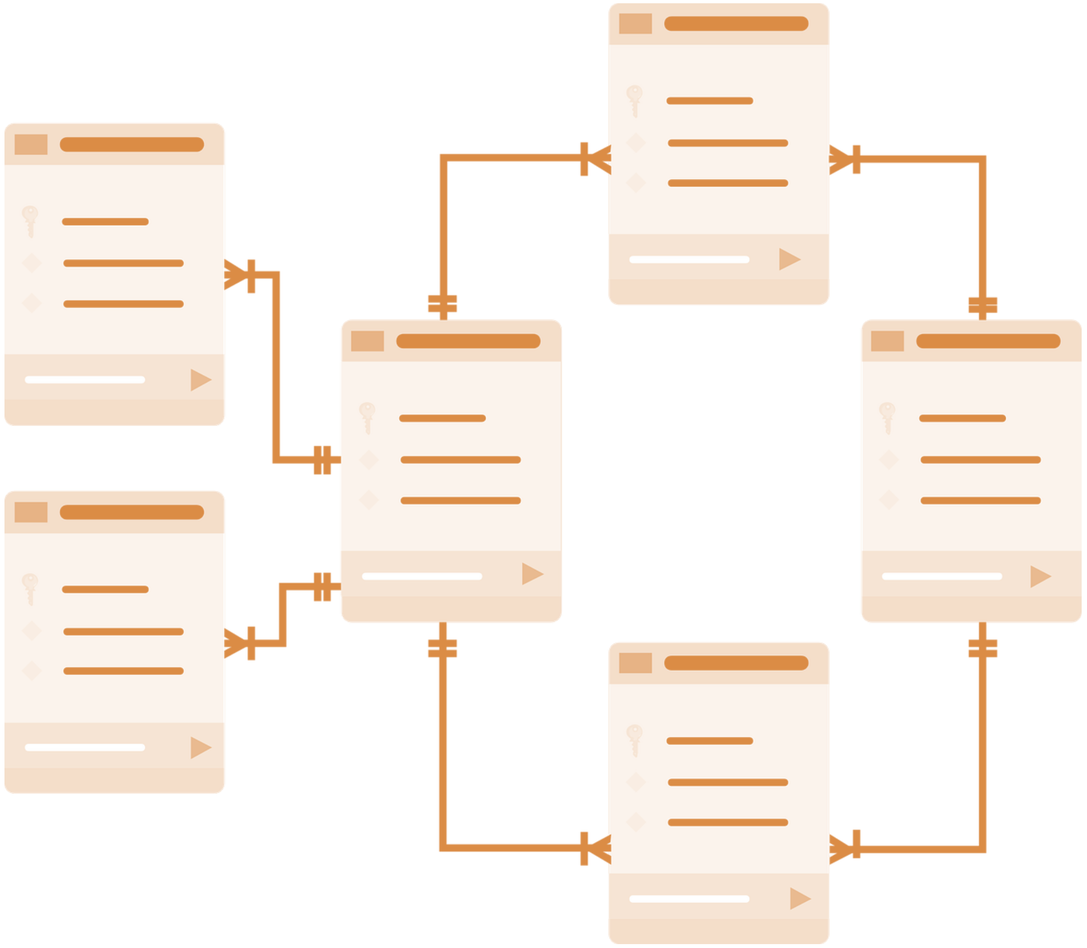
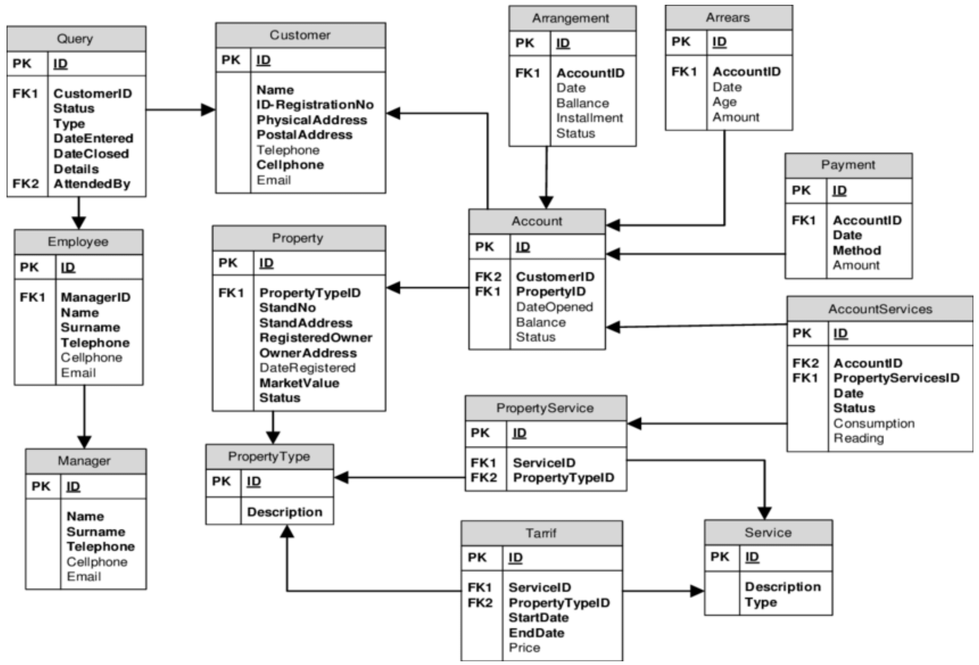
Relational Data
Datos Relacionales



