Data Manipulation

Data Merging

Merging Data

- One of the challenges in data analytics is merging two datasets that share at least one common column
- Merging data involves in
  - Adding Columns
  - Adding Rows

Adding Rows

- To join two data frames (datasets) vertically, use the *rbind* function

- The two data frames must have the same variables, but they do not have to be in the same order.

Syntax

total <- rbind(data frameA, data frameB)

Adding Rows

Consider a data frame df1:

	key	field1
1	aaa	3
2	bbb	1
3	ccc	4

Data frame df2

	field1	key
1	2	aaa
2	1	ccc
3	7	eee
4	8	bbb

Adding Rows

```
>rbind(df1,df2)
o/p:
        key     field1
1       aaa     3
2       bbb     1
3       ccc     4
4        aaa    2
5       ccc     1
6        eee    7
7       bbb     8
```

Adding Rows

If data frameA has variables that data frameB does not, then either:

- Delete the extra variables in data frameA or
- Create the additional variables in data frameB and set them to NA (missing) before joining them with rbind( )

Adding Columns

- You can append two data frames using *cbind()* function
- Number of observations in two data frames must be same
- Appending is different from merging

Adding Columns

```
> df1 <- data.frame(key = c('aaa','bbb','ccc'), field1 = c(3,1,4))
> df2 <- data.frame(key = c('aaa','ccc','eee'), field2 = c(2,1,7))
> df1
  key field1
1 aaa      3
2 bbb      1
3 ccc      4
> df2
  key field2
1 aaa      2
2 ccc      1
3 eee      7
> cbind(df1,df2)
  key field1 key field2
1 aaa      3 aaa      2
2 bbb      1 ccc      1
3 ccc      4 eee      7
> rbind(df1,df2)
Error in match.names(clabs, names(xi)) :
  names do not match previous names
```

