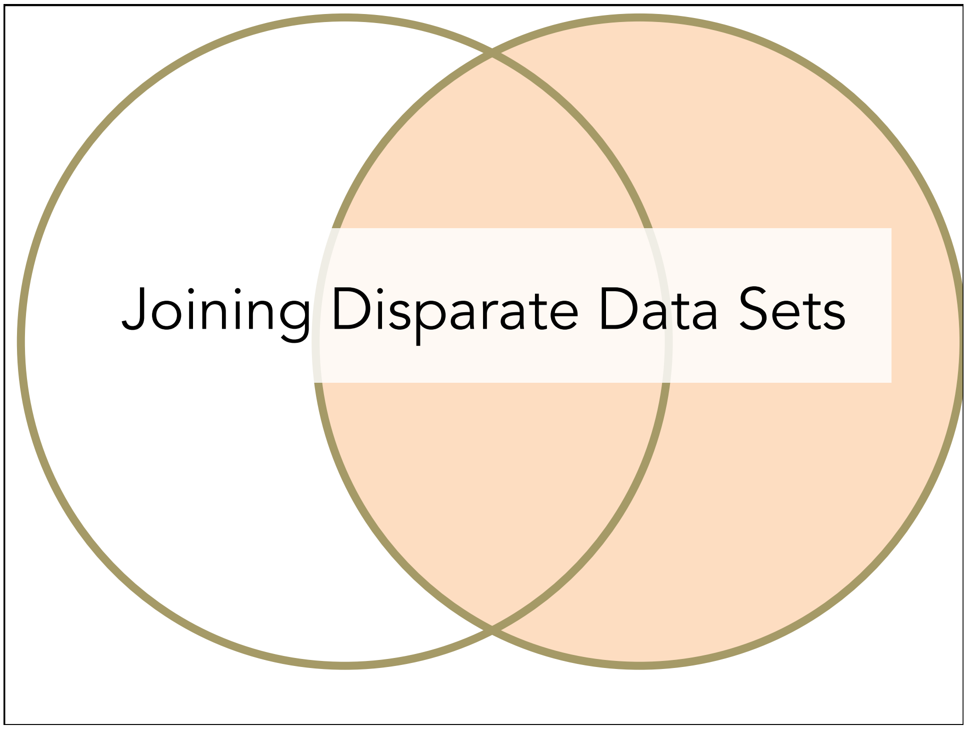
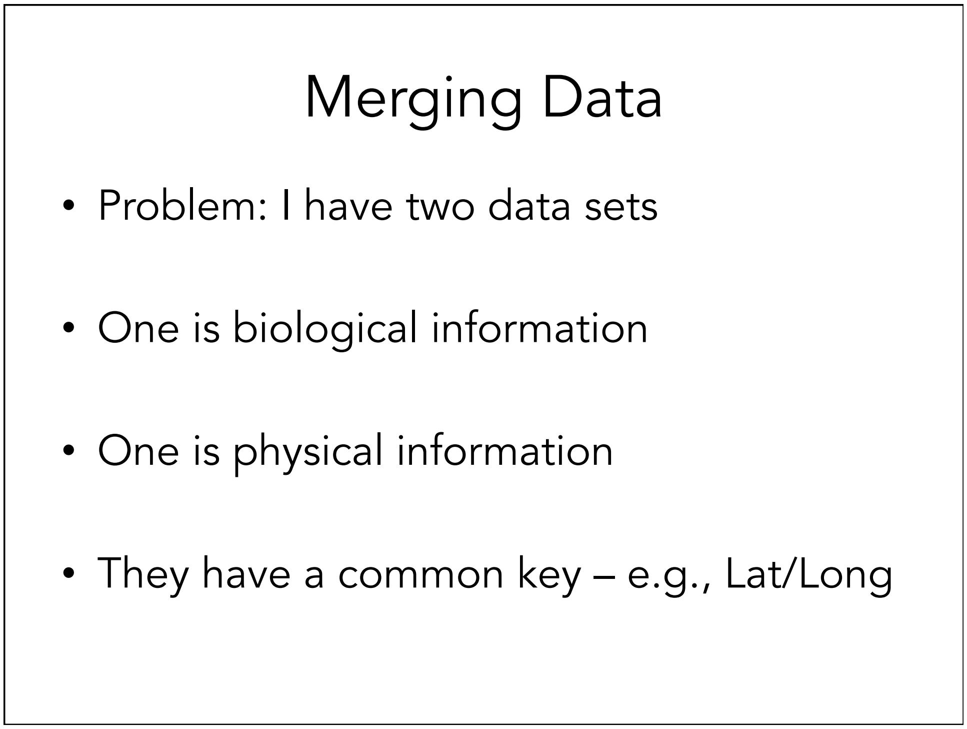
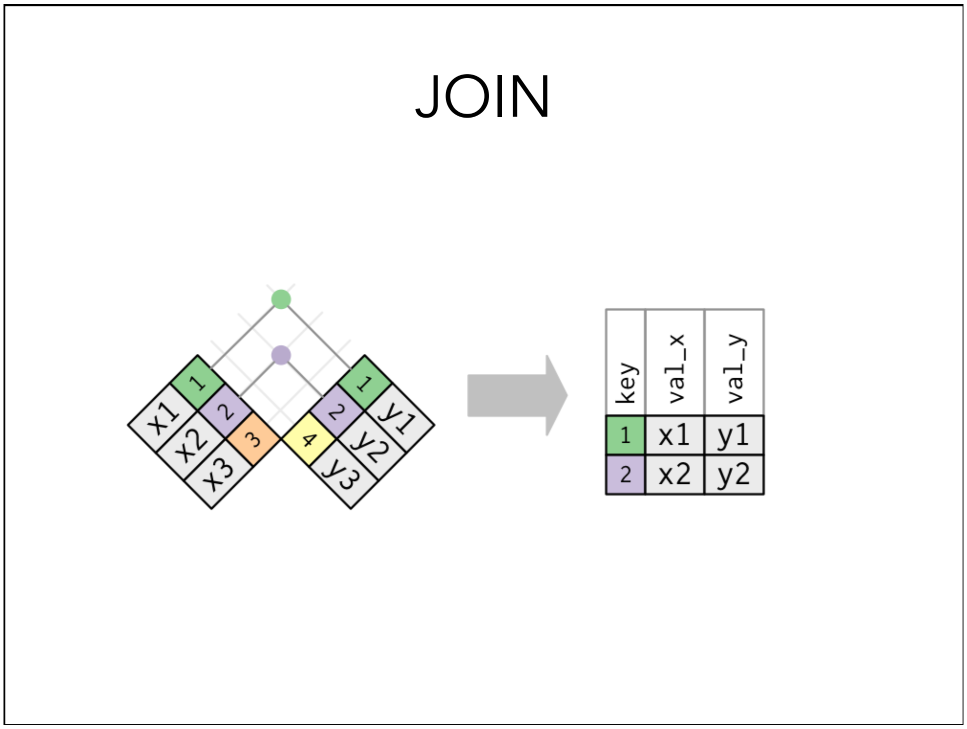


