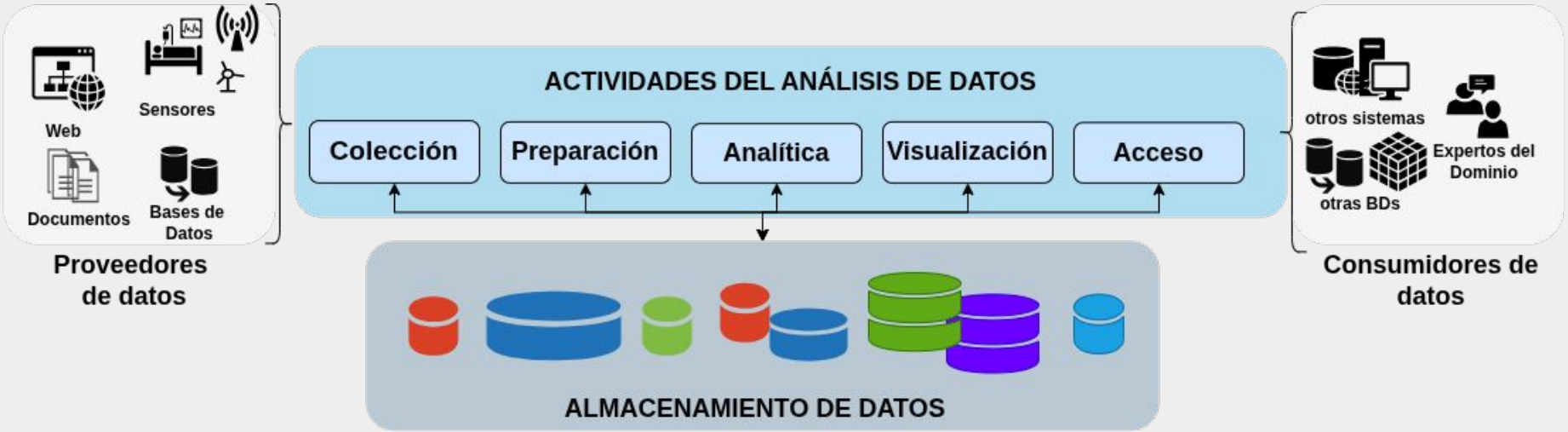
Especialización en Inteligencia de Datos Aplicada

Facultad de Informática
Universidad Nacional del Comahue

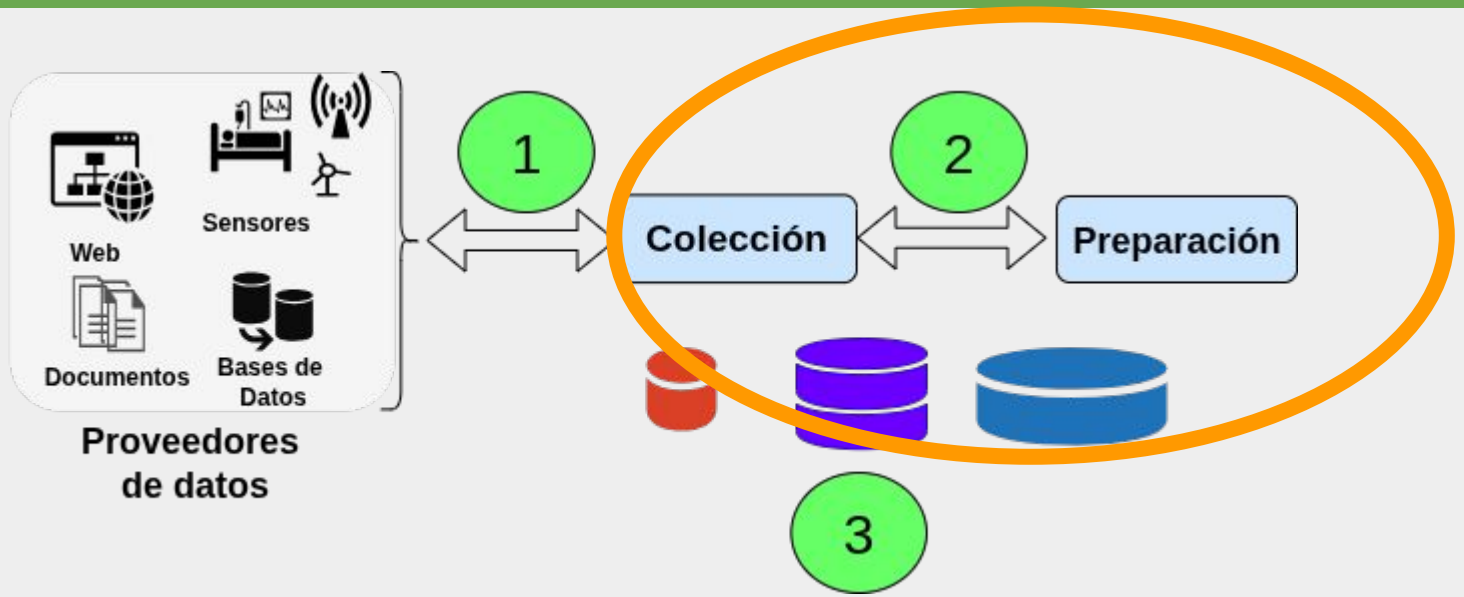
EXTRACCIÓN, PREPARACIÓN Y ALMACENAMIENTO DE LOS DATOS

Segundo Cuatrimestre
Unidad II

PROCESO DE ANÁLISIS DE DATOS
Actividades

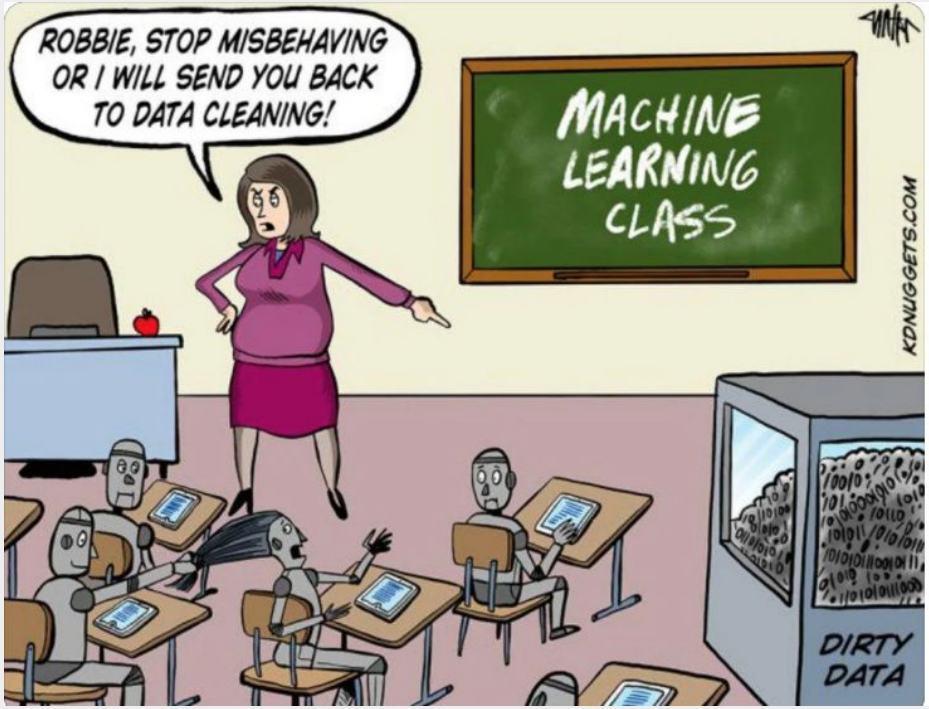


PROCESO DE ANÁLISIS DE DATOS
y ahora?

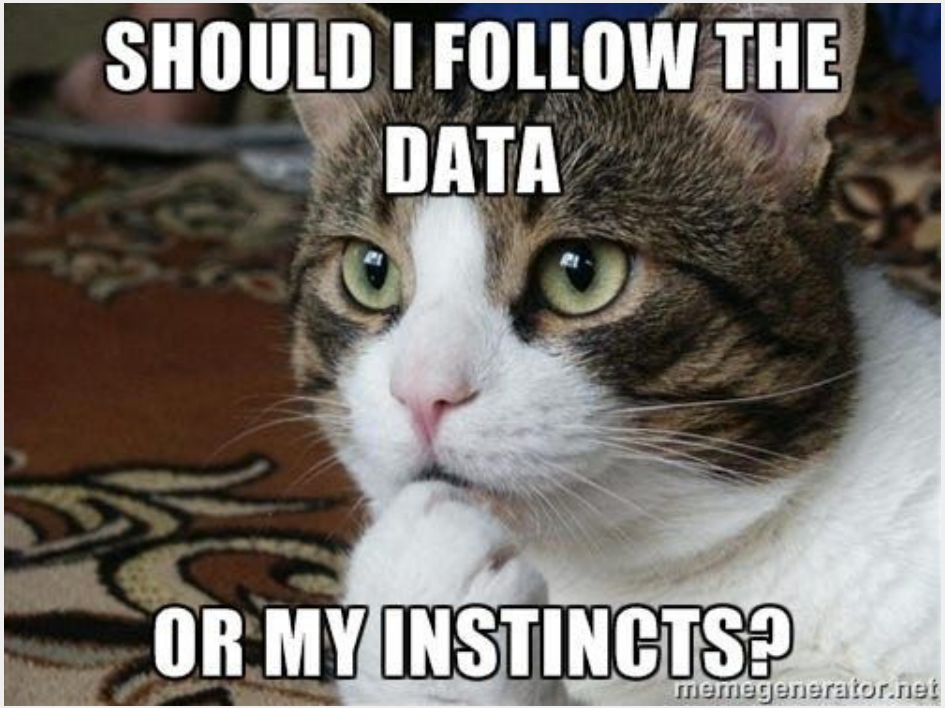


PROCESO DE ANÁLISIS DE DATOS
Preparación de los Datos

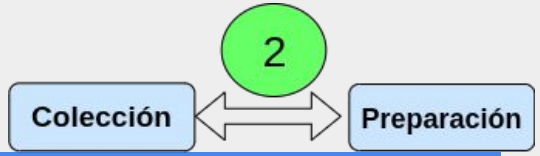
PROCESO DE ANÁLISIS DE DATOS
Preparación de los Datos



PROCESO DE ANÁLISIS DE DATOS
Preparación de los Datos

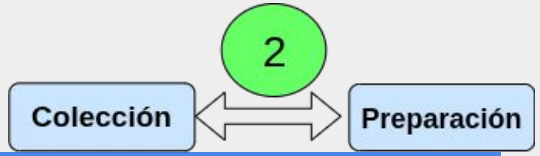


PROCESO DE ANÁLISIS DE DATOS
Preparación de Datos



- Es una tarea **compleja** en donde la información de las fuentes de datos seleccionadas debe ser procesada para:
  - *limpiarlas*, en el sentido de eliminar redundancias innecesarias como datos duplicados, inconsistencias, etc.
  - *reorganizarlas*, en caso de ser necesario, en información útil de acuerdo a los requerimientos del sistema de análisis.
- **Puede ocupar más del 80% de todo el proceso**

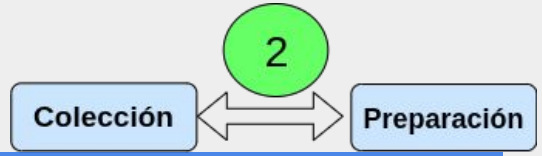
PROCESO DE ANÁLISIS DE DATOS
Preparación de Datos

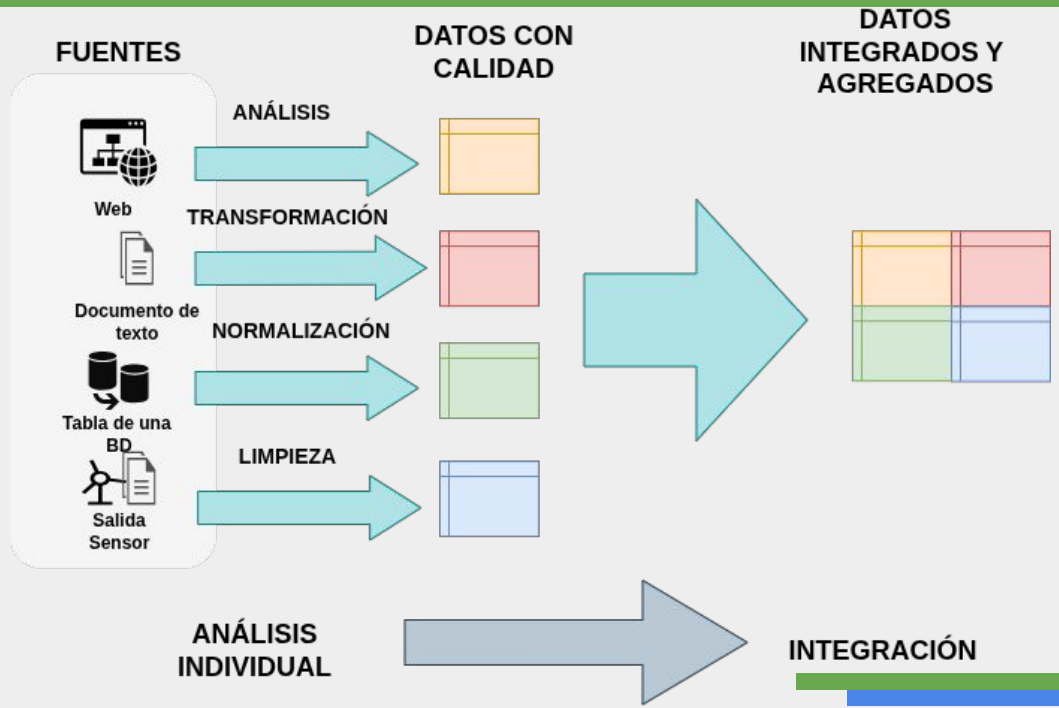


- Tenemos muchas fuentes de inform
  entonces debemos separar en dos actividades:
  - ***Analizar cada fuente por separado***: limpieza, normalización, filtrado, etc.
  - ***Integrar las fuentes***: relacionar los datos dentro de una misma estructura.

