THE UNIVERSITY OF
MELBOURNE

COMP20008
Elements of Data Processing

Semester 1 2020

Lecture 20: Differential Privacy – Local and Global

Plan today

- An introduction to differential privacy

What to do?

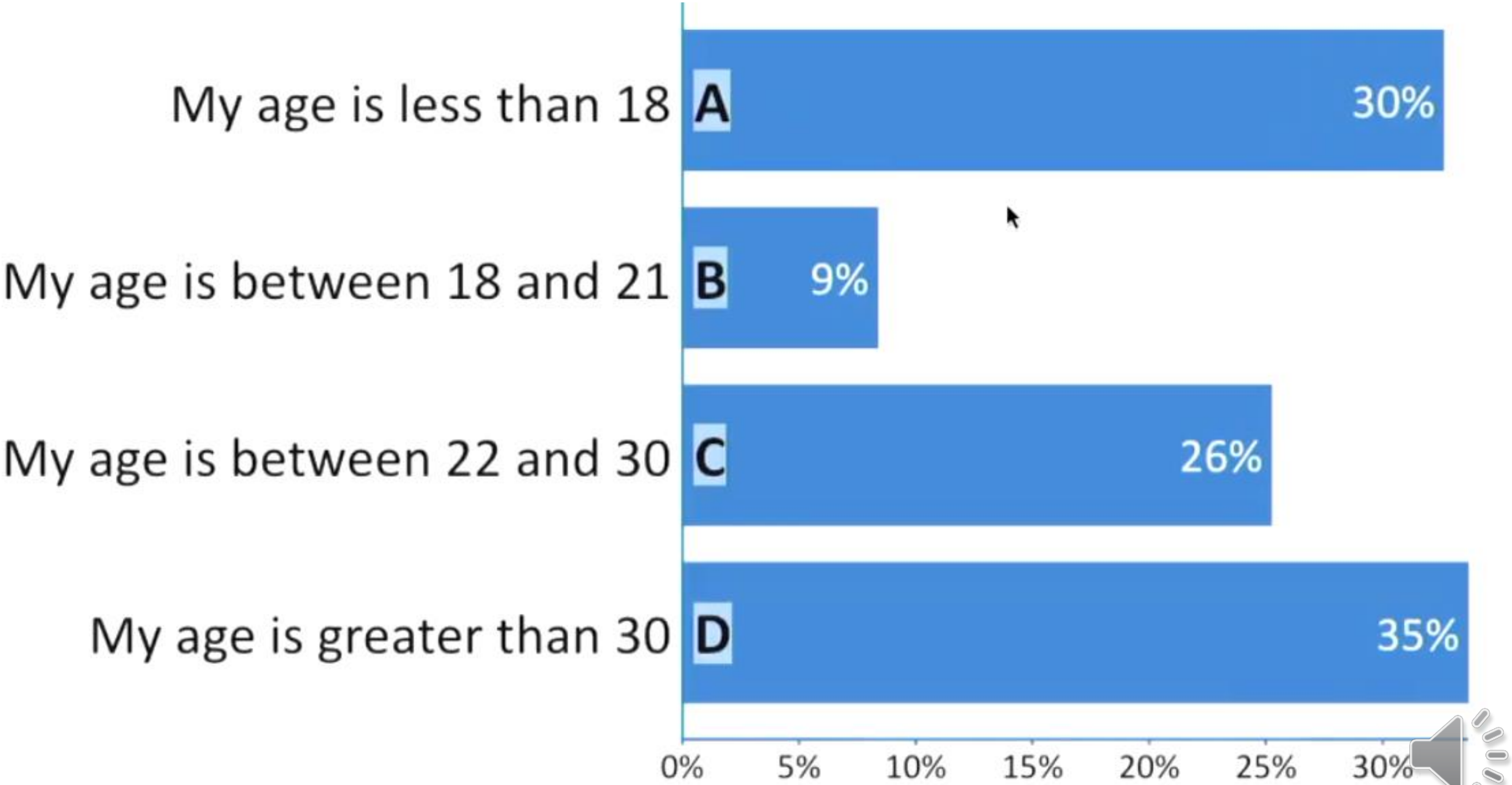
"The future of privacy is lying"

– (April 10 2013, Matt Buchanan, New Yorker)

A Simple Example

- Negative data survey – ask people to lie, and then make inferences based on the aggregate answers

Randomly select an option which is not true for you



Negative data surveys

- Participants select a choice that does not fit their situation
- Providing more choices provides more privacy
- May be challenging to design appropriate questions
- Reliance on honesty of the respondents
- This is an example of a local type of privacy, each person responsible for adding noise to their data

Differential privacy: Local and global

- **Global**: We have a sensitive dataset, a trusted data owner Alice and a researcher Bob. Alice does analysis on the raw data, adds noise to the answers, and reports the (noisy) answers to Bob

