Joining Data Sets with the Merge() Function

Dan Tofan
SOFTWARE ENGINEER, PHD

@dan_tofan www.programmingwithdan.com

Joining Data Sets: Why Care?

Practical

Different data sources

Transferable

To/from R

Overview

Data frames

Inner joins

Left, right and full outer joins

Data frame keys

Data frame relationships

Summary

Data Frames

Similar to

- R matrix + columns of different types
- Excel worksheet
- Table in a database

Create a data frame

Load a CSV file into a data frame

Inner Joins

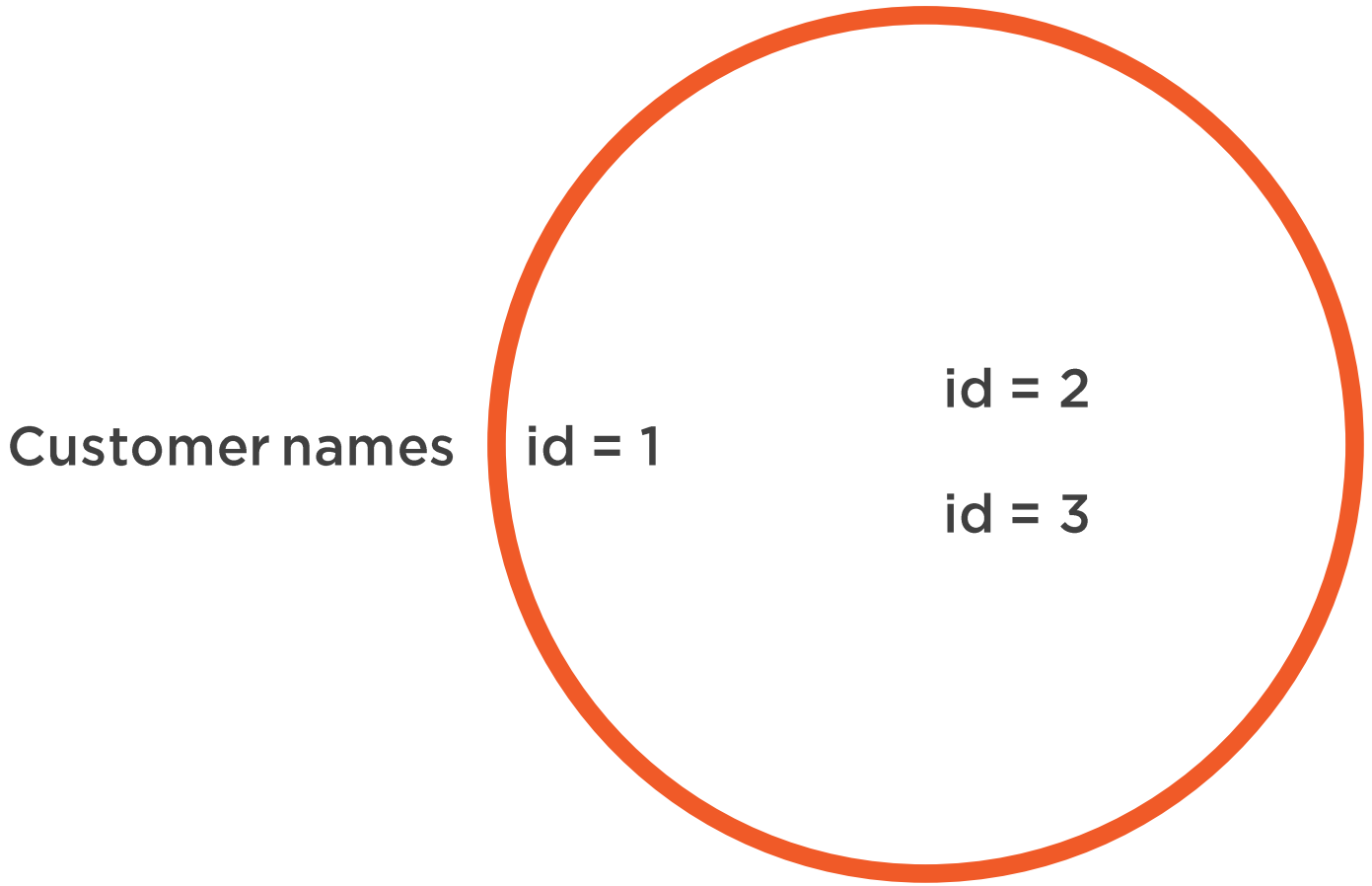
Example

Inner join with merge()

