

# Bases de Datos

Clase 12: Privacidad

El uso de datos personales tiene riesgos

Los datos pueden contener  
información sensible o confidencial  
de individuos

Y su uso puede filtrar (parte de)  
esta información sensible

# El uso de datos personales tiene riesgos

Los datos pueden contener  
información sensible o confidencial  
de individuos



Y su uso puede **filtrar** (parte de)  
esta información sensible

# El caso de Netflix

# El caso de Netflix

## Competencia Netflix 2007

Competencia abierta para mejorar el sistema de recomendación:

**PREMIO:** 1.000.000 USD

**PARTICIPANTES:** acceso a un dataset de entrenamiento **anonimizado**



# El caso de Netflix

## Competencia Netflix 2007

Competencia abierta para mejorar el sistema de recomendación:

**PREMIO:** 1.000.000 USD

**PARTICIPANTES:** acceso a un dataset de entrenamiento **anonimizado**

“To protect customer privacy, all personal information identifying individual customers has been removed and all customer ids [sic] have been replaced by randomly assigned ids [sic].”



Netflix anonymised dataset		
X		
X		
X		
X		

# El caso de Netflix

## Competencia Netflix 2007

Competencia abierta para mejorar el sistema de recomendación:

**PREMIO:** 1.000.000 USD

**PARTICIPANTES:** acceso a un dataset de entrenamiento **anonimizado**

“To protect customer privacy, all personal information identifying individual customers has been removed and all customer ids [sic] have been replaced by randomly assigned ids [sic].”



Netflix anonymised dataset		
X		
X		
X		
X		



IMDB public (non-anonymised) records		
Alice		
Bob		
Caty		
Tim		

# El caso de Netflix

## Competencia Netflix 2007

Competencia abierta para mejorar el sistema de recomendación:

**PREMIO:** 1.000.000 USD

**PARTICIPANTES:** acceso a un dataset de entrenamiento **anonimizado**

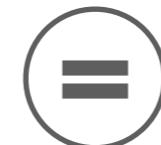


"To protect customer privacy, all personal information identifying individual customers has been removed and all customer ids [sic] have been replaced by randomly assigned ids [sic]."

Netflix anonymised dataset		
X		
X		
X		
X		



IMDB public (non-anonymised) records		
Alice		
Bob		
Caty		
Tim		



Netflix <u>deanonymised</u> dataset		
Alice		
Bob		
—		
Tim		

# El caso de Netflix

## Competencia Netflix 2007

Competencia abierta para mejorar el sistema de recomendación:

**PREMIO:** 1.000.000 USD

**PARTICIPANTES:** acceso a un dataset de entrenamiento **anonimizado**

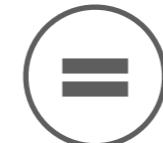


"To protect customer privacy, all personal information identifying individual customers has been removed and all customer ids [sic] have been replaced by randomly assigned ids [sic]."

Netflix anonymised dataset		
X		
X		
X		
X		



IMDB public (non-anonymised) records		
Alice		
Bob		
Caty		
Tim		



Netflix <u>deanonymised</u> dataset		
Alice		
Bob		
—		
Tim		

Ataque de asociación de registros

# El caso del gobernador de Massachusetts

# El caso del gobernador de Massachusetts

En 1997, William Weld, gobernador de MA aprobó la liberación de los registros médicos de funcionarios públicos, post anonimización:



# El caso del gobernador de Massachusetts

En 1997, William Weld, gobernador de MA aprobó la liberación de los registros médicos de funcionarios públicos, post anonimización:

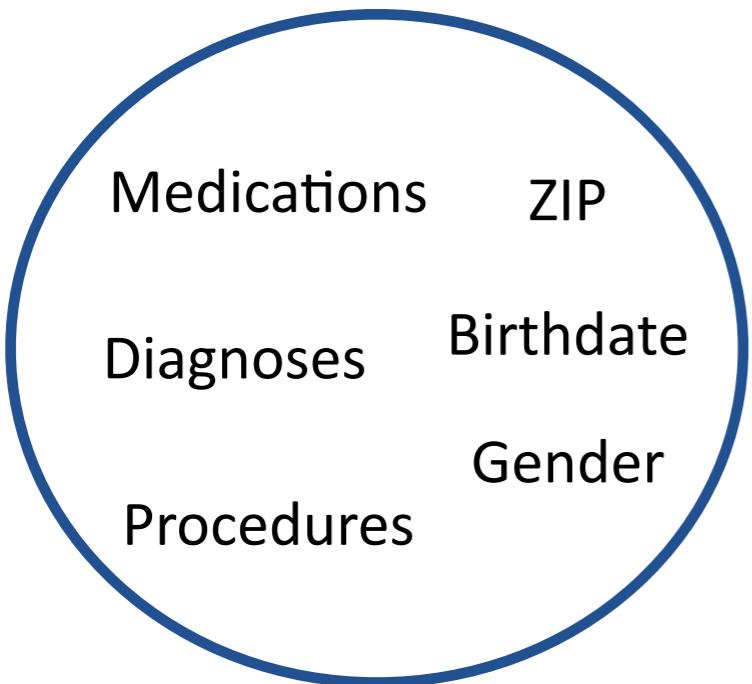


Dos días después, Latanya Sweeney, una estudiante de doctorado del MIT, le envió un correo con sus registros médicos



# El caso del gobernador de Massachusetts

En 1997, William Weld, gobernador de MA aprobó la liberación de los registros médicos de funcionarios públicos, post anonimización:

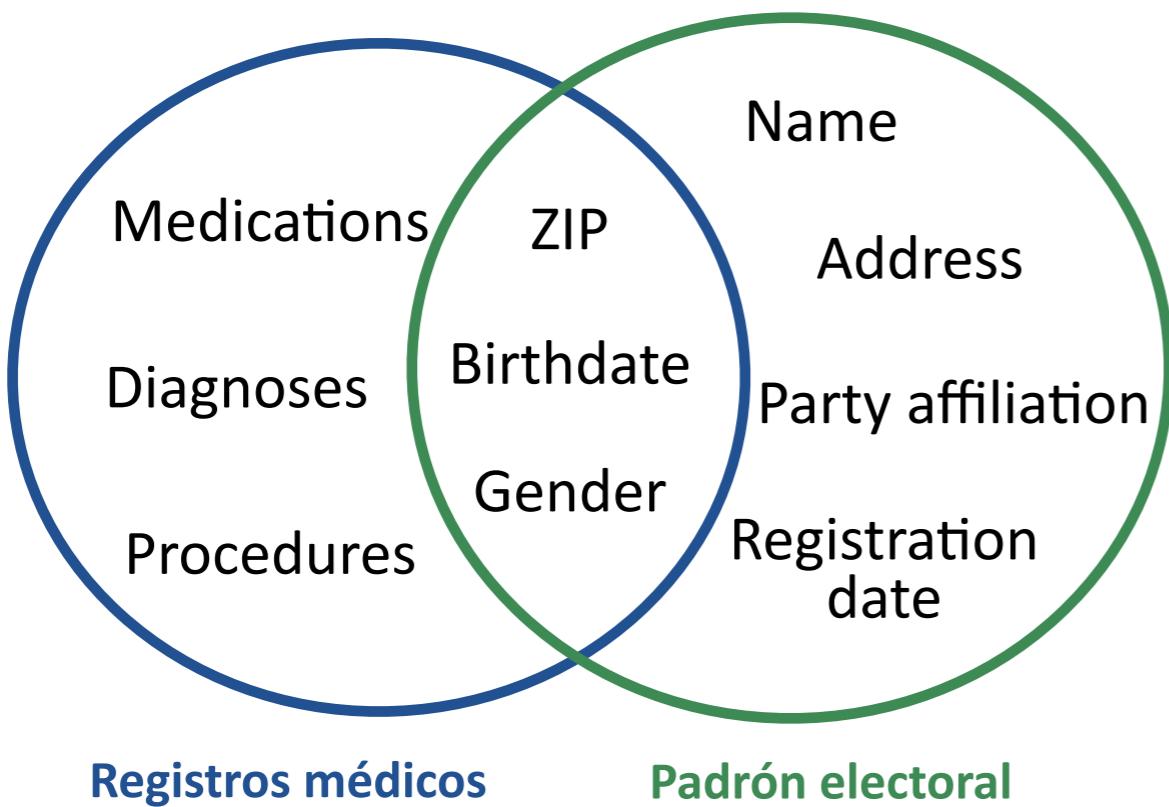


Dos días después, Latanya Sweeney, una estudiante de doctorado del MIT, le envió un correo con sus registros médicos



# El caso del gobernador de Massachusetts

En 1997, William Weld, gobernador de MA aprobó la liberación de los registros médicos de funcionarios públicos, post anonimización:



Dos días después, Latanya Sweeney, una estudiante de doctorado del MIT, le envió un correo con sus registros médicos



# Agenda

# Agenda

- **Informalmente:** Evitar que los datos de un individuo se hagan públicos o conocidos por terceros se llama *privacidad*

# Agenda

- **Informalmente:** Evitar que los datos de un individuo se hagan públicos o conocidos por terceros se llama *privacidad*
- Dada la gran cantidad de datos que existe hoy se hace necesario asegurar la privacidad de las bases de datos y sus usos

# Agenda

- **Informalmente:** Evitar que los datos de un individuo se hagan públicos o conocidos por terceros se llama *privacidad*
- Dada la gran cantidad de datos que existe hoy se hace necesario asegurar la privacidad de las bases de datos y sus usos
- Para intentar preservar privacidad hay distintas técnicas. Esta clase veremos algunas:
  - De-identificación / Anonimización
    - $k$ -anonimato
    - $\ell$ -diversidad
  - Privacidad diferencial

# Privacidad

# Privacidad

Un análisis de datos **preserva privacidad** si:

1. Aprendes algo útil del análisis
2. El análisis no viola la privacidad de algún individuo

# Privacidad

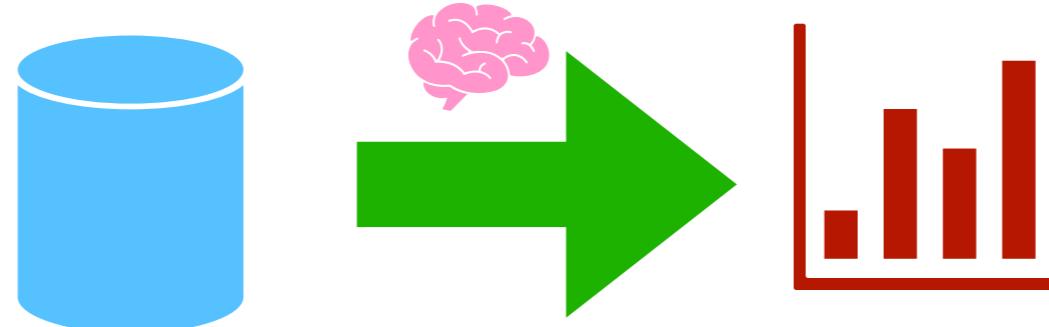
# Privacidad

La **privacidad** de un individuo A es **violada** si:

# Privacidad

La **privacidad** de un individuo A es **violada** si:

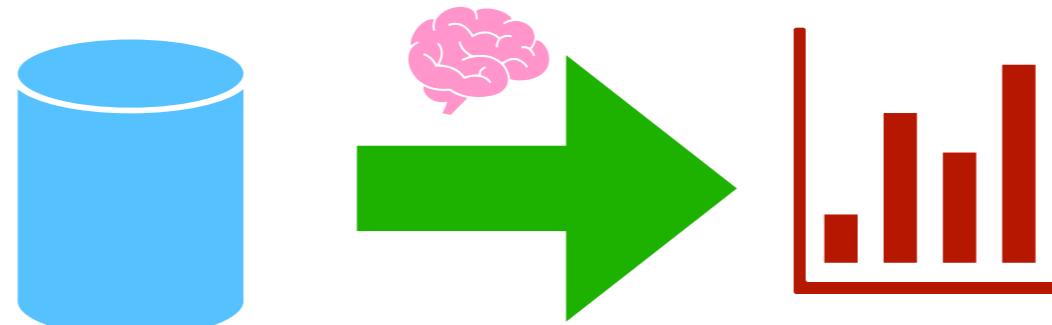
1. El analista aprende B si no cuenta con los datos del individuo A



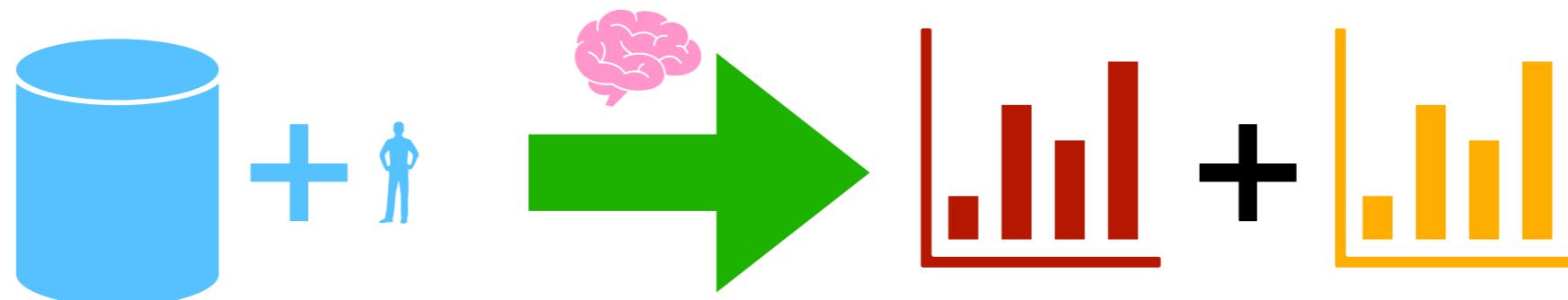
# Privacidad

La **privacidad** de un individuo A es **violada** si:

1. El analista aprende B si no cuenta con los datos del individuo A



2. El analista ahora aprende B+C si agrega los datos de A al conjunto de datos

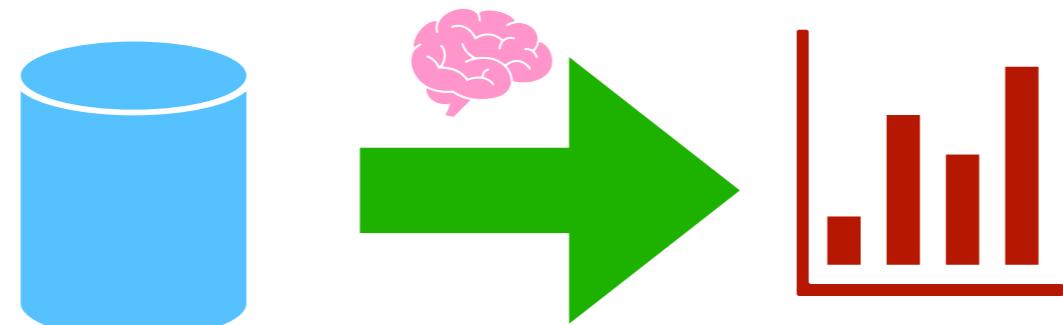


# Privacidad

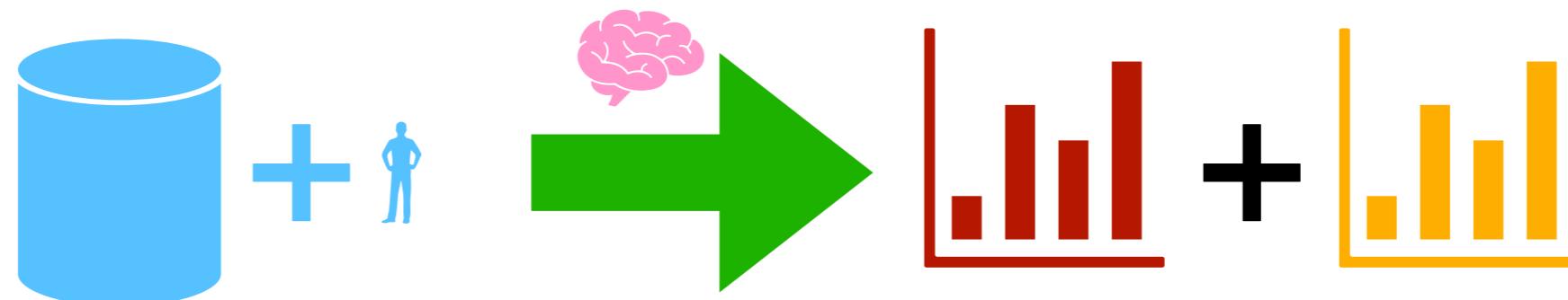
Un análisis preserva privacidad de un individuo si no importa si sus datos son considerados en el análisis.

La **privacidad** de un individuo A es **violada** si:

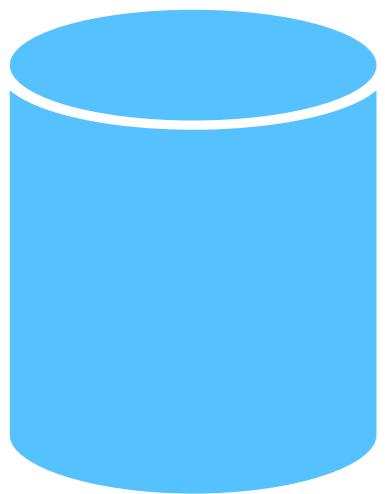
1. El analista aprende B si no cuenta con los datos del individuo A



2. El analista ahora aprende B+C si agrega los datos de A al conjunto de datos



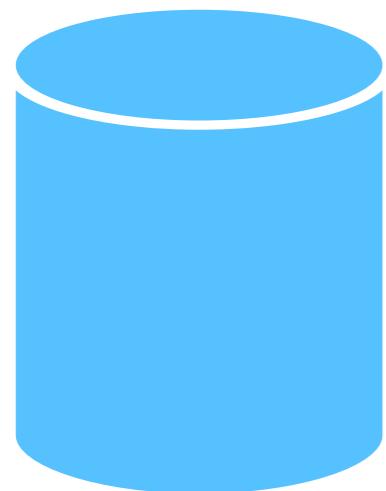
# Privacidad vs Seguridad



# Privacidad vs Seguridad



**Seguridad** es acerca del acceso a la información.

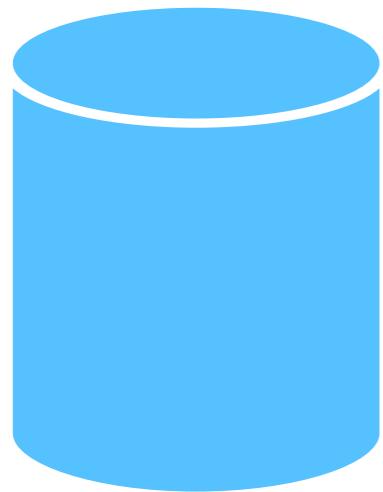


# Privacidad vs Seguridad



**Seguridad** es acerca del acceso a la información.

Confidencialidad: ¿quién  
puede ver los datos?



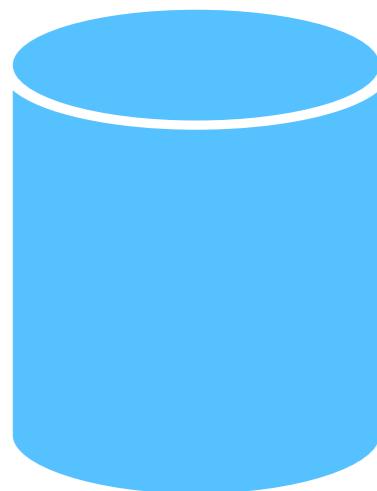
# Privacidad vs Seguridad



**Seguridad** es acerca del acceso a la información.

Confidencialidad: ¿quién puede ver los datos?

Integridad: ¿quién puede modificar los datos?



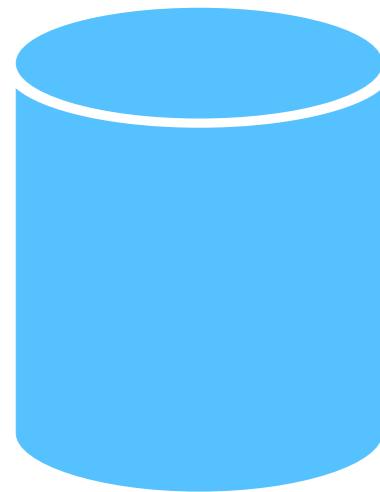
# Privacidad vs Seguridad



**Seguridad** es acerca del acceso a la información.

Confidencialidad: ¿quién puede ver los datos?

Integridad: ¿quién puede modificar los datos?



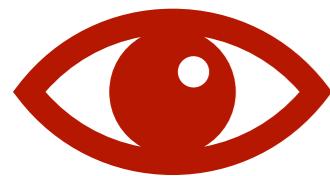
# Privacidad vs Seguridad



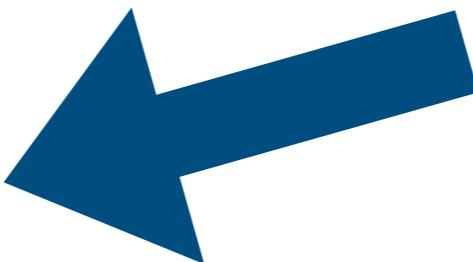
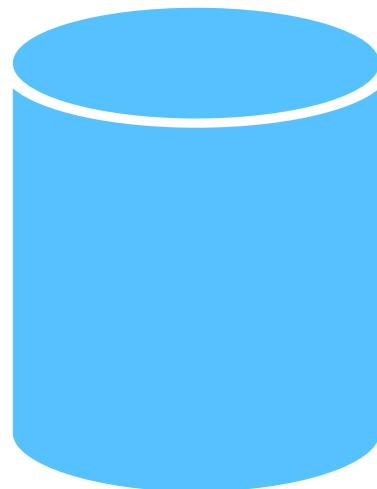
**Seguridad** es acerca del acceso a la información.

