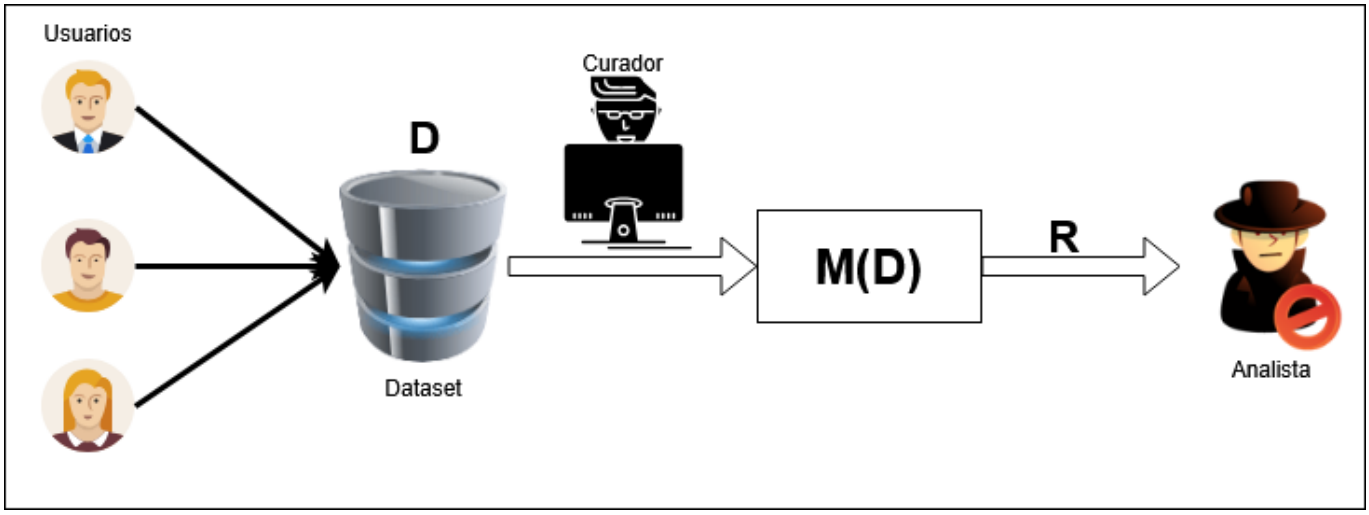
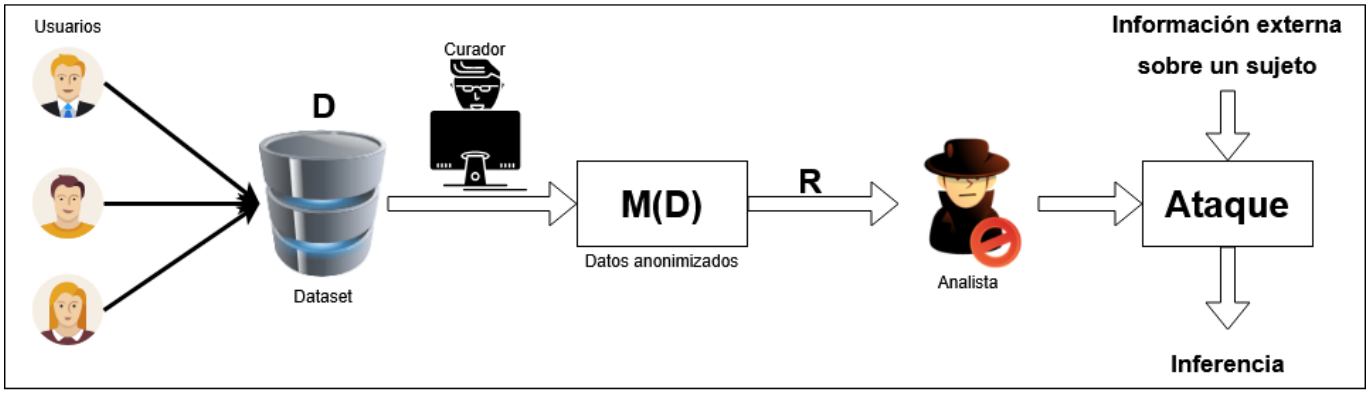
Privacidad Diferencial

