Introduction to R for Biologists

Tidyverse ecosystem & making data tidy with tidyr

*Several slides have been contributed by or modified from slides by Dr. Claus Wilke

Day 1 Outline

1. How to get set up using R
2. How and why to use RStudio & R Markdown (.Rmd)
3. Basics of programming
   - Data types
   - Functions
   - Troubleshooting
4. Intro to the Tidyverse
   - Tidy vs untidy data
   - Tidyverse-specific functions

Day 1 Outline

1. How to get set up using R
2. How and why to use RStudio & R Markdown (.Rmd)
3. Basics of programming
4. Intro to the Tidyverse
    - Tidy vs untidy data
    - Tidyverse-specific functions

Tidy data

“Tidy datasets are all alike but every messy dataset is messy in its own way” — Hadley Wickham

Tidy data

Three rules:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Example: Contingency table

	survived	died
drug	15	3
placebo	4	12

not tidy

Example: Contingency table

	survived	died
drug	15	3
placebo	4	12

not tidy

tidy

treatment	outcome	count
drug	survived	15
drug	died	3
placebo	survived	4
placebo	died	12

Example: Contingency table

	survived	died
drug	15	3
placebo	4	12

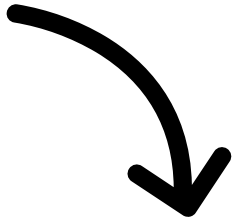
not tidy

tidy

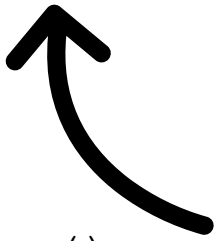
patient	treatment	outcome
1	drug	survived
2	drug	died
3	drug	survived
4	placebo	died
⋮	⋮	⋮

`tidyr` library provides functions for transforming tables

	survived	died
drug	15	3
placebo	4	12



`pivot_longer()`



`pivot_wider()`

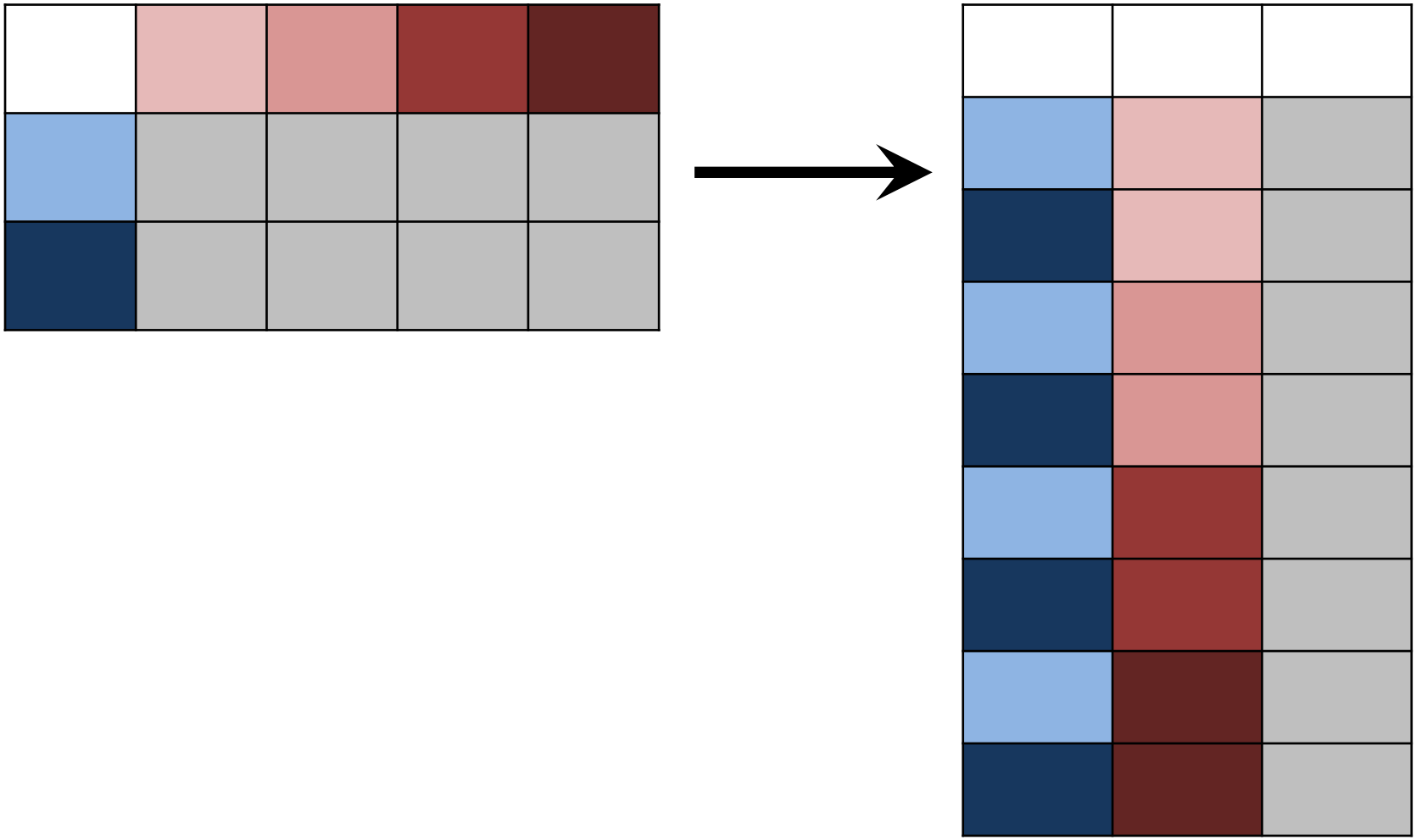
patient	treatment	outcome
1	drug	survived
2	drug	died
3	drug	survived
4	placebo	died
⋮		