PROCESO DE ANÁLISIS DE DATOS

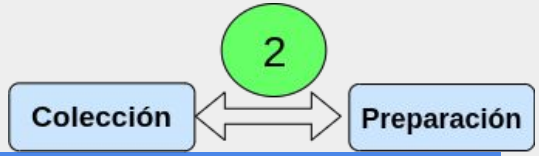
Preparación de Datos

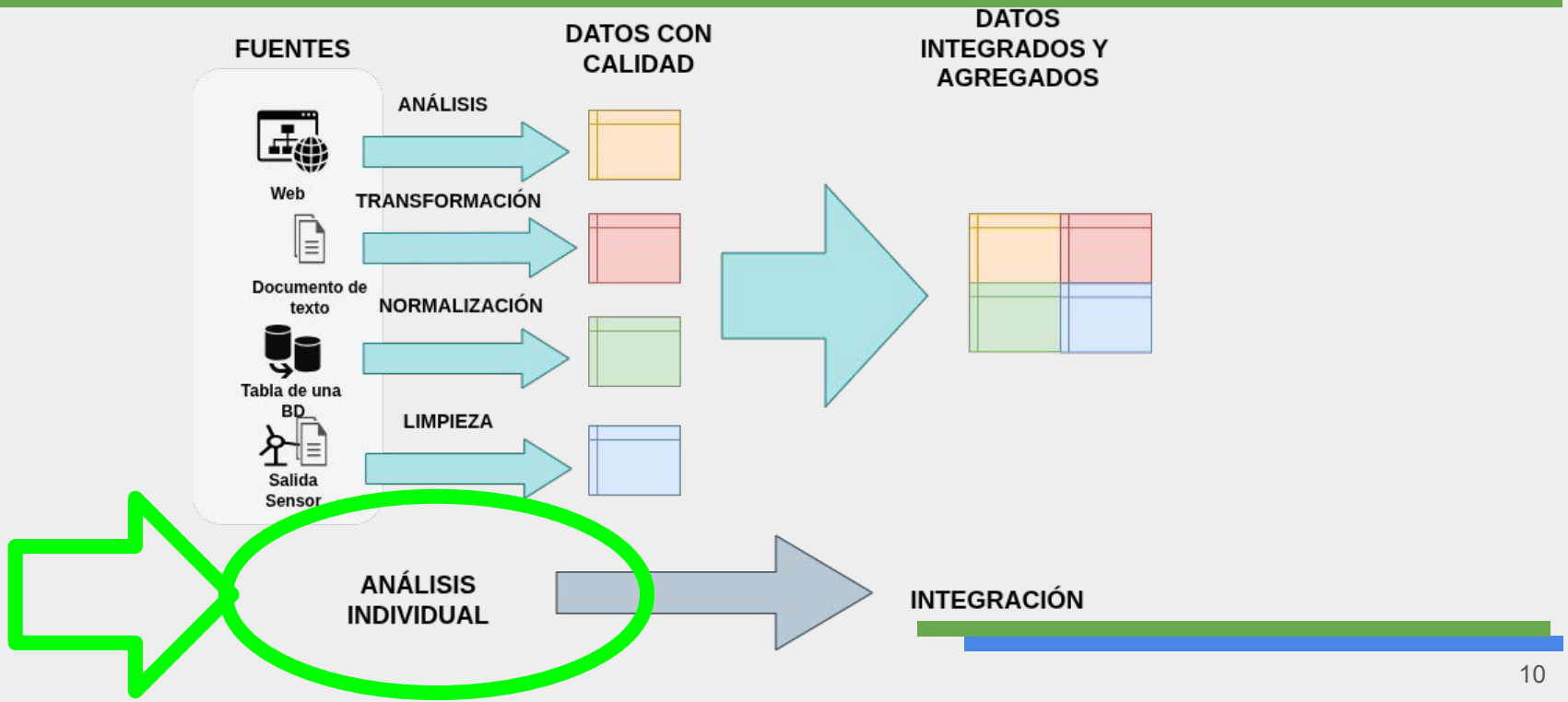




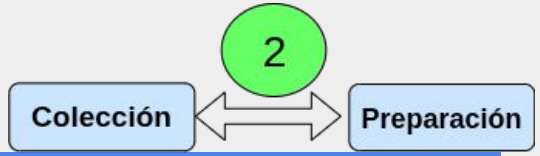
PROCESO DE ANÁLISIS DE DATOS

Preparación de Datos



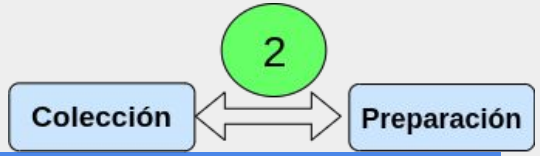


PROCESO DE ANÁLISIS DE DATOS
Calidad de Datos



- Garantizar o mejorar la **calidad de los datos** obtenidos de diferentes fuentes es una ***tarea compleja*** ya que requiere analizar cada fuente de datos, en el formato en que está provista y mejorarla o acondicionarla a los requerimientos del sistema de análisis a construir.
- ***Debemos***:
  - ***Traducir*** las fuentes de información a nuevas estructuras que sean consistentes

PROCESO DE ANÁLISIS DE DATOS
Calidad de Datos

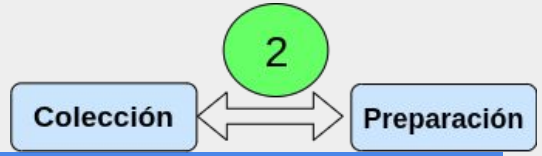


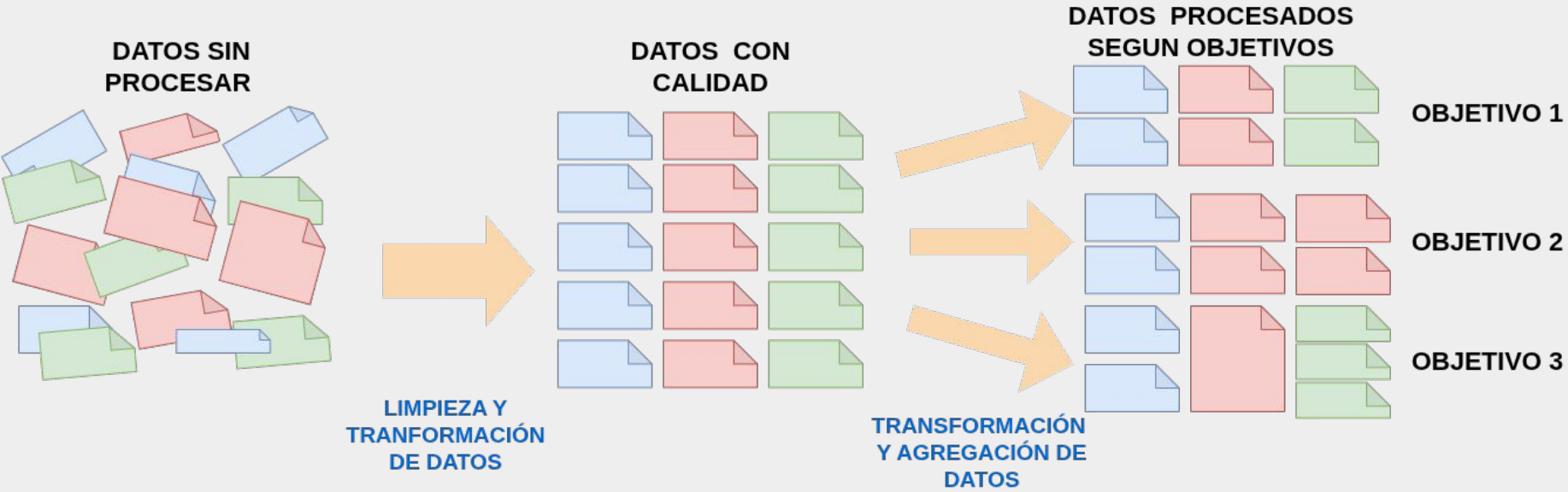
- ***Debemos***:
  - *Traducir….*
  - ***Transformar*** las fuentes en otras que posean agregaciones y/o datos ya procesados.

La calidad de los datos es la capacidad de complir con su propósito en un contexto determinado

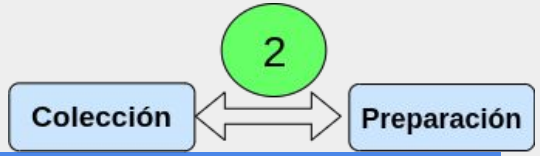
PROCESO DE ANÁLISIS DE DATOS

Calidad de Datos



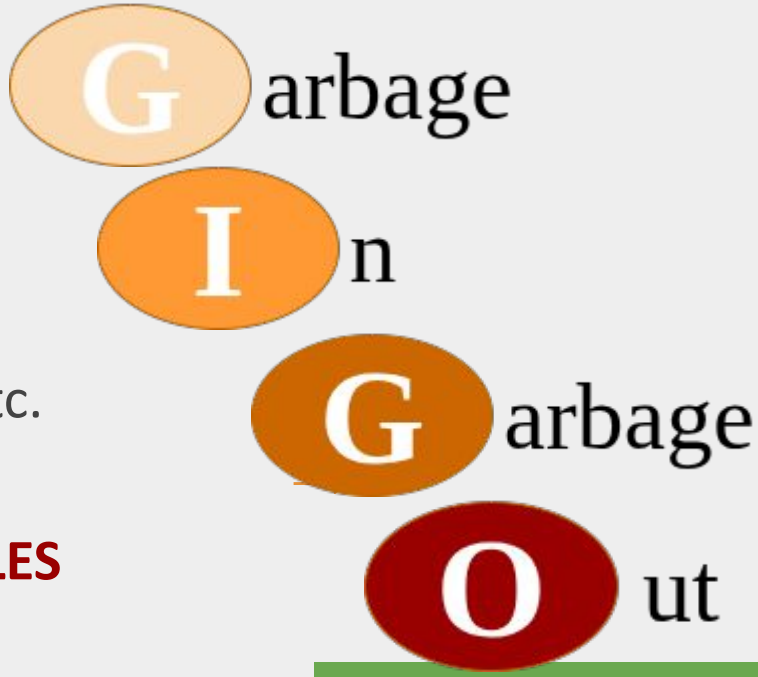


PROCESO DE ANÁLISIS DE DATOS
Calidad de Datos

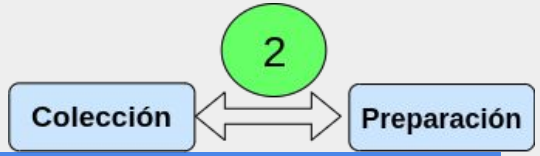


- Nos vamos a encontrar con:
  - datos incompletos
  - datos inexactos
  - datos inconsistentes
  - datos desbalanceados, etc.
- datos sucios

  resultados NO CONFIABLES

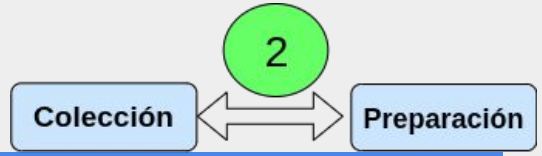


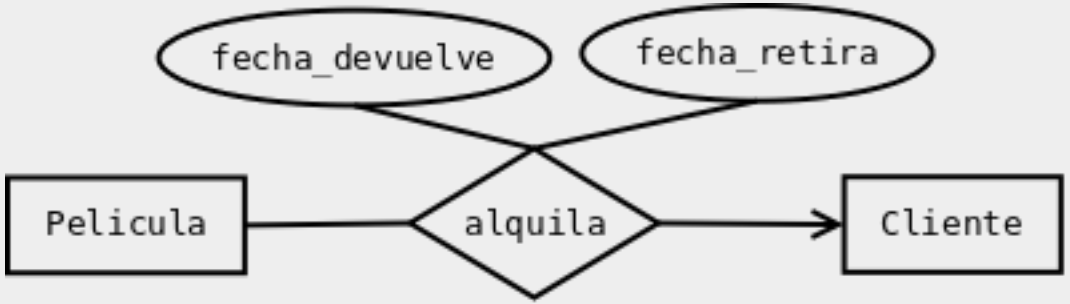
PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos



- **Limpieza de los Datos**.
  - realizar ***rutinas de limpieza*** para ***"limpiar"*** los datos completando los valores faltantes, suavizando los datos ruidosos, identificando o eliminando los valores atípicos y resolviendo las inconsistencias.

