

Data Wrangling with the Tidyverse



Your Turn

Form groups of 2-4 people. Introduce yourself and prepare for some group quizzes. Tell them:

1. Who you are
2. What you do with data
3. How long you have been using R

05:00

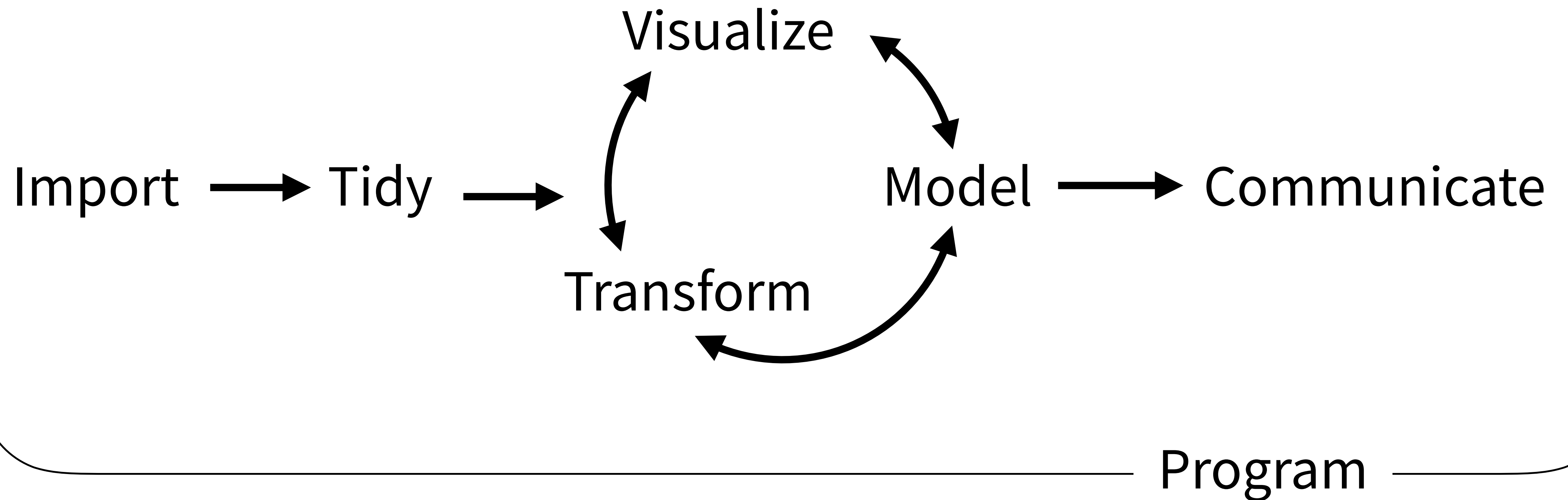
HELLO

my name is

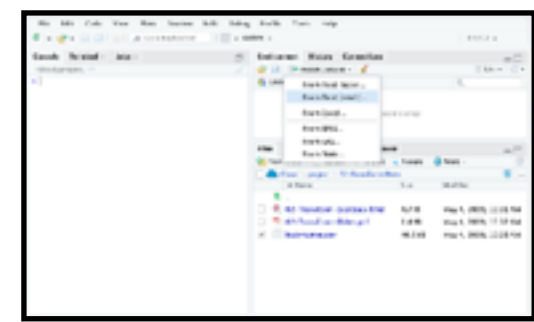
Garrett