Making data sets longer or wider

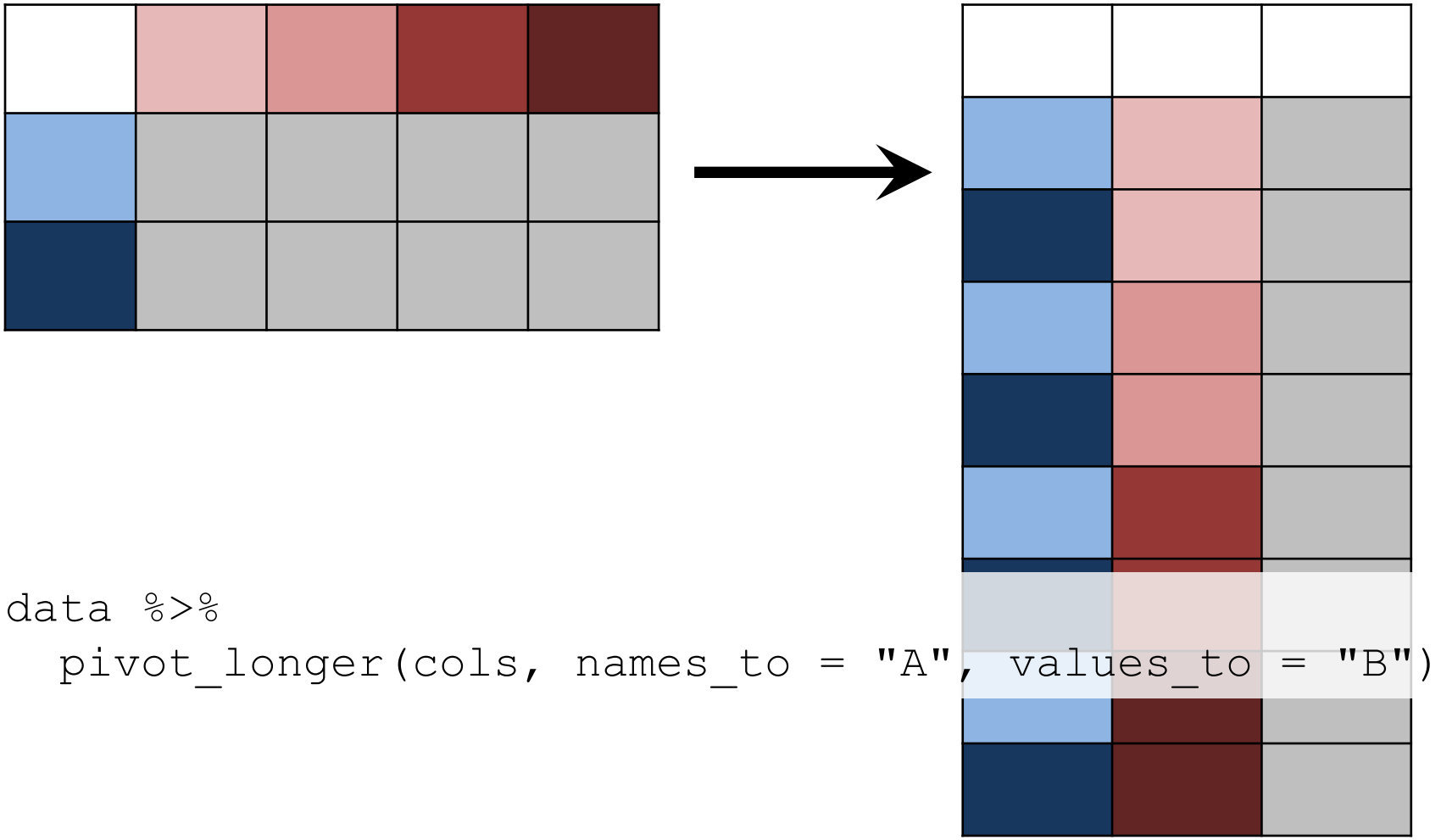
We'll be discussing two functions:

- `pivot_longer()` — make a wide table long
- `pivot_wider()` — make a long table wide

pivot_longer()