Joining Disparate Data Sets



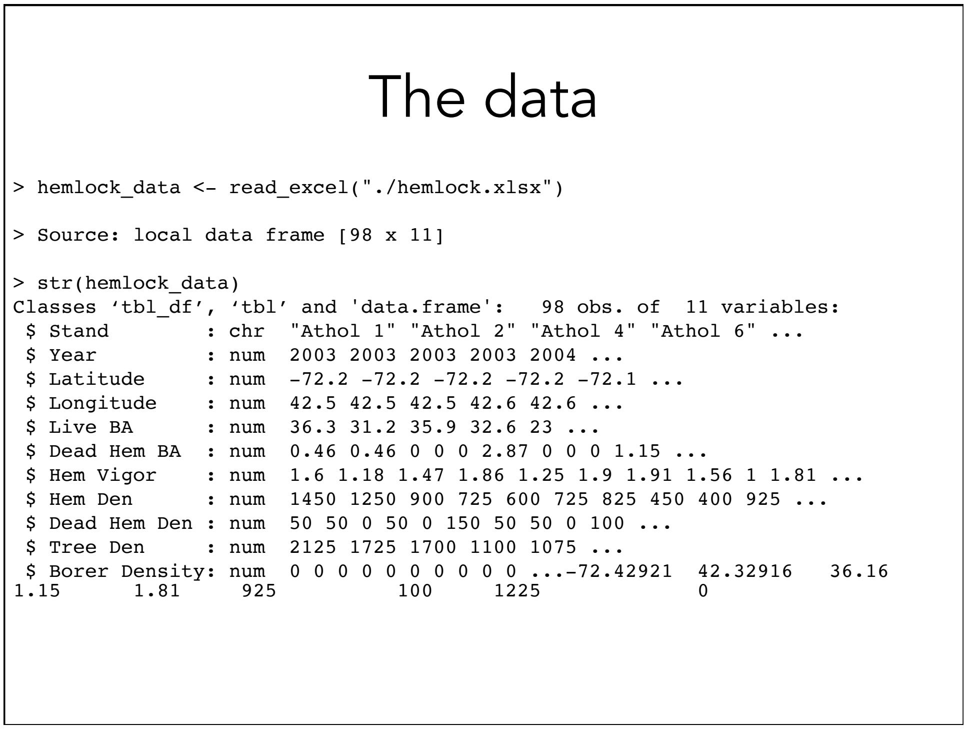
Merging Data

- Problem: I have two data sets
- One is biological information
- One is physical information
- They have a common key – e.g., Lat/Long