 @StatGarrett

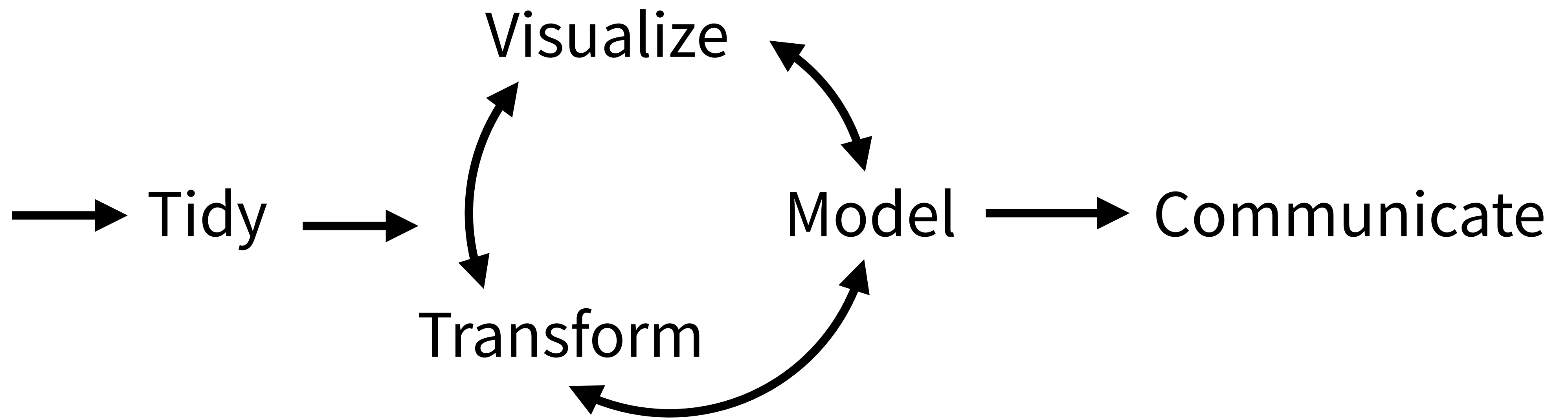
(Applied) Data Science



(Applied) Data Science

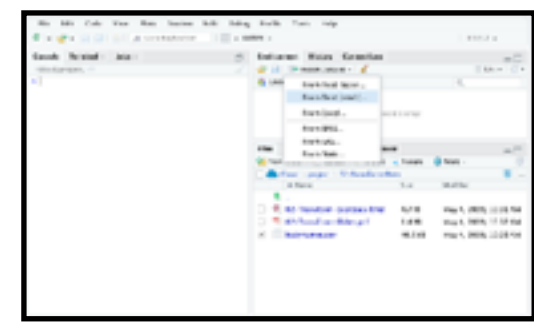


R Studio®

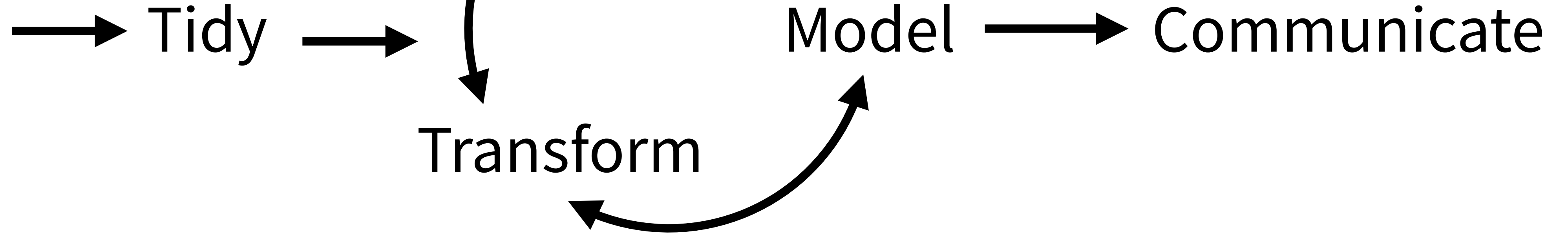


Program

(Applied) Data Science

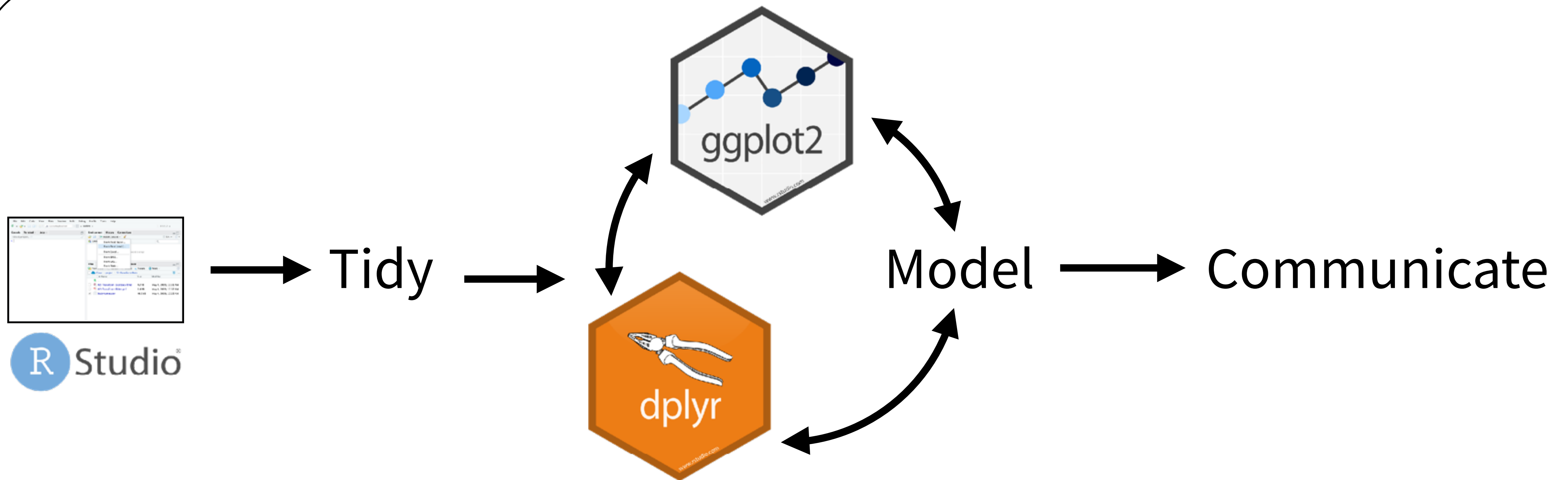


R Studio®



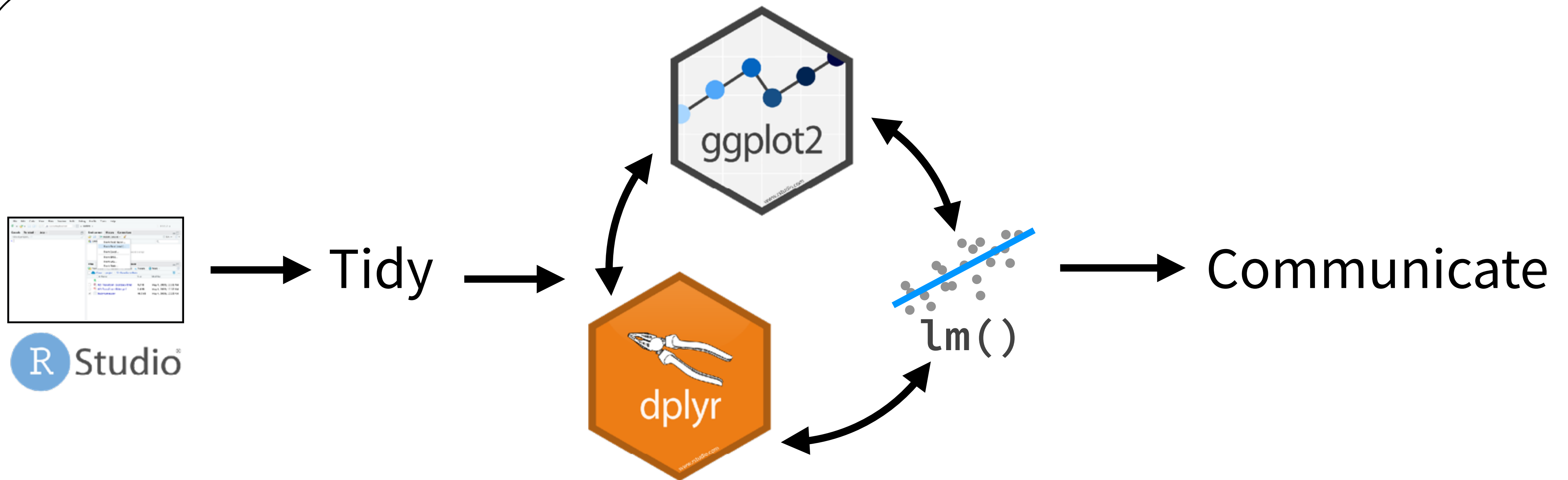