Confidencialidad: ¿quién puede ver los datos?

Integridad: ¿quién puede modificar los datos?



**Privacidad** es acerca de la información personal.



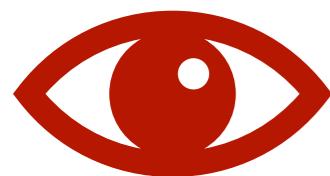
# Privacidad vs Seguridad



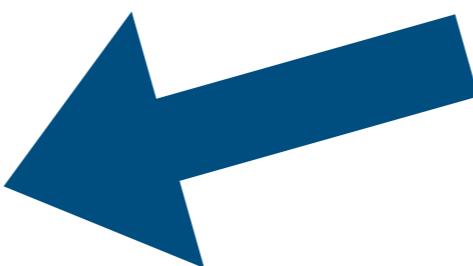
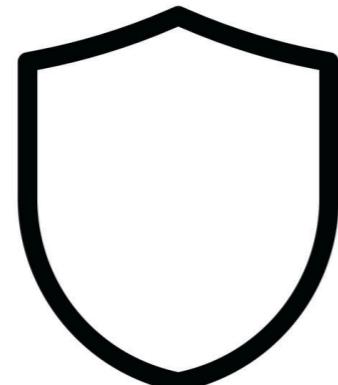
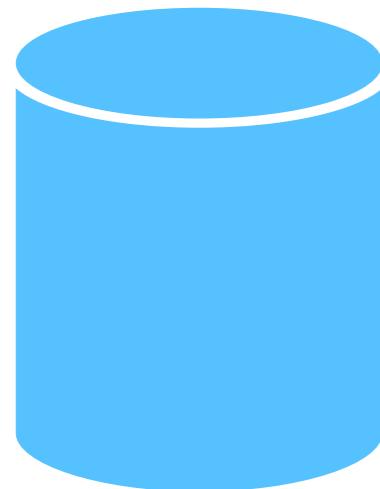
**Seguridad** es acerca del acceso a la información.

Confidencialidad: ¿quién puede ver los datos?

Integridad: ¿quién puede modificar los datos?



**Privacidad** es acerca de la información personal.



¿qué podemos aprender de los datos (el contenido)?

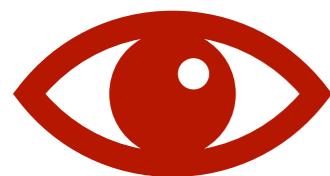
# Privacidad vs Seguridad



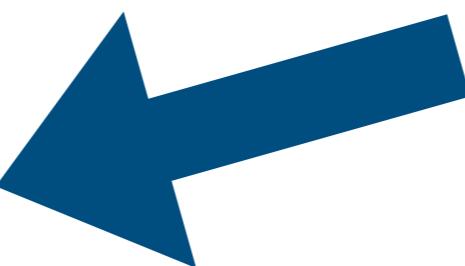
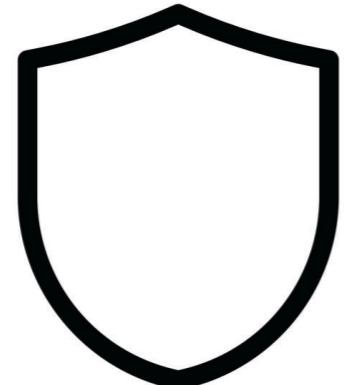
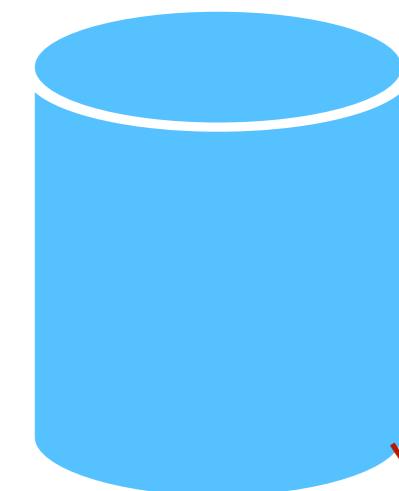
**Seguridad** es acerca del acceso a la información.

Confidencialidad: ¿quién puede ver los datos?

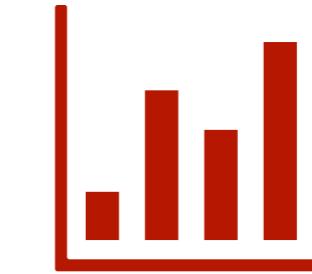
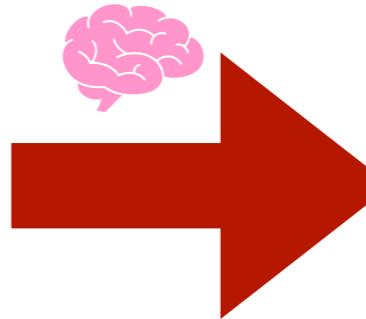
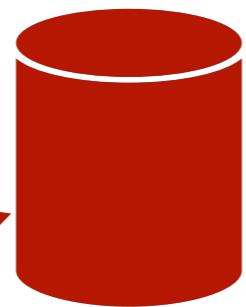
Integridad: ¿quién puede modificar los datos?



**Privacidad** es acerca de la información personal.



¿qué podemos aprender de los datos (el contenido)?



# ALGUNAS DE TÉCNICAS DE PRIVACIDAD

# Tipos de datos

# Tipos de datos

**Los datos sintéticos** son datos generados artificialmente, como forma opuesta a recolectarlos del mundo real.

# Tipos de datos

**Los datos sintéticos** son datos generados artificialmente, como forma opuesta a recolectarlos del mundo real.

# Tipos de datos

**Los datos sintéticos** son datos generados artificialmente, como forma opuesta a recolectarlos del mundo real.

**Las respuestas a consultas** se refiere a datos agregados de alguna forma

# Datos sintéticos vs respuesta de consultas

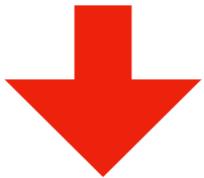
Los datos sintéticos *lucen* como los datos originales:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825

# Datos sintéticos vs respuesta de consultas

Los datos sintéticos *lucen* como los datos originales:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



# Datos sintéticos vs respuesta de consultas

Los datos sintéticos *lucen* como los datos originales:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



Nombre	Fecha Nac.	Sexo	Coding Postal
*****	26/02/2018	M	73909
*****	08/05/2018	M	14890
*****	11/05/2018	F	58821
*****	31/03/2019	F	465113
*****	26/04/2018	F	90825

# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825

# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



¿Cuántas personas nacieron en el 2018?

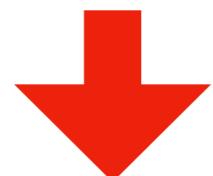
# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



¿Cuántas personas nacieron en el 2018?



# Datos sintéticos vs respuesta de consultas

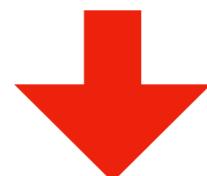
Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



¿Cuántas personas nacieron en el 2018?

¿Cuántas personas hay de cada sexo?



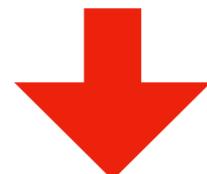
# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825

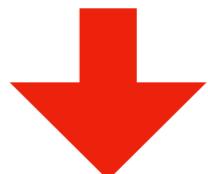


¿Cuántas personas nacieron en el 2018?



4

¿Cuántas personas hay de cada sexo?



M=2 y F=3

# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825

# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



¿Cuántas personas nacieron por año?

# Datos sintéticos vs respuesta de consultas

Las respuestas a consultas *requieren* una consulta en particular:

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



¿Cuántas personas nacieron por año?

Fecha Nac.	Cantidad
2018	4
2019	1

# Algunas técnicas de privacidad

Técnica	Funcionalidad
Anonimización	Datos sintéticos
k-Anonimato	Datos sintéticos
l-diversidad	Datos sintéticos
Privacidad diferencial	Respuestas a consultas

# Datos sintéticos vs respuesta de consultas

# Datos sintéticos vs respuesta de consultas

## Datos sintéticos

1. Permite re-utilizar análisis de datos existentes
2. Funciona para todas las consultas
3. Facilita las cosas a los analistas
4. **Imposible de obtener perfecta utilidad y fuerte privacidad**

# Datos sintéticos vs respuesta de consultas

## Datos sintéticos

1. Permite re-utilizar análisis de datos existentes
2. Funciona para todas las consultas
3. Facilita las cosas a los analistas
4. **Imposible** de obtener perfecta **utilidad** y fuerte **privacidad**

## Respuestas de consultas

1. Requiere a veces análisis de datos modificados
2. Depende de cada consulta
3. Dificulta las cosas para los analistas
4. La especialización a *una consulta* permite obtener **mejor utilidad** y **privacidad**.

# Utilidad

# Utilidad

**“Qué tan útil es la respuesta”**

# Utilidad

**“Qué tan útil es la respuesta”**

(depende de en que será usada la respuesta)

# Utilidad

**“Qué tan útil es la respuesta”**

(depende de en que será usada la respuesta)

¿Cuántas personas se llaman Ford?

# Utilidad

**“Qué tan útil es la respuesta”**

(depende de en que será usada la respuesta)

**¿Cuántas personas se llaman Ford?**

Datos  
anonimizados

# Utilidad

**“Qué tan útil es la respuesta”**

(depende de en que será usada la respuesta)

**¿Cuántas personas se llaman Ford?**

Datos  
anonimizados

Privacidad  
diferencial

# Utilidad

**“Qué tan útil es la respuesta”**

(depende de en que será usada la respuesta)

¿Cuántas personas se llaman Ford?

Datos  
anonimizados

Impossible de responder

Privacidad  
diferencial

# Utilidad

**“Qué tan útil es la respuesta”**

(depende de en que será usada la respuesta)

¿Cuántas personas se llaman Ford?

Datos  
anonimizados

Impossible de responder

Privacidad  
diferencial

Podría responder  $\pm 1$  personas

# ANONIMIZACIÓN / DE-IDENTIFICACIÓN

# Definiciones y convenciones básicas

Los atributos de una tabla se clasifican en 4 categorías disjuntas

<b>Identificador explícito (IdE)</b>	conjunto de atributos que individualmente identifican de manera explícita al “dueño” de un registro. Por ejemplo, nombre y RUT.
<b>Cuasi identificador (CId)</b>	Datos conjunto de atributos que colectivamente tienen el potencial de identificar al dueño de un registro. Por ejemplo, día de nacimiento y dirección.
<b>Atributos sensibles (AS)</b>	conjunto de atributos cuyo valor se desea proteger. Por ejemplo, religión, etnia, enfermedad.
<b>Atributos no-sensibles (AnS)</b>	conjunto de atributos que no tienen un carácter privado (y por lo tanto no se busca proteger). Por ejemplo, equipo de fútbol o película favorita.

# Definiciones y convenciones básicas

Los atributos de una tabla se clasifican en 4 categorías disjuntas

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
Rashad	26/02/2018	M	73909	Resfrío
Alyssa	08/05/2018	M	14890	Hepatitis
Myra Ford	11/05/2018	F	58821	Hepatitis
Meredith	31/03/2019	F	465113	VIH
Aimee	26/04/2018	F	90825	Bronquitis

# Definiciones y convenciones básicas

Los atributos de una tabla se clasifican en 4 categorías disjuntas

Identificador

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
Rashad	26/02/2018	M	73909	Resfrío
Alyssa	08/05/2018	M	14890	Hepatitis
Myra Ford	11/05/2018	F	58821	Hepatitis
Meredith	31/03/2019	F	465113	VIH
Aimee	26/04/2018	F	90825	Bronquitis

# Definiciones y convenciones básicas

Los atributos de una tabla se clasifican en 4 categorías disjuntas

Identificador		Cuasi Identificador		
Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
Rashad	26/02/2018	M	73909	Resfrío
Alyssa	08/05/2018	M	14890	Hepatitis
Myra Ford	11/05/2018	F	58821	Hepatitis
Meredith	31/03/2019	F	465113	VIH
Aimee	26/04/2018	F	90825	Bronquitis

# Definiciones y convenciones básicas

Los atributos de una tabla se clasifican en 4 categorías disjuntas

Identificador		Cuasi Identificador		Sensible
Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
Rashad	26/02/2018	M	73909	Resfrío
Alyssa	08/05/2018	M	14890	Hepatitis
Myra Ford	11/05/2018	F	58821	Hepatitis
Meredith	31/03/2019	F	465113	VIH
Aimee	26/04/2018	F	90825	Bronquitis

# Definiciones y convenciones básicas

Los atributos de una tabla se clasifican en 4 categorías disjuntas

Identificador		Cuasi Identificador		Sensible
Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
Rashad	26/02/2018	M	73909	Resfrío
Alyssa	08/05/2018	M	14890	Hepatitis
Myra Ford	11/05/2018	F	58821	Hepatitis
Meredith	31/03/2019	F	465113	VIH
Aimee	26/04/2018	F	90825	Bronquitis
...				

No Sensibles

# De-identificación

De-identificación es el proceso que remueve la asociación entre una persona y un conjunto de datos.

## Objetivos:

- Reducir el riesgo de violación de privacidad
- Maximizar la utilidad de los datos

## Técnicas:

- Supresión (remover datos)
- Variación (“revolver” los datos)
- Enmascaramiento
- Etc.

# De-identificación: Ejemplo

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825

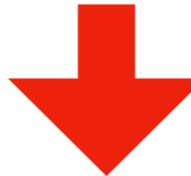
# De-identificación: Ejemplo

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



# De-identificación: Ejemplo

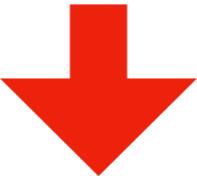
Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



Nombre	Fecha Nac.	Sexo	Coding Postal
*****	26/02/2018	M	73909
*****	08/05/2018	M	14890
*****	11/05/2018	F	58821
*****	31/03/2019	F	465113
*****	26/04/2018	F	90825

# De-identificación: Ejemplo

Nombre	Fecha Nac.	Sexo	Coding Postal
Rashad Arnold	26/02/2018	M	73909
Alyssa Cherry	08/05/2018	M	14890
Myra Ford	11/05/2018	F	58821
Meredith Perry	31/03/2019	F	465113
Aimee Thornton	26/04/2018	F	90825



Nombre	Fecha Nac.	Sexo	Coding Postal
*****	26/02/2018	M	73909
*****	08/05/2018	M	14890
*****	11/05/2018	F	58821
*****	31/03/2019	F	465113
*****	26/04/2018	F	90825

En estos datos, los nombres han sido enmascarados

# Re-identificación: Ataque de asociación de registros

La re-identificación es el proceso que re-asocia una persona con un conjunto de datos.

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
*****	26/02/2018	M	73909	Resfrío
*****	08/05/2018	M	14890	Hepatitis
*****	11/05/2018	F	58821	Hepatitis
*****	31/03/2019	F	465113	VIH
*****	26/04/2018	F	90825	Bronquitis

# Re-identificación: Ataque de asociación de registros

La re-identificación es el proceso que re-asocia una persona con un conjunto de datos.

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
*****	26/02/2018	M	73909	Resfrío
*****	08/05/2018	M	14890	Hepatitis
*****	11/05/2018	F	58821	Hepatitis
*****	31/03/2019	F	465113	VIH
*****	26/04/2018	F	90825	Bronquitis



# Re-identificación: Ataque de asociación de registros

La re-identificación es el proceso que re-asocia una persona con un conjunto de datos.

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
*****	26/02/2018	M	73909	Resfrío
*****	08/05/2018	M	14890	Hepatitis
*****	11/05/2018	F	58821	Hepatitis
*****	31/03/2019	F	465113	VIH
*****	26/04/2018	F	90825	Bronquitis



Depende de datos auxiliares

Nombre	Fecha Nac.
Rashad Arnold	08/05/2018

# Re-identificación: Ataque de asociación de registros

La re-identificación es el proceso que re-asocia una persona con un conjunto de datos.

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
*****	26/02/2018	M	73909	Resfrío
*****	08/05/2018	M	14890	Hepatitis
*****	11/05/2018	F	58821	Hepatitis
*****	31/03/2019	F	465113	VIH
*****	26/04/2018	F	90825	Bronquitis



Depende de datos auxiliares

Nombre	Fecha Nac.
Rashad Arnold	08/05/2018



# Re-identificación: Ataque de asociación de registros

La re-identificación es el proceso que re-asocia una persona con un conjunto de datos.

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
*****	26/02/2018	M	73909	Resfrío
*****	08/05/2018	M	14890	Hepatitis
*****	11/05/2018	F	58821	Hepatitis
*****	31/03/2019	F	465113	VIH
*****	26/04/2018	F	90825	Bronquitis



Depende de datos auxiliares

Nombre	Fecha Nac.
Rashad Arnold	08/05/2018



Nombre	Fecha Nac.	Sexo	Coding Postal	Enfermedad
Rashad Arnold	08/05/2018	M	14890	Hepatitis

# Re-identificación: Ataque de asociación de registros

La re-identificación es el proceso que re-asocia una persona con un conjunto de datos.

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
*****	26/02/2018	M	73909	Resfrío
*****	08/05/2018	M	14890	Hepatitis
*****	11/05/2018	F	58821	Hepatitis
*****	31/03/2019	F	465113	VIH
*****	26/04/2018	F	90825	Bronquitis



Depende de datos auxiliares

Nombre	Fecha Nac.
Rashad Arnold	08/05/2018



Pudimos re-identificar a Rashad!

Nombre	Fecha Nac.	Sexo	Coding Postal	Enfermedad
Rashad Arnold	08/05/2018	M	14890	Hepatitis

# Re-identificación: Ataque de asociación de registros

La re-identificación es el proceso que re-asocia una persona con un conjunto de datos.

Nombre	Fecha Nac.	Sexo	Coding	Enfermedad
*****	26/02/2018	M	73909	Resfrío
*****	08/05/2018	M	14890	Hepatitis
*****	11/05/2018	F	58821	Hepatitis
*****	31/03/2019	F	465113	VIH
*****	26/04/2018	F	90825	Bronquitis



Depende de datos auxiliares

También llamado ataque de asociación de registros

Nombre	Fecha Nac.
Rashad Arnold	08/05/2018



Pudimos re-identificar a Rashad!

Nombre	Fecha Nac.	Sexo	Coding Postal	Enfermedad
Rashad Arnold	08/05/2018	M	14890	Hepatitis

# Re-identificación: Ataque de diferenciación

El problema puede ser peor cuando uno puede diseñar sus propias consultas.

```
SELECT sum(age) FROM person;
```

# Re-identificación: Ataque de diferenciación

El problema puede ser peor cuando uno puede diseñar sus propias consultas.

Una consulta de la “suma” en un grupo grande puede parecer adecuada:

```
SELECT sum(age) FROM person;
```

# Re-identificación: Ataque de diferenciación

El problema puede ser peor cuando uno puede diseñar sus propias consultas.

Una consulta de la “suma” en un grupo grande puede parecer adecuada:

```
SELECT sum(age) FROM person;
```

1256257

# Re-identificación: Ataque de diferenciación

El problema puede ser peor cuando uno puede diseñar sus propias consultas.

Una consulta de la “suma” en un grupo grande puede parecer adecuada:

```
SELECT sum(age) FROM person;
```

1256257

Podemos hacer otra consulta en un grupo grande

```
SELECT sum(age) FROM person  
WHERE nombre<>'Alan Brito';
```

# Re-identificación: Ataque de diferenciación

El problema puede ser peor cuando uno puede diseñar sus propias consultas.

Una consulta de la “suma” en un grupo grande puede parecer adecuada:

```
SELECT sum(age) FROM person;
```

1256257

Podemos hacer otra consulta en un grupo grande

```
SELECT sum(age) FROM person  
WHERE nombre<>'Alan Brito';
```

1256218

# Re-identificación: Ataque de diferenciación

El problema puede ser peor cuando uno puede diseñar sus propias consultas.

Una consulta de la “suma” en un grupo grande puede parecer adecuada:

```
SELECT sum(age) FROM person;
```

1256257

Podemos hacer otra consulta en un grupo grande

```
SELECT sum(age) FROM person  
WHERE nombre<>'Alan Brito';
```

1256218

$$1256257 - 1256218 = 39$$

# Re-identificación: Ataque de diferenciación

El problema puede ser peor cuando uno puede diseñar sus propias consultas.

Una consulta de la “suma” en un grupo grande puede parecer adecuada:

```
SELECT sum(age) FROM person;
```

1256257

Podemos hacer otra consulta en un grupo grande

```
SELECT sum(age) FROM person  
WHERE nombre<>'Alan Brito';
```

1256218

Descubrimos que Alan tiene 39 años!

$$1256257 - 1256218 = 39$$

# Resumen (I)

- Un ataque de asociación implica combinar datos auxiliares con datos no identificados para volver a identificar a las personas.
- En el caso más simple, un ataque de enlace se puede realizar a través de una combinación de dos tablas que contienen estos conjuntos de datos.

