Clase 10:
Unión de bases de datos

Universidad de los Andes
Facultad de Economía

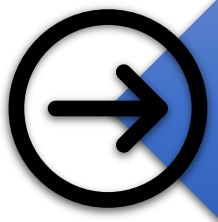
Contenido

1. Conceptos iniciales
   a. Left (Master), Right (Using) y Llave
   b. Tipos de unión
2. Unión vertical de bases
3. Unión horizontal de bases
   a. Inner join
   b. Left join
   c. Right join
   d. Outer join
   e. Semi join
   f. Anti join

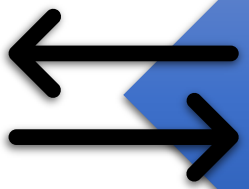
1. Conceptos iniciales

Conceptos iniciales

Left: es la base de datos que recibirá la información. Conocida como *master*

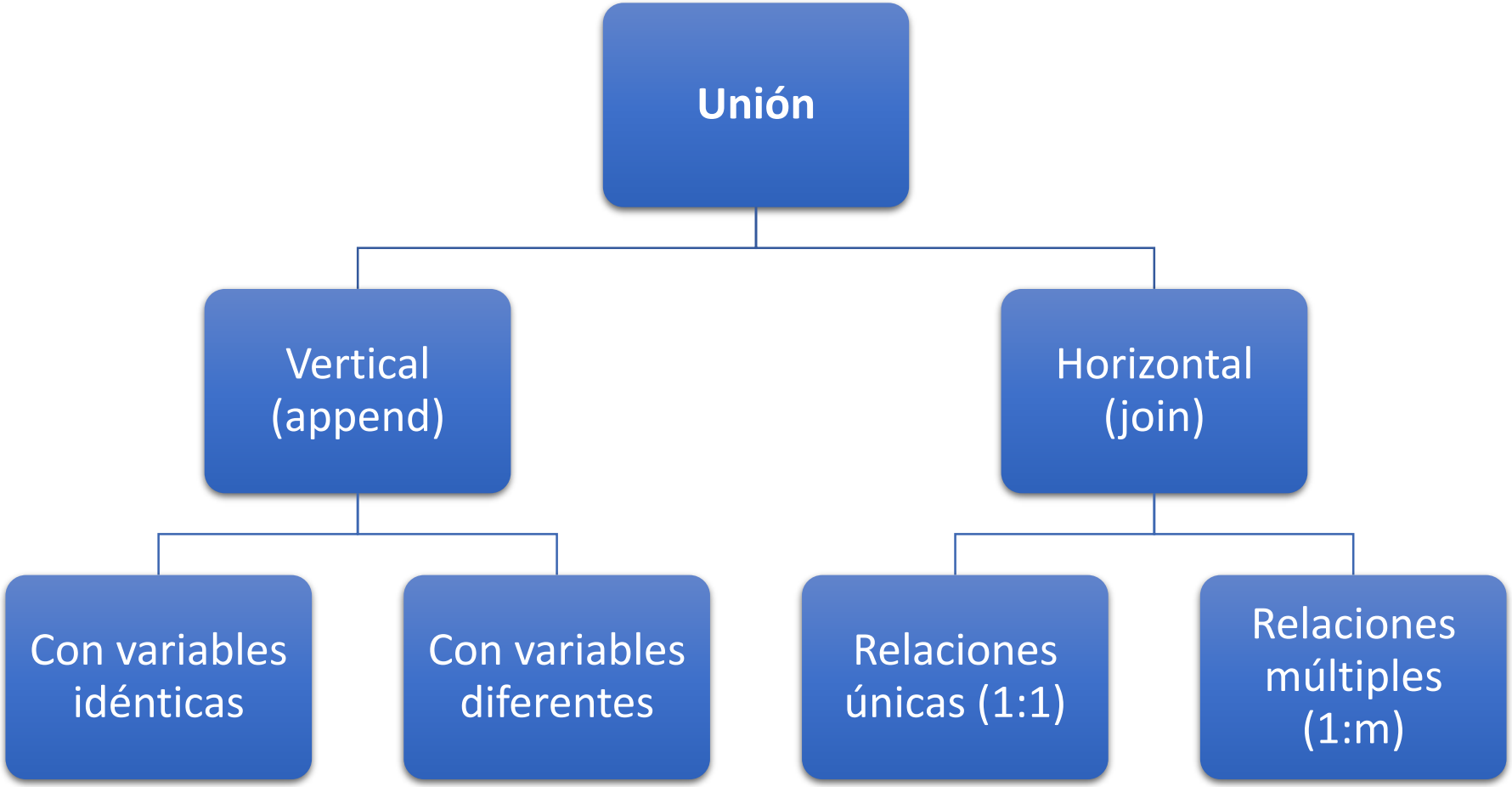


Right: es la base de datos que recibirá la información. Conocida como *using*



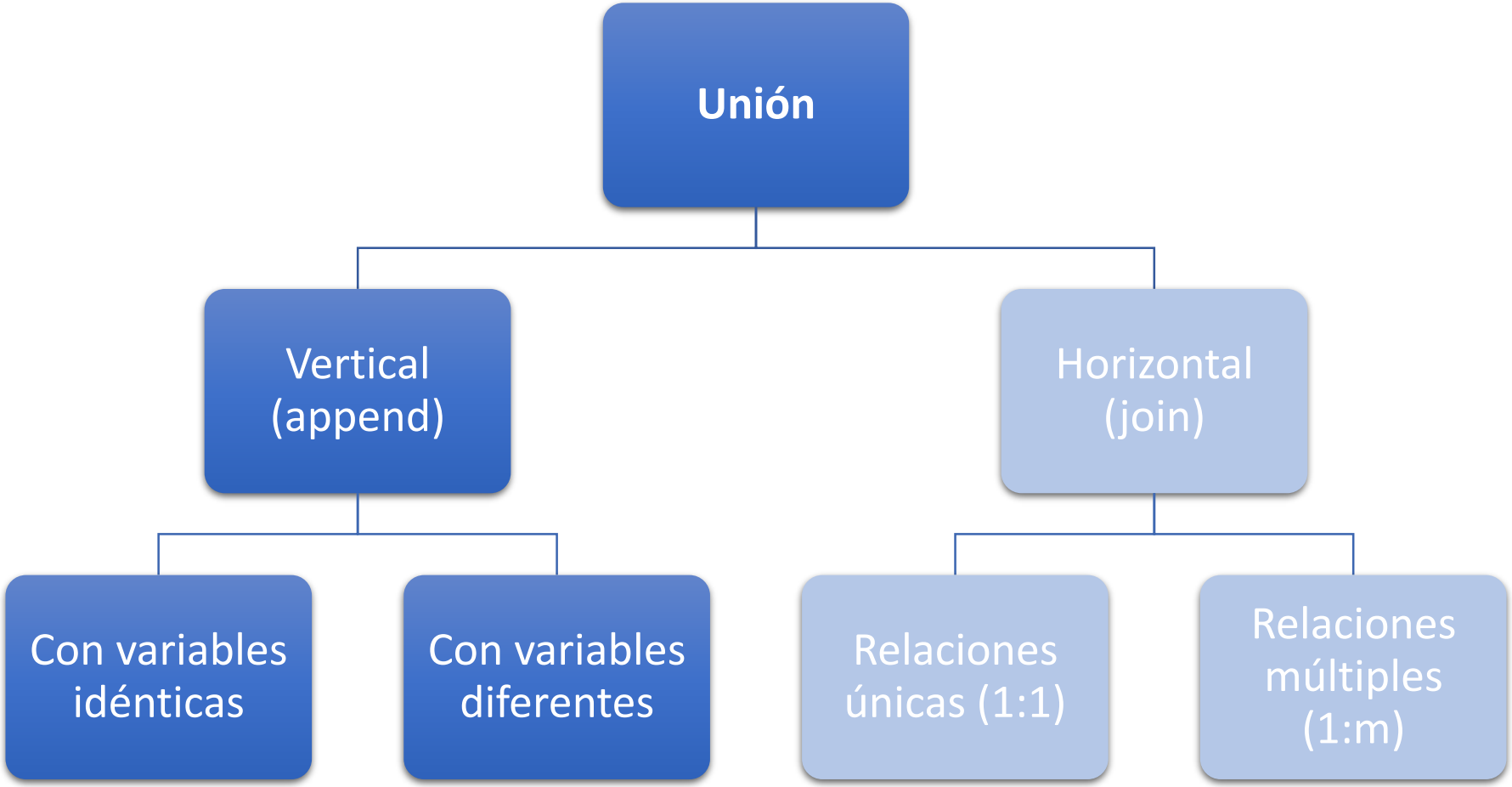
Llave (identificador): es la variable (o variables) que nos permite hacer el emparejamiento. Puede ser una variable numérica o de texto