Program

(Applied) Data Science



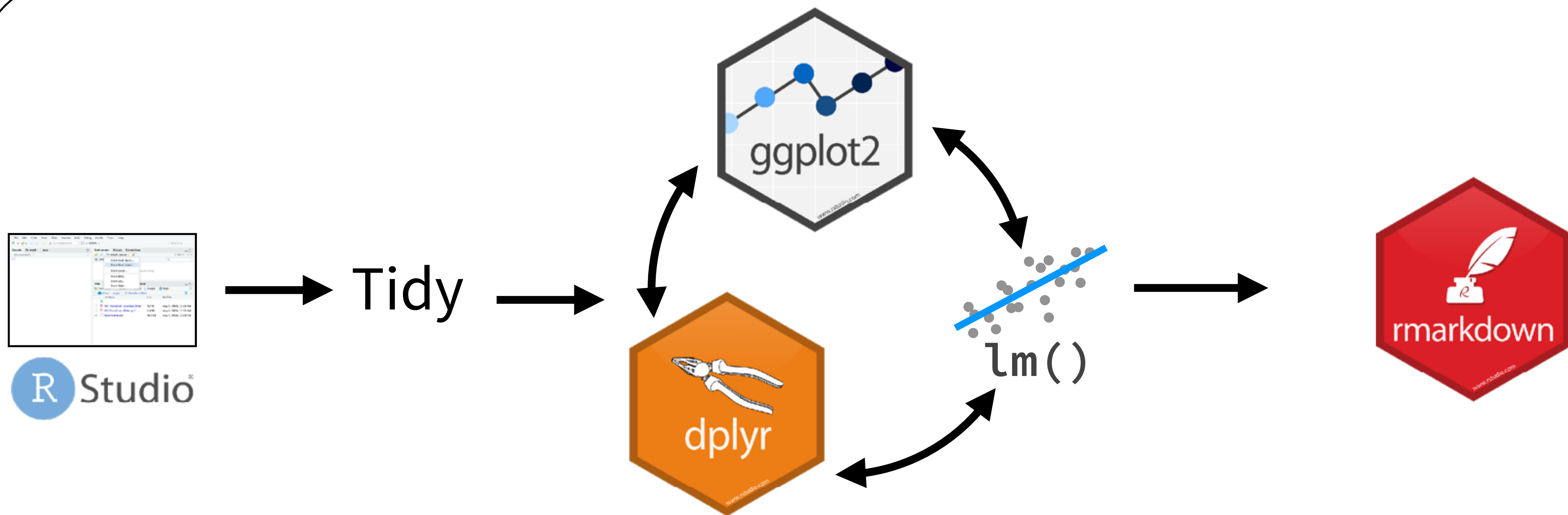
Program

(Applied) Data Science



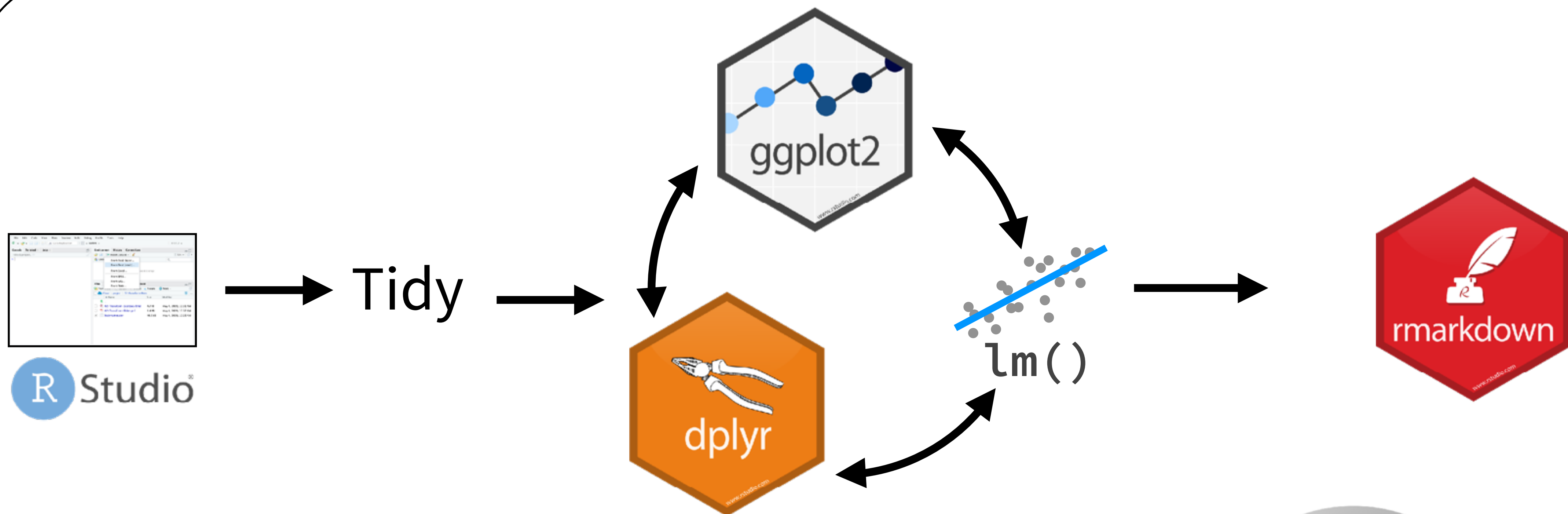
Program

(Applied) Data Science

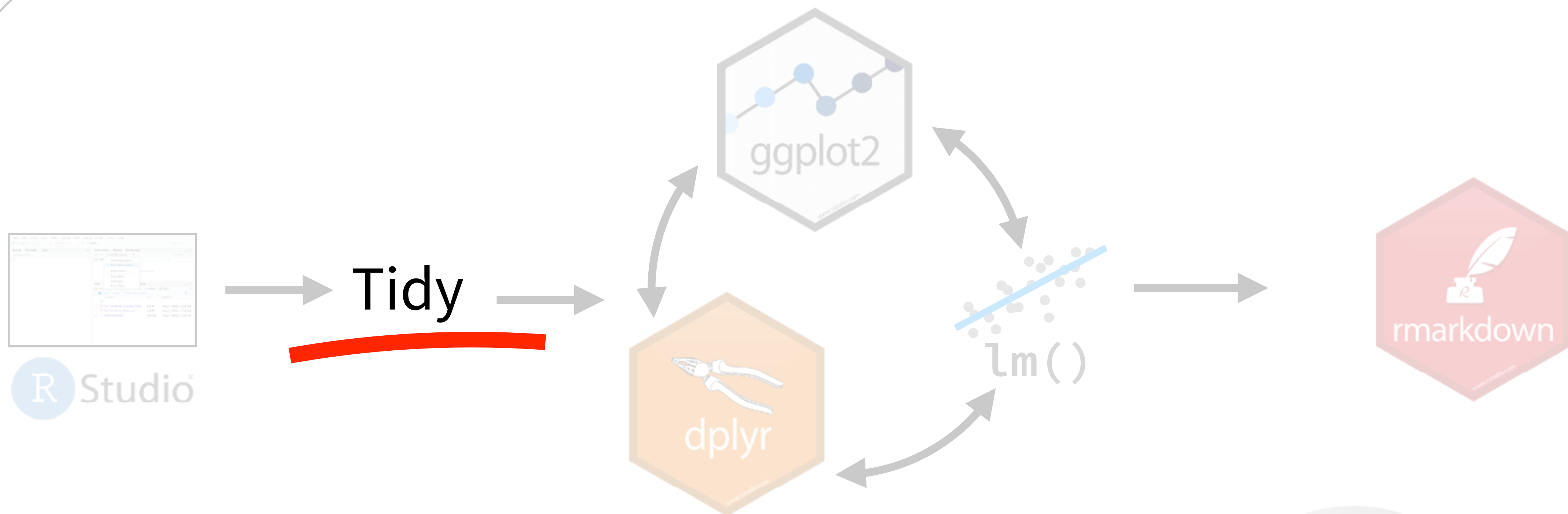


Program

(Applied) Data Science



(Applied) Data Science



Quiz

What types of data are in this data set?

