Importing and combining data sets

Akshay Bareja

Importing data into R

The `readr` package (found in the `tidyverse` collection) contains a number of useful functions of the form `read_*` to import data. For example, if you have a `.csv` file, you would use the `read_csv` function

Download a file from uniprot.org

After selecting some columns of interest, click the Download button and download as a compressed Text file

Rename the file to something simple (yet informative!), like `uniprot` and make sure the extension is `.tsv`

To import into RStudio, run the following

```
uniprot <- read_tsv("uniprot.tsv")
```

```
## Parsed with column specification:
## cols(
##   Entry = col_character(),
##   `Gene names` = col_character(),
##   Length = col_double()
## )
```

You can also use the `readr` package to import data from a URL

For example, to load a dataset from the (very useful) Tidy Tuesday series, run the following

```
pizza <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-
```

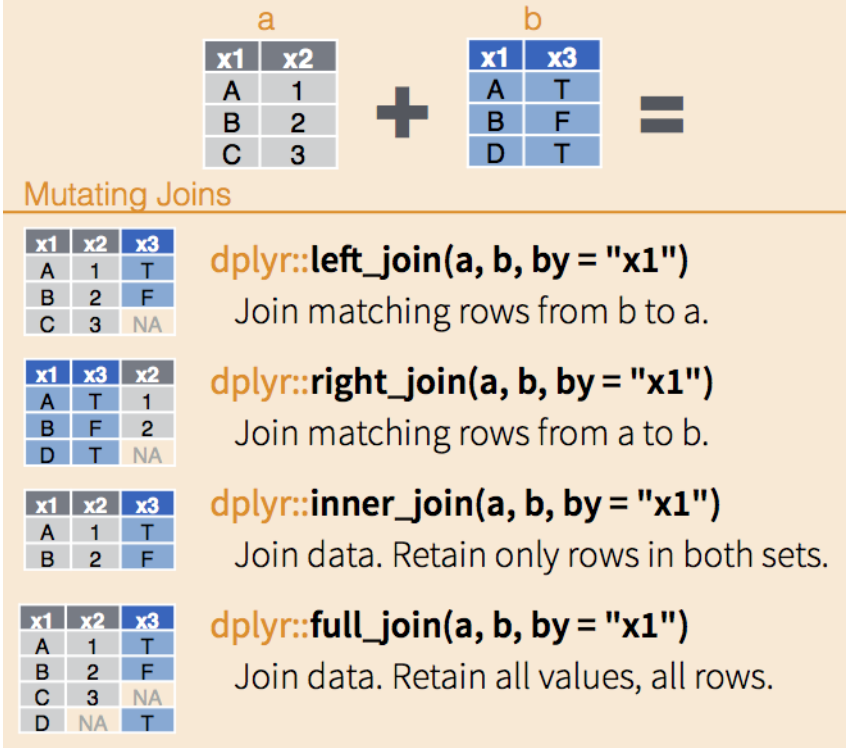
This data set contains ratings of various pizzerias in Manhattan

Combining datasets

There are many times when you have two or more overlapping datasets that you would like to combine

The `dplyr` package has a number of `*_join` functions for this purpose

---



`left_join`

Returns all rows from a, and all columns from a and b

Rows in a with no match in b will have NA values in the new columns

If there are multiple matches between a and b, all combinations of the matches are returned

First, load the two datasets needed for this example - `proteins` and `mitocarta`

```
library(proteins)
library(mitocarta)
```

Take a look at the variables in each dataset

`gene_name` in `proteins` and `symbol` in `mitocarta` contain the gene IDs for each gene. They can therefore serve as a common variable

