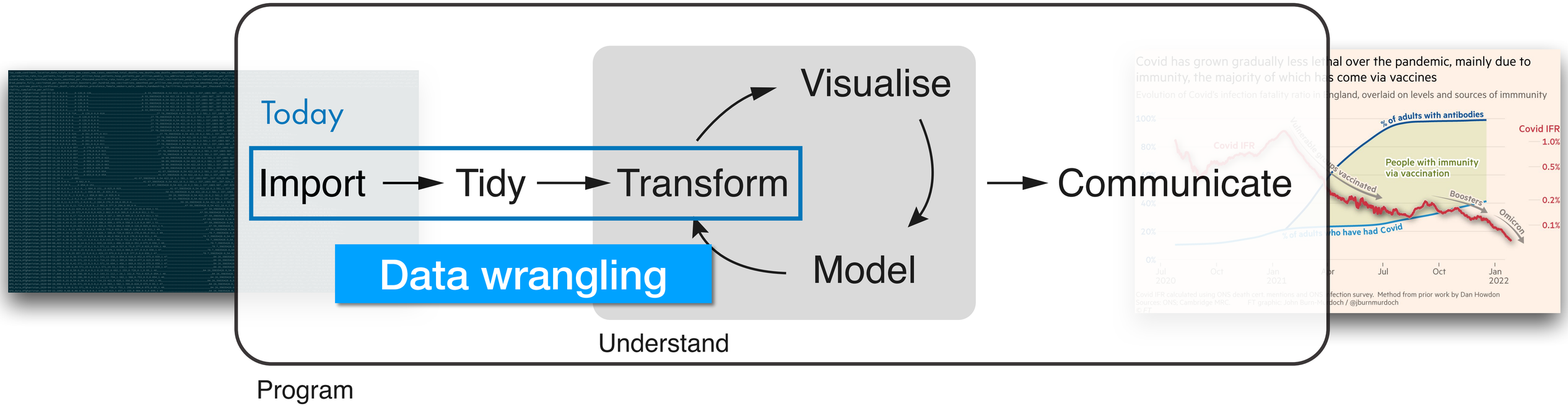
https://r4ds.had.co.nz/introduction.html

A data science project workflow



https://r4ds.had.co.nz/introduction.html

A data science project workflow



https://r4ds.had.co.nz/introduction.html

When data is messy ...

species	habitat	weight	length	latitude/longitude	date
Alligator mississippiensis	swamp	431 lb	4 ft 2	29.531,-82.184	Sept 15, 2015
Puma concolor	forest	125 lb	2.2m	29.125,-81.682	08/10/2015
Ursus americanus	forest	88 kg	133 cm	N29°7'30"/W81°40'55.2"	07-13-2015

In which ways is this dataset "untidy"?

Hart et al (2016) PLoS Comp Biol (https://doi.org/10.1371/journal.pcbi.1005097)

When data is messy … we tidy it up

species	habitat	weight	length	latitude/longitude	date
Alligator mississippiensis	swamp	431 lb	4 ft 2	29.531,-82.184	Sept 15, 2015
Puma concolor	forest	125 lb	2.2m	29.125,-81.682	08/10/2015
Ursus americanus	forest	88 kg	133 cm	N29°7'30"/W81°40'55.2"	07-13-2015

meta-data

data

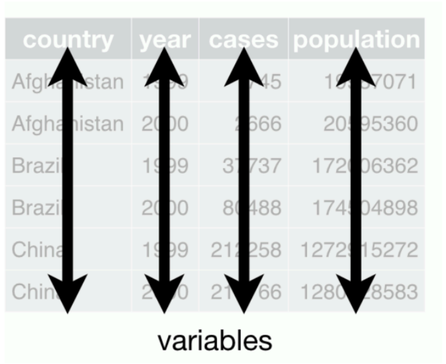
species_code	date	station_code	weight_kg	length_cm
TSN 551771	2015-09-15	1	196	127
TSN 55247	2015-08-10	2	57	220
TSN 180544	2015-07-13	2	88	133

station_code	habitat	latitude	longitude
1	swamp	29.531	-82.184
2	forest	29.125	-81.682

species_code	class	genus	species
TSN 551771	Reptilia	Alligator	mississippiensis
TSN 55247	Mammalia	Puma	concolor
TSN 180544	Mammalia	Ursus	americanus

