STAT 209
Merging Datasets

March 7, 2018

Colin Reimer Dawson

Reminders

- HW5: First DataCamp on `dplyr` due *right now*
- Lab5 due *right now*
- Lab6 due by class time Friday
- HW6: Second DataCamp on `dplyr` due Monday
- Today's lab (Lab7) also due Monday

Today

- Merging data from two tables
- New verbs:
  - `inner_join()`
  - `left_join()`
- Lab7

Coming Up...

- Friday: Reshaping data and the "tidy" format
- Next week: Some CS nuts and bolts (functions, loops) as applied to R

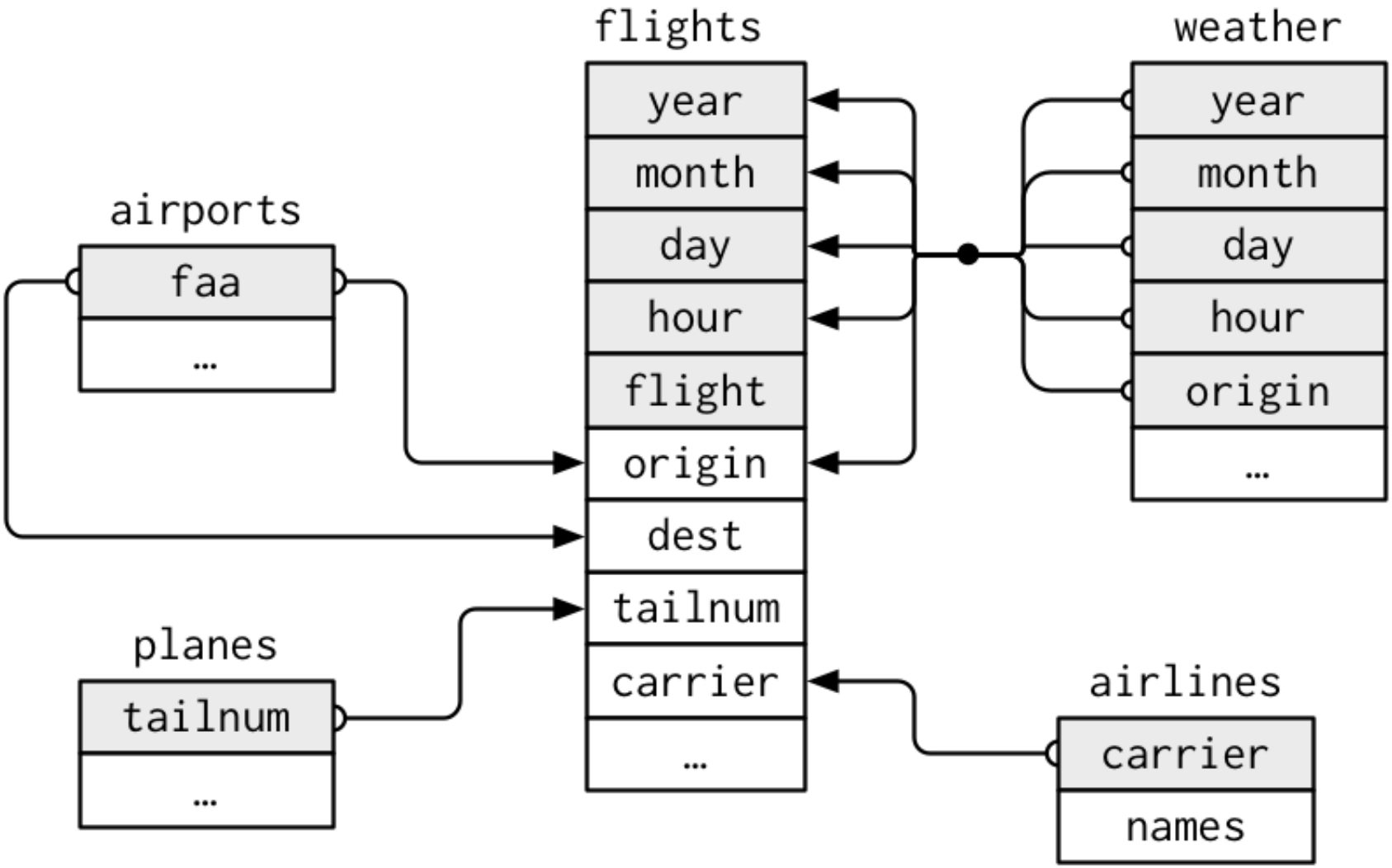
Outline

Relational Data

- "Relational data" is data from two or more tables that provide information about (some of) the same entities
- Example: `nycflights13` package contains datasets
  - `flights`
  - `airports`
  - `airlines`
  - `planes`
  - `weather`

  which are (in some sense) different views of the same objects and events ("entities")

