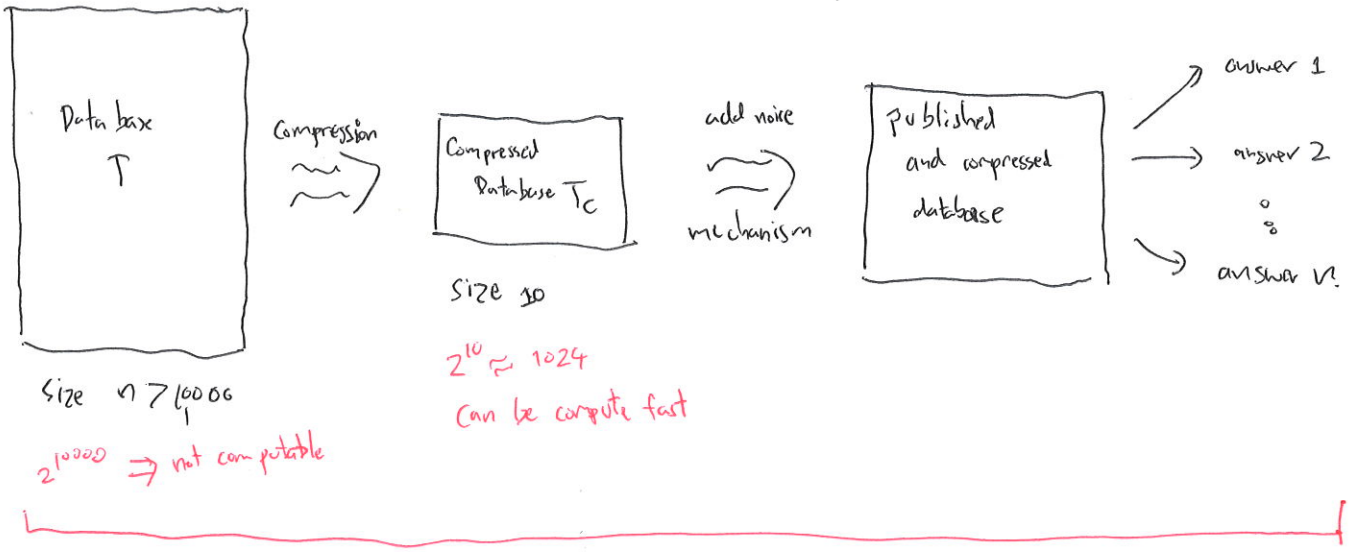


Assumption : 1. We know the query set in advance. It is $f_1, \dots, f_m$. (So the results should be $f_1(T), \dots, f_m(T)$

(we actually do not need this assumption)

2. Fuction f_i can be written in the form

$$f_i(T) = \frac{\sum_{j=1}^{q} f_i(T_j)}{\|T\|}$$

T_j → record j of Table T

$\|T\|$ → size of T (#records)

$f_i(T_j)$ must be between 0 and 1

Example

Name	Chosen Party	Age < 40
Alice	Melon-pan	30
Bob	Melon-pan	40
Charles	Gyu-don	20
Doe	Gyu-don	20