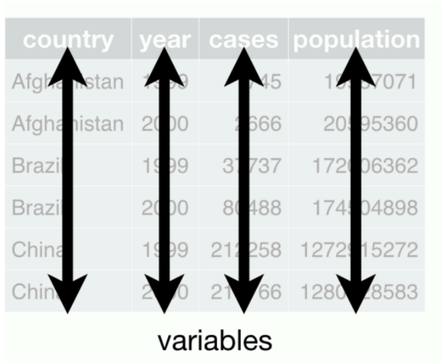
Hart et al (2016) PLoS Comp Biol (https://doi.org/10.1371/journal.pcbi.1005097)

Principles of tidy data

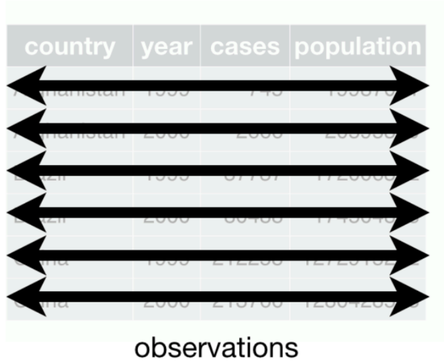


1. Each variable has its own column

Principles of tidy data

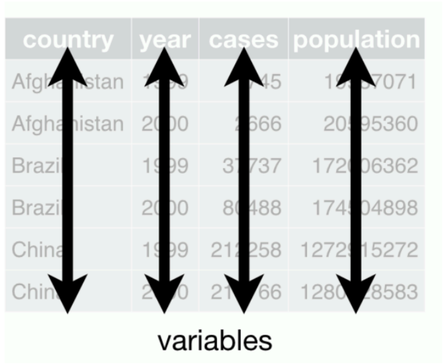


1. Each variable has its own column

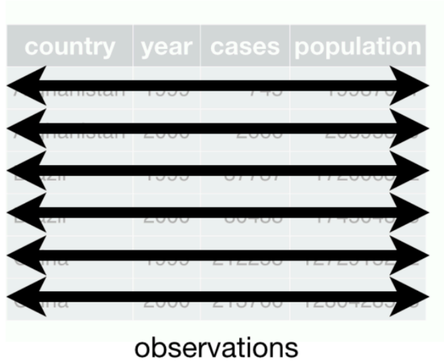


2. Each observation has its own row

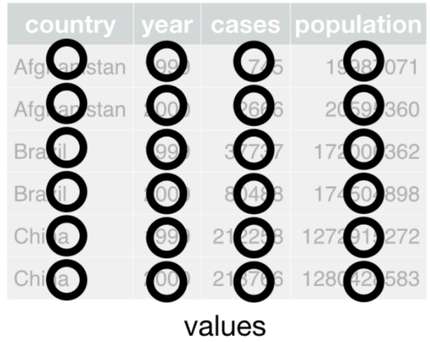
Principles of tidy data



1. Each variable has its own column



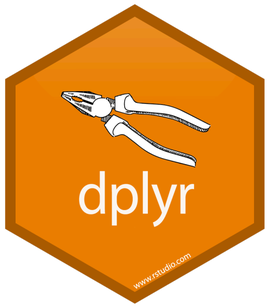
2. Each observation has its own row

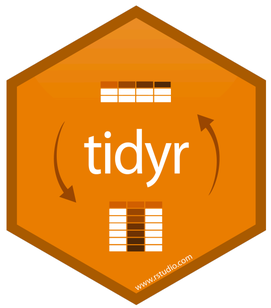


3. Each values has its own cell

Data transformation

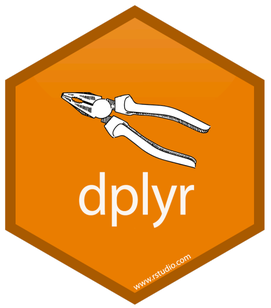
Data tidying





Data transformation

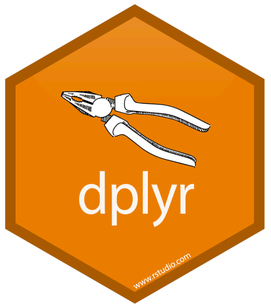
Data tidying



Data transformation

Data tidying

select(.data, …**)** Extract columns as a table.
select(mtcars, mpg, wt)