	time_hour	name	air_time	distance	day	delayed
1	2013-01-01 05:00:00	United Air Lines Inc.	13620s (~3.78 hours)	1400	Tuesday	TRUE
2	2013-01-01 05:00:00	United Air Lines Inc.	13620s (~3.78 hours)	1416	Tuesday	TRUE
3	2013-01-01 05:00:00	American Airlines Inc.	9600s (~2.67 hours)	1089	Tuesday	TRUE
4	2013-01-01 05:00:00	JetBlue Airways	10980s (~3.05 hours)	1576	Tuesday	FALSE
5	2013-01-01 06:00:00	Delta Air Lines Inc.	6960s (~1.93 hours)	762	Tuesday	FALSE
6	2013-01-01 05:00:00	United Air Lines Inc.	9000s (~2.5 hours)	719	Tuesday	TRUE
7	2013-01-01 06:00:00	JetBlue Airways	9480s (~2.63 hours)	1065	Tuesday	TRUE
8	2013-01-01 06:00:00	ExpressJet Airlines Inc.	3180s (~53 minutes)	229	Tuesday	FALSE
9	2013-01-01 06:00:00	JetBlue Airways	8400s (~2.33 hours)	944	Tuesday	FALSE
10	2013-01-01 06:00:00	American Airlines Inc.	8280s (~2.3 hours)	733	Tuesday	TRUE
11	2013-01-01 06:00:00	JetBlue Airways	8940s (~2.48 hours)	1028	Tuesday	FALSE

Quiz

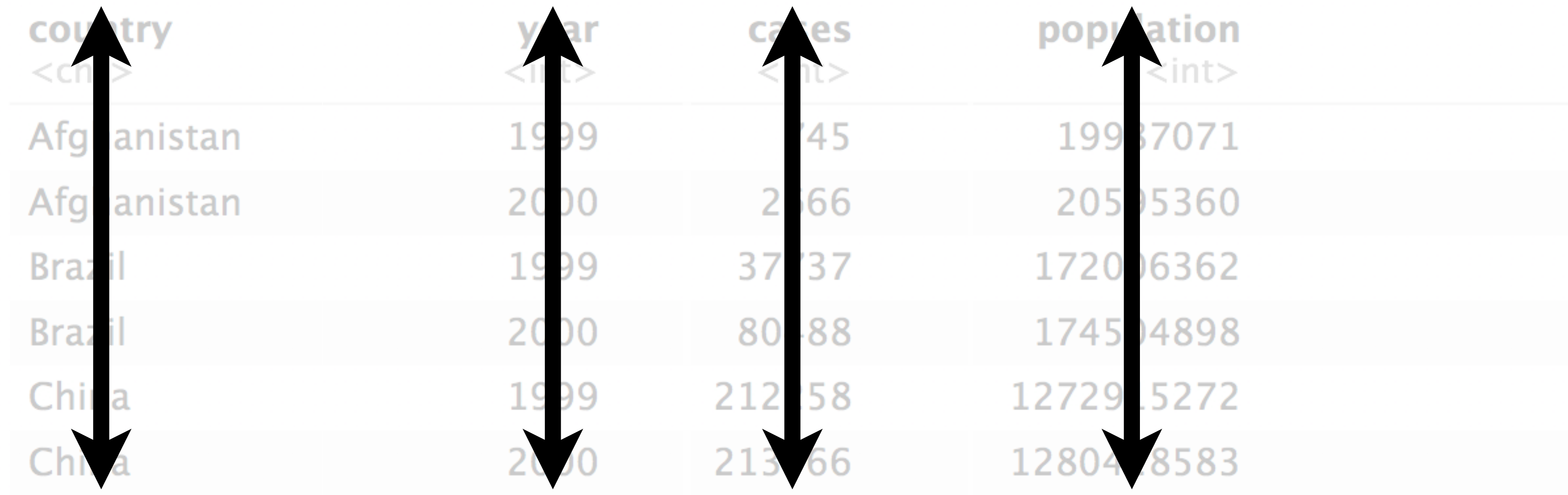
What types of data is this?

