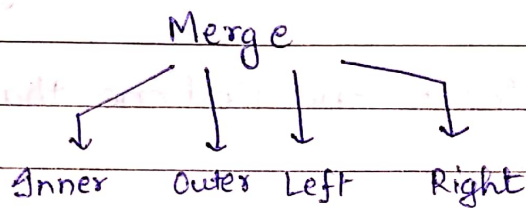
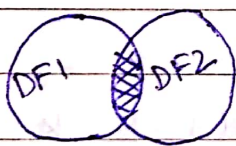


## Mission Name:- Data Cleaning Walkthrough: Combining the data

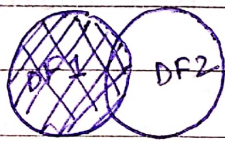
- \* For merging 2 data sets, we must remove any duplicate column value which is common for both datasets else we'll run into an error.
- \* Whenever we aggregate a dataset, the column upon which we aggregate becomes the index. To reset it, we need to use `pandas.DataFrame.reset-index()`.



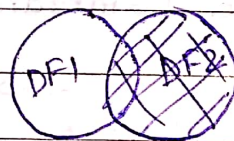
- \* Merge type should be selected wisely as it will affect the overall result.



Inner Join



Left Join



Right Join



Outer Join

\* So, different joins must be used as per requirement

\* When there are many datasets to handle with, the prime focus should be not to lose essential data during merging.

→ So, to avoid this problem, first we should identify the most important dataset and its merging column.

→ Then check the column with all other dataframes/dataset column.

→ The dataset with most numbers of common merge values should be joined with the important dataset.

→ Rest datasets can be joined accordingly.

\* FOR NOW:-

Missing values can be handled by imputing with mean of every column.

Eg:- `means = df.mean()`  
`df = df.fillna(means)`

\* Even after the step, NaN/null remains, then replace by 0.