Joining Datasets

Joining Datasets

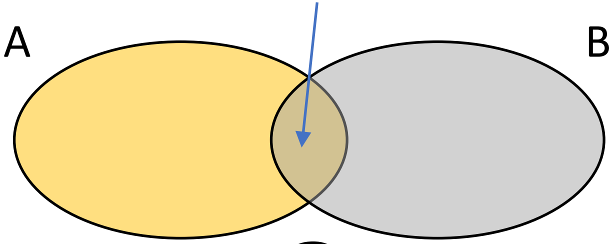
Overview

- Many ways to join two tables
- Follows a SQL syntax
- What you typically need is an index for joining
- Joining checks for matches or non-matches

Joining Datasets

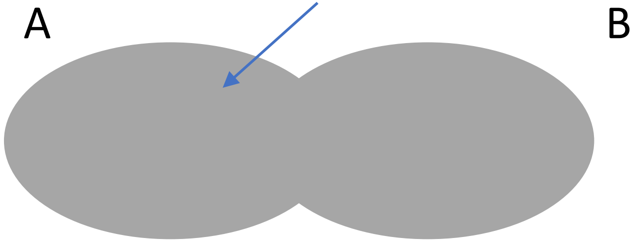
Set Operations

intersect()



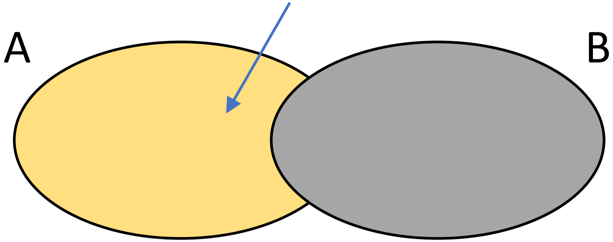
$A \bigcap B$

union()



$\mathrm{A} \cup B$

setdiff()

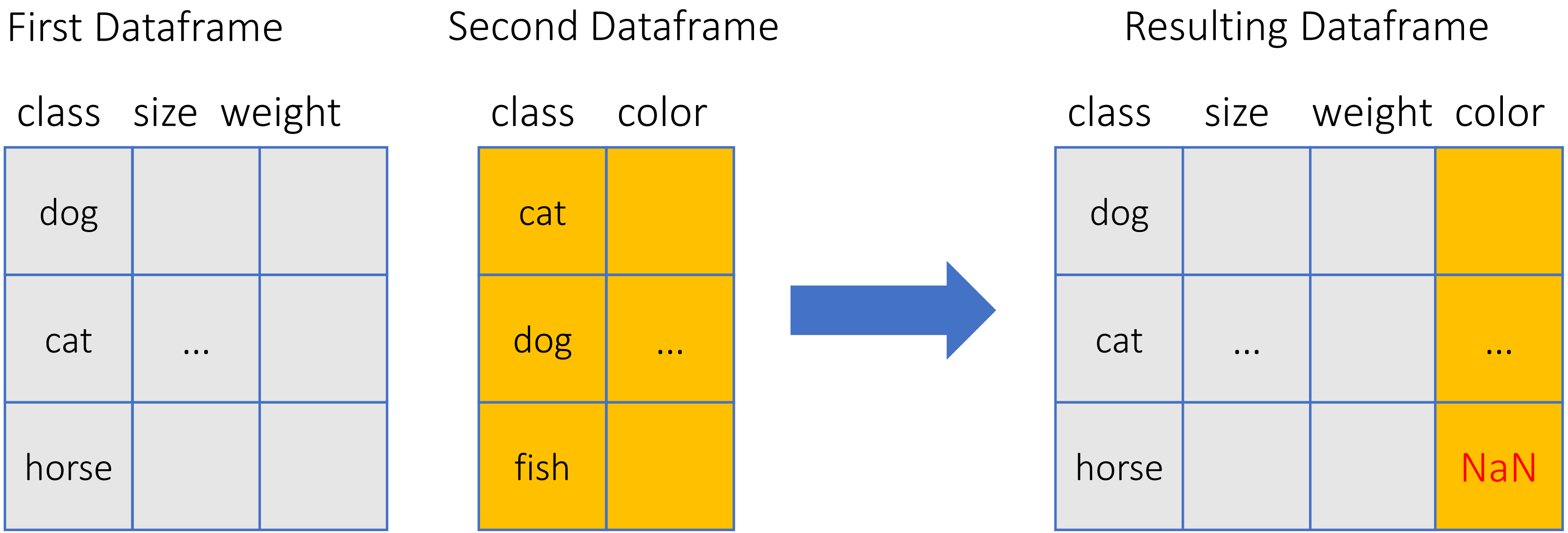


A \ B

Joining Datasets

Left Join

- Left Join keeps the full dataframe at the LHS
- Adds information from RHS, if possible, otherwise adds NA
- Default for merge method



