Pandas merge

SHARP SIGHT

What you'll learn

- How to use the Pandas `merge()` function

- How to "merge" data
  - i.e., combine by looking for a match on a variable

- How to perform different types of merges
  - left merge
  - right merge
  - inner merge

Pandas merge overview

What is a merge?

- A merge combines records from 2 datasets

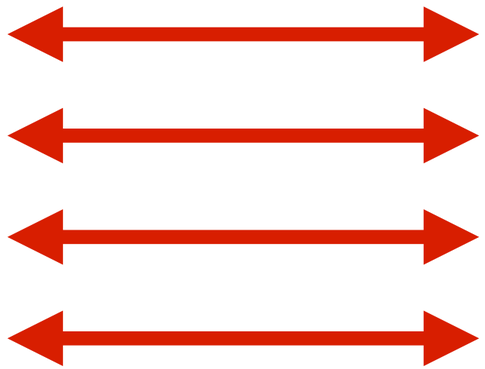
- Merges are essentially the same as:
  - merge (SAS)
  - join (R's dplyr)
  - vlookup (Excel)
  - SQL (join)

High level example of a merge

```
pd.merge(emp_fst_nm, emp_lst_nm, on = 'employee_id')
```

emp_fst_nm

* employee_id	first_name
101	John
102	Paul
103	George
104	Ringo
900	Stewart



emp_lst_nm

* employee_id	last_name
101	Lennon
102	McCartney
103	Harrison
104	Starr
901	Best

High level example of a merge

```
pd.merge(emp_fst_nm, emp_lst_nm, on = 'employee_id')
```

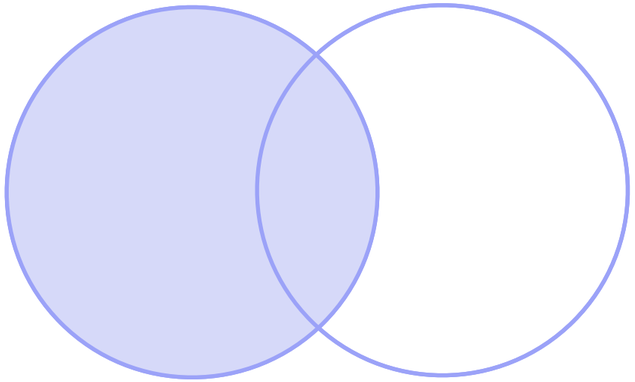
OUT:

employee_id	first_name	last_name
101	John	Lennon
102	Paul	McCartney
103	George	Harrison
104	Ringo	Starr

2 MERGE TYPES YOU'LL LEARN

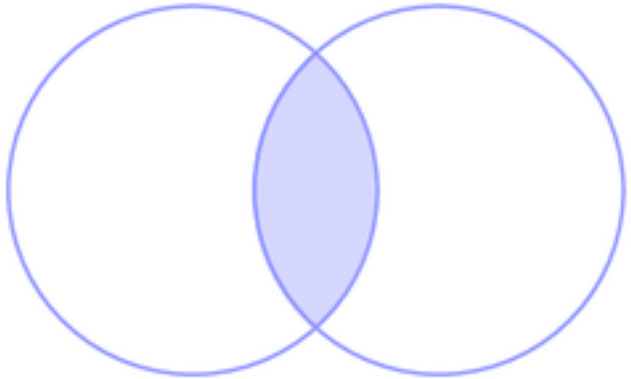
Left merge

(keep all records from "left" dataset)



Inner merge

(only keep records that are in BOTH datasets)



Why inner merge & left merge?