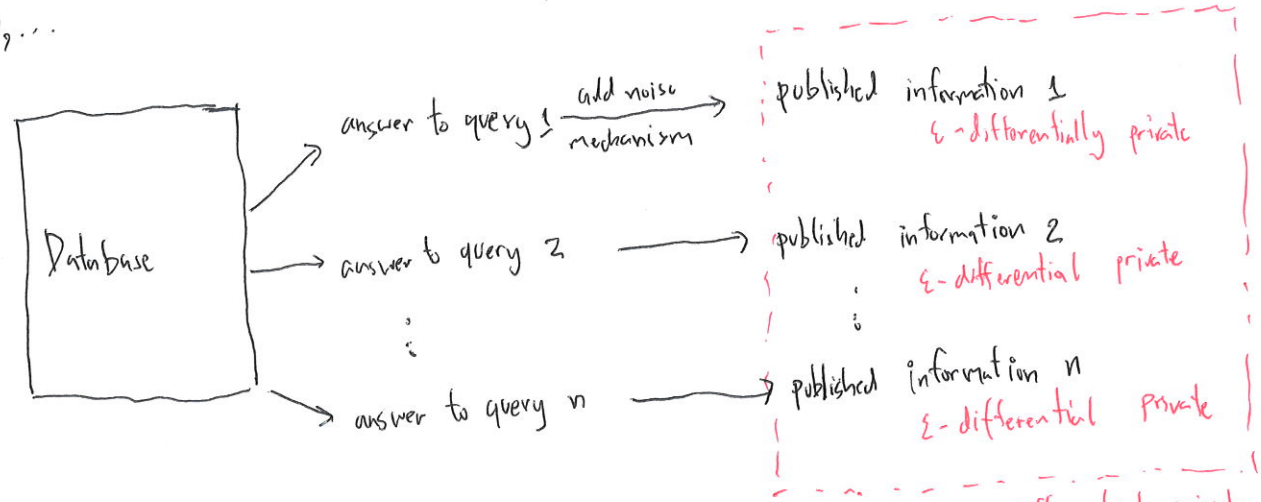
How many persons "choose Gyu-don" and "age smaller than 30" ?

$$f_?((N, P, A)) = \begin{cases} 1 & \text{then } P = \text{Gyu-don and } A \le 30 \\ 0 & \text{otherwise} \end{cases}$$

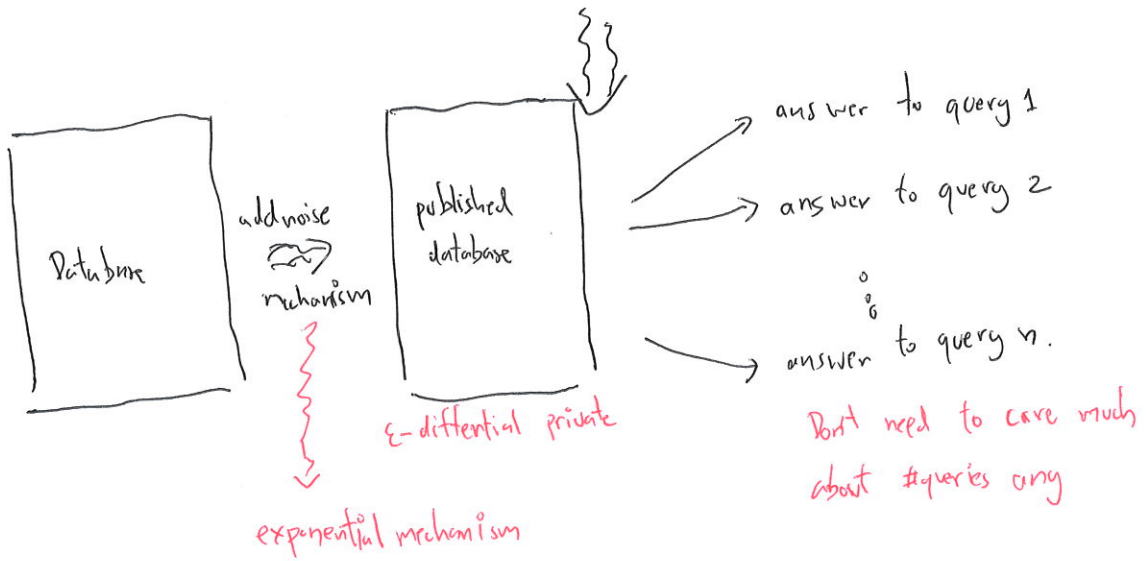
$$f_?(T) = \frac{\#\text{persons who "choose Gyu-don" and "age smaller than 30"}}{\|T\|} \quad \|T\| \to \#\text{records}$$

