Merging datasets

Many datasets are split across different tables/files and need to be combined during analysis.

The `join` family of functions provides all the functionality you need.

When merging two tables, first decide what columns define a **match** between rows.

```
blah_join(x, y, by = <columns>)
```

Pick from six join types, which differ in what they do with **non-matching** rows.

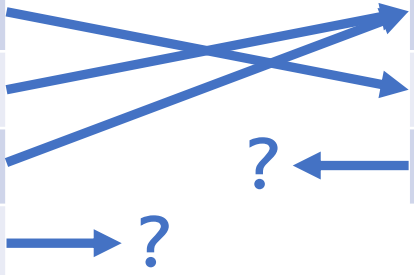
Join type	Rows of x	Columns of x	Columns of y
`inner`	> 1 match in y	all	all
`left`	all	all	all
`right`	> 1 match in y	all	all
`full`	all	all	all
`semi`	> 1 match in y	all	none
`anti`	0 match in y	all	none

sightings

location	poke_id
ORC	2
ORC	1
Sloan	1
Sloan	4

pokedex

ID	name
1	Bulbasaur
2	Squirtle
3	Charmander



We would like to match `poke_id` on the LHS to `ID` on the RHS, so we specify:

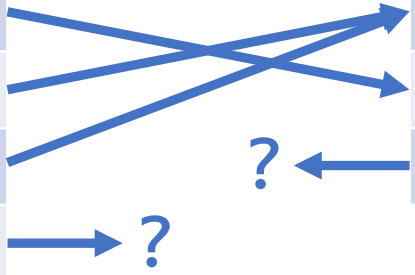
```
xxxx_join(sightings, pokedex, by = c("poke_id" = "ID"))
```

sightings

location	poke_id
ORC	2
ORC	1
Sloan	1
Sloan	4

pokedex

ID	name
1	Bulbasaur
2	Squirtle
3	Charmander




```
sightings %>%
  inner_join(pokedex, by = c("poke_id" = "ID"))
```

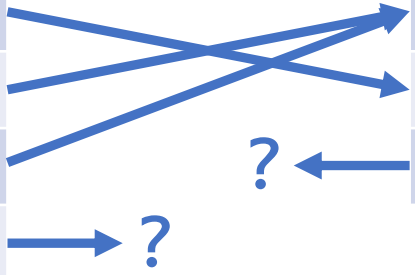
location	poke_id	name
ORC	2	Squirtle
ORC	1	Bulbasaur
Sloan	1	Bulbasaur

sightings

location	poke_id
ORC	2
ORC	1
Sloan	1
Sloan	4

pokedex

ID	name
1	Bulbasaur
2	Squirtle
3	Charmander




```
sightings %>%
   left_join(pokedex, by = c("poke_id" = "ID"))
```

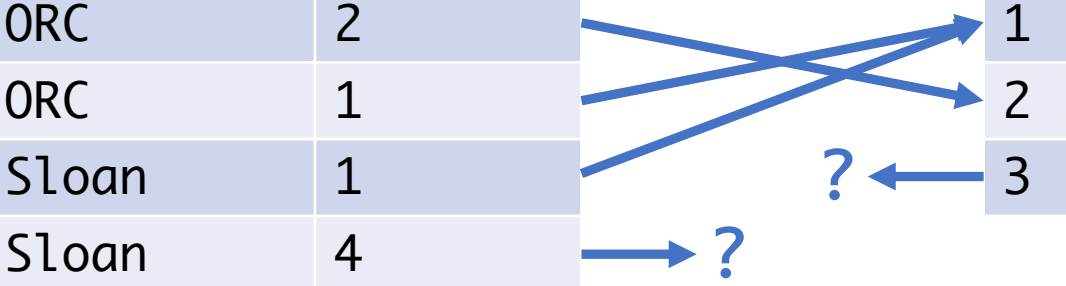
location	poke_id	name
ORC	2	Squirtle
ORC	1	Bulbasaur
Sloan	1	Bulbasaur
Sloan	4	NA

sightings

location	poke_id
ORC	2
ORC	1
Sloan	1
Sloan	4

pokedex

ID	name
1	Bulbasaur
2	Squirtle
3	Charmander




```
sightings %>%
  right_join(pokedex, by = c("poke_id" = "ID"))
```