# Resumen (II)

- Los ataques de enlace simples son sorprendentemente efectivos:
  - Una solo ClId es suficiente para reducir las cosas a unos pocos registros.
  - Dos ClId a menudo son lo suficientemente buenos como para volver a identificar a una gran fracción de la población en un conjunto de datos en particular
  - Tres ClId (sexo, código postal y fecha de nacimiento) identifican de forma exclusiva al 87% de las personas en los EE.UU.

*k*-ANONIMATO

# ¿Qué es $k$ -Anonimato?

Sea  $T(A_1, \dots, A_n)$  una tabla y  $\mathcal{Q}$  los Cls asociados a ella.

Se dice que  $T$  satisface  **$k$ -anonimato** si y solo si cada grupo de Cls  $\mathcal{Q}_i \in \mathcal{Q}$  aparece al menos  $k$  veces en  $T$

# ¿Qué es $k$ -Anonimato?

Garantía formal

Sea  $T(A_1, \dots, A_n)$  una tabla y  $\mathcal{Q}$  los Cls asociados a ella.

Se dice que  $T$  satisface  **$k$ -anonimato** si y solo si cada grupo de Cls  $\mathcal{Q}_i \in \mathcal{Q}$  aparece al menos  $k$  veces en  $T$

# ¿Qué es $k$ -Anonimato?

Garantía formal

Sea  $T(A_1, \dots, A_n)$  una tabla y  $\mathcal{Q}$  los Cls asociados a ella.

Se dice que  $T$  satisface  **$k$ -anonimato** si y solo si cada grupo de Cls  $\mathcal{Q}_i \in \mathcal{Q}$  aparece al menos  $k$  veces en  $T$

Asegura que ningún individuo es únicamente identificable de un grupo de tamaño  $k$

# ¿Qué es $k$ -Anonimato?

Garantía formal

Sea  $T(A_1, \dots, A_n)$  una tabla y  $\mathcal{Q}$  los CIds asociados a ella.

Se dice que  $T$  satisface  **$k$ -anonimato** si y solo si cada grupo de CIds  $\mathcal{Q}_i \in \mathcal{Q}$  aparece al menos  $k$  veces en  $T$

Asegura que ningún individuo es únicamente identificable de un grupo de tamaño  $k$

```
SELECT CIds, count(*) FROM T  
GROUP BY CIds
```

$\geq k$

# ¿Qué es $k$ -Anonimato?

Garantía formal

Sea  $T(A_1, \dots, A_n)$  una tabla y  $\mathcal{Q}$  los CIds asociados a ella.

Se dice que  $T$  satisface  **$k$ -anonimato** si y solo si cada grupo de CIds  $\mathcal{Q}_i \in \mathcal{Q}$  aparece al menos  $k$  veces en  $T$

Asegura que ningún individuo es únicamente identificable de un grupo de tamaño  $k$

Aún requiere identificar a todos los cuasi-identificadores

```
SELECT CIds, count(*) FROM T
GROUP BY CIds
```

 $\geq k$

# $k$ -Anonimato: Ejemplo

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

# $k$ -Anonimato: Ejemplo

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

```
SELECT age, eduction_num, count(*) as c FROM person  
GROUP BY age, eduction_num ORDER BY c ASC;
```

# $k$ -Anonimato: Ejemplo

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

a	e	c
81	15	1
73	14	1
30	1	1
68	2	1
71	1	1
...	...	...
33	9	313
35	9	320
19	10	329
21	10	372
20	10	413



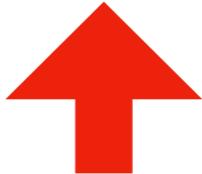
```
SELECT age, eduction_num, count(*) as c FROM person  
GROUP BY age, eduction_num ORDER BY c ASC;
```

# $k$ -Anonimato: Ejemplo

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

a	e	c
81	15	1
73	14	1
30	1	1
68	2	1
71	1	1
...	...	...
33	9	313
35	9	320
19	10	329
21	10	372
20	10	413

1-anonima!



```
SELECT age, eduction_num, count(*) as c FROM person  
GROUP BY age, eduction_num ORDER BY c ASC;
```

# $k$ -Anonimato: Generalización

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

# $k$ -Anonimato: Generalización

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

```
SELECT (age/10)*10 as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY (age/10)*10, (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Generalización

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

a	en	c
90	0	1
80	0	2
70	0	7
70	10	9
10	10	12
60	10	13
...	...	...
30	0	122
30	10	148
20	10	157
40	10	160



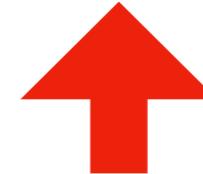
```
SELECT (age/10)*10 as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY (age/10)*10, (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Generalización

Mejoró pero...

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

a	en	c
90	0	1
80	0	2
70	0	7
70	10	9
10	10	12
60	10	13
...	...	...
30	0	122
30	10	148
20	10	157
40	10	160



```
SELECT (age/10)*10 as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY (age/10)*10, (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Generalización

Mejoró pero...

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

a	en	c
90	0	1
80	0	2
70	0	7
70	10	9
10	10	12
60	10	13
...	...	...
30	0	122
30	10	148
20	10	157
40	10	160

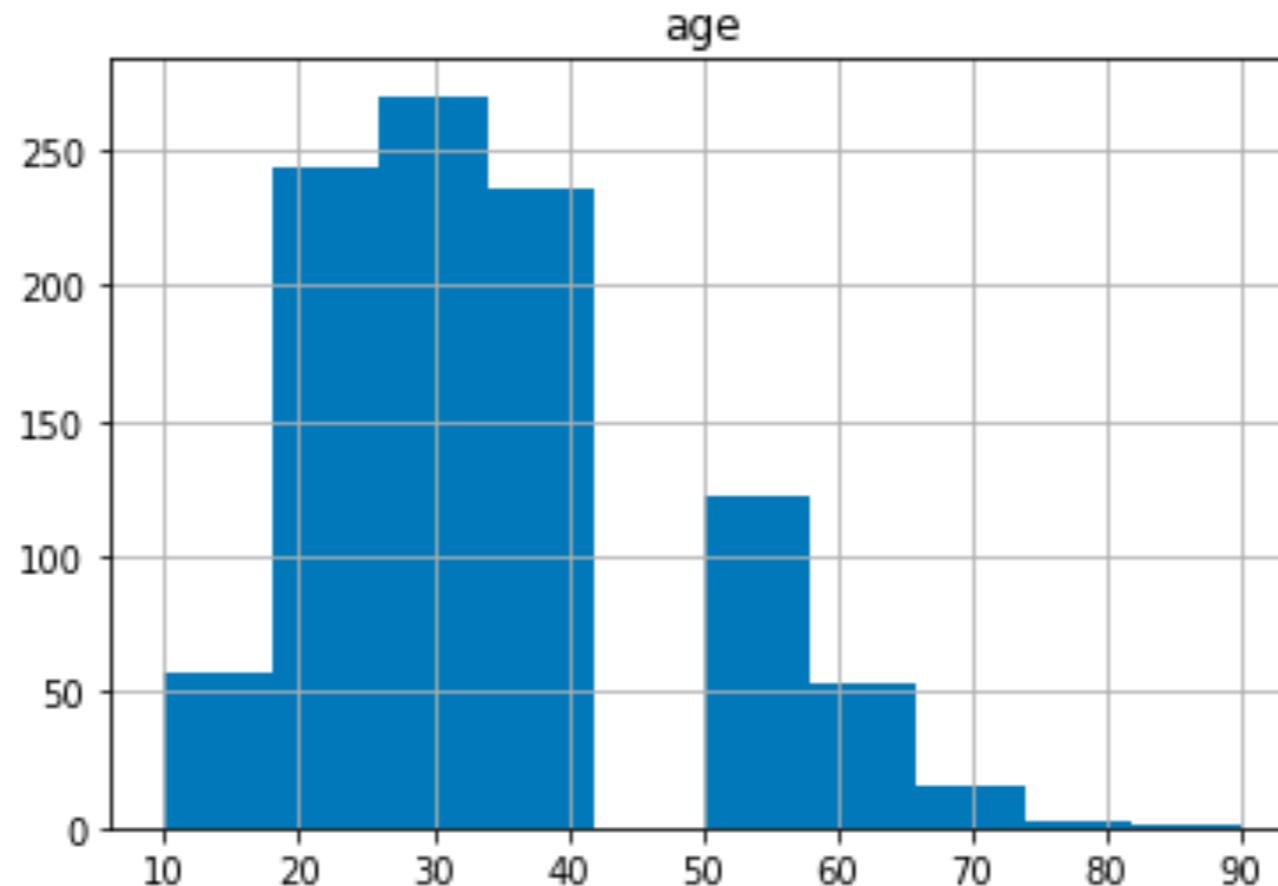
Sigue siendo  
1-anonima!



```
SELECT (age/10)*10 as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY (age/10)*10, (eduction_num/10)*10  
ORDER BY c ASC;
```

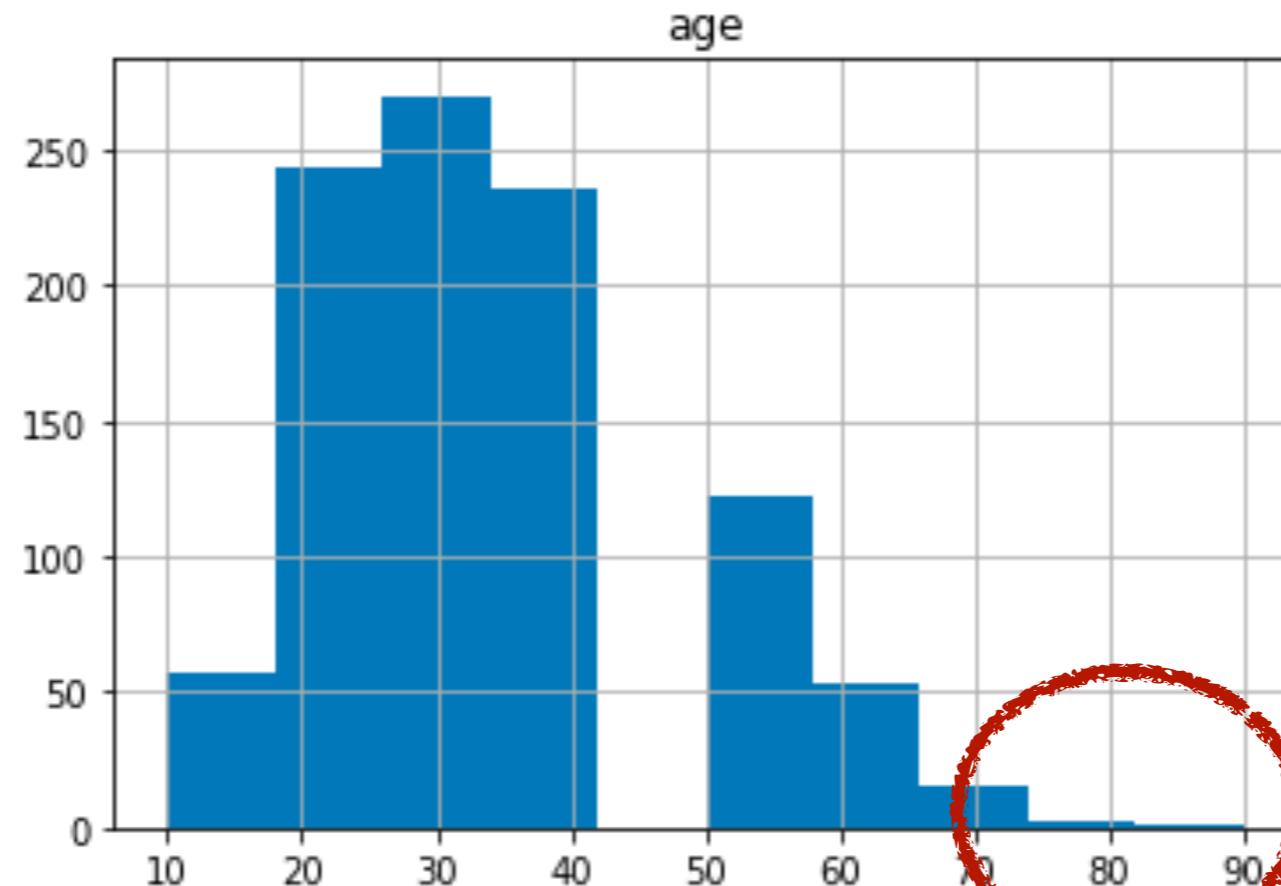
# $k$ -Anonimato: Valores atípicos

a	en	c
90	0	1
80	0	2
70	0	7
70	10	9
10	10	12
60	10	13
...	...	...
30	0	122
30	10	148
20	10	157
40	10	160



# $k$ -Anonimato: Valores atípicos

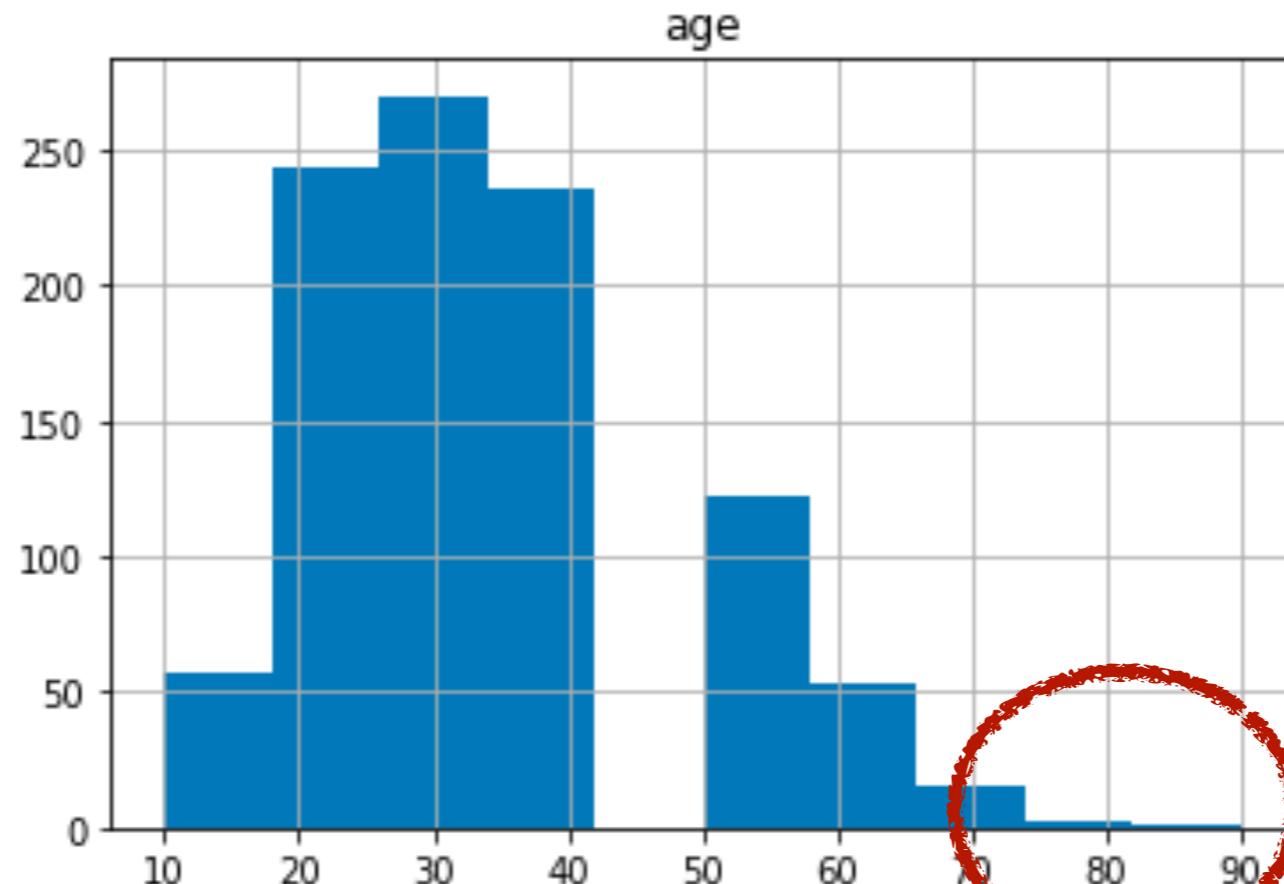
a	en	c
90	0	1
80	0	2
70	0	7
70	10	9
10	10	12
60	10	13
...	...	...
30	0	122
30	10	148
20	10	157
40	10	160



El problema son los  
valores atípicos (outliers)

# $k$ -Anonimato: Valores atípicos

a	en	c
90	0	1
80	0	2
70	0	7
70	10	9
10	10	12
60	10	13
...	...	...
30	0	122
30	10	148
20	10	157
40	10	160



Idea: eliminar outliers

El problema son los valores atípicos (outliers)

# $k$ -Anonimato: Eliminando outliers

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

# $k$ -Anonimato: Eliminando outliers

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

```
SELECT LEAST( (age/10)*10, 60) as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY LEAST( (age/10)*10, 60), (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Eliminando outliers

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

Truncamos valores  
mayores que 60 a 60

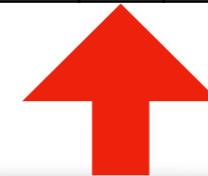
```
SELECT LEAST( (age/10)*10, 60) as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY LEAST( (age/10)*10, 60), (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Eliminando outliers

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

Truncamos valores  
mayores que 60 a 60

a	e	c
10	10	12
60	10	22
10	0	45
60	0	50
50	10	53
50	0	69
40	0	76
20	0	86
30	0	122
30	10	148
20	10	157
40	10	160



```
SELECT LEAST( (age/10)*10, 60) as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY LEAST( (age/10)*10, 60), (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Eliminando outliers

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

Truncamos valores  
mayores que 60 a 60

a	e	c
10	10	12
60	10	22
10	0	45
60	0	50
50	10	53
50	0	69
40	0	76
20	0	86
30	0	122
30	10	148
20	10	157
40	10	160

Logramos 12-anonimato!

```
SELECT LEAST( (age/10)*10, 60) as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY LEAST( (age/10)*10, 60), (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Agregando mas datos

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

# $k$ -Anonimato: Agregando mas datos

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

```
SELECT LEAST( (age/10)*10, 60) as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY LEAST( (age/10)*10, 60), (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Agregando mas datos

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

a	e	c
10	10	455
60	10	1172
10	0	1202
60	0	1472
50	0	2172
50	10	2246
40	0	2795
20	0	3380
30	0	3733
40	10	4380
20	10	4674
30	10	4880



```
SELECT LEAST( (age/10)*10, 60) as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY LEAST( (age/10)*10, 60), (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Agregando mas datos

name	ssn	age	eduction_num	...	target
Karrie Trusslove	732-14-61	39	13	...	<=50K
Brandise Tripony	150-19-27	50	13	...	<=50K
Brenn McNeely	725-59-98	38	9	...	<=50K
Dorry Poter	659-57-49	53	7	...	<=50K
Dick Honnan	220-93-38	28	13	...	<=50K
...	...	...	...	...	...
Ardyce Golby	212-61-83	27	12	...	<=50K
Jean O'Connor	737-32-29	40	9	...	>50K
Reuben Skrzynski	314-48-02	58	9	...	<=50K
Caye Biddle	647-75-35	22	9	...	<=50K
Hortense Hardesty	690-42-56	52	9	...	>50K

a	e	c
10	10	455
60	10	1172
10	0	1202
60	0	1472
50	0	2172
50	10	2246
40	0	2795
20	0	3380
30	0	3733
40	10	4380
20	10	4674
30	10	4880

Logramos 455-anonimato!