First Dataframe

class	size	weight
dog		
cat	…	
horse		

Second Dataframe

class	color
cat	
dog	…
fish	

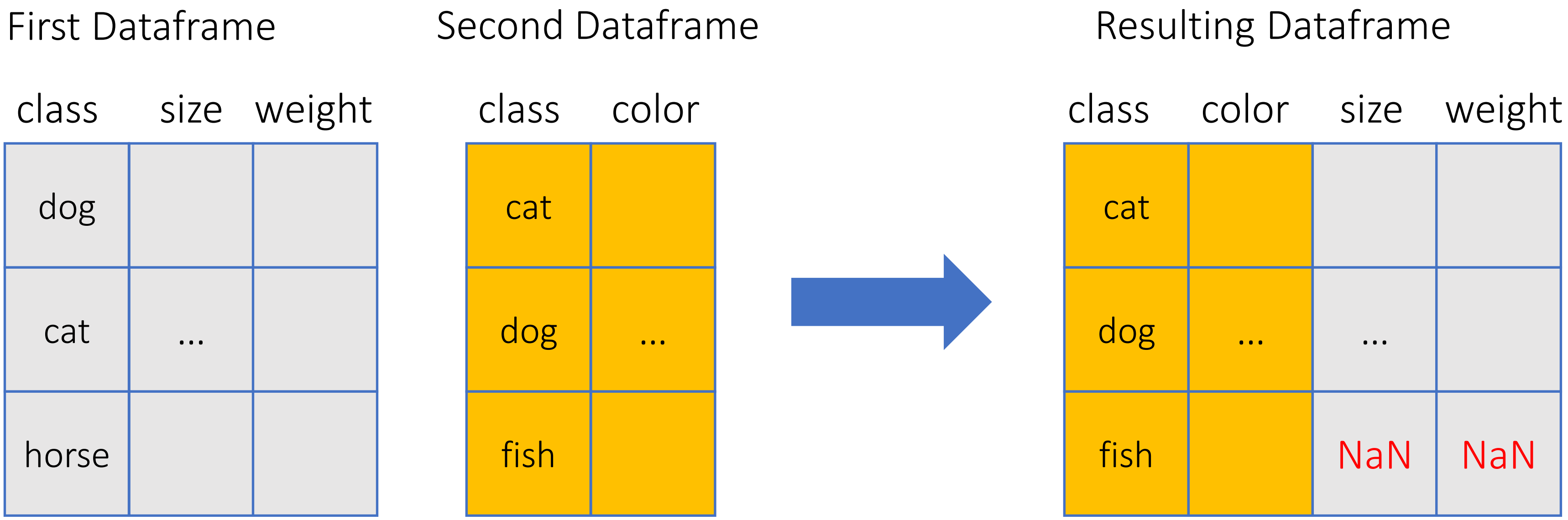
Resulting Dataframe

class	size	weight	color
dog			
cat	…		…
horse			NaN

Joining Datasets

Right Join

- Right Join keeps the full dataframe at the RHS
- Adds information from LHS, if possible, otherwise adds NA