Tenemos un dataset D que contiene datos de usuarios, siendo cada fila los datos de un usuario. El curador, una entidad de confianza para los usuarios, publica algunos datos usando un mecanismo M que da como resultado $R = M(D)$. El adversario trata de realizar inferencias sobre los datos D contenidos en R.



- **Usuario**: Fila en la BBDD
- **Curador**: Procesa los datos para responder a un analista
  - Publica mediante un mecanismo ($M(D)$), un algoritmo que oculta los datos a porteger.
- **Analista**: Intenta hacer inferencias de la BBDD con lo que se publica

Contra que protege la privacidad diferencial



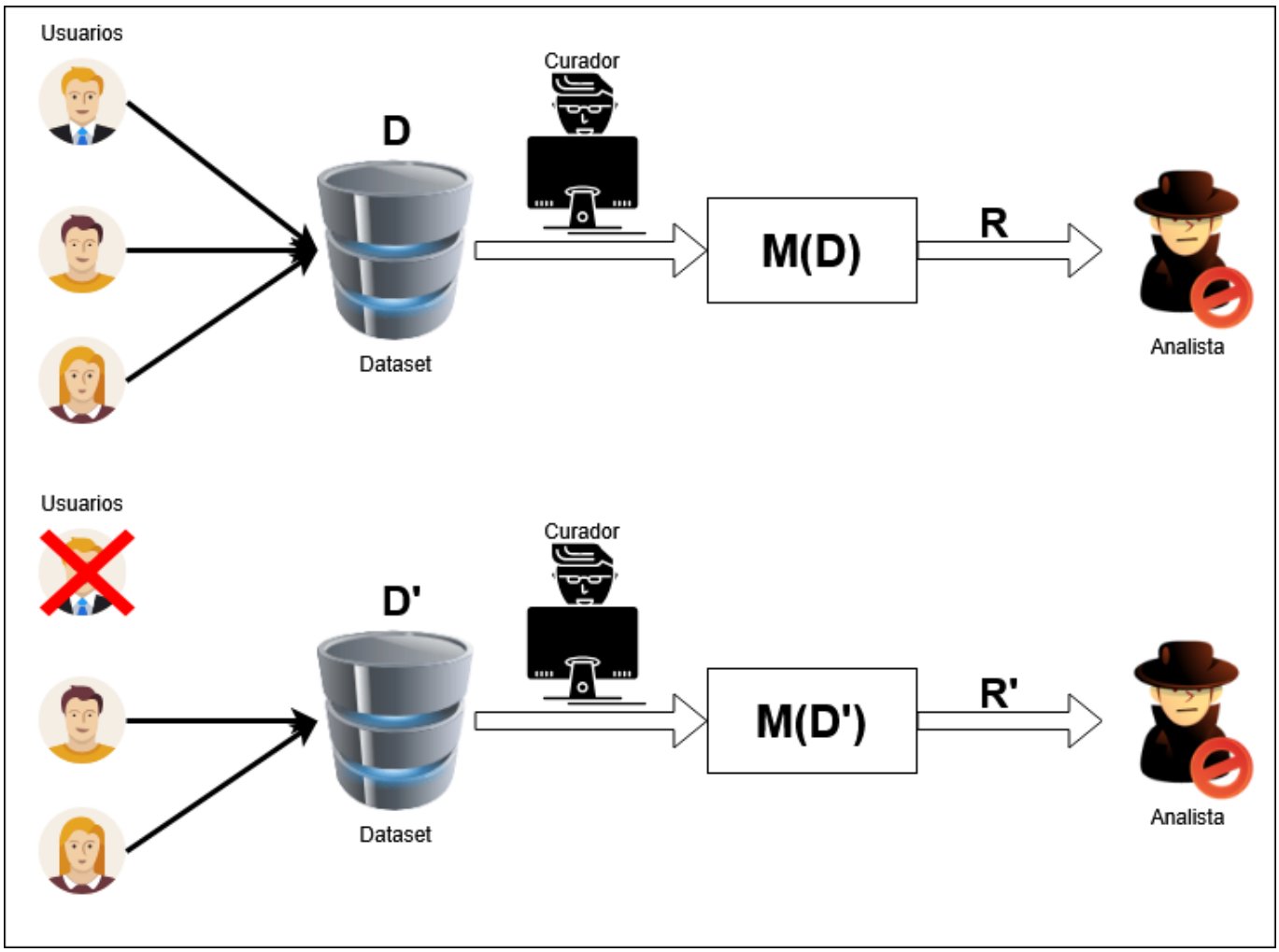
La privacidad diferencial protege contra el riesgo de conocimiento de información de un sujeto mediante el uso de inferencia con información externa sobre dicho sujeto. De esta forma, observando la respuesta R no se puede cambiar lo que el adversario puede saber.