- Típicamente el análisis de de un cierto problema involucra más de un *dataset* / tabla de datos

  Se le llama **datos relacionales** a esas múltiples tablas de datos, ya que sus relaciones, y no solo los conjuntos de datos individuales, son importantes.

Las relaciones siempre se definen sobre un par de tablas



El típico proceso de integración de datos consiste en:

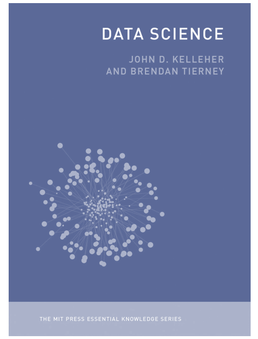
-Extracción: ¿diferentes interfaces?

-Limpieza: clase anterior (formato de mismos atributos, tratamiento de datos faltantes con mismo criterio).

-Estandarización: evaluar los atributos de la misma forma en todos los datasets.

-Transformación: realizar las transformaciones necesarias para que la nueva tabla de datos sume al análisis de nuestro problema

-Integración



aerolineas

```
#> # A tibble: 16 × 2
#>   aerolinea nombre
#>   <chr>     <chr>
#> 1 9E        Endeavor Air Inc.
#> 2 AA        American Airlines Inc.
#> 3 AS        Alaska Airlines Inc.
#> 4 B6        JetBlue Airways
#> 5 DL        Delta Air Lines Inc.
#> 6 EV        ExpressJet Airlines Inc.
```

