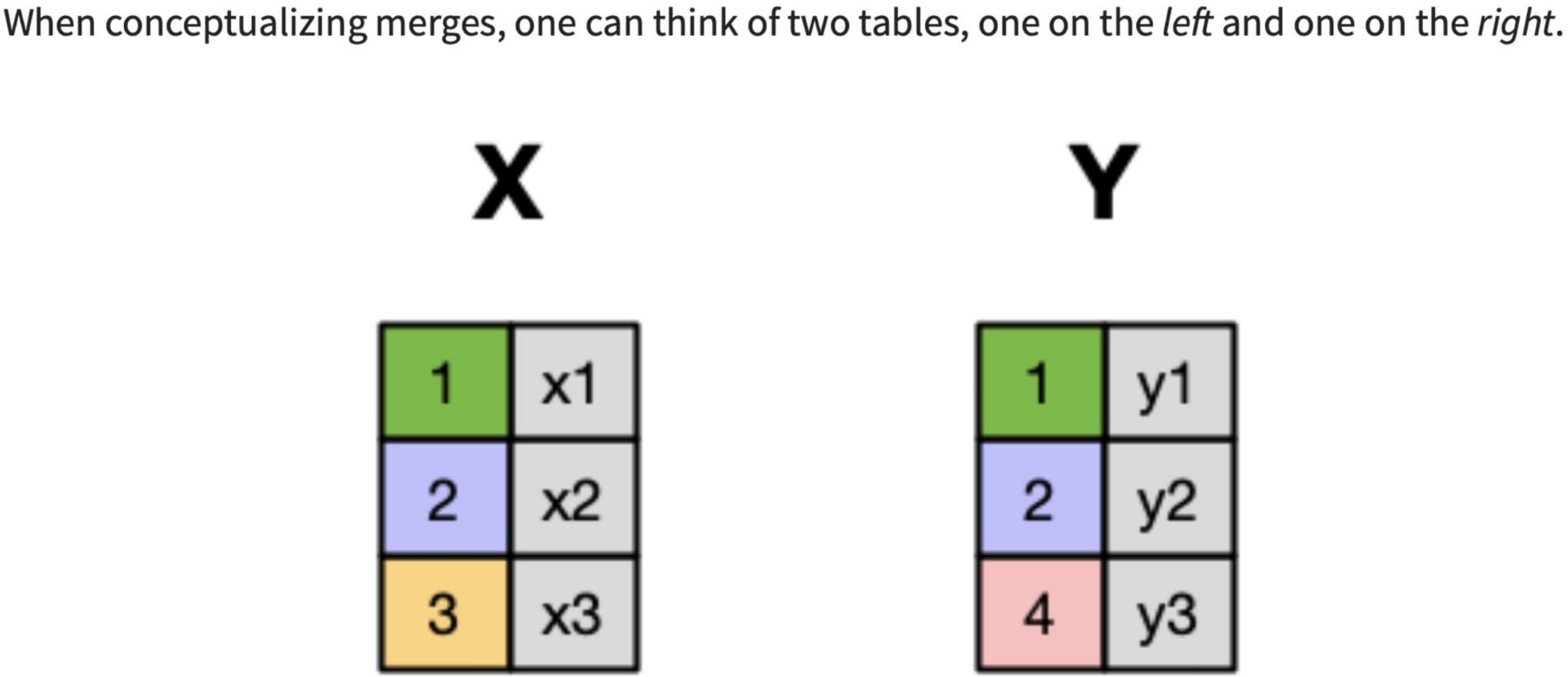
Merging Data

Instructor: Carmen Galaz García
UCSB Bren School EDS 220 - Fall 2023
https://carmengg.github.io/eds-220-book/



When conceptualizing merges, one can think of two tables, one on the *left* and one on the *right*.