$$\#\text{persons} = f_?(T) \cdot \|T\|$$

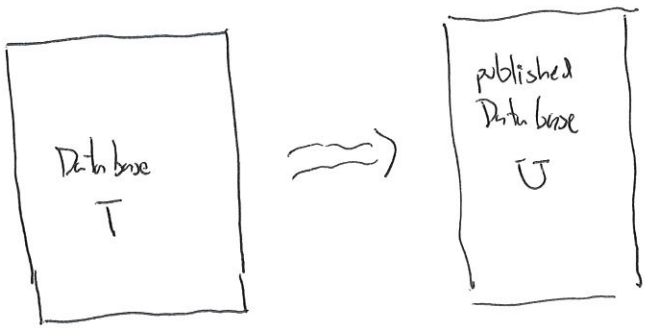
Until now, ...



can be risky when n is large ⟺ Can-differential private



Don't need to care much about #queries any



$Utility(T, U)$ → better large if U is similar to T
small otherwise

$$h_T(U) = \exp\left(\frac{\varepsilon \cdot Utility(T, U)}{2\,\Delta Utility}\right)$$

$$Pr_T[Out(T) = U] = \frac{h_T(U)}{\sum_{\text{all possible databases } U'} h_T(U')}$$

→ takes a lot of time to compute.

too many possible databases

at least 2^n possible databases for n persons!!

Average Age ?

$$f_i((\text{name}, A)) = \frac{A}{200}$$

$$f_i(T) = \frac{\sum_r A_i/200}{\|T\|} = \frac{\sum_i A_i}{\|T\|} \cdot \frac{1}{200}$$

Average height $= f_i(T) \cdot 200$

Small Database (SmallDB) Algorithm — Use Exponential Mechanism

$P := \{$ Database with $\frac{m}{\alpha}$ $\}$ → parameter to define later

P: possible report

$$\text{Utility}(T, U) := -\max_i |f_i(T) - f_i(U)|$$

↳ Different in query results

↳ take the maximum difference

→ the larger difference, the smaller utility.

Chose $U \in P$ using exponential mechanism

Privacy: ε-differentially private because of the exponential mechanism.

Chernoff Bound Let $X_1, \ldots, X_n$ be independent random variables. Let

$S = \frac{1}{n} \sum_{i=1}^{n} X_i$ and μ is expected value of S.

averaged values

$\mu = \mu_1 + \ldots + \mu_n$ → expected value of X_i.

We have

$$\Pr[S > \mu + \varepsilon] \leq e^{-2m\varepsilon^2}$$

→ larger m = smaller probability

→ larger ε = smaller probability

$$\Pr[S < \mu - \varepsilon] \leq e^{-2m\varepsilon^2}$$

$$\Pr[|S - \mu| > \varepsilon] \leq 2 \cdot e^{-2m\varepsilon^2}$$

Theorem: For any table T and function $f_1, \dots, f_m$, there exists table U with $\log(m)/\alpha^2$ records ↓

such that

$$\max_i |f_i(T) - f_i(U)| \le \alpha \longrightarrow \text{larger } \alpha = \text{smaller \# records}$$

Proof: We will construct U as follows:

Each record U_p is randomly picked from Table T, i.e.

$$\Pr[U_i = T_j] = \frac{1}{|T|} \quad \text{for all } i, j$$

$f_i(U_j)$ has expected value $\sum_j f_i(T_j)/n = f_i(T)$.

Let $X_j = f_i(U_j)/|U|$. The expected value of X_j is $f_i(T)/|U|$

$$S = f_i(U) = \sum_j \frac{f_i(U_j)}{|U|} = \sum_j X_j$$

Expected value of $S = f_i(T)$.

