STAT 209
Combining Data From Multiple Sources

July 1, 2021

Colin Reimer Dawson

Today

- Merging data from two tables
- New verbs:
  - `inner_join()`
  - `left_join()` and `right_join()`
  - `full_join()`

Coming Up...

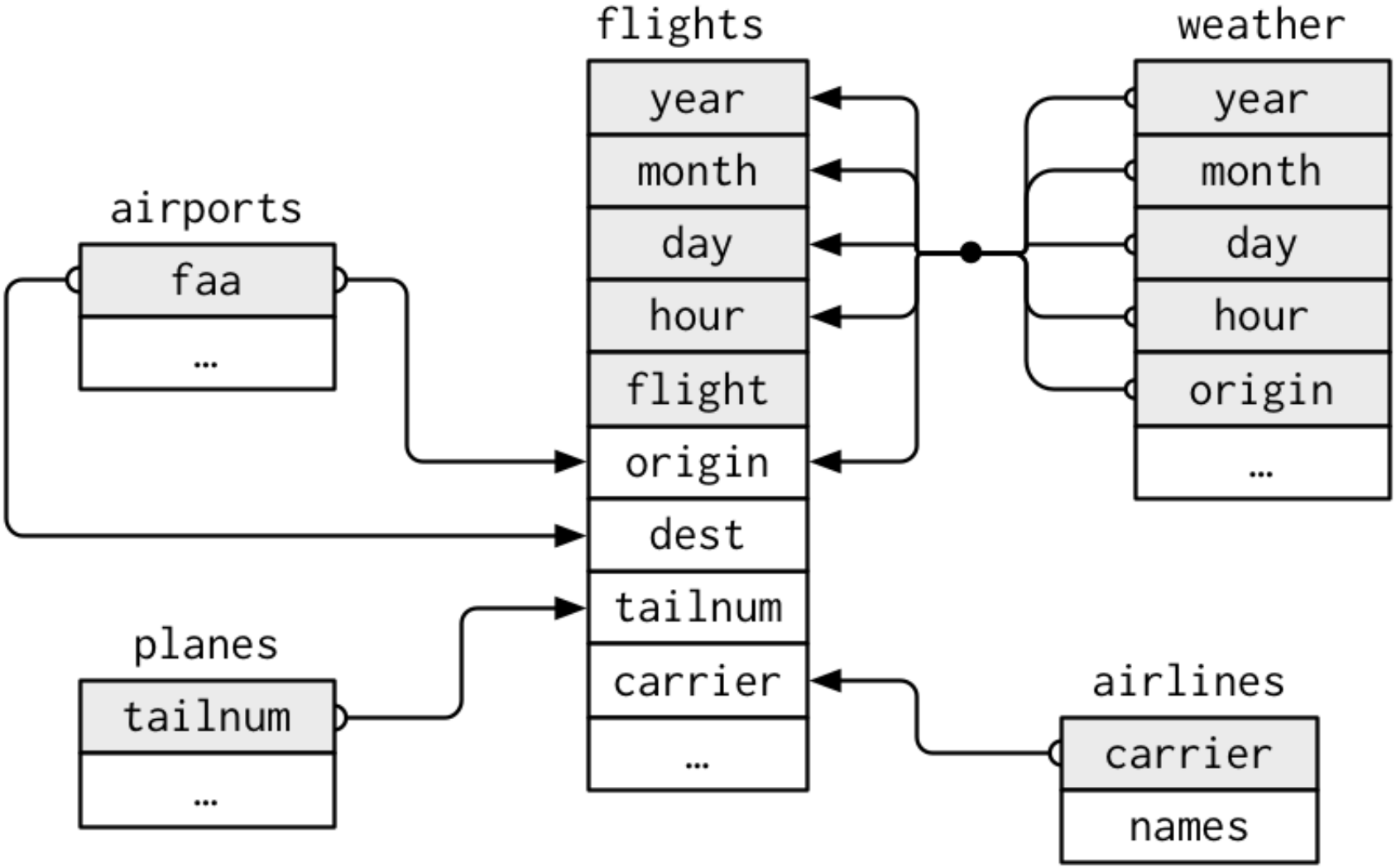
- Tuesday: Reshaping data and the "tidy" format
- Next Thursday: Some CS nuts and bolts (custom functions, loops) in R