No protege contra inferencias que se puedan hacer fuera de la Base de Datos Suponemos que existen 2 Bases de Datos con la única diferencia de que una tiene a Alice y la otra no.

- Se modifica una fila de una Base de datos a Otra
- Existen unas salidas con Alice y otras sin ella.

- La privacidad diferencial busca que estas 2 salidas NO se puedan distinguir
- El mecanismo $M(D)$ no puede dar un resultado diferente cuando Alice está y cuando no.
- El curador va a tomar las respuestas y les va a añadir ruido para que no se puedan distiguir.
- Las distribuciones deben ser lo más parecidas posibles
    - Las 2 distribuciones se tienen que parecer para que la diferencia entre una y otra no sea mayor de un valor P que establecemos.
- So dos bases de datos difieren en una sola fila son bases de datos vecinas.

En resumidas cuentas, la clave para dificultar que un adversario pueda identificar datos sobre un sujeto es crear dos salidas $R = M(D)$ y $R' = M(D')$, siendo D y D' Bases de datos vecinas, de forma que ambas respuestas no puedan ser distinguidas. Para hacer esto se diseña un mecanismo M el cual no puede ser determinístico, tiene que ser probabilístico.



Como definir distribuciones similares

Definición tentativa de privacidad con parámetro P

Un mecanismo M es privado si para todas las posibles salidas de R y todos los pares de Bases de datos Vecinas $(D, 'D)$:

$Pr(M(D') = R) - p < Pr(M(D) = R) < Pr(M(D') = R) + p$

El valor de P debe ser uno que no facilite que se pueda identificar cuando Alice no está en la Base de datos. Con esta primera definición nos encontramos con el problema de existen ciertas salidas de R que solo pueden ocurrir cuando la entrada es D', permitiendo diferenciar los dos datasets. Para corregir esto, se realiza la siguiente definición:

Definición tentativa de privacidad 2 con parámetro P

Un mecanismo M es privado si para todas las posibles salidas de R y para todos los pares de las bases de datos vecinas (D, D'):

$\frac{Pr(M(D')=R)}{p}≤Pr(M(D)=R)≤Pr(M(D)=R)*p$

Cuanto más alto sea el valor de P menor es la privacidad, por lo tanto, si $P=∞$ no hay privacidad.

