Relational data

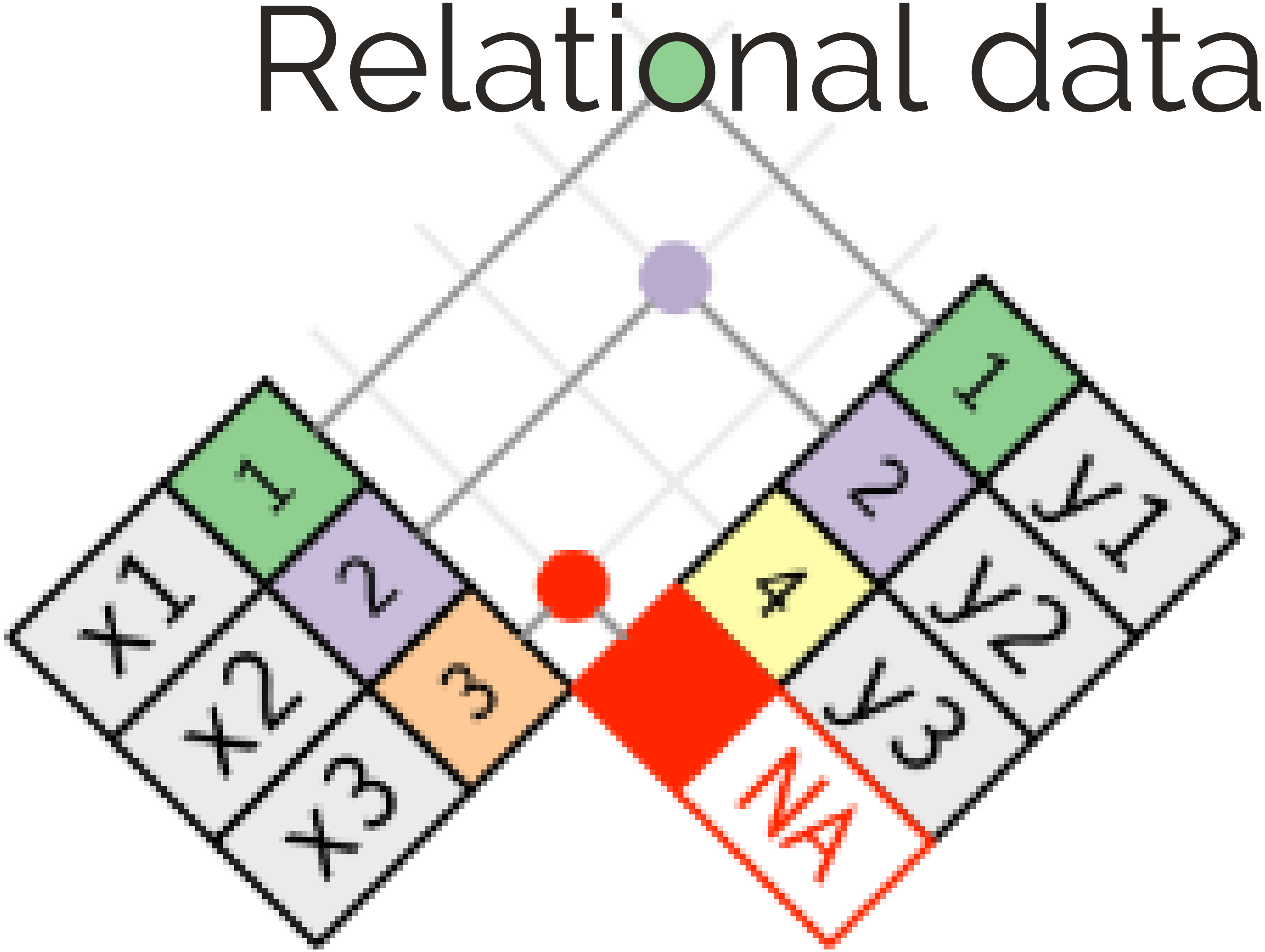


Image: R for Data Science

Andrew Jones | Strategy Unit

Relational data

It’s rare to find all the data you need for an analysis in a single table.

