

The left_join verb

JOINING DATA WITH DPLYR



Chris Cardillo

Data Scientist

Batmobile vs. Batwing



Recall: inner join

```
inventory_parts_joined <- inventories %>%  
  inner_join(inventory_parts, by = c("id" = "inventory_id")) %>%  
  select(-id, -version) %>%  
  arrange(desc(quantity))
```

```
inventory_parts_joined
```

```
# A tibble: 258,958 x 4  
  set_num part_num color_id quantity  
  <chr>    <chr>      <dbl>    <dbl>  
1 40179-1 3024         72      900  
2 40179-1 3024         15      900  
3 40179-1 3024          0      900  
4 40179-1 3024         71      900  
5 40179-1 3024         14      900  
6 k34434-1 3024         15      810  
7 21010-1 3023        320      771  
8 k34431-1 3024          0      720  
9 42083-1 2780          0      684  
10 k34434-1 3024          0      540  
# ... with 258,948 more rows
```

Filter for LEGO sets

```
batmobile <- inventory_parts_joined %>%  
  filter(set_num == "7784-1") %>%  
  select(-set_num)
```

```
batwing <- inventory_parts_joined %>%  
  filter(set_num == "70916-1") %>%  
  select(-set_num)
```

Comparing tables

batmobile

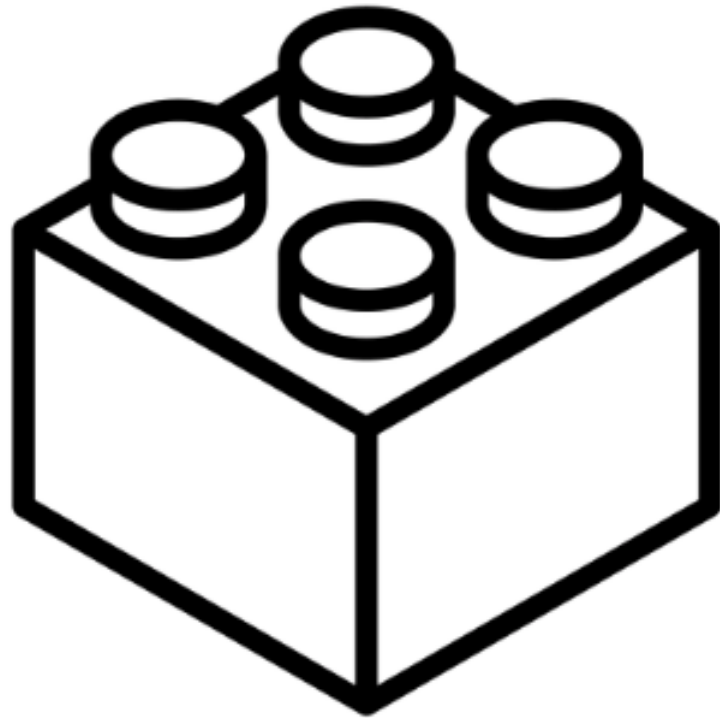
```
# A tibble: 173 x 3
  part_num color_id quantity
  <chr>      <dbl>     <dbl>
1 3023         72         62
2 2780          0         28
3 50950         0         28
4 3004         71         26
5 43093          1         25
6 3004          0         23
7 3010          0         21
8 30363         0         21
9 32123b        14         19
10 3622          0         18
# ... with 163 more rows
```

batwing

```
# A tibble: 309 x 3
  part_num color_id quantity
  <chr>      <dbl>     <dbl>
1 3023          0         22
2 3024          0         22
3 3623          0         20
4 11477         0         18
5 99207         71         18
6 2780          0         17
7 3666          0         16
8 22385         0         14
9 3710          0         14
10 99563         0         13
# ... with 299 more rows
```