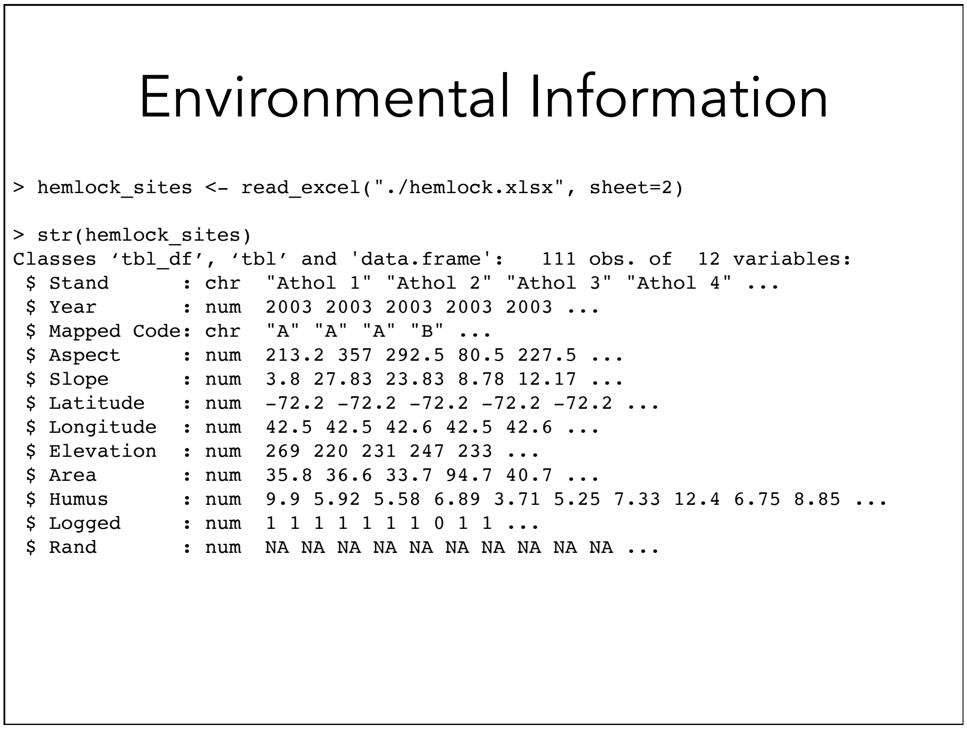
JOIN

key	val_x	val_y
1	x1	y1
2	x2	y2



The data

```
> hemlock_data <- read_excel("./hemlock.xlsx")

> Source: local data frame [98 x 11]

> str(hemlock_data)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':   98 obs. of  11 variables:
 $ Stand        : chr  "Athol 1" "Athol 2" "Athol 4" "Athol 6" ...
 $ Year         : num  2003 2003 2003 2003 2004 ...
 $ Latitude     : num  -72.2 -72.2 -72.2 -72.2 -72.1 ...
 $ Longitude    : num  42.5 42.5 42.5 42.6 42.6 ...
 $ Live BA      : num  36.3 31.2 35.9 32.6 23 ...
 $ Dead Hem BA  : num  0.46 0.46 0 0 0 2.87 0 0 0 1.15 ...
 $ Hem Vigor    : num  1.6 1.18 1.47 1.86 1.25 1.9 1.91 1.56 1 1.81 ...
 $ Hem Den      : num  1450 1250 900 725 600 725 825 450 400 925 ...
 $ Dead Hem Den : num  50 50 0 50 0 150 50 50 0 100 ...
 $ Tree Den     : num  2125 1725 1700 1100 1075 ...
 $ Borer Density: num  0 0 0 0 0 0 0 0 0 0 ...-72.42921  42.32916   36.16
1.15      1.81     925          100     1225            0
```



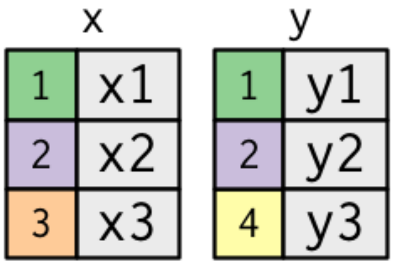
Environmental Information

```
> hemlock_sites <- read_excel("./hemlock.xlsx", sheet=2)

> str(hemlock_sites)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':   111 obs. of  12 variables:
 $ Stand      : chr  "Athol 1" "Athol 2" "Athol 3" "Athol 4" ...
 $ Year       : num  2003 2003 2003 2003 2003 ...
 $ Mapped Code: chr  "A" "A" "A" "B" ...
 $ Aspect     : num  213.2 357 292.5 80.5 227.5 ...
 $ Slope      : num  3.8 27.83 23.83 8.78 12.17 ...
 $ Latitude   : num  -72.2 -72.2 -72.2 -72.2 -72.2 ...
 $ Longitude  : num  42.5 42.5 42.6 42.5 42.6 ...
 $ Elevation  : num  269 220 231 247 233 ...
 $ Area       : num  35.8 36.6 33.7 94.7 40.7 ...
 $ Humus      : num  9.9 5.92 5.58 6.89 3.71 5.25 7.33 12.4 6.75 8.85 ...
 $ Logged     : num  1 1 1 1 1 1 1 0 1 1 ...
 $ Rand       : num  NA NA NA NA NA NA NA NA NA NA ...
```