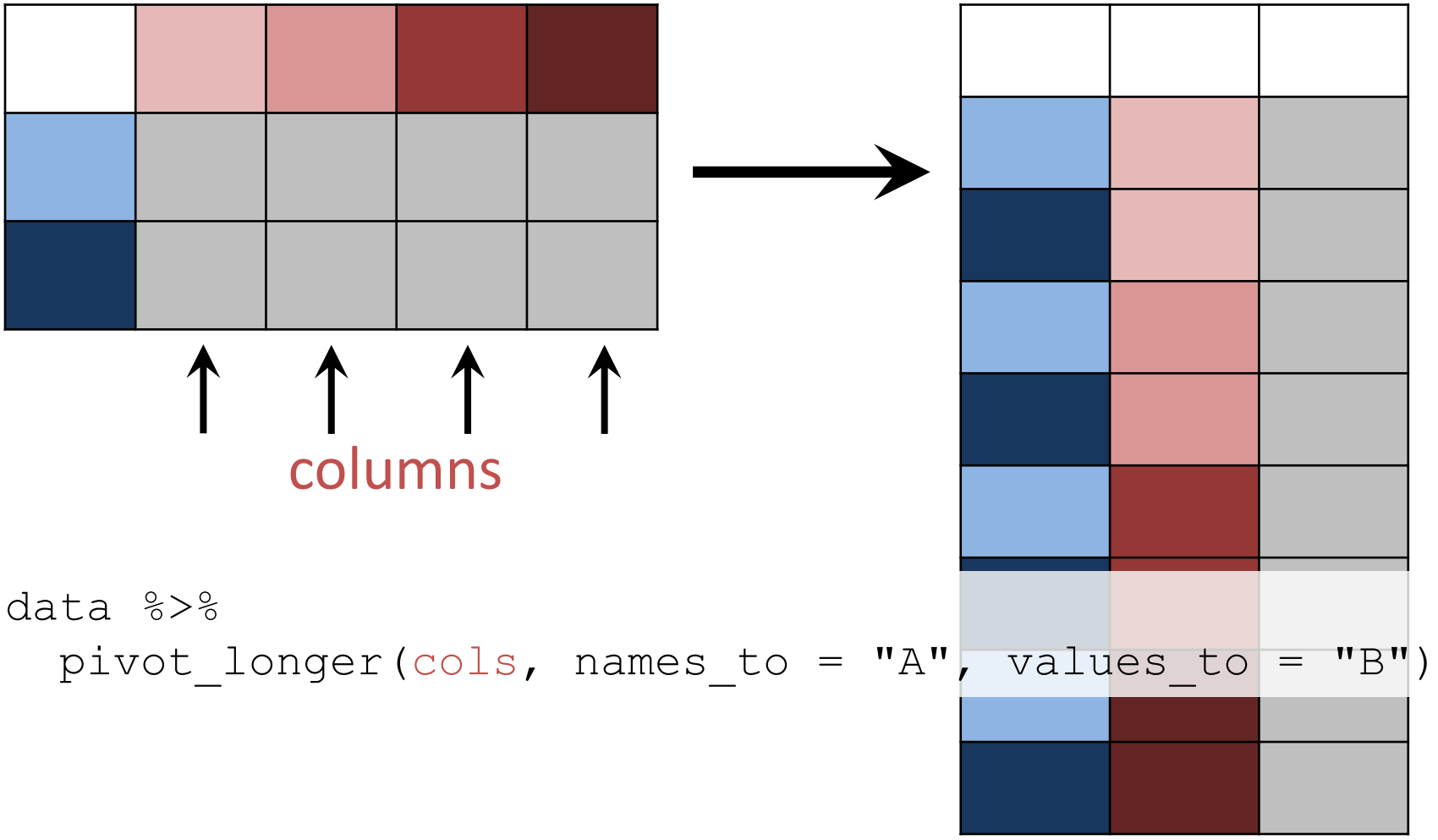


pivot_longer()