Tipos de unión



2. Unión vertical de bases

Tipos de unión



Unión vertical: ejemplo (1)

Levantamiento de información en enero

Id	Mes	Edad	Mujer	Trabaja	Ingreso
P1	01	30	1	1	10
P2	01	24	0	1	15
P3	01	32	1	0	13

Levantamiento de información en julio

Id	Mes	Edad	Mujer	Trabaja	Ingreso
P1	07	27	0	0	18
P2	07	23	0	0	7
P3	07	29	1	1	20

Id	Mes	Edad	Sexo	Trabaja	Ingreso
P1	01	30	1	1	10
P2	01	24	0	1	15
P3	01	32	1	0	13
P1	07	27	0	0	18
P2	07	23	0	0	7
P3	07	29	1	1	20

Unión vertical: ejemplo (2)

Levantamiento de información en enero

Id	Mes	Edad	Mujer	Trabaja	Ingreso
P1	01	30	1	1	10
P2	01	24	0	1	15
P3	01	32	1	0	13

Levantamiento de información en julio

Id	Mes	Edad	Mujer	Trabaja	Covid
P1	07	27	0	0	1
P2	07	23	0	0	1
P3	07	29	1	1	0

Id	Mes	Edad	Sexo	Trabaja	Ingreso	Covid
P1	01	30	1	1	10	.
P2	01	24	0	1	15	.
P3	01	32	1	0	13	.
P1	07	27	0	0	.	1
P2	07	23	0	0	.	1
P3	07	29	1	1	.	0

Código

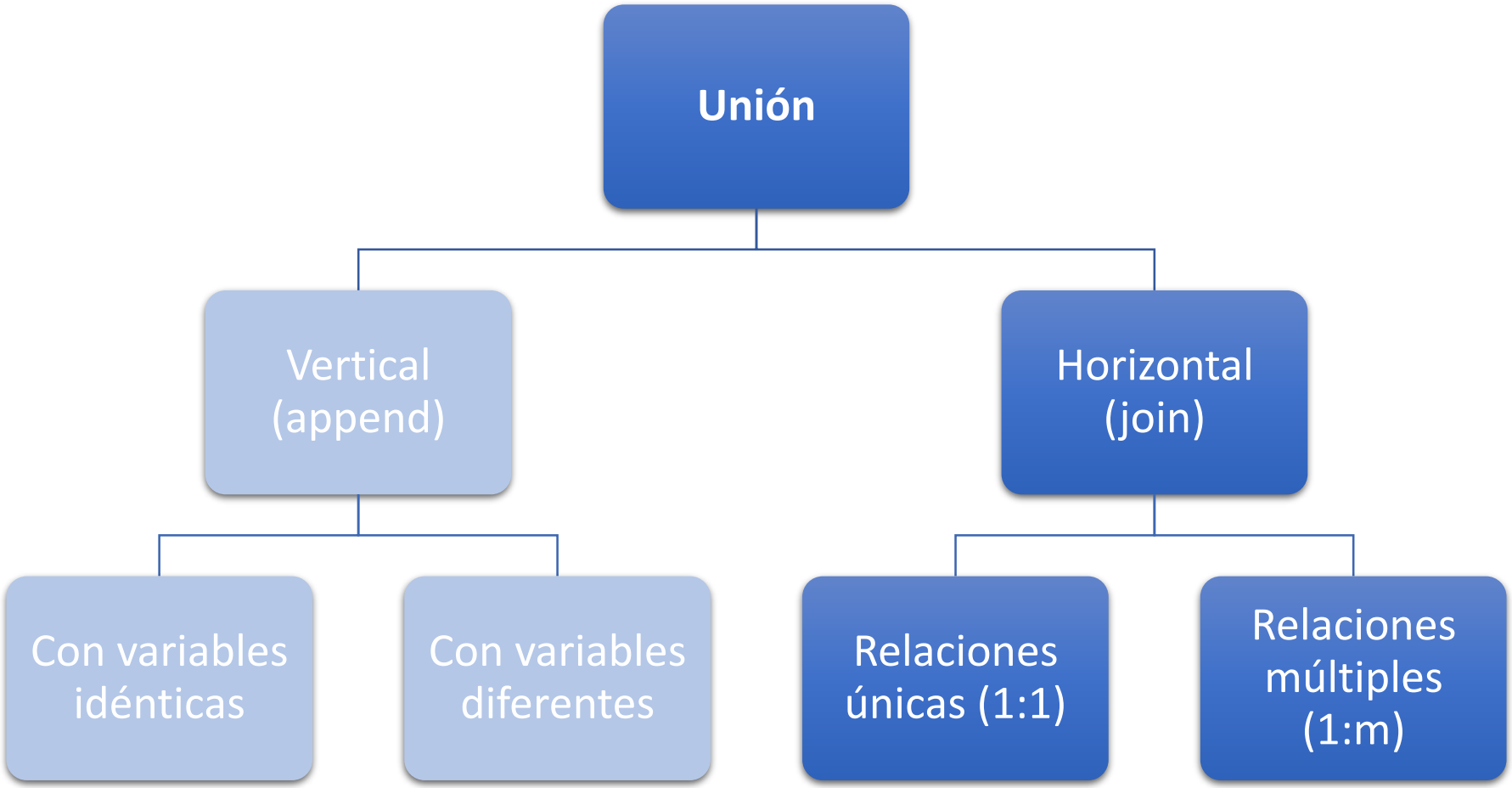
```
append using df2
```

```
data_appended = row_binds(df1, df2)
```

```
data_appended = df1.append(df2)
```