Typically, you’ll have to link two (or more) tables together by matching on common “key” variable(s).

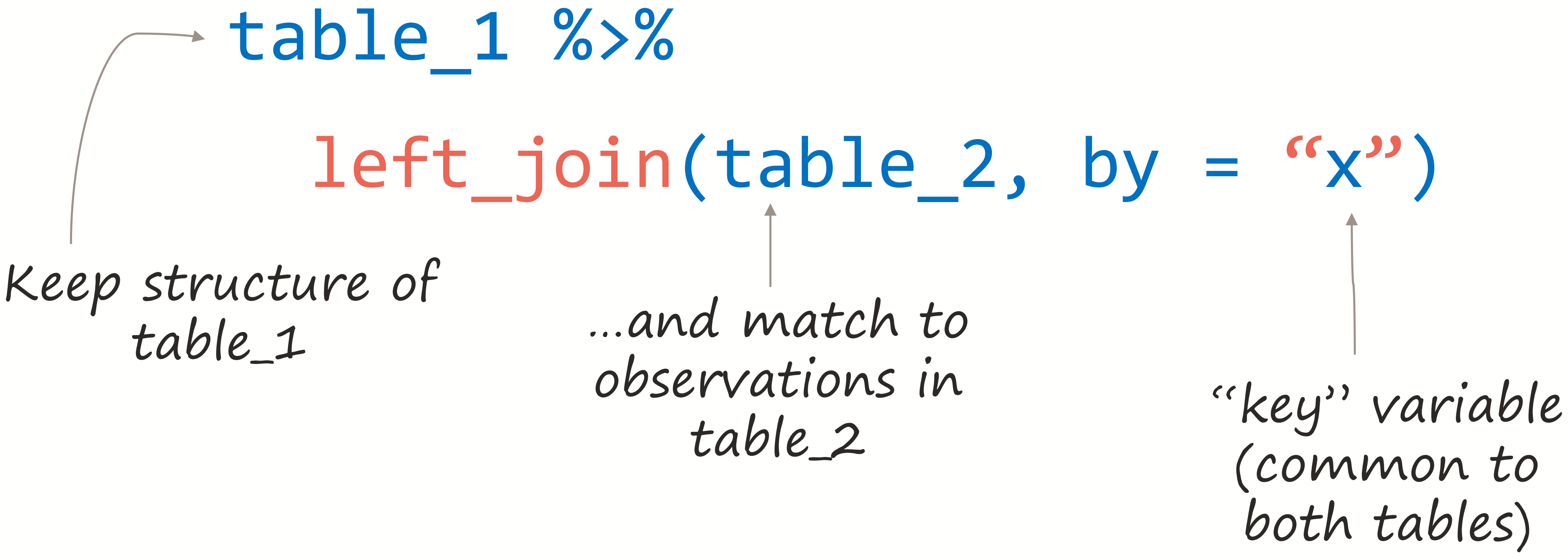
We use joins in SQL *or* R *(or VLOOKUP in* Excel)

Relational data

Here, we'll focus on left (outer) joins.

The syntax is similar for other types of join.