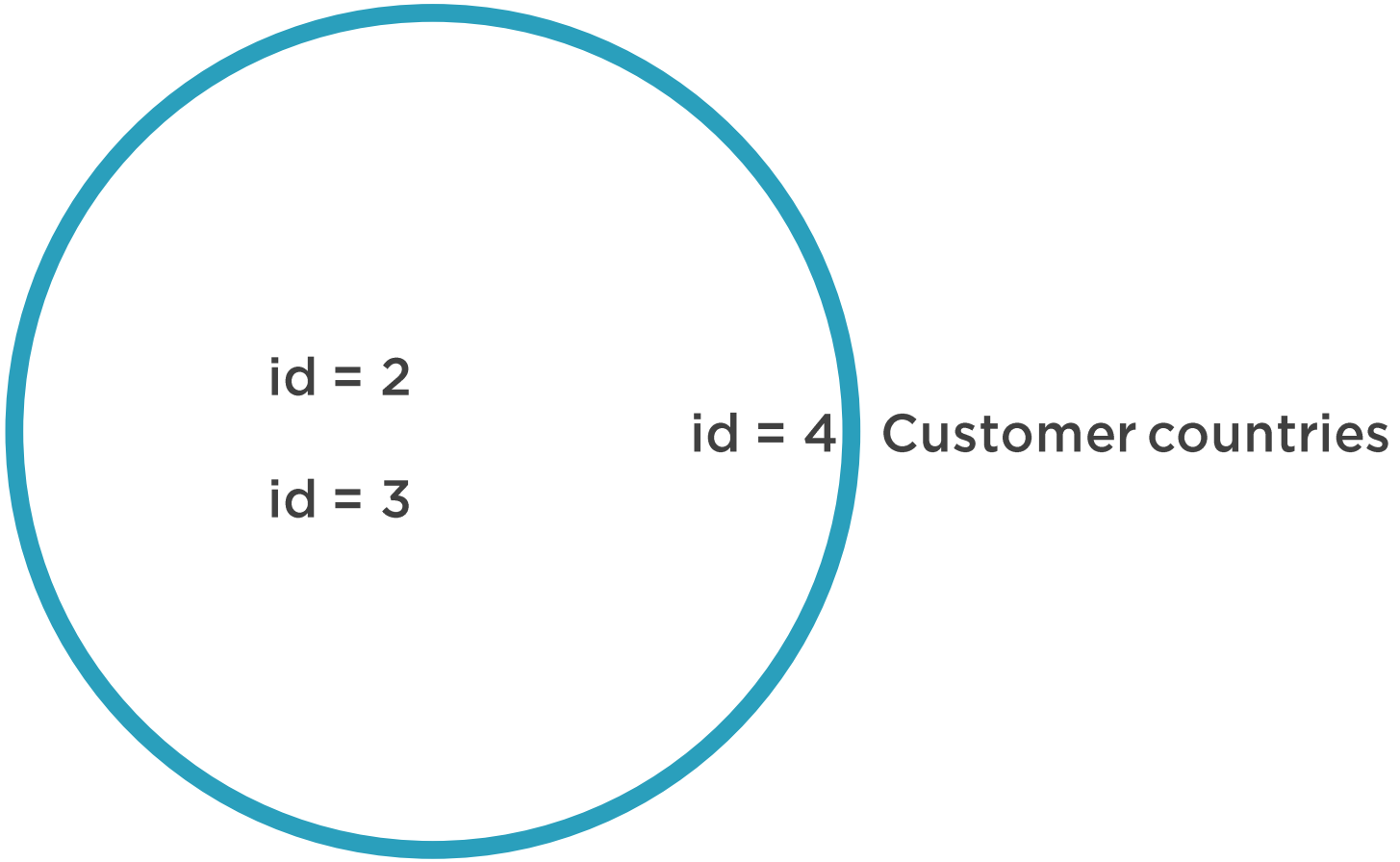
Good habits

- Use unique ids
- Avoid natural joins

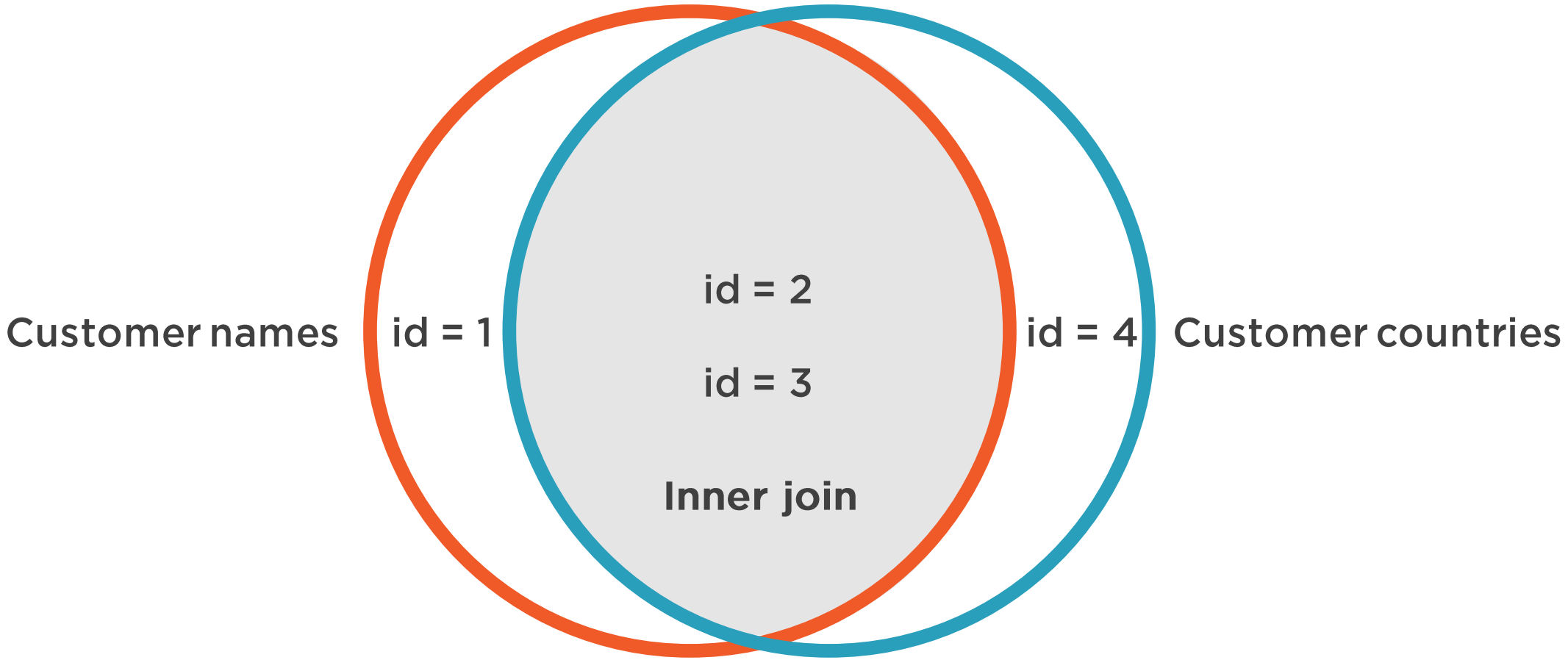
Inner Join by Id Variable



Inner Join by Id Variable



Inner Join by Id Variable