The problem

```
> nrow(hemlock_data)
[1] 98

> nrow(hemlock_sites)
[1] 111
```

Mismatched Data Sets with Common Keys



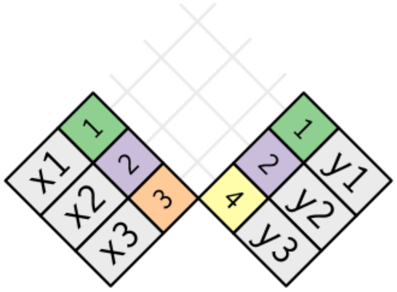
x

1	x1
2	x2
3	x3

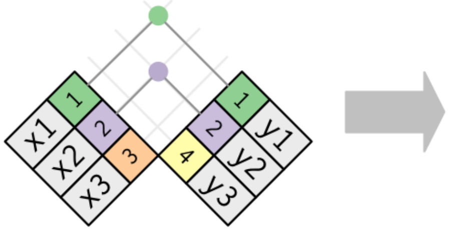
y

1	y1
2	y2
4	y3

Mismatched Data Sets with Common Keys

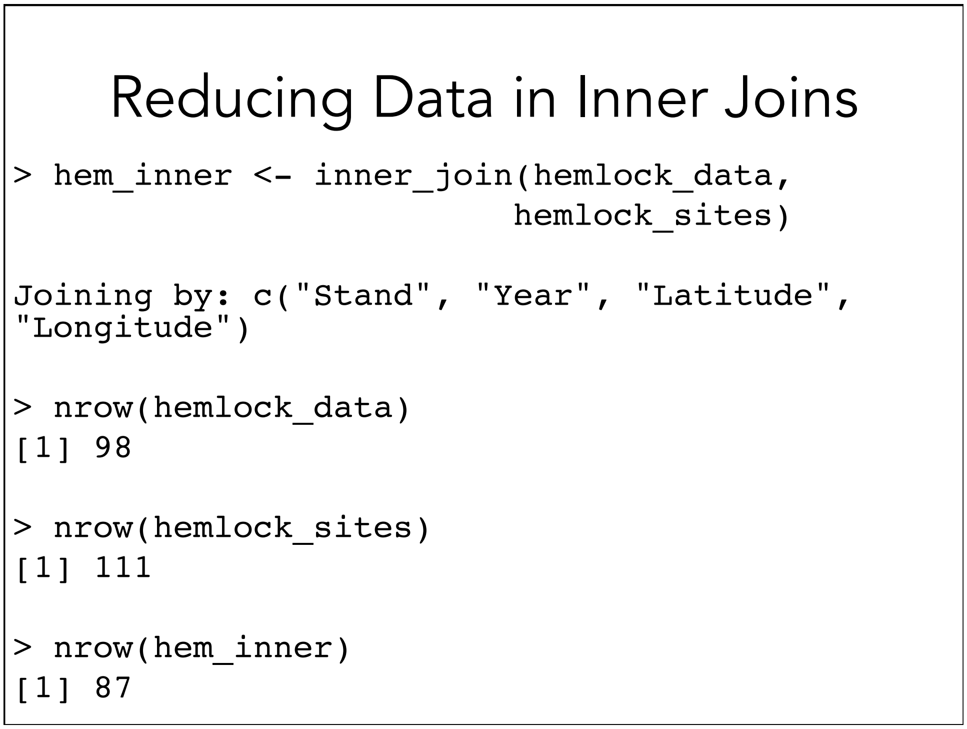


Inner Join



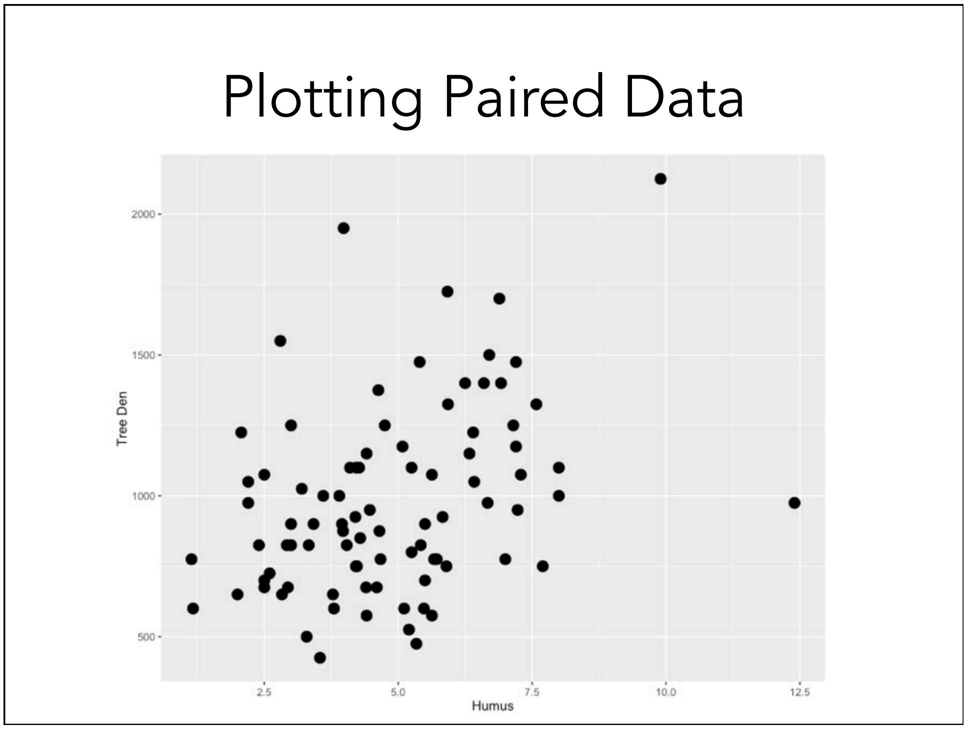
key	val1_x	val1_y
1	x1	y1
2	x2	y2

Creates new Data with rows that exist in both data sets



Reducing Data in Inner Joins

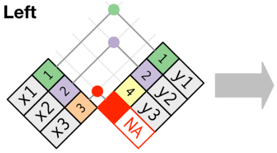
```
> hem_inner <- inner_join(hemlock_data,
                          hemlock_sites)

Joining by: c("Stand", "Year", "Latitude",
"Longitude")

> nrow(hemlock_data)
[1] 98

> nrow(hemlock_sites)
[1] 111