- **Local**: Each person is responsible for adding noise to their own data. Classic survey example each person has to answer question "Do you use drugs?"
  - They flip a coin in secret and answer "Yes" if it comes up heads, but tell the truth otherwise.
  - Plausible deniability about a "Yes" answer

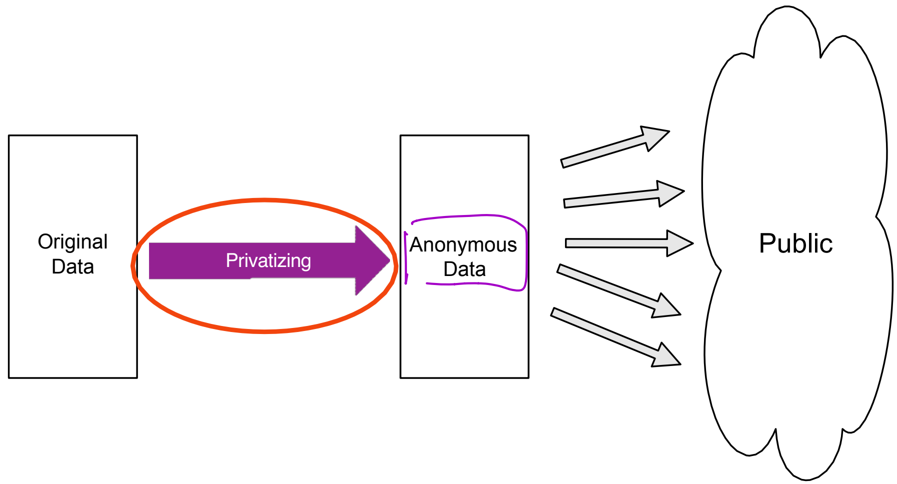
- We will next be looking further at the global case (global systems generally more accurate, and less noise is needed)

Differential privacy: Where?

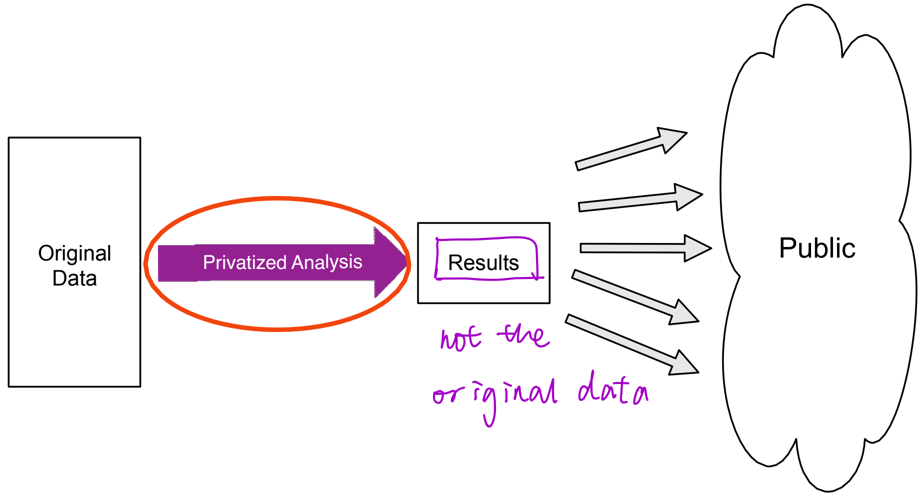
- Since its introduction in 2006:
  - US Census Bureau in 2012: *On The Map* project
    - Where people are employed and where they live
  - Apple in 2016: iOS 10
    - User data collection, e.g. for emoji suggestions
    - https://images.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
  - NSW Department of Transport open release of 2016 Opal ticketing system data
    - https://opendata.transport.nsw.gov.au/sites/default/files/resources/Open%20Opal%20Data%20Documentation%20170728.pdf

Global differential privacy: Our focus

k-anonymity
l-diversity



Differential privacy



What is being protected?

- Imagine a survey is asking you:
  - Are you a smoker?
    - Result: Number of smokers will be reported

Would you take part in it?