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")
- For example, **flights** involve particular **planes**, take off and land at **airports**, and occur during particular **weather** conditions

A Relational Diagram



`flights`: Detailed flight-by-flight information

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

```
# A tibble: 336,776 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>
 1  2013     1     1      517            515         2      830
 2  2013     1     1      533            529         4      850
 3  2013     1     1      542            540         2      923
 4  2013     1     1      544            545        -1     1004
 5  2013     1     1      554            600        -6      812
 6  2013     1     1      554            558        -4      740
 7  2013     1     1      555            600        -5      913
 8  2013     1     1      557            600        -3      709
 9  2013     1     1      557            600        -3      838
10  2013     1     1      558            600        -2      753
# ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
#   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
#   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#   minute <dbl>, time_hour <dttm>
```

airlines: Look-Up Table for Airline Name by Code

```
airlines

    # A tibble: 16 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
     5 DL      Delta Air Lines Inc.
     6 EV      ExpressJet Airlines Inc.
     7 F9      Frontier Airlines Inc.
     8 FL      AirTran Airways Corporation
     9 HA      Hawaiian Airlines Inc.
    10 MQ      Envoy Air
    11 OO      SkyWest Airlines Inc.
    12 UA      United Air Lines Inc.
    13 US      US Airways Inc.
    14 VX      Virgin America
    15 WN      Southwest Airlines Co.
```

Joining `flights` with `airlines`

```
flights %>% select(flight, carrier) %>% slice_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) %>% slice_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`

```
flightsWithCarrierNames <- flights %>%
    inner_join(airlines, by = "carrier")
flightsWithCarrierNames %>%
    select(flight, carrier, name, dep_time, sched_dep_time)
```

```
# A tibble: 336,776 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
 5    461 DL      Delta Air Lines Inc.          554            600
 6   1696 UA      United Air Lines Inc.         554            558
 7    507 B6      JetBlue Airways               555            600
 8   5708 EV      ExpressJet Airlines Inc.      557            600
 9     79 B6      JetBlue Airways               557            600
10    301 AA      American Airlines Inc.        558            600
# ... with 336,766 more rows
```