By Chernoff Bound

$$\Pr\{|f_i(U) - f_i(T)| > \alpha\} \le 2e^{-2|U|\alpha^2} \quad \text{for any } i$$

($f_i(U)$ = S)

$$\Pr\{\text{There exist } i \text{ such that } |f_i(U) - f_i(T)| > \alpha\} \le 2 \cdot m \cdot e^{-2|U|\alpha^2}$$

$$= 2 \cdot m \cdot \exp\left(-2 \cdot \frac{\log m}{\alpha^2} \cdot \alpha^2\right)$$

$$= 2 \cdot m \cdot [\exp(\log m)]^{-2}$$

($\exp(\log m) = m$)

$$= 2 \cdot m \cdot \frac{1}{m^2} \quad < 1 \text{ when } m > 2$$

~~$\Pr\{\text{There exist } i \text{ such that}$~~

$$\Pr\{|f_i(U) - f_i(T)| \le \alpha \text{ for all } i\} > 0$$

↑ It is possible to have such U

∎

There is table $\bar{U}$ such that

$$\max_i |f_i(T) - f_i(\bar{U})| \le \alpha$$

$$OPT \ge Utility[T, \bar{U}] = -\max_i |f_i(T) - f_i(\bar{U})| \ge -\alpha$$

$$OPT \ge -\alpha.$$

Suppose that T and T' are neighboring tables different at record j^*

$$f(T) = \sum_{j=1}^{n} f_i(T_j) / \|T\|$$

$$f(T') = \sum_{j=1}^{n} f_i(T'_j) / \|T'\|$$

$f_i \in [0,1]$

$$f(T) - f(T') = \frac{f_i(T_{j^*}) - f_i(T'_{j^*})}{\|T\|} \le 1/\|T\|$$

$$Utility(T, \bar{U}) - Utility(T', U) = |f_i(T) - f_i(\bar{U})| - |f(T') - f(\bar{U})|$$

$$\le |(f(T) - f(\bar{U})) - (f(T') - f(\bar{U}))|$$

$$= |f(T) - f(T')|$$

$$\le 1/\|T\|$$

$$\Delta Utility = 1/\|T\|$$

From last week,

$$Pr[E \le OPT - \frac{2\,\Delta Utility}{\varepsilon}(\ln(\#choices) + t)] \le e^{-t} = \beta$$

$\ln(\#choices) = \ln[(\#possible\ values)^{\frac{m}{\alpha^2}}]$

$-t = \ln\beta$

$t = -\ln\beta$

$$= \|T\| \cdot \ln(\#possible\ values) \cdot \ln m / \alpha^2$$

$$Pr[E \le -\alpha - \frac{2}{\varepsilon \cdot \|T\|}(\|T\| \cdot \ln(\#possible\ values)\frac{\ln m}{\alpha^2} - \ln\beta)] \le \beta$$

→ gap will be smaller when we have a larger database

$$\Pr\left[E \le -\alpha - \frac{2}{\varepsilon\|T\|}\left(\ln\frac{m}{\alpha^2} \cdot \ln(\#\text{possible values}) - \ln\beta\right)\right] \le \beta$$

(α → $\alpha/2$; α^2 → $\alpha^2/2$)

$-\alpha$

β

$$+\alpha = +\frac{\alpha}{2} + \frac{2}{\varepsilon\|T\|}\left(\frac{\ln m}{\alpha^2/4} \cdot \ln(\#\text{possible values}) - \ln\beta\right)$$

$$\frac{\alpha}{2} = \frac{2}{\varepsilon\|T\|}\left(\frac{\ln m}{\alpha^2/4} \cdot \ln(\#\text{possible values}) - \ln\beta\right)$$

$$\|T\| = \frac{4}{\alpha}\left(\frac{\ln m}{\alpha^2/4} \cdot \ln(\#\text{possible values}) - \ln\beta\right) \quad N$$

We will have $\Pr[E \le -\alpha] \le \beta$, when $\|T\| \ge N$.

Theorem. Using exponential mechanism over $\frac{\log m}{(2\alpha/2)^2}$-record table, we will have

ε-differential private; and, when $\|T\| \ge N$, we have

$$\Pr[E \le -\alpha] \le \beta.$$