Let's join on this

`left_join` example

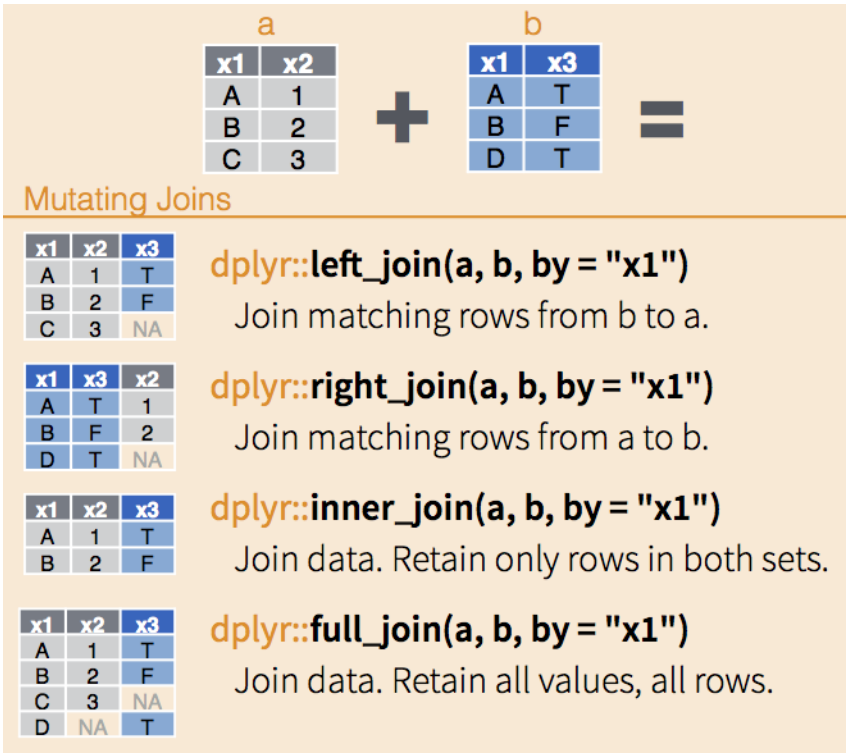
`left_join` `proteins` with `mitocarta` and assign the output to a new object called `pm_left`

```
pm_left <- left_join(proteins, mitocarta, by = c("gene_name" = "symbol"))
pm_left %>% head(1)
```

```
## # A tibble: 1 x 50
##   uniprot_id gene_name gene_name_alt protein_name protein_name_alt sequence
##   <chr>      <chr>     <chr>         <chr>        <chr>            <chr>
## 1 P04217     A1BG      <NA>          "Alpha-1B-g~ Alpha-1-B glyco~ MSMLVVF~
## # ... with 44 more variables: length <dbl>, mass <dbl>, training_dataset <chr>,
## #   human_gene_id <dbl>, mouse_ortholog_gene_id <dbl>, synonyms <chr>,
## #   description <chr>, ensembl_gene_id <chr>, protein_length <dbl>,
```

```
## #   target_p_score <dbl>, mito_domain_score <chr>,
## #   coexpression_gnf_n50_score <dbl>, pgc_induction_score <dbl>,
## #   yeast_mito_homolog_score <chr>, rickettsia_homolog_score <chr>,
## #   msms_score <chr>, mcarta2_score <dbl>, mcarta2_fdr <dbl>,
## #   mcarta2_list <dbl>, mcarta2_evidence <chr>, hg19_chromosome <fct>,
## #   hg19_start <dbl>, hg19_stop <dbl>, msms_num_tissues <dbl>,
## #   msms_num_peptides_unique <dbl>, msms_num_spectra <dbl>,
## #   msms_total_intensity <dbl>, msms_percent_coverage <dbl>, tissues <chr>,
## #   cerebrum_total_peak_intensity_log10 <dbl>,
## #   cerebellum_total_peak_intensity_log10 <dbl>,
## #   brainstem_total_peak_intensity_log10 <dbl>,
## #   spinalcord_total_peak_intensity_log10 <dbl>,
## #   kidney_total_peak_intensity_log10 <dbl>,
## #   liver_total_peak_intensity_log10 <dbl>,
## #   heart_total_peak_intensity_log10 <dbl>,
## #   skeletalmuscle_total_peak_intensity_log10 <dbl>,
## #   adipose_total_peak_intensity_log10 <dbl>,
## #   smallintestine_total_peak_intensity_log10 <dbl>,
## #   largeintestine_total_peak_intensity_log10 <dbl>,
## #   stomach_total_peak_intensity_log10 <dbl>,
## #   placenta_total_peak_intensity_log10 <dbl>,
## #   testis_total_peak_intensity_log10 <dbl>,
## #   hpa_primary_subcellular_localization_2015 <chr>
```