Definición de Privacidad Diferencial (PD)

Es similar a las dos definiciones antes realizadas, con la diferencia de que se sustituye p con $e^ε$:

Un mecanismo $M: D$ → *R* es ε-privadamente diferencial ($ε-DP$) si para todas las posibles salidas $R ∈ $ *R* y todos los pares de las Bases de Datos Vecinas $D,D'∈ D$:

$Pr(M(D) = R) ≤ Pr(M(D') = R) * e^ε$

- Usamos $e^ε$ en vez de $ε$ ya que hace más fácil formular algunos teoremas útiles
- $ε ∈ [0,∞)$ asegura que $e^ε ∈ [1,∞)$
- Cuanto más pequeño es el valor de $ε$, mayor es la privacidad
- La privacidad perfecta se da cuando $ε = 0$, pero en cambio, la salida que se obtiene es completamente inútil.
- No existe consenso sobre como de pequeño debe ser el valor de $ε$, pero debe tener un valor que evite que la salida del mecanismo sea inútil.

Si tomamos logaritmos naturales, aparece la siguiente definición alternativa:

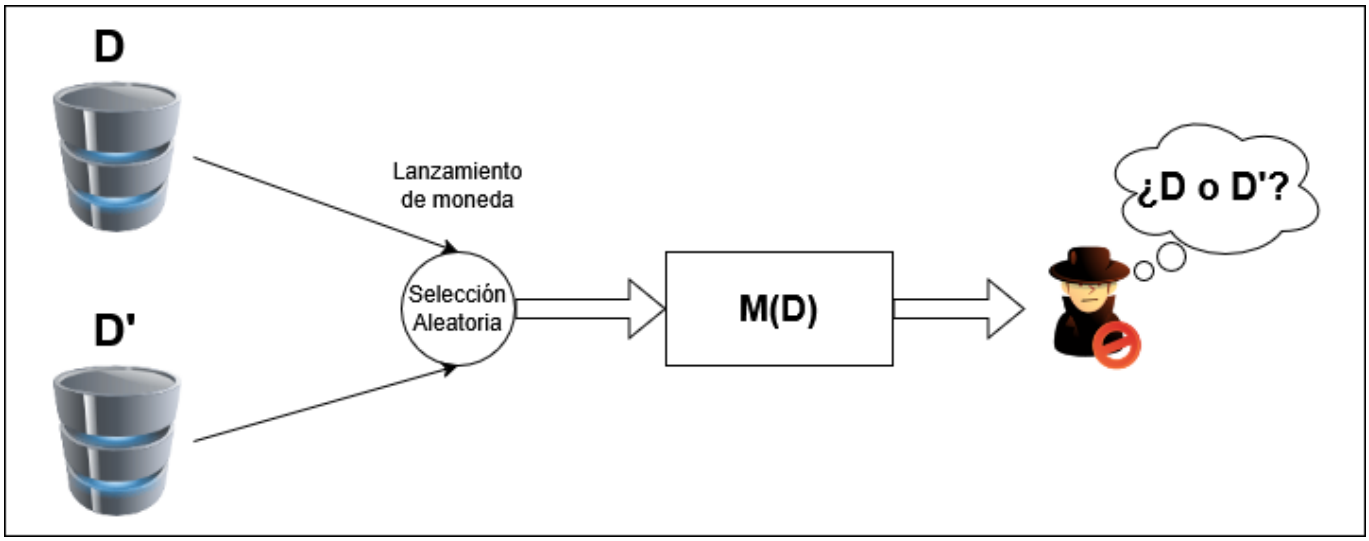
Privacidad diferencial: Definición alternativa

A un mecanismo $M : D$ → *R* es ε-privadamente diferencial (ε-PD) si para todas las posibles salidas $R ∈$ *R* y todos los pares de las bases de datos vecinas $D,D' ∈$ *D*:

$$|log(Pr(M(D) = R)) - log(Pr(M(D') = R))| ≤ ε$$

Privacidad diferencial como un juego de decisión estadística

Uno de los problemas de la privacidad diferencial es que es difícil de interpretar, por ejemplo, considerando el siguiente caso:



- Teniendo en cuenta este caso, sabemos que $Pr(D) = Pr(D') = 0.5$
- El adversario tiene que decidir D si $Pr(D|R) > Pr(D'|R')$, en caso contrario decide D'
- Al hacer eso, existe una probabilidad P_{err} de que el adversario se equivoque.
- Las probabilidades pueden ser calculadas usando el teorema de Bayes:

$$Pr(D|R) = \frac{Pr(R|D) * Pr(D)}{Pr(R)} = \frac{Pr(R|D)}{Pr(R|D) + Pr(R|D')}$$

- Lo que es equivalente a:

$$Pr(D|R) = \frac{1}{1+\frac{Pr(R|D)}{Pr(R|D)}}$$

- Pero si el mecanismo es $ε-DP$, entonces:

$$\frac{1}{1+e^ε} ≤ Pr(D|R) ≤ \frac{1}{1+e^{-ε}}$$

Sobre la Privacidad Diferencial y el rendimiento de un ataque empírico

La privacidad diferencial asegura la protección incluso contra adversarios poderosos que saben que la entrada es D o D'. En la práctica un algoritmo que provee $ε = 10$ puede proveer alta protección empírica contra ataques existentes. En este punto el problema es que el peor caso teórico no importa ya que uno puede usar algo que no de privacidad diferencial pero se obtiene un mejor rendimiento empírico.

El problema con ε-DP

Considerando el mecanismo laplaciano siendo X_i valores del dataset:

$$r = M(D) = \frac{1}{n}Σ_{i=1}^n X_i + y$$

Donde y es una muestra de una distribución laplaciana con media 0 y escalka b:

$$f(y) = \frac{1}{2b}e^{-\frac{|y|}{b}} = Lap(b)$$

Este mecanismo provee 1/b-privacidad diferencial. La salida de este mecanismo debe tener la siguiente distribución:

$$ f(r|D) = \frac{1}{2b} \exp \left(- \frac{|r - \frac{1}{n} \sum_{i=1}^{n} x_i|}{b} \right)
$$f(r|D') = \frac{1}{2b} \exp \left(- \frac{|r - \frac{1}{n} \sum_{i=1}^{n} x'_i|}{b} \right) $

Estas dos distribuciones difieren en la media. Tomando $b=1$ obtenemos una $ε-DP$ con $ε=1$. Supongamos que trucamos el laplaciano en $y > 1000$. El mecanismo es prácticamente el mismo ya que:

$$Pr[Lap(1) > 1000] = \frac{e^{-1000}}{2} ≈ 10^{-43}$$

De todas formas, truncando vamos de $ε = 1$ a $ε = ∞$, por lo que pasamos de tener una

privacidad muy buena a no tener nada de privacidad.

Privacidad diferencial aproximada

Un mecanismo $M : D$ → *R* es (ε,δ)-privadamente diferencial ($(ε,δ)-DP$) si para todas las posibles salidas $R ⊂$ *R* y los pares de bases de datos vecinas $D,D' ∈ D$:

```
$Pr(M(D) ∈ R) ≤ Pr(M(D') ∈ R) * e^ε + δ$
```