Parts and pieces

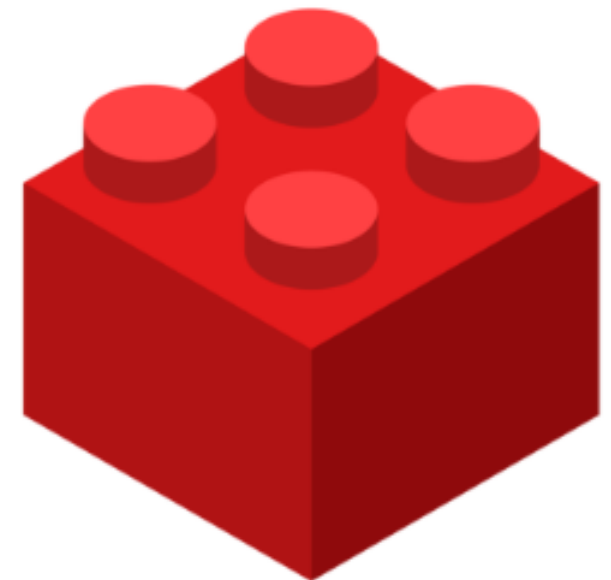
part



color



piece



Joining with multiple columns

```
batmobile %>%  
  inner_join(batwing, by = c("part_num", "color_id"), suffix = c("_batmobile", "_batwing"))
```

```
# A tibble: 45 x 4  
  part_num color_id quantity_batmobile quantity_batwing  
  <chr>      <dbl>          <dbl>          <dbl>  
1 2780         0           28             17  
2 50950        0           28              2  
3 3004        71           26              2  
4 43093         1           25              6  
5 3004         0           23              4  
6 3622         0           18              2  
7 4286         0           16              1  
8 3039         0           12              2  
9 4274        71           12              7  
10 3001         0           11              4  
# ... with 35 more rows
```

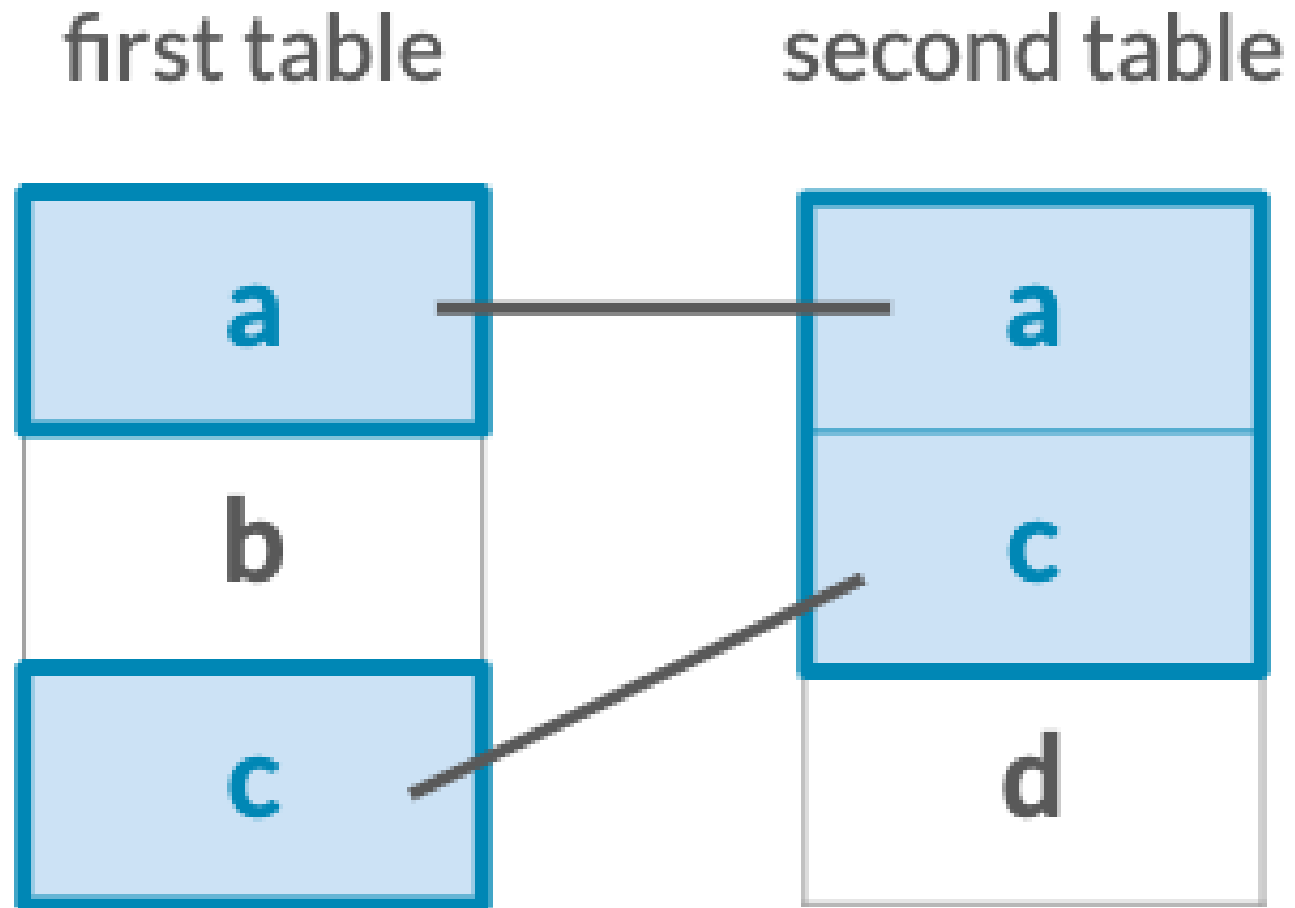
The left join

```
batmobile %>%  
  left_join(batwing, by = c("part_num", "color_id"), suffix = c("_batmobile", "_batwing"))
```

