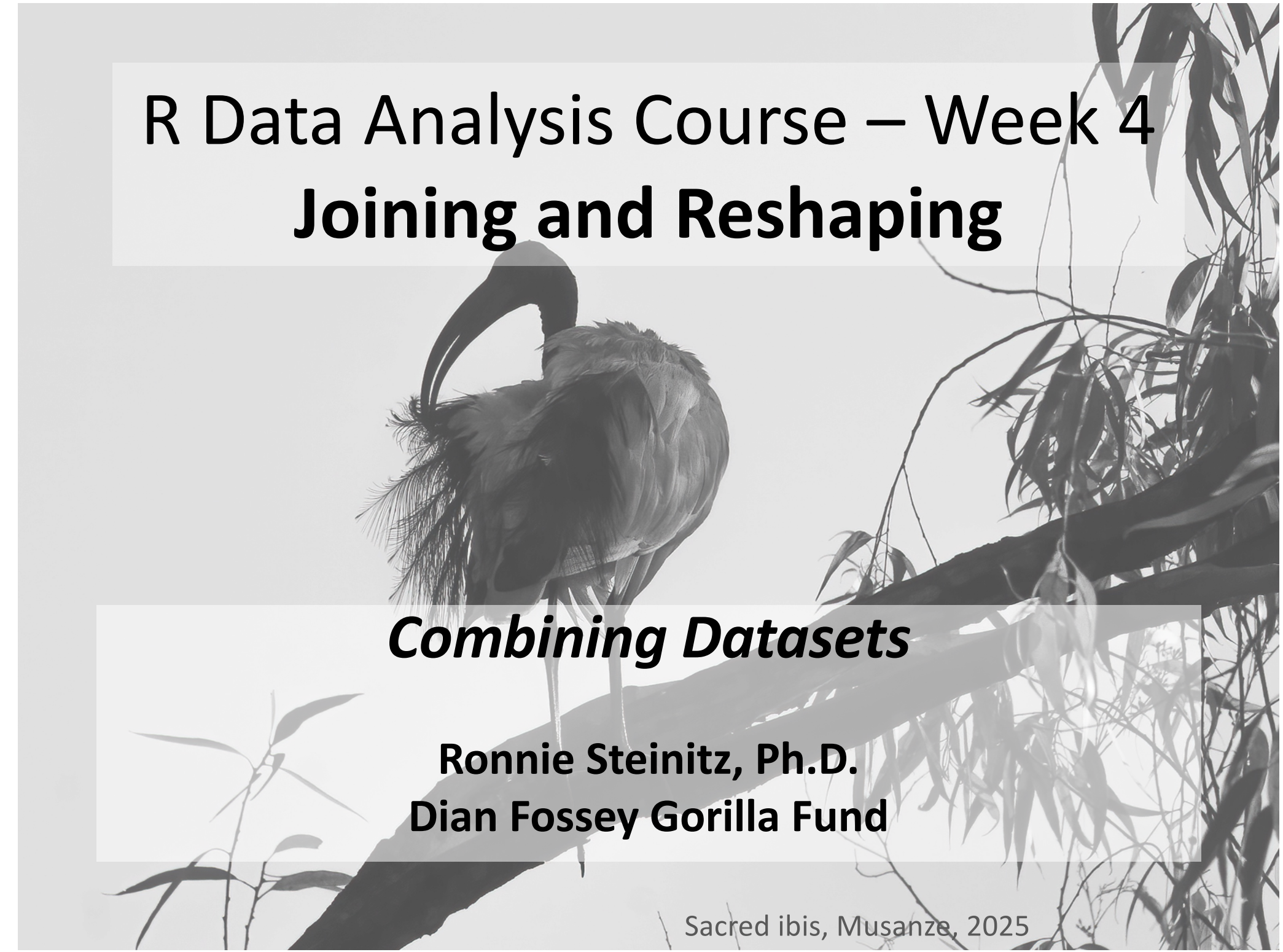
R Data Analysis Course – Week 4

Joining and Reshaping

Combining Datasets

Ronnie Steinitz, Ph.D.
Dian Fossey Gorilla Fund

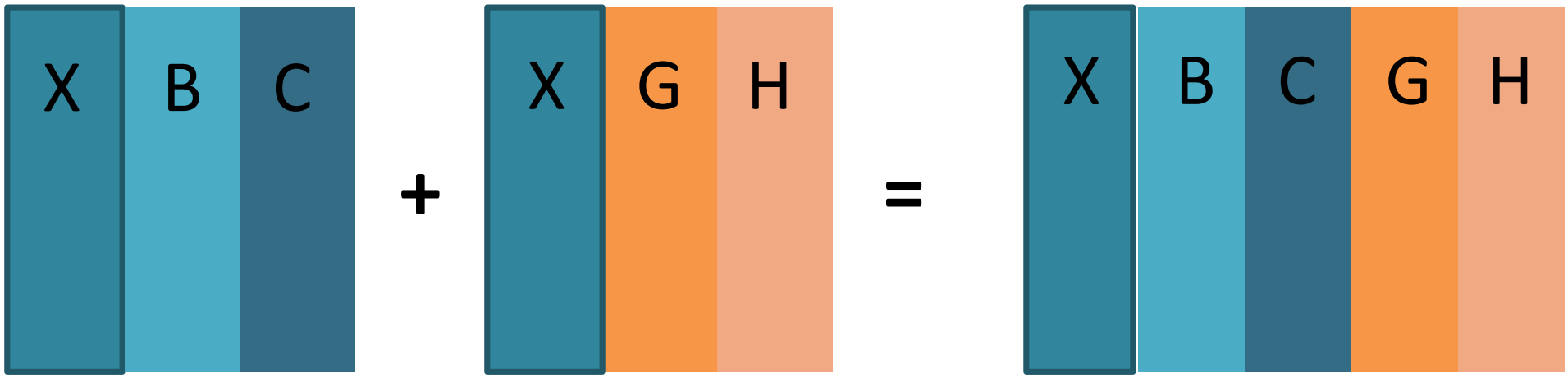


Sacred ibis, Musanze, 2025

Joining

- Combine columns from related tables
  - by keys, a variable of choice
- left_join( ) keeps all rows from the “left" table
- Example:

```
dataset1 %>% left_join(dataset2, by = ”common variable")
```



```
combined <- animals %>%
            left_join(species_weight, by = "species")
```

dataset1 = animals

species	speed_mps
Zebra	17
Lion	20
Hyena	14
Hyena	13
Gazelle	26
Lion	19
Zebra	18
Gazelle	23
Gazelle	25
Zebra	16
Lion	21
Gazelle	24