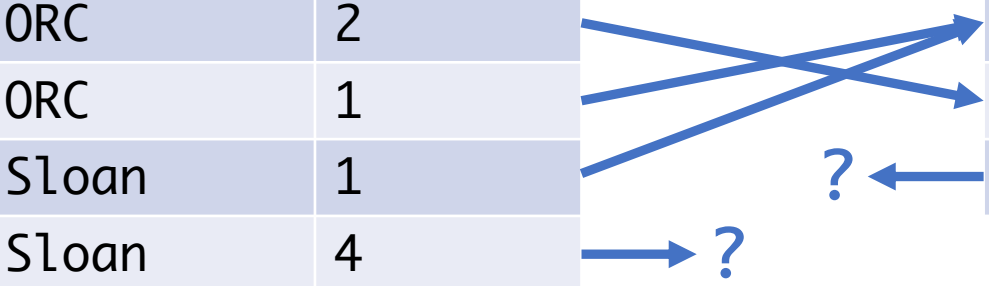
location	poke_id	name
ORC	2	Squirtle
ORC	1	Bulbasaur
Sloan	1	Bulbasaur
NA	3	Charmander

sightings

location	poke_id
ORC	2
ORC	1
Sloan	1
Sloan	4

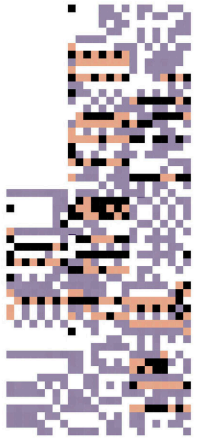
pokedex

ID	name
1	Bulbasaur
2	Squirtle
3	Charmander




```
sightings %>%
   full_join(pokedex, by = c("poke_id" = "ID"))
```

location	poke_id	name
ORC	2	Squirtle
ORC	1	Bulbasaur
Sloan	1	Bulbasaur
Sloan	4	NA
NA	3	Charmander

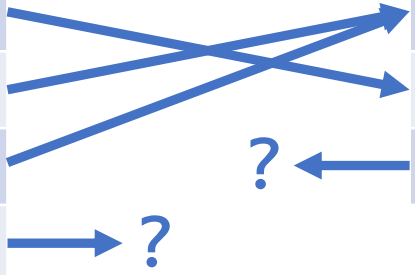


sightings

location	poke_id
ORC	2
ORC	1
Sloan	1
Sloan	4

pokedex

ID	name
1	Bulbasaur
2	Squirtle
3	Charmander




```
sightings %>%
   semi_join(pokedex, by = c("poke_id" = "ID"))
```