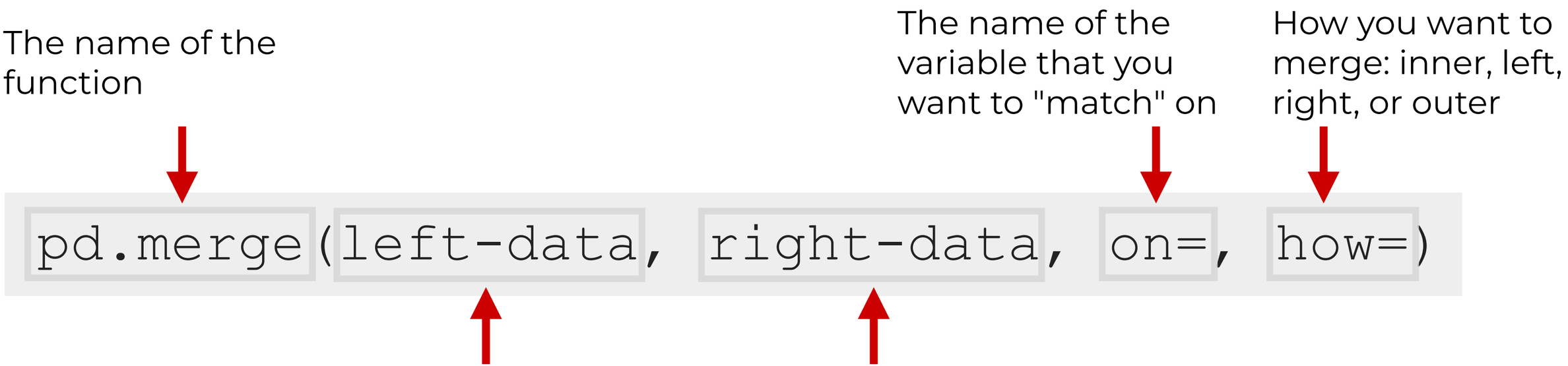
- These two are most common

- There are two other merge types
  - you'll probably never use them
  - so, focus on most common

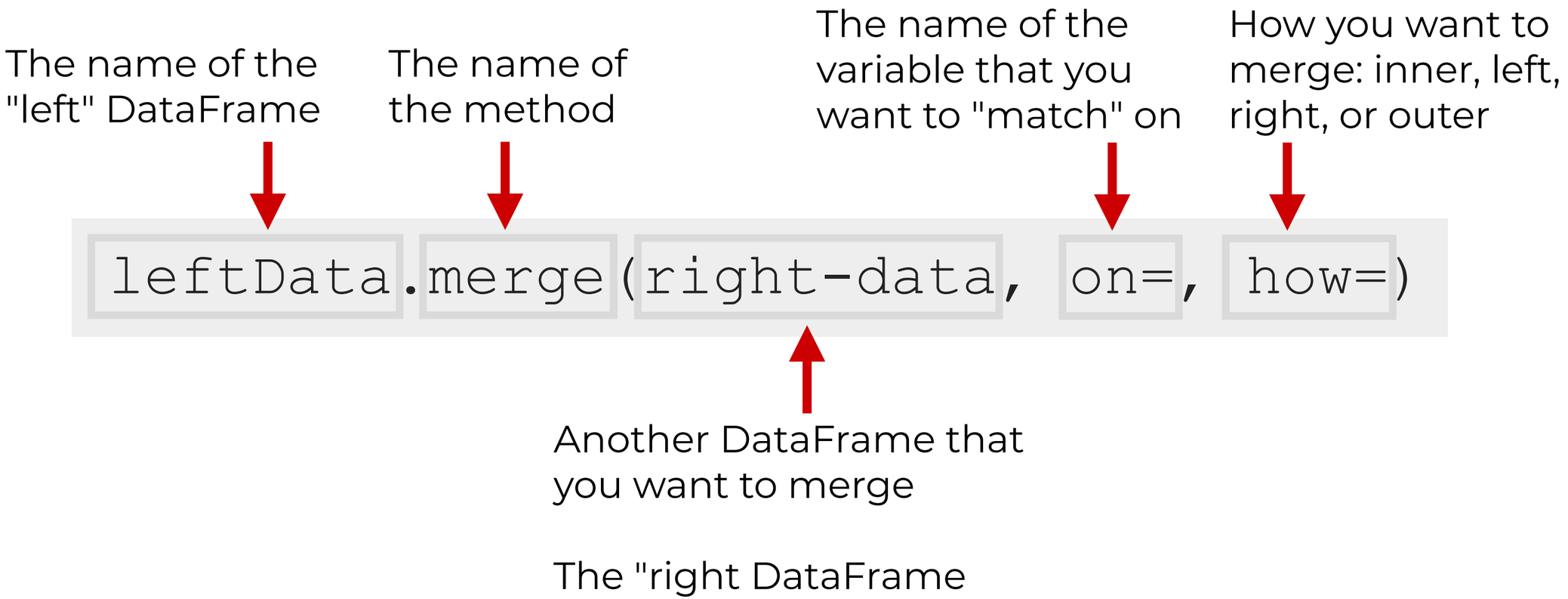
- Right merge is "inverted" version of left merge
  - once you know left, you can do right