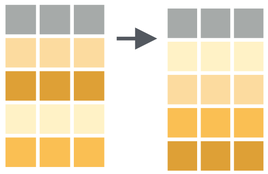
Simple transformations

https://rstudio.github.io/cheatsheets/data-transformation.pdf

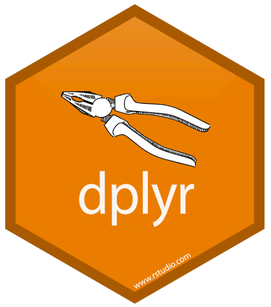
Data transformation

Data tidying

select(.data, …**)** Extract columns as a table.
select(mtcars, mpg, wt)



arrange(.data, …, .by_group = FALSE**)** Order rows by values of a column or columns (low to high), use with **desc()** to order from high to low.
arrange(mtcars, mpg)
arrange(mtcars, desc(mpg))



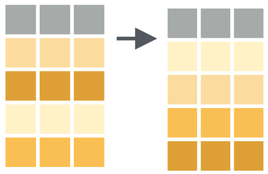
Simple transformations

https://rstudio.github.io/cheatsheets/data-transformation.pdf

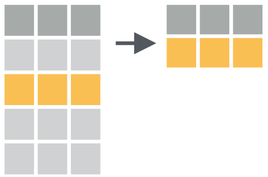
Data transformation

Data tidying

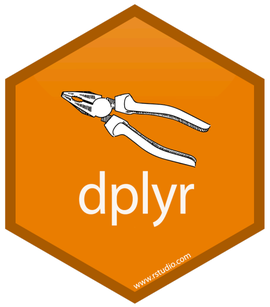
select(.data, …**)** Extract columns as a table.
select(mtcars, mpg, wt)



arrange(.data, …, .by_group = FALSE**)** Order rows by values of a column or columns (low to high), use with **desc()** to order from high to low.
arrange(mtcars, mpg)
arrange(mtcars, desc(mpg))



filter(.data, …, .preserve = FALSE**)** Extract rows that meet logical criteria.
filter(mtcars, mpg > 20)



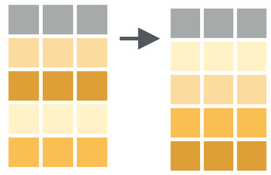
Simple transformations

https://rstudio.github.io/cheatsheets/data-transformation.pdf

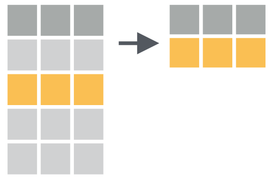
Data transformation

Data tidying

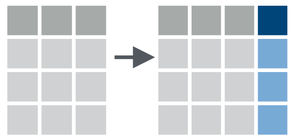
select(.data, …**)** Extract columns as a table.
select(mtcars, mpg, wt)



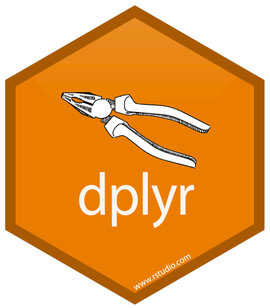
arrange(.data, …, .by_group = FALSE**)** Order rows by values of a column or columns (low to high), use with **desc()** to order from high to low.
arrange(mtcars, mpg)
arrange(mtcars, desc(mpg))



filter(.data, …, .preserve = FALSE**)** Extract rows that meet logical criteria.
filter(mtcars, mpg > 20)



mutate(.data, …, .keep = "all", .before = NULL, .after = NULL**)** Compute new column(s). Also **add_column(), add_count(),** and **add_tally().**
mutate(mtcars, gpm = 1 / mpg)