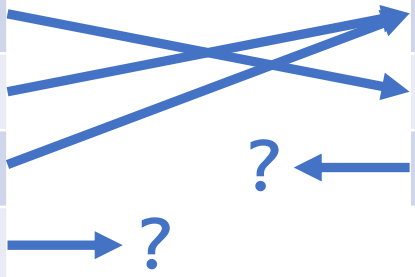
location	poke_id
ORC	2
ORC	1
Sloan	1

sightings

location	poke_id
ORC	2
ORC	1
Sloan	1
Sloan	4

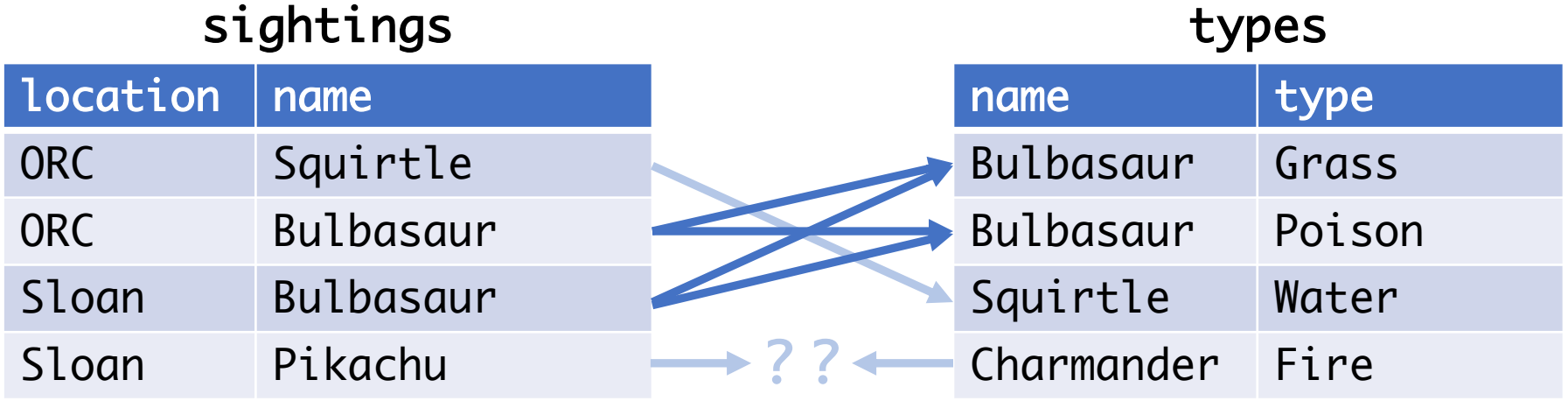
pokedex

ID	name
1	Bulbasaur
2	Squirtle
3	Charmander




```
sightings %>%
   anti_join(pokedex, by = c("poke_id" = "ID"))
```

location	poke_id
Sloan	4



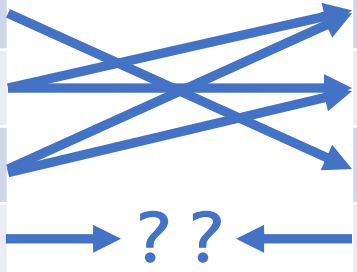
If there are **multiple** matches to a row, the result will have a row for **every combination**.

sightings

location	name
ORC	Squirtle
ORC	Bulbasaur
Sloan	Bulbasaur
Sloan	Pikachu

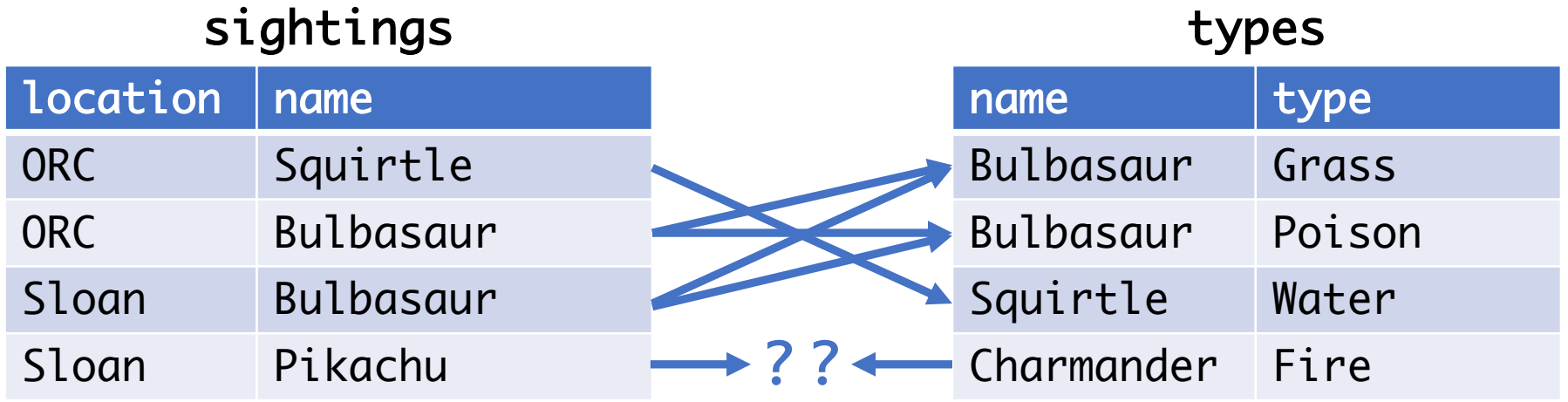
types

name	type
Bulbasaur	Grass
Bulbasaur	Poison
Squirtle	Water
Charmander	Fire




```
sightings %>%
  inner_join(types, by = "name")
```

location	name	type
ORC	Squirtle	Water
ORC	Bulbasaur	Grass
ORC	Bulbasaur	Poison
Sloan	Bulbasaur	Grass
Sloan	Bulbasaur	Poison




```
sightings %>%
   left_join(types, by = “name”)
```