"2013-01-01 05:00:00"

Quiz

What are the variables in this data set?

table1



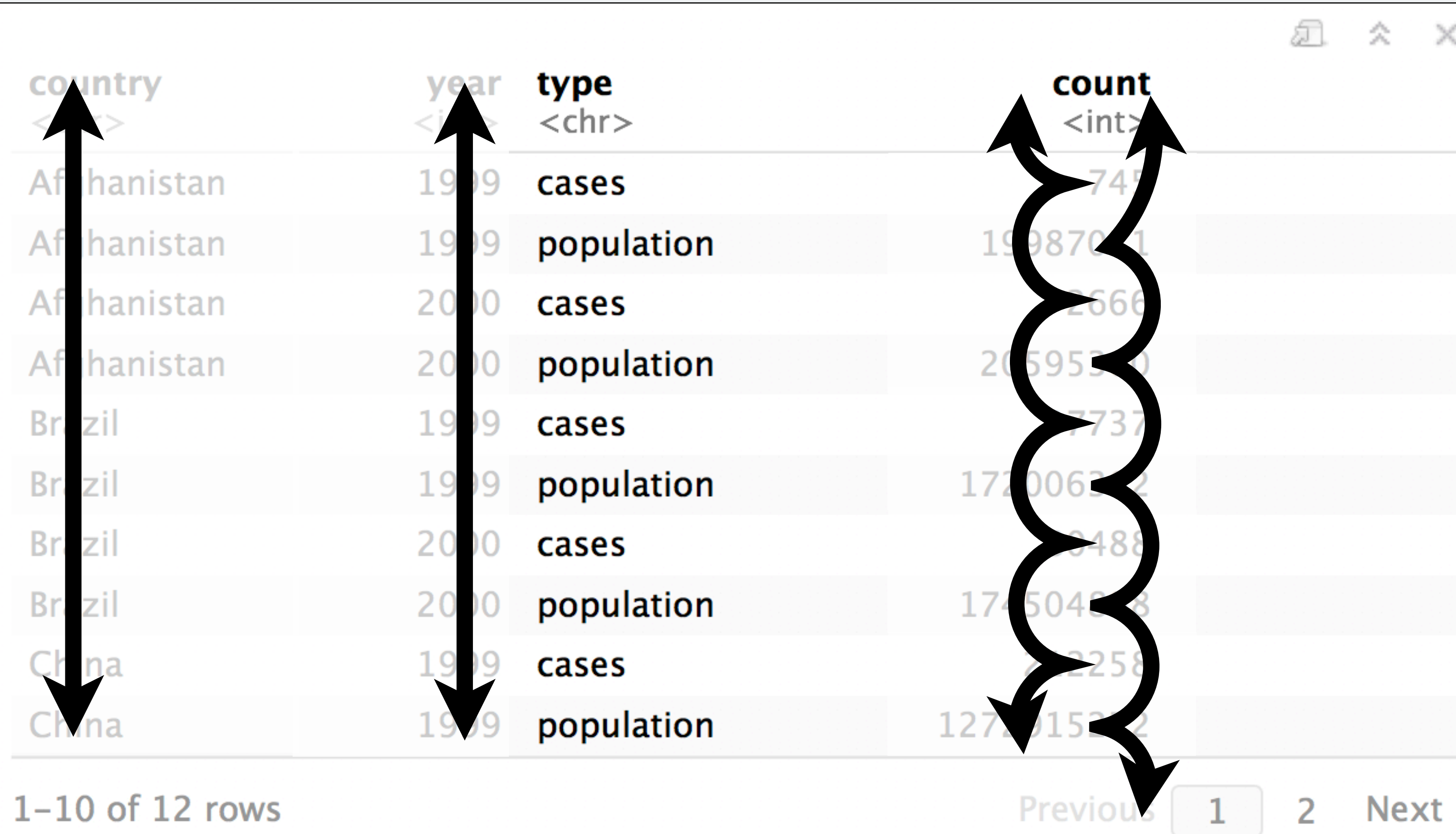
country <chr>	year <int>	cases <int>	population <int>
Afghanistan	1999	745	19987071
Afghanistan	2000	266	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212158	1272915272
China	2000	213766	128048583