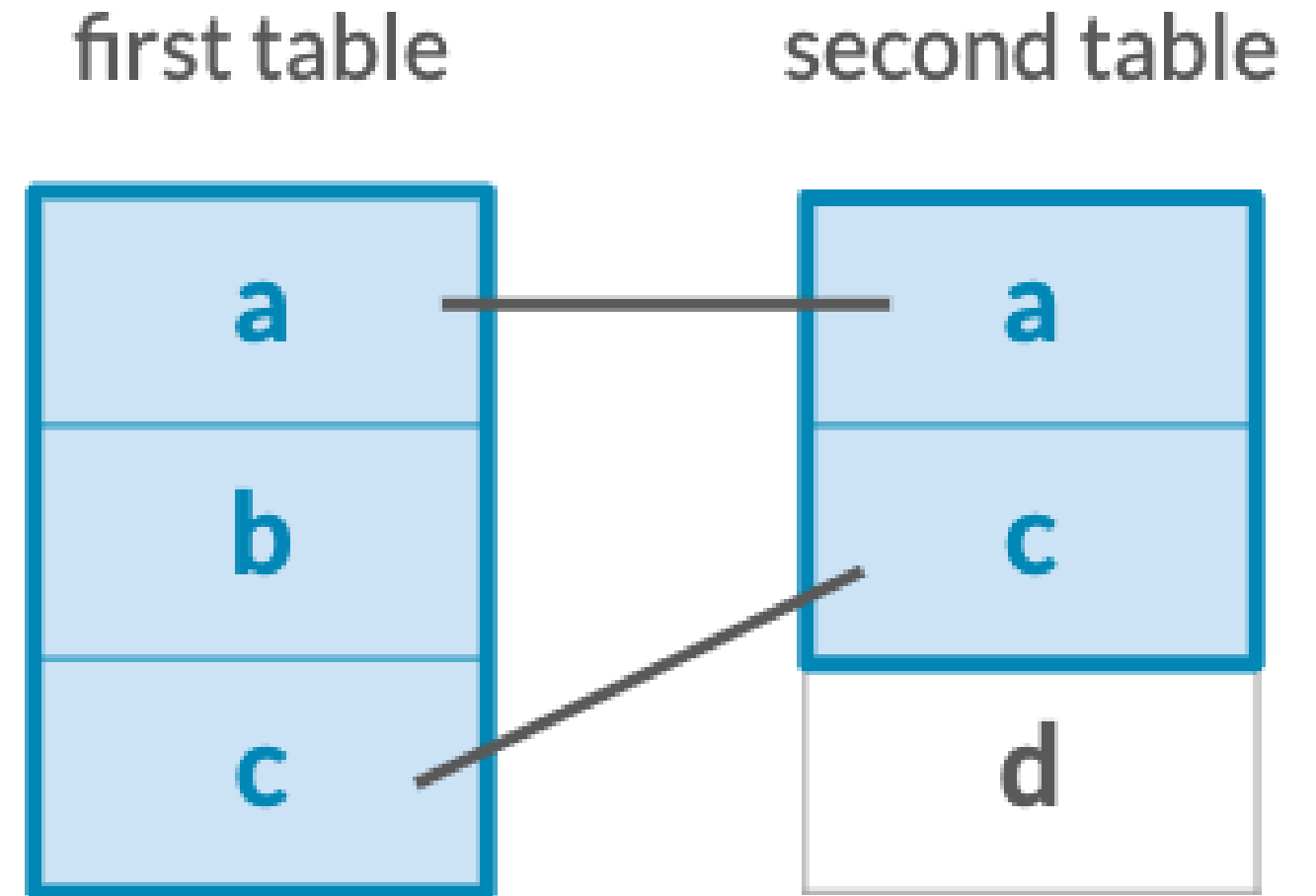
```
# A tibble: 173 x 4  
  part_num color_id quantity_batmobile quantity_batwing  
  <chr>      <dbl>          <dbl>          <dbl>  
1 3023         72           62             NA  
2 2780          0           28             17  
3 50950         0           28              2  
4 3004         71           26              2  
5 43093          1           25              6  
6 3004          0           23              4  
7 3010          0           21             NA  
8 30363         0           21             NA  
9 32123b        14           19             NA  
10 3622          0           18              2  
# ... with 163 more rows
```


Join review

Inner join



Left join



Let's practice!