First Dataframe

class	size	weight
dog		
cat	…	
horse		

Second Dataframe

class	color
cat	
dog	…
fish	

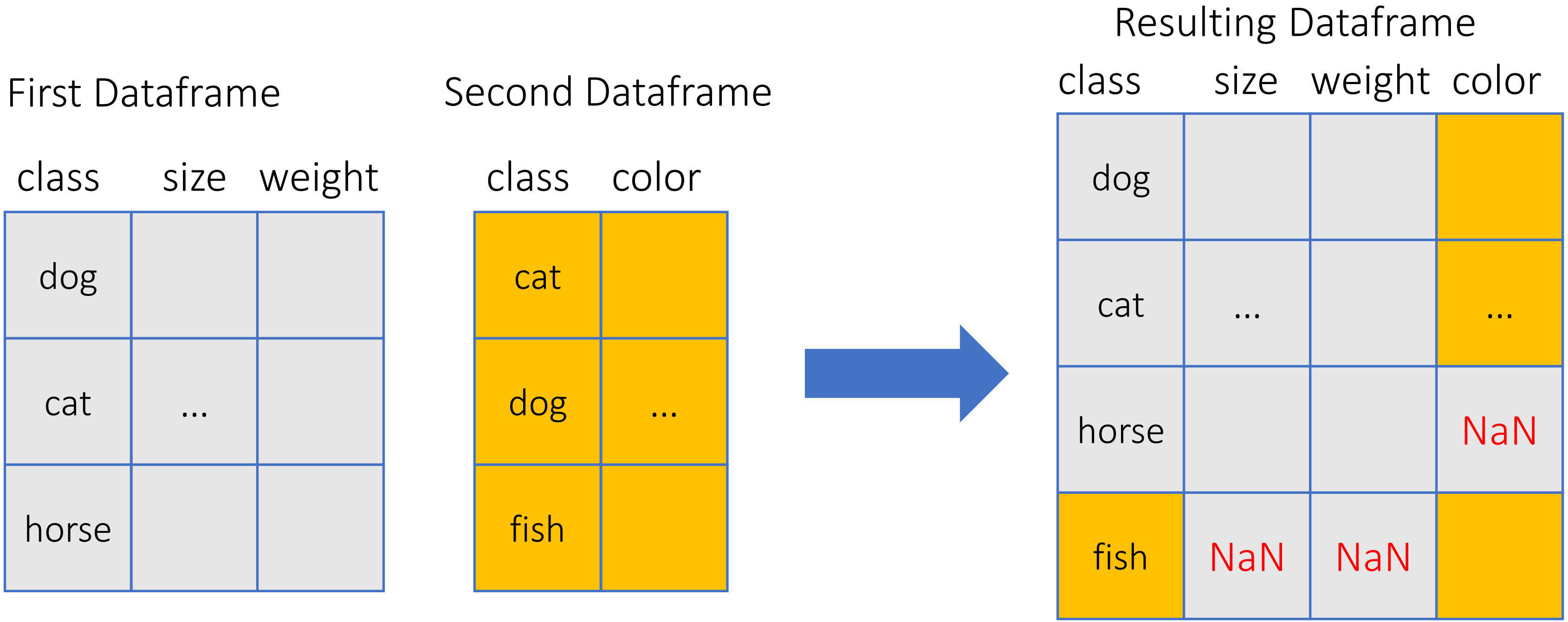
Resulting Dataframe

class	color	size	weight
cat			
dog	…	…	
fish		NaN	NaN

Joining Datasets

Outer Join

- Outer Join keeps the full dataframe at the LHS and RHS
- NAs are added in cells with missing information