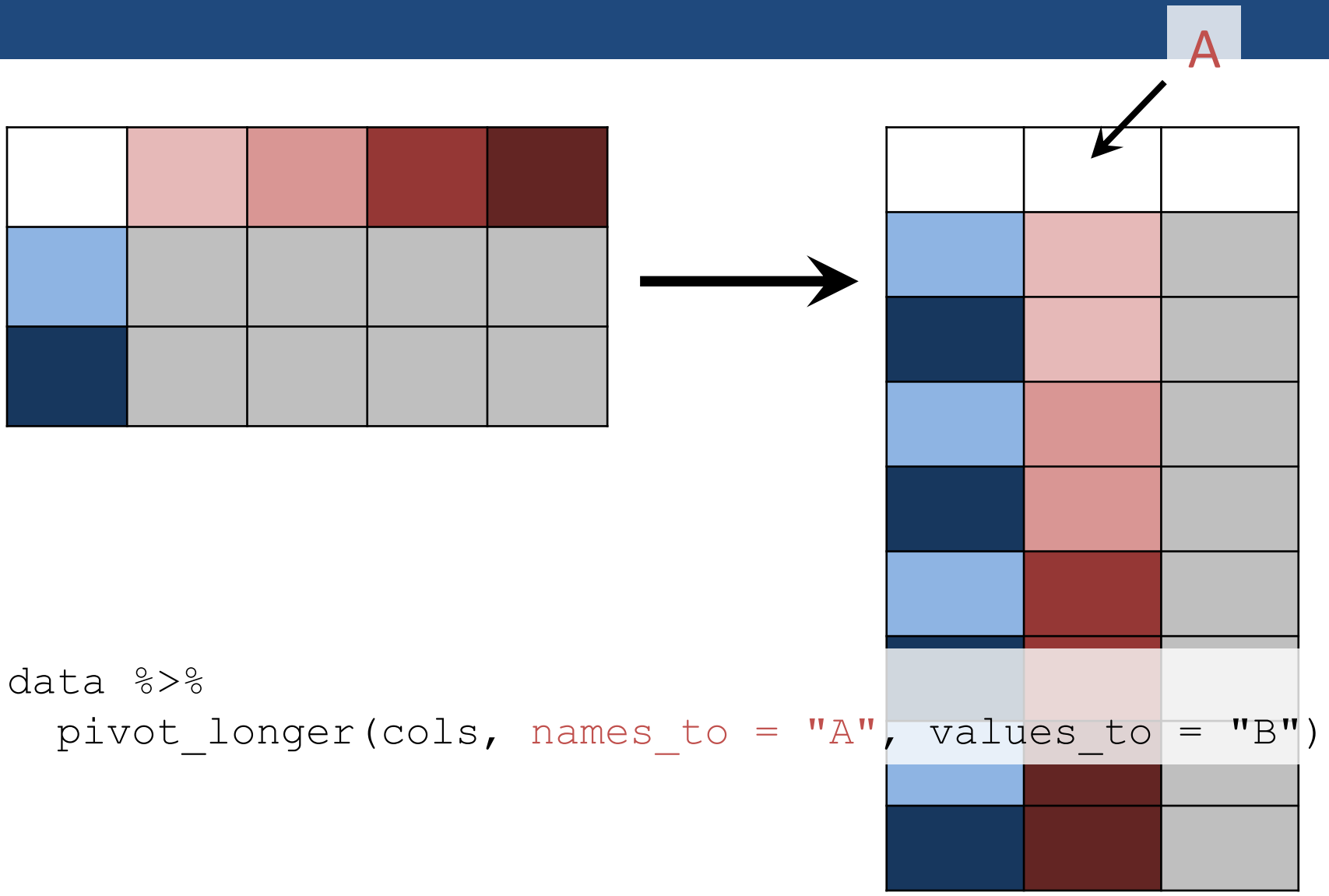
```
data %>%
  pivot_longer(cols, names_to = "A", values_to = "B")
```

pivot_longer()



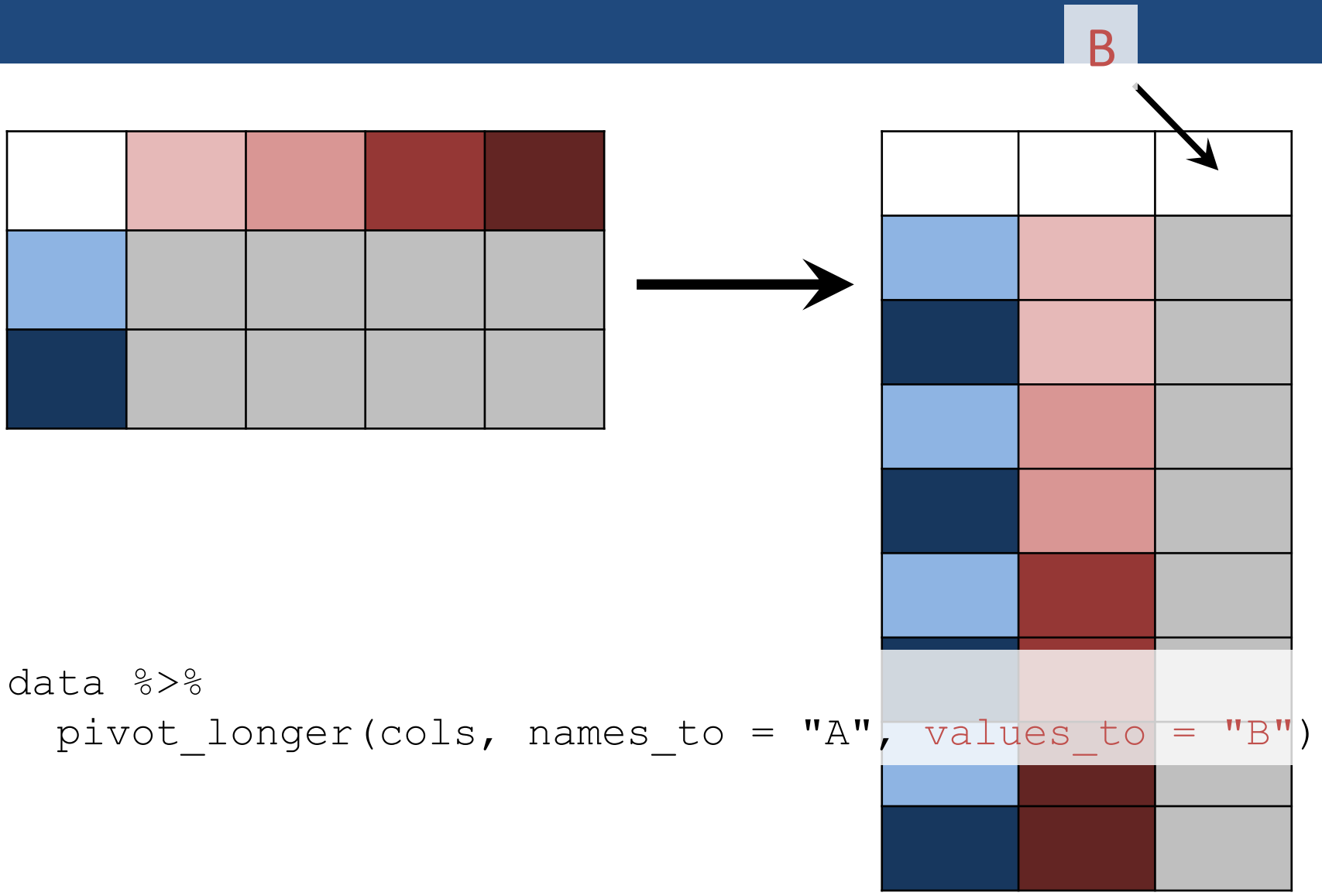

```
data %>%
  pivot_longer(cols, names_to = "A", values_to = "B")
```