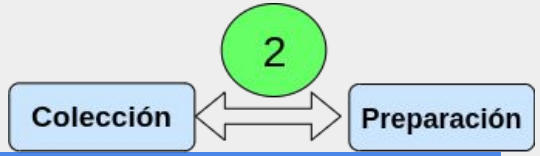
PROCESO DE ANÁLISIS DE DATOS
Calidad de Datos





Id_Pelicula	Id_Cliente	fecha_retira	fecha_devuelve
1	C1	10/11/2012	07/12/2012
2	null	28/11/2012	05/12/2012
2	C3	30/11/2012	null
3	C2	15/09/2012	null

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos



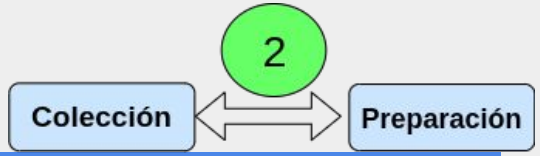
- ***Valores perdidos o faltantes (missing values):***

  TENER CUIDADO!

  Un valor faltante puede no implicar un error en los datos

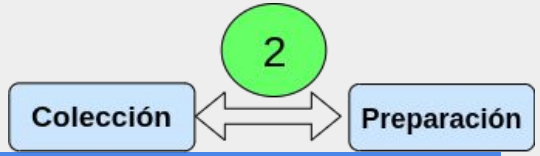
Id_Pelicula	Id_Cliente	fecha_retira	fecha_devuelve
1	C1	10/11/2012	07/12/2012
2	null	28/11/2012	05/12/2012
2	C3	30/11/2012	null
3	C2	15/09/2012	null

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos



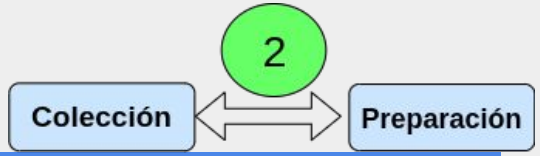
- Qué son los **valores nulos ("null")**?
  - no es la cadena vacía, no es 0, ni otro valor, tiene un significado en sí mismo.
  - Sin embargo… es significado es **ambiguo**

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos



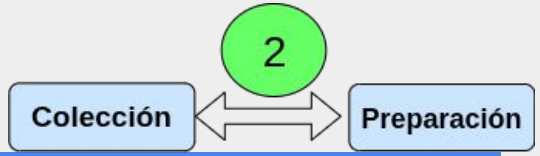
- Problemas con los **valores nulos ("null")**:
  - ***Valor desconocido:*** El dato existe, pero no se conoce en este momento. Ejemplo: Se registró una persona pero aún no se sabe su número de teléfono.
  - ***No aplica:*** El dato no es relevante para ese registro. Ejemplo: piso y depto en una dirección.

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos



- Problemas con los **valores nulos ("null")**:
  - ***Aún no ingresado/Pendiente*:** El valor debe ser completado más adelante. Ejemplo: fecha de devolución
  - ***Valor perdido:*** El dato existía, pero se perdió o no se pudo recuperar. Ejemplo: un mal ingreso, mala importación de datos, una limpieza mal efectuada.

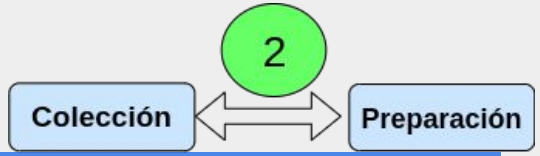
PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos



- Tratamiento de ***valores perdidos o faltantes (missing values):***
  - ***Ignorar la tupla o eliminarla:*** Este método no es muy efectivo, a menos que la tupla contenga varios atributos con valores faltantes. No es efectiva tampoco cuando el porcentaje de valores faltantes por atributo varía considerablemente. Al ignorar la tupla, no hacemos uso de los valores de los atributos restantes en la tupla. Tales datos podrían haber sido útiles para los objetivos planteados.

Id_Pelicula	Id_Cliente	fecha_retira	fecha_devuelve
1	C1	10/11/2012	07/12/2012
2	null	28/11/2012	05/12/2012
2	C3	30/11/2012	null
3	C2	15/09/2012	null

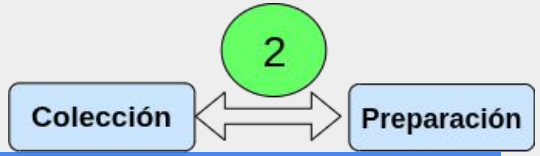
PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Null



- Tratamiento de ***valores perdidos o faltantes (missing values):***
  - ***Rellenar el valor que falta manualmente:*** en general, este enfoque requiere mucho tiempo y puede no ser factible dado un gran conjunto de datos con muchos valores que faltan.

Id_Pelicula	Id_Cliente	fecha_retira	fecha_devuelve
1	C1	10/11/2012	07/12/2012
2	null	28/11/2012	05/12/2012
2	C3	30/11/2012	null
3	C2	15/09/2012	null

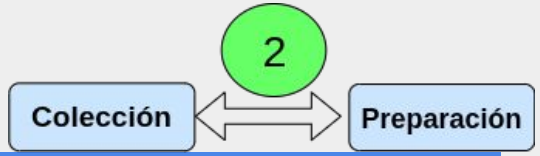
PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Null



- Tratamiento de ***valores perdidos o faltantes (missing values):***
  - ***Utilizar una constante global para completar el valor faltante***: reemplazar todos los valores de atributos faltantes por la misma constante, como con una etiqueta "**Desconocido**" o **−∞**. Aquí la técnica de análisis de datos que utilicemos luego podrá procesar erróneamente ese valor, ya que se repite muchas veces.

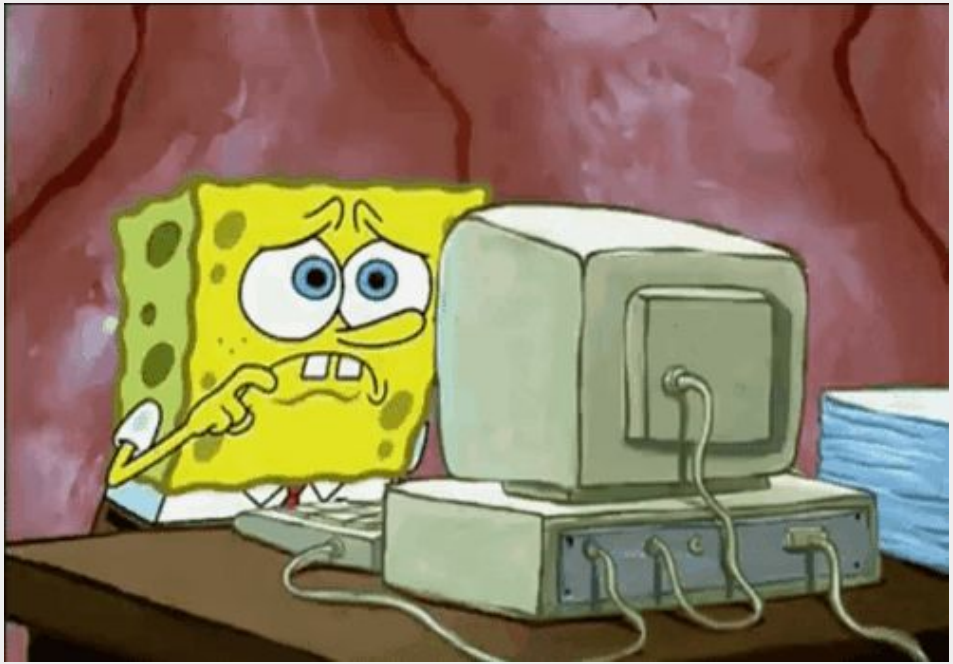
Id_Pelicula	Id_Cliente	fecha_retira	fecha_devuelve
1	C1	10/11/2012	07/12/2012
2	null	28/11/2012	05/12/2012
2	C3	30/11/2012	null
3	C2	15/09/2012	null

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Null

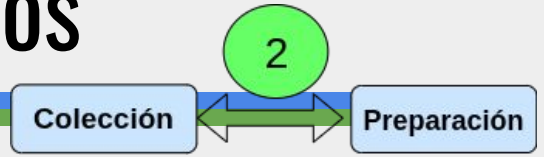


- Tratamiento de ***valores perdidos o faltantes (missing values):***
  - ***Utilizar una medida de tendencia central para el atributo (p. ej., la media o la mediana) para completar los valores que faltan***: Dependerá de cuantos datos debo reemplazar.
  - ***Reemplazar con el valor más probable para completarlo***: con técnicas más sofisticadas como regresión o árboles de decisión.

PROCESO DE ANÁLISIS DE DATOS
A TRABAJAR!!



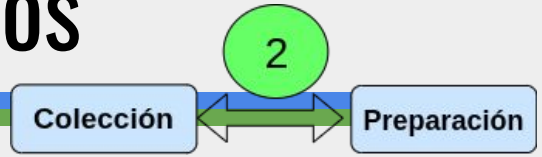
PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Atípicos



- **Datos con ruido (noisy data) y valores atípicos (outliers values):**: Son conceptos diferentes, que generalmente se detectan visualmente y/o con técnicas estadísticas.
  - ***Datos con ruido (noisy data)***: representan errores en los datos, que pueden surgir debido a un mal funcionamiento de un sensor, un dato mal ingresado por el usuario, etc. Son algo que nunca puede existir.

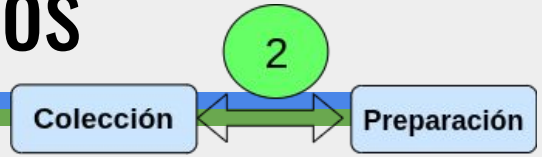
https://letsdatascience.com/outlier-treatment/

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Atípicos

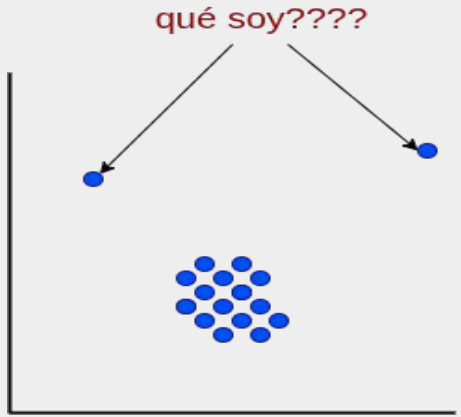


- **Datos con ruido (noisy data) y valores atípicos (outliers values):**
  - ***Valores atípicos***: son valores que caen lejos de la mayoría de los valores existentes, por ejemplo, un sensor de temperatura que marque 35°C en invierno en Neuquén. Un valor atípico puede ser válido o puede ser un error (noisy).