JOINING DATA WITH DPLYR

The right-join verb

JOINING DATA WITH DPLYR



Chris Cardillo

Data Scientist

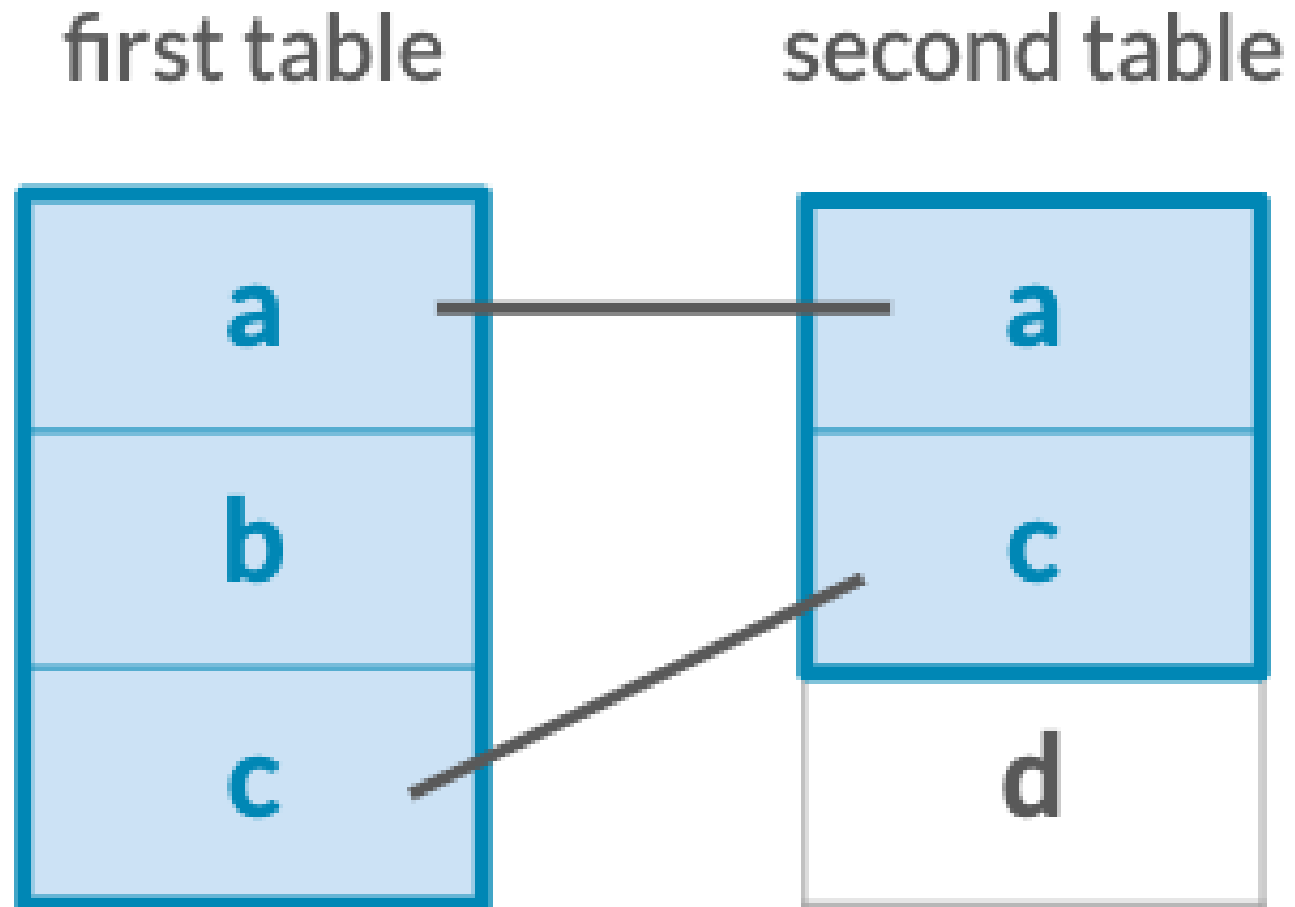
The right join

```
batmobile %>%  
  left_join(batwing, by = c("part_num", "color_id"), suffix = c("_batmobile", "_batwing"))
```