pivot_longer()




```
data %>%
  pivot_longer(cols, names_to = "A", values_to = "B")
```

pivot_longer()




```
data %>%
  pivot_longer(cols, names_to = "A", values_to = "B")
```

Example: Let's recreate the sitka data from a wide table

```
> head(sitka_wide)
  tree treat t152 t174 t201 t227 t258
1    1 ozone 4.51 4.98 5.41 5.90 6.15
2    2 ozone 4.24 4.20 4.68 4.92 4.96
3    3 ozone 3.98 4.36 4.79 4.99 5.03
4    4 ozone 4.36 4.77 5.10 5.30 5.36
5    5 ozone 4.34 4.95 5.42 5.97 6.28
6    6 ozone 4.59 5.08 5.36 5.76 6.00
```

Example: Let's recreate the sitka data from a wide table

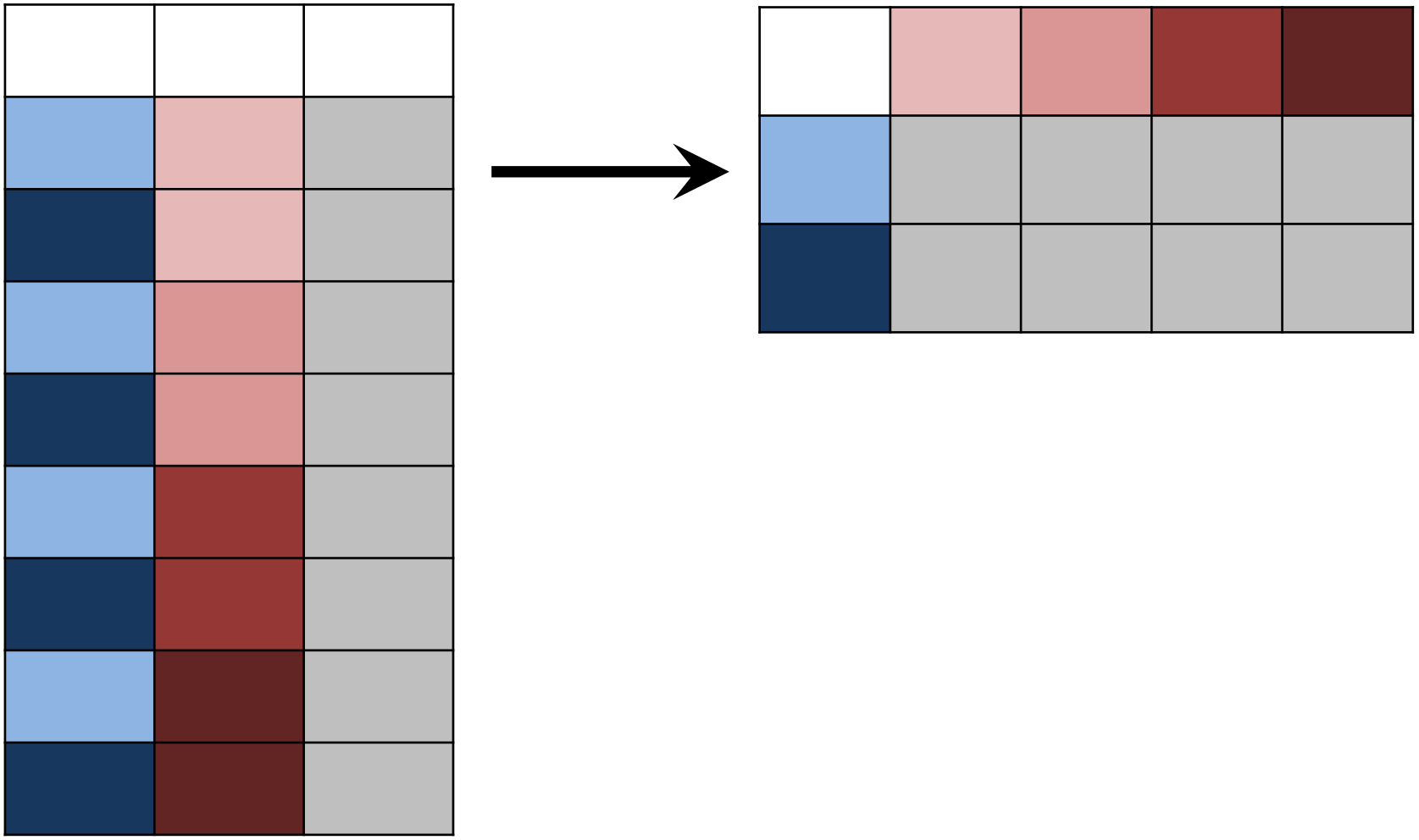
```
> head(sitka_wide)
  tree treat t152 t174 t201 t227 t258
1    1 ozone 4.51 4.98 5.41 5.90 6.15
2    2 ozone 4.24 4.20 4.68 4.92 4.96
3    3 ozone 3.98 4.36 4.79 4.99 5.03
4    4 ozone 4.36 4.77 5.10 5.30 5.36
5    5 ozone 4.34 4.95 5.42 5.97 6.28
6    6 ozone 4.59 5.08 5.36 5.76 6.00
```

```
sitka_wide %>%
  pivot_longer(
    t152:t258, names_to = "time", values_to = "size"
  )
```