Syntax: pandas merge

Syntax: pandas merge function



Note: there's also a method version



Parameters of Pandas merge

The parameters of the Pandas merge function

Parameter	What it does	Format	Default	Required?
`left-data`	The the dataset on the left side of the syntax that you want to merge	`DataFrame`		Yes
`right-data`	The the dataset on the right side of the syntax that you want to merge	`DataFrame`		Yes
`on=`	Which variable upon which you want to look for "matches" for the merge	A column or list of columns	Intersection of the columns in both DataFrames. (You should always provide an argument to `on`)	No
`how=`	How you want to merge the data ... inner, left, right, or outer	One of the following: `inner`, `left`, `right`, `outer`	`inner`	No

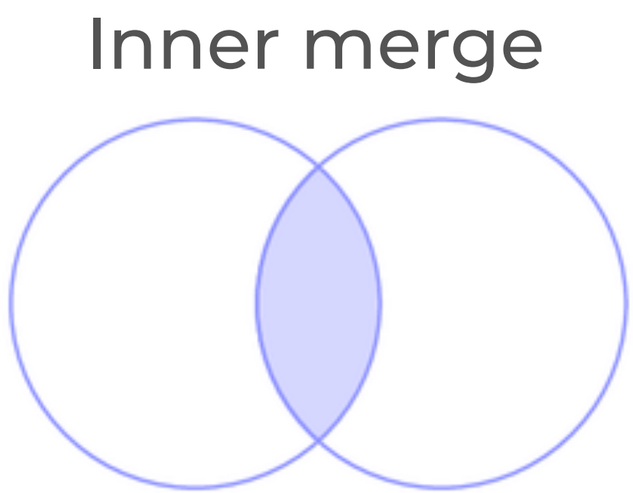
The output of Pandas merge

- The output of pd.merge() is a DataFrame
  - Contains the merged data
  - Merged as specified in your syntax