aviones

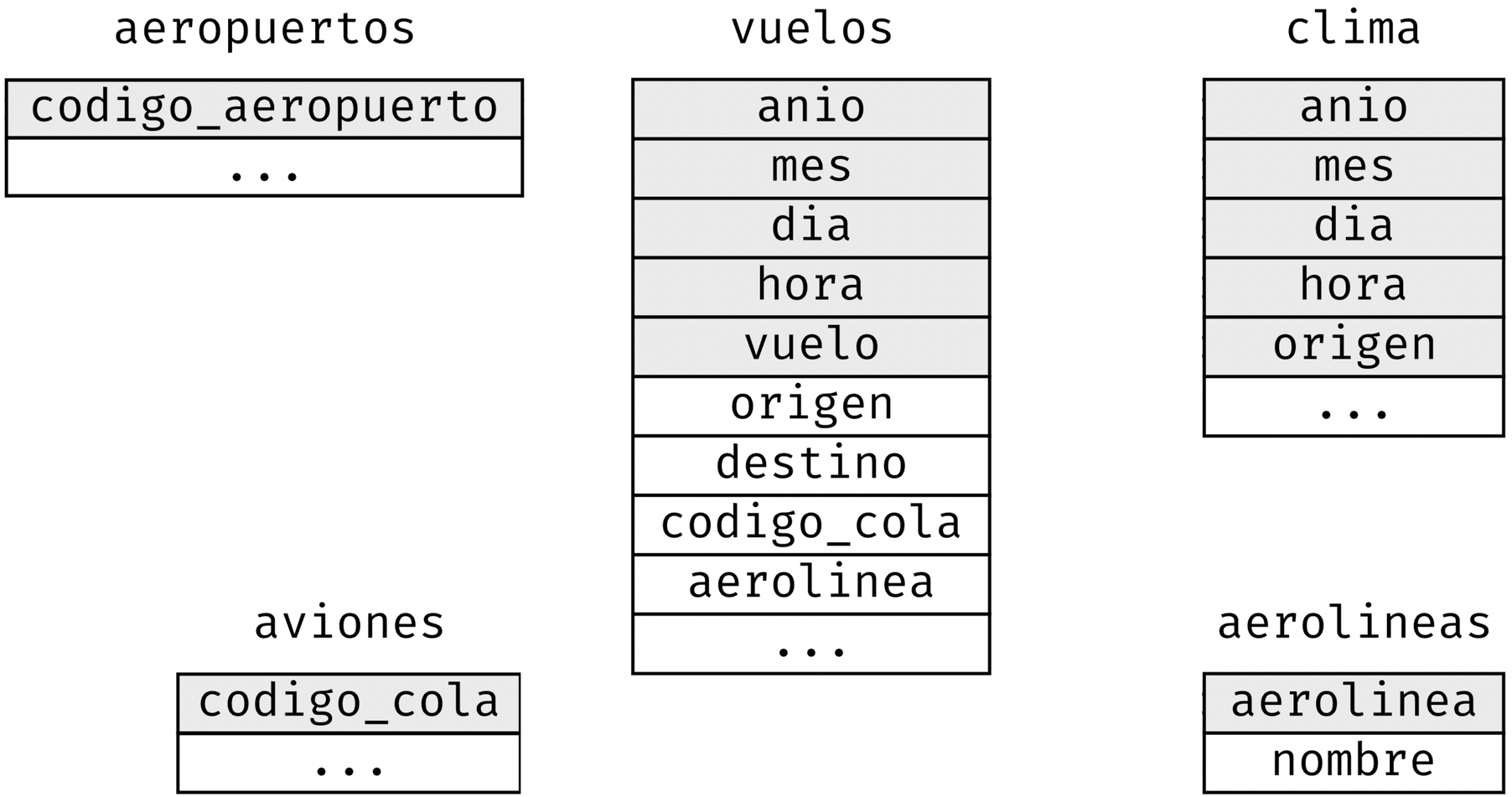
```
#> # A tibble: 3,322 × 9
#>   codigo_cola  anio tipo         fabricante   modelo  motores asientos velocidad
#>   <chr>       <int> <chr>        <chr>        <chr>     <int>    <int>     <int>
#> 1 N10156       2004 Ala fija mu… EMBRAER      EMB-14…       2       55        NA
#> 2 N102UW       1998 Ala fija mu… AIRBUS INDU… A320-2…       2      182        NA
#> 3 N103US       1999 Ala fija mu… AIRBUS INDU… A320-2…       2      182        NA
#> 4 N104UW       1999 Ala fija mu… AIRBUS INDU… A320-2…       2      182        NA
#> 5 N10575       2002 Ala fija mu… EMBRAER      EMB-14…       2       55        NA
#> 6 N105UW       1999 Ala fija mu… AIRBUS INDU… A320-2…       2      182        NA
```