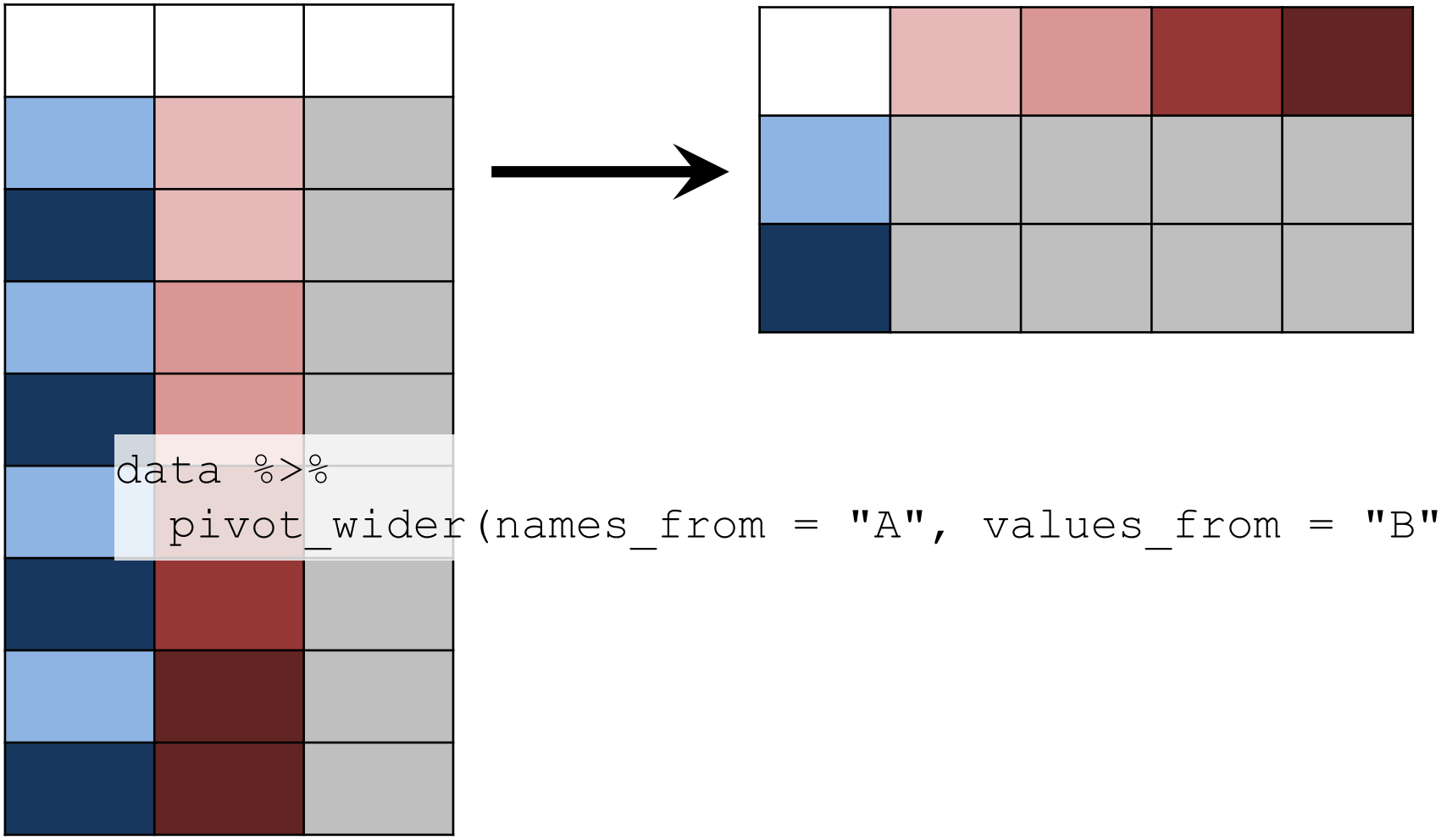
Example: Let's recreate the sitka data from a wide table

```
> sitka_wide %>%
  pivot_longer(
    t152:t258, names_to = "time", values_to = "size"
  )
# A tibble: 395 x 4
    tree treat time   size
   <int> <fct> <chr> <dbl>
 1     1 ozone t152   4.51
 2     1 ozone t174   4.98
 3     1 ozone t201   5.41
 4     1 ozone t227   5.9
 5     1 ozone t258   6.15
 6     2 ozone t152   4.24
 7     2 ozone t174   4.2
 8     2 ozone t201   4.68
 9     2 ozone t227   4.92
10     2 ozone t258   4.96
```

pivot_wider()