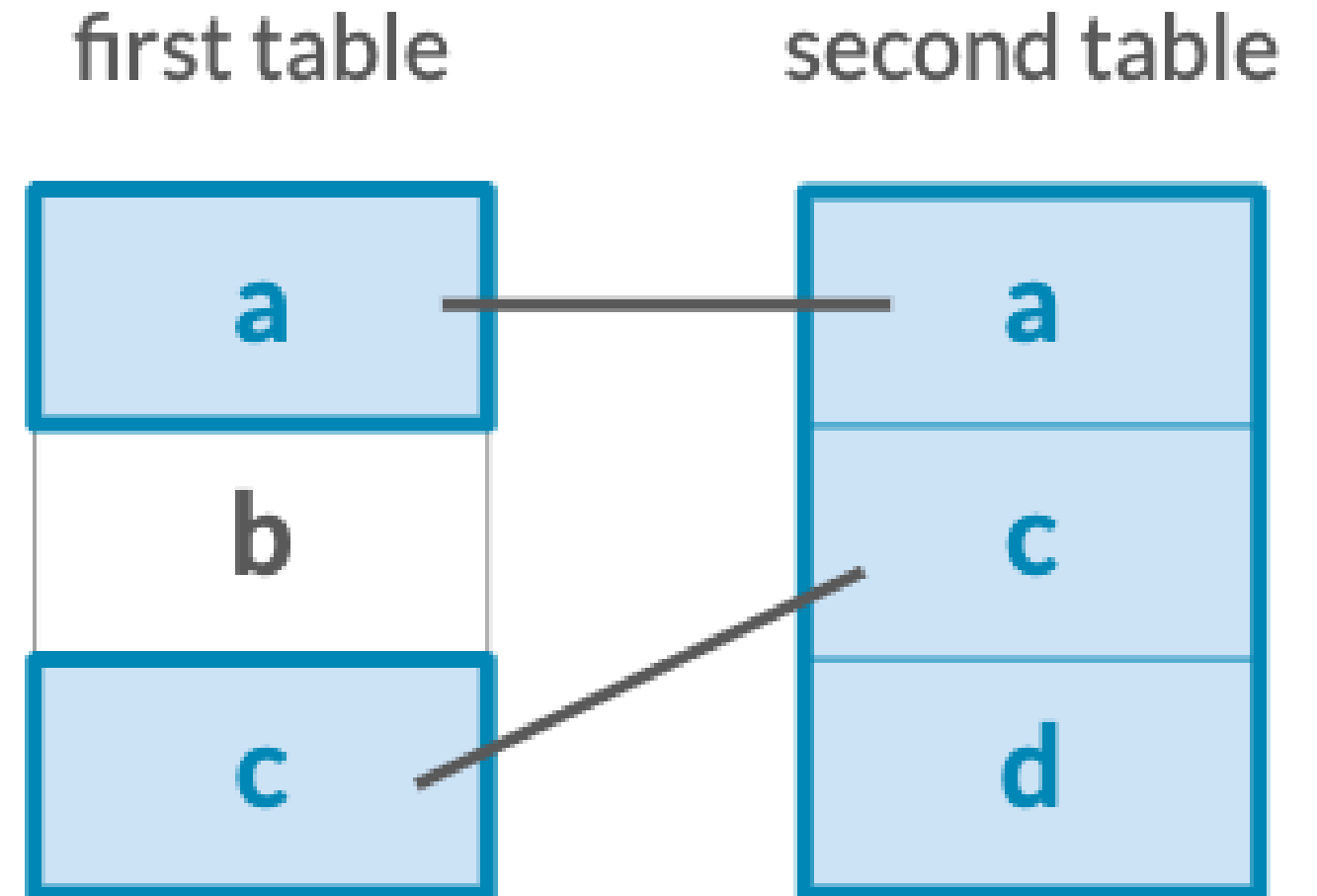
```
# A tibble: 173 x 4  
  part_num color_id quantity_batmobile quantity_batwing  
  <chr>      <dbl>          <dbl>          <dbl>  
1 3023         72           62             NA  
2 2780          0           28             17  
3 50950         0           28              2  
4 3004         71           26              2  
5 43093          1           25              6  
6 3004          0           23              4  
7 3010          0           21             NA  
8 30363         0           21             NA  
9 32123b        14           19             NA  
10 3622          0           18              2  
# ... with 163 more rows
```