ID	Age	Sex	Smoker
sdhj5vbg	20	Male	False
wu234u4	25	Female	True
hi384yrh	17	Female	False
po92okwj	50	Male	False

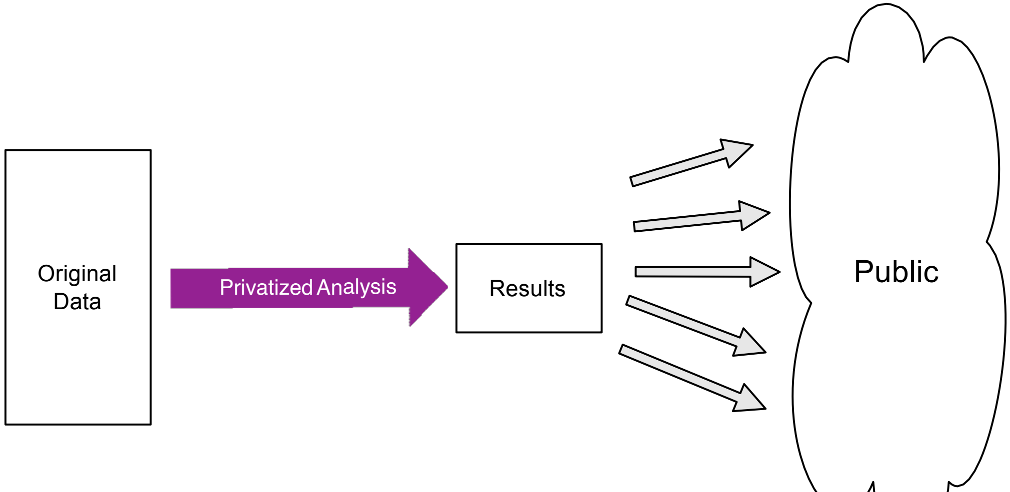
What is being protected?

I would feel safe submitting the survey if:

I know the chance that the privatized result would be R is nearly the same, whether or not I take part in the survey.

- Does this mean that an individual's answer has no impact on the released result?

Overview of the process: Global differential privacy



- The privatized analysis comprises two steps:
  - Query the data and obtain the real result, e.g., how many female students are in the survey?
  - Add random noise to hide the presence/absence of any individual. Release noisy result to the user.

$$D_{original} \xrightarrow{\text{Query}} R_{real} + \text{Noise} \longrightarrow R_{released}$$

The released results will be different each time (different amount of noised added)

- Query: How many females in the dataset? (true result = 32)
- Generate some random values, according to a distribution with mean value 0: {1,2,-2,-1,0,-3,1,0}, add to true result and release
  - 1st query: Released result=33 (32+1)
  - 2nd query: Released result=34 (32+2)
  - 3rd query: Released result=30 (32-2)
  - 4th query: Released result=31 (32-1)
  - 5th query: Released result=32 (32+0)
  - 6th query: Released result=29 (32-3)
  - 7th query: Released result=33 (32+1)
  - 8th query: Released result=32 (32,0)
  - …
- On average, the released result will be 32, but observing a single released result doesn't give the adversary exact knowledge

Emoji scenario and use of differential privacy

- A developer wants to understand which emoji's are popular, in order to make better recommendations. There is a database like

User	Emoji used today
Bob	
Alice	
Sarah	
Rudolph	
Cameron	

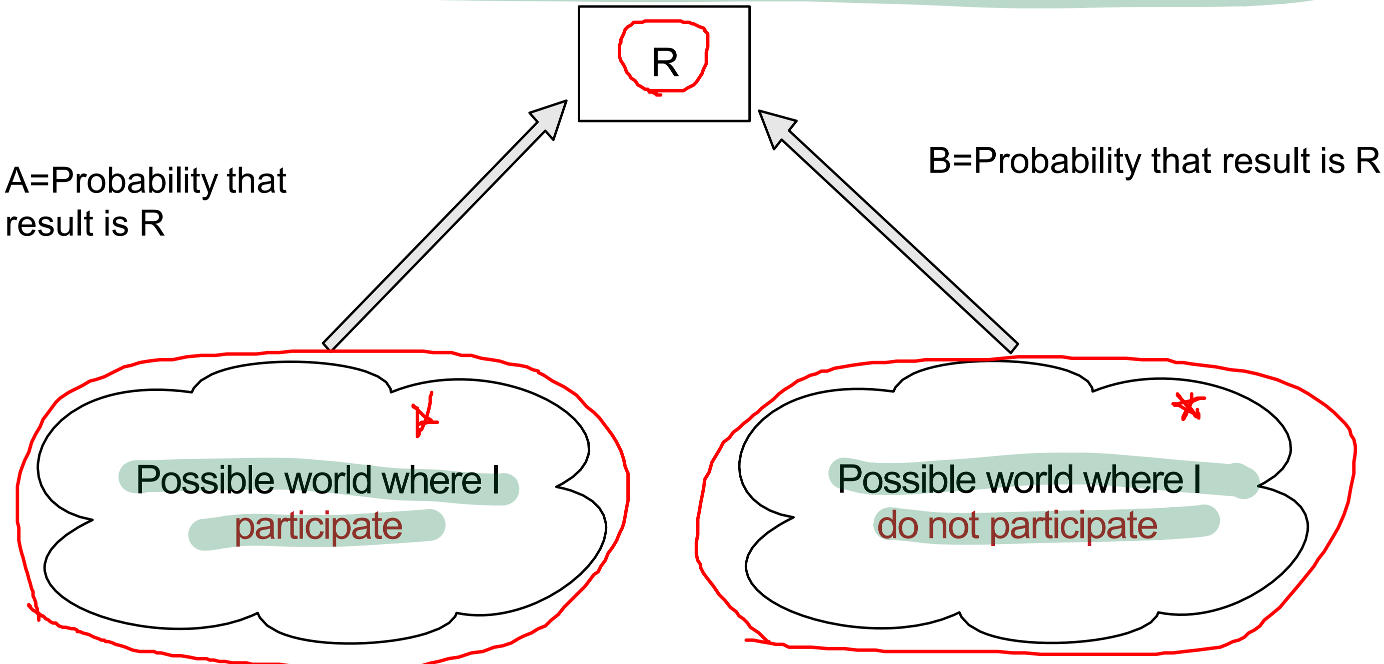
- Query from developer: How many times was  used today?