X

1	x1
2	x2
3	x3

Y

1	y1
2	y2
4	y3

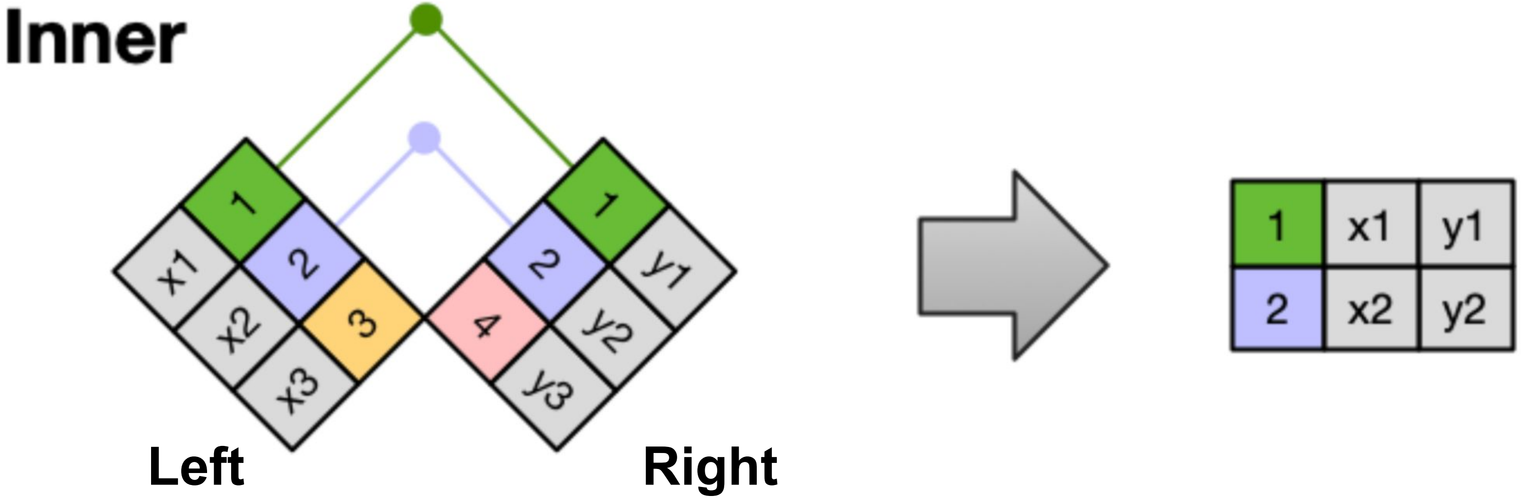
Left dataframe

Right dataframe

Image source: Data Modeling Essentials, NCEAS Learning Hub

Inner Join

An *INNER JOIN* is when you merge the subset of rows that have matches in both the left table and the right table.