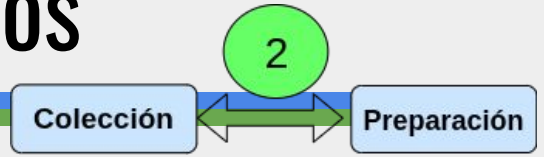
PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Atípicos



- **Datos con ruido (noisy data) y valores atípicos (outliers values):**
  - *Identificarlos con técnicas como el método percentil, el método del rango intercuartil (IQR), el método Z-Score, clustering.*

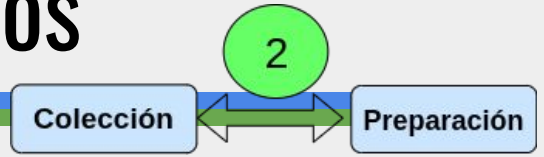


PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Atípicos



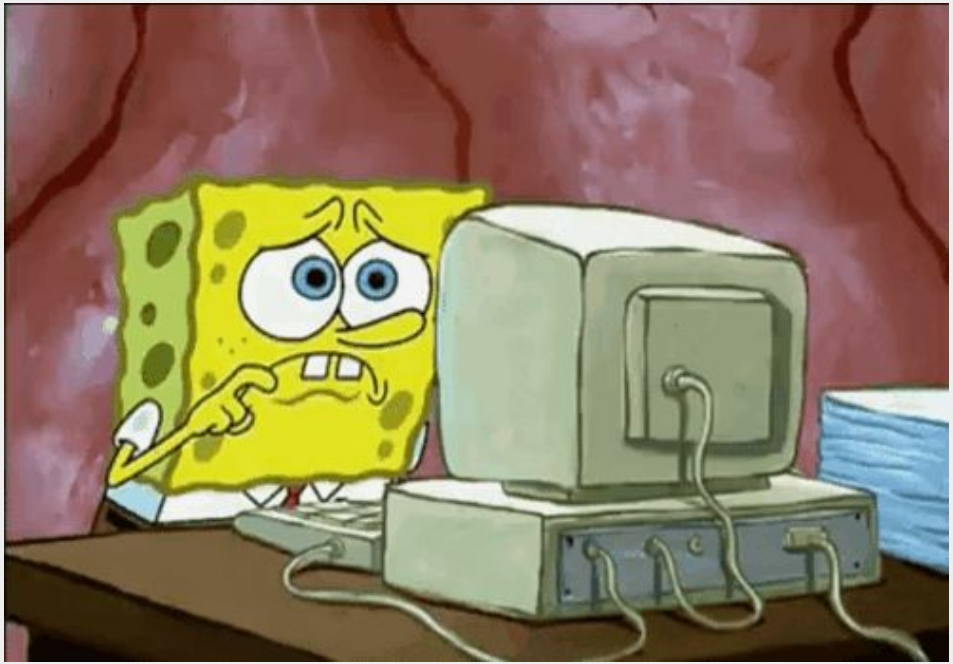
- Tratamiento de ***valores atípicos:***
  - ***Eliminar el valor***: Es peligroso porque el valor podría haber sido debido a un evento significativo y no a un error.
  - ***Limitar el valor a un rango***: Si el valor es muy grande, se puede reducir su valor hasta cierto límite. Este límite puede basarse en el IQR o la desviación estándar.

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Atípicos

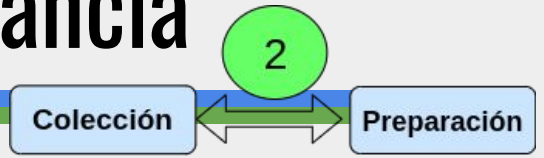


- Tratamiento de ***valores atípicos:***
  - **Imputar el valor:** Reemplazar con la media o la mediana o con técnicas de aprendizaje automático como KNN.
  - ***Agrupar los atípicos:*** También podemos agrupar nuestros datos en grupos y tratar todo el grupo que contiene el valor atípico como uno solo.
  - y varios mas….

PROCESO DE ANÁLISIS DE DATOS
A TRABAJAR!!

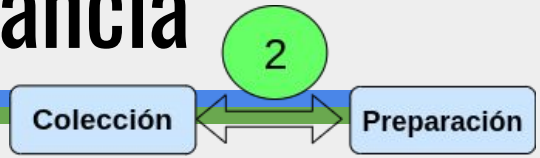


PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Redundancia



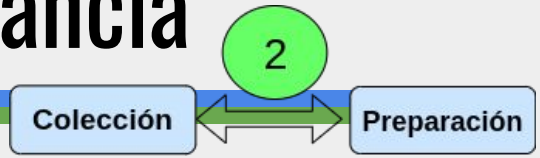
- **Redundancia de Datos:** Son datos que se repiten en forma exacta, o parcialmente, pero tienen el mismo significado o representan la misma realidad.
  - Es la duplicación innecesaria de los datos.
  - Genera inconsistencias
  - En general ocurren cuando hacemos una **integración de fuentes de datos**.

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Redundancia



- **Redundancia de Datos**
  - *Redundancia lógica*: la misma información se almacena en varios lugares (documentos, tablas) generando inconsistencias en los datos y anomalías en las actualizaciones
  - *Redundancia estructural*: repetición de datos dentro de una sola tabla/documento, causada por un modelado de datos ineficiente.

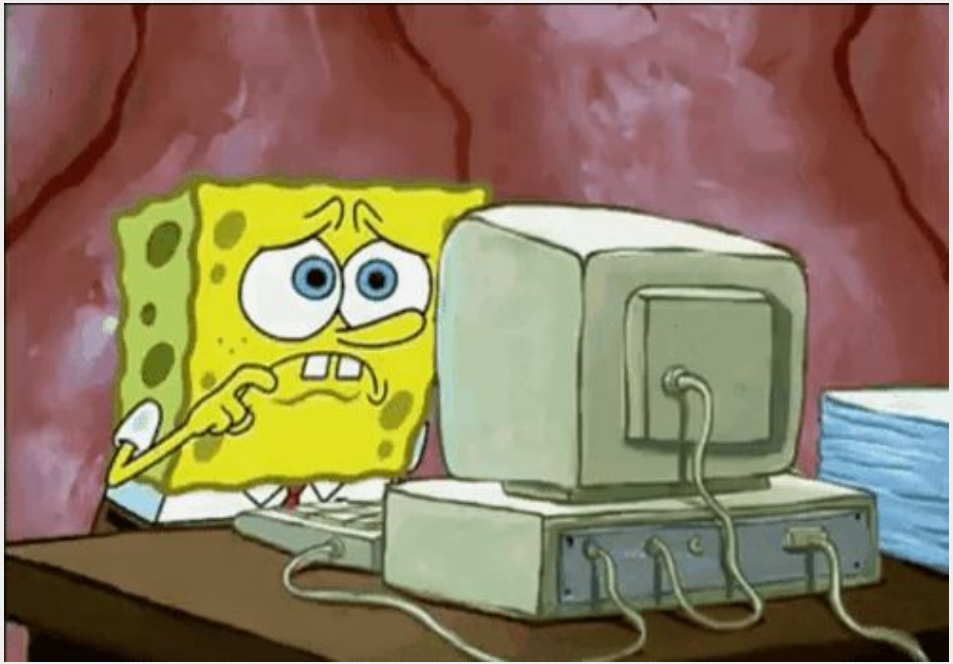
PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Redundancia



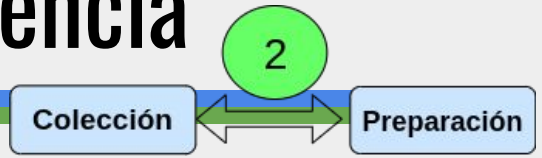
- **Redundancia de Datos**
  - *Redundancia de control*: duplicación de datos para garantizar la tolerancia a fallas y la recuperación de datos en caso de fallas del sistema. Disponibilidad y eficiencia vs inconsistencia.

REDUNDANCIA vs INCONSISTENCIA

PROCESO DE ANÁLISIS DE DATOS
A TRABAJAR!!

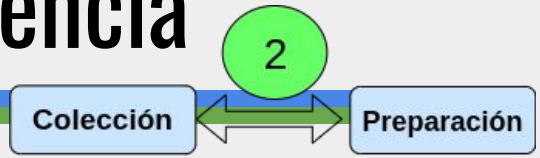


PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Consistencia



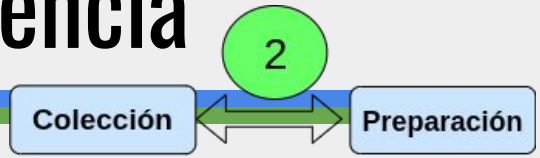
- **Consistencia de Datos:** Es un concepto MUY amplio ya que abarca diferentes niveles:
  - desde el diseño de los datos (BDs) en sí mismo (redundancia-inconsistencia)
  - desde el punto de vista de transacciones (principios ACID)
  - desde los tipos y formatos de los datos (mismas medidas de longitud, fechas fuera de rango, etc.)
  - desde la semántica y relación del dato (un abogado de 5 años)

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Consistencia



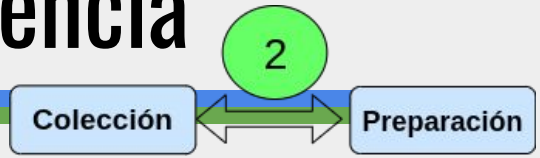
	nombre	genero	altura	peso	fecha_nacimiento	fecha_registro
0	Ana	F	1.70 m	70 kg	1990-05-01	2021-01-01
1	Luis	masculino	170 cm	154 lbs	1880-01-01	2022-01-01
2	Carlos	M	5.9 ft	65kg	2025-01-01	2020-01-01
3	Marta	femenina	1,80 metros	setenta	2000-06-15	2019-01-01
4	Sofía	FEM	1.65	60 kilogramos	2015-10-10	2017-01-01
5	Juan	H	300 cm	0 kg	2050-12-31	2010-01-01
6	Andrea	no binario	160	52	2002-02-29	2020-05-05
7	Pedro	MASC	1.55m	80kg	1985-07-20	2021-12-31

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Consistencia



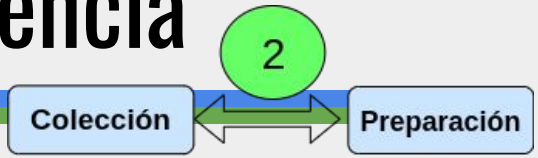
	nombre	genero	altura	peso	fecha_nacimiento	fecha_registro
0	Ana	F	1.70 m	70 kg	1990-05-01	2021-01-01
1	Luis	masculino	170 cm	154 lbs	1880-01-01	2022-01-01
2	Carlos	M	5.9 ft	65kg	2025-01-01	2020-01-01
3	Marta	femenina	1,80 metros	setenta	2000-06-15	2019-01-01
4	Sofía	FEM	1.65	60 kilogramos	2015-10-10	2017-01-01
5	Juan	H	300 cm	0 kg	2050-12-31	2010-01-01
6	Andrea	no binario	160	52	2002-02-29	2020-05-05
7	Pedro	MASC	1.55m	80kg	1985-07-20	2021-12-31

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Consistencia



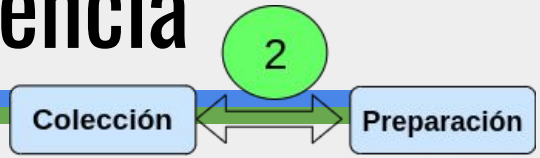
	nombre	genero	altura	peso	fecha_nacimiento	fecha_registro
0	Ana	F	1.70 m	70 kg	1990-05-01	2021-01-01
1	Luis	masculino	170 cm	154 lbs	1880-01-01	2022-01-01
2	Carlos	M	5.9 ft	65kg	2025-01-01	2020-01-01
3	Marta	femenina	,80 metros	setenta	2000-06-15	2019-01-01
4	Sofía	FEM	1.65	60 kilogramos	2015-10-10	2017-01-01
5	Juan	H	300 cm	0 kg	2050-12-31	2010-01-01
6	Andrea	no binario	160	52	2002-02-29	2020-05-05
7	Pedro	MASC	1.55m	80kg	1985-07-20	2021-12-31

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Consistencia



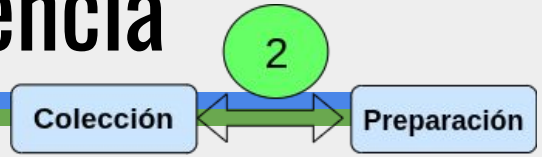
	nombre	genero	altura	peso	fecha_nacimiento	fecha_registro
0	Ana	F	1.70 m	70 kg	1990-05-01	2021-01-01
1	Luis	masculino	170 cm	154 lbs	1880-01-01	2022-01-01
2	Carlos	M	5.9 ft	65kg	2025-01-01	2020-01-01
3	Marta	femenina	1,80 metros	setenta	2000-06-15	2019-01-01
4	Sofía	FEM	1.65	60 kilogramos	2015-10-10	2017-01-01
5	Juan	H	300 cm	0 kg	2050-12-31	2010-01-01
6	Andrea	no binario	160	52	2002-02-29	2020-05-05
7	Pedro	MASC	1.55m	80kg	1985-07-20	2021-12-31