Left, Right, and Full Outer Joins

Left outer joins

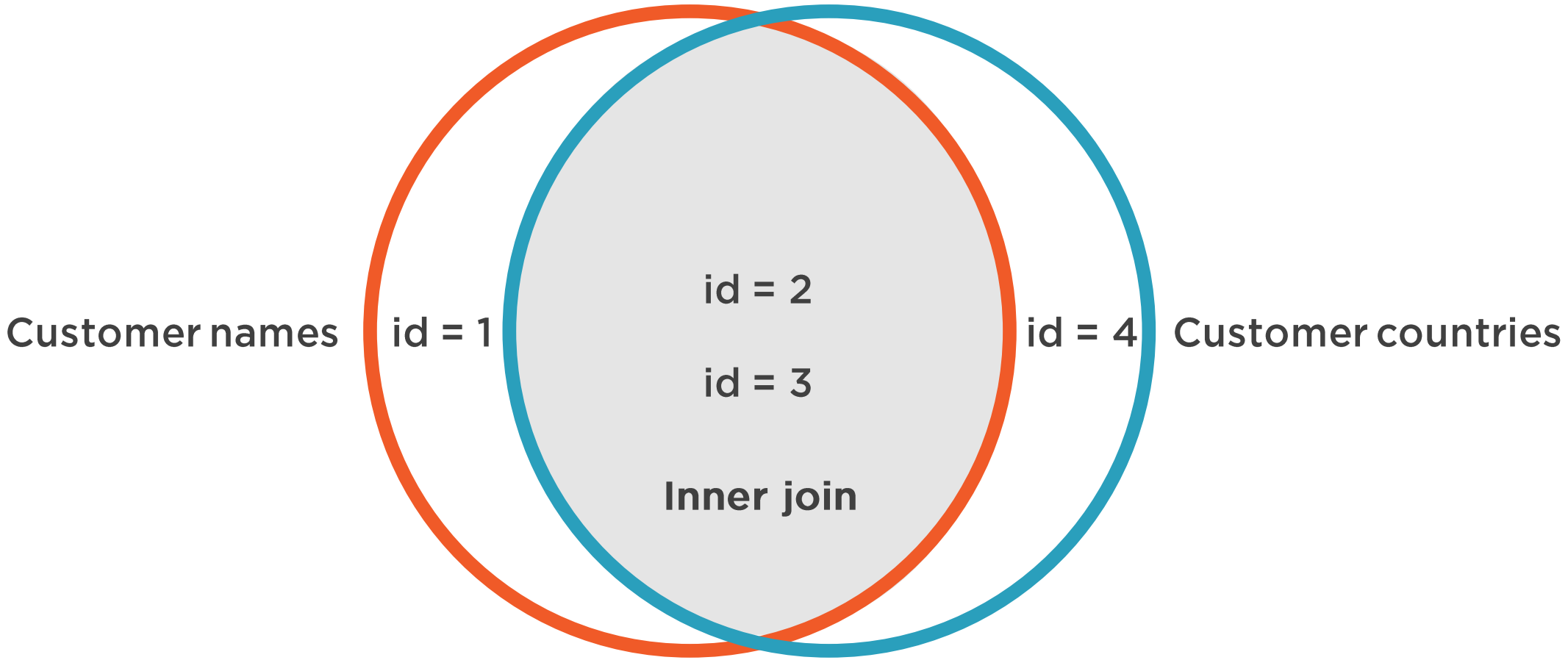
Right outer joins

Full outer joins

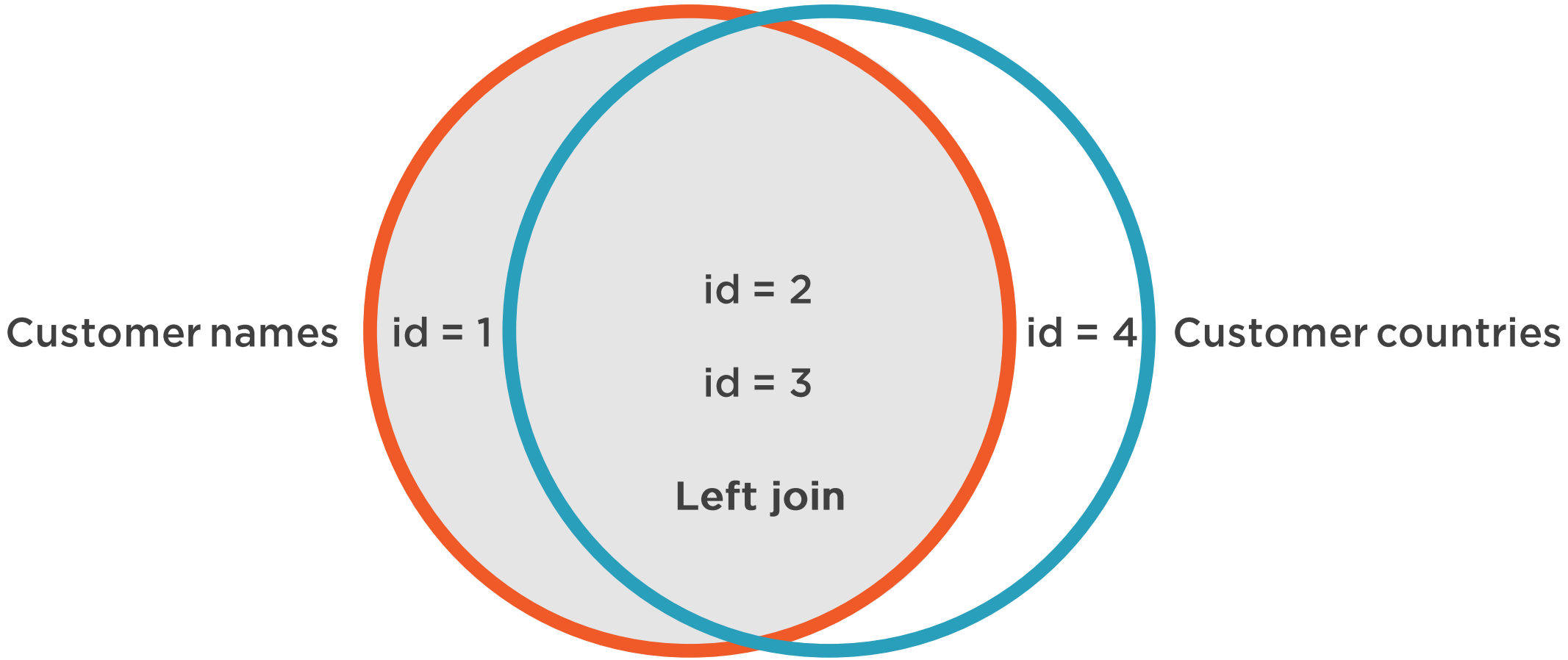
OIO mnemonic

- **O**uter left
- **I**nner