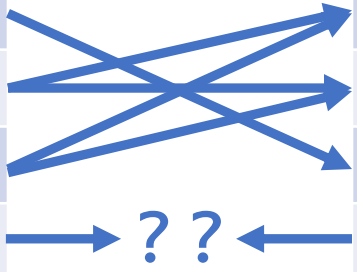
location	name	type
ORC	Squirtle	Water
ORC	Bulbasaur	Grass
ORC	Bulbasaur	Poison
Sloan	Bulbasaur	Grass
Sloan	Bulbasaur	Poison
Sloan	Pikachu	NA

sightings

location	name
ORC	Squirtle
ORC	Bulbasaur
Sloan	Bulbasaur
Sloan	Pikachu

types

name	type
Bulbasaur	Grass
Bulbasaur	Poison
Squirtle	Water
Charmander	Fire




```
sightings %>%
  inner_join(types, by = "name")
```

location	name	type
ORC	Squirtle	Water
ORC	Bulbasaur	Grass
ORC	Bulbasaur	Poison
Sloan	Bulbasaur	Grass
Sloan	Bulbasaur	Poison

```
%>% count(type)
```

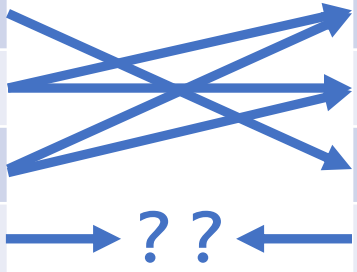
type	n
Water	1
Grass	2
Poison	2

sightings

location	name
ORC	Squirtle
ORC	Bulbasaur
Sloan	Bulbasaur
Sloan	Pikachu

types

name	type
Bulbasaur	Grass
Bulbasaur	Poison
Squirtle	Water
Charmander	Fire




```
sightings %>%
  inner_join(types, by = "name")
```

location	name	type
ORC	Squirtle	Water
ORC	Bulbasaur	Grass
ORC	Bulbasaur	Poison
Sloan	Bulbasaur	Grass
Sloan	Bulbasaur	Poison

```
%>% count(name)
```

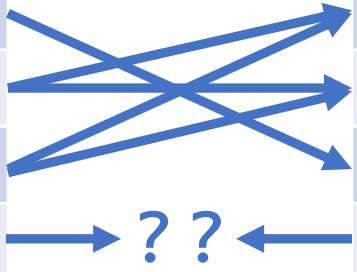
type	n
Squirtle	1
Bulbasaur	4

sightings

location	name
ORC	Squirtle
ORC	Bulbasaur
Sloan	Bulbasaur
Sloan	Pikachu

types

name	type
Bulbasaur	Grass
Bulbasaur	Poison
Squirtle	Water
Charmander	Fire




```
sightings %>%
  inner_join(types, by = "name")
```

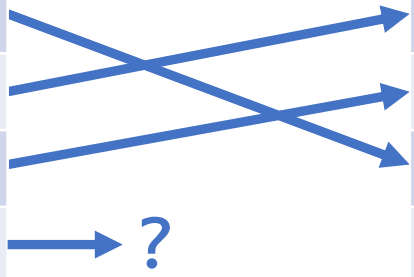
location	name	type
ORC	Squirtle	Water
ORC	Bulbasaur	Grass
ORC	Bulbasaur	Poison
Sloan	Bulbasaur	Grass
Sloan	Bulbasaur	Poison

Moral

Always be aware of row duplication!!!

sightings

location	name
ORC	Squirtle
ORC	Bulbasaur
Sloan	Bulbasaur
Sloan	Pikachu



catch

name	location	catch_pct
Bulbasaur	ORC	50
Bulbasaur	Sloan	75
Squirtle	ORC	90
Squirtle	Sloan	100
Pikachu	ORC	42

A match can be defined by more than one column.

```
sightings %>%
  left_join(catch, by = c("name" = "name", "location" = "location"))
```

location	name	catch_pct
ORC	Squirtle	90
ORC	Bulbasaur	50
Sloan	Bulbasaur	75
Sloan	Pikachu	NA