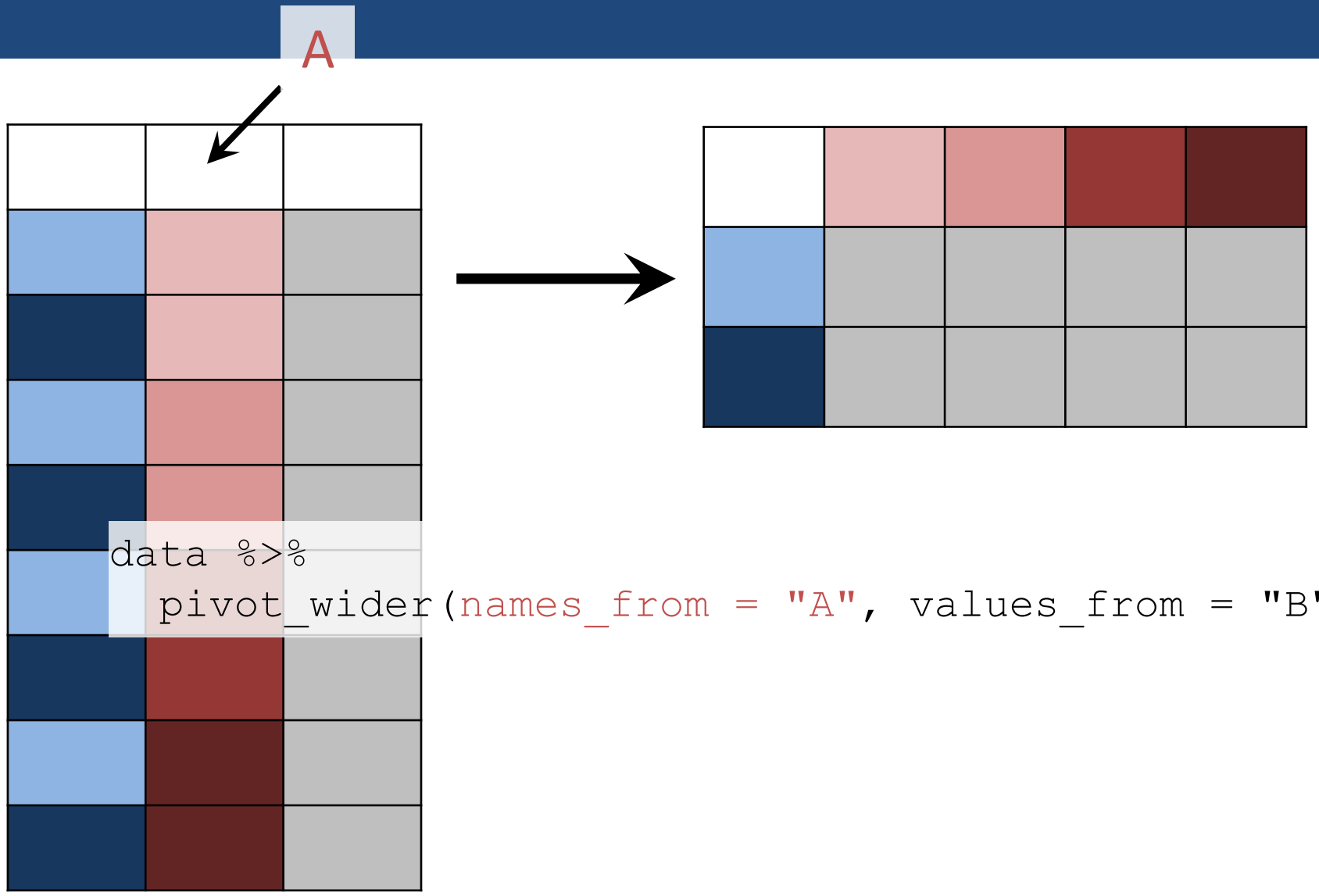


pivot_wider()



