

# Welcome to the course

JOINING DATA WITH DATA.TABLE IN R



**Scott Ritchie**

Postdoctoral Researcher in Systems  
Genomics

# Joining data.tables

- Combine information from two data.tables into a single data.table

demographics:

name	gender	age
Trey	NA	54
Matthew	M	43
Angela	F	39

+

shipping:

name	address
Trey	12 High street
Matthew	7 Mill road
Angela	33 Pacific boulevard



name	gender	age	address
Trey	NA	54	12 High street
Matthew	M	43	7 Mill road
Angela	F	39	33 Pacific boulevard

# Course overview

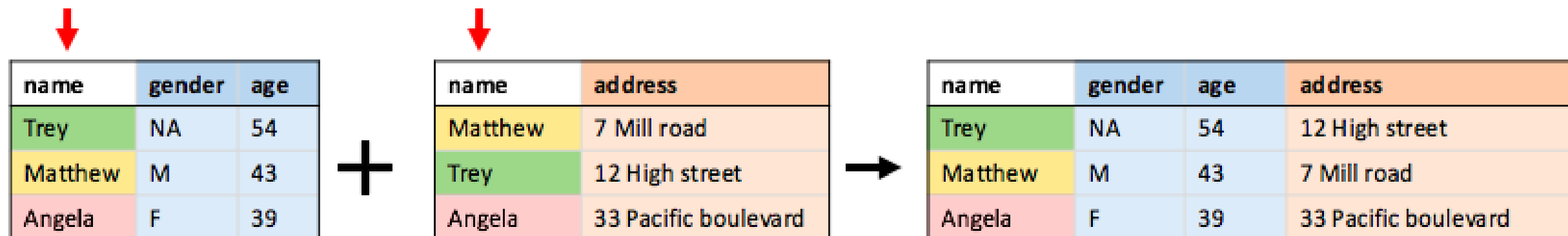
- Chapter 1: Joining data with `merge()`
- Chapter 2: Joins in the `data.table` workflow
- Chapter 3: Troubleshooting joins
- Chapter 4: Concatenating and reshaping `data.table` s

# Table keys

Columns that link information across two tables

```
library(data.table)
demographics <- data.table(name = c("Trey", "Matthew", "Angela"),
                           gender = c(NA, "M", "F"),
                           age = c(54, 43, 39))

shipping <- data.table(name = c("Matthew", "Trey", "Angela"),
                      address = c("7 Mill road", "12 High street",
                                "33 Pacific boulevard"))
```



# Inspecting `data.tables` in your R session

The `tables()` function will show you all `data.tables` loaded in your R session

```
tables()
```

```
      NAME NROW NCOL MB      COLS KEY
1: demographics    3    3  0 name,gender,age
2:      shipping    3    2  0      name,address
Total: 0MB
```

# Inspecting `data.tables` in your R session

The `str()` will show you the type of each column in a single `data.table`

```
str(demographics)
```

```
Classes 'data.table' and 'data.frame':   3 obs. of  3 variables:
 $ name   : chr  "Trey" "Matthew" "Angela"
 $ gender: chr   NA  "M"  "F"
 $ age    : num  54  43  39
- attr(*, ".internal.selfref")=<externalptr>
```

# Inspecting `data.tables` in your R session

```
demographics_all
```

```
   name sex age
1:   Trey NA  54
2: Matthew M  43
3:  Angela F  39
4: Michelle F  63
5: Mohamed M  26
---
102: Patrick M  27
103:   Wei  F  41
104:  Adam  M  33
105: Somchai M  53
106:  Alma  F  19
```

# Let's practice!

JOINING DATA WITH DATA.TABLE IN R



# The merge function

JOINING DATA WITH DATA.TABLE IN R



**Scott Ritchie**

Postdoctoral Researcher in Systems  
Genomics

# Joins

- Concept of joins come from database query languages (e.g. SQL).
- Four standard joins:
  - inner
  - full
  - left
  - right
- All four can be done using `merge()`