A Relational Diagram



Joining flights with airlines

```
library(tidyverse); library(nycflights13)
names(flights)

 [1] "year"           "month"          "day"            "dep_time"
 [5] "sched_dep_time" "dep_delay"      "arr_time"       "sched_arr_time"
 [9] "arr_delay"      "carrier"        "flight"         "tailnum"
[13] "origin"         "dest"           "air_time"       "distance"
[17] "hour"           "minute"         "time_hour"

dim(flights)

[1] 336776     19

names(airlines)

[1] "carrier" "name"

dim(airlines)

[1] 16  2
```

Joining `flights` with `airlines`

```
flights %>% select(flight, carrier) %>% head(n = 4)
```

```
# A tibble: 4 x 2
  flight carrier
   <int>   <chr>
1   1545      UA
2   1714      UA
3   1141      AA
4    725      B6
```

```
airlines %>% select(carrier, name) %>% head(n = 4)
```

```
# A tibble: 4 x 2
  carrier                name
    <chr>               <chr>
1      9E      Endeavor Air Inc.
2      AA American Airlines Inc.
3      AS   Alaska Airlines Inc.
4      B6         JetBlue Airways
```

Joining flights with airlines

```
merged.table <- flights %>%
    inner_join(airlines, by = "carrier")
merged.table %>%
    select(flight, carrier, name, dep_time, sched_dep_time) %>%
    head(n = 4)
```

```
# A tibble: 4 x 5
  flight carrier                    name dep_time sched_dep_time
   <int>   <chr>                   <chr>    <int>          <int>
1   1545      UA  United Air Lines Inc.      517            515
2   1714      UA  United Air Lines Inc.      533            529
3   1141      AA American Airlines Inc.      542            540
4    725      B6        JetBlue Airways      544            545
```

```
dim(merged.table)
```

```
[1] 336776     20
```

```r
## Same result, but ordered by carrier as in airlines
airlines %>%
    inner_join(flights, by = "carrier") %>%
    select(flight, carrier, name, dep_time, sched_dep_time) %>%
    head(n = 4)
```

```
# A tibble: 4 x 5
  flight carrier name                dep_time sched_dep_time
   <int>   <chr>   <chr>               <int>          <int>
1   3538      9E Endeavor Air Inc.       810            810
2   4105      9E Endeavor Air Inc.      1451           1500
3   3295      9E Endeavor Air Inc.      1452           1455
4   3843      9E Endeavor Air Inc.      1454           1500
```