Hyena	15
Hyena	16

dataset2 = species_weight

species	mean_species_weight
Gazelle	*30*
Hyena	*60*
Lion	*160*
Zebra	*350*

```
combined <- animals %>%
    left_join(species_weight, by = "species")
```

species	speed_mps
Zebra	17
Lion	20
Hyena	14
Hyena	13
Gazelle	26
Lion	19
Zebra	18
Gazelle	23
Gazelle	25
Zebra	16
Lion	21
Gazelle	24
Hyena	15
Hyena	16

species	mean_species_weight
Gazelle	30
Hyena	60
Lion	160
Zebra	350

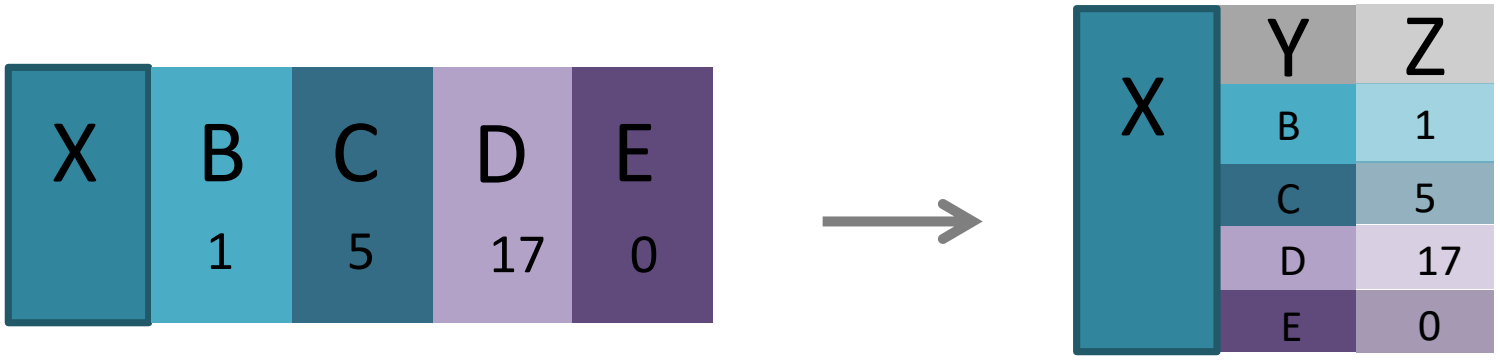
combined

species	speed_mps	mean_species_weight
Zebra	17	350
Lion	20	160
Hyena	14	60
Hyena	13	60
Gazelle	26	30
Lion	19	160
Zebra	18	350
Gazelle	23	30
Gazelle	25	30
Zebra	16	350
Lion	21	160
Gazelle	24	30
Hyena	15	60
Hyena	16	60