6 rows

Quiz

What are the variables in this data set?

table2




The image shows a screenshot of a data table with four columns: country, year, type, and count. The table contains 10 rows of data. Hand-drawn arrows indicate the variables in the data set: a straight arrow for 'country', a straight arrow for 'year', and a wavy arrow for 'count'.

country <chr>	year <int>	type <chr>	count <int>
Afghanistan	1999	cases	745
Afghanistan	1999	population	1998701
Afghanistan	2000	cases	2666
Afghanistan	2000	population	2059530
Brazil	1999	cases	7737
Brazil	1999	population	17200632
Brazil	2000	cases	9488
Brazil	2000	population	17450488
China	1999	cases	2258
China	1999	population	12720152

1-10 of 12 rows

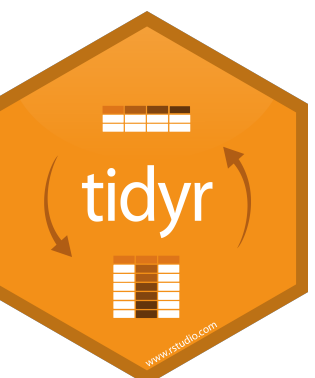
Previous 1 2 Next



country <chr>	year <int>	cases <int>	population <int>
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

6 rows

```
table1$country
table1$year
table1$cases
table1$population
```



📄 ⬆ ✕

country <chr>	year <int>	type <chr>	count <int>
Afghanistan	1999	cases	745
Afghanistan	1999	deaths	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	deaths	95360
Brazil	1999	cases	737
Brazil	1999	deaths	62
Brazil	2000	cases	88
Brazil	2000	deaths	8
China	1999	cases	58
China	1999	deaths	72

1–10 of 12 rows

Previous 1 2 Next

```
table2$country
table2$year
table2$count[c(1,3,5,7,9,11)]
table2$count[c(2,4,6,8,10,12)]
```


"Better experimental design = simpler statistics.
Better data model = simpler analysis."

- Jenny Bryan

Outline

**Introduction and
Joining Data**

8:00 - 9:45

Morning Break

9:45 - 10:15

Data Types

10:15 - 12:00

Lunch

12:00 - 1:30

Tidy Data

1:30 - 3:15

Afternoon Break

3:15 - 3:45

Lists

3:45 - 5:30

Your Turn

Go here and log in for the class materials

<https://rstudio.cloud/project/385945>

Navigate up to the 02-Join folder.

Open 02-Join-Exercises.Rmd

03:00