- **O**uter right

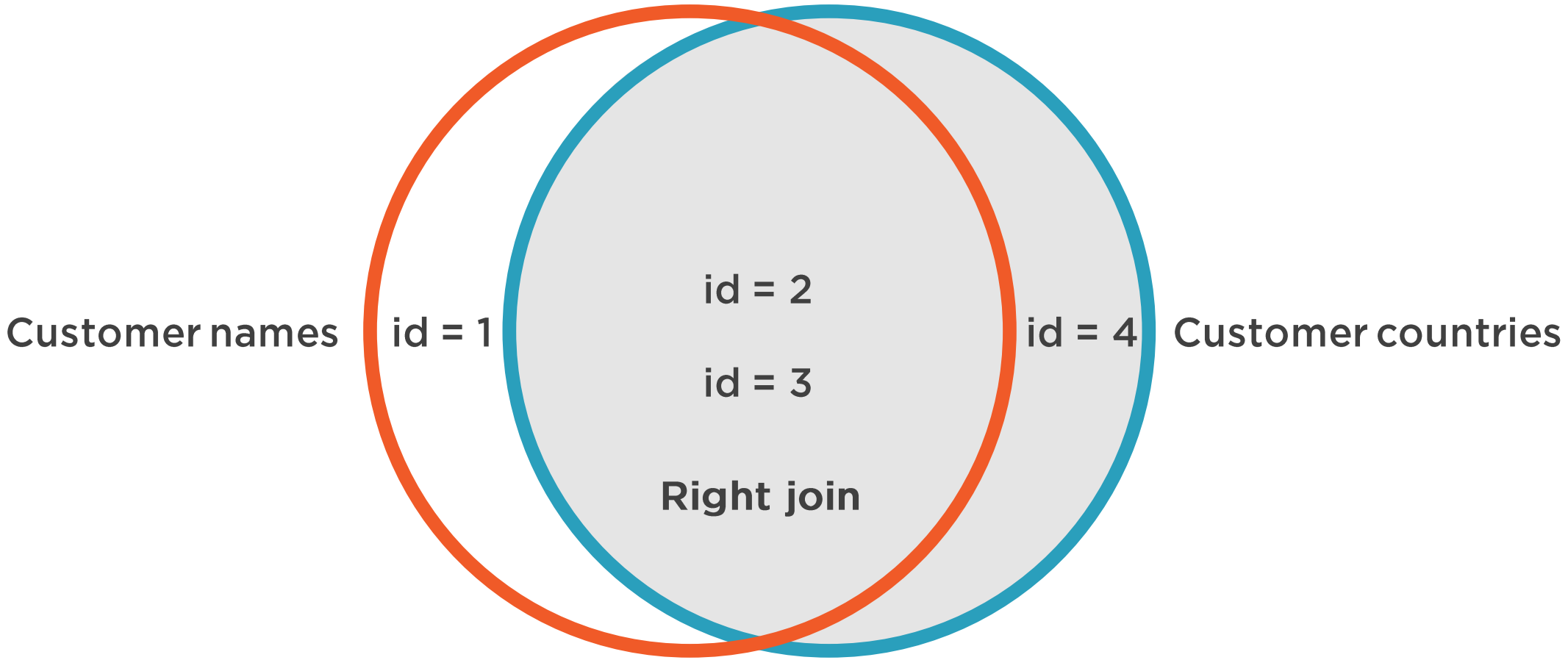
Inner Join by Id Variable



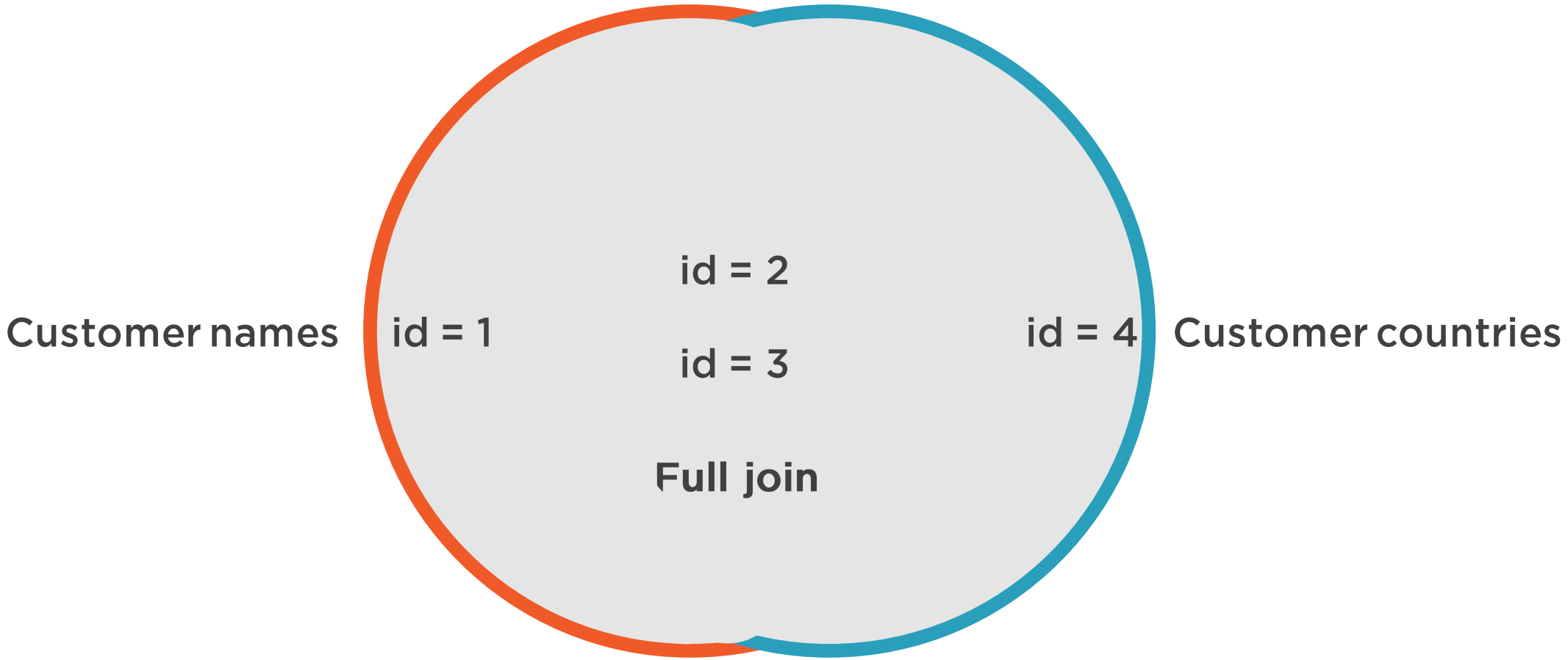
Left Join by Id Variable



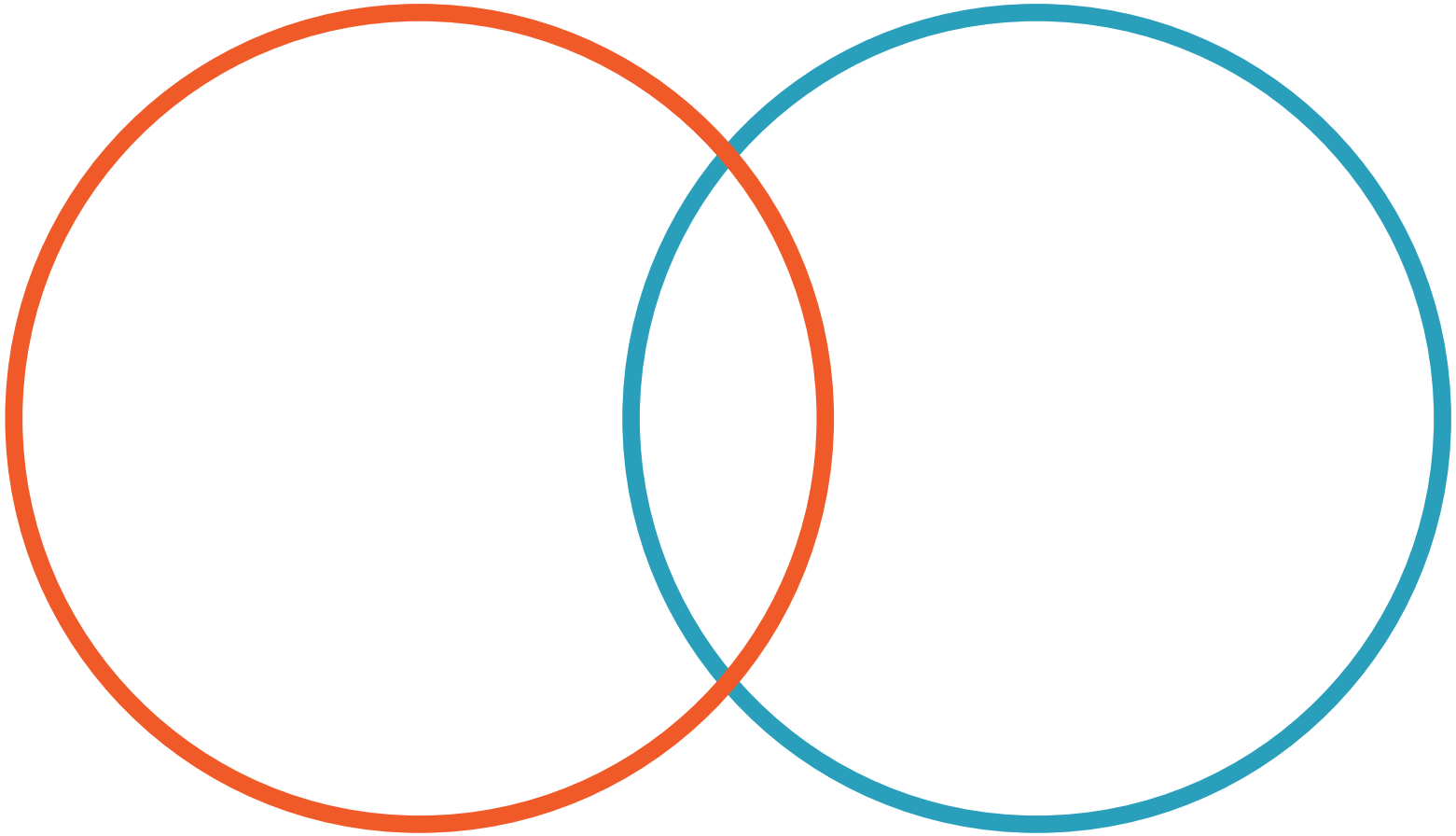
Right Join by Id Variable