> nrow(hem_inner)
[1] 87
```



Plotting Paired Data

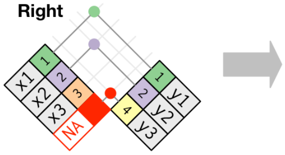
Outer Joins

Left



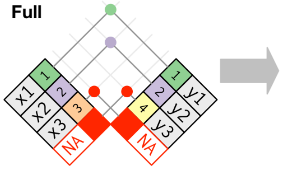
key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA

Right



key	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

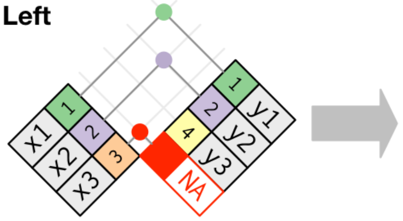
Full



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

Left Join: Retain Rows with NAs in First Dataset

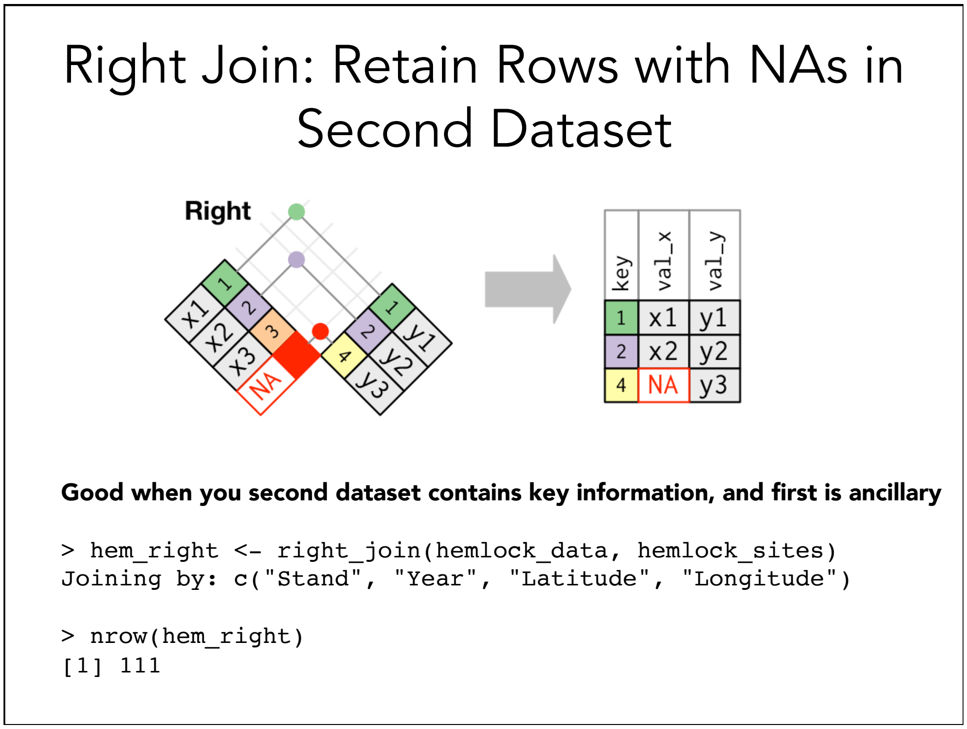
Left



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA

Good when too much descriptive information available

```
> hem_left <- left_join(hemlock_data, hemlock_sites)
Joining by: c("Stand", "Year", "Latitude", "Longitude")

> nrow(hem_left)
[1] 98
```