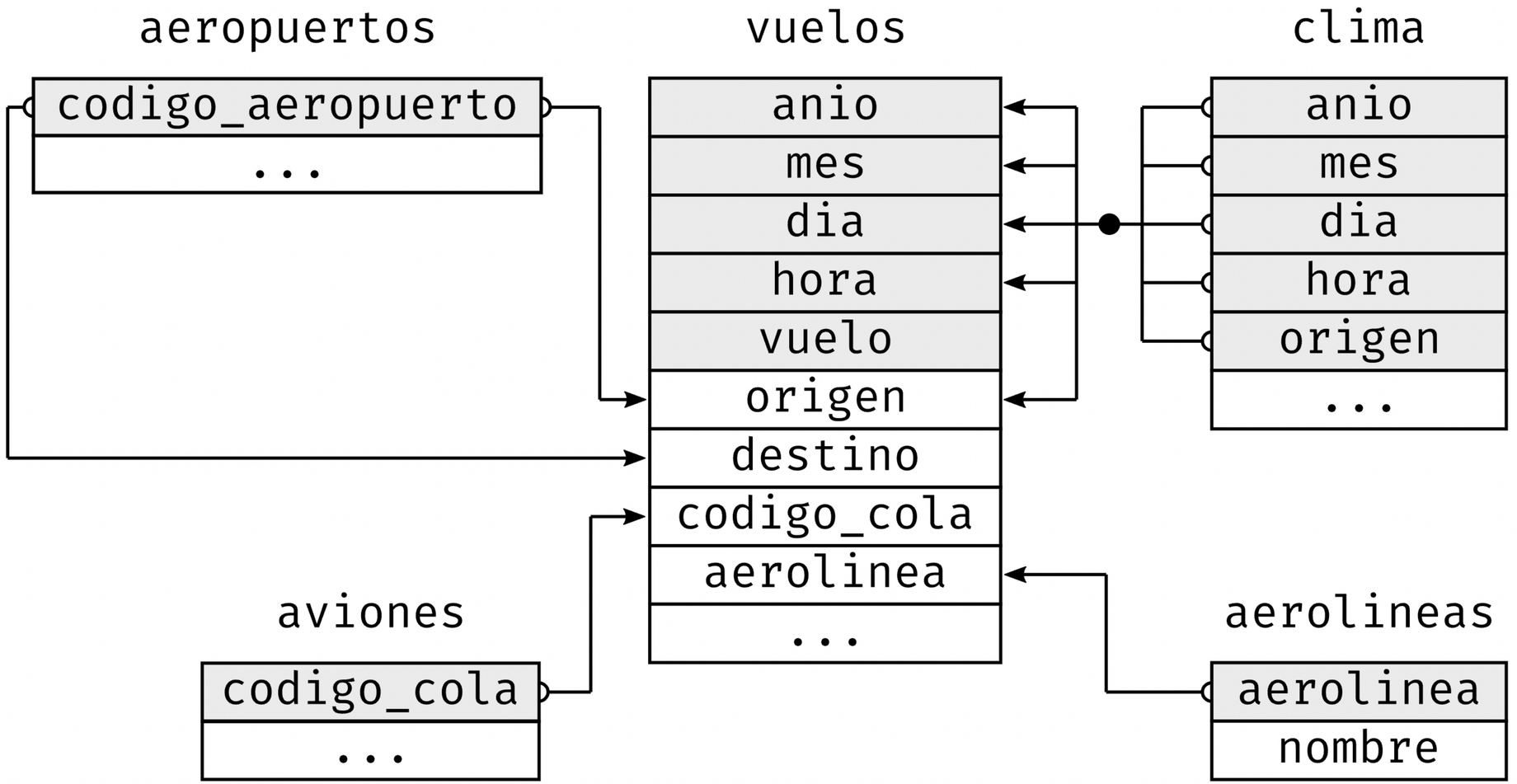
clima

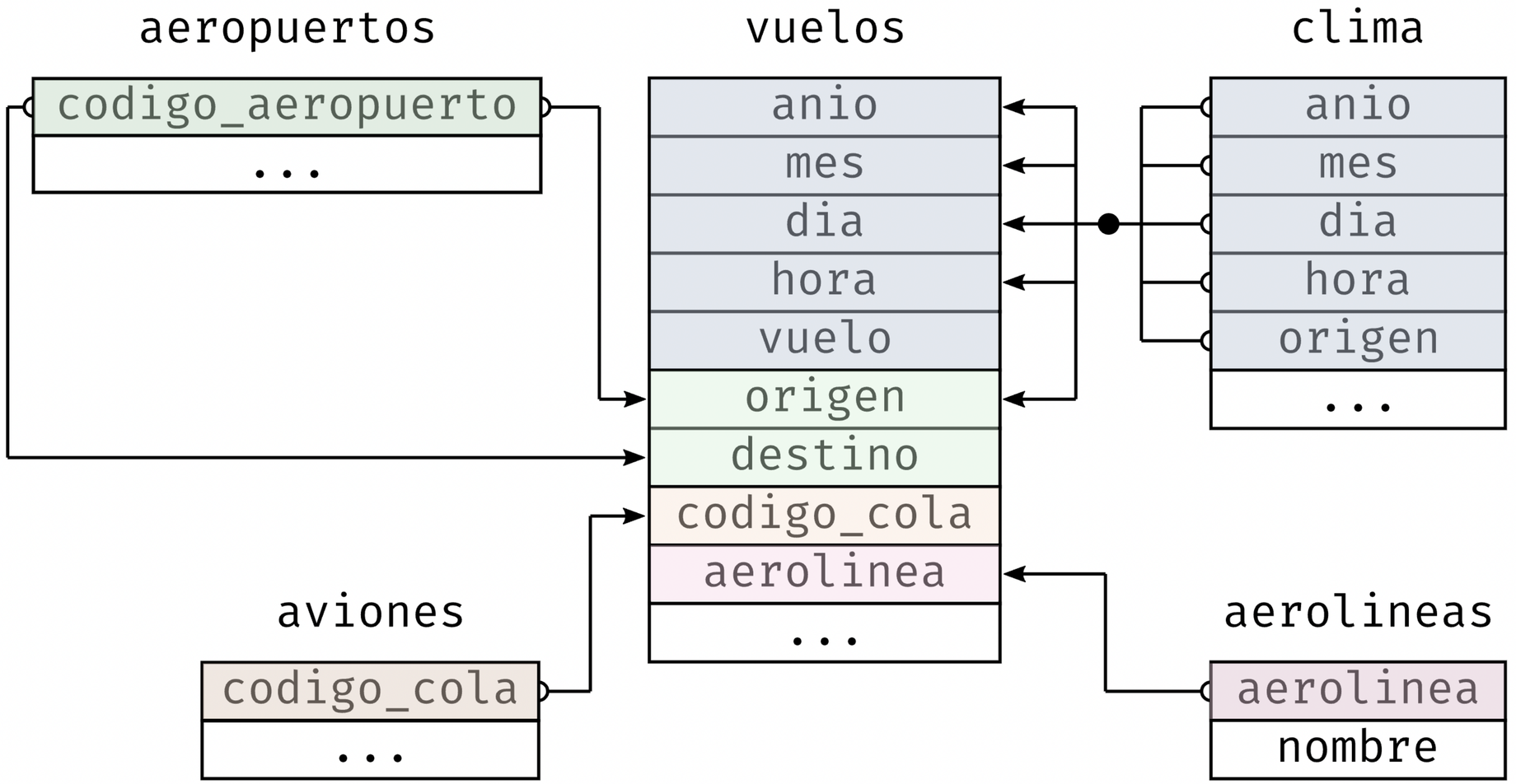
```
#> # A tibble: 26,115 × 15
#>   origen  anio   mes   dia  hora temperatura punto_rocio humedad
#>   <chr>  <int> <int> <int> <int>       <dbl>       <dbl>   <dbl>
#> 1 EWR     2013     1     1     1        39.0        26.1    59.4
#> 2 EWR     2013     1     1     2        39.0        27.0    61.6
#> 3 EWR     2013     1     1     3        39.0        28.0    64.4
#> 4 EWR     2013     1     1     4        39.9        28.0    62.2
#> 5 EWR     2013     1     1     5        39.0        28.0    64.4
#> 6 EWR     2013     1     1     6        37.9        28.0    67.2
```

aeropuertos

```
#> # A tibble: 1,458 × 8
#>   codigo_aeropuerto nombre   latitud longitud altura zona_horaria horario_verano
#>   <chr>             <chr>      <dbl>    <dbl>  <dbl>        <dbl> <chr>
#> 1 04G               Lansdow…    41.1    -80.6   1044           -5 A
#> 2 06A               Moton F…    32.5    -85.7    264           -6 A
#> 3 06C               Schaumb…    42.0    -88.1    801           -6 A
#> 4 06N               Randall…    41.4    -74.4    523           -5 A
#> 5 09J               Jekyll …    31.1    -81.4     11           -5 A
#> 6 0A9               Elizabe…    36.4    -82.2   1593           -5 A
```