```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)
```

```
# A tibble: 336,776 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
 5   3792 9E      Endeavor Air Inc.       1507           1515
 6   3369 9E      Endeavor Air Inc.       1530           1530
 7   3338 9E      Endeavor Air Inc.       1546           1540
 8   3372 9E      Endeavor Air Inc.       1550           1550
 9   3459 9E      Endeavor Air Inc.       1552           1600
10   3331 9E      Endeavor Air Inc.       1554           1600
# ... with 336,766 more rows
```

Types of Joins: Toy Example

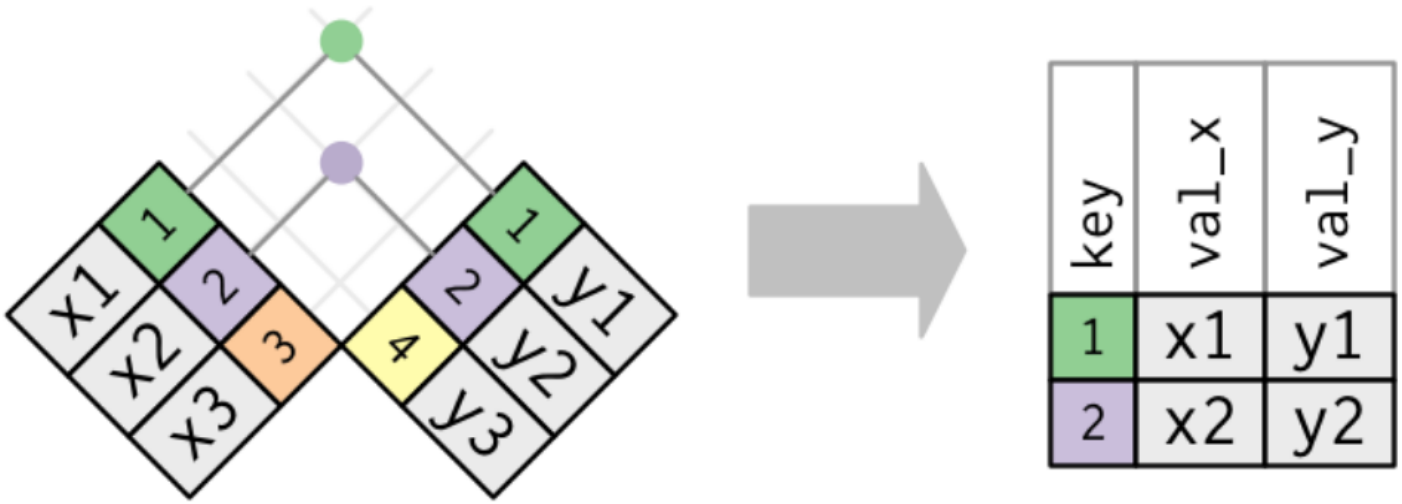
x

1	x1
2	x2
3	x3

y

1	y1
2	y2
4	y3

Inner Join