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Consistencia



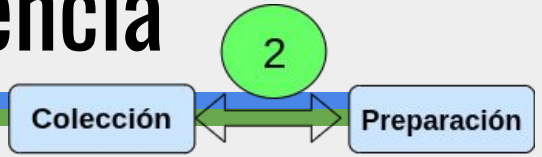
	nombre	genero	altura	peso	fecha_nacimiento	fecha_registro
0	Ana	F	1.70 m	70 kg	1990-05-01	2021-01-01
1	Luis	masculino	170 cm	154 lbs	1880-01-01	2022-01-01
2	Carlos	M	5.9 ft	65kg	2025-01-01	2020-01-01
3	Marta	femenina	1,80 metros	setenta	2000-06-15	2019-01-01
4	Sofía	FEM	1.65	60 kilogramos	2015-10-10	2017-01-01
5	Juan	H	300 cm	0 kg	2050-12-31	2010-01-01
6	Andrea	no binario	160	52	2002-02-29	2020-05-05
7	Pedro	MASC	1.55m	80kg	1985-07-20	2021-12-31

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Consistencia



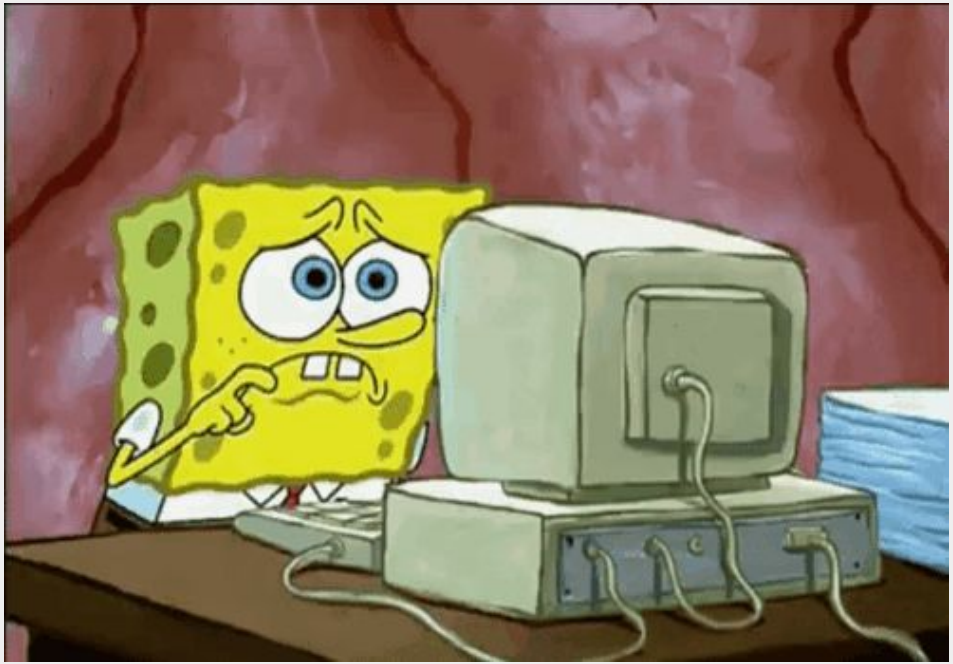
- Detección de ***inconsistencias:***
  - Como son muy variadas y pueden venir desde diferentes orígenes, se deben definir reglas sobre los datos
    - las medidas deben ser en metros
    - la fecha de nacimiento debe ser menor a la fecha de registro
    - las fechas de nacimiento válidas son aquellas menores al año 2020

PROCESO DE ANÁLISIS DE DATOS
Limpieza de Datos - Consistencia



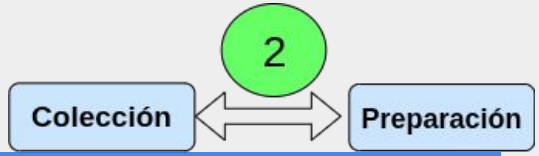
- Tratamiento de ***inconsistencias:***
  - En base a las reglas que se usaron para detectar las inconsistencias, de deben definir procedimientos para transformar los datos:
    - Eliminar la tupla
    - Poner nulos
    - Reemplazar/Predecir con algún valor
    - etc.

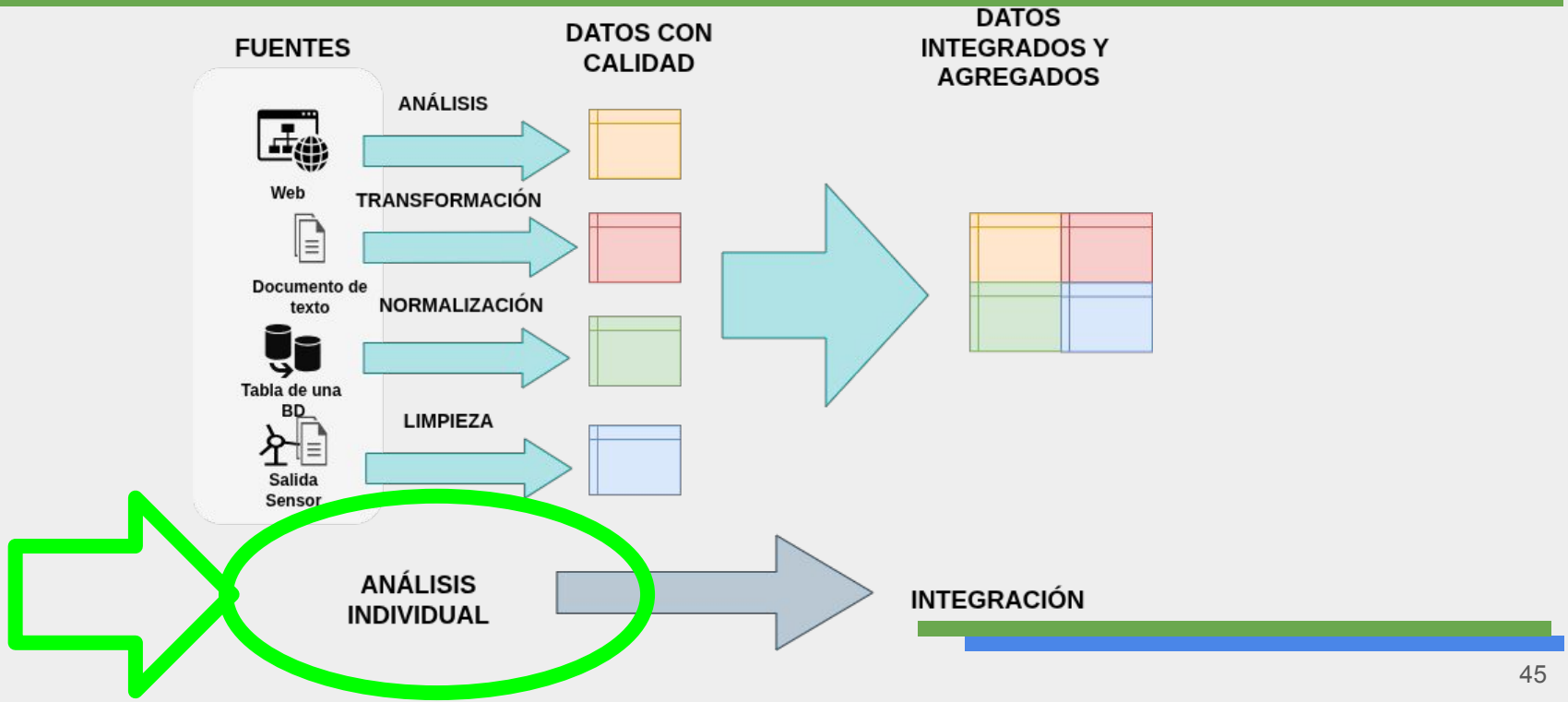
PROCESO DE ANÁLISIS DE DATOS
A TRABAJAR!!



PROCESO DE ANÁLISIS DE DATOS

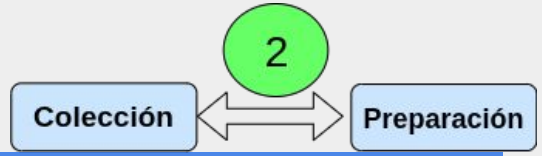
Preparación de Datos

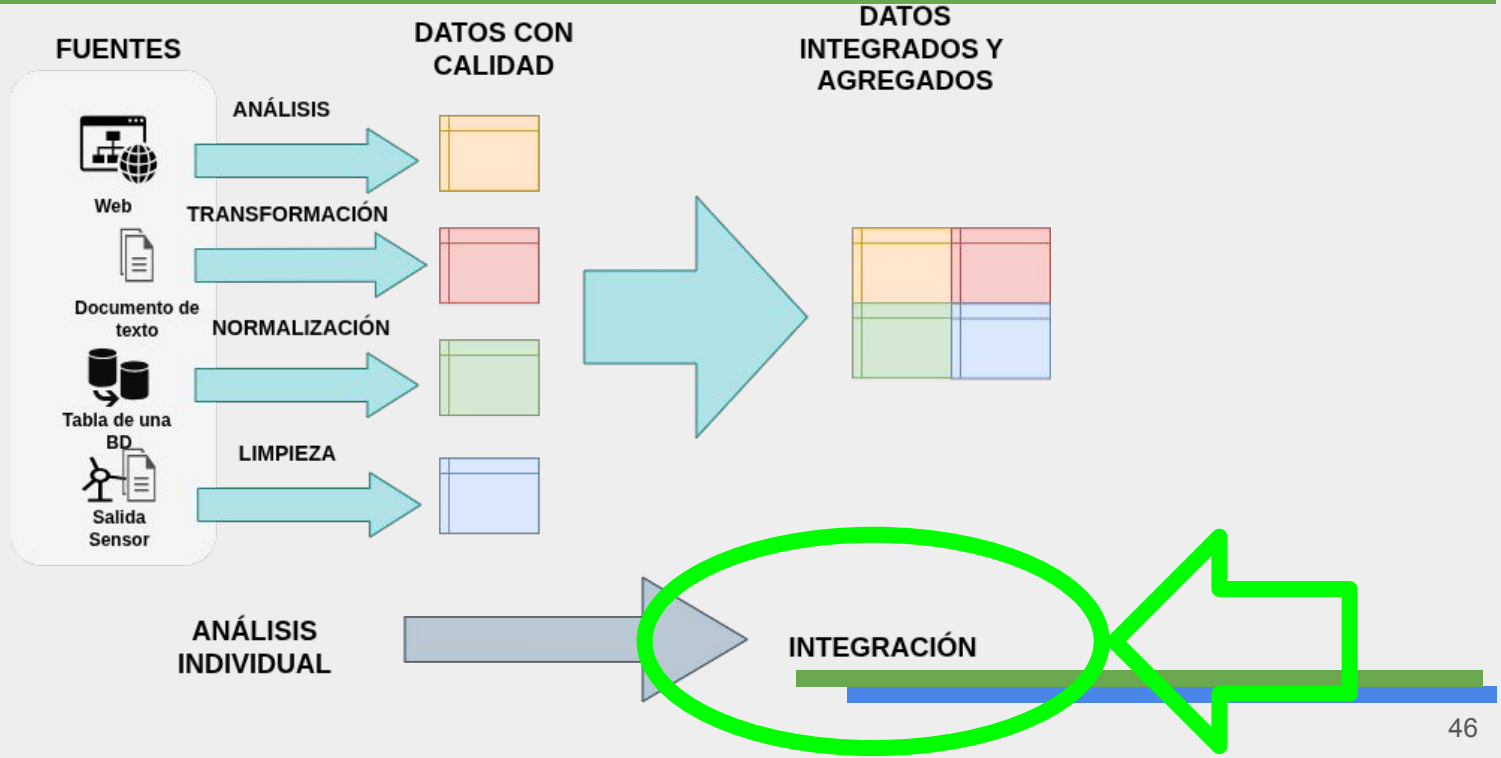




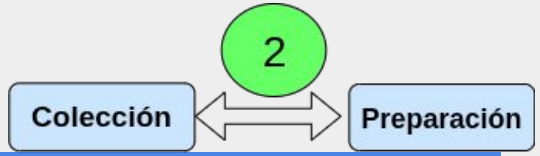
PROCESO DE ANÁLISIS DE DATOS

Preparación de Datos



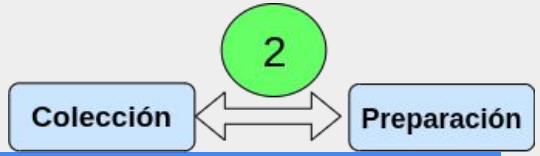


PROCESO DE ANÁLISIS DE DATOS
Integración de Datos



- La **integración de datos** se refiere a la unión de los datos provenientes de las diferentes fuentes.
  - Se unen para cumplir algún objetivo del sistema
  - Es una tarea muy compleja
  - Los datos presentan muchas heterogeneidades
  - Hay mucha redundancia y posibles inconsistencias

PROCESO DE ANÁLISIS DE DATOS
Integración de Datos

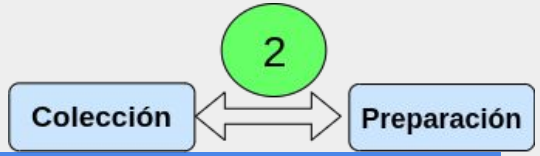


- La **heterogeneidad** se refiere a la diferenciación de los datos en cuanto a si representan los mismos objetos del mundo real, son especializaciones del mismo concepto, están relacionados de alguna manera, etc.