The left and right join

Left join



Right join



Mirror images

```
batmobile %>%  
  right_join(batwing, by = c("part_num", "color_id"), suffix = c("_batmobile", "_batwing"))
```

```
# A tibble: 312 x 4  
  part_num color_id quantity_batmobile quantity_batwing  
  <chr>      <dbl>          <dbl>          <dbl>  
1 3023         0           NA             22  
2 3024         0             2             22  
3 3623         0          10             20  
4 11477        0           NA             18  
5 99207        71           NA             18  
6 2780         0          28             17  
7 2780         0             1             17  
8 3666         0           NA             16  
9 22385        0           NA             14  
10 3710         0           NA             14  
# ... with 302 more rows
```

Count and sort

```
sets %>%  
  count(theme_id, sort = TRUE)
```

```
# A tibble: 569 x 2  
  theme_id      n  
    <dbl> <int>  
1     501   122  
2     494   111  
3     435    94  
4     505    94  
5     632    93  
6     371    89  
7     497    86  
8     503    82  
9     516    78  
10    220    72  
# ... with 559 more rows
```

Inner join

```
sets %>%  
  count(theme_id, sort = TRUE) %>%  
  inner_join(themes, by = c("theme_id" = "id"))
```

```
# A tibble: 569 x 4  
  theme_id      n name      parent_id  
    <dbl> <int> <chr>      <dbl>  
1     501   122 Gear         NA  
2     494   111 Friends       NA  
3     435    94 Ninjago       NA  
4     505    94 Basic Set    504  
5     632    93 Town       504  
6     371    89 Supplemental 365  
7     497    86 Books         NA  
8     503    82 Key Chain    501  
9     516    78 Duplo and Explore 507  
10    220    72 City       217  
# ... with 559 more rows
```


Right join

```
sets %>%  
  count(theme_id, sort = TRUE) %>%  
  right_join(themes, by = c("theme_id" = "id"))
```

```
# A tibble: 665 x 4  
  theme_id      n name      parent_id  
    <dbl> <int> <chr>      <dbl>  
1         1    58 Technic          NA  
2         2     1 Arctic Technic        1  
3         3     4 Competition        1  
4         4    13 Expert Builder        1  
5         5     6 Model          1  
6         6     7 Airport          5  
7         7    20 Construction        5  
8         8    NA Farm          5  
9         9     2 Fire          5  
10        10     3 Harbor          5  
# ... with 655 more rows
```

Replace NAs

```
library(tidyr)
```

```
sets %>%  
  count(theme_id, sort = TRUE) %>%  
  right_join(themes, by = c("theme_id" = "id")) %>%  
  replace_na(list(n = 0))
```

```
# A tibble: 665 x 4  
  theme_id      n name      parent_id  
    <dbl> <dbl> <chr>      <dbl>  
1         1    58 Technic         NA  
2         2     1 Arctic Technic         1  
3         3     4 Competition         1  
4         4    13 Expert Builder         1  
5         5     6 Model         1  
6         6     7 Airport         5  
7         7    20 Construction         5  
8         8     0 Farm         5  
9         9     2 Fire         5  
10        10     3 Harbor         5  
# ... with 655 more rows
```