aeropuertos
codigo_aeropuerto
...
vuelos
anio
mes
dia
hora
vuelo
origen
destino
codigo_cola
aerolinea
...
clima
anio
mes
dia
hora
origen
...
aviones
codigo_cola
...
aerolineas
aerolinea
nombre


aeropuertos
codigo_aeropuerto
...
vuelos
anio
mes
dia
hora
vuelo
origen
destino
codigo_cola
aerolinea
...
clima
anio
mes
dia
hora
origen
...
aviones
codigo_cola
...
aerolineas
aerolinea
nombre



La clave para entender estos diagramas es recordar que cada relación siempre involucra un par de tablas.

Clave o key: Variable (o conjunto de variables) que identifican **de manera única una observación**

Clave Primaria

Identifica únicamente una observación en su propia tabla.
Por ejemplo, **aviones$codigo_cola** es una clave primaria, ya que identifica de manera única cada avión en la tabla aviones.

Clave Foránea

Únicamente identifica una observación en otra tabla.
Por ejemplo, **vuelos$codigo_cola** es una clave foránea, ya que aparece en la tabla vuelos, en la que une cada vuelo con un único avión.

aviones	código_cola	año	tipo	fabricante	modelo	motores	asientos	velocidad	...	
vuelos	año	mes	día	hora	vuelo	origen	destino	código_cola	aerolínea	...

Una clave primaria y su correspondiente clave foránea en otra tabla forman una relación.
Las relaciones son típicamente **uno-a-muchos**.

Chequear siempre que la clave identifique efectivamente de forma única cada observación.

¿Se les ocurre otro ejemplo de variable que en un dataset sea clave primaria y en otro clave foránea?

¿Qué pasa con el dataset vuelos?

aerolínea		**airports**		**aviones**		**clima**		**vuelos**
aerolínea		código_aeropuerto		código_cola		origen		año
nombre		nombre		año		año		mes
		lat		tipo		mes		día
		lon		fabricante		día		hora
		alt		modelo		hora		vuelo
		zona_horaria		motores		temperatura		origen
		horario_verano		asientos		punto_rocío		destino
		...		velocidad		humedad		código_cola
				...		...		aerolínea
								...

¿Se les ocurre otro ejemplo de variable que en un dataset sea clave primaria y en otro clave foránea?

¿Qué pasa con el dataset vuelos?

aerolínea	airports	aviones	clima	vuelos
aerolínea	código_aeropuerto	código_cola	origen	año
nombre	nombre	año	año	mes
	lat	tipo	mes	día
	lon	fabricante	día	hora
	alt	modelo	hora	vuelo
	zona_horaria	motores	temperatura	origen
	horario_verano	asientos	punto_rocío	destino
	...	velocidad	humedad	código_cola
		...	...	aerolínea
				...

En caso de que no exista una variable que identifique únicamente cada observación puede ser útil agregar una: **clave subrogada**

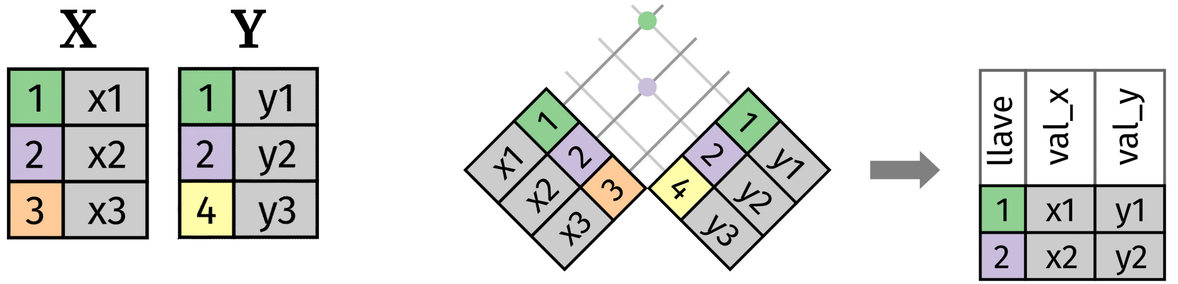
Unión de transformación

Una unión de transformación te permite combinar variables a partir de dos tablas.
B**usca coincidencias de observaciones** de acuerdo a sus claves y luego copia todas las variables de una tabla en la otra.