Simple transformations

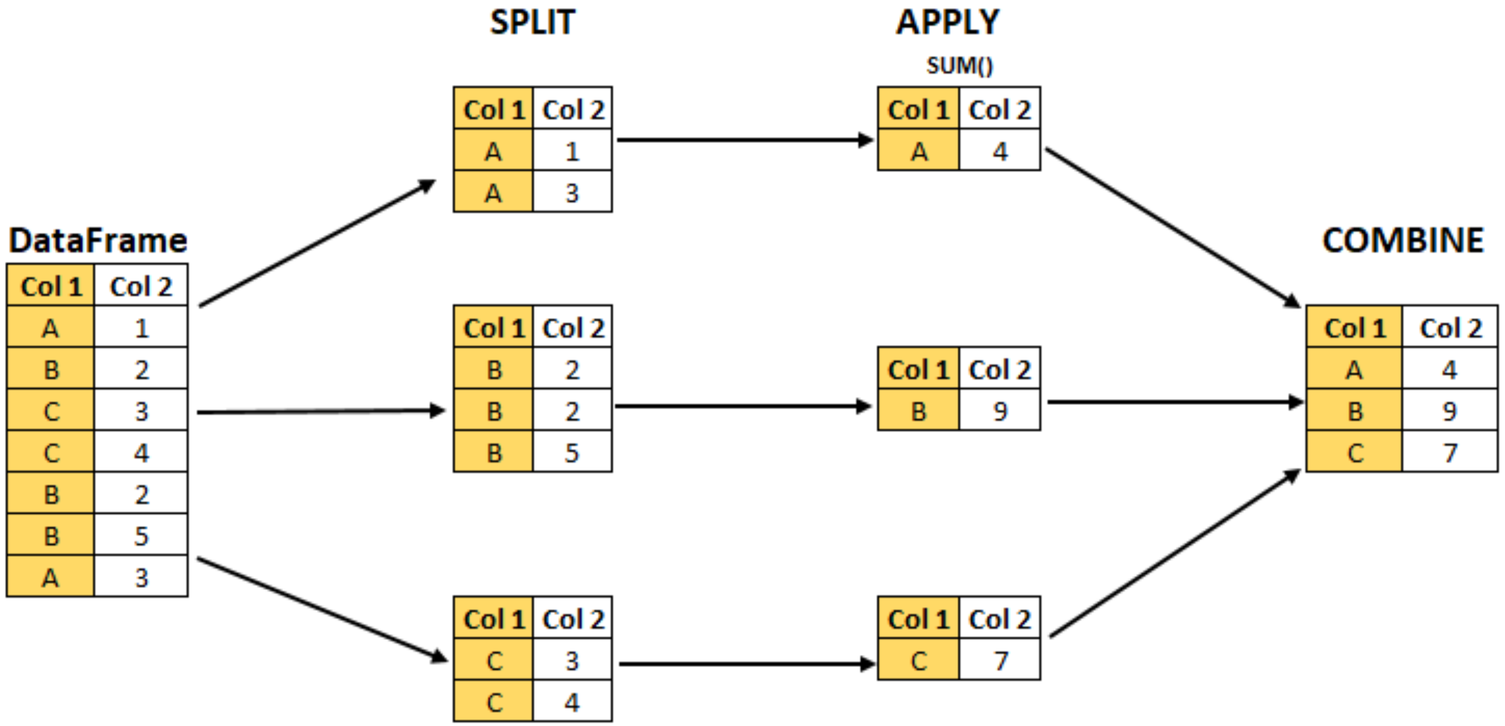
https://rstudio.github.io/cheatsheets/data-transformation.pdf

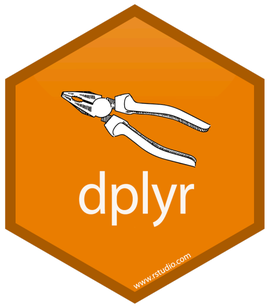
The pipe


```
do_another_thing(do_something(data))

# versus

data %>%
    do_something() %>%
    do_another_thing()
```