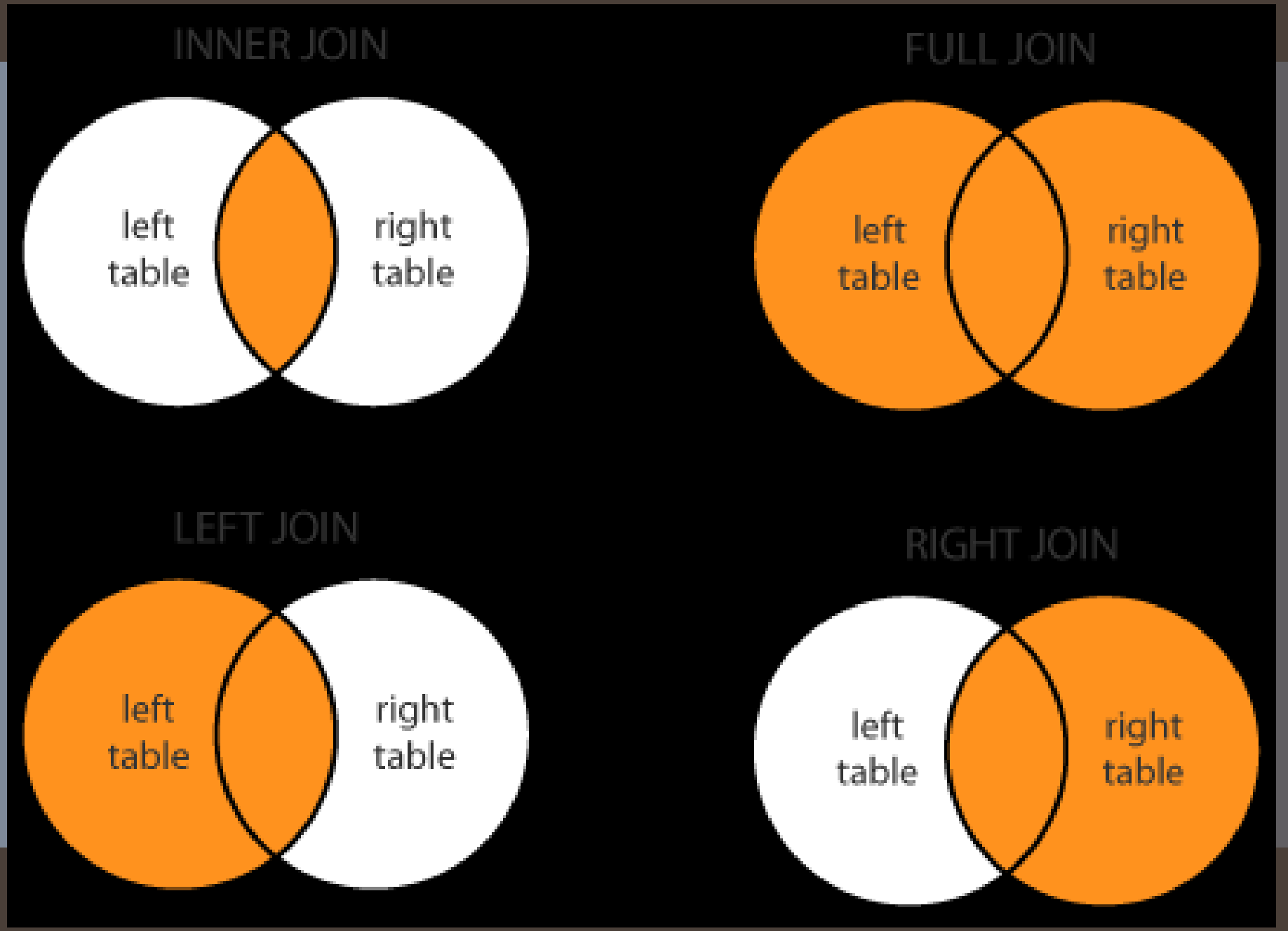
Merging Data

- Merging is joining two datasets that share at least one common column
- Merging is also known as join
- Joins are of mainly three types
  - Inner Join or Join
  - Outer Join or Full Join
  - Cross Join
- Outer join is of two types
  - Left Outer Join or Left Join
  - Right Outer Join or Right Join