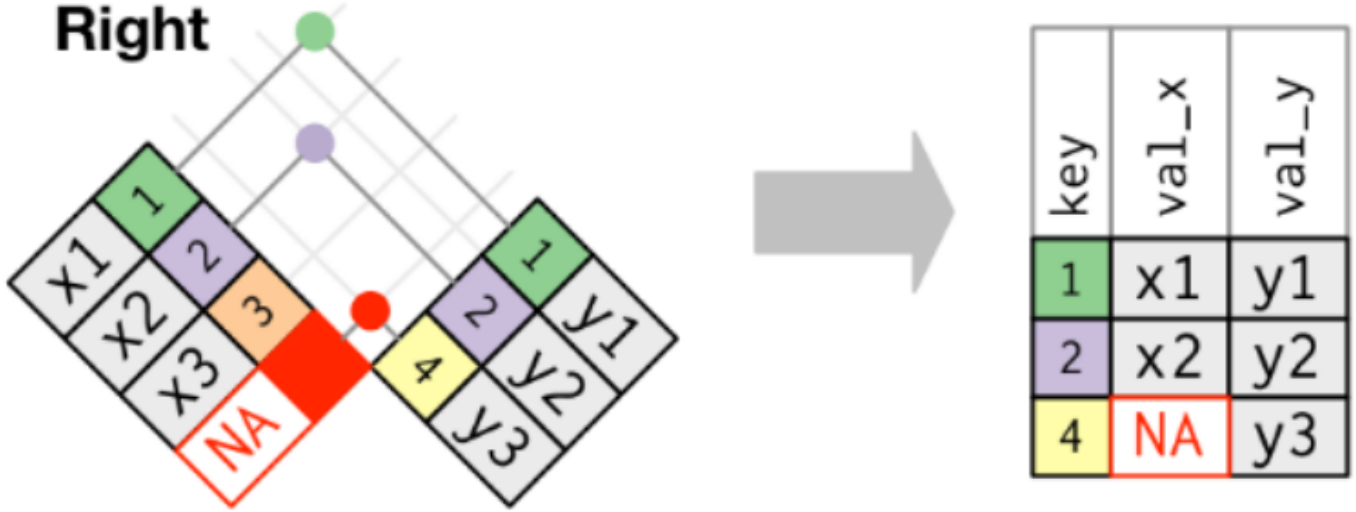


key	val_x	val_y
1	x1	y1
2	x2	y2

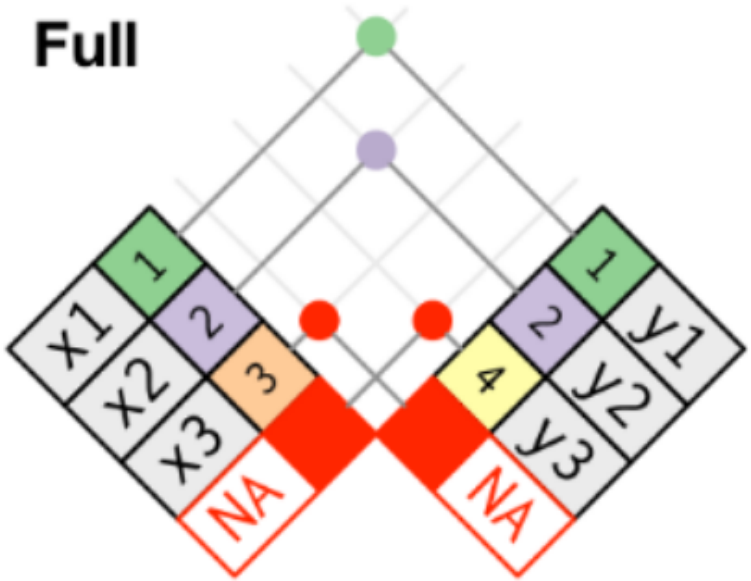
Left Join

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

Right Join

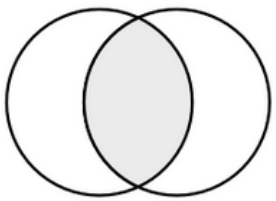

Right
x1
x2
x3
NA
y1
y2
y3
1
2
3
4
1
2
4
key
val_x
val_y
1
x1
y1
2
x2
y2
4
NA
y3

Full Join

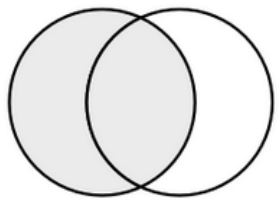


key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

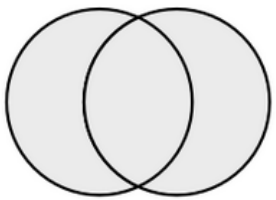
Venn Diagrams



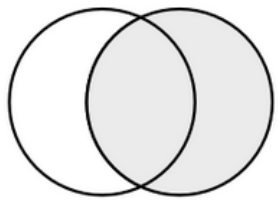
`inner_join(x, y)`



`left_join(x, y)`

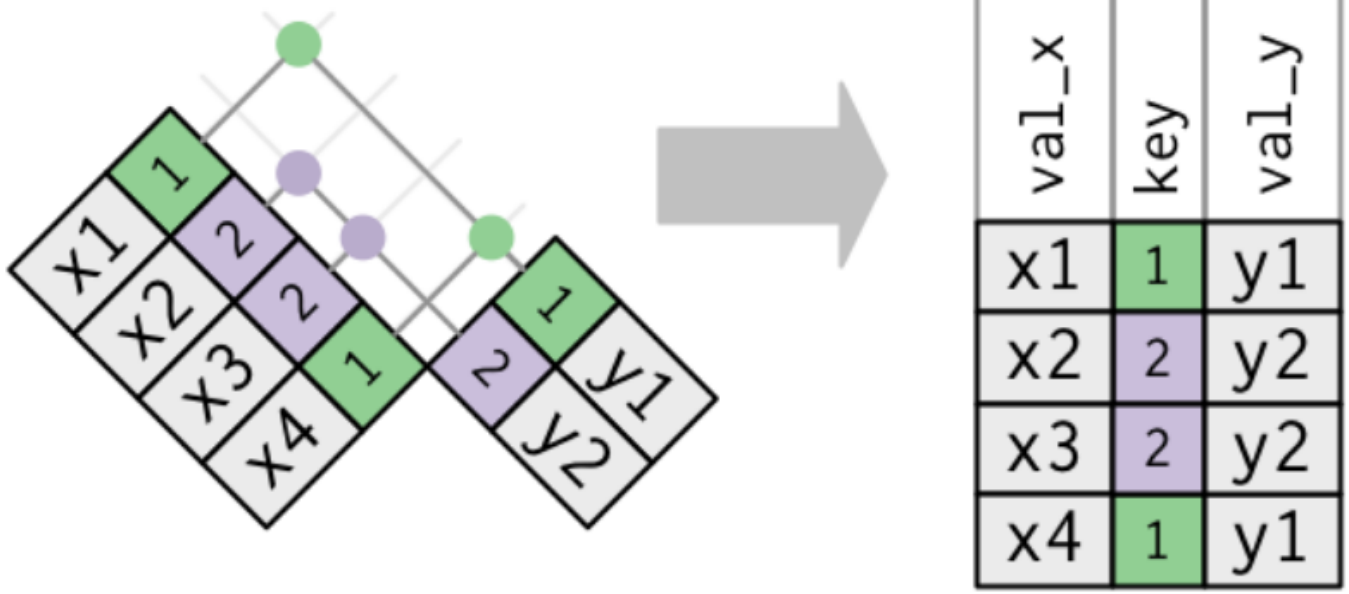


`full_join(x, y)`



`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