Reshaping

- Wide → Long so columns becomes a real variable
- **pivot_longer( )** choose columns to reshape
- Example:

```
data %>%
pivot_longer(cols = -X,   # all columns except X
             names_to = "Y",   # grouping variable
             values_to = "Z")   # value
```



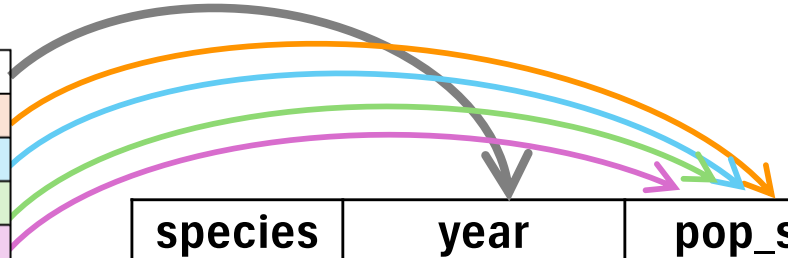
dataset 3 = species_census

species	site_A	site_B	site_C	site_D
Gazelle	502	564	613	648
Lion	23	28	35	27
Zebra	1127	1473	1285	1942
Hyena	48	52	51	49

Mean population size according to census at each site

dataset 3 = species_census

species	site_A	site_B	site_C	site_D
Gazelle	502	564	613	648
Lion	23	28	35	27
Zebra	1127	1473	1285	1942
Hyena	48	52	51	49



species	year	pop_size
Gazelle	site_A	502
	site_B	564
	site_C	613
	site_D	648
Lion	site_A	23
	site_B	28
	site_C	35
	site_D	27
Zebra	site_A	1127
	site_B	1473
	site_C	1285
	site_D	1942
Hyena	site_A	48
	site_B	52
	site_C	51
	site_D	49

Mean population size according to census at each site

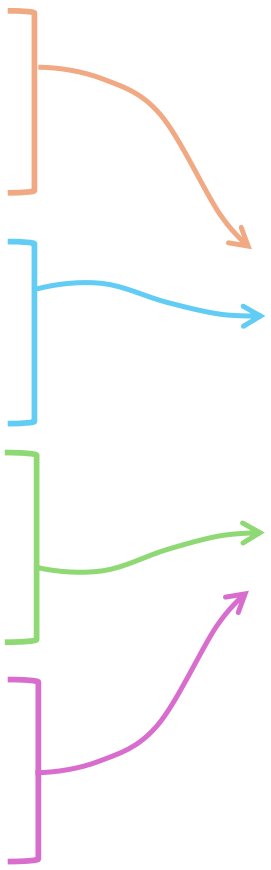
```
species_census %>%
    pivot_longer(cols = -species,
                 names_to = ”year",
                 values_to = ”pop_size")
```

dataset 3 = species_census

```
mean_pop <- species_census %>%
        group_by(species) %>%
        summarize(mean_pop_size = mean(pop_size))
```

species_census

species	year	pop_size
Gazelle	site_A	502
	site_B	564
	site_C	613
	site_D	648
Lion	site_A	23
	site_B	28
	site_C	35
	site_D	27
Zebra	site_A	1127
	site_B	1473
	site_C	1285
	site_D	1942
Hyena	site_A	48
	site_B	52
	site_C	51
	site_D	49



mean_pop

species	mean_pop_size
Gazelle	582
Lion	28
Zebra	1457
Hyena	50

Mean population size
across all sites

```
combined_new <- combined %>%
        left_join(mean_pop, by = "species")
```

mean_pop

species	mean_pop_size
Gazelle	582
Lion	28
Zebra	1457
Hyena	50

combined_new

combined

species	speed_mps	mean_species_weight	mean_pop_size
Zebra	17	350	1457
Lion	20	160	28
Hyena	14	60	50
Hyena	13	60	50
Gazelle	26	30	582
Lion	19	160	28
Zebra	18	350	1457
Gazelle	23	30	582
Gazelle	25	30	582
Zebra	16	350	1457
Lion	21	160	28
Gazelle	24	30	582
Hyena	15	60	50
Hyena	16	60	50

Skills Learning

code-along lecture

Village weaver, Musanze 2025

Skills Application

Laboratory exercise

Village weaver, Musanze 2025

Week 4 - Skills Application

Download **Skills Application Instructions** from *Google Drive folder* > *Week 4*

If you are working on **your own** dataset:

- Keep working on your ***Lastname_Firstname_Data.Rmd***

If you are working on **sample** dataset:

- create a Markdown file just for this week.
- Name it: ***Lastname_Firstname_Week3.Rmd*** **in `/Week 4`**

In your new Markdown file: Use code blocks (Ctrl+Alt+i) to load packages, import dataset into the environment, and save data as an object (e.g., `*data <-* `). Then follow ***instructions***.