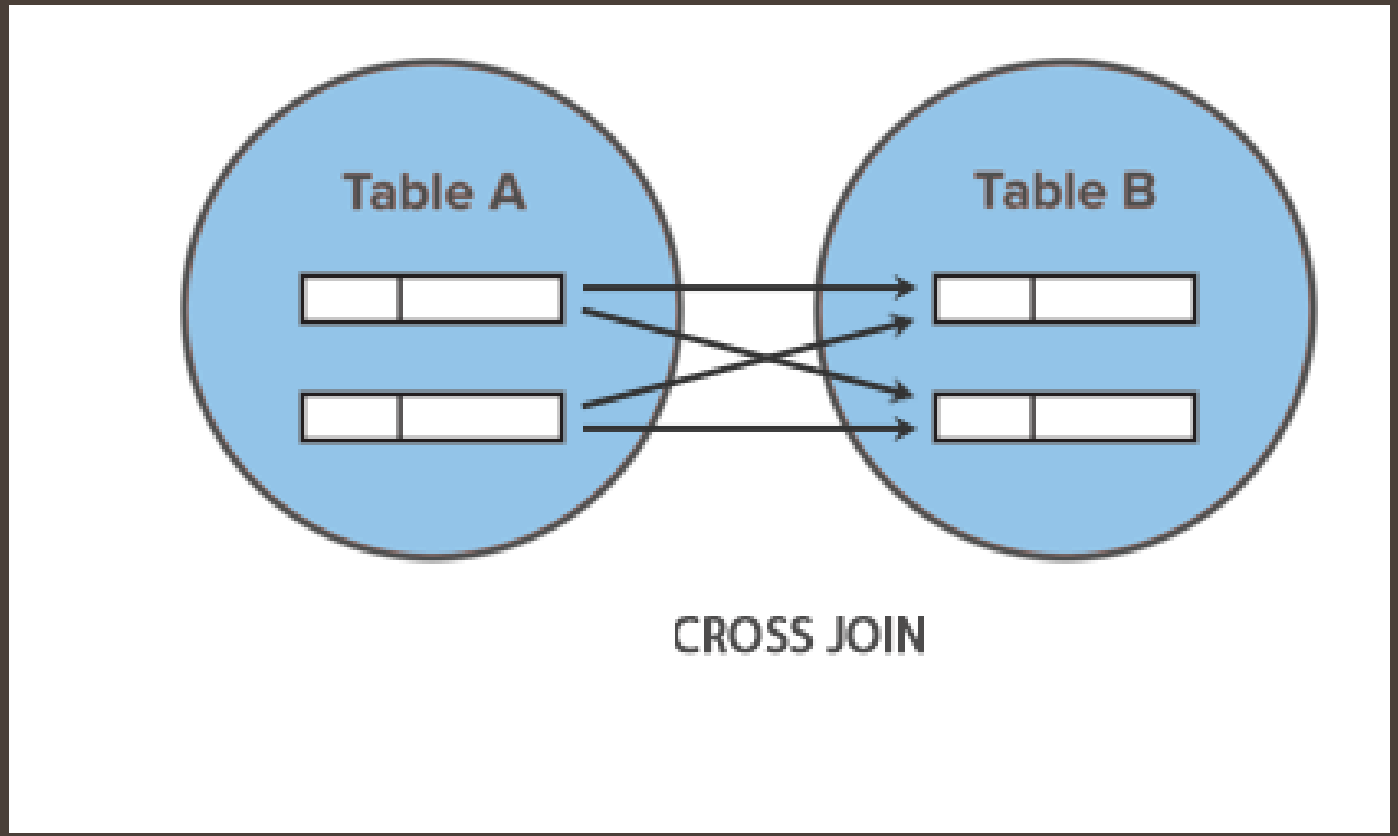
Types of Joins

- **(INNER) JOIN:** Returns records that have matching values in both tables
- **LEFT (OUTER) JOIN:** Return all records from the left table, and the matched records from the right table
- **RIGHT (OUTER) JOIN:** Return all records from the right table, and the matched records from the left table
- **FULL (OUTER) JOIN:** Return all records when there is a match in either left or right table

Types of Joins



Types of Joins



Example for Merging Data

```
df1 = data.frame(CustomerId = c(1:6), Product = c(rep("Toaster", 3), rep("Radio", 3)))
df2 = data.frame(CustomerId = c(2, 4, 6), State = c(rep("Alabama", 2), rep("Ohio", 1)))

df1
#  CustomerId Product
#           1 Toaster
#           2 Toaster
#           3 Toaster
#           4   Radio
#           5   Radio
#           6   Radio

df2
#  CustomerId   State
#           2 Alabama
#           4 Alabama
#           6    Ohio
```

Merging Data

Outer join: `merge(x = df1, y = df2, by = "CustomerId", all = TRUE)`

Left outer: `merge(x = df1, y = df2, by = "CustomerId", all.x = TRUE)`

Right outer: `merge(x = df1, y = df2, by = "CustomerId", all.y = TRUE)`

Cross join: `merge(x = df1, y = df2, by = NULL)`

Outer Join

```
> merge(x = df1, y = df2, by = "CustomerId", all = TRUE)
  CustomerId Product   State
1          1 Toaster    <NA>
2          2 Toaster Alabama
3          3 Toaster    <NA>
4          4   Radio Alabama
5          5   Radio    <NA>
6          6   Radio    Ohio
```

Left Outer Join

```
> merge(x = df1, y = df2, by = "CustomerId", all.x = TRUE)
  CustomerId Product   State
1          1 Toaster    <NA>
2          2 Toaster Alabama
3          3 Toaster    <NA>
4          4   Radio Alabama
5          5   Radio    <NA>
6          6   Radio    Ohio
```

Right Outer Join

```
> merge(x = df1, y = df2, by = "CustomerId", all.y = TRUE)
  CustomerId Product   State
1          2 Toaster Alabama
2          4   Radio Alabama
3          6   Radio    Ohio
```

Cross Join

```
> merge(x = df1, y = df2, by = NULL)
   CustomerId.x Product CustomerId.y   State
1             1 Toaster            2 Alabama
2             2 Toaster            2 Alabama
3             3 Toaster            2 Alabama
4             4   Radio            2 Alabama
5             5   Radio            2 Alabama
6             6   Radio            2 Alabama
7             1 Toaster            4 Alabama
8             2 Toaster            4 Alabama
9             3 Toaster            4 Alabama
10            4   Radio            4 Alabama
11            5   Radio            4 Alabama
12            6   Radio            4 Alabama
13            1 Toaster            6    Ohio
14            2 Toaster            6    Ohio
15            3 Toaster            6    Ohio
16            4   Radio            6    Ohio
17            5   Radio            6    Ohio
18            6   Radio            6    Ohio
```

Inner Join

```
> merge(x = df1, y = df2)
  CustomerId Product   State
1          2 Toaster Alabama
2          4   Radio Alabama
3          6   Radio    Ohio
> merge(x = df1, y = df2, by = "CustomerId")
  CustomerId Product   State
1          2 Toaster Alabama
2          4   Radio Alabama
3          6   Radio    Ohio
```