```
SELECT LEAST( (age/10)*10, 60) as a, (eduction_num/10)*10 as e,  
count(*) as c  
FROM person  
GROUP BY LEAST( (age/10)*10, 60), (eduction_num/10)*10  
ORDER BY c ASC;
```

# $k$ -Anonimato: Otro ejemplo

zip	age	nationality	disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cáncer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cáncer
13053	37	Indian	Cáncer
13068	36	Japanese	Cáncer
13068	32	American	Cáncer

# $k$ -Anonimato: Otro ejemplo

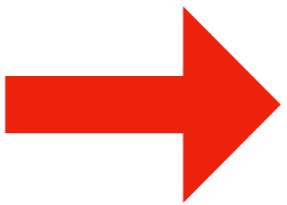
zip	age	nationality	disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cáncer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cáncer
13053	37	Indian	Cáncer
13068	36	Japanese	Cáncer
13068	32	American	Cáncer

1-anónimo

# $k$ -Anonimato: Otro ejemplo

zip	age	nationality	disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cáncer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cáncer
13053	37	Indian	Cáncer
13068	36	Japanese	Cáncer
13068	32	American	Cáncer

1-anónimo

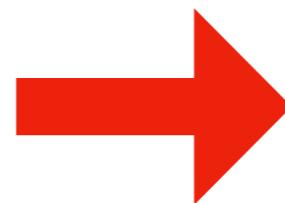


Generalizamos

# $k$ -Anonimato: Otro ejemplo

zip	age	nationality	disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cáncer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cáncer
13053	37	Indian	Cáncer
13068	36	Japanese	Cáncer
13068	32	American	Cáncer

1-anónimo



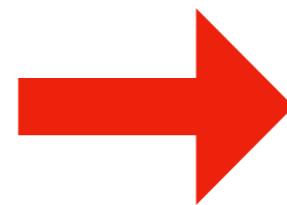
Generalizamos

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# $k$ -Anonimato: Otro ejemplo

zip	age	nationality	disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cáncer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cáncer
13053	37	Indian	Cáncer
13068	36	Japanese	Cáncer
13068	32	American	Cáncer

1-anónimo



Generalizamos

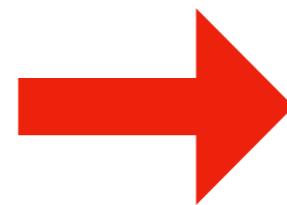
zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

4-anónimo

# $k$ -Anonimato: Otro ejemplo

zip	age	nationality	disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cáncer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cáncer
13053	37	Indian	Cáncer
13068	36	Japanese	Cáncer
13068	32	American	Cáncer

1-anónimo



Generalizamos

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

4-anónimo

Hay algún problema?

# Ataque de $k$ -Anonimato #1: Homogeneidad

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $k$ -Anonimato #1: Homogeneidad

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $k$ -Anonimato #1: Homogeneidad

name	zip	age	nationality
Bob	13053	35	??



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $k$ -Anonimato #1: Homogeneidad

name	zip	age	nationality
Bob	13053	35	??



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

Aprendemos que Bob tiene cáncer

# Ataque de $k$ -Anonimato #2: Data auxiliar

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $k$ -Anonimato #2: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $k$ -Anonimato #2: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

Los japoneses tiene una muy baja  
incidencia de enfermedades al corazón

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $k$ -Anonimato #2: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

Los japoneses tiene una muy baja incidencia de enfermedades al corazón

# Ataque de $k$ -Anonimato #2: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

Los japoneses tiene una muy baja incidencia de enfermedades al corazón

Aprendemos que Umeko está resfriada

# $\ell$ -DIVERSIDAD

# ¿Qué es $\ell$ -Diversidad?

Adicionalmente a lo que requiere  $k$ -anonimato:  
Una tabla  $T$  es  $\ell$ -diversa si para cada grupo de filas con los mismos Clids, por cada atributo sensible  $S$ , existen al menos  $\ell$  valores distintos.

# ¿Qué es $\ell$ -Diversidad?

Adicionalmente a lo que requiere  $k$ -anonimato:  
Una tabla  $T$  es  $\ell$ -diversa si para cada grupo de filas con los mismos Clids, por cada atributo sensible  $S$ , existen al menos  $\ell$  valores distintos.

↑  
Previene ataque #1  
(homogeneidad)

# ¿Qué es $\ell$ -Diversidad?

Adicionalmente a lo que requiere  $k$ -anonimato:  
Una tabla  $T$  es  $\ell$ -diversa si para cada grupo de filas con los mismos CIds, por cada atributo sensible  $S$ , existen al menos  $\ell$  valores distintos.

↑  
Previene ataque #1  
(homogeneidad)

```
SELECT CIds, count(DISTINCT S)  
FROM T GROUP BY CIds
```

 $\geq \ell$

# ¿Qué es $\ell$ -Diversidad?

Adicionalmente a lo que requiere  $k$ -anonimato:  
Una tabla  $T$  es  $\ell$ -diversa si para cada grupo de filas con los mismos CIds, por cada atributo sensible  $S$ , existen al menos  $\ell$  valores distintos.

↑  
Previene ataque #1  
(homogeneidad)

↑  
Incrementa la resistencia frente  
al ataque #2 (data auxiliar)