3. Unión horizontal de bases

Tipos de unión



Unión horizontal: ejemplo (rel. única)

Información en enero, módulo demografía

Id	Mes	Edad	Mujer	Trabaja	Ingreso
P1	01	30	1	1	10
P2	01	24	0	1	15
P3	01	32	1	0	13

Información en enero, módulo gastos

Id	Mes	G. fijos	G. ocio
P1	01	20	7
P2	01	19	4
P3	01	23	6

Id	Mes	Edad	Sexo	Trabaja	Ingreso	G. fijos	G. ocio
P1	01	30	1	1	10	20	7
P2	01	24	0	1	15	19	4
P3	01	32	1	0	13	23	6

Unión horizontal: ejemplo (rel. múltiple)

Información en enero, módulo demografía

Id	Mes	Edad	Mujer	Trabaja	Ingreso
P1	01	30	1	1	10
P2	01	24	0	1	15
P3	01	32	1	0	13

Información en enero, módulo vivienda

Id	Mes	V. Propia	Estrato
V1	01	1	2
V1	01	1	2
V2	01	0	1

IdV	IdP	Mes	V. Propia	Estrato
V1	P1	01	1	2
V1	P2	01	1	2
V2	P3	01	0	1

Inner join (keep 3)

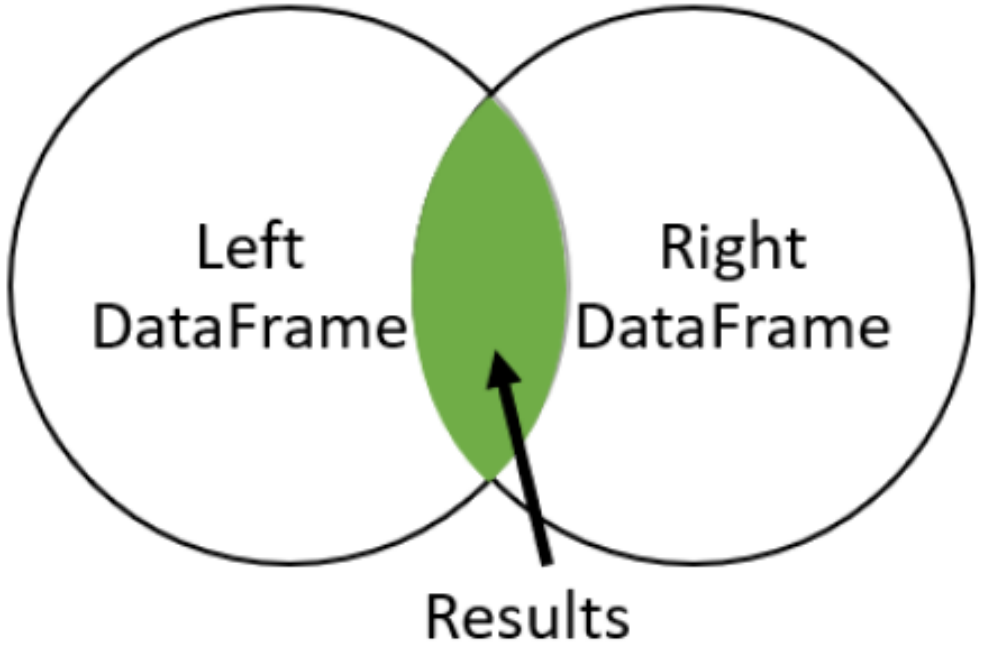
Llave	Var1
1	A
2	B
3	C

⟷

Llave	Var2
1	X
3	Y
4	Z

=

Llave	Var1	Var2
1	A	X
3	C	Y



¿Cuando debo usarlo?

Si nos interesa únicamente las observaciones en común entre dos bases de datos.

Código

```
merge 1:m key using df2, keep(3)
```

```
inner = inner_join(df1, df2, by = “key")
```

```
inner = df1.merge(df2, on=“key")
```

Left join (keep 1, 3)

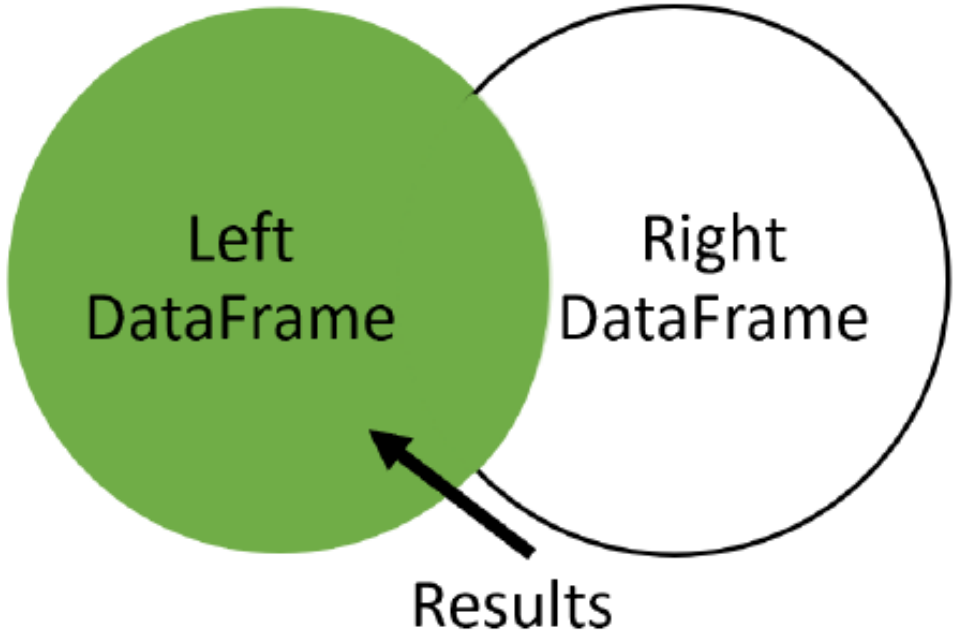
Llave	Var1
1	A
2	B
3	C

⟷

Llave	Var2
1	X
3	Y
4	Z

=

Llave	Var1	Var2
1	A	X
2	B	-
3	C	Y



¿Cuando debo usarlo?

Si nos interesa complementar las observaciones que están en la base de datos del lado izquierdo con las observaciones del lado derecho.