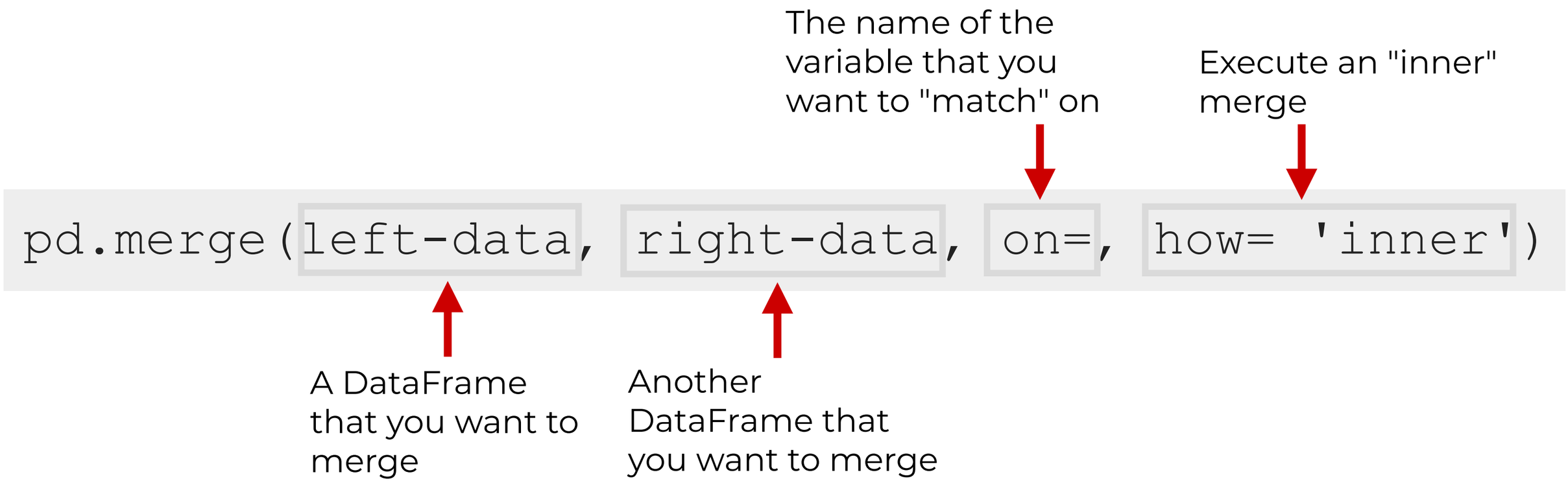
Example: inner merge

Inner merge

- Only keep records that are in BOTH datasets



Syntax: pandas inner merge



Example: We have two separate datasets

- Want to combine them
  - keep only records that are in both

`emp_fst_nm`

employee_id	first_name
101	John
102	Paul
103	George
104	Ringo
900	Stewart

`emp_lst_nm`

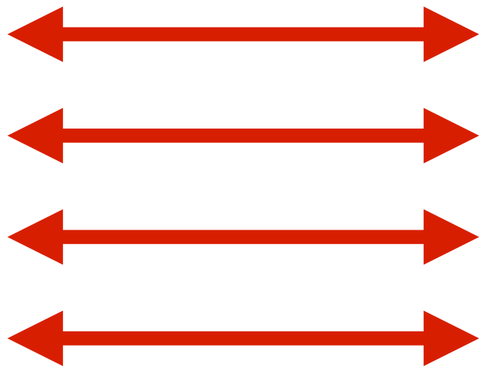
employee_id	last_name
101	Lennon
102	McCartney
103	Harrison
104	Starr
901	Best

Example: inner merge

```
pd.merge(emp_fst_nm, emp_lst_nm, on = 'employee_id', how = 'inner')
```

emp_fst_nm

* employee_id	first_name
101	John
102	Paul
103	George
104	Ringo
~~900~~	~~Stewart~~



emp_lst_nm

* employee_id	last_name
101	Lennon
102	McCartney
103	Harrison
104	Starr
~~901~~	~~Best~~

Inner merge results

```
pd.merge(emp_fst_nm, emp_lst_nm, on = 'employee_id', how = 'inner')
```

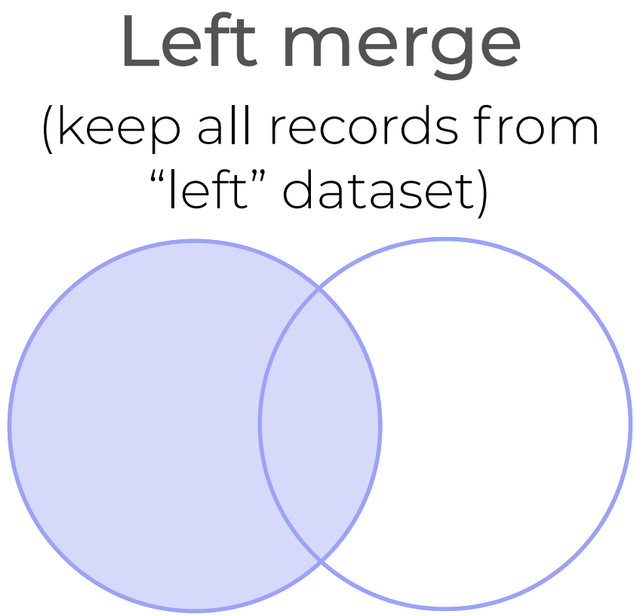
- Result is a combined dataset
  - Columns from both data sets
  - Excludes rows that were not in both

employee_id	first_name	last_name
101	John	Lennon
102	Paul	McCartney
103	George	Harrison
104	Ringo	Starr