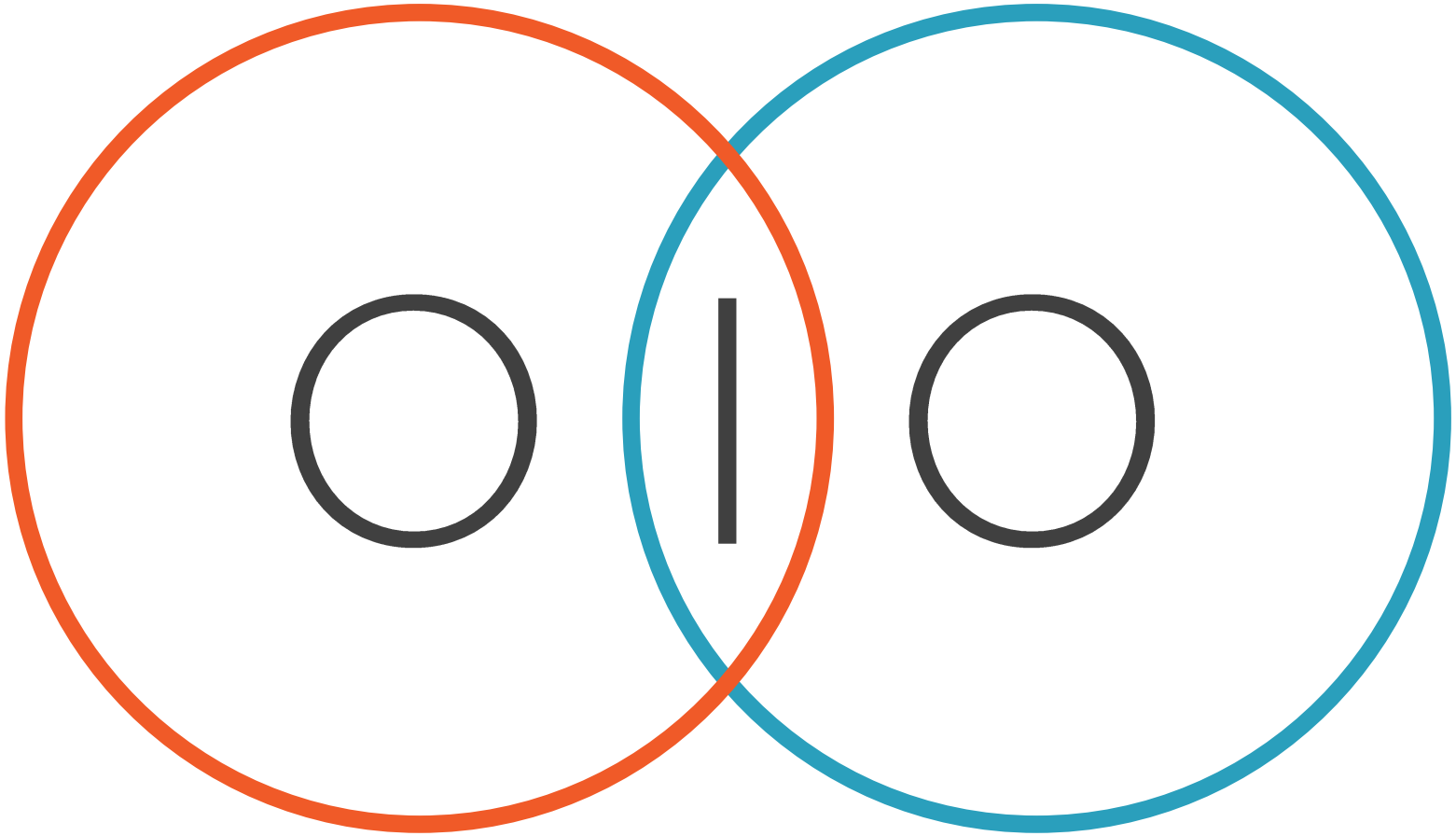
Full Join by Id Variable



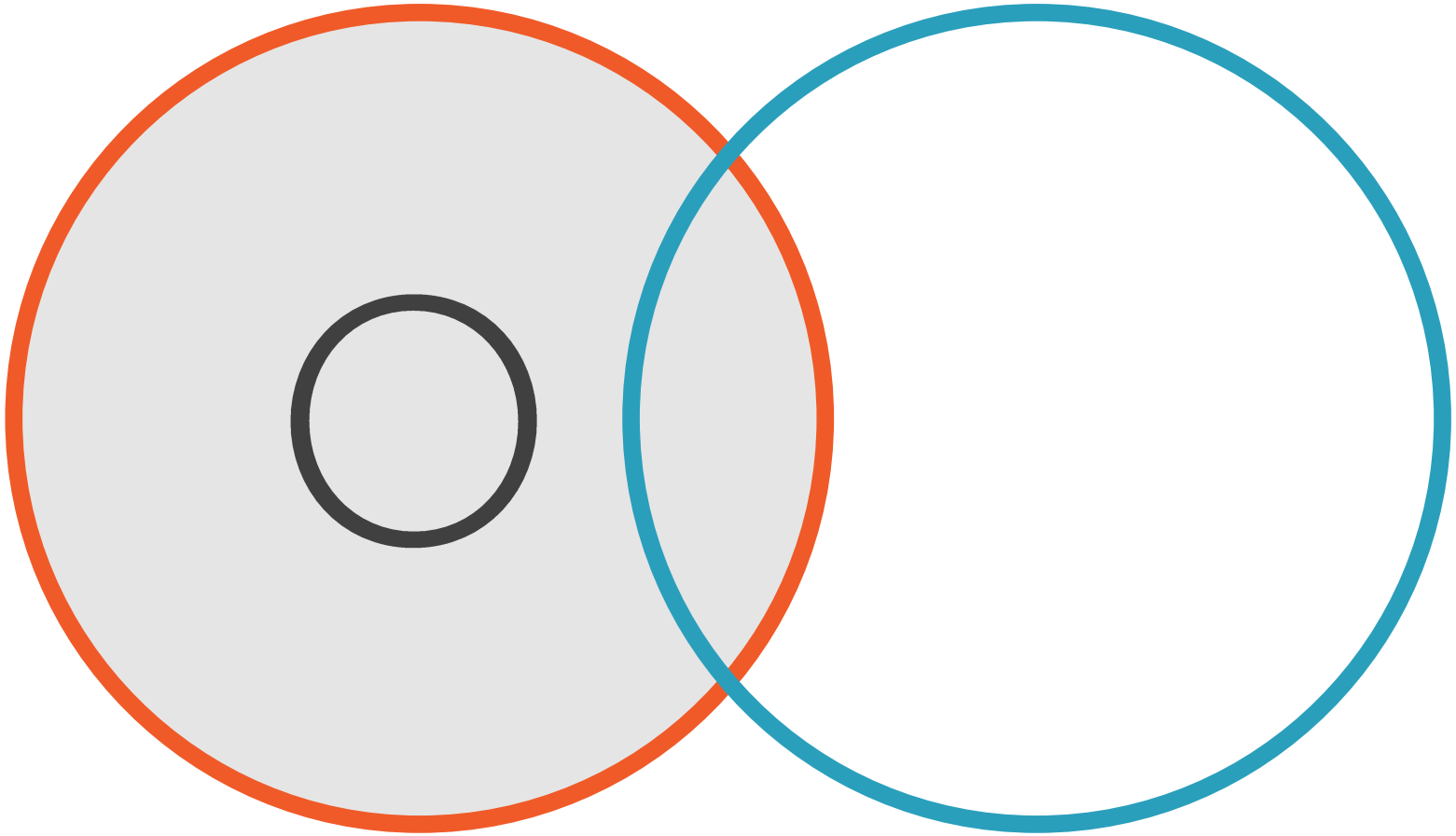
OIO Mnemonic



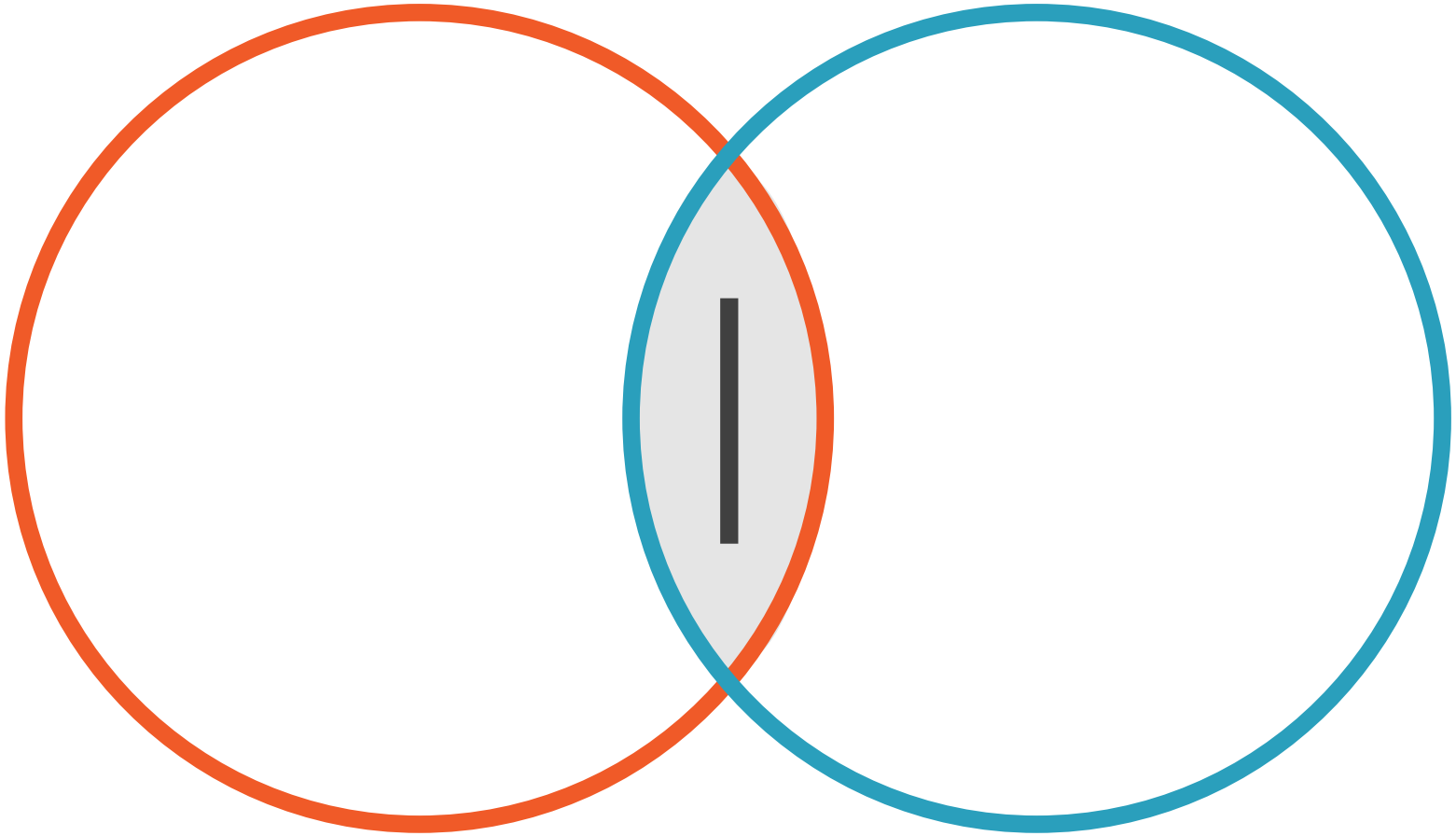
OIO Mnemonic



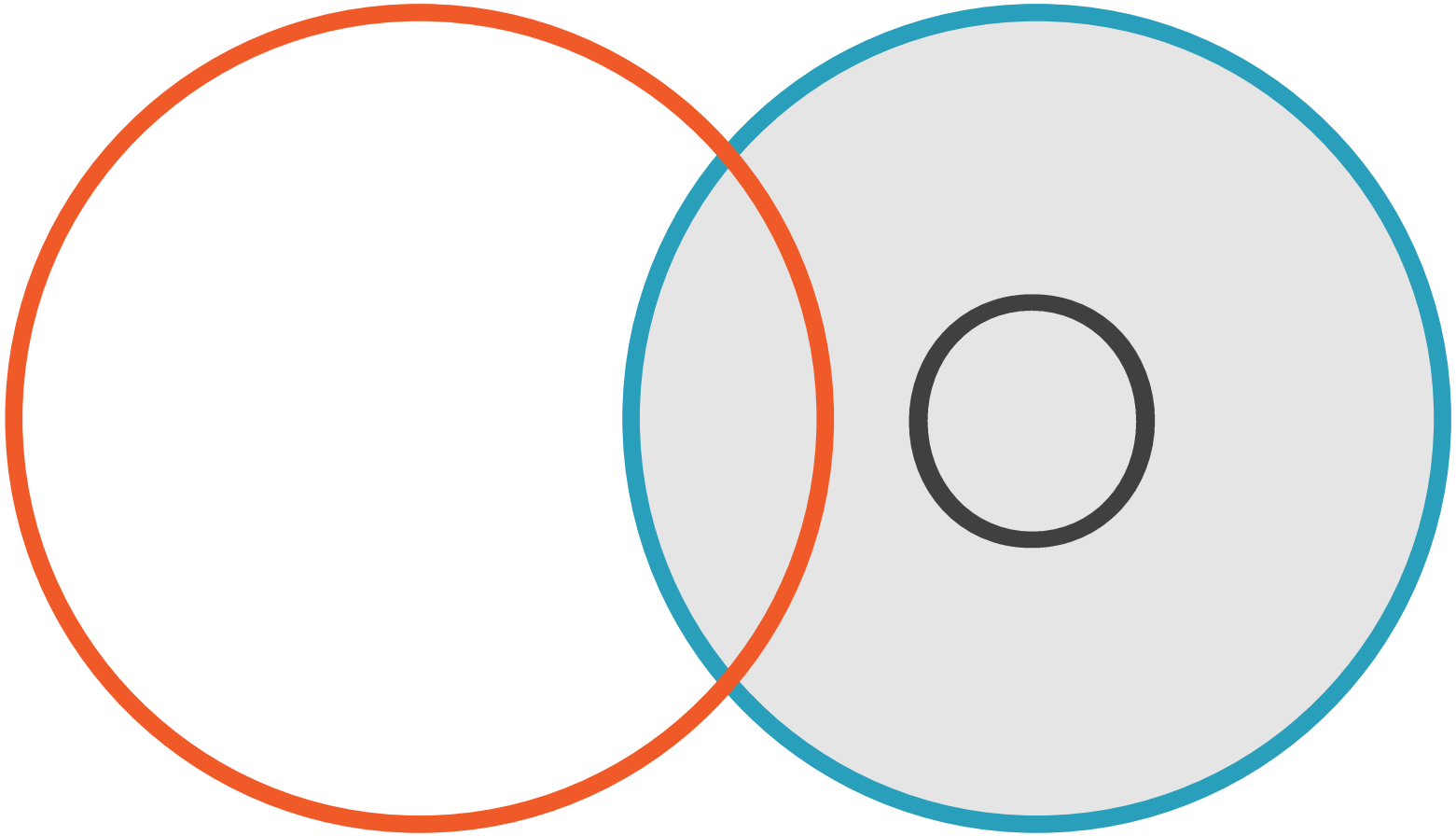
