Exploring data 2

Nesting and mapping

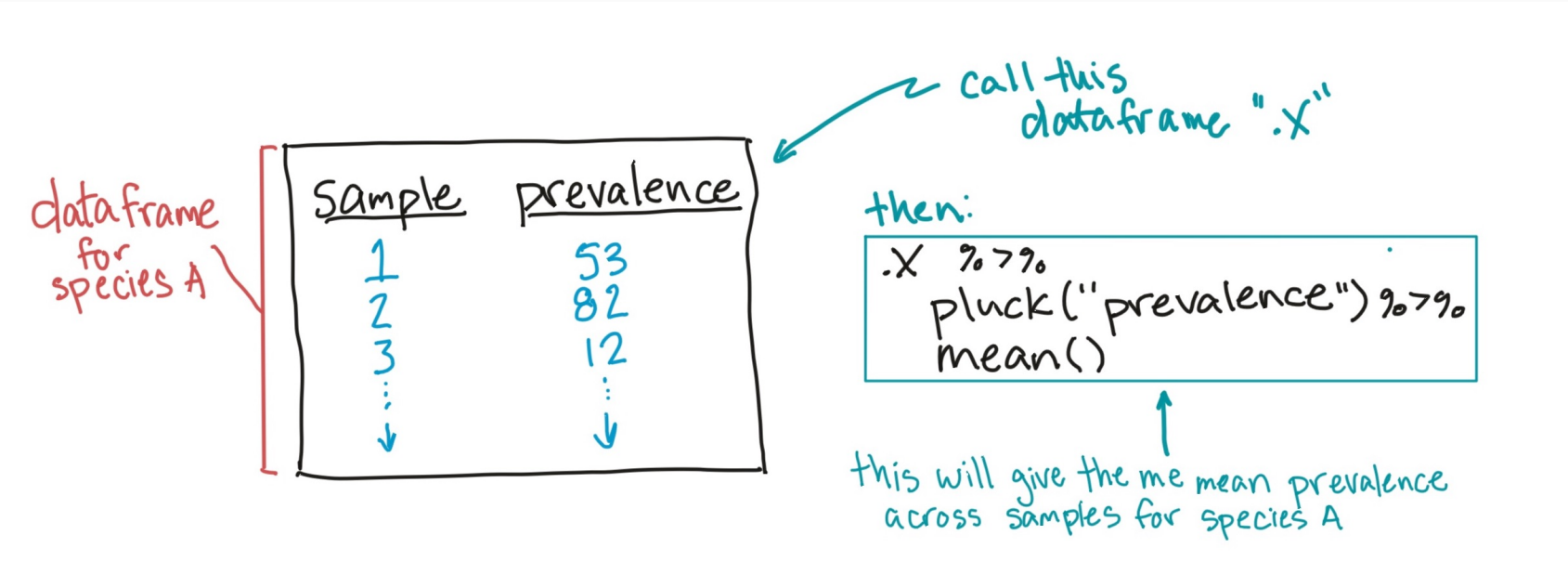
Nesting and mapping

The first step, with nesting and mapping, is to decide what you'd do to a subsample—the dataframe that you'd get if you filtered to the rows just for one grouping factor (for example, bacteria species).

Nesting and mapping

As a simple example, say you want to get the mean of each bacteria's prevalence across all samples.

Start by thinking about how you would calculate the mean prevalence for **one** species of bacteria if you had a subset of the dataframe rows for just that species.



Nesting and mapping

For example, say that you created a subset of the data that only had the rows for the species "Allistipes et rel.":

```
allistipes <- tidy_samples %>%
  filter(species == "Allistipes et rel.")

allistipes %>%
  slice(1:5)
```

```
## # A tibble: 5 x 3
##   species             sample   prevalence
##   <chr>               <chr>         <dbl>
## 1 Allistipes et rel. Sample-1         72
## 2 Allistipes et rel. Sample-2        127
## 3 Allistipes et rel. Sample-3         34
## 4 Allistipes et rel. Sample-4        344
## 5 Allistipes et rel. Sample-5         50
```

Nesting and mapping

You could determine the mean of prevalence by "pulling" the column measuring prevalence and then taking the mean of that vector:

```
allistipes %>%
  pull("prevalence") %>%
  mean()
```

```
## [1] 199.5248
```

Nesting and mapping

Once you've figured out this "recipe", you can **nest** the full dataframe by the grouping factor (e.g., bacteria species) and then **map** this recipe across the subsetted dataframe for each value of the grouping factor.