Types of Joins: Toy Example

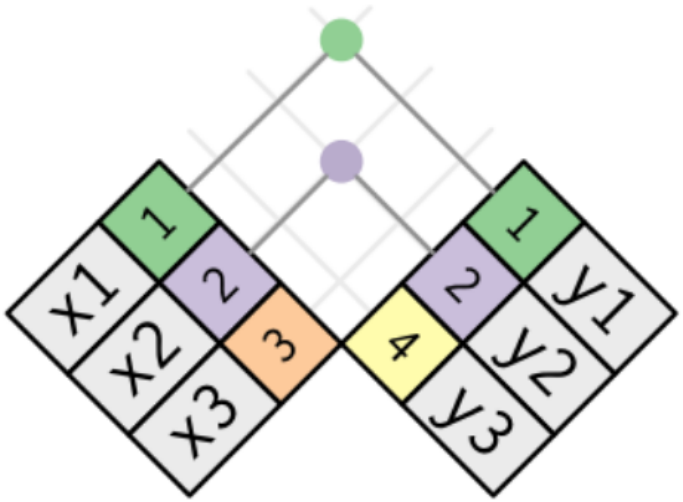
x

1	x1
2	x2
3	x3

y

1	y1
2	y2
4	y3

Inner Join

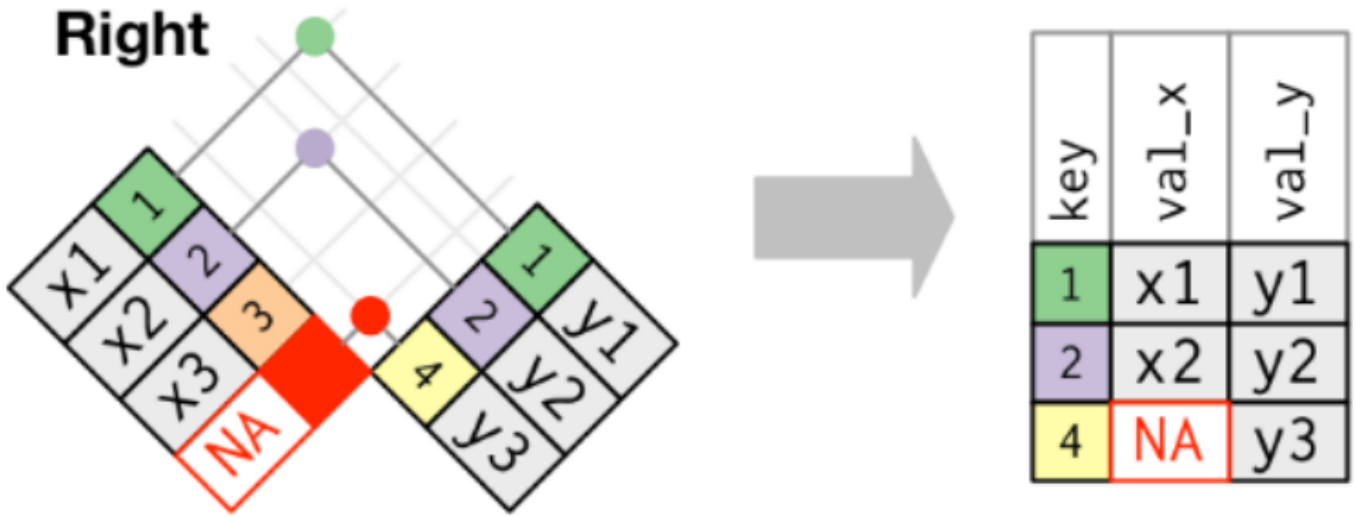


key	val_x	val_y
1	x1	y1
2	x2	y2

Left Join

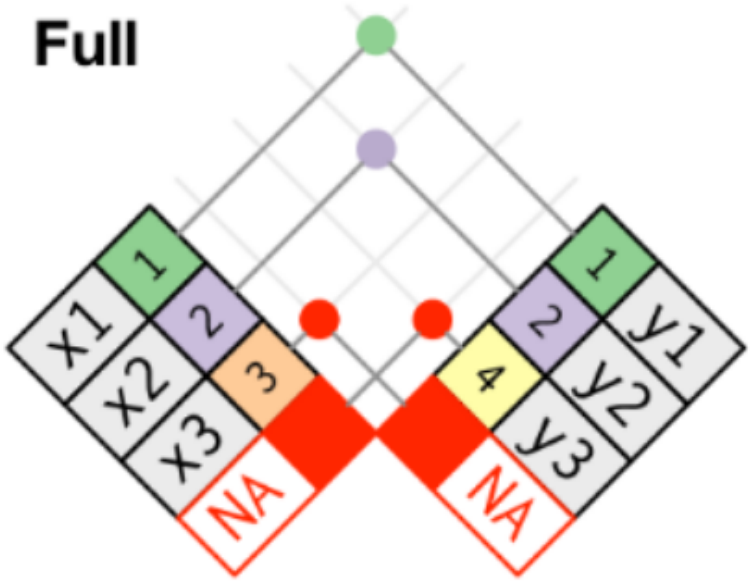

Left
x1
x2
x3
1
2
3
4
y1
y2
y3
NA
1
2
key
val_x
val_y
1
x1
y1
2
x2
y2
3
x3
NA

Right Join



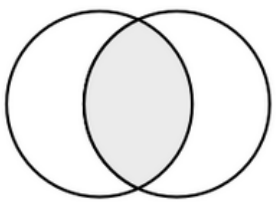
key	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

Full Join

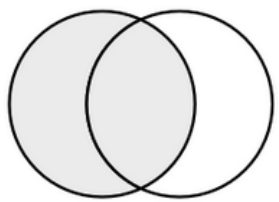


key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

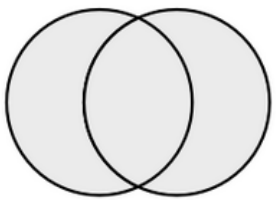
Venn Diagrams



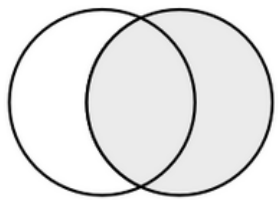
`inner_join(x, y)`



`left_join(x, y)`

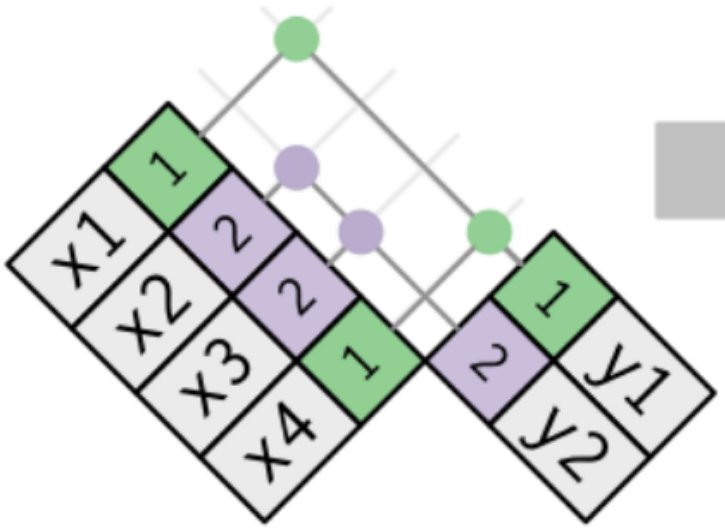


`full_join(x, y)`



`right_join(x, y)`

Many-to-One Mapping



val_x	key	val_y
x1	1	y1
x2	2	y2
x3	2	y2
x4	1	y1