Now you have one dataset with additional useful information



`right_join`

Returns all rows from b, and all columns from a and b

Rows in b with no match in a will have NA values in the new columns

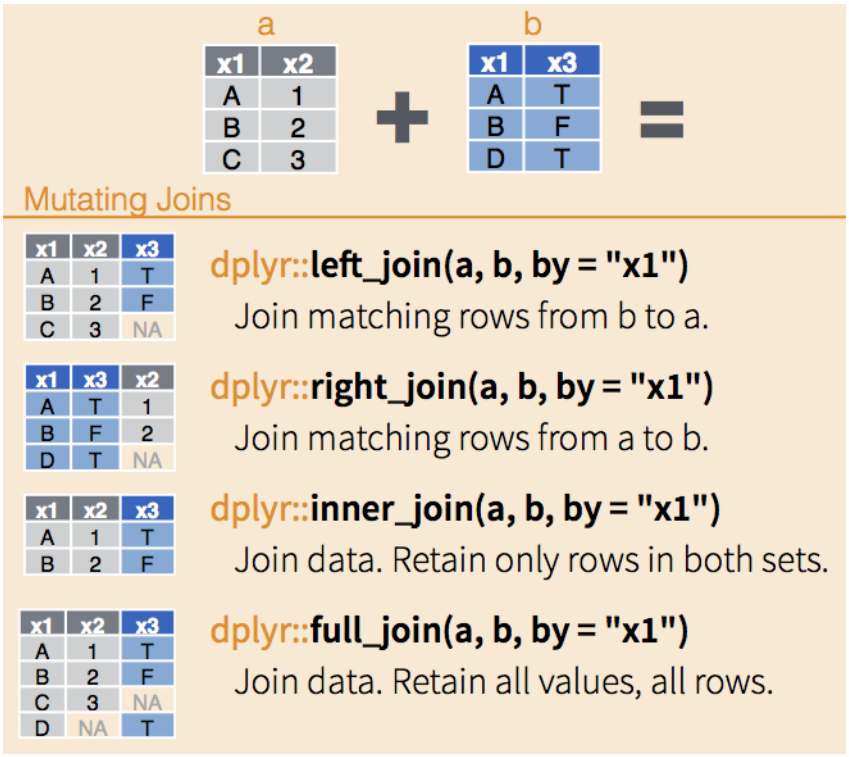
If there are multiple matches between a and b, all combinations of the matches are returned

`right_join` example

`right_join` `proteins` with `mitocarta` and assign the output to a new object called `pm_right`

```
pm_right <- right_join(proteins, mitocarta, by = c("gene_name" = "symbol"))
pm_right %>% head(1)
```

```
## # A tibble: 1 x 50
##   uniprot_id gene_name gene_name_alt protein_name protein_name_alt sequence
##   <chr>      <chr>     <chr>         <chr>        <chr>            <chr>
## 1 P04217     A1BG      <NA>          "Alpha-1B-g~ Alpha-1-B glyco~ MSMLVVF~
## # ... with 44 more variables: length <dbl>, mass <dbl>, training_dataset <chr>,
## #   human_gene_id <dbl>, mouse_ortholog_gene_id <dbl>, synonyms <chr>,
## #   description <chr>, ensembl_gene_id <chr>, protein_length <dbl>,
## #   target_p_score <dbl>, mito_domain_score <chr>,
## #   coexpression_gnf_n50_score <dbl>, pgc_induction_score <dbl>,
## #   yeast_mito_homolog_score <chr>, rickettsia_homolog_score <chr>,
## #   msms_score <chr>, mcarta2_score <dbl>, mcarta2_fdr <dbl>,
## #   mcarta2_list <dbl>, mcarta2_evidence <chr>, hg19_chromosome <fct>,
## #   hg19_start <dbl>, hg19_stop <dbl>, msms_num_tissues <dbl>,
## #   msms_num_peptides_unique <dbl>, msms_num_spectra <dbl>,
## #   msms_total_intensity <dbl>, msms_percent_coverage <dbl>, tissues <chr>,
## #   cerebrum_total_peak_intensity_log10 <dbl>,
## #   cerebellum_total_peak_intensity_log10 <dbl>,
## #   brainstem_total_peak_intensity_log10 <dbl>,
## #   spinalcord_total_peak_intensity_log10 <dbl>,
## #   kidney_total_peak_intensity_log10 <dbl>,
## #   liver_total_peak_intensity_log10 <dbl>,
## #   heart_total_peak_intensity_log10 <dbl>,
## #   skeletalmuscle_total_peak_intensity_log10 <dbl>,
## #   adipose_total_peak_intensity_log10 <dbl>,
## #   smallintestine_total_peak_intensity_log10 <dbl>,
## #   largeintestine_total_peak_intensity_log10 <dbl>,
## #   stomach_total_peak_intensity_log10 <dbl>,
## #   placenta_total_peak_intensity_log10 <dbl>,
## #   testis_total_peak_intensity_log10 <dbl>,
## #   hpa_primary_subcellular_localization_2015 <chr>
```