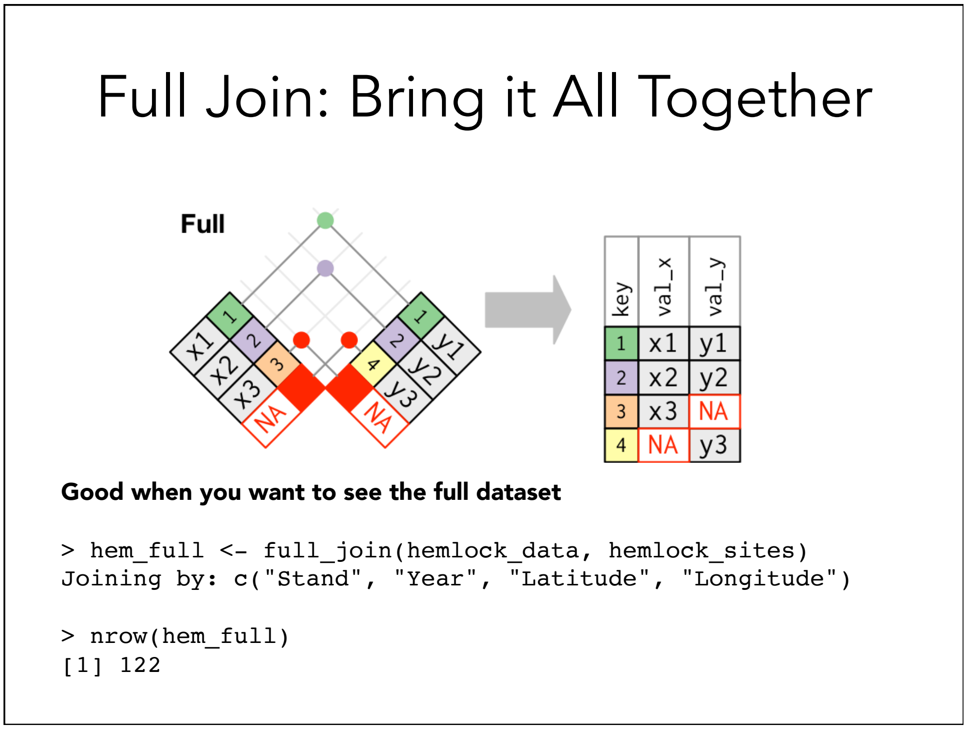


Right Join: Retain Rows with NAs in Second Dataset

key	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

Good when you second dataset contains key information, and first is ancillary

```
> hem_right <- right_join(hemlock_data, hemlock_sites)
Joining by: c("Stand", "Year", "Latitude", "Longitude")

> nrow(hem_right)
[1] 111
```



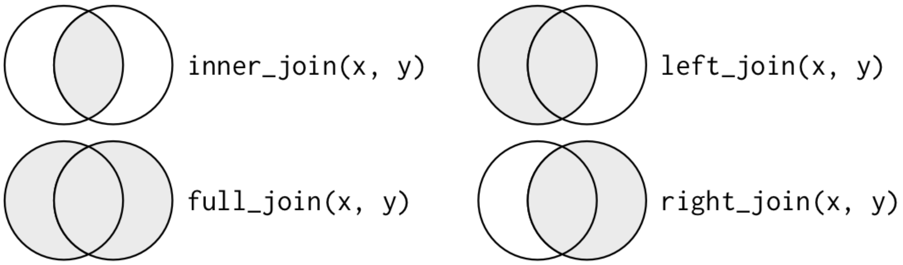
Full Join: Bring it All Together

key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

Good when you want to see the full dataset

```
> hem_full <- full_join(hemlock_data, hemlock_sites)
Joining by: c("Stand", "Year", "Latitude", "Longitude")

> nrow(hem_full)
[1] 122
```

The Joins



inner_join(x, y)