Nested dataframe

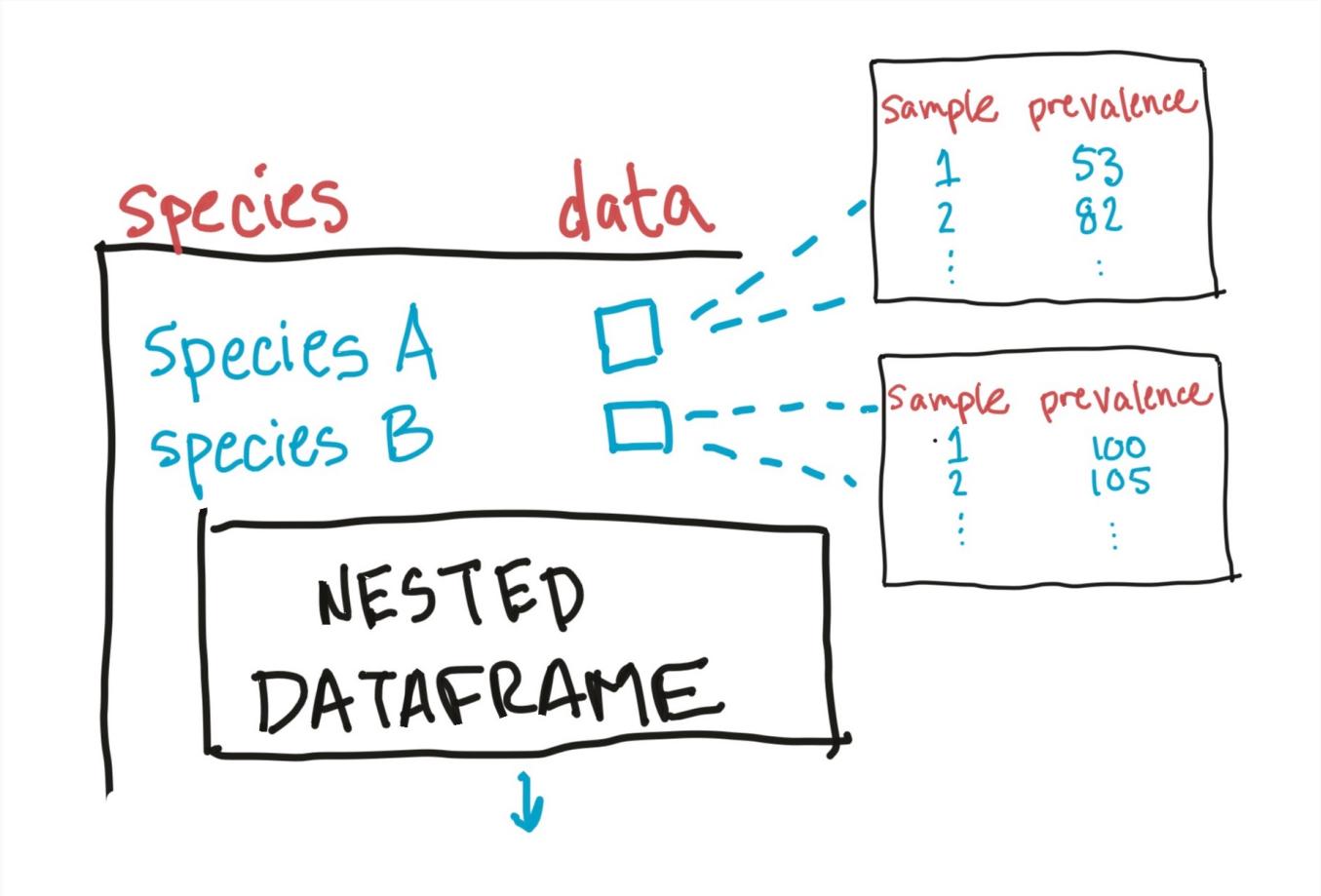
A nested dataframe is a fancy type of tibble.

For classic dataframes, each column must be a **vector**. For a nested dataframe, some of the columns can be **list-columns**, where each element is a more complex object than just a vector.

The elements in one of these list-columns can be a dataframe or a statistical model output object (or any other kind of list).

Nested dataframe

Here's an example where the list-column ("data") contains a dataframe for each bacterial species, with the prevalence measured for each sample for that bacteria.



Nested dataframe

Because a list-column packs in a lot more than a typical column, it will print out a little differently in R. For example, here the "data" column stores a dataframe for each bacteria sample:

```
## # A tibble: 3 x 2
## # Groups:   species [3]
##   species          data
##   <chr>            <list>
## 1 Actinomycetaceae <tibble [1,151 x 2]>
## 2 Aerococcus       <tibble [1,151 x 2]>
## 3 Aeromonas        <tibble [1,151 x 2]>
```

You can see that this element is a dataframe and its dimensions, but not values in it.