```
SELECT CIds, count(DISTINCT S)  
FROM T GROUP BY CIds
```

 $\geq \ell$

# ¿Qué es $\ell$ -Diversidad?

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# ¿Qué es $\ell$ -Diversidad?

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

4-anónima, pero  
1-diversa

# ¿Qué es $\ell$ -Diversidad?

zip	age	nationality	disease
130**	<=40	*	Heart
130**	<=40	*	Heart
130**	<=40	*	Flu
130**	<=40	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	<=40	*	Cáncer
130**	<=40	*	Cáncer
130**	<=40	*	Cáncer
130**	<=40	*	Cáncer

# ¿Qué es $\ell$ -Diversidad?

zip	age	nationality	disease
130**	<=40	*	Heart
130**	<=40	*	Heart
130**	<=40	*	Flu
130**	<=40	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	<=40	*	Cáncer
130**	<=40	*	Cáncer
130**	<=40	*	Cáncer
130**	<=40	*	Cáncer

Ahora logramos  
3-diversad!

# Ataque de $\ell$ -Diversidad: Data auxiliar

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $\ell$ -Diversidad: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $\ell$ -Diversidad: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

Umeko no tiene cáncer

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $\ell$ -Diversidad: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

Umeko no tiene cáncer

Umeko no tiene una  
enfermedad al corazón

zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $\ell$ -Diversidad: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

Umeko no tiene cáncer

Umeko no tiene una  
enfermedad al corazón



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $\ell$ -Diversidad: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

Umeko no tiene cáncer

Umeko no tiene una  
enfermedad al corazón



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

Umeko puede tener diabetes o estar  
resfriada

# Ataque de $\ell$ -Diversidad: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

Umeko no tiene cáncer

Umeko no tiene una  
enfermedad al corazón



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $\ell$ -Diversidad: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

Umeko no tiene cáncer

Umeko no tiene una  
enfermedad al corazón

Umeko no tiene diabetes



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

# Ataque de $\ell$ -Diversidad: Data auxiliar

name	zip	age	nationality
Umeko	13068	24	Japan

Umeko no tiene cáncer

Umeko no tiene una  
enfermedad al corazón

Umeko no tiene diabetes



zip	age	nationality	disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cáncer
130**	<30	*	Flu
1485*	>40	*	Cáncer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer
130**	30-40	*	Cáncer

Aprendemos que Umeko está resfriada

# Resumen

- Enfoques formales y sistemáticos de de-identificación
- Una mejora sustentable con respecto a enfoques ad-hoc
- Aún así susceptible a ataques
  - La protección de privacidad depende de la **información auxiliar del adversario/atacante.**
- Automatización tiene un **alto costo computacional:**
  - Dada una tabla  $T$ , encontrar una tabla  $T'$  que satisfaga  $k$ -anonimidad y maximize la utilidad
  - NP-Complejo (Meyerson & Williams, 2004)

# OPERACIONES DE ANONIMIZACIÓN

# Operaciones de Anonimización

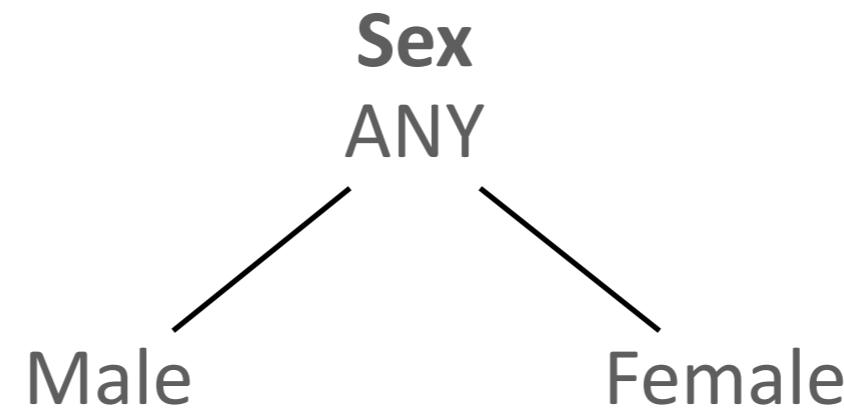
- Generalización
- Supresión
- Anatomización
- Permutación
- Perturbación aleatoria

# Generalización

Esconder algunos detalles de los cuasi-identificadores (CIds).

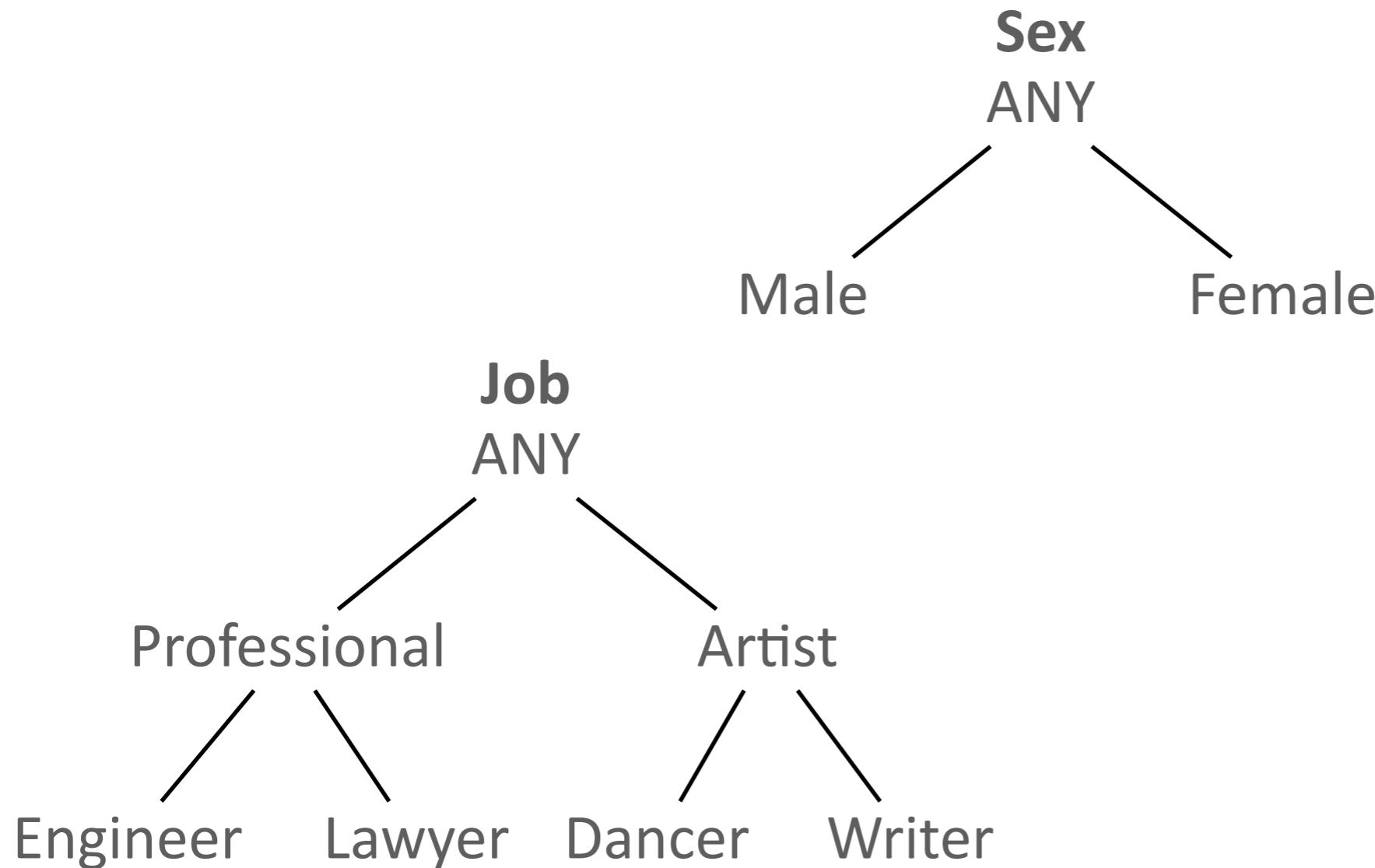
# Generalización

Esconder algunos detalles de los cuasi-identificadores (CIDs).



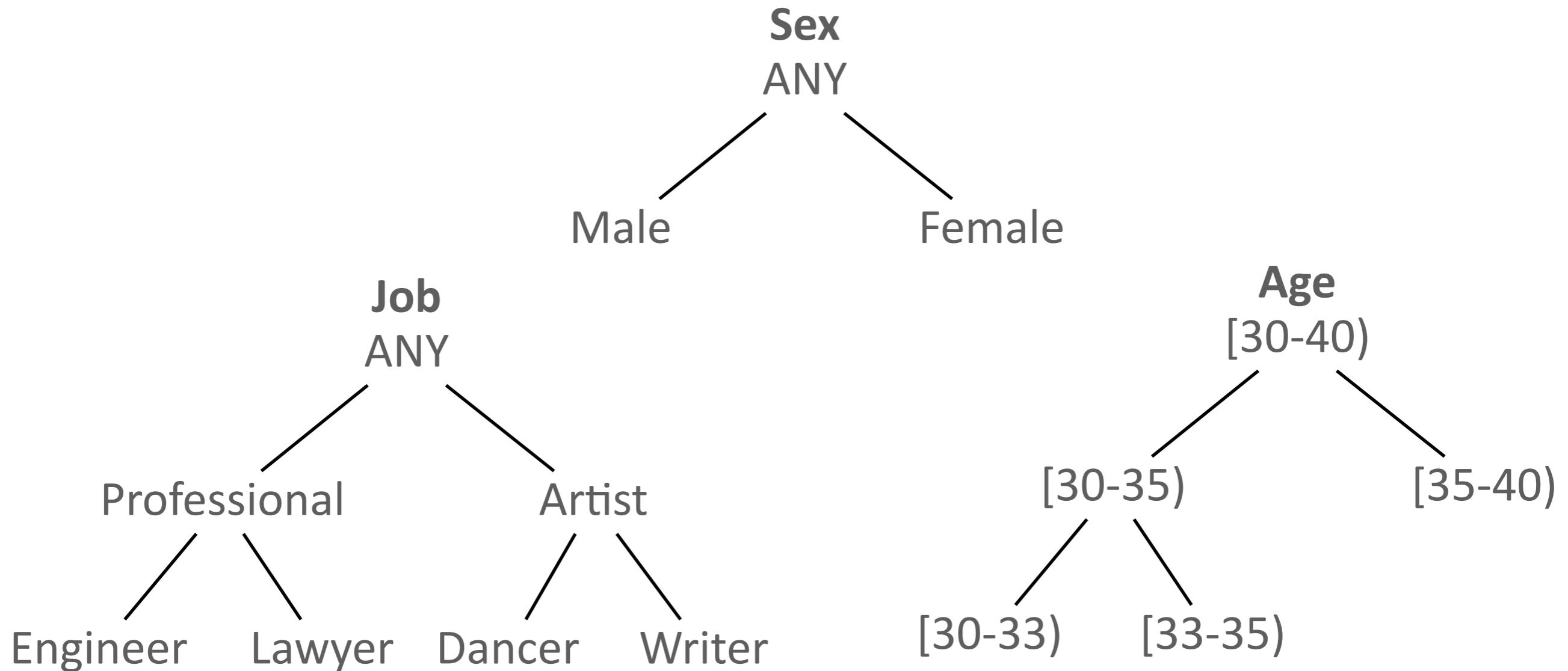
# Generalización

Esconder algunos detalles de los cuasi-identificadores (CIDs).



# Generalización

Esconder algunos detalles de los cuasi-identificadores (CIDs).

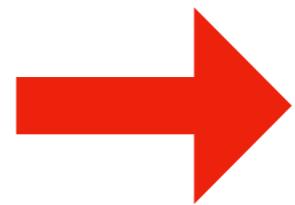


# Generalización

Name	Job	Sex	Age
Alice	Writer	Female	30
Cathy	Engineer	Female	35
Bob	Writer	Male	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Dancer	Male	38
Gladys	Engineer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

# Generalización

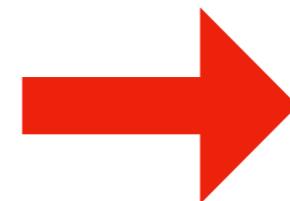
Name	Job	Sex	Age
Alice	Writer	Female	30
Cathy	Engineer	Female	35
Bob	Writer	Male	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Dancer	Male	38
Gladys	Engineer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32



Generalizamos

# Generalización

Name	Job	Sex	Age
Alice	Writer	Female	30
Cathy	Engineer	Female	35
Bob	Writer	Male	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Dancer	Male	38
Gladys	Engineer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

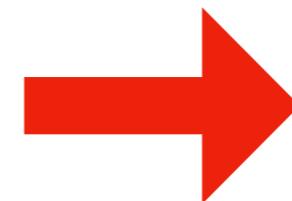


Generalizamos

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Cathy	Professional	Female	[35-40)
Bob	Artist	Male	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Artist	Male	[35-40)
Gladys	Professional	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

# Generalización

Name	Job	Sex	Age
Alice	Writer	Female	30
Cathy	Engineer	Female	35
Bob	Writer	Male	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Dancer	Male	38
Gladys	Engineer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

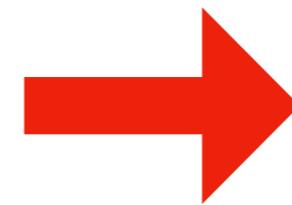


Generalizamos

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Cathy	Professional	Female	[35-40)
Bob	Artist	Male	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Artist	Male	[35-40)
Gladys	Professional	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

# Generalización

Name	Job	Sex	Age
Alice	Writer	Female	30
Cathy	Engineer	Female	35
Bob	Writer	Male	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Dancer	Male	38
Gladys	Engineer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32



Generalizamos

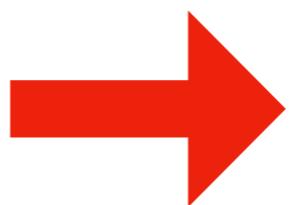
Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Cathy	Professional	Female	[35-40)
Bob	Artist	Male	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Artist	Male	[35-40)
Gladys	Professional	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

# Supresión

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Cathy	Professional	Female	[35-40)
Bob	Artist	Male	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Artist	Male	[35-40)
Gladys	Professional	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

# Supresión

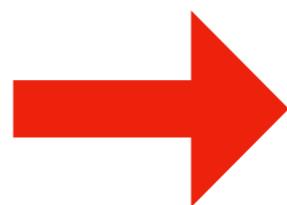
Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Cathy	Professional	Female	[35-40)
Bob	Artist	Male	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Artist	Male	[35-40)
Gladys	Professional	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)



Suprimimos

# Supresión

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Cathy	Professional	Female	[35-40)
Bob	Artist	Male	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Artist	Male	[35-40)
Gladys	Professional	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

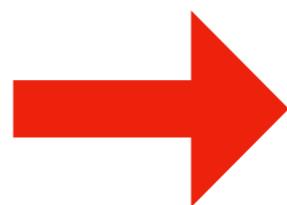


Suprimimos

Name	Job	Sex	Age
****	Artist	Female	[30-35)
****	Professional	Female	[35-40)
****	Artist	Male	[30-35)
****	Professional	Male	[35-40)
****	Artist	Female	[30-35)
****	Artist	Male	[35-40)
****	Professional	Female	[30-35)
****	Professional	Male	[35-40)
****	Artist	Female	[30-35)

# Supresión

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Cathy	Professional	Female	[35-40)
Bob	Artist	Male	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Artist	Male	[35-40)
Gladys	Professional	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)



Suprimimos

Name	Job	Sex	Age
****	Artist	Female	[30-35)
****	Professional	Female	[35-40)
****	Artist	Male	[30-35)
****	Professional	Male	[35-40)
****	Artist	Female	[30-35)
****	Artist	Male	[35-40)
****	Professional	Female	[30-35)
****	Professional	Male	[35-40)
****	Artist	Female	[30-35)

# Anatomización

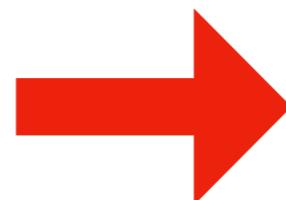
Idea: agrupar los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart

# Anatomización

Idea: agrupar los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart



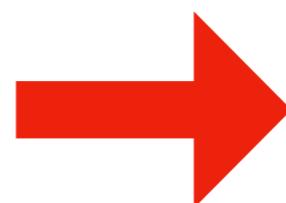
Generalizamos

Hasta que al menos  $\frac{1}{\ell}$  de  
los registros tengan la  
misma enfermedad

# Anatomización

Idea: agrupar los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart



Generalizamos

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

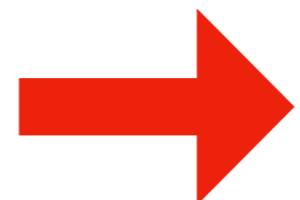
1  
Hasta que al menos  $\frac{1}{\ell}$  de  
los registros tengan la  
misma enfermedad

# Anatomización

Idea: agrupar los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart

Generalizamos



Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

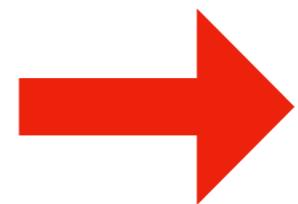
1  
Hasta que al menos  $\frac{1}{\ell}$  de  
los registros tengan la  
misma enfermedad

# Anatomización

Idea: agrupar los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart

Generalizamos



Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

1  
Hasta que al menos  $\frac{1}{\ell}$  de  
los registros tengan la  
misma enfermedad

# Anatomización

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

# Anatomización

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

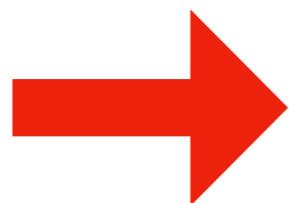
Grupo 1

Grupo 2

# Anatomización

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

Grupo 1



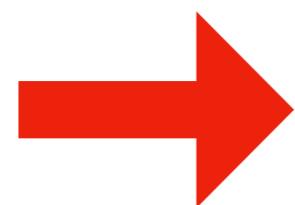
Generamos

Grupo 2

# Anatomización

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

Grupo 1



Generamos

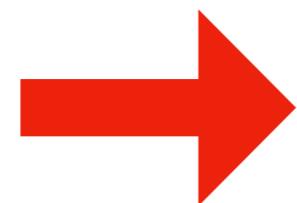
Age	Sex	Grupoid
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

Grupo 2

# Anatomización

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

Grupo 1



Generamos

¿Cuántas personas de 38 años tienen enfermedades del corazón? (2)

Grupo 2

$$2 \times \frac{1}{5} = 0.4$$

$$3 \times \frac{2}{4} = 1.5$$

Age	Sex	Grupoid
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

Grupoid	Disease	Count
1	Hepatitis	3
1	HIV	3
2	COVID	2
2	Heart	2

# Permutación

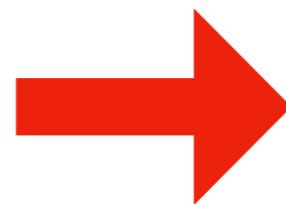
Idea: revolver los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart

# Permutación

Idea: revolver los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart

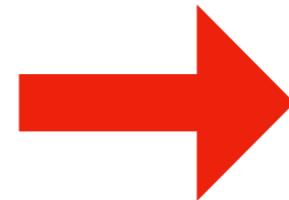


Generalizamos

# Permutación

Idea: revolver los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart



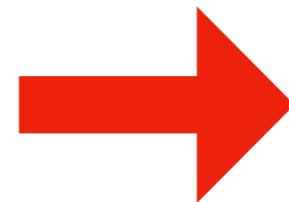
Generalizamos

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

# Permutación

Idea: revolver los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart



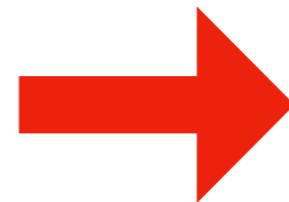
Generalizamos

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

# Permutación

Idea: revolver los valores sensibles.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart



Generalizamos

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

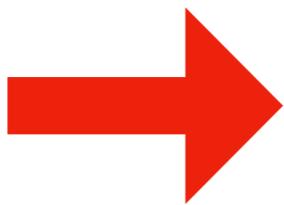
# Permutación

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart

Age	Sex	Disease
30	Male	HIV
30	Male	HIV
30	Male	Hepatitis
32	Male	HIV
32	Male	Hepatitis
32	Male	Hepatitis
36	Female	Heart
38	Female	COVID
38	Female	COVID
38	Female	Heart

# Permutación

Age	Sex	Disease
[30-35)	Male	Hepatitis
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	Hepatitis
[30-35)	Male	HIV
[30-35)	Male	HIV
[35-40)	Female	COVID
[35-40)	Female	COVID
[35-40)	Female	Heart
[35-40)	Female	Heart



Revolvemos

Age	Sex	Disease
30	Male	HIV
30	Male	HIV
30	Male	Hepatitis
32	Male	HIV
32	Male	Hepatitis
32	Male	Hepatitis
36	Female	Heart
38	Female	COVID
38	Female	COVID
38	Female	Heart

# Perturbación aleatoria

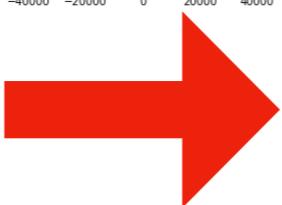
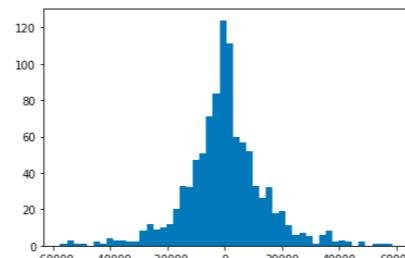
Idea: generar data sintética a partir de información estadística.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart

# Perturbación aleatoria

Idea: generar data sintética a partir de información estadística.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart

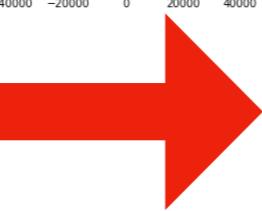
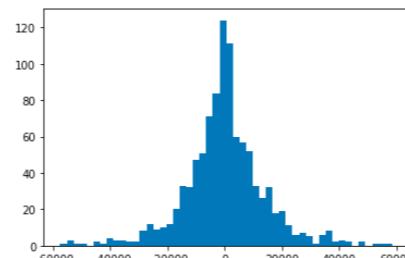


Generamos

# Perturbación aleatoria

Idea: generar data sintética a partir de información estadística.

Age	Sex	Disease
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	COVID
38	Female	COVID
38	Female	Heart
38	Female	Heart



Generamos

Age	Sex	Disease
31	Male	HIV
29	Male	HIV
31	Male	Hepatitis
33	Male	HIV
35	Male	Hepatitis
31	Male	Hepatitis
36	Female	Heart
37	Female	COVID
40	Female	COVID
38	Female	Heart

# PRIVACIDAD DIFERENCIAL

# ¿Qué es privacidad diferencial?

# ¿Qué es privacidad diferencial?

$$f: \text{DB} \rightarrow R$$

# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f: \text{Stack of coins} \rightarrow R$$

# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{Stack of coins} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{Stack of coins} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

$x$

# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{Stack of coins} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

$x$

$y$

# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{Stack of coins} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

$x \xleftarrow{\text{at “distance” } \leq 1} y$

# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{BD} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

Idénticas BD excepto  
por 1 individuo

$x$  ← at “distance”  $\leq 1$  →  $y$

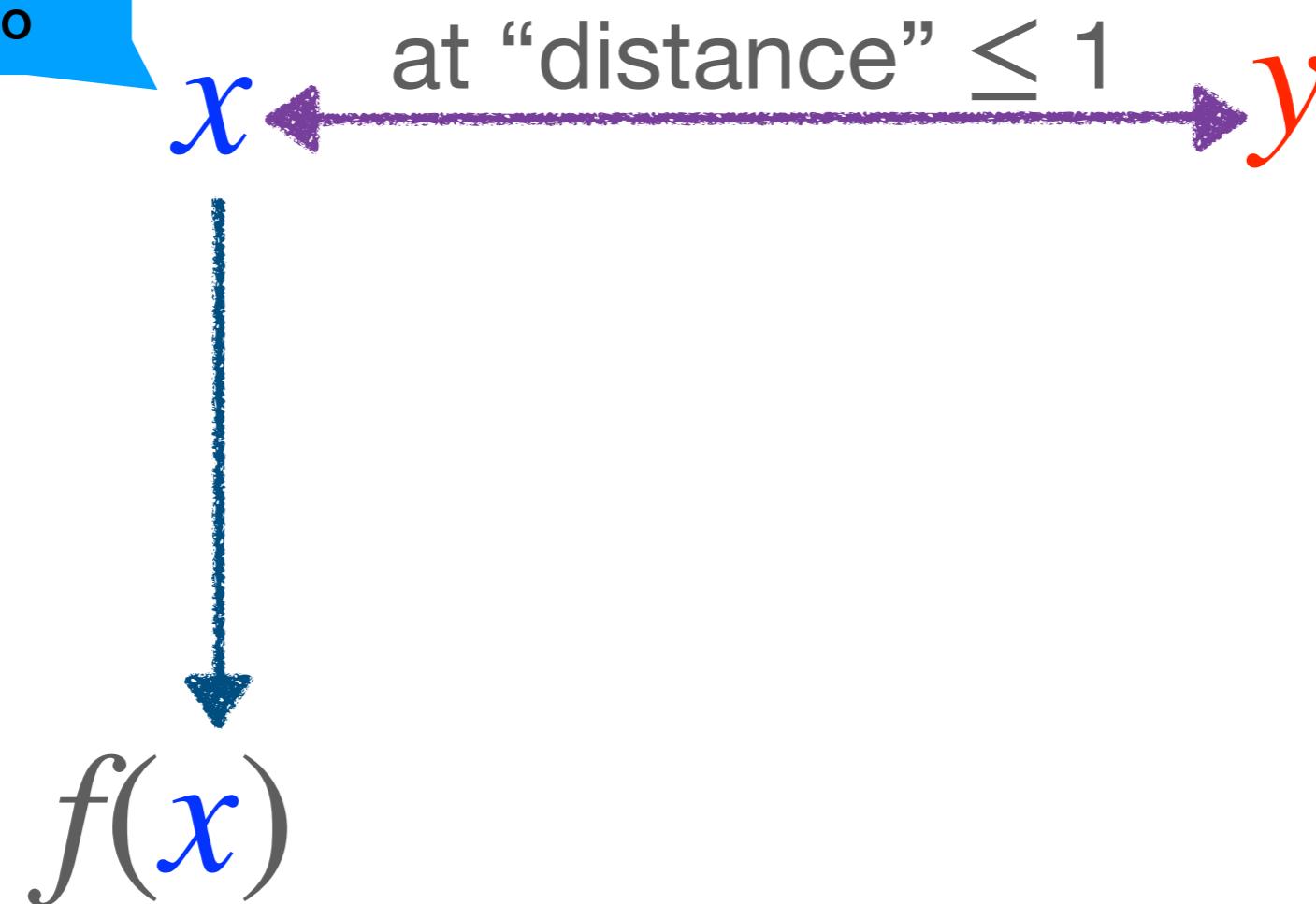
# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{BD} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

Idénticas BD excepto  
por 1 individuo



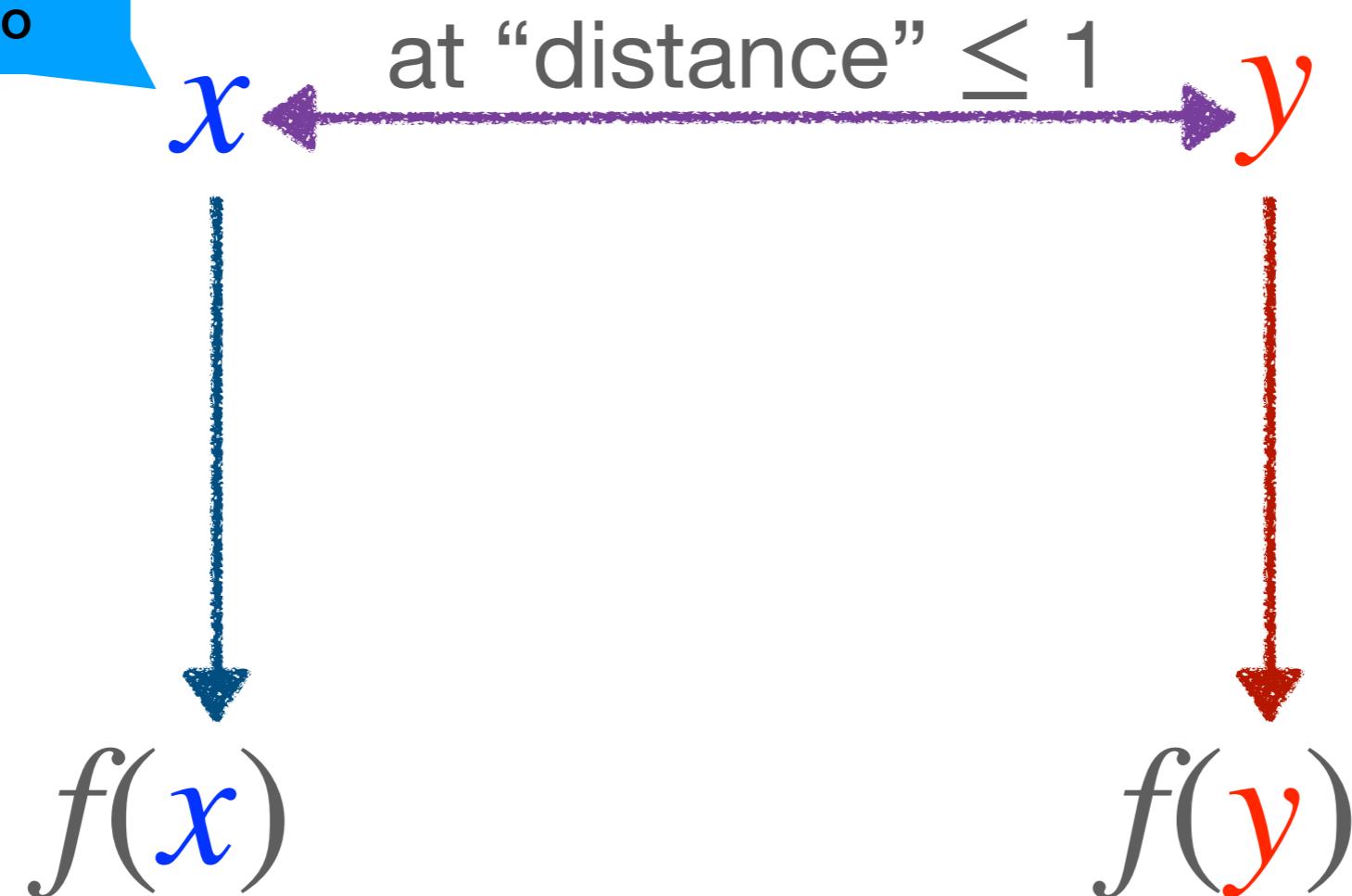
# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{BD} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

Idénticas BD excepto  
por 1 individuo



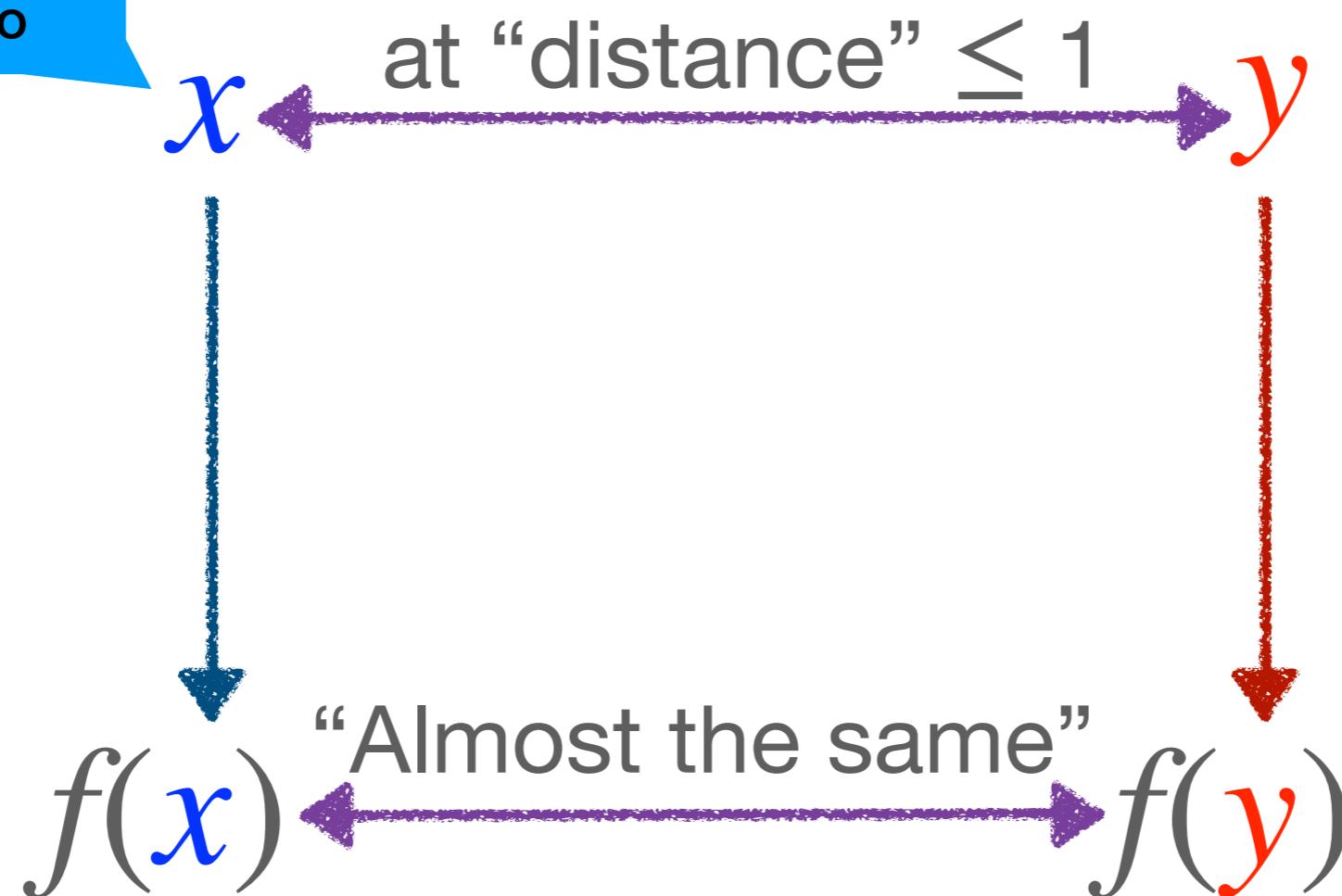
# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{BD} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

Idénticas BD excepto  
por 1 individuo



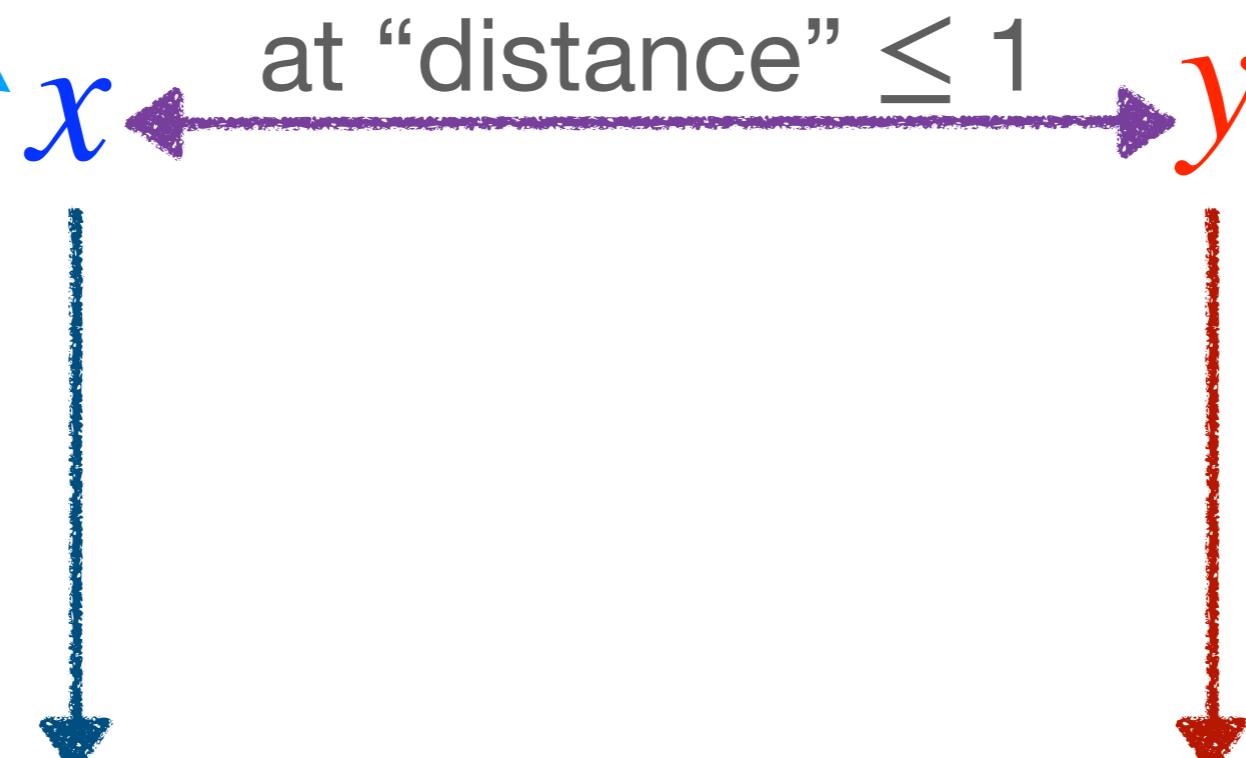
# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{BD} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

Idénticas BD excepto  
por 1 individuo



$$\Pr[f(\textcolor{blue}{x}) \in S] \leq e^\epsilon \Pr[f(\textcolor{red}{y}) \in S]$$

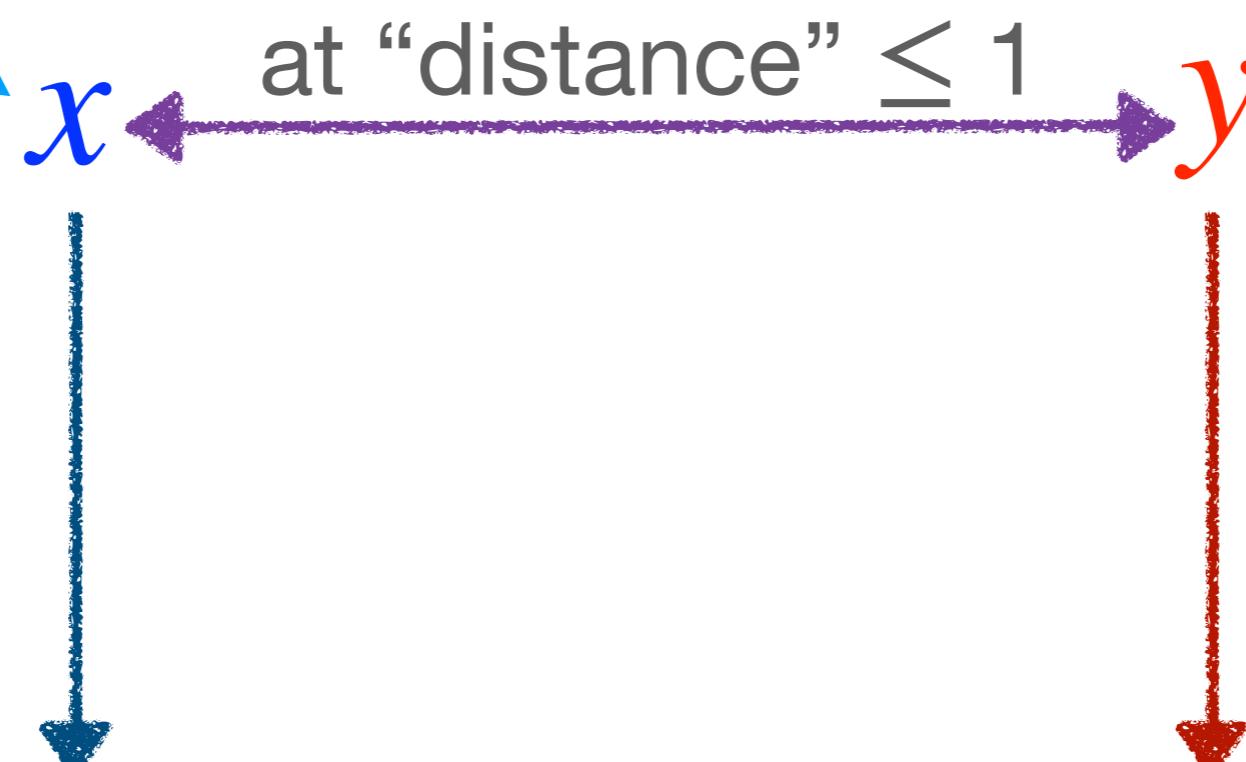
# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{BD} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

Idénticas BD excepto  
por 1 individuo



$$\Pr[f(\textcolor{blue}{x}) \in S] \leq e^\epsilon \Pr[f(\textcolor{red}{y}) \in S]$$

$$S \in R$$

# ¿Qué es privacidad diferencial?

Algoritmo aleatorio

$$f : \text{BD} \rightarrow R$$

Decimos que  $f$  es  
diferencialmente privada  
si....

Idénticas BD excepto  
por 1 individuo

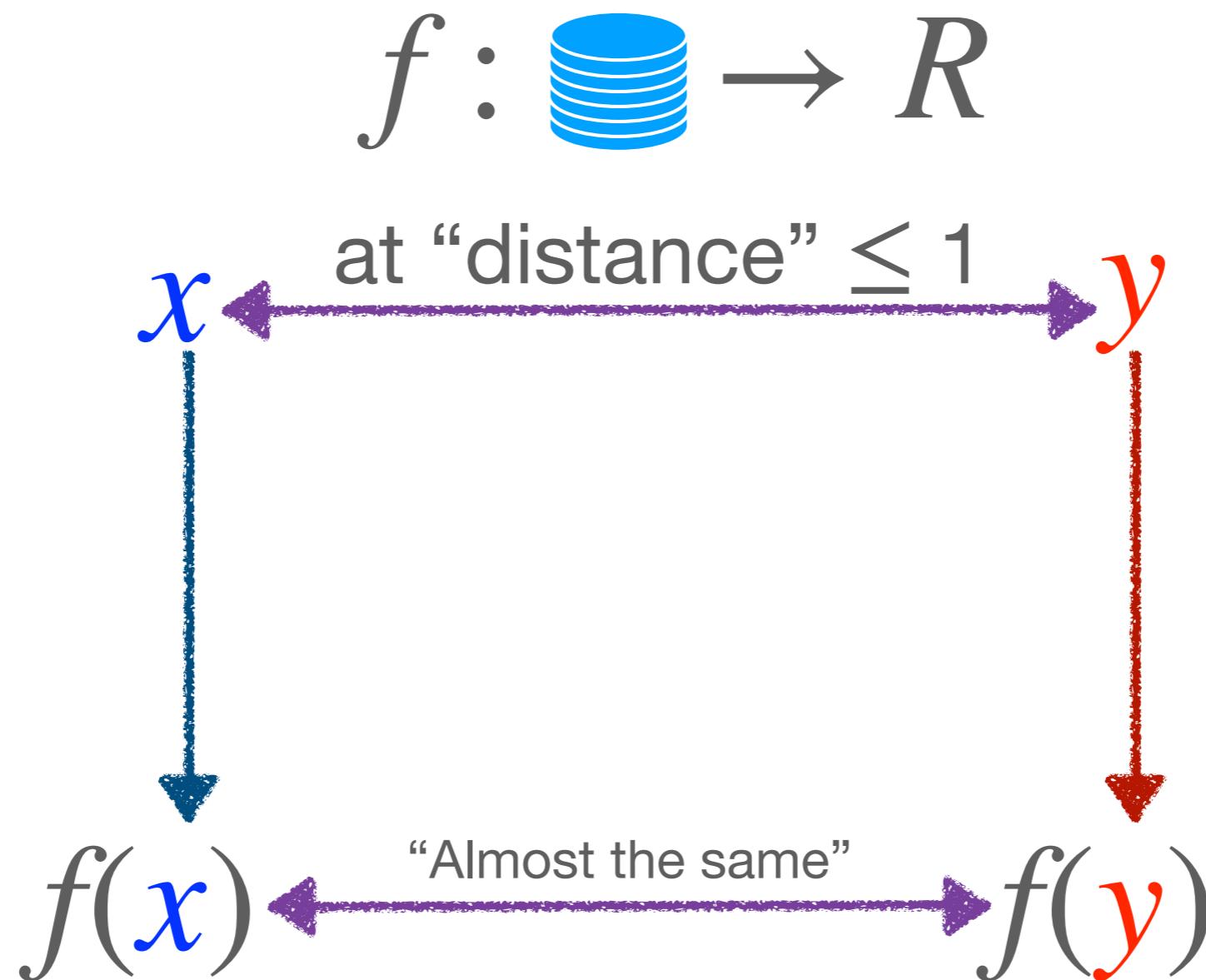
$x$  ←→  $y$   
at “distance”  $\leq 1$

Decimos que  $f$  es  
 $\epsilon$ -diferencialmente  
privada

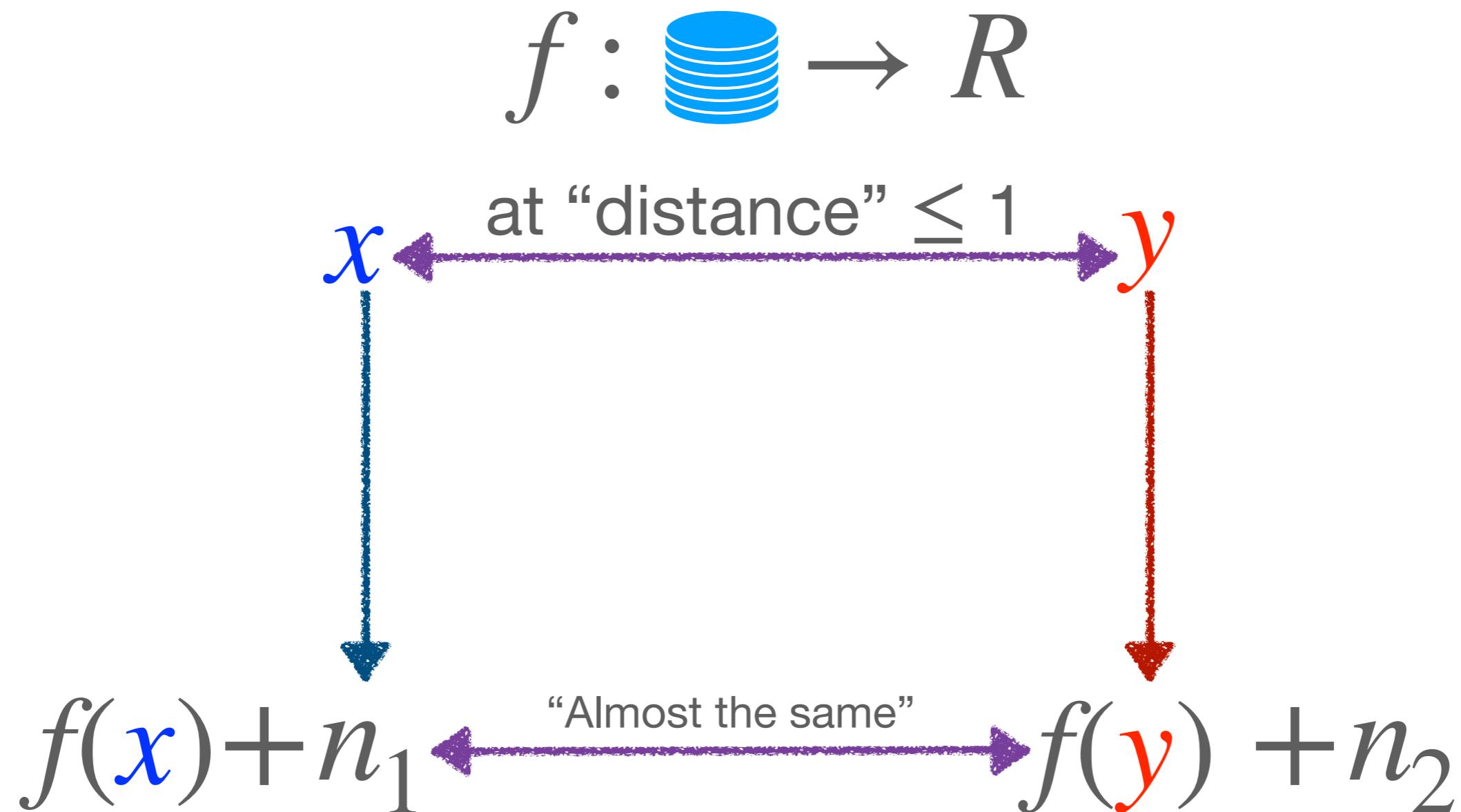
$$\Pr[f(x) \in S] \leq e^\epsilon \Pr[f(y) \in S]$$

$$S \in R$$

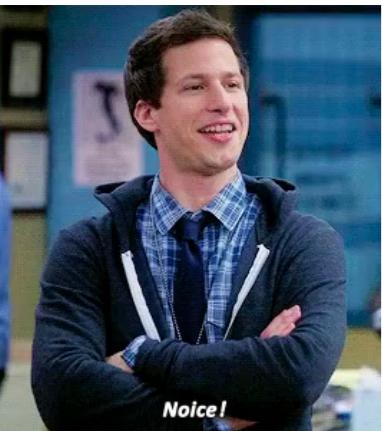
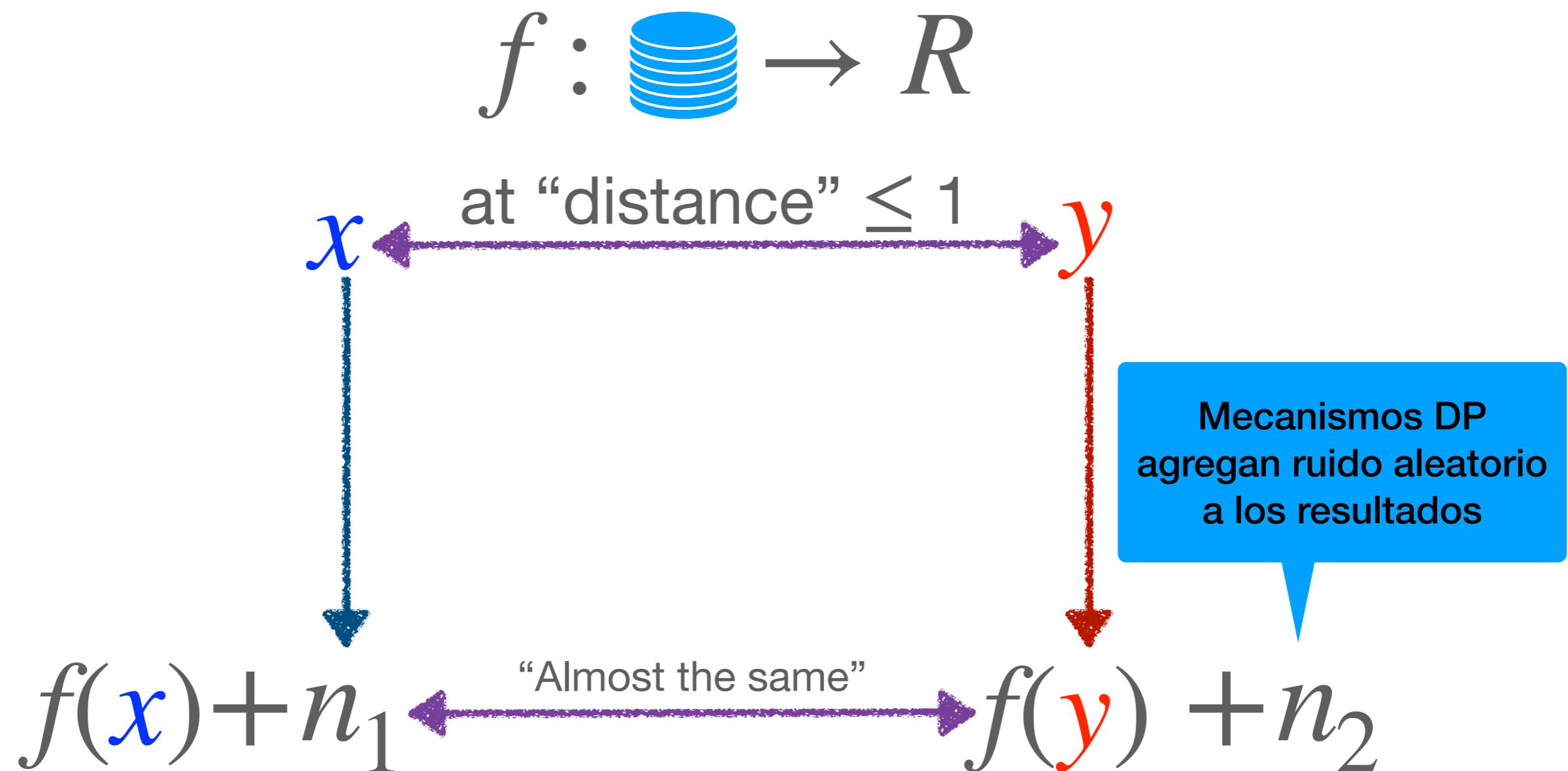
# ¿Qué es privacidad diferencial?



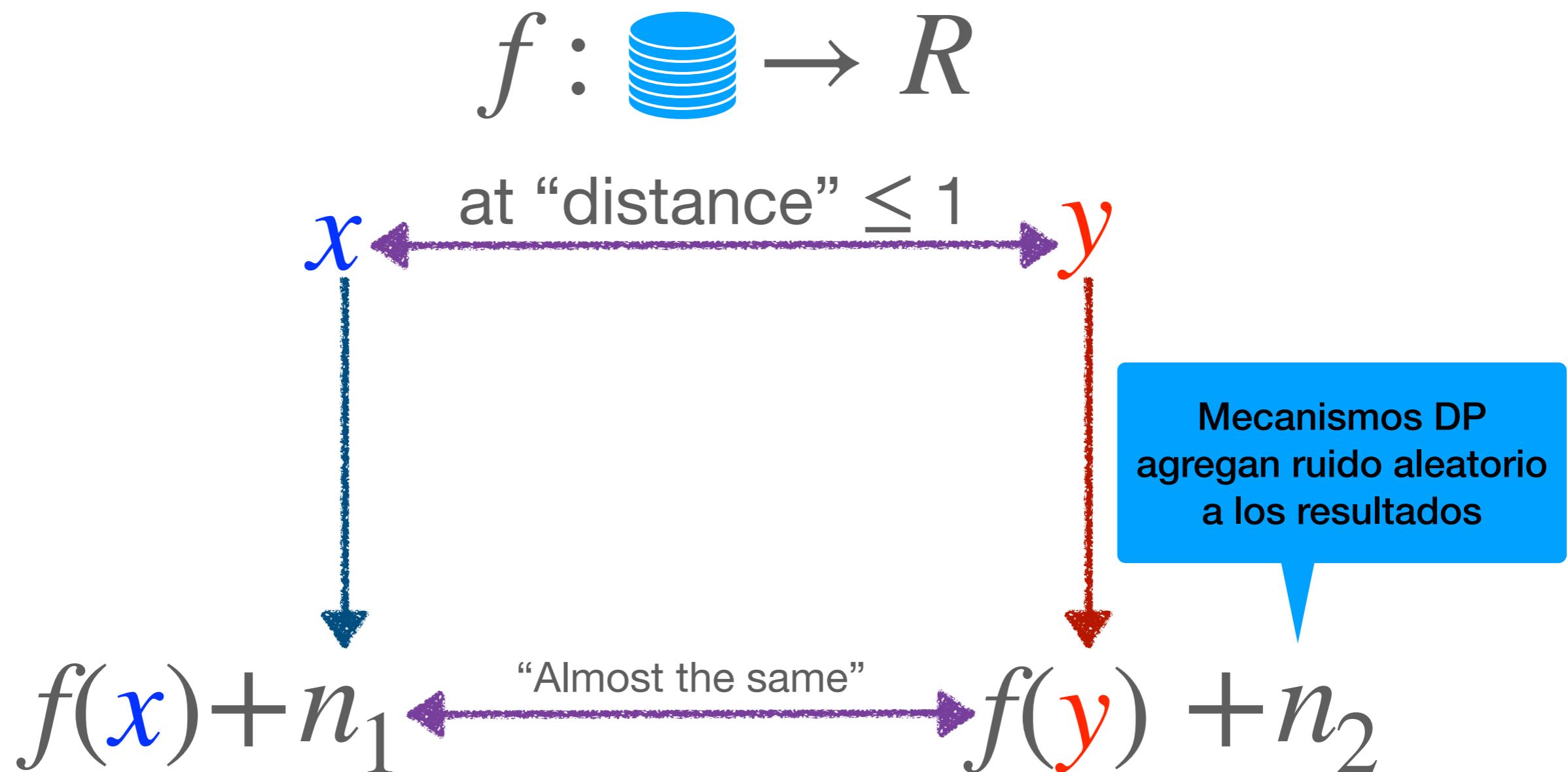
# ¿Qué es privacidad diferencial?



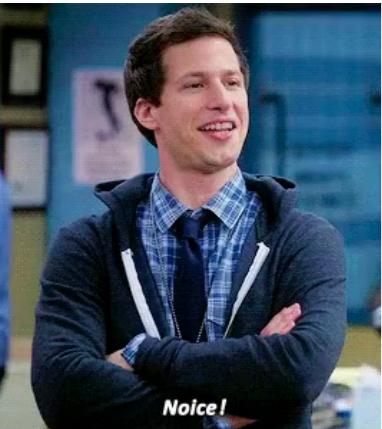
# ¿Qué es privacidad diferencial?



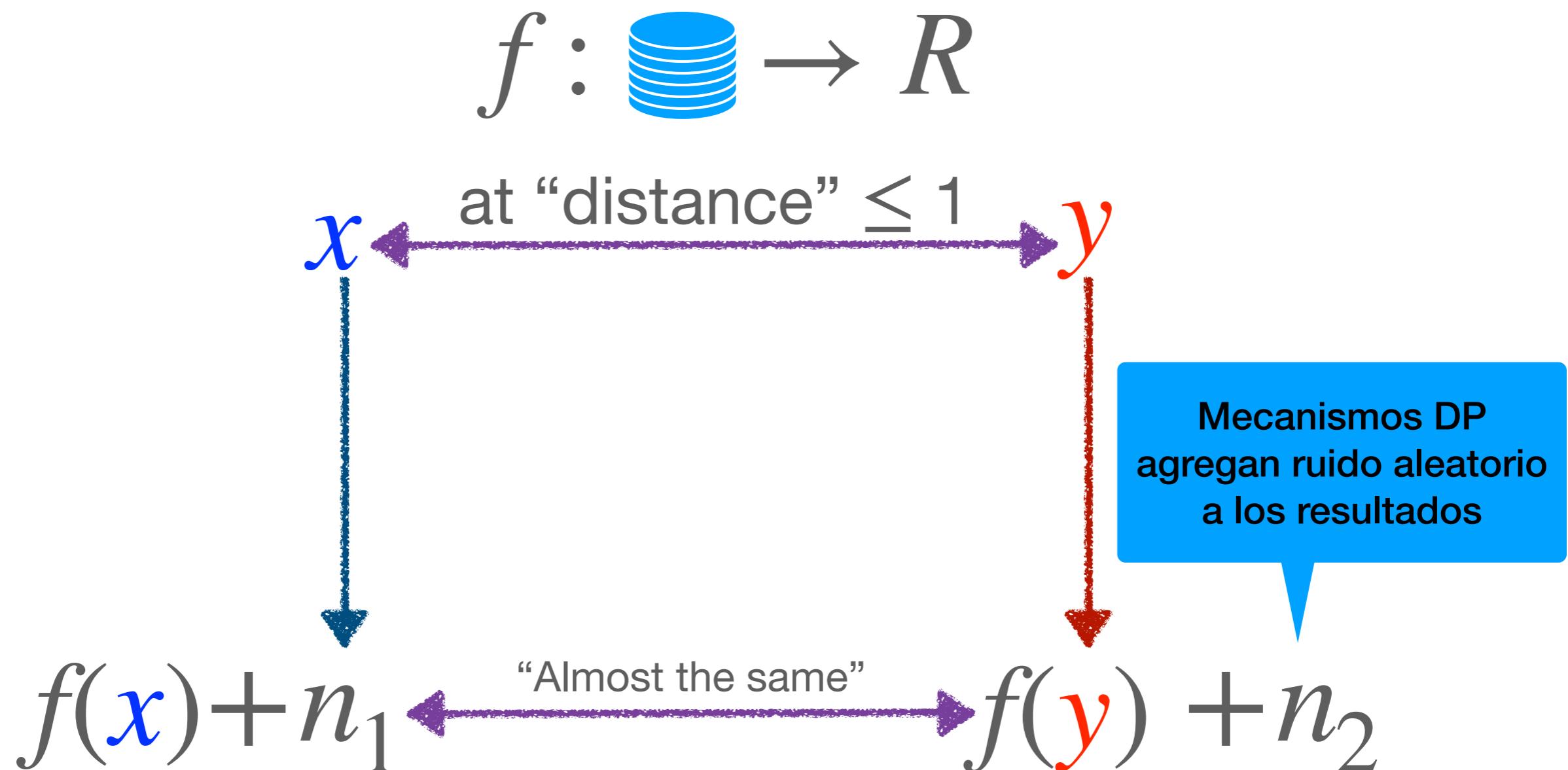
# ¿Qué es privacidad diferencial?



Pero cuánto ruido?

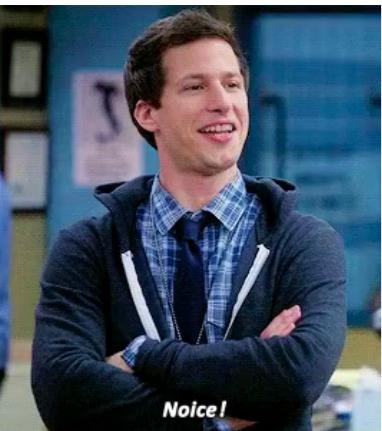


# ¿Qué es privacidad diferencial?



Pero cuánto ruido?

Depende en la sensibilidad de la consulta



# Privacidad diferencia: Ejemplo

¿Cuántas personas en el conjunto de datos tienen 40 años o más?

# Privacidad diferencia: Ejemplo

¿Cuántas personas en el conjunto de datos tienen 40 años o más?