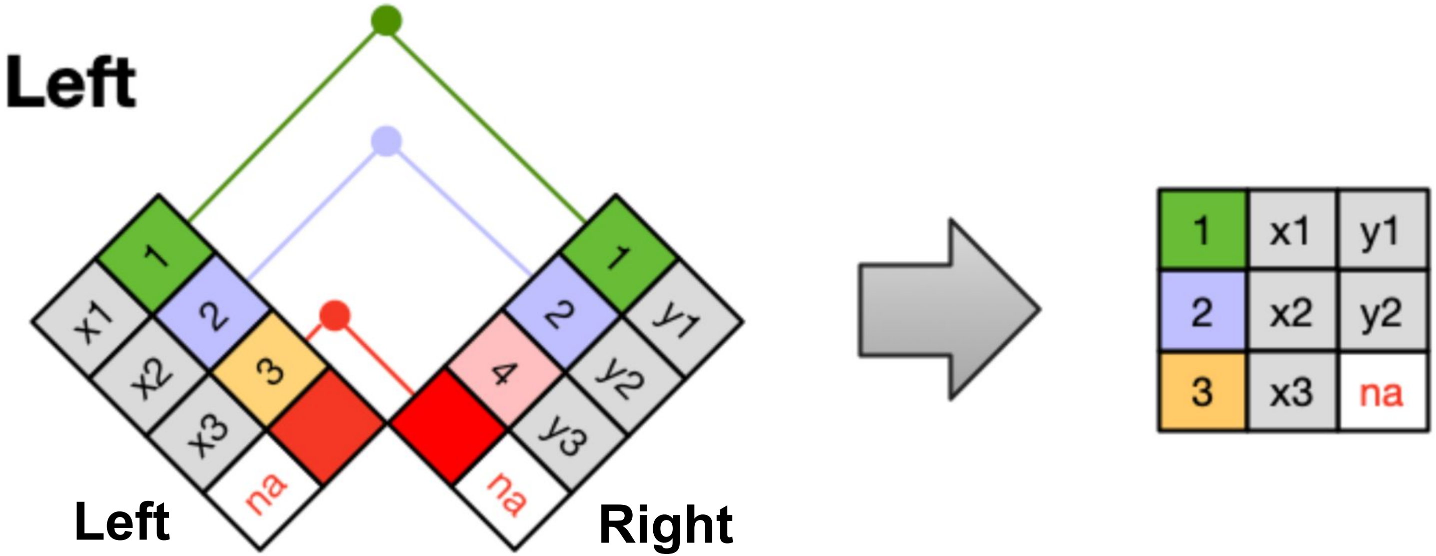


1	x1	y1
2	x2	y2

Image source: Data Modeling Essentials, NCEAS Learning Hub

Left Join

A *LEFT JOIN* takes all of the rows from the left table, and merges on the data from matching rows in the right table. Keys that don't match from the left table are still provided with a missing value (na) from the right table.



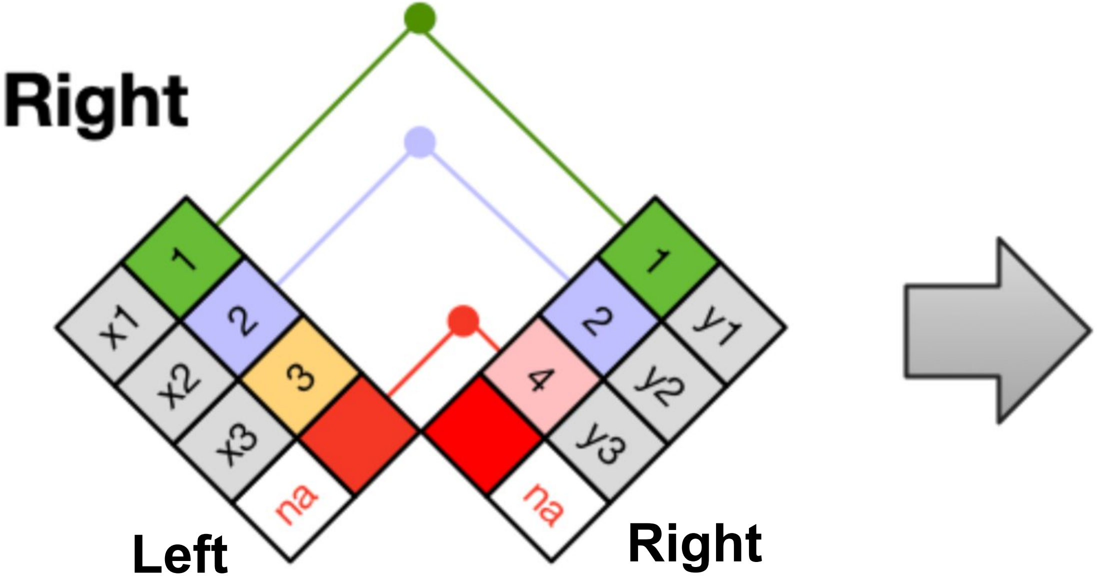
1	x1	y1
2	x2	y2
3	x3	na

Image source: Data Modeling Essentials, NCEAS Learning Hub

Instructor: Carmen Galaz García
UCSB Bren School EDS 220 - Fall 2023
https://carmengg.github.io/eds-220-book/

Right Join

A *RIGHT JOIN* is the same as a left join, except that all of the rows from the right table are included with matching data from the left, or a missing value. Notice that left and right joins can ultimately be the same depending on the positions of the tables