# Inner join

Only keep observations that have information in both `data.tables`

```
merge(x = demographics, y = shipping,  
      by.x = "name", by.y = "name")
```

demographics:

name	gender	age
Trey	NA	54
Matthew	M	43
Angela	F	39
Michelle	F	63

+

shipping:

name	address
Matthew	7 Mill road
Trey	12 High street
Abdullah	3a Union street
Angela	33 Pacific boulevard



name	gender	age	address
Angela	F	39	33 Pacific boulevard
Matthew	M	43	7 Mill road
Trey	M	NA	12 High street

# The by argument

Use `by` to avoid repeated typing of the same column name

```
merge(x = demographics, y = shipping,  
      by = "name")
```

demographics:

name	gender	age
Trey	NA	54
Matthew	M	43
Angela	F	39
Michelle	F	63

+

shipping:

name	address
Matthew	7 Mill road
Trey	12 High street
Abdullah	3a Union street
Angela	33 Pacific boulevard



name	gender	age	address
Angela	F	39	33 Pacific boulevard
Matthew	M	43	7 Mill road
Trey	M	NA	12 High street

# Full join

Keep all observations that are in either `data.table`

```
merge(x = demographics, y = shipping,  
      by = "name", all = TRUE)
```

demographics:

name	gender	age
Trey	NA	54
Matthew	M	43
Angela	F	39
Michelle	F	63

+

shipping:

name	address
Matthew	7 Mill road
Trey	12 High street
Abdullah	3a Union street
Angela	33 Pacific boulevard



name	gender	age	address
Abdullah	NA	NA	3a Union street
Angela	F	39	33 Pacific boulevard
Matthew	M	43	7 Mill road
Michelle	F	63	NA
Trey	M	NA	12 High street

# Let's practice!

JOINING DATA WITH DATA.TABLE IN R

# Left and right joins

JOINING DATA WITH DATA.TABLE IN R



**Scott Ritchie**

Postdoctoral Researcher in Systems  
Genomics

# Left joins

Add information **from** the right `data.table` **to** the left `data.table`

```
merge(x = demographics, y = shipping, by = "name", all.x = TRUE)
```

demographics:

name	gender	age
Trey	NA	54
Matthew	M	43
Angela	F	39
Michelle	F	63

+

shipping:

name	address
Matthew	7 Mill road
Trey	12 High street
Abdullah	3a Union street
Angela	33 Pacific boulevard



name	gender	age	address
Angela	F	39	33 Pacific boulevard
Matthew	M	43	7 Mill road
Michelle	F	63	NA
Trey	M	NA	12 High street



# Right joins

Add information **from** the left `data.table` to the **right** `data.table`

```
merge(x = demographics, y = shipping, by = "name", all.y = TRUE)
```

demographics:

name	gender	age
Trey	NA	54
Matthew	M	43
Angela	F	39
Michelle	F	63

+

shipping:

name	address
Matthew	7 Mill road
Trey	12 High street
Abdullah	3a Union street
Angela	33 Pacific boulevard



name	gender	age	address
Abdullah	NA	NA	3a Union street
Angela	F	39	33 Pacific boulevard
Matthew	M	43	7 Mill road
Trey	M	NA	12 High street

# Right joins - Left joins

```
# Right join
```

```
merge(x = demographics, y = shipping, by = "name", all.y = TRUE)
```

```
# Same as
```

```
merge(x = shipping, y = demographics, by = "name", all.x = TRUE)
```

# Default values

- Default values for `all`, `all.x` and `all.y` are `FALSE` in the `merge()` function
- Look up function argument defaults using `help("merge")`

# Exercise instructions

Left **join** `shipping` to `demographics` :

```
merge(demographics, shipping, by = "name", all.x = TRUE)
```

Right **join** `shipping` to `demographics` :

```
merge(demographics, shipping, by = "name", all.y = TRUE)
```

# Let's practice!

JOINING DATA WITH DATA.TABLE IN R