```
SELECT count(*) FROM person WHERE age>=40
```

# Privacidad diferencia: Ejemplo

¿Cuántas personas en el conjunto de datos tienen 40 años o más?

```
SELECT count(*) FROM person WHERE age>=40
```

14237

# Privacidad diferencia: Ejemplo

¿Cuántas personas en el conjunto de datos tienen 40 años o más?

```
SELECT count(*) FROM person WHERE age>=40
```

14237

¿Cuál es la sensibilidad de esta consulta?

# Privacidad diferencia: Ejemplo

¿Cuántas personas en el conjunto de datos tienen 40 años o más?

```
SELECT count(*) FROM person WHERE age>=40
```

14237

¿Cuál es la sensibilidad de esta consulta?

¿En cuánto puede variar (máximo) el resultado si saco o agrego un individuo?

# Privacidad diferencia: Ejemplo

¿Cuántas personas en el conjunto de datos tienen 40 años o más?

```
SELECT count(*) FROM person WHERE age>=40
```

14237

¿Cuál es la sensibilidad de esta consulta?

¿En cuánto puede variar (máximo) el resultado si saco o agrego un individuo?

En 1!

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

## El mecanismo de Laplace

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

El mecanismo de Laplace

Sensibilidad

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

El mecanismo de Laplace

Sensibilidad

1

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

## El mecanismo de Laplace

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

Sensibilidad

1

Presupuesto  
de privacidad

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

## El mecanismo de Laplace

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

Sensibilidad

1

Presupuesto  
de privacidad

0.1

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

El mecanismo de Laplace

Sensibilidad

1

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

Presupuesto  
de privacidad

0.1

$$F(x) = 14211.230613$$

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

El mecanismo de Laplace

Sensibilidad

1

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

Presupuesto  
de privacidad

0.1

$$F(x) = 14211.230613$$

$$F(x) = 14265.917434$$

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

El mecanismo de Laplace

Sensibilidad

1

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

Presupuesto  
de privacidad

0.1

$$F(x) = 14211.230613$$

$$F(x) = 14265.770017$$

$$F(x) = 14265.917434$$

# Privacidad diferencia: Ejemplo

```
SELECT count(*) FROM person WHERE age>=40
```

14237

El mecanismo de Laplace

Sensibilidad

1

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$$

Presupuesto  
de privacidad

0.1

$$F(x) = 14211.230613$$

$$F(x) = 14265.770017$$

$$F(x) = 14265.917434$$

$$F(x) = 14209.093826$$

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa



# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

 Ruido

-0.688716

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

 Ruido

18.475044

-0.688716

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

 Ruido

18.475044

15.158466

-0.688716

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

 Ruido

18.475044

15.158466

-0.688716

-0.203654

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

 Ruido

18.475044

15.158466

-0.688716

-2.115105

-0.203654

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

 Ruido

18.475044

15.158466

-0.688716

-2.115105

-0.203654

13.167499

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

+ Ruido

Es realmente útil la consulta?

18.475044

15.158466

-0.688716

-2.115105

-0.203654

13.167499

# ¿Cuánto ruido es suficiente?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

Consulta maliciosa

+ Ruido

Es realmente útil la consulta?

18.475044

15.158466

-0.688716

-2.115105

-0.203654

13.167499

No, pero no importa porque la consulta era maliciosa (muy selectiva)

# ¿Qué pasa si repetimos la consulta muchas veces?

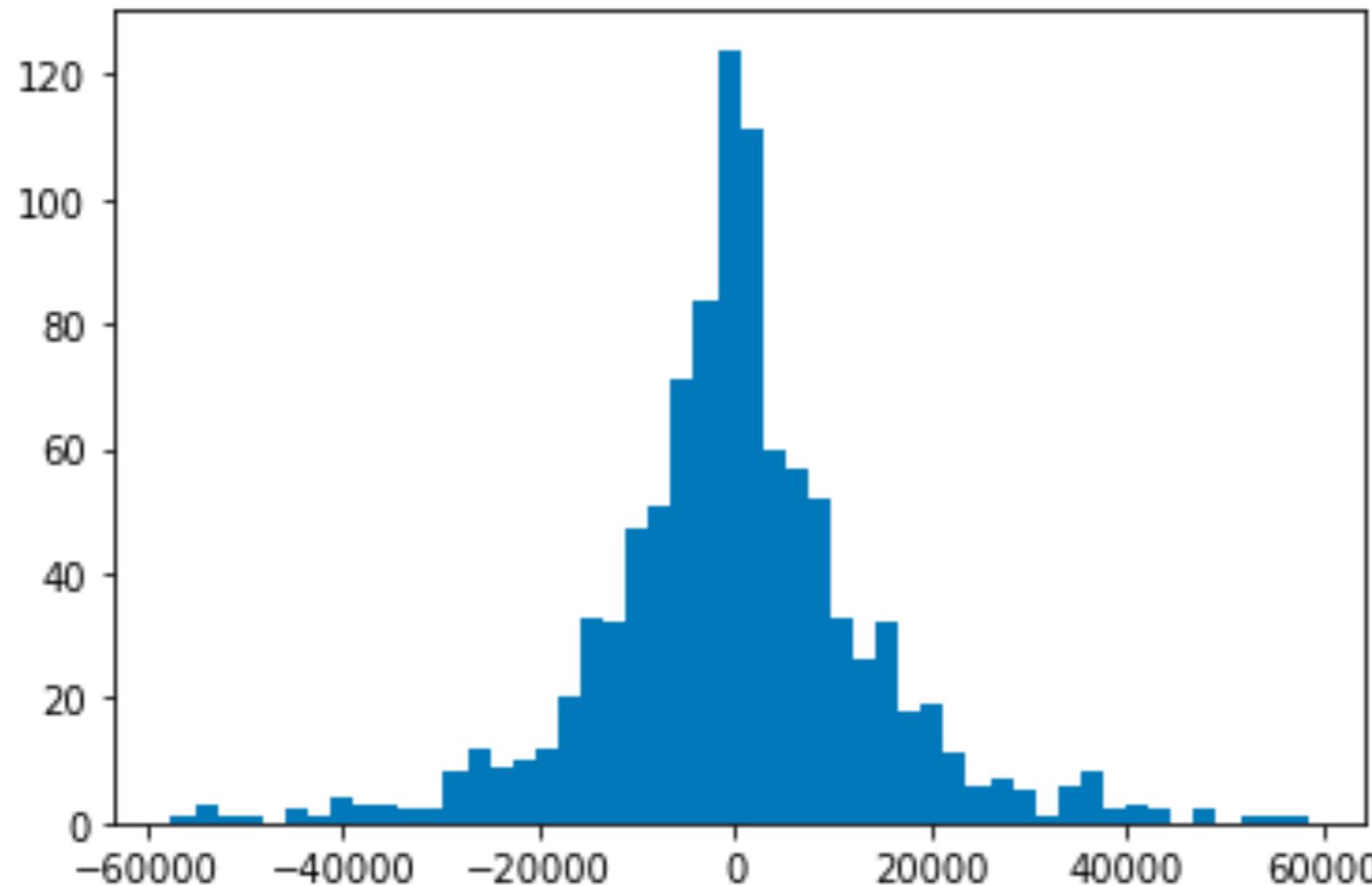
```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

# ¿Qué pasa si repetimos la consulta muchas veces?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

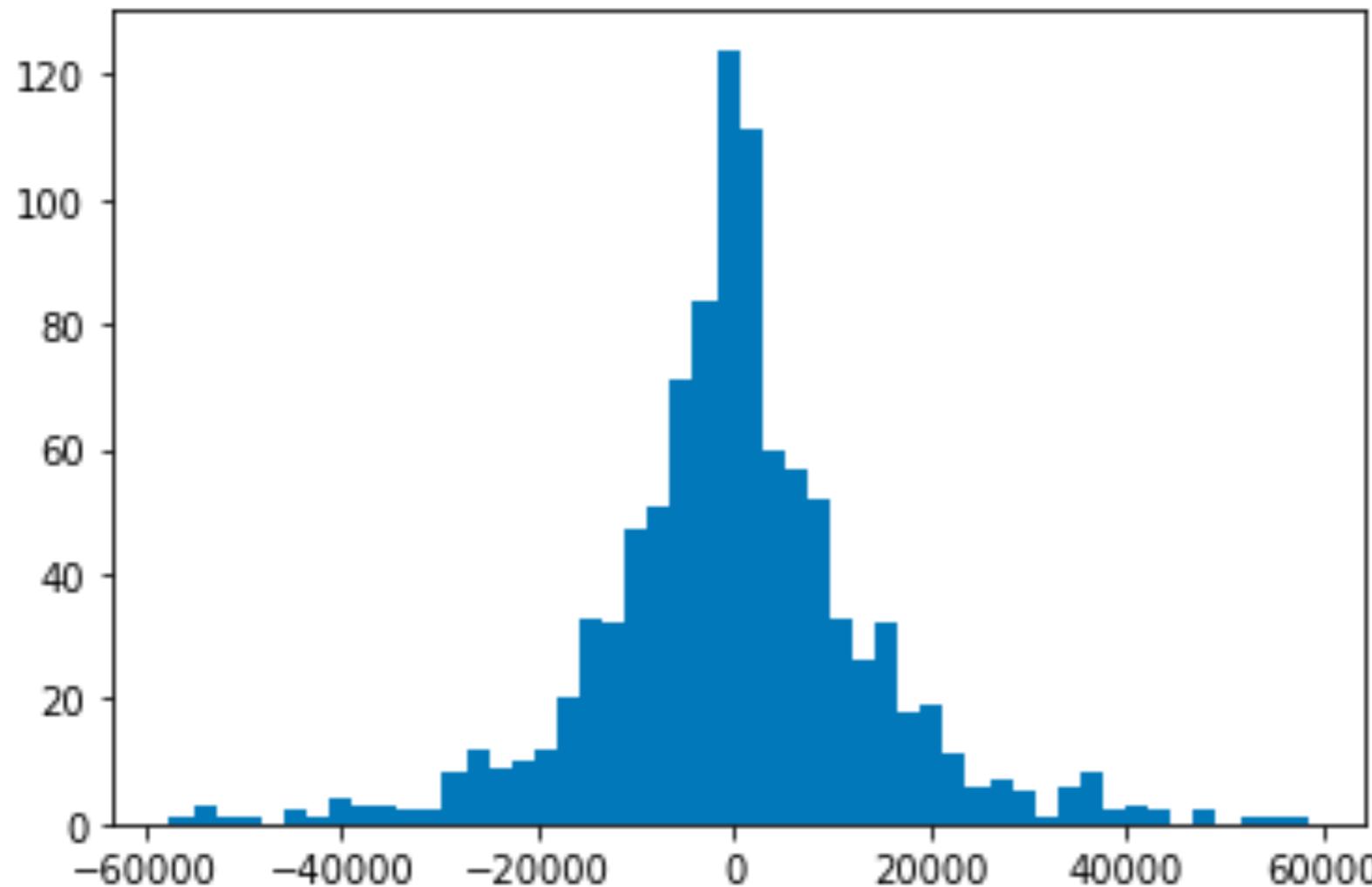
1



# ¿Qué pasa si repetimos la consulta muchas veces?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1

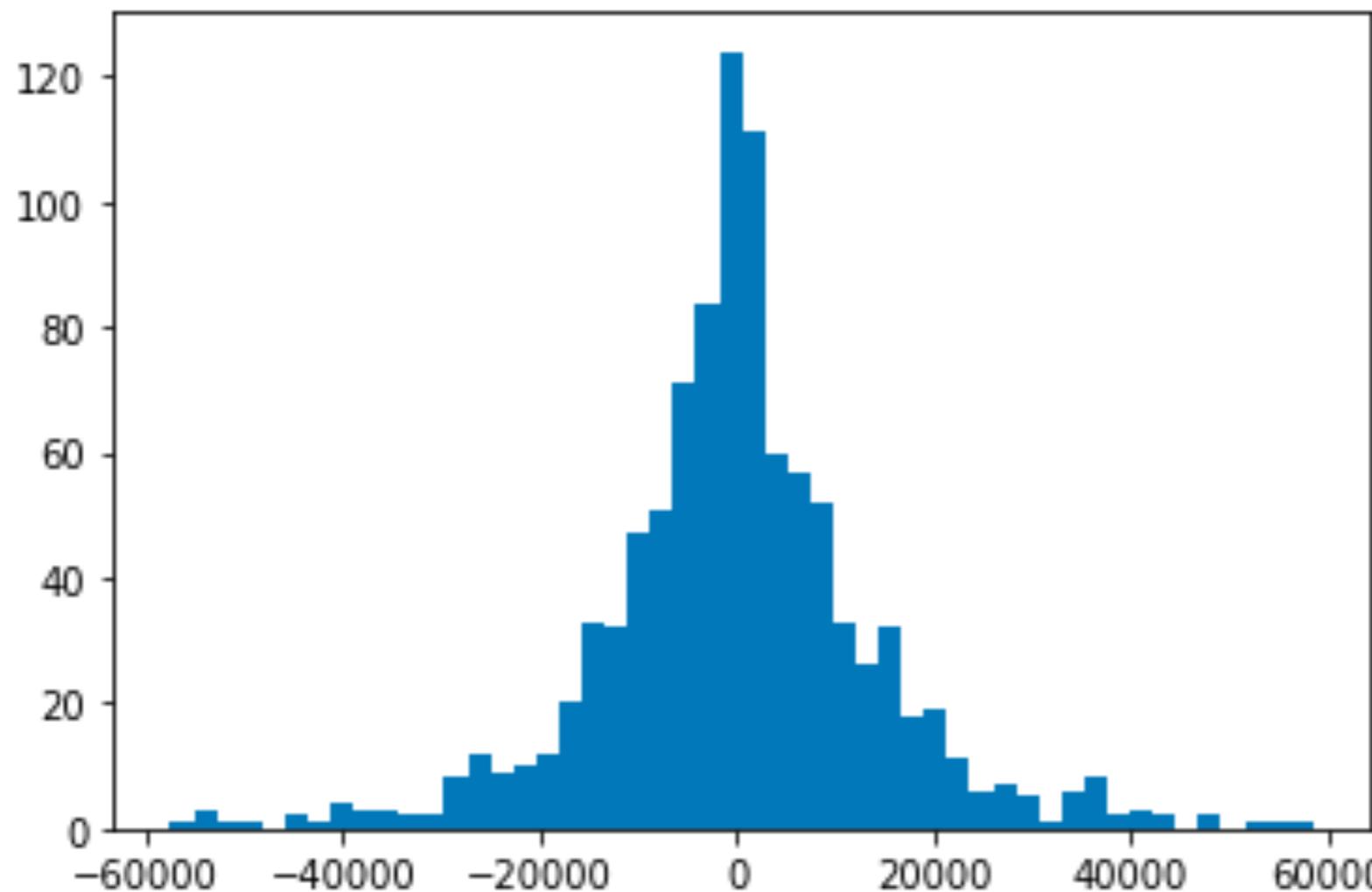


Podemos deducir que  
es 1!

# ¿Qué pasa si repetimos la consulta muchas veces?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1



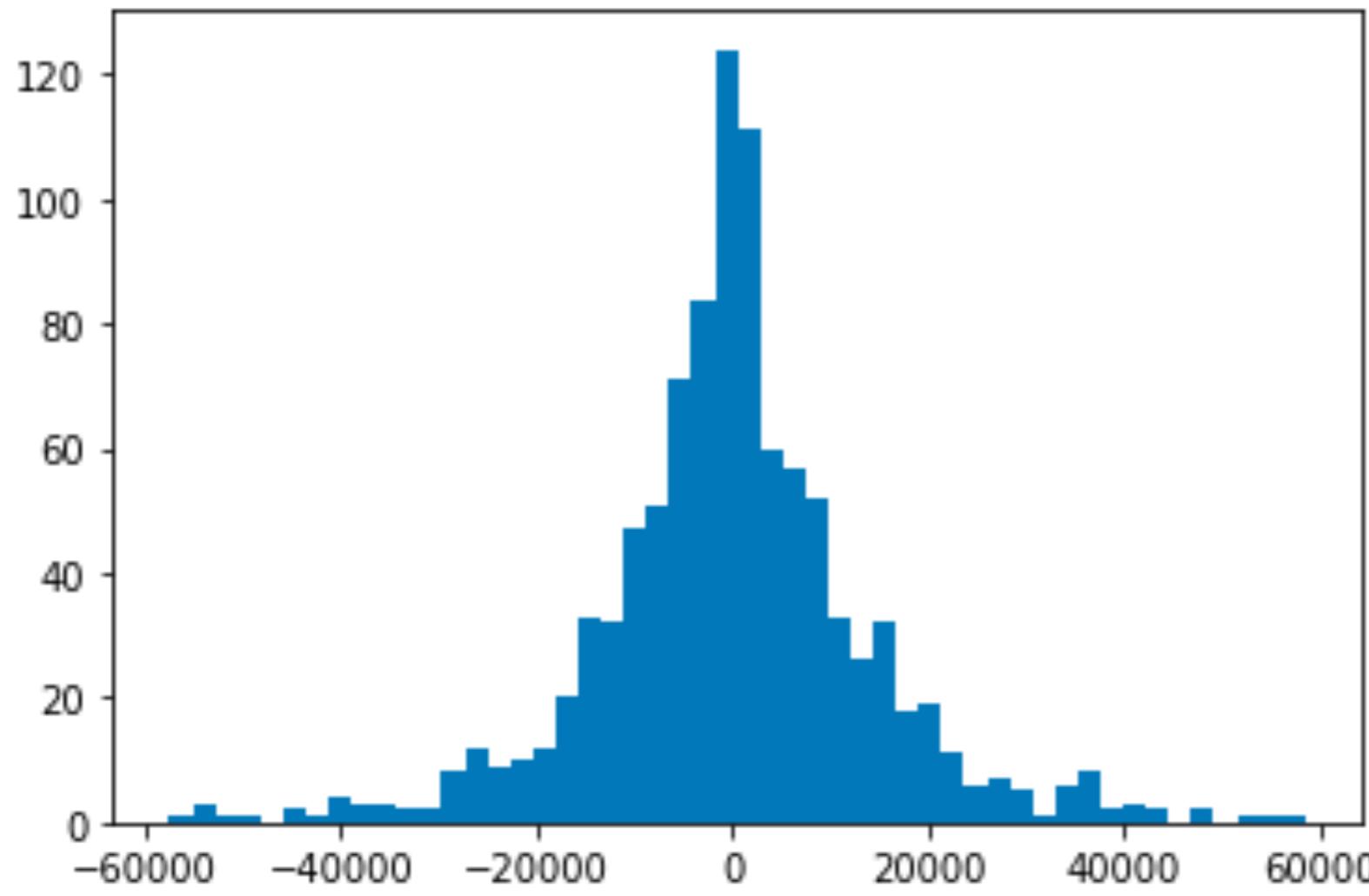
Podemos deducir que es 1!

Es por eso que  $\epsilon$  se le llama presupuesto

# ¿Qué pasa si repetimos la consulta muchas veces?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

1



Podemos deducir que es 1!