ESTUDIANTE				
DNI	Nombre y Apellido	Apellido	Dirección	Provincia

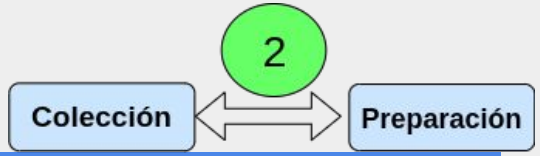
PERSONA				
CUIT	Nombre	Apellido	Dirección	Ciudad

PROCESO DE ANÁLISIS DE DATOS
Integración de Datos



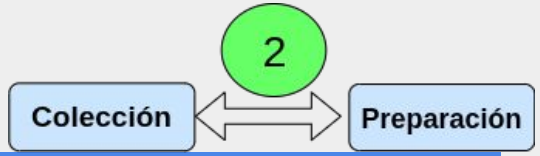
- Hay algunas herramientas que ayudan al proceso de integración y solucionar los problemas de **heterogeneidad**, sin embargo hay mucho trabajo manual
  - Se pueden usar librerías que ayudan a comprender la semántica de una palabra y su relación léxica con otras.
  - Por ejemplo librerías que consultan WordNet (https://wordnet.princeton.edu/) para sinónimos, homónimos, hiperónimos, etc. (https://en-word.net/lemma/house)

PROCESO DE ANÁLISIS DE DATOS
Integración de Datos



- De acuerdo con **Pandas**, hay dos formas generales de integrar datos:
  - unirlos ***sin*** considerar su ***semántica*** (método concat())
  - unirlos ***considerando*** su ***semántica*** (método merge())

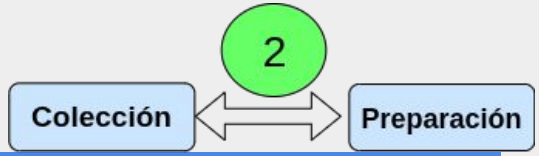
PROCESO DE ANÁLISIS DE DATOS
Concatenar Datos



- Unir ***sin*** considerar su ***semántica*** (método **concat()**)

	nombre_de_perro	fecha_nacimiento	raza	peso	dni_duenio
0	Luna	2020-05-10	Labrador	25.4	29543218
1	Max	2019-11-22	Bulldog Francés	12.1	32219876
2	Simba	2021-03-15	Golden Retriever	30.2	30987654
3	Coco	2018-08-03	Caniche	7.5	29876543
4	Rocky	2022-01-27	Beagle	10.3	28881234
5	Toby	2020-12-14	Doberman	34.7	31587964
6	Mila	2019-07-30	Border Collie	18.2	30214567
7	Bruno	2023-02-05	Boxer	28.6	28456789
8	Lola	2021-09-19	Shih Tzu	6.8	28881234
9	Negrita	2010-08-03	Pastor Belga	40.3	29876543
10	Bruna	2014-12-13	Mestizo	13.2	<NA>
11	Toreto	2022-11-27	Caniche	9.3	<NA>
12	Osi	2019-12-23	Boxer	25.8	29543218

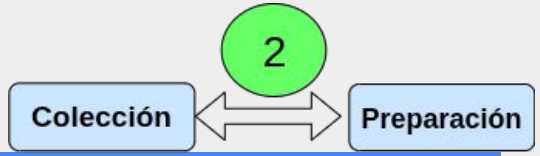
PROCESO DE ANÁLISIS DE DATOS
Concatenar Datos



- Unir ***sin*** considerar su ***semántica*** (método **concat()**)

	dni_duenio	nombre	apellido	fecha_nacimiento	direccion	ciudad
0	30214567	Susana	Horia	1985-04-12	Avenida Siempre Viva 123	Córdoba
1	28456789	Oscar	Acol	1979-09-23	Calle Falsa 456	Neuquén
2	31587964	Elena	Morado	1990-12-01	Pasaje Los Andes 789	Neuquén
3	29876543	Carlos	Fernández	1982-07-19	Belgrano 321	La Plata
4	32765432	Lucila	Nave	1995-11-30	Santa Fe 654	Salta
5	30987654	Paloma	Blanca	1988-06-15	Rivadavia 102	Bahía Blanca
6	29543218	Zoyla	Vaca	1977-03-05	Catamarca 89	Salta
7	32219876	Esteban	Quito	1993-08-27	Mitre 12	Salta
8	28881234	Elsa	Capunta	1984-10-10	9 de Julio 456	Neuquén
9	19333465	Aquiles	Muestro	1964-03-05	Av Argentina 233	Neuquén
10	32465340	Zacarias	Flores	1995-11-15	Av San Martin 323	Neuquén

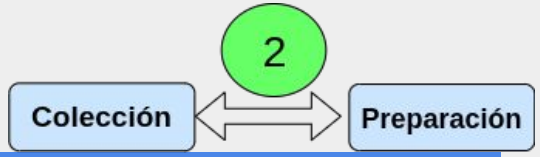
PROCESO DE ANÁLISIS DE DATOS
Concatenar Datos



- **Concatenar** agrega datos de otra tabla como **FILAS manteniendo todas (outer)**

	nombre_de_perro	fecha_nacimiento	raza	peso	dni_duenio	nombre	apellido	direccion	ciudad
0	Luna	2020-05-10	Labrador	25.4	29543218.0	NaN	NaN	NaN	NaN
1	Max	2019-11-22	Bulldog Francés	12.1	32219876.0	NaN	NaN	NaN	NaN
...									
12	Osi	2019-12-23	Boxer	25.8	29543218.0	NaN	NaN	NaN	NaN
0	NaN	1985-04-12	NaN	NaN	30214567.0	Susana	Horia	Avenida Siempre Viva 123	Córdoba
1	NaN	1979-09-23	NaN	NaN	28456789.0	Oscar	Acol	Calle Falsa 456	Neuquén
...									
9	NaN	1964-03-05	NaN	NaN	19333465.0	Aquiles	Muestro	Av Argentina 233	Neuquén
10	NaN	1995-11-15	NaN	NaN	32465340.0	Zacarias	Flores	Av San Martin 323	Neuquén

PROCESO DE ANÁLISIS DE DATOS
Concatenar Datos



- **Concatenar** agrega datos de otra tabla como **FILAS manteniendo solo las iguales (inner) por nombre de índice**

	fecha_nacimiento	dni_duenio
0	2020-05-10	29543218.0
1	2019-11-22	32219876.0
2	2021-03-15	30987654.0
3	2018-08-03	29876543.0
4	2022-01-27	28881234.0
5	2020-12-14	31587964.0
6	2019-07-30	30214567.0
7	2023-02-05	28456789.0

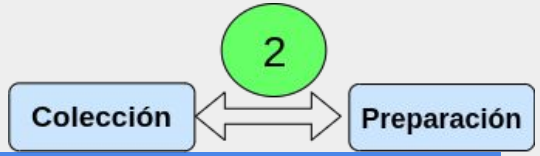
7	2023-02-05	28456789.0
8	2021-09-19	28881234.0
9	2010-08-03	29876543.0
10	2014-12-13	NaN
11	2022-11-27	NaN
12	2019-12-23	29543218.0
0	1985-04-12	30214567.0
1	1979-09-23	28456789.0
2	1990-12-01	31587964.0

2	1990-12-01	31587964.0
3	1982-07-19	29876543.0
4	1995-11-30	32765432.0
5	1988-06-15	30987654.0
6	1977-03-05	29543218.0
7	1993-08-27	32219876.0
8	1984-10-10	28881234.0
9	1964-03-05	19333465.0
10	1995-11-15	32465340.0

- **Concatenar** agrega datos de otra tabla como **COLUMNAS manteniendo todas (outer)**

	nombre_de_perro	fecha_nacimiento	raza	peso	dni_duenio	dni_duenio	nombre	apellido	fecha_nacimiento	direccion	ciudad
0	Luna	2020-05-10	Labrador	25.4	29543218.0	30214567.0	Susana	Horia	1985-04-12	Avenida Siempre Viva 123	Córdoba
1	Max	2019-11-22	Bulldog Francés	12.1	32219876.0	28456789.0	Oscar	Acol	1979-09-23	Calle Falsa 456	Neuquén
2	Simba	2021-03-15	Golden Retriever	30.2	30987654.0	31587964.0	Elena	Morado	1990-12-01	Pasaje Los Andes 789	Neuquén
3	Coco	2018-08-03	Caniche	7.5	29876543.0	29876543.0	Carlos	Fernández	1982-07-19	Belgrano 321	La Plata
4	Rocky	2022-01-27	Beagle	10.3	28881234.0	32765432.0	Lucila	Nave	1995-11-30	Santa Fe 654	Salta
5	Toby	2020-12-14	Doberman	34.7	31587964.0	30987654.0	Paloma	Blanca	1988-06-15	Rivadavia 102	Bahía Blanca
6	Mila	2019-07-30	Border Collie	18.2	30214567.0	29543218.0	Zoyla	Vaca	1977-03-05	Catamarca 89	Salta
7	Bruno	2023-02-05	Boxer	28.6	28456789.0	32219876.0	Esteban	Quito	1993-08-27	Mitre 12	Salta
8	Lola	2021-09-19	Shih Tzu	6.8	28881234.0	28881234.0	Elsa	Capunta	1984-10-10	9 de Julio 456	Neuquén
9	Negrita	2010-08-03	Pastor Belga	40.3	29876543.0	19333465.0	Aquiles	Muestro	1964-03-05	Av Argentina 233	Neuquén
10	Bruna	2014-12-13	Mestizo	13.2	NaN	32465340.0	Zacarias	Flores	1995-11-15	Av San Martin 323	Neuquén
11	Toreto	2022-11-27	Caniche	9.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
12	Osi	2019-12-23	Boxer	25.8	29543218.0	NaN	NaN	NaN	NaN	NaN	NaN

PROCESO DE ANÁLISIS DE DATOS
Mezclar Datos



- Unir considerando su ***semántica*** (método **merge()**)
  - Es un join de bases de datos
  - En donde se puede implementar **'inner', 'left', 'right', 'outer'**

- **inner join on dni_duenio**

	nombre_de_perro	fecha_nacimiento_x	raza	peso	dni_duenio	nombre	apellido	fecha_nacimiento_y	direccion	ciudad
0	Luna	2020-05-10 00:00:00	Labrador	25.400000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
1	Max	2019-11-22 00:00:00	Bulldog Francés	12.100000	32219876	Esteban	Quito	1993-08-27 00:00:00	Mitre 12	Salta
2	Simba	2021-03-15 00:00:00	Golden Retriever	30.200000	30987654	Paloma	Blanca	1988-06-15 00:00:00	Rivadavia 102	Bahía Blanca
3	Coco	2018-08-03 00:00:00	Caniche	7.500000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
4	Rocky	2022-01-27 00:00:00	Beagle	10.300000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
5	Toby	2020-12-14 00:00:00	Doberman	34.700000	31587964	Elena	Morado	1990-12-01 00:00:00	Pasaje Los Andes 789	Neuquén
6	Mila	2019-07-30 00:00:00	Border Collie	18.200000	30214567	Susana	Horia	1985-04-12 00:00:00	Avenida Siempre Viva 123	Córdoba
7	Bruno	2023-02-05 00:00:00	Boxer	28.600000	28456789	Oscar	Acol	1979-09-23 00:00:00	Calle Falsa 456	Neuquén
8	Lola	2021-09-19 00:00:00	Shih Tzu	6.800000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
9	Negrita	2010-08-03 00:00:00	Pastor Belga	40.300000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
10	Osi	2019-12-23 00:00:00	Boxer	25.800000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta

- **left join on dni_duenio**

	nombre_de_perro	fecha_nacimiento_x	raza	peso	dni_duenio	nombre	apellido	fecha_nacimiento_y	direccion	ciudad
0	Luna	2020-05-10 00:00:00	Labrador	25.400000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
1	Max	2019-11-22 00:00:00	Bulldog Francés	12.100000	32219876	Esteban	Quito	1993-08-27 00:00:00	Mitre 12	Salta
2	Simba	2021-03-15 00:00:00	Golden Retriever	30.200000	30987654	Paloma	Blanca	1988-06-15 00:00:00	Rivadavia 102	Bahía Blanca
3	Coco	2018-08-03 00:00:00	Caniche	7.500000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
4	Rocky	2022-01-27 00:00:00	Beagle	10.300000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
5	Toby	2020-12-14 00:00:00	Doberman	34.700000	31587964	Elena	Morado	1990-12-01 00:00:00	Pasaje Los Andes 789	Neuquén
6	Mila	2019-07-30 00:00:00	Border Collie	18.200000	30214567	Susana	Horia	1985-04-12 00:00:00	Avenida Siempre Viva 123	Córdoba
7	Bruno	2023-02-05 00:00:00	Boxer	28.600000	28456789	Oscar	Acol	1979-09-23 00:00:00	Calle Falsa 456	Neuquén
8	Lola	2021-09-19 00:00:00	Shih Tzu	6.800000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
9	Negrita	2010-08-03 00:00:00	Pastor Belga	40.300000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
10	Bruna	2014-12-13 00:00:00	Mestizo	13.200000	<NA>	nan	nan	NaT	nan	nan
11	Toreto	2022-11-27 00:00:00	Caniche	9.300000	<NA>	nan	nan	NaT	nan	nan
12	Osi	2019-12-23 00:00:00	Boxer	25.800000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta

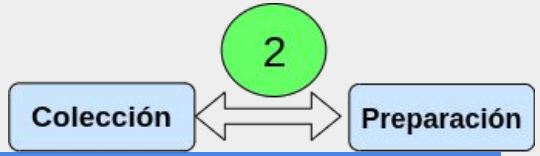
- **right join on dni_duenio**

	nombre_de_perro	fecha_nacimiento_x	raza	peso	dni_duenio	nombre	apellido	fecha_nacimiento_y	direccion	ciudad
0	Mila	2019-07-30 00:00:00	Border Collie	18.200000	30214567	Susana	Horia	1985-04-12 00:00:00	Avenida Siempre Viva 123	Córdoba
1	Bruno	2023-02-05 00:00:00	Boxer	28.600000	28456789	Oscar	Acol	1979-09-23 00:00:00	Calle Falsa 456	Neuquén
2	Toby	2020-12-14 00:00:00	Doberman	34.700000	31587964	Elena	Morado	1990-12-01 00:00:00	Pasaje Los Andes 789	Neuquén
3	Coco	2018-08-03 00:00:00	Caniche	7.500000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
4	Negrita	2010-08-03 00:00:00	Pastor Belga	40.300000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
5	nan	NaT	nan	nan	32765432	Lucila	Nave	1995-11-30 00:00:00	Santa Fe 654	Salta
6	Simba	2021-03-15 00:00:00	Golden Retriever	30.200000	30987654	Paloma	Blanca	1988-06-15 00:00:00	Rivadavia 102	Bahía Blanca
7	Luna	2020-05-10 00:00:00	Labrador	25.400000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
8	Osi	2019-12-23 00:00:00	Boxer	25.800000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
9	Max	2019-11-22 00:00:00	Bulldog Francés	12.100000	32219876	Esteban	Quito	1993-08-27 00:00:00	Mitre 12	Salta
10	Rocky	2022-01-27 00:00:00	Beagle	10.300000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
11	Lola	2021-09-19 00:00:00	Shih Tzu	6.800000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
12	nan	NaT	nan	nan	19333465	Aquiles	Muestro	1964-03-05 00:00:00	Av Argentina 233	Neuquén
13	nan	NaT	nan	nan	32465340	Zacarias	Flores	1995-11-15 00:00:00	Av San Martin 323	Neuquén

- **outer join on dni_duenio**

	nombre_de_perro	fecha_nacimiento_x	raza	peso	dni_duenio	nombre	apellido	fecha_nacimiento_y	direccion	ciudad
0	nan	NaT	nan	nan	19333465	Aquiles	Muestro	1964-03-05 00:00:00	Av Argentina 233	Neuquén
1	Bruno	2023-02-05 00:00:00	Boxer	28.600000	28456789	Oscar	Acol	1979-09-23 00:00:00	Calle Falsa 456	Neuquén
2	Rocky	2022-01-27 00:00:00	Beagle	10.300000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
3	Lola	2021-09-19 00:00:00	Shih Tzu	6.800000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
4	Luna	2020-05-10 00:00:00	Labrador	25.400000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
5	Osi	2019-12-23 00:00:00	Boxer	25.800000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
6	Coco	2018-08-03 00:00:00	Caniche	7.500000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
7	Negrita	2010-08-03 00:00:00	Pastor Belga	40.300000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
8	Mila	2019-07-30 00:00:00	Border Collie	18.200000	30214567	Susana	Horia	1985-04-12 00:00:00	Avenida Siempre Viva 123	Córdoba
9	Simba	2021-03-15 00:00:00	Golden Retriever	30.200000	30987654	Paloma	Blanca	1988-06-15 00:00:00	Rivadavia 102	Bahía Blanca
10	Toby	2020-12-14 00:00:00	Doberman	34.700000	31587964	Elena	Morado	1990-12-01 00:00:00	Pasaje Los Andes 789	Neuquén
11	Max	2019-11-22 00:00:00	Bulldog Francés	12.100000	32219876	Esteban	Quito	1993-08-27 00:00:00	Mitre 12	Salta
12	nan	NaT	nan	nan	32465340	Zacarias	Flores	1995-11-15 00:00:00	Av San Martin 323	Neuquén
13	nan	NaT	nan	nan	32765432	Lucila	Nave	1995-11-30 00:00:00	Santa Fe 654	Salta
14	Bruna	2014-12-13 00:00:00	Mestizo	13.200000	<NA>	nan	nan	NaT	nan	nan
15	Toreto	2022-11-27 00:00:00	Caniche	9.300000	<NA>	nan	nan	NaT	nan	nan

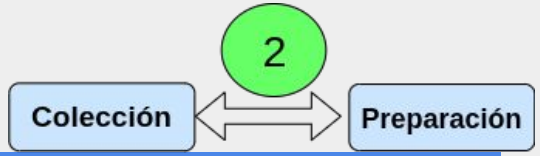
PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos



- Además de la integración, muchas veces es necesario realizar una **agregación de datos**
- La **agregación** es el proceso de reunir, resumir o combinar datos individuales en una forma más compacta y significativa
- Es transformar datos detallados en datos resumidos

Integración = juntar datos de distintas fuentes.
Agregación = resumir datos para análisis.

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos



- **Funciones Agregadas**: realizan un cálculo sobre un conjunto de valores y devuelve un único valor.
  - **MIN (atributo)**: devuelve el valor mínimo entre los valores de una columna (según orden numérico o alfabético).
  - **MAX (atributo)**: devuelve el valor máximo entre los valores de una columna (según orden numérico o alfabético).
  - **SUM (atributo)**: devuelve la suma total de los valores de una columna numérica.
  - **AVG (atributo)**: devuelve el promedio de los valores de una columna numérica.
  - **COUNT (atributo)**: devuelve la cantidad de registros que cumplen un criterio.

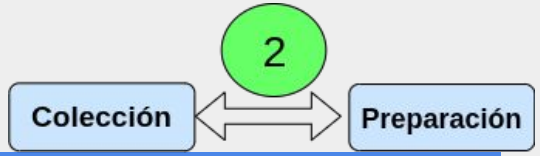
también existe el
COUNT-DISTINCT

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos

	nombre_de_perro	fecha_nacimiento	raza	peso	dni_duenio
0	Luna	2020-05-10	Labrador	25.4	29543218.0
1	Max	2019-11-22	Bulldog Francés	12.1	32219876.0
2	Simba	2021-03-15	Golden Retriever	30.2	30987654.0
3	Coco	2018-08-03	Caniche	7.5	29876543.0
4	Rocky	2022-01-27	Beagle	10.3	28881234.0
5	Toby	2020-12-14	Doberman	34.7	31587964.0
6	Mila	2019-07-30	Border Collie	18.2	30214567.0
7	Bruno	2023-02-05	Boxer	28.6	28456789.0
8	Lola	2021-09-19	Shih Tzu	6.8	28881234.0
9	Negrita	2010-08-03	Pastor Belga	40.3	29876543.0
10	Bruna	2014-12-13	Mestizo	13.2	NaN
11	Toreto	2022-11-27	Caniche	9.3	NaN
12	Osi	2019-12-23	Boxer	25.8	29543218.0

1. Cuánto pesan todos los perros?
sum(pesos)
2. Cuantos perros tengo?
count(nombre_perro)
3. Cuál es el perro más chiquito?
min(fecha_nacimiento)
4. Cuál es el perro más grande?
max(fecha_nacimiento)
5. Cuál es el perro más liviano?
min(peso)

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos



Cláusula GROUP BY (1)

4

- **GROUP BY** se usa cuando algún atributo tiene valores que se repiten en varios registros, y se quiere agrupar los registros.
- El resultado tendrá una fila por cada agrupamiento de registros.
- ~~Es muy útil para usar con~~ las **Funciones Agregadas**.

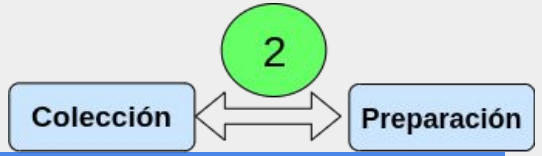
Curso BDs!!

Se usa con….

- **inner join on dni_duenio**

	nombre_de_perro	fecha_nacimiento_x	raza	peso	dni_duenio	nombre	apellido	fecha_nacimiento_y	direccion	ciudad
0	Luna	2020-05-10 00:00:00	Labrador	25.400000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
1	Max	2019-11-22 00:00:00	Bulldog Francés	12.100000	32219876	Esteban	Quito	1993-08-27 00:00:00	Mitre 12	Salta
2	Simba	2021-03-15 00:00:00	Golden Retriever	30.200000	30987654	Paloma	Blanca	1988-06-15 00:00:00	Rivadavia 102	Bahía Blanca
3	Coco	2018-08-03 00:00:00	Caniche	7.500000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
4	Rocky	2022-01-27 00:00:00	Beagle	10.300000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
5	Toby	2020-12-14 00:00:00	Doberman	34.700000	31587964	Elena	Morado	1990-12-01 00:00:00	Pasaje Los Andes 789	Neuquén
6	Mila	2019-07-30 00:00:00	Border Collie	18.200000	30214567	Susana	Horia	1985-04-12 00:00:00	Avenida Siempre Viva 123	Córdoba
7	Bruno	2023-02-05 00:00:00	Boxer	28.600000	28456789	Oscar	Acol	1979-09-23 00:00:00	Calle Falsa 456	Neuquén
8	Lola	2021-09-19 00:00:00	Shih Tzu	6.800000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
9	Negrita	2010-08-03 00:00:00	Pastor Belga	40.300000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
10	Osi	2019-12-23 00:00:00	Boxer	25.800000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos



1. Listar la cantidad de perros de cada persona. Se debe visualizar el nombre, apellido y cantidad.

```
SELECT nombre, apellido, count(nombre_de_perro)
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido
```

qué está mal en esta consulta?

1. Listar la cantidad de perros de cada persona. Se debe visualizar el nombre, apellido y cantidad.

```
SELECT nombre, apellido, count(nombre_de_perro)
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido
```

nombre	apellido	dni_duenio	nombre_de_perro
Carlos	Fernández	29876543	2
Elena	Morado	31587964	1
Elsa	Capunta	28881234	2
Esteban	Quito	32219876	1
Oscar	Acol	28456789	1
Paloma	Blanca	30987654	1
Susana	Horia	30214567	1
Zoyla	Vaca	29543218	2

qué está mal en esta consulta?