Unión interior (*inner join*)

Mantiene las observaciones que aparecen en ambas tablas.

inner_join(vuelos, aerolíneas, by = "aerolínea")



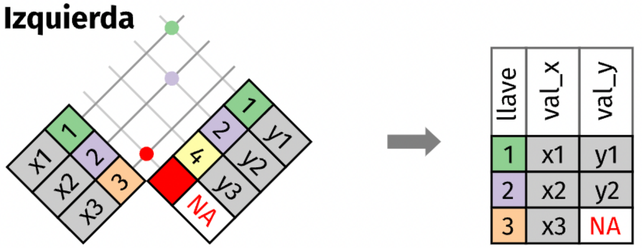
Las filas no coincidentes no se incluyen en el resultado

by = NULL usa todas las variables que aparecen en ambas tablas, lo que se conoce como unión natural.

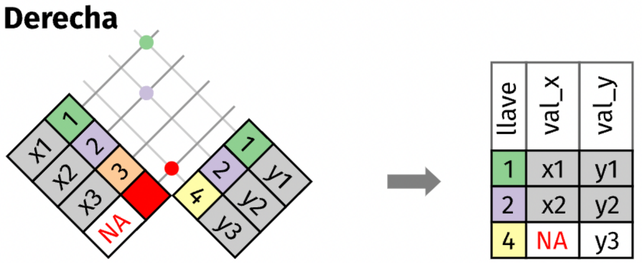
Unión exterior (*outer join*)

Mantiene las observaciones que aparecen en al menos una de las tablas.

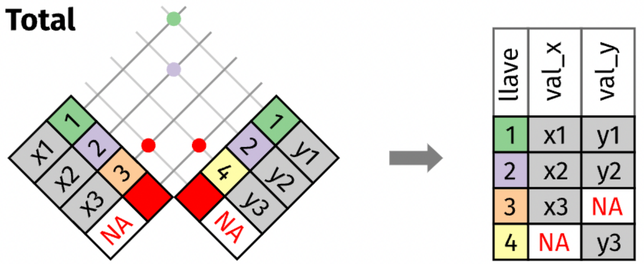
left_join(vuelos, aerolíneas, by = "aerolínea")



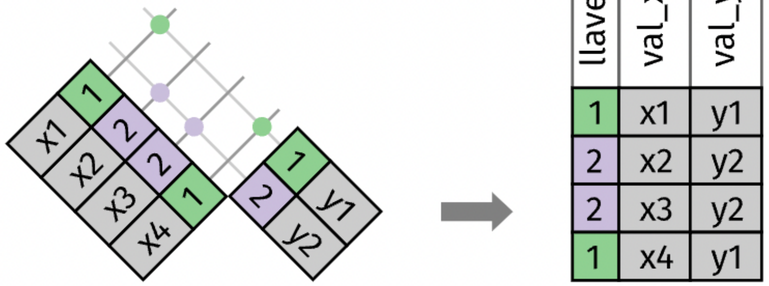
right_join(vuelos, aerolíneas, by = "aerolínea")



full_join(vuelos, aerolíneas, by = "aerolínea")

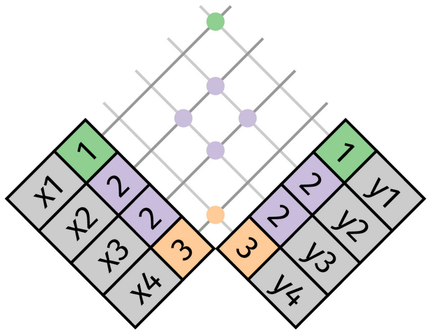


Duplicados



llave	val_x	val_y
1	x1	y1
2	x2	y2
2	x3	y2
1	x4	y1

Típico cuando queremos agregar información adicional da una relación uno a muchos (ej: vuelos y aerolíneas).