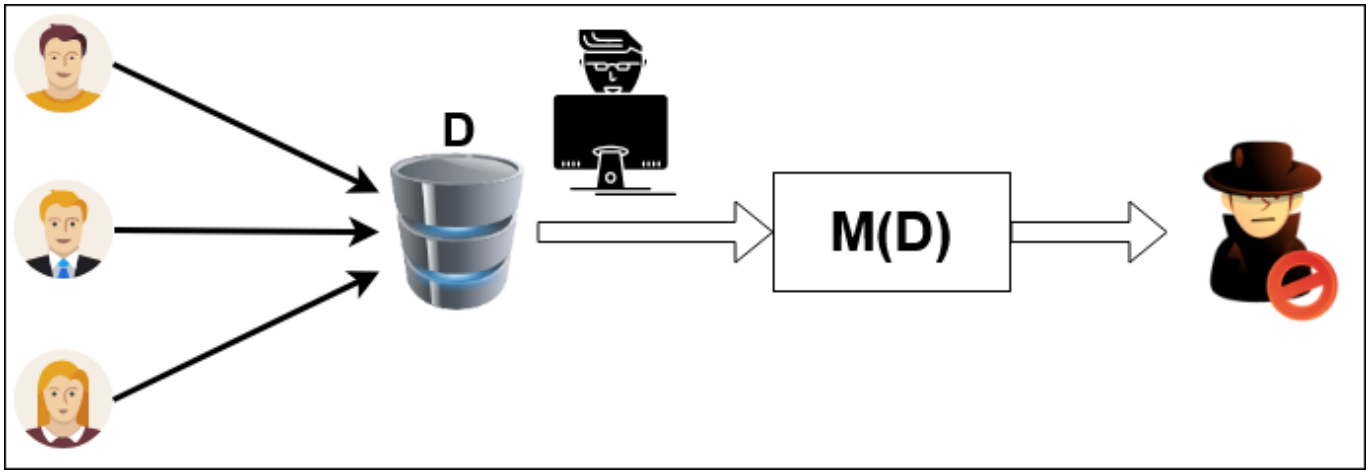
Esta definición es una relajación de la de Privacidad diferencial que permite cierta tolerancia. Si $δ = 0$, entonces tenemos el mismo caso que $ε-DP$

Escenarios de Privacidad Diferencial

Dependiendo de donde se ejecuta el mecanismo $M(D)$ hay 2 modelos generales para la privacidad diferencial:

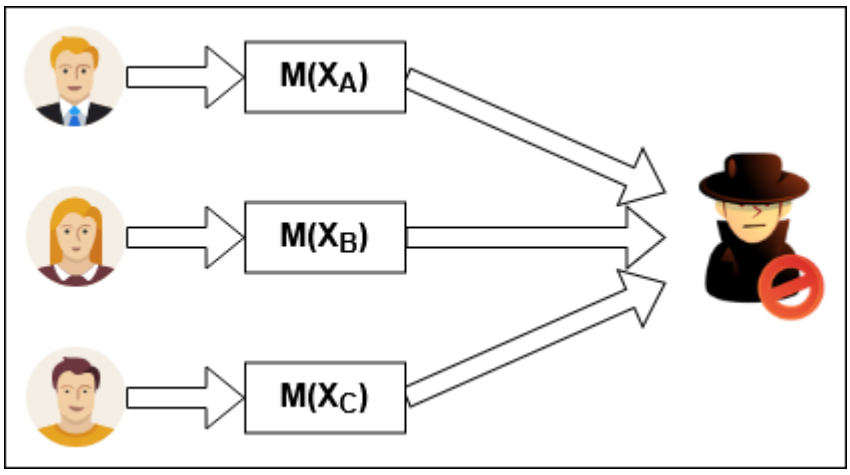
- **Privacidad diferencial central**: Hay un agregador de confianza centralizado.
  - Un mecanismo $M : D$ → *R* es ε-privadamente diferencial si para todas las salidas $R ⊂$ *R* y todos los pares de bases de datos vecinas $D,D' ∈ D$:

```
$Pr(M(D) ∈ R) ≤ Pr(M(D') ∈ R) * e^ε$
```



- **Privacidad diferencial local**: Cada usuario ejecuta el mecanismo y reporta el resultado al analista
  - Un mecanismo $M : D$ → *R* es ε-privadamente diferencial si para todas las salidas $R ⊂$ *R* y todos los pares de inputs X,X':

$Pr(M(x) \in R) \leq Pr(M(x') \in R) * e^\varepsilon$



La definición de la privacidad diferencial se asegura de que correr el mecanismo $M(D)$ en una base de datos vecina producen resultados similares. Normalmente hay 2 definiciones principales para como se definen las bases de datos vecinas en el modelo central:

- **Privacidad Diferencial Acotada**: $D y D'$ tienen el mismo número de entradas, pero difieren en el valor de una de ellas (Se modifica el valor de una fila)
- **Privacidad Diferencial No Acotada**: D' es obtenida tras eliminar una entrada de D (Se elimina una fila)