OIO Mnemonic



OIO Mnemonic



OIO Mnemonic



Data Frame Keys

Primary keys	Foreign keys
Column (or combination)	Column (or combination)
Relationship	Relationship
Duplicates are NOT allowed	Duplicates are allowed
Missing values are NOT allowed	Missing values are allowed

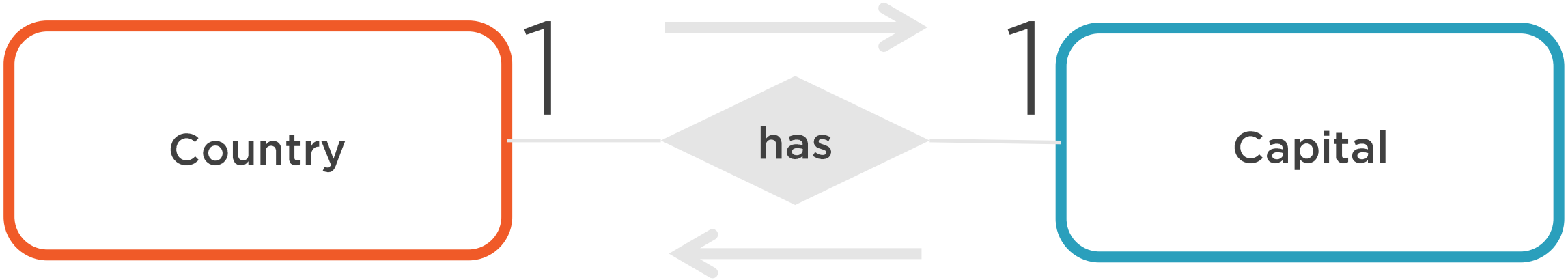
Data Frame Relationships

One-to-one

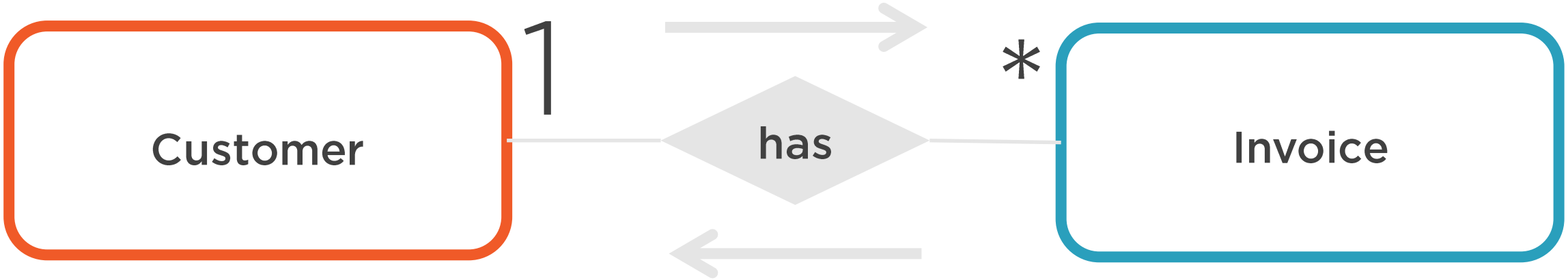
One-to-many

Many-to-many

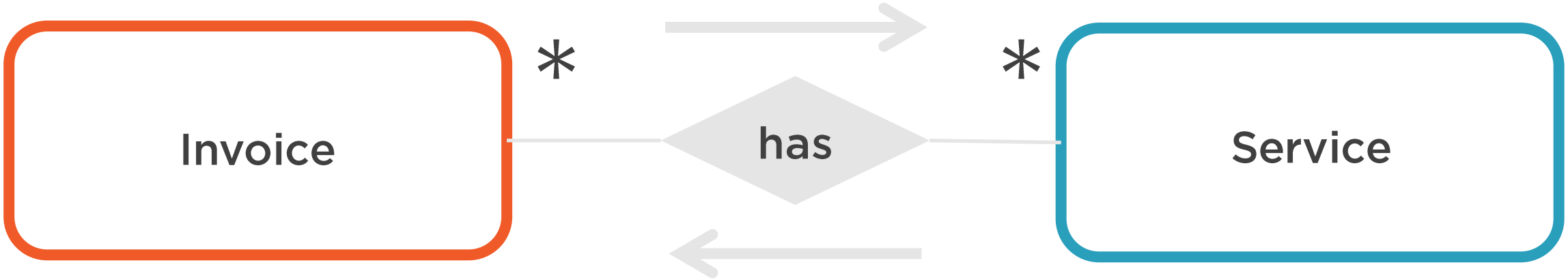
One-to-One Relationship



One-to-Many Relationship



Many-to-Many Relationship



Summary

Data frames

Inner joins

Left, right and full outer joins

Data frame keys

Data frame relationships

Summary