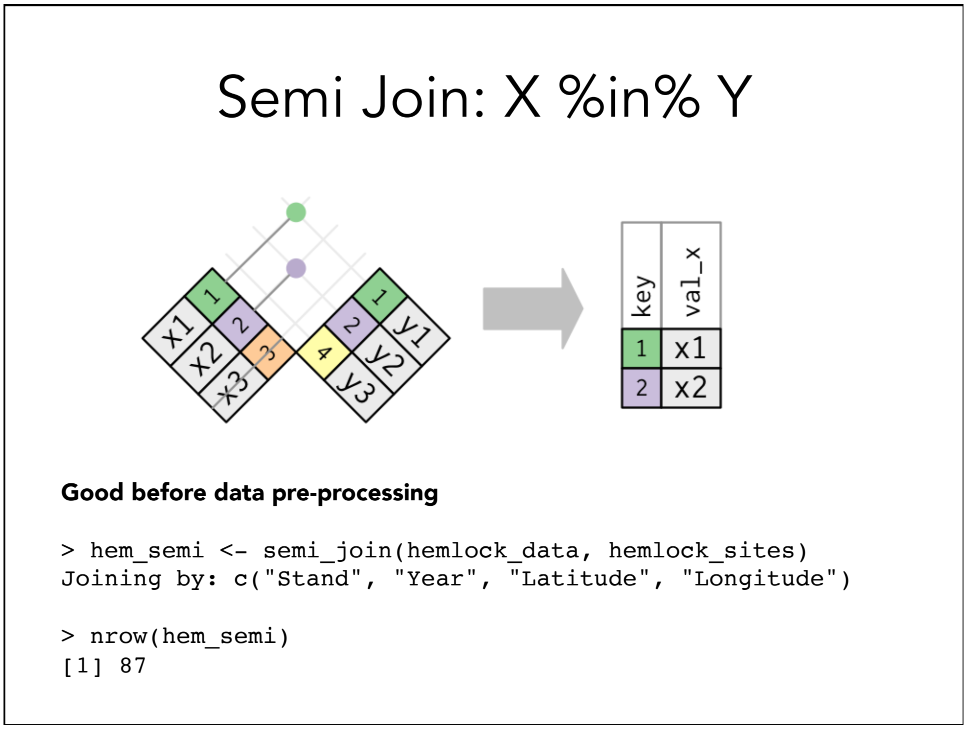
left_join(x, y)

full_join(x, y)

right_join(x, y)

Filtering Joins

- I only want data that matches a set of criteria
- Like outer joins with a second na.omit step

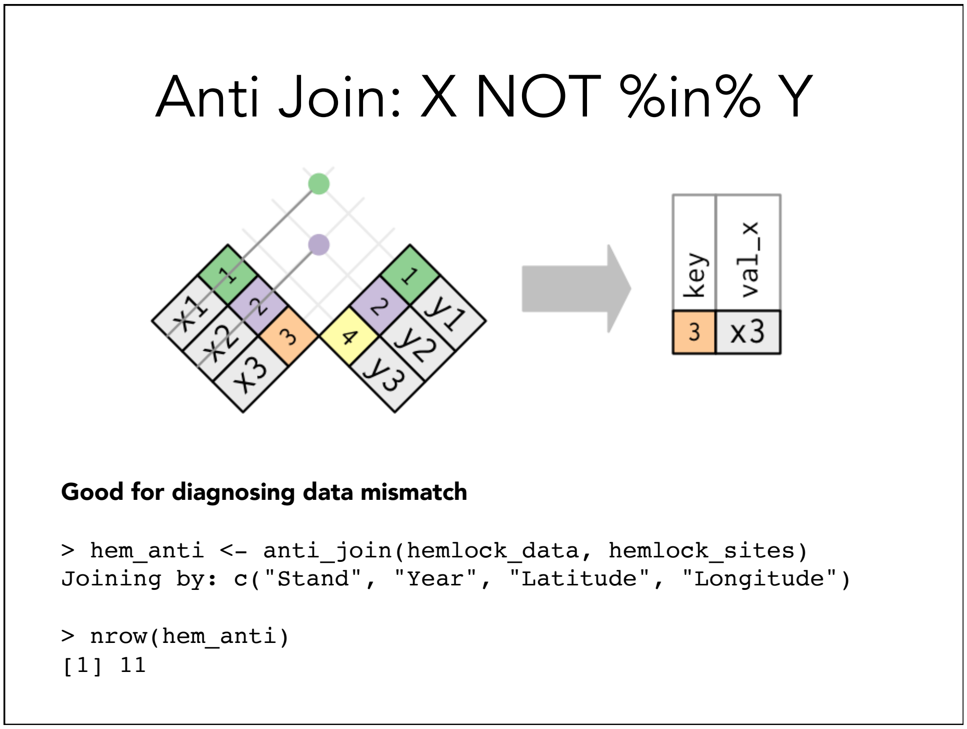


Semi Join: X %in% Y

key	val_x
1	x1
2	x2

Good before data pre-processing

```
> hem_semi <- semi_join(hemlock_data, hemlock_sites)
Joining by: c("Stand", "Year", "Latitude", "Longitude")

> nrow(hem_semi)
[1] 87
```