`inner_join`

Returns all rows from a where there are matching values in b, and all columns from a and b

If there are multiple matches between a and b, all combination of the matches are returned

`inner_join` example

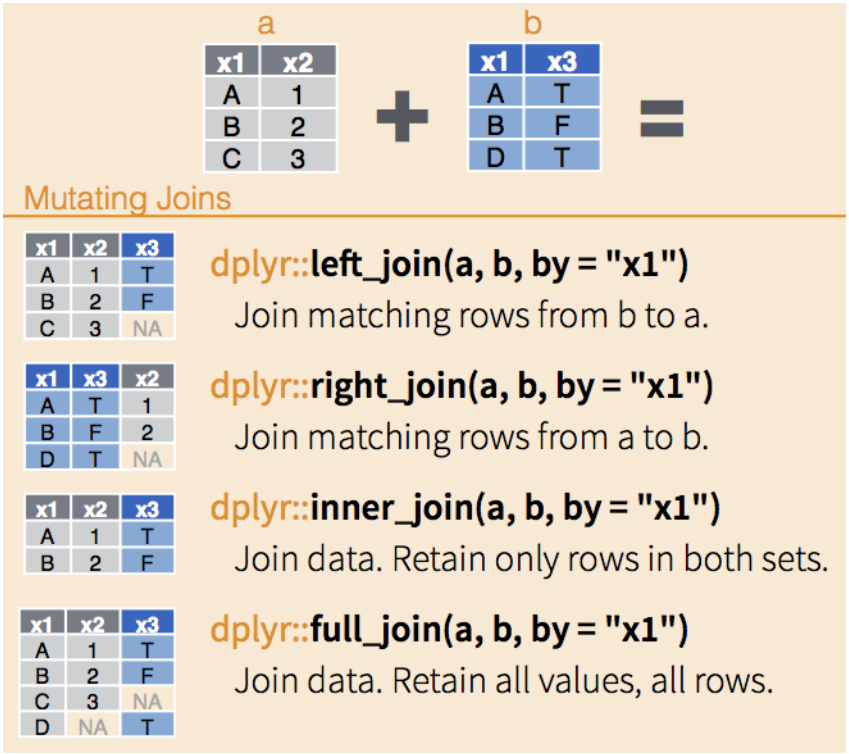
`inner_join` `proteins` with `mitocarta` and assign the output to a new object called `pm_inner`

```
pm_inner <- inner_join(proteins, mitocarta, by = c("gene_name" = "symbol"))
pm_inner %>% head(1)
```

```
## # A tibble: 1 x 50
##   uniprot_id gene_name gene_name_alt protein_name protein_name_alt sequence
##   <chr>      <chr>     <chr>         <chr>        <chr>            <chr>
## 1 P04217     A1BG      <NA>          "Alpha-1B-g~ Alpha-1-B glyco~ MSMLVVF~
## # ... with 44 more variables: length <dbl>, mass <dbl>, training_dataset <chr>,
## #   human_gene_id <dbl>, mouse_ortholog_gene_id <dbl>, synonyms <chr>,
## #   description <chr>, ensembl_gene_id <chr>, protein_length <dbl>,
## #   target_p_score <dbl>, mito_domain_score <chr>,
## #   coexpression_gnf_n50_score <dbl>, pgc_induction_score <dbl>,
## #   yeast_mito_homolog_score <chr>, rickettsia_homolog_score <chr>,
## #   msms_score <chr>, mcarta2_score <dbl>, mcarta2_fdr <dbl>,
## #   mcarta2_list <dbl>, mcarta2_evidence <chr>, hg19_chromosome <fct>,
## #   hg19_start <dbl>, hg19_stop <dbl>, msms_num_tissues <dbl>,
## #   msms_num_peptides_unique <dbl>, msms_num_spectra <dbl>,
## #   msms_total_intensity <dbl>, msms_percent_coverage <dbl>, tissues <chr>,
## #   cerebrum_total_peak_intensity_log10 <dbl>,
## #   cerebellum_total_peak_intensity_log10 <dbl>,
## #   brainstem_total_peak_intensity_log10 <dbl>,
```

```
## #   spinalcord_total_peak_intensity_log10 <dbl>,
## #   kidney_total_peak_intensity_log10 <dbl>,
## #   liver_total_peak_intensity_log10 <dbl>,
## #   heart_total_peak_intensity_log10 <dbl>,
## #   skeletalmuscle_total_peak_intensity_log10 <dbl>,
## #   adipose_total_peak_intensity_log10 <dbl>,
## #   smallintestine_total_peak_intensity_log10 <dbl>,
## #   largeintestine_total_peak_intensity_log10 <dbl>,
## #   stomach_total_peak_intensity_log10 <dbl>,
## #   placenta_total_peak_intensity_log10 <dbl>,
## #   testis_total_peak_intensity_log10 <dbl>,
## #   hpa_primary_subcellular_localization_2015 <chr>
```