Código

```
merge 1:m key using df2, keep(1, 3)
```

```
left = left_join(df1, df2, by = “key")
```

```
left = df1.merge(df2, on=“key", how=“left")
```

Right join (keep 2 3)

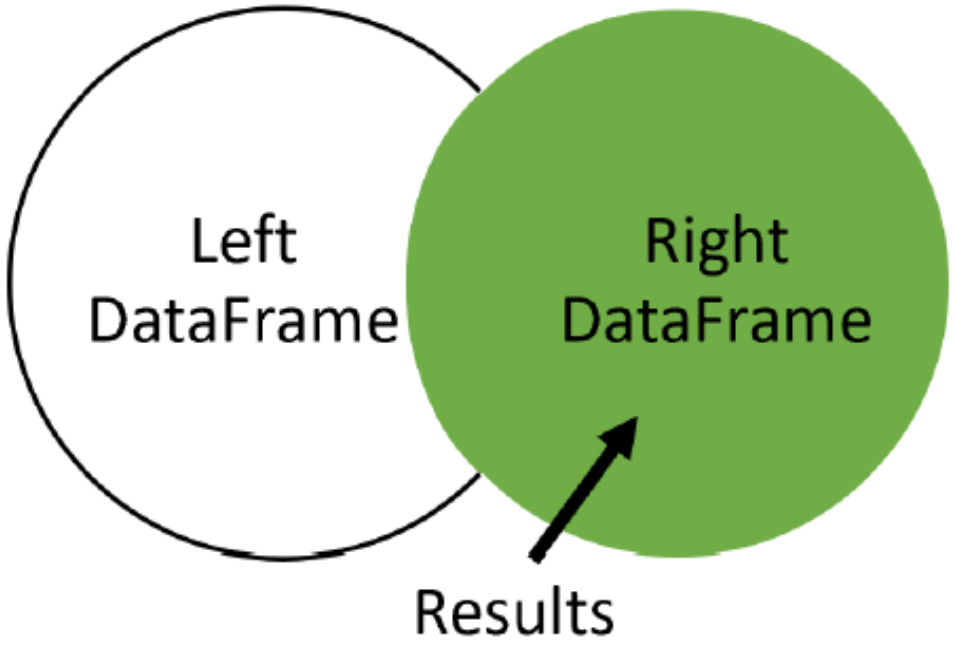
Llave	Var1
1	A
2	B
3	C

⟷

Llave	Var2
1	X
3	Y
4	Z

=

Llave	Var2	Var1
1	X	A
3	Y	C
4	Z	-



¿Cuando debo usarlo?

Si nos interesa complementar las observaciones que están en la base de datos del lado derecho con las observaciones del lado izquierdo

Código

```
merge 1:m key using df2, keep(2, 3)
```

```
right = right_join(df1, df2, by = “key")
```

```
right = df1.merge(df2, on=“key", how=“right")
```

Outer join (keep 1 2 3)