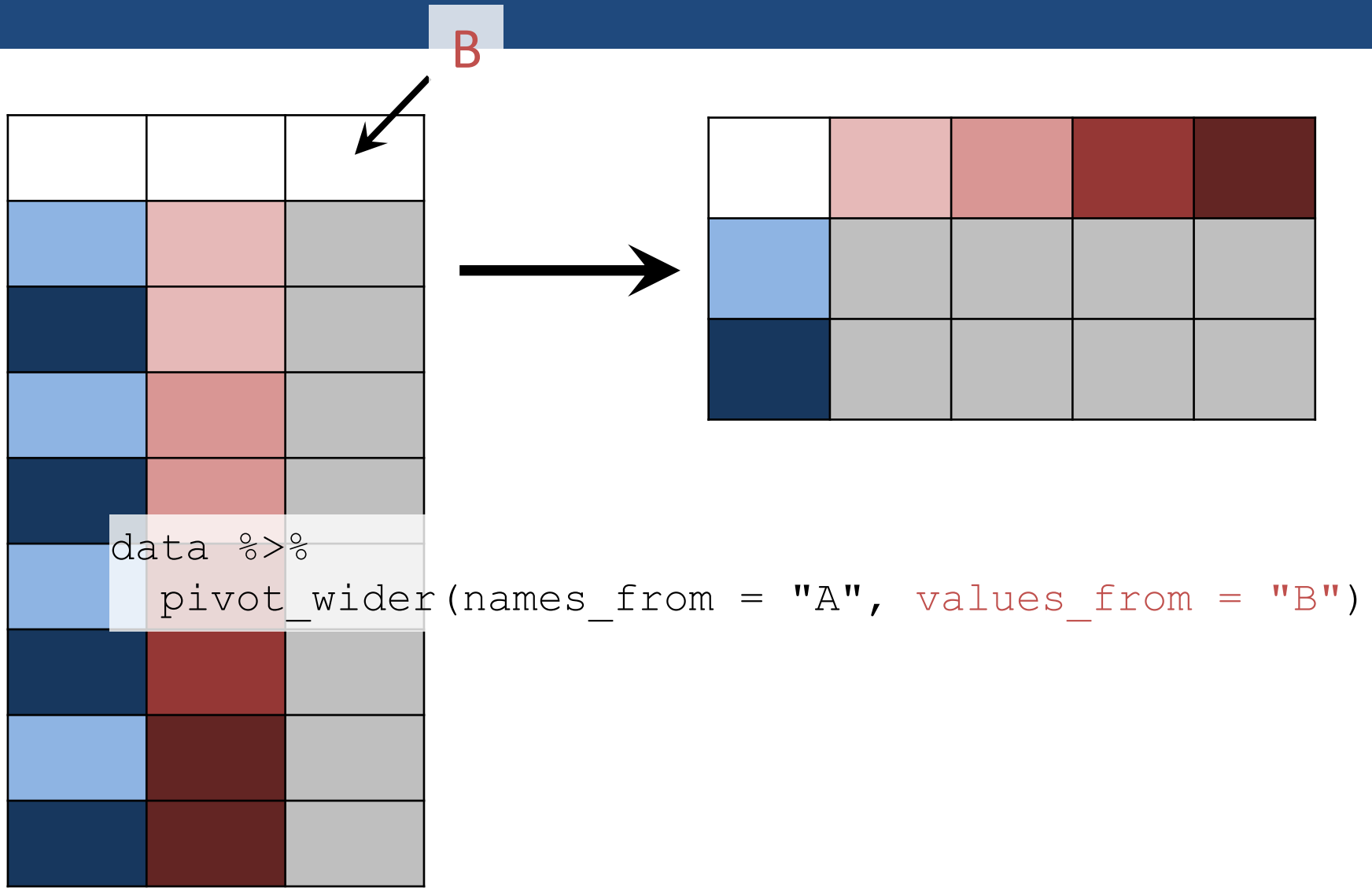
```
data %>%
  pivot_wider(names_from = "A", values_from = "B")
```

pivot_wider()




```
data %>%
  pivot_wider(names_from = "A", values_from = "B")
```

pivot_wider()




```
data %>%
  pivot_wider(names_from = "A", values_from = "B")
```

Example: Let's turn the sitka data into a wide table

```
> head(sitka)
   size Time tree treat
1 4.51   152    1 ozone
2 4.98   174    1 ozone
3 5.41   201    1 ozone
4 5.90   227    1 ozone
5 6.15   258    1 ozone
6 4.24   152    2 ozone
```

```
sitka %>%
  pivot_wider(names_from="Time", values_from="size")
```

Example: Let's turn the Sitka data into a wide table

```
> sitka %>%
  pivot_wider(names_from="Time", values_from="size")

# A tibble: 79 x 7
    tree treat `152` `174` `201` `227` `258`
   <int> <fct> <dbl> <dbl> <dbl> <dbl> <dbl>
 1     1 ozone  4.51  4.98  5.41  5.9   6.15
 2     2 ozone  4.24  4.2   4.68  4.92  4.96
 3     3 ozone  3.98  4.36  4.79  4.99  5.03
 4     4 ozone  4.36  4.77  5.1   5.3   5.36
 5     5 ozone  4.34  4.95  5.42  5.97  6.28
 6     6 ozone  4.59  5.08  5.36  5.76  6
 7     7 ozone  4.41  4.56  4.95  5.23  5.33
 8     8 ozone  4.24  4.64  4.95  5.38  5.48
 9     9 ozone  4.82  5.17  5.76  6.12  6.24
10    10 ozone  3.84  4.17  4.67  4.67  4.8
# … with 69 more rows
```

Working with tidy data in R: tidyverse

Fundamental actions on data tables:

- make new columns — `mutate()`
- combine tables, adding columns — `left_join()`
- combine tables, adding rows — `bind_rows()`
- choose rows — `filter()`
- choose columns — `select()`
- arrange rows — `arrange()`
- calculate summary statistics — `summarize()`
- work on groups of data — `group_by()`