First Dataframe

class	size	weight
dog		
cat	...	
horse		

Second Dataframe

class	color
cat	
dog	...
fish	

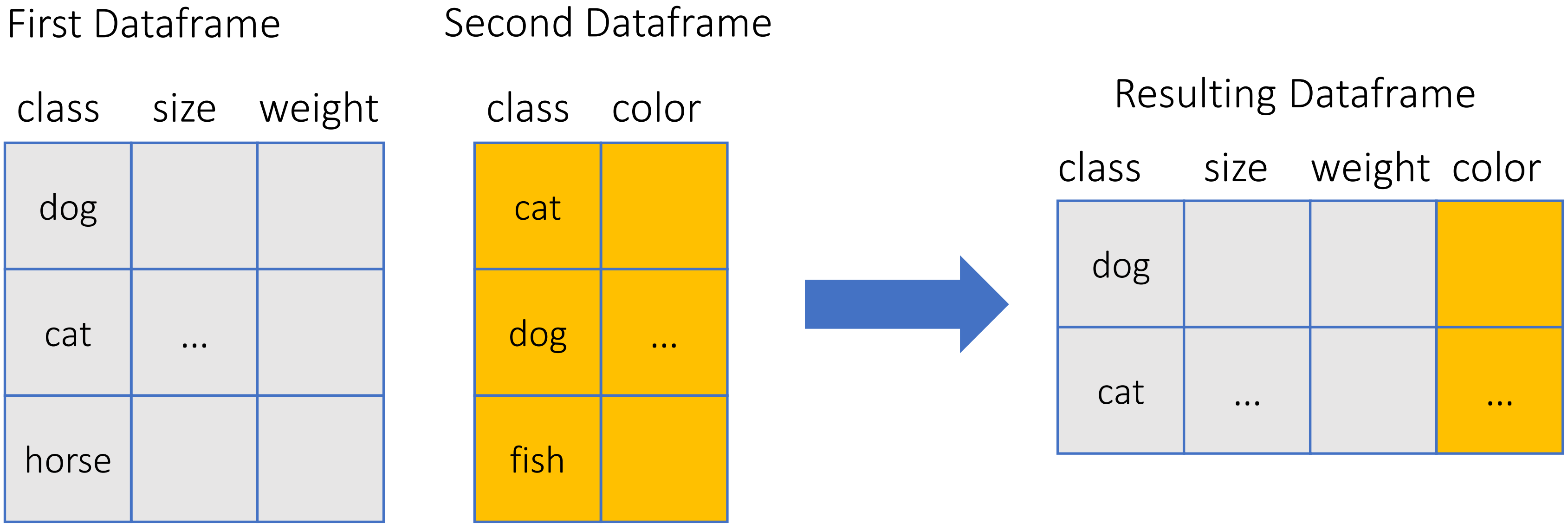
Resulting Dataframe

class	size	weight	color
dog			
cat	...		...
horse			NaN
fish	NaN	NaN	

Joining Datasets

Inner Join

- Inner Join keeps the rows of the dataframe where identical indices at LHS and RHS are available



First Dataframe

class	size	weight
dog		
cat	...	
horse		

Second Dataframe

class	color
cat	
dog	...
fish	

Resulting Dataframe

class	size	weight	color
dog			
cat	...		...

Joining Datasets

Cheatsheet

Source Dataframes

class	size
dog	
cat	...
horse	

class	color
cat	
dog	...
fish	

Left Join

class	size	color
dog		
cat	...	...
horse		NaN

Right Join

class	color	size
cat		
dog	...	...
fish		NaN

Inner Join

class	size	color
dog		
cat	...	...

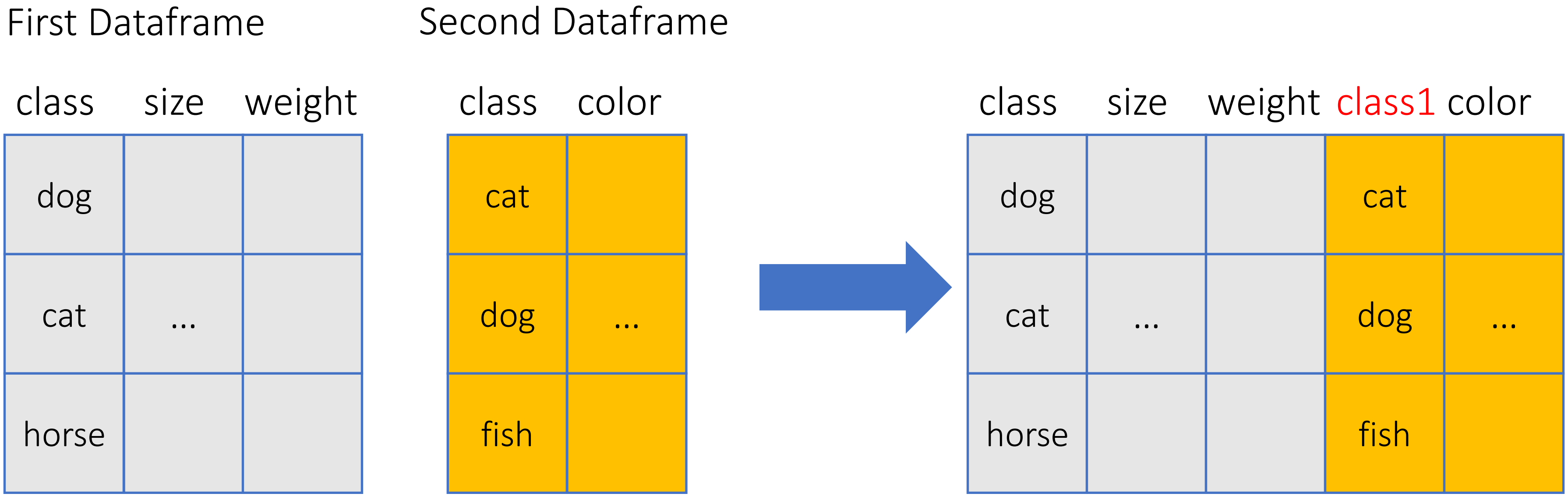
Outer Join

class	size	color
dog		
cat	...	...
horse		NaN
fish	NaN	

Joining Datasets

Bind Columns

- Bind columns pastes tables next to each other (as they are)
- pd.concat([df1, df2])