Let's practice!
JOINING DATA WITH DPLYR

Joining tables to themselves

JOINING DATA WITH DPLYR



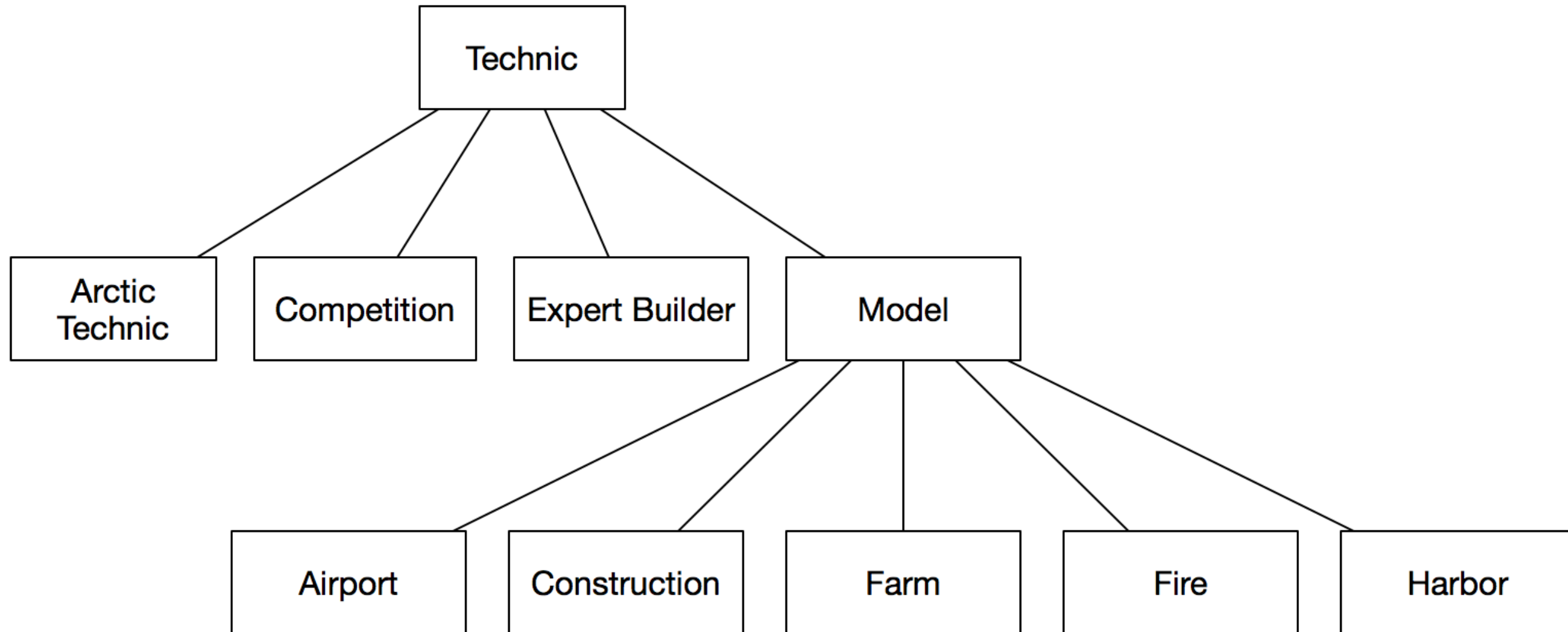
Chris Cardillo
Data Scientist

The themes table

themes

```
# A tibble: 665 x 3
  id name      parent_id
  <dbl> <chr>      <dbl>
1     1 Technic          NA
2     2 Arctic Technic     1
3     3 Competition        1
4     4 Expert Builder      1
5     5 Model                1
6     6 Airport              5
7     7 Construction         5
8     8 Farm                  5
9     9 Fire                  5
10    10 Harbor              5
# ... with 655 more rows
```

The hierarchy of themes



Child-parent table

```
themes %>%  
  inner_join(themes, by = c("parent_id" = "id"))
```

```
# A tibble: 544 x 5  
   id name.x      parent_id name.y parent_id.y  
  <dbl> <chr>      <dbl> <chr>      <dbl>  
1     2 Arctic Technic         1 Technic         NA  
2     3 Competition         1 Technic         NA  
3     4 Expert Builder         1 Technic         NA  
4     5 Model         1 Technic         NA  
5     6 Airport         5 Model           1  
6     7 Construction         5 Model           1  
7     8 Farm         5 Model           1  
8     9 Fire         5 Model           1  
9    10 Harbor         5 Model           1  
10   11 Off-Road         5 Model           1  
# ... with 534 more rows
```

Adding a suffix

```
themes %>%  
  inner_join(themes, by = c("parent_id" = "id"), suffix = c("_child", "_parent"))
```

```
# A tibble: 544 x 5  
   id name_child parent_id name_parent parent_id_parent  
  <dbl> <chr>      <dbl> <chr>      <dbl>  
1     2 Arctic Technic         1 Technic         NA  
2     3 Competition         1 Technic         NA  
3     4 Expert Builder         1 Technic         NA  
4     5 Model         1 Technic         NA  
5     6 Airport         5 Model           1  
6     7 Construction         5 Model           1  
7     8 Farm         5 Model           1  
8     9 Fire         5 Model           1  
9    10 Harbor         5 Model           1  
10   11 Off-Road         5 Model           1  
# ... with 534 more rows
```


Lord of the Rings themes: parent

```
themes %>%  
  inner_join(themes, by = c("parent_id" = "id"), suffix = c("_child", "_parent")) %>%  
  filter(name_child == "The Lord of the Rings")
```