llave	val_x	val_y
1	x1	y1
2	x2	y2
2	x2	y3
2	x3	y2
2	x3	y3
3	x4	y4

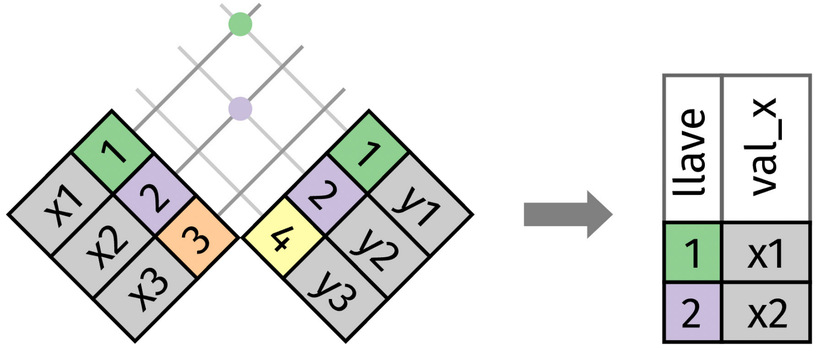
Esto es usualmente un error debido a que en ninguna de las tablas las claves identifican de manera única una observación.

Resultado: se obtienen todas las posibles combinaciones

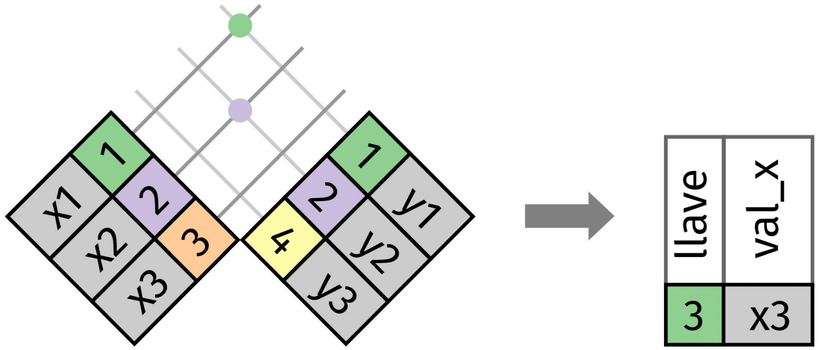
Uniones de filtros (*filter unions*)

Uniones afectando las observaciones en sí, y no a las variables

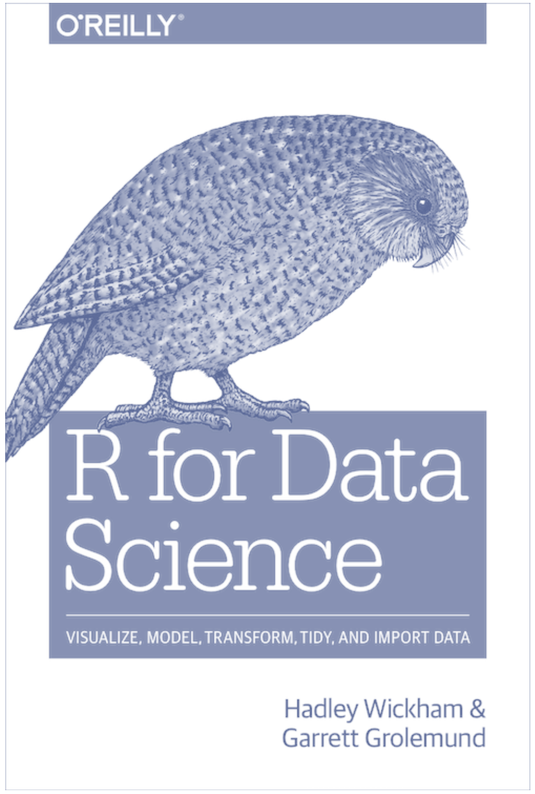
semijoin(x, y, by = "")



antijoin(x, y, by = "")



Todo esto y mucho más...



Cap. 13