- **inner join on dni_duenio**

	nombre_de_perro	fecha_nacimiento_x	raza	peso	dni_duenio	nombre	apellido	fecha_nacimiento_y	direccion	ciudad
0	Luna	2020-05-10 00:00:00	Labrador	25.400000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
1	Max	2019-11-22 00:00:00	Bulldog Francés	12.100000	32219876	Esteban	Quito	1993-08-27 00:00:00	Mitre 12	Salta
2	Simba	2021-03-15 00:00:00	Golden Retriever	30.200000	30987654	Paloma	Blanca	1988-06-15 00:00:00	Rivadavia 102	Bahía Blanca
3	Coco	2018-08-03 00:00:00	Caniche	7.500000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
4	Rocky	2022-01-27 00:00:00	Beagle	10.300000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
5	Toby	2020-12-14 00:00:00	Doberman	34.700000	31587964	Elena	Morado	1990-12-01 00:00:00	Pasaje Los Andes 789	Neuquén
6	Mila	2019-07-30 00:00:00	Border Collie	18.200000	30214567	Susana	Horia	1985-04-12 00:00:00	Avenida Siempre Viva 123	Córdoba
7	Bruno	2023-02-05 00:00:00	Boxer	28.600000	28456789	Oscar	Acol	1979-09-23 00:00:00	Calle Falsa 456	Neuquén
8	Lola	2021-09-19 00:00:00	Shih Tzu	6.800000	28881234	Elsa	Capunta	1984-10-10 00:00:00	9 de Julio 456	Neuquén
9	Negrita	2010-08-03 00:00:00	Pastor Belga	40.300000	29876543	Carlos	Fernández	1982-07-19 00:00:00	Belgrano 321	La Plata
10	Osi	2019-12-23 00:00:00	Boxer	25.800000	29543218	Zoyla	Vaca	1977-03-05 00:00:00	Catamarca 89	Salta
	Frida	2019-12-23 00:00:00	Boxer	24.8	44432228	Zoyla	Vaca	2000-03-05 00:00:00	Rioja 23	Salta

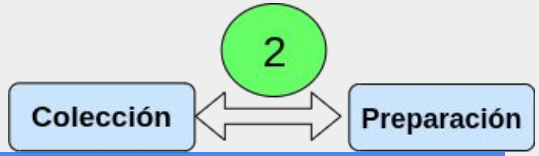
1. Listar la cantidad de perros de cada persona. Se debe visualizar el nombre, apellido y cantidad.

```
SELECT nombre, apellido, count(nombre_de_perro)
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido
```

nombre	apellido	nombre_de_perro
Carlos	Fernández	2
Elena	Morado	1
Elsa	Capunta	2
Esteban	Quito	1
Oscar	Acol	1
Paloma	Blanca	1
Susana	Horia	1
Zoyla	Vaca	3

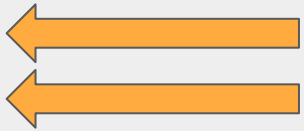
qué está mal en esta consulta?

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos

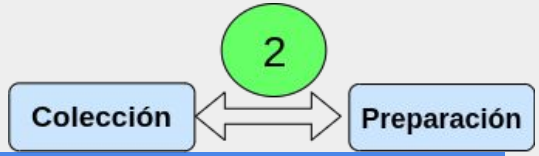



```
SELECT nombre, apellido, count(nombre_de_perro)
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido, d.dni_duenio
```

	nombre	apellido	dni_duenio	nombre_de_perro
0	Carlos	Fernández	29876543	2
1	Elena	Morado	31587964	1
2	Elsa	Capunta	28881234	2
3	Esteban	Quito	32219876	1
4	Oscar	Acol	28456789	1
5	Paloma	Blanca	30987654	1
6	Susana	Horia	30214567	1
7	Zoyla	Vaca	29543218	2
8	Zoyla	Vaca	44432228	1



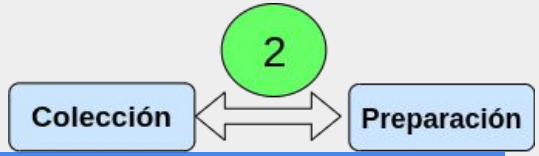
PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos




```sql
SELECT nombre, apellido,
count(nombre_de_perro)
FROM perros p INNER JOIN
duenios d ON  p.dni_duenio =
d.dni_duenio
GROUP BY nombre, apellido,
d.dni_duenio
```

```python
df = (df_join_duenio
      .groupby(['nombre','apellido','dni_duenio'])
      [['nombre_de_perro']]
      .count()
      .reset_index() #para que quede flat
     )
```

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos

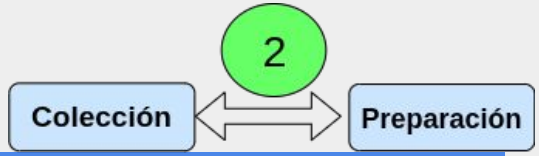



```sql
SELECT nombre, apellido,
count(nombre_de_perro)
FROM perros p INNER JOIN
duenios d ON  p.dni_duenio =
d.dni_duenio
GROUP BY nombre, apellido,
d.dni_duenio
```

XXX

```python
df = (df_join_duenio
      .groupby(['nombre','apellido','dni_duenio'])
      [['nombre_de_perro']]
      .count()
      .reset_index() #para que quede flat
     )
```

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos

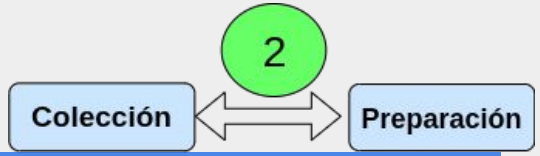



```sql
SELECT nombre, apellido,
count(nombre_de_perro)
FROM perros p INNER JOIN
duenios d ON  p.dni_duenio =
d.dni_duenio
GROUP BY nombre, apellido,
d.dni_duenio
```

```python
df = (df_join_duenio
      .groupby(['nombre','apellido','dni_duenio'])
      [['nombre_de_perro']]
      .count()
      .reset_index() #para que quede flat
     )
```

```python
df[['nombre','apellido','nombre_de_perro']]
```

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos



2. Listar la *cantidad de perros* con la *suma total de sus pesos* de cada persona.
Se debe visualizar el nombre, apellido, cantidadDePerros y sumaPesos.

```
SELECT nombre, apellido, count(nombre_de_perro) as cantidadDePerros,
sum(peso) as sumaPesos
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido, dni_duenio
```

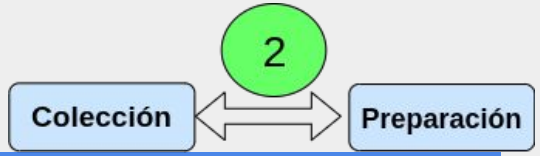
```
SELECT nombre, apellido, count(nombre_de_perro) as cantidadDePerros,
sum(peso) as sumaPesos
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido, dni_duenio
```

	nombre	apellido	dni_duenio	cantidadDePerros	sumaPesos
0	Carlos	Fernández	29876543	2	47.8
1	Elena	Morado	31587964	1	34.7
2	Elsa	Capunta	28881234	2	17.1
3	Esteban	Quito	32219876	1	12.1
4	Oscar	Acol	28456789	1	28.6
5	Paloma	Blanca	30987654	1	30.2
6	Susana	Horia	30214567	1	18.2
7	Zoyla	Vaca	29543218	2	51.2
8	Zoyla	Vaca	44432228	1	24.8

```sql
SELECT nombre, apellido, count(nombre_de_perro) as cantidadDePerros,
sum(peso) as sumaPesos
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido, dni_duenio
```

```python
df = (df_join_duenio
      .groupby(['nombre','apellido','dni_duenio'])
      .agg(
          cantidadDePerros = ('nombre_de_perro', 'count'),
          sumaPesos = ('peso', 'sum')
      )
     .reset_index()
     )
```

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos



3. Listar la *cantidad de perros* con la *suma total de sus pesos* de cada persona.
Se debe visualizar el nombre, apellido, cantidadDePerros y sumaPesos, siempre y cuando las personas tengan más de 1 perro.

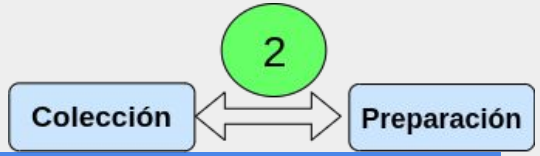
```
SELECT nombre, apellido, count(nombre_de_perro) as cantidadDePerros,
sum(peso) as sumaPesos
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido, dni_duenio
HAVING count(nombre_de_perro)>1
```

```sql
SELECT nombre, apellido, count(nombre_de_perro) as cantidadDePerros,
sum(peso) as sumaPesos
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY nombre, apellido, dni_duenio
HAVING count(nombre_de_perro)>1
```

```python
df = (df_join_duenio
      .groupby(['nombre','apellido','dni_duenio'])
      .agg(
          cantidadDePerros = ('nombre_de_perro', 'count'),
          sumaPesos = ('peso', 'sum')
      )
     .reset_index()
     ).query('cantidadDePerros >1')
```

	nombre	apellido	dni_duenio	cantidadDePerros	sumaPesos
0	Carlos	Fernández	29876543	2	47.8
2	Elsa	Capunta	28881234	2	17.1
7	Zoyla	Vaca	29543218	2	51.2

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos



4. Listar la *cantidad de perros* nacidos por año. Se debe visualizar el año y la cantidadDePerros.

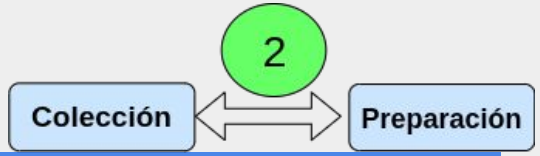
```
SELECT YEAR(fecha_nacimiento_x), count(nombre_de_perro) as
cantidadDePerros
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY (fecha_nacimiento_x)
```

```sql
SELECT YEAR(fecha_nacimiento_x), count(nombre_de_perro) as cantidadDePerros
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
GROUP BY (fecha_nacimiento_x)
```

```python
df = (df_join_duenio
      .groupby([df_join_duenio['fecha_nacimiento_x'].dt.year])
      .agg(
          cantidadDePerros = ('nombre_de_perro', 'count'))
     .reset_index()
     )
```

	fecha_nacimiento_x	cantidadDePerros
0	2010	1
1	2018	1
2	2019	4
3	2020	2
4	2021	2
5	2022	1
6	2023	1

PROCESO DE ANÁLISIS DE DATOS
Agregación de Datos



4. Listar la *cantidad de perros* nacidos por año, **siempre y cuando su dueño viva en Salta o Neuquén**. Se debe visualizar el año y la cantidadDePerros.

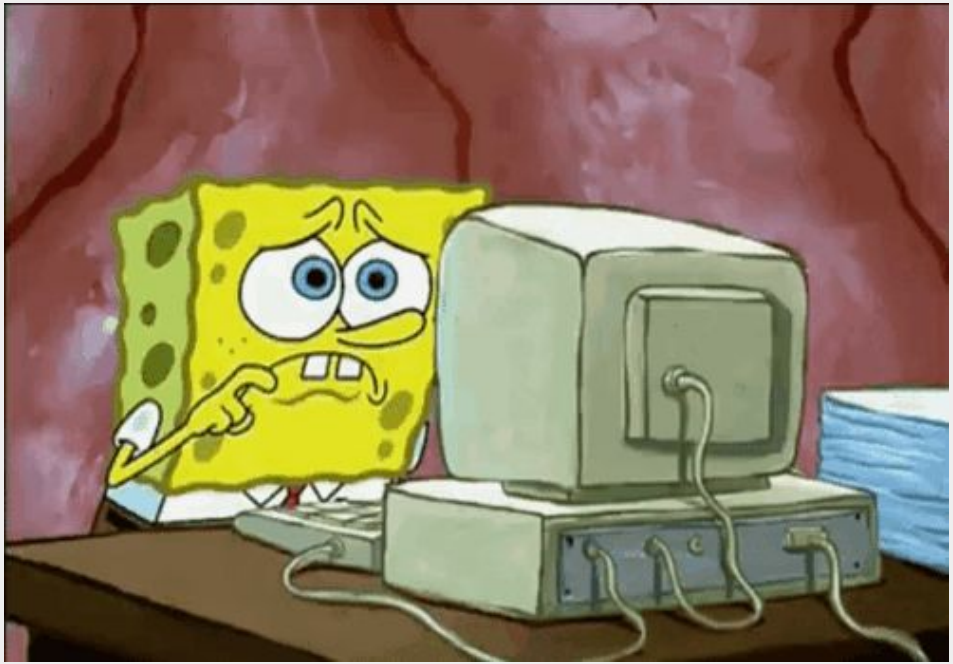
```
SELECT YEAR(fecha_nacimiento_x), count(nombre_de_perro) as
cantidadDePerros
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
WHERE d.ciudad LIKE 'Salta' OR 'Neuquén'
GROUP BY (fecha_nacimiento_x)
```

```sql
SELECT YEAR(fecha_nacimiento_x), count(nombre_de_perro) as cantidadDePerros
FROM perros p INNER JOIN duenios d ON  p.dni_duenio = d.dni_duenio
WHERE d.ciudad LIKE 'Salta' OR 'Neuquén'
GROUP BY (fecha_nacimiento_x)
```

```python
df = (df_join_duenio.loc[(df_join_duenio['ciudad'] == 'Salta') | (df_join_duenio['ciudad'] == 'Neuquén')]
      .groupby([df_join_duenio['fecha_nacimiento_x'].dt.year])
      .agg(
          cantidadDePerros = ('nombre_de_perro', 'count'))
     .reset_index()
     )
```

	fecha_nacimiento_x	cantidadDePerros
0	2019	3
1	2020	2
2	2021	1
3	2022	1
4	2023	1

PROCESO DE ANÁLISIS DE DATOS
A TRABAJAR!!



PROCESO DE ANÁLISIS DE DATOS
Con qué seguimos?