Es por eso que  $\epsilon$  se le llama presupuesto

- Las consultas no son independientes!
- El presupuesto debe distribuirse por cada una de las consultas  
⇒ resultados más ruidosos

# ¿Qué pasa si repetimos la consulta muchas veces?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

# ¿Qué pasa si repetimos la consulta muchas veces?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

Si ejecutamos 4 veces la consulta dividimos  $\epsilon$  en 4

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

$$f(x) + \text{Lap}\left(\frac{1}{\left(\frac{\epsilon}{4}\right)}\right)$$

# ¿Qué pasa si repetimos la consulta muchas veces?

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

Si ejecutamos 4 veces la consulta dividimos  $\epsilon$  en 4

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

$$f(x) + \text{Lap}\left(\frac{1}{\left(\frac{\epsilon}{4}\right)}\right)$$

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

El ruido será mayor en cada consulta

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

# Composición secuencial

# Composición secuencial

- Si  $F_1(x)$  satisface  $\epsilon_1$ -privacidad diferencial

# Composición secuencial

- Si  $F_1(x)$  satisface  $\epsilon_1$ -privacidad diferencial
- Si  $F_2(x)$  satisface  $\epsilon_2$ -privacidad diferencial

# Composición secuencial

- Si  $F_1(x)$  satisface  $\epsilon_1$ -privacidad diferencial
- Si  $F_2(x)$  satisface  $\epsilon_2$ -privacidad diferencial
- Entonces el mecanismo  
 $G(x) = (F_1(x), F_2(x))$  que libera ambos resultados satisface  $(\epsilon_1 + \epsilon_2)$ -privacidad diferencial

# Composición secuencial

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

$$f(x) + \text{Lap}\left(\frac{1}{\left(\frac{\epsilon}{4}\right)}\right)$$

# Composición secuencial

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

```
SELECT count(*) FROM person  
WHERE name='Alan Brito' AND target='<=50k'
```

$$f(x) + \text{Lap}\left(\frac{1}{\left(\frac{\epsilon}{4}\right)}\right)$$

La **publicación de estas cuatro consultas** satisface

$$\left(\frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4}\right) -\text{privacidad diferencial.}$$

# Composición paralela

# Composición paralela

- Si  $F(x)$  satisface  $\epsilon_1$ -privacidad diferencial

# Composición paralela

- Si  $F(x)$  satisface  $\epsilon_1$ -privacidad diferencial
- Y dividimos un conjunto de datos  $X$  en  $k$  pedazos distintos de modo que  $x_1 \cup \dots \cup x_k = X$

# Composición paralela

- Si  $F(x)$  satisface  $\epsilon_1$ -privacidad diferencial
- Y dividimos un conjunto de datos  $X$  en  $k$  pedazos distintos de modo que  $x_1 \cup \dots \cup x_k = X$
- Luego, el mecanismo que libera todos los resultados  $F(x_1), \dots, F(x_k)$  satisface  $\epsilon$ -privacidad diferencial

# Composición paralela

- Si  $F(x)$  satisface  $\epsilon_1$ -privacidad diferencial
- Y dividimos un conjunto de datos  $X$  en  $k$  pedazos distintos de modo que  $x_1 \cup \dots \cup x_k = X$
- Luego, el mecanismo que libera todos los resultados  $F(x_1), \dots, F(x_k)$  satisface  $\epsilon$ -privacidad diferencial

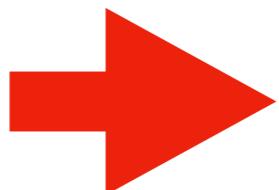
Este límite es mucho mejor que lo que daría la composición secuencial:  
 $k\epsilon$ -privacidad diferencial

# Composición paralela

<b>Education</b>	<b>Female</b>	<b>Male</b>
<b>10th</b>	295	638
<b>11th</b>	432	743
<b>12th</b>	144	289
<b>1st-4th</b>	46	122
<b>5th-6th</b>	84	249
<b>7th-8th</b>	160	486
<b>9th</b>	144	370
<b>Assoc-acdm</b>	421	646
<b>Assoc-voc</b>	500	882
<b>Bachelors</b>	1619	3736
<b>Doctorate</b>	86	327
<b>HS-grad</b>	3390	7111
<b>Masters</b>	536	1187
<b>Preschool</b>	16	35
<b>Prof-school</b>	92	484
<b>Some-college</b>	2806	4485

# Composición paralela

<b>Education</b>	<b>Female</b>	<b>Male</b>
<b>10th</b>	295	638
<b>11th</b>	432	743
<b>12th</b>	144	289
<b>1st-4th</b>	46	122
<b>5th-6th</b>	84	249
<b>7th-8th</b>	160	486
<b>9th</b>	144	370
<b>Assoc-acdm</b>	421	646
<b>Assoc-voc</b>	500	882
<b>Bachelors</b>	1619	3736
<b>Doctorate</b>	86	327
<b>HS-grad</b>	3390	7111
<b>Masters</b>	536	1187
<b>Preschool</b>	16	35
<b>Prof-school</b>	92	484
<b>Some-college</b>	2806	4485



<b>Education</b>	<b>Female</b>	<b>Male</b>
<b>10th</b>	295.591316	638.730833
<b>11th</b>	433.991320	742.901816
<b>12th</b>	144.546311	289.543775
<b>1st-4th</b>	45.877242	123.385749
<b>5th-6th</b>	86.657895	249.939610
<b>7th-8th</b>	161.499974	485.533936
<b>9th</b>	143.441692	367.792887
<b>Assoc-acdm</b>	420.879967	644.833957
<b>Assoc-voc</b>	500.347331	881.176781
<b>Bachelors</b>	1618.321878	3736.108821
<b>Doctorate</b>	83.983045	325.787201
<b>HS-grad</b>	3390.710973	7112.123852
<b>Masters</b>	537.215908	1185.515231
<b>Preschool</b>	15.782764	32.575344
<b>Prof-school</b>	90.655808	483.799082
<b>Some-college</b>	2804.440558	4485.117936

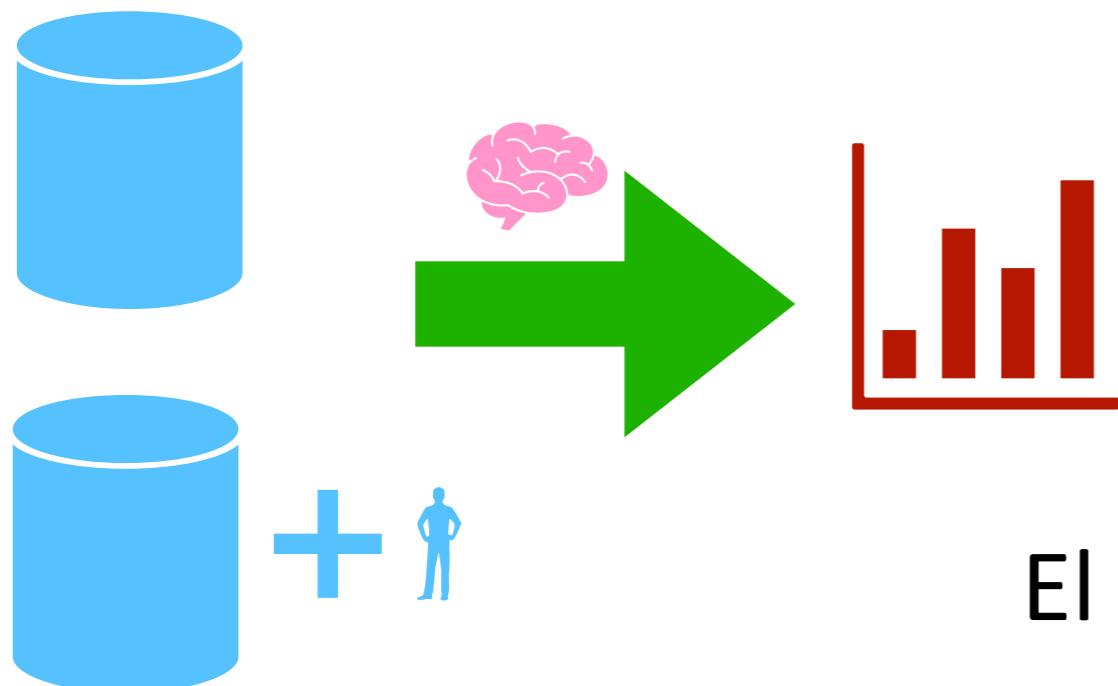
# Composición paralela

```
SELECT education, sex,  
       count(*)+laplace(—)  
              ^  
             ε  
FROM person  
GROUP BY education, sex
```

Education	Female	Male
10th	295.591316	638.730833
11th	433.991320	742.901816
12th	144.546311	289.543775
1st-4th	45.877242	123.385749
5th-6th	86.657895	249.939610
7th-8th	161.499974	485.533936
9th	143.441692	367.792887
Assoc-acdm	420.879967	644.833957
Assoc-voc	500.347331	881.176781
Bachelors	1618.321878	3736.108821
Doctorate	83.983045	325.787201
HS-grad	3390.710973	7112.123852
Masters	537.215908	1185.515231
Preschool	15.782764	32.575344
Prof-school	90.655808	483.799082
Some-college	2804.440558	4485.117936

# ¿Cuales son las garantías de PD?

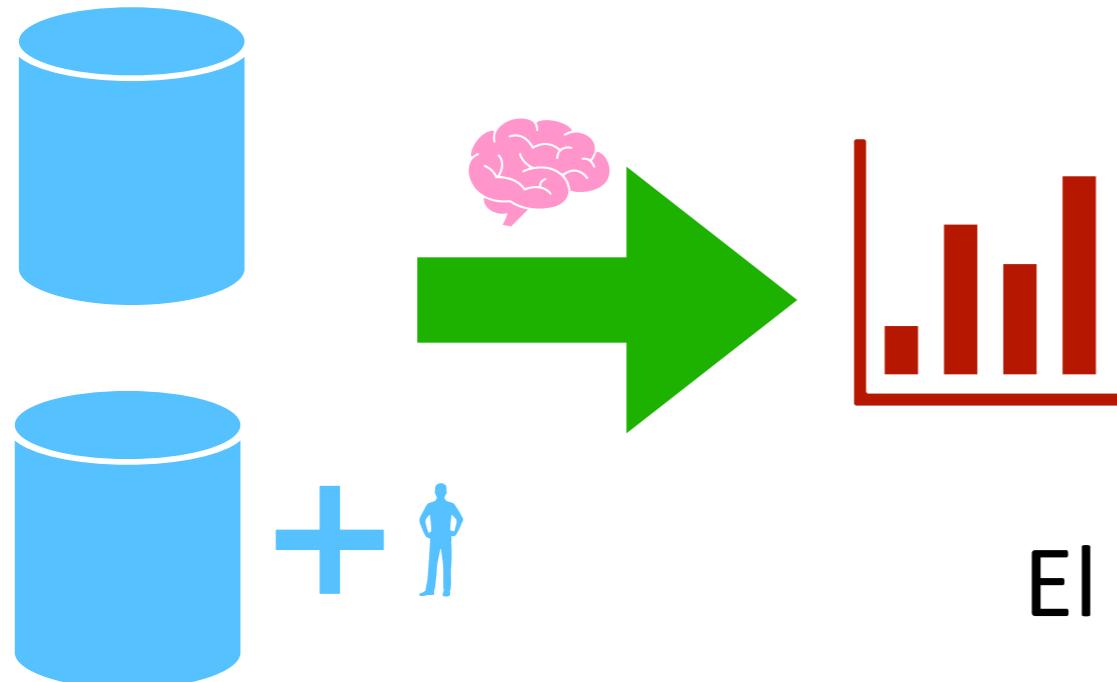
Se mantiene independiente de  
los datos auxiliares que pudiera  
tener un adversario!



El resultado del mecanismo no deja  
que un adversario pueda distinguir  
entre las dos bases de datos

# ¿Cuales son las garantías de PD?

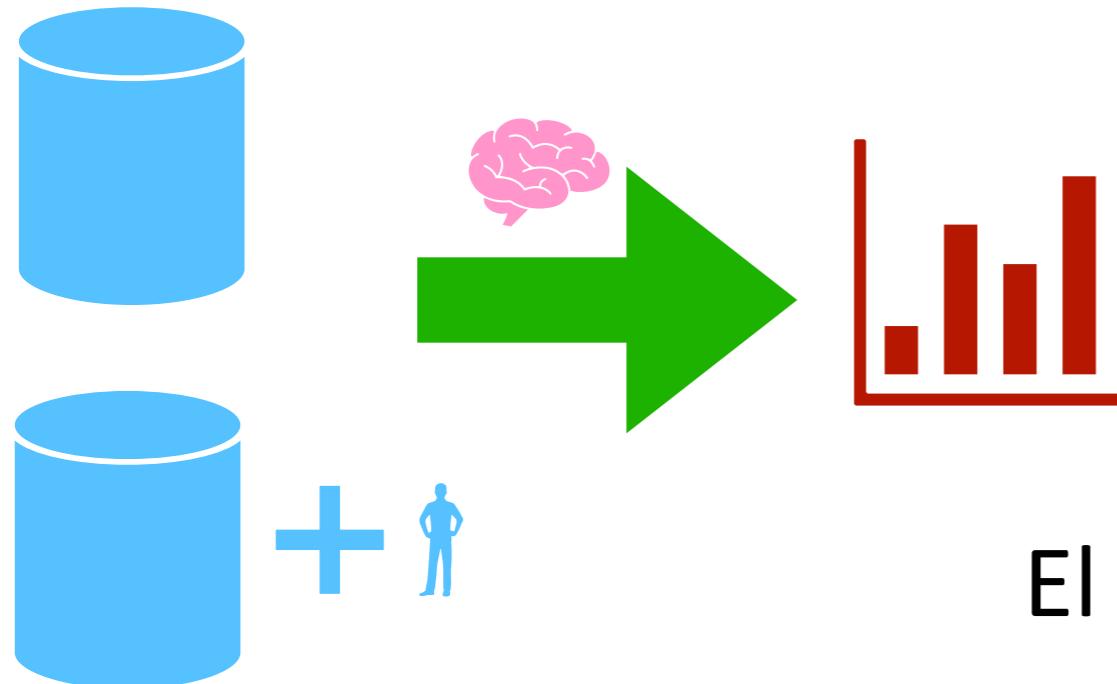
Se mantiene independiente de  
los datos auxiliares que pudiera  
tener un adversario!



El resultado del mecanismo no deja  
que un adversario pueda distinguir  
entre las dos bases de datos

# ¿Cuales son las garantías de PD?

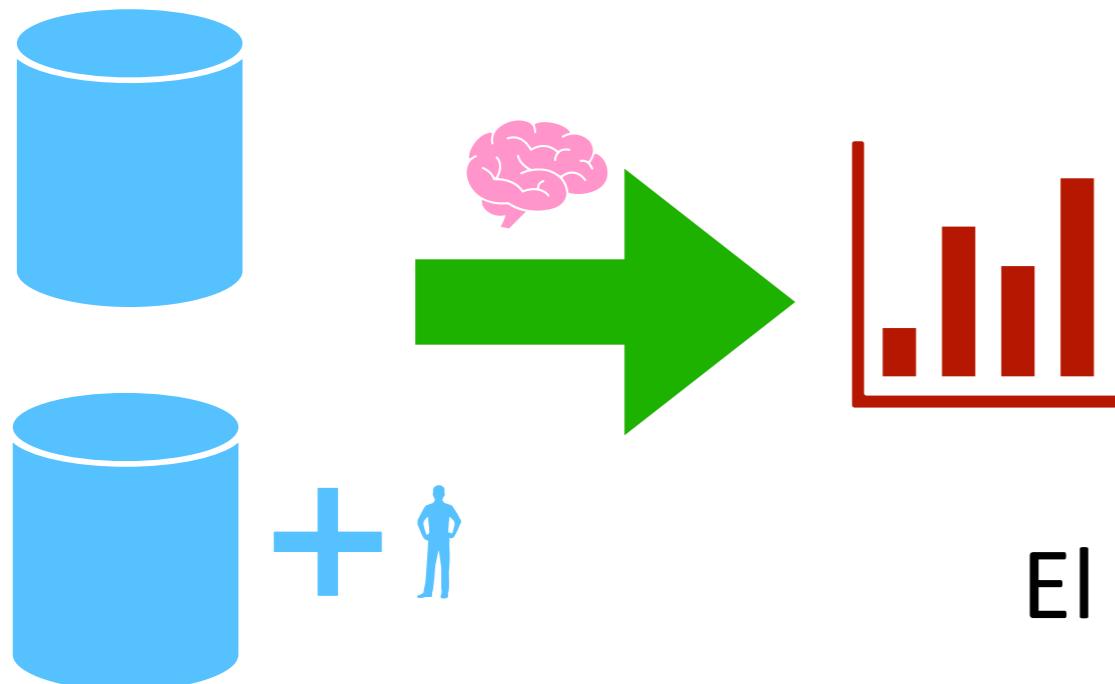
Se mantiene independiente de  
los datos auxiliares que pudiera  
tener un adversario!



El resultado del mecanismo no deja  
que un adversario pueda distinguir  
entre las dos bases de datos

# ¿Cuales son las garantías de PD?

Se mantiene independiente de  
los datos auxiliares que pudiera  
tener un adversario!



El resultado del mecanismo no deja  
que un adversario pueda distinguir  
entre las dos bases de datos

# ¿Cuales son las garantías de PD?

# ¿Cuales son las garantías de PD?

- Calza con una buena intuición de privacidad: “nada malo me pasa a mí como resultado de mi participación en el análisis”
  - Si algo malo pasa, hubiera pasado *aun si* yo no hubiera participado

# ¿Cuales son las garantías de PD?

- Calza con una buena intuición de privacidad: “nada malo me pasa a mí como resultado de mi participación en el análisis”
  - Si algo malo pasa, hubiera pasado *aun si* yo no hubiera participado
- Definiciones formales permite *probar* que el mecanismo satisface privacidad diferencial

# ¿Cuales son las garantías de PD?

- Calza con una buena intuición de privacidad: “nada malo me pasa a mí como resultado de mi participación en el análisis”
  - Si algo malo pasa, hubiera pasado *aun si* yo no hubiera participado
- Definiciones formales permite *probar* que el mecanismo satisface privacidad diferencial
- **Se mantiene independiente de los datos auxiliares que pudiera tener un adversario!**
  - Incluido el caso donde el adversario conoce toda la base de datos excepto la fila objetivo
  - Previene ataques de asociación en  $k$ -anonimato y  $\ell$ -diversidad
  - La única forma conocida que se acerca a una “anonimización verdadera”

¿Que es lo malo?

# ¿Qué es lo malo?

- Es complejo para datos sintéticos:

# ¿Que es lo malo?

- Es complejo para datos sintéticos:

Nombre	Fecha Nac.	Sexo	Coding Postal
*****	26/02/2018	M	73909
*****	08/05/2018	M	14890
*****	11/05/2018	F	58821
*****	31/03/2019	F	465113
*****	26/04/2018	F	90825

# ¿Qué es lo malo?

- Es complejo para datos sintéticos:

Nombre	Fecha Nac.	Sexo	Coding Postal
*****	26/02/2018	M	73909
*****	08/05/2018	M	14890
*****	11/05/2018	F	58821
*****	31/03/2019	F	465113
*****	26/04/2018	F	90825

- Garantía difícil de interpretar

# ¿Qué es lo malo?

- Es complejo para datos sintéticos:

Nombre	Fecha Nac.	Sexo	Coding Postal
*****	26/02/2018	M	73909
*****	08/05/2018	M	14890
*****	11/05/2018	F	58821
*****	31/03/2019	F	465113
*****	26/04/2018	F	90825

- Garantía difícil de interpretar

$$\text{Lap}(\frac{s}{\epsilon})$$

# ¿Qué es lo malo?

- Es complejo para datos sintéticos:

Nombre	Fecha Nac.	Sexo	Coding Postal
*****	26/02/2018	M	73909
*****	08/05/2018	M	14890
*****	11/05/2018	F	58821
*****	31/03/2019	F	465113
*****	26/04/2018	F	90825

- Garantía difícil de interpretar

$$\text{Lap}(\frac{s}{\epsilon})$$


# RESUMEN

# Resumen (1)

# Resumen (1)

- **De-identificación / Anonimización**
  - Elimina los identificadores exclusivos para reducir el riesgo de re-identificación
  - Enfoques ad-hoc se traducen en un alto riesgo de errores
  - Técnica más usada

# Resumen (1)

- **De-identificación / Anonimización**
  - Elimina los identificadores exclusivos para reducir el riesgo de re-identificación
  - Enfoques ad-hoc se traducen en un alto riesgo de errores
  - Técnica más usada
- **$k$ -Anonimato**
  - Formaliza de-identificación sistemática
  - Requiere grupos de tamaño al menos  $k$
  - Sujeto a ataques de homogeneidad y datos auxiliares

# Resumen (2)

# Resumen (2)

- $\ell$ -Diversidad
  - Requiere grupos *diversos*
  - Previene ataques de homogeneidad
  - Previene ataques de datos auxiliares cuando el adversario conoce menos de  $\ell - 2$  datos negativos acerca del grupo

# Resumen (2)

- $\ell$ -Diversidad
  - Requiere grupos *diversos*
  - Previene ataques de homogeneidad
  - Previene ataques de datos auxiliares cuando el adversario conoce menos de  $\ell - 2$  datos negativos acerca del grupo
- Privacidad diferencial
  - Propiedad formal de un *mecanismo* (algoritmo o análisis)
  - Corresponde a la noción de indistinguibilidad: **mismo resultado**, ya sea si participé o no
  - La garantía se mantiene **independiente de los datos auxiliares del adversario**

**¿Preguntas?**