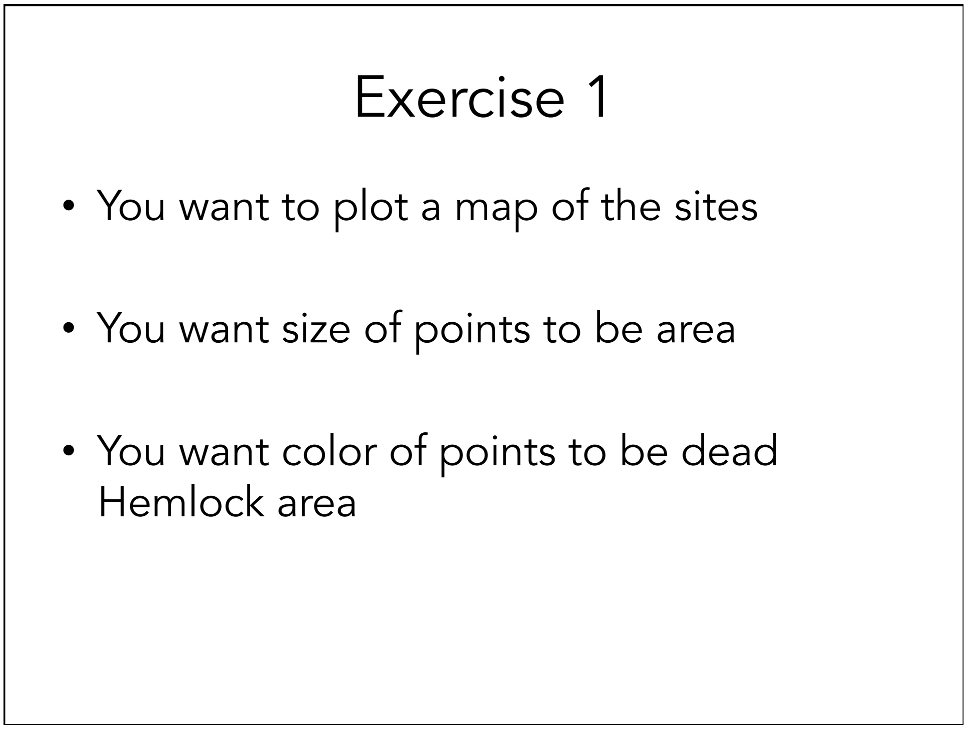


Anti Join: X NOT %in% Y

key	val_x
3	x3

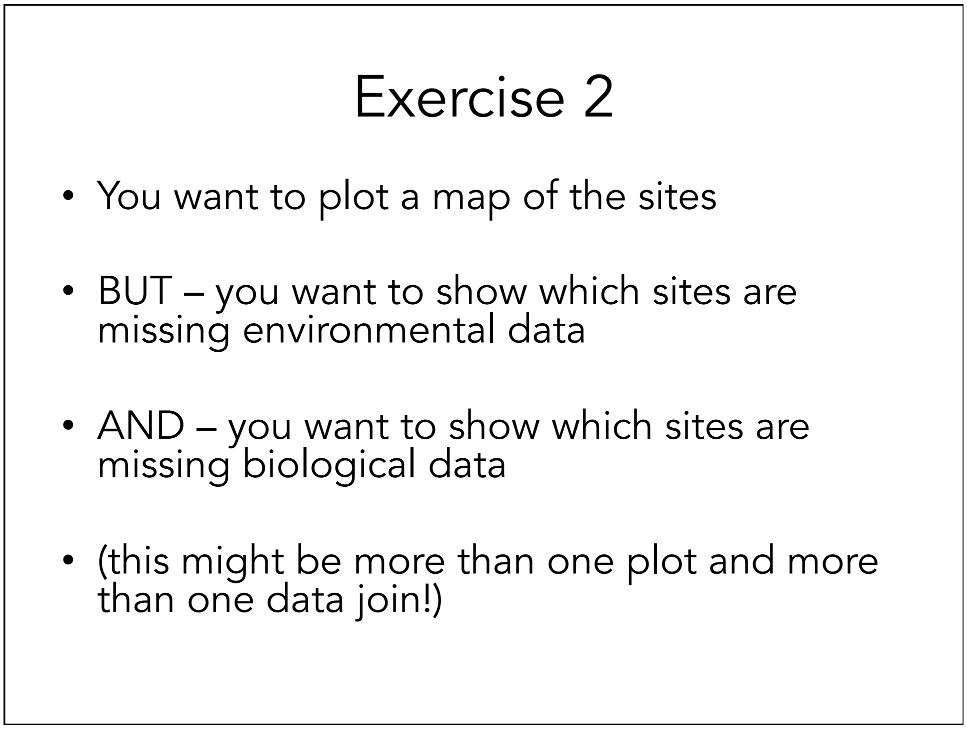
Good for diagnosing data mismatch

```
> hem_anti <- anti_join(hemlock_data, hemlock_sites)
Joining by: c("Stand", "Year", "Latitude", "Longitude")

> nrow(hem_anti)
[1] 11
```



Exercise 1

- You want to plot a map of the sites
- You want size of points to be area
- You want color of points to be dead Hemlock area



Exercise 2

- You want to plot a map of the sites
- BUT – you want to show which sites are missing environmental data
- AND – you want to show which sites are missing biological data
- (this might be more than one plot and more than one data join!)