Why might this type of join be useful?



`full_join`

Returns all rows and all columns from both a and b

Where there are no matching values, returns NA for the one missing

`full_join` example

`full_join` `proteins` with `mitocarta` and assign the output to a new object called `pm_full`

```
pm_full <- full_join(proteins, mitocarta, by = c("gene_name" = "symbol"))
pm_full %>% head(1)
```

```
## # A tibble: 1 x 50
##   uniprot_id gene_name gene_name_alt protein_name protein_name_alt sequence
```

```
##   <chr>      <chr>     <chr>         <chr>        <chr>            <chr>
## 1 P04217     A1BG      <NA>          "Alpha-1B-g~ Alpha-1-B glyco~ MSMLVVF~
## # ... with 44 more variables: length <dbl>, mass <dbl>, training_dataset <chr>,
## #   human_gene_id <dbl>, mouse_ortholog_gene_id <dbl>, synonyms <chr>,
## #   description <chr>, ensembl_gene_id <chr>, protein_length <dbl>,
## #   target_p_score <dbl>, mito_domain_score <chr>,
## #   coexpression_gnf_n50_score <dbl>, pgc_induction_score <dbl>,
## #   yeast_mito_homolog_score <chr>, rickettsia_homolog_score <chr>,
## #   msms_score <chr>, mcarta2_score <dbl>, mcarta2_fdr <dbl>,
## #   mcarta2_list <dbl>, mcarta2_evidence <chr>, hg19_chromosome <fct>,
## #   hg19_start <dbl>, hg19_stop <dbl>, msms_num_tissues <dbl>,
## #   msms_num_peptides_unique <dbl>, msms_num_spectra <dbl>,
## #   msms_total_intensity <dbl>, msms_percent_coverage <dbl>, tissues <chr>,
## #   cerebrum_total_peak_intensity_log10 <dbl>,
## #   cerebellum_total_peak_intensity_log10 <dbl>,
## #   brainstem_total_peak_intensity_log10 <dbl>,
## #   spinalcord_total_peak_intensity_log10 <dbl>,
## #   kidney_total_peak_intensity_log10 <dbl>,
## #   liver_total_peak_intensity_log10 <dbl>,
## #   heart_total_peak_intensity_log10 <dbl>,
## #   skeletalmuscle_total_peak_intensity_log10 <dbl>,
## #   adipose_total_peak_intensity_log10 <dbl>,
## #   smallintestine_total_peak_intensity_log10 <dbl>,
## #   largeintestine_total_peak_intensity_log10 <dbl>,
## #   stomach_total_peak_intensity_log10 <dbl>,
## #   placenta_total_peak_intensity_log10 <dbl>,
## #   testis_total_peak_intensity_log10 <dbl>,
## #   hpa_primary_subcellular_localization_2015 <chr>
```