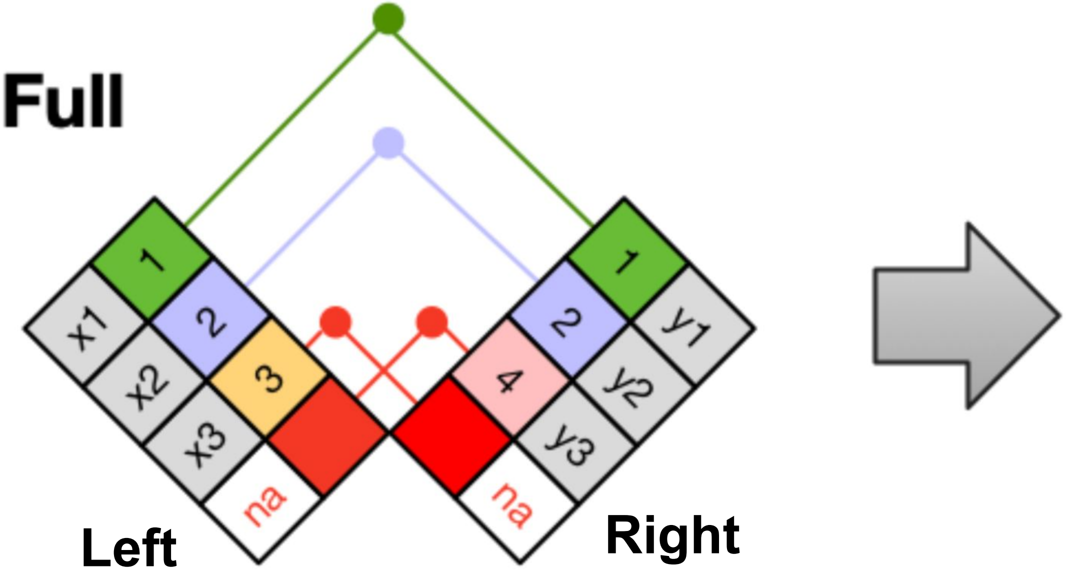


1	x1	y1
2	x2	y2
4	na	y3

Image source: Data Modeling Essentials, NCEAS Learning Hub

Full Outer Join

Finally, a *FULL OUTER JOIN* includes all data from all rows in both tables, and includes missing values wherever necessary.



1	x1	y1
2	x2	y2
3	x3	na
4	na	y3

Image source: Data Modeling Essentials, NCEAS Learning Hub

Sometimes people represent joins as Venn diagrams, showing which parts of the left and right tables are included in the results for each join. This representation is useful, however, they miss part of the story related to where the missing value comes from in each result.

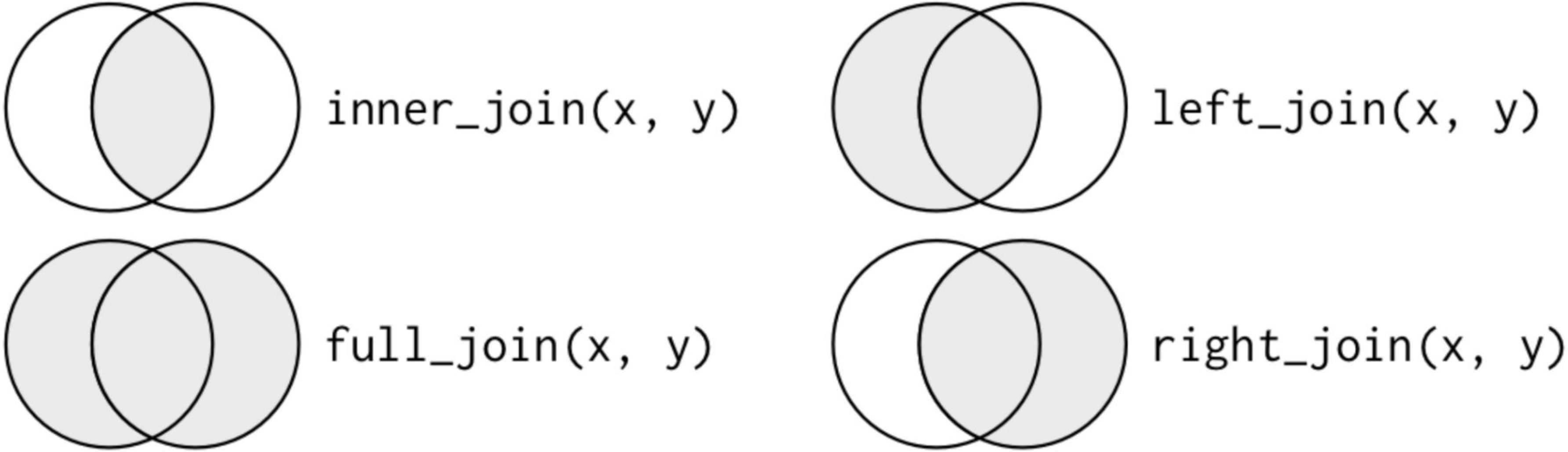


Image source: R for Data Science, Wickham & Grolemund.