left_join



```
table_1 %>%
  left_join(table_2, by = “x”)
```

Keep structure of table_1

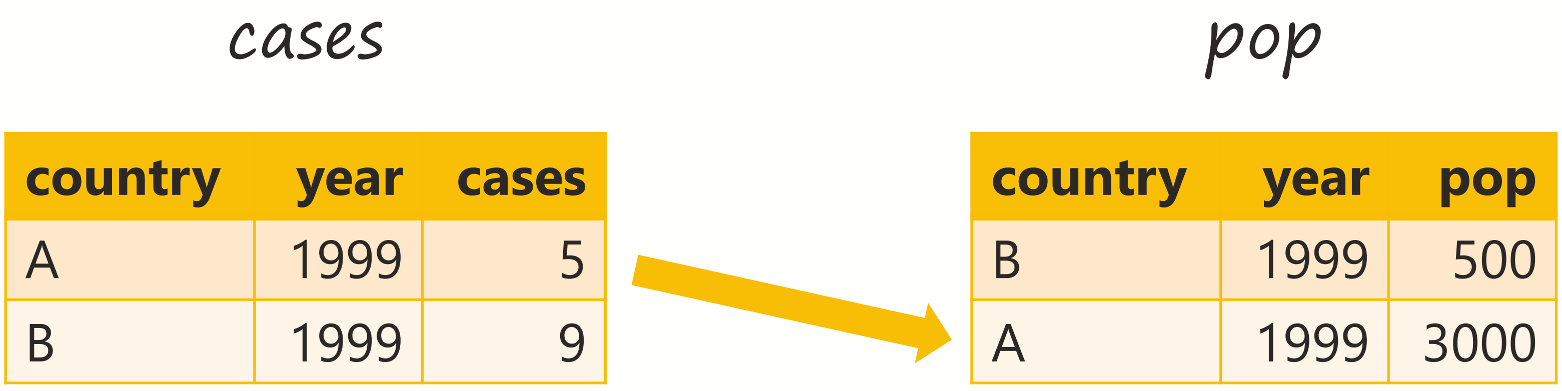
…and match to observations in table_2

“key” variable (common to both tables)

Relational Data

We’re going to join two tables - one with cases of tuberculosis by country, one with population by country.

From this new table we can derive a rate.



cases

country	year	cases
A	1999	5
B	1999	9

pop

country	year	pop
B	1999	500
A	1999	3000

Please Import

tb_cases.csv

and

tb_pop.csv

left_join

Keep the original structure of
the tb_cases data frame

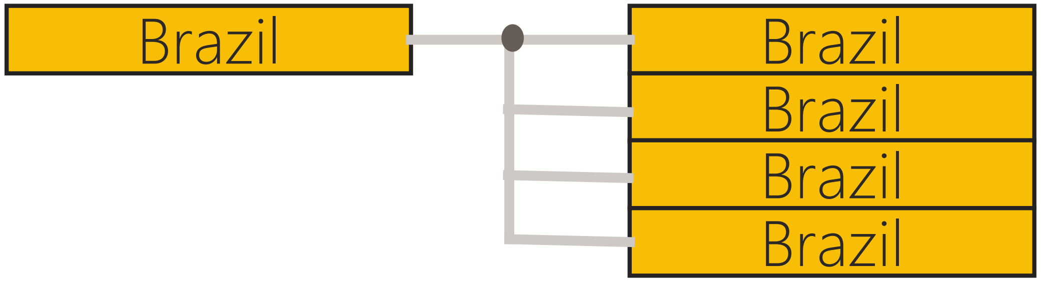
```
tb_cases %>%
  left_join(tb_pop, by = “country”)
```

...then match
to rows in
tb_pop

based on
“country”
value

Duplicates!

```
tb_cases %>%
  left_join(tb_pop, by = “country”)
```



Brazil

Brazil
Brazil
Brazil
Brazil

For every value of Brazil in tb_cases, there are 4 in tb_pop…

Join on multiple rows

match on two variables

```
tb_cases %>%
  left_join(tb_pop, by = c("country" , "year"))
```

c stands for ‘combine’

Joining with different names

If two tables have different names for same variable:

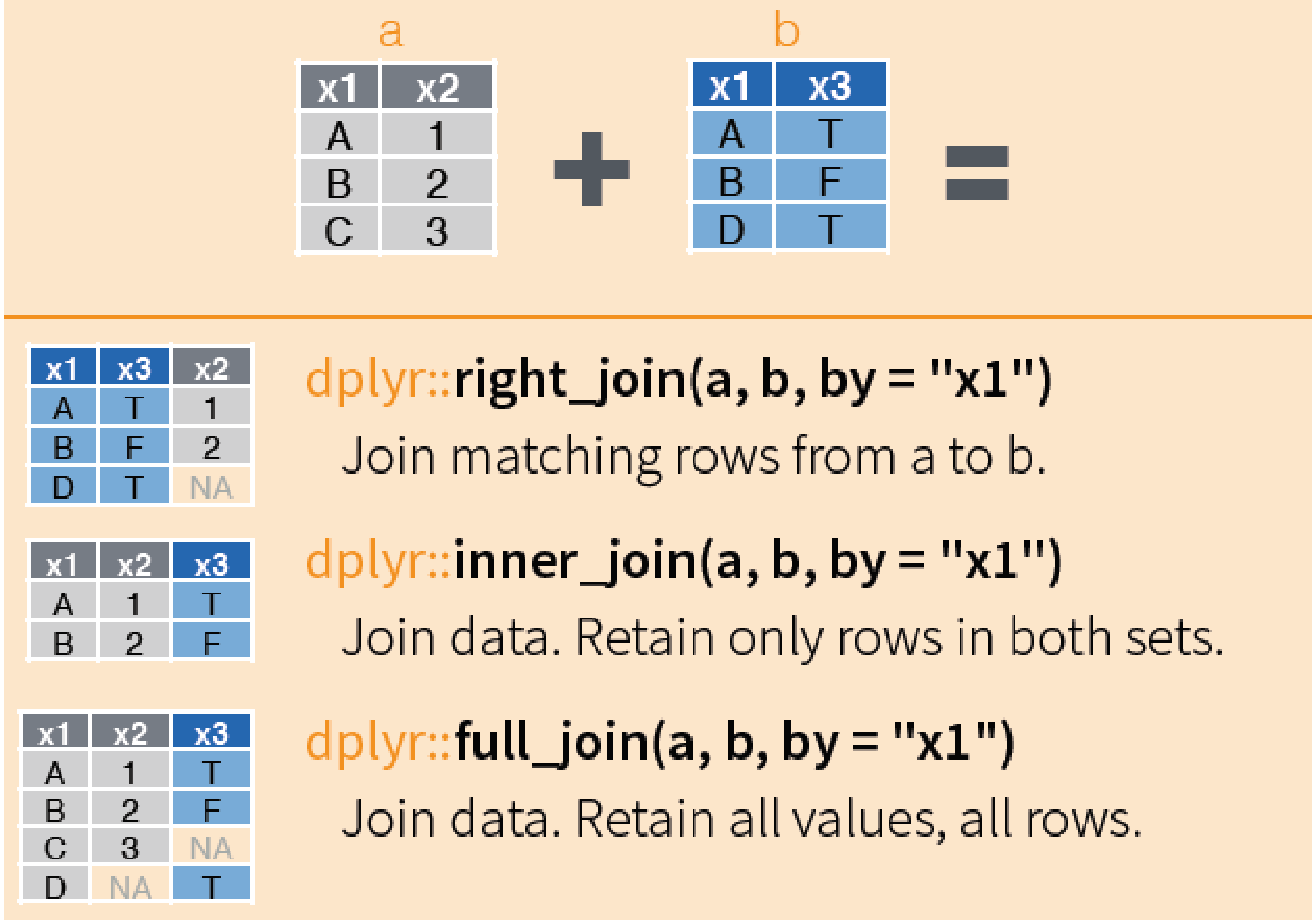
imaginary table

```
tb_cases %>%
   left_join(bad_names,
   by = c(“country” = “Place” , “year” = “Yr”))
```

name in cases

name in bad_names

Some other dplyr joins



a

x1	x2
A	1
B	2
C	3

+

b

x1	x3
A	T
B	F
D	T

=

x1	x3	x2
A	T	1
B	F	2
D	T	NA

dplyr::**right_join(a, b, by = "x1")**

Join matching rows from a to b.

x1	x2	x3
A	1	T
B	2	F

dplyr::**inner_join(a, b, by = "x1")**

Join data. Retain only rows in both sets.

x1	x2	x3
A	1	T
B	2	F
C	3	NA
D	NA	T

dplyr::**full_join(a, b, by = "x1")**

Join data. Retain all values, all rows.

Image taken from: https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

This work is licensed as

End