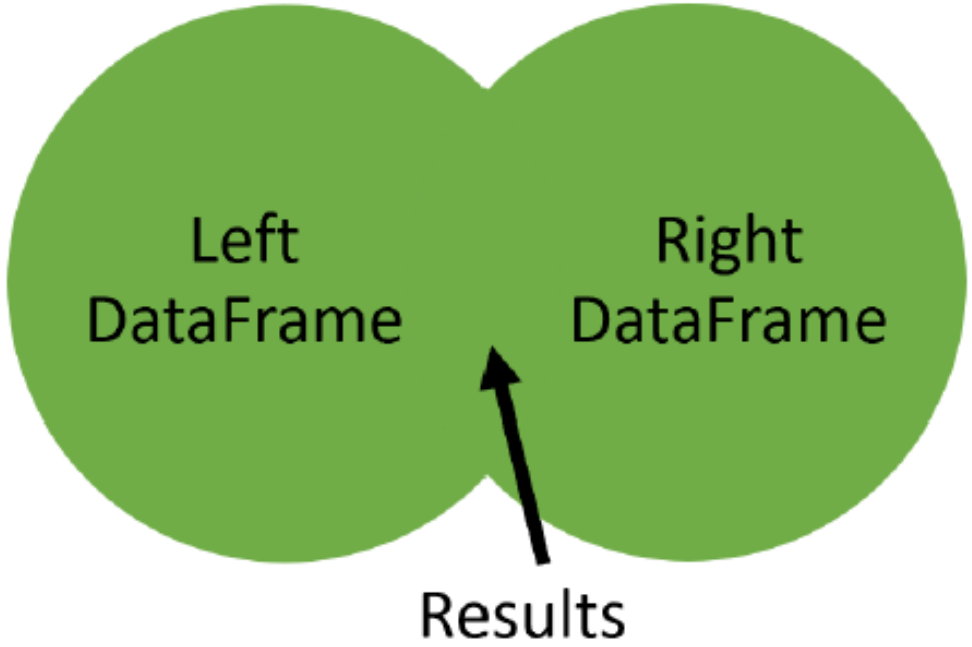
Llave	Var1
1	A
2	B
3	C

⟷

Llave	Var2
1	X
3	Y
4	Z

=

Llave	Var1	Var2
1	A	X
2	B	-
3	C	Y
4	-	Z



¿Cuando debo usarlo?
Si nos interesa unir las observaciones de dos bases de datos.

Código

```
merge 1:m key using df2, keep(1,2, 3)
```

```
outer = full_join(df1, df2, by = “key")
```

```
outer = df1.merge(df2, on=“key", how=“outer")
```

Semi join

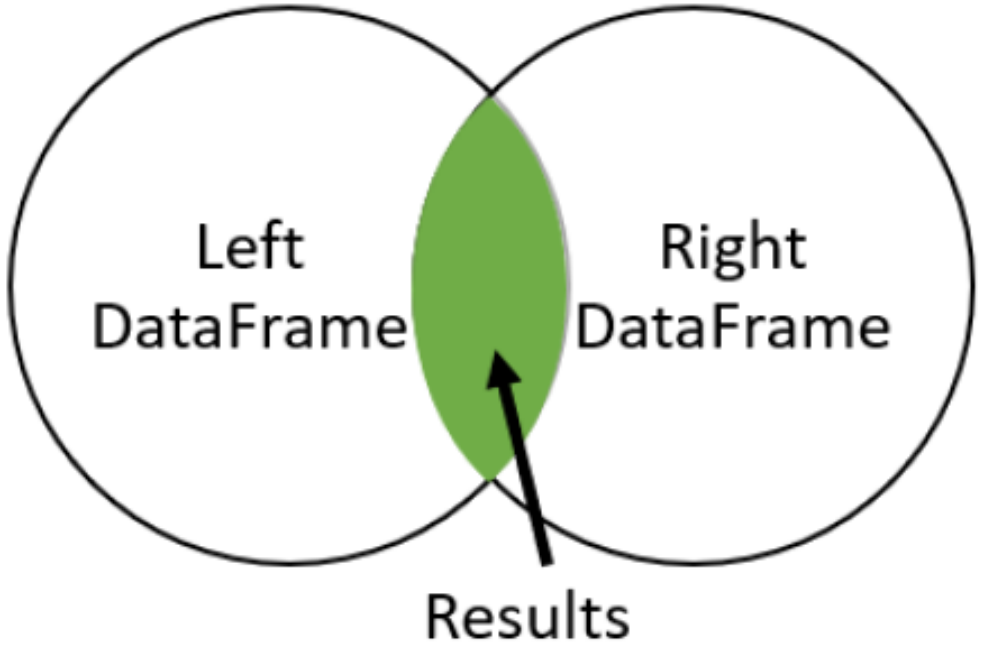
Llave	Var1
1	A
2	B
3	C

⟷

Llave	Var2
1	X
3	Y
4	Z

=

Llave	Var1	Var2
1	A	X
3	C	Y



¿Cuando debo usarlo?

Si nos interesa quedarnos con las observaciones del lado izquierdo solo si dichas observaciones son comunes entre dos bases de datos. A diferencia del *inner join*, *semi join* no genera duplicados

Código

```
merge 1:m key using df2, keep(3) keepusing(key)
```

```
semi = semi_join(df1, df2, by = “key")
```

```
semi = df1.merge(df2, on=“key”)
df1[df1[“key”].isin(semi[“key”])]
```