First Dataframe

class	size	weight
dog		
cat	…	
horse		

Second Dataframe

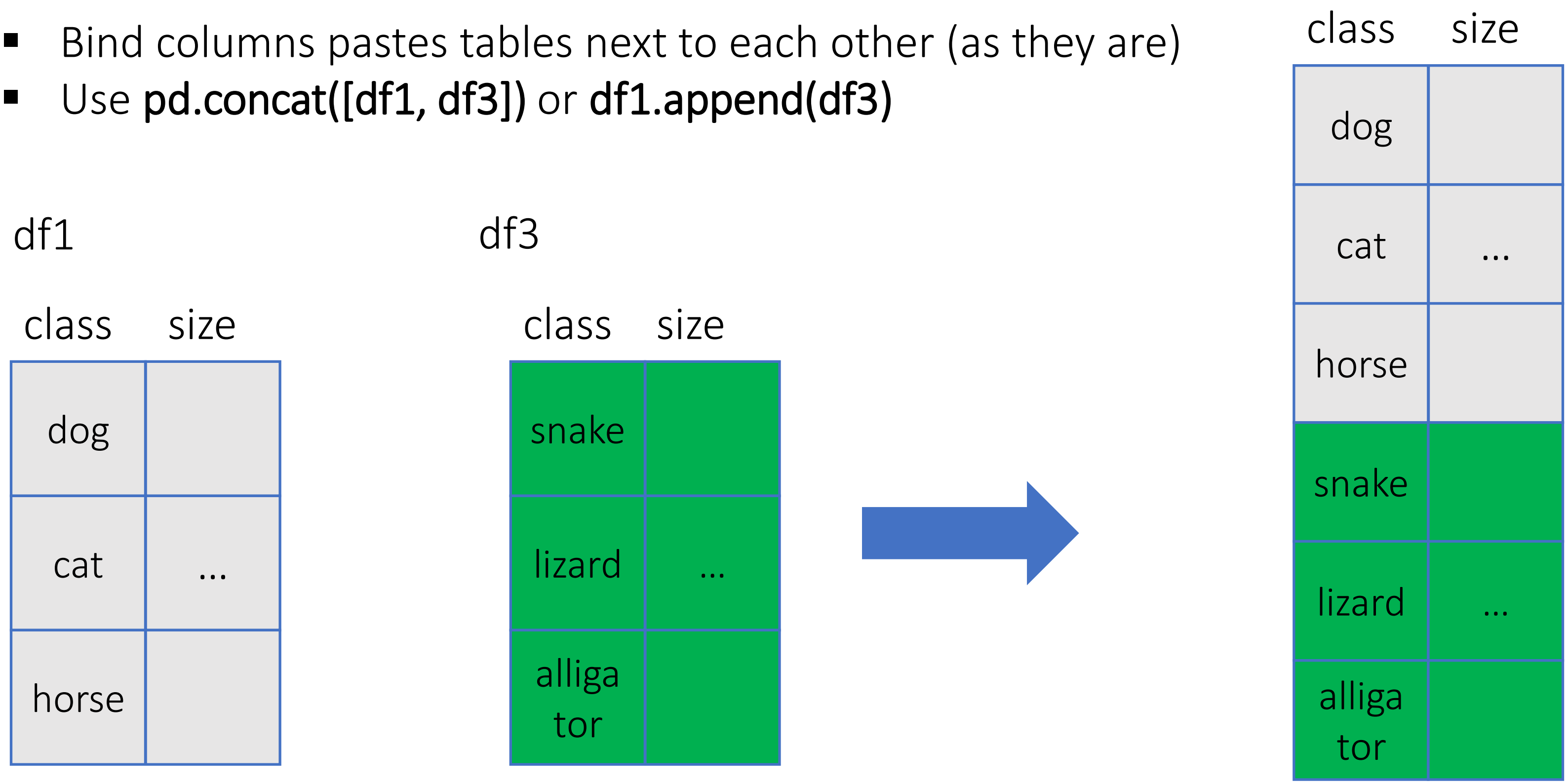
class	color
cat	
dog	…
fish	

class	size	weight	class1	color
dog			cat	
cat	…		dog	…
horse			fish	

Joining Datasets

Bind Rows

- Bind columns pastes tables next to each other (as they are)
- Use **pd.concat([df1, df3])** or **df1.append(df3)**



df1

class	size
dog	
cat	...
horse	

df3

class	size
snake	
lizard	...
alligator	

class	size
dog	
cat	...
horse	
snake	
lizard	...
alligator	

Joining Datasets

Further Options

- Bind rows: returns tables one on top of the other
- Semi Join: returns rows of LHS that find a match at RHS
- Anti Join: returns rows of LHS that do not find a match at RHS