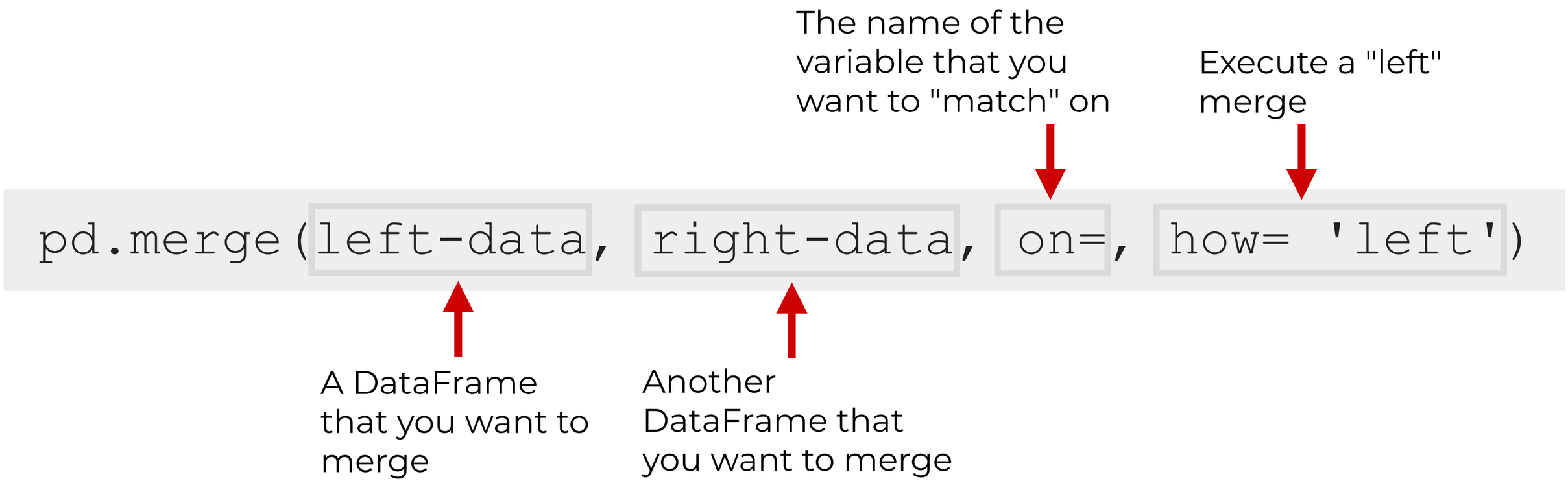
Example: left merge

LEFT MERGE

- Keep all records in the LEFT dataset
  - drop records in the right dataset that do not match



Syntax: pandas left merge



Example: We have two separate datasets

- Keep all records in the LEFT dataset
  - drop records in the right dataset that do not match

emp_fst_nm

employee_id	first_name
101	John
102	Paul
103	George
104	Ringo
900	Stewart

emp_lst_nm

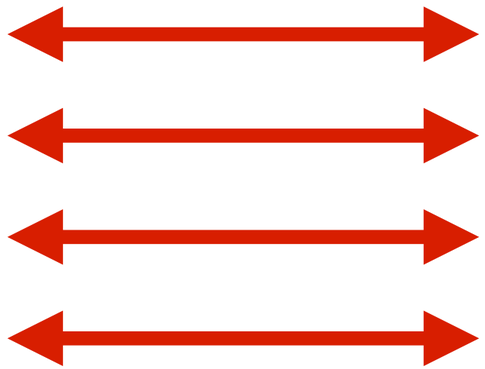
employee_id	last_name
101	Lennon
102	McCartney
103	Harrison
104	Starr
901	Best

Example: left merge

```
pd.merge(emp_fst_nm, emp_lst_nm, on = 'employee_id', how = 'left')
```

emp_fst_nm

* employee_id	first_name
101	John
102	Paul
103	George
104	Ringo
900	Stewart



emp_lst_nm

* employee_id	last_name
101	Lennon
102	McCartney
103	Harrison
104	Starr
~~901~~	~~Best~~

Left merge results

```
pd.merge(emp_fst_nm, emp_lst_nm, on = 'employee_id', how = 'left')
```

- Result is a combined dataset
  - Include all rows in left dataset
  - Excludes rows from right dataset where no match was found
  - Replace missing data with missing value

employee_id	first_name	last_name
101	John	Lennon
102	Paul	McCartney
103	George	Harrison
104	Ringo	Starr
900	Stewart	NaN

Recap

Recap of what we learned

- You can "merge" DataFrames together with the `pd.merge()` function

- How to perform different types of merges
  - inner
  - left

- **Next Steps**: Watch the code walkthrough video for step-by-step examples of how to merge data