Anti join

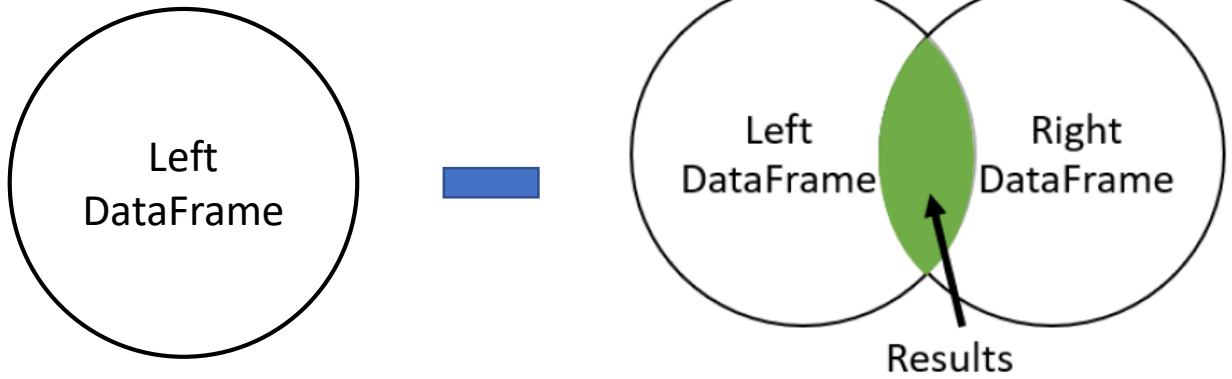
Llave	Var1
1	A
2	B
3	C

⟷

Llave	Var2
1	X
3	Y
4	Z

=

Llave	Var1
2	B



¿Cuando debo usarlo?

Si nos interesa quedarnos con las observaciones del lado izquierdo que no sean comunes entre dos bases de datos. Retorna solo las columnas de la base de la izquierda

Código

```
merge 1:m key using df2, , keep(1) keepusing(key)
```

```
anti = anti_join(df1, df2)
```

```
anti = df1.merge(df2, on="key", how="left", indicator=True)
list=anti.loc[anti["_merge"] == "left_only", "key"]
df1[df1["key"].isin(list)]
```

Validaciones

- Es necesario verificar que la unión sea consistente
- Se puede hacer de forma manual verificando los emparejamientos
- También es posible verificar con ayuda del programa
- La uniones *many to many* no son recomendadas puesto que pueden crear bases de datos inconsistentes

Ejercicio

Usted es el asesor económico principal de la presidenta de Colombia. Ella le proporciona las siguientes bases de datos:

Nombre	¿Qué mide?	Identificador	Tamaño
GEIH_personas	Características generales de las personas	IdV IdP	$(6{,}100)$
GEIH_ocupados_área	Características de personas con trabajo en zonas urbanas	IdV IdP	$(7{,}100)$
GEIH_ocupados_cabecera	Características de personas con trabajo en centros poblados	IdV IdP	$(7{,}100)$
GEIH_viviendas	Características de las viviendas	IdV	$(5{,}100)$

Son una submuestra de la Gran encuesta integrada de hogares (GEIH) de marzo de 2021.

Para cada una de las siguientes situaciones:

1. Identifique cuales es la llave (o las llaves) que utilizaría para unir las bases

2. Indique el tipo de unión que realizaría: horizontal o vertical

3. Si la unión es horizontal indique cual es el procedimiento a aplicar: inner, left, right, outer, semi, anti join.

4. Intente calcular el número de columnas que tendría la base después de la unión

Para cada una de las siguientes situaciones:

1. La presidenta quiere conocer el número de personas que tienen acceso a acueducto para proponer una política pública de acceso al agua

2. La presidenta quiere conocer el número de mujeres que trabajan en los centros poblados para proponer una política pública que cierre la brecha del mercado laboral para las mujeres

3. La presidenta quiere conocer el promedio de ingresos de quienes trabajan en los centros poblados comparado con los ingresos de las zonas urbanas para proponer una política pública para cerrar la brecha salarial entre campo y ciudad

Referencias útiles

- R for STATA users: https://www.matthieugomez.com/statar/join-and-reshape.html
- Stata to Python Equivalents: http://www.danielmsullivan.com/pages/tutorial_stata_to_python.html#merging-and-joining
- Merge/Append using Stata: https://www.princeton.edu/~otorres/Merge101.pdf