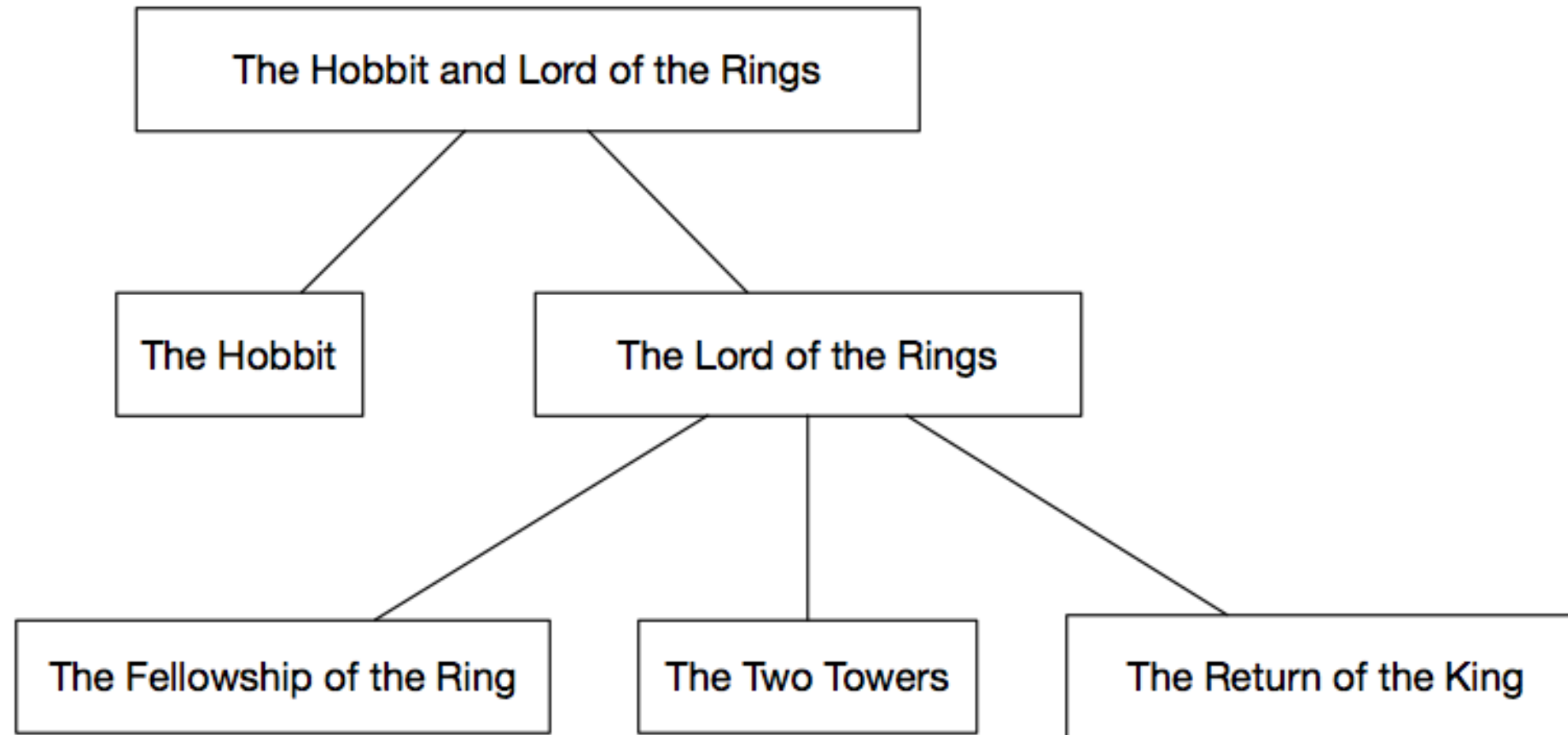
```
# A tibble: 1 x 5  
   id name_child parent_id name_parent parent_id_parent  
<dbl> <chr>      <dbl> <chr>      <dbl>  
1  566 The Lord of the Rings  561 The Hobbit and Lord of the Rings  NA
```

Lord of the Rings themes: children

```
themes %>%  
  inner_join(themes, by = c("parent_id" = "id"), suffix = c("_child", "_parent")) %>%  
  filter(name_parent == "The Lord of the Rings")
```

```
# A tibble: 3 x 5  
   id name_child parent_id name_parent parent_id_parent  
  <dbl> <chr>      <dbl> <chr>      <dbl>  
1  567 The Fellowship of the Ring    566 The Lord of the Rings    561  
2  568 The Two Towers                566 The Lord of the Rings    561  
3  569 The Return of the King         566 The Lord of the Rings    561
```

The Lord of the Rings trilogy



Let's practice!
JOINING DATA WITH DPLYR