Data transformation

Data tidying





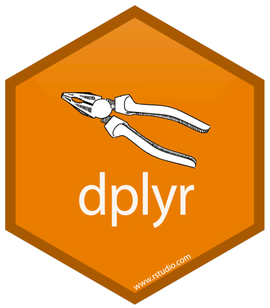
split-apply-combine

https://medium.com/analytics-vidhya/split-apply-combine-strategy-for-data-mining-4fd6e2a0cc99

Data transformation

Data tidying

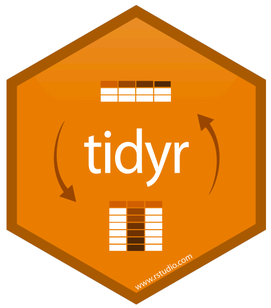
```
mtcars  %>%
  group_by(cyl) %>%
  summarise(avg = mean(mpg))
```



Group and summarise

https://medium.com/analytics-vidhya/split-apply-combine-strategy-for-data-mining-4fd6e2a0cc99

Data tidying

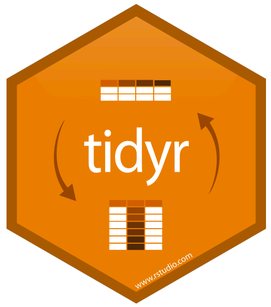


Data tidying

country	1999	2000
A	0.7K	2K
B	37K	80K
C	212K	213K

country	year	cases
A	1999	0.7K
B	1999	37K
C	1999	212K
A	2000	2K
B	2000	80K
C	2000	213K

pivot_longer(data, cols, names_to = "name", values_to = "value", values_drop_na = FALSE**)**

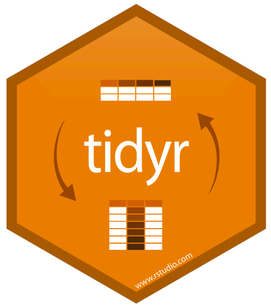


Reshape “wide” data to “long” data

Data tidying

country	year	type	count
A	1999	cases	0.7K
A	1999	pop	19M
A	2000	cases	2K
A	2000	pop	20M
B	1999	cases	37K
B	1999	pop	172M
B	2000	cases	80K
B	2000	pop	174M
C	1999	cases	212K
C	1999	pop	1T
C	2000	cases	213K
C	2000	pop	1T

country	year	cases	pop
A	1999	0.7K	19M
A	2000	2K	20M
B	1999	37K	172M
B	2000	80K	174M
C	1999	212K	1T
C	2000	213K	1T

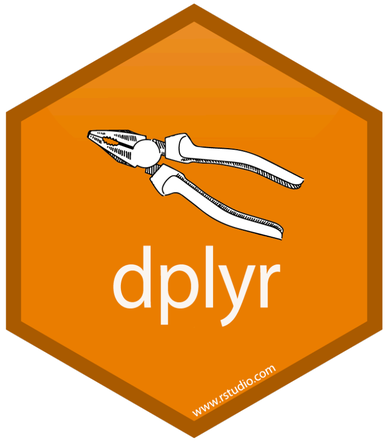


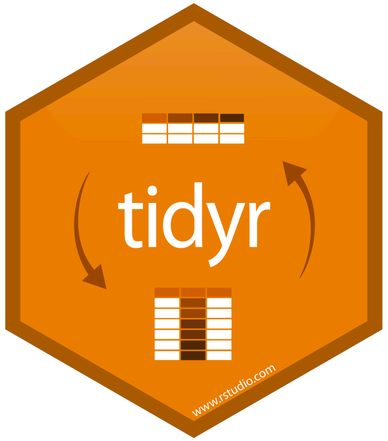
pivot_wider(data, names_from = "name", values_from = "value"**)**

Reshape “long” data to “wide” data

Programme for today

- Introductory lecture
- Hands-on tutorial on data wrangling
- Individual exercises





At the end of the day you should be familiar with transforming and tidying data in R using the dplyr and tidy packages

https://ucph.padlet.org/martinsikora4/data_analysis_2025_wrangle