- System will release a noisy result to developer, to protect customer privacy

THE UNIVERSITY OF MELBOURNE

The promise of differential privacy

- The chance that the noisy released result will be R is nearly the same, whether or not an individual participates in the dataset.



- If we can guarantee A≅B (A is very close to B), then no one can guess which possible world resulted in R.

The promise of differential privacy

- Does this mean that the attacker cannot learn anything sensitive about individuals from the released results?

Differential privacy: How?

- How much noise should we add to the result? This depends on
  - **Privacy loss budget:** How private we want the result to be (how hard for the attacker to guess the true result)
  - **Global sensitivity:** How much difference the presence or absence of an individual could make to the result.

the more noice add → the harder for people to find out private information
→ the less useful for the predict

Global sensitivity

- Global sensitivity of a query Q is the maximum difference in answers that adding or removing any individual from the dataset can cause (maximum effect of an individual)

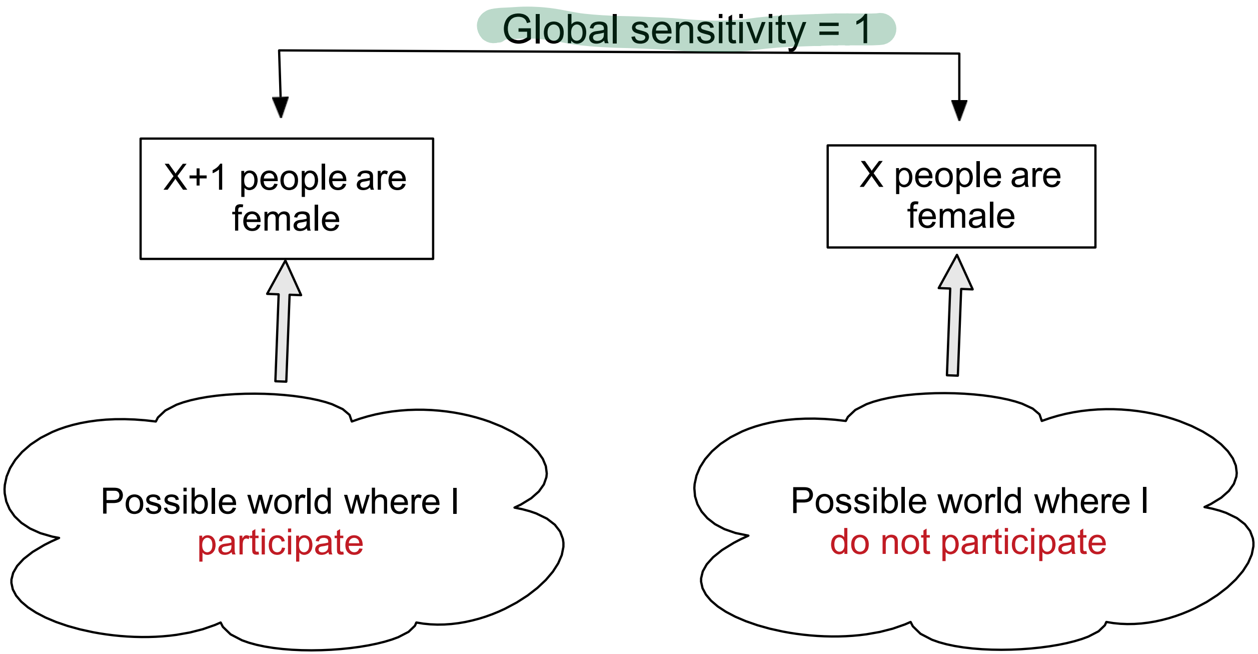
- Intuitively, we want to consider the worst case scenario

what is the maximum effect when add an individual)

- If asking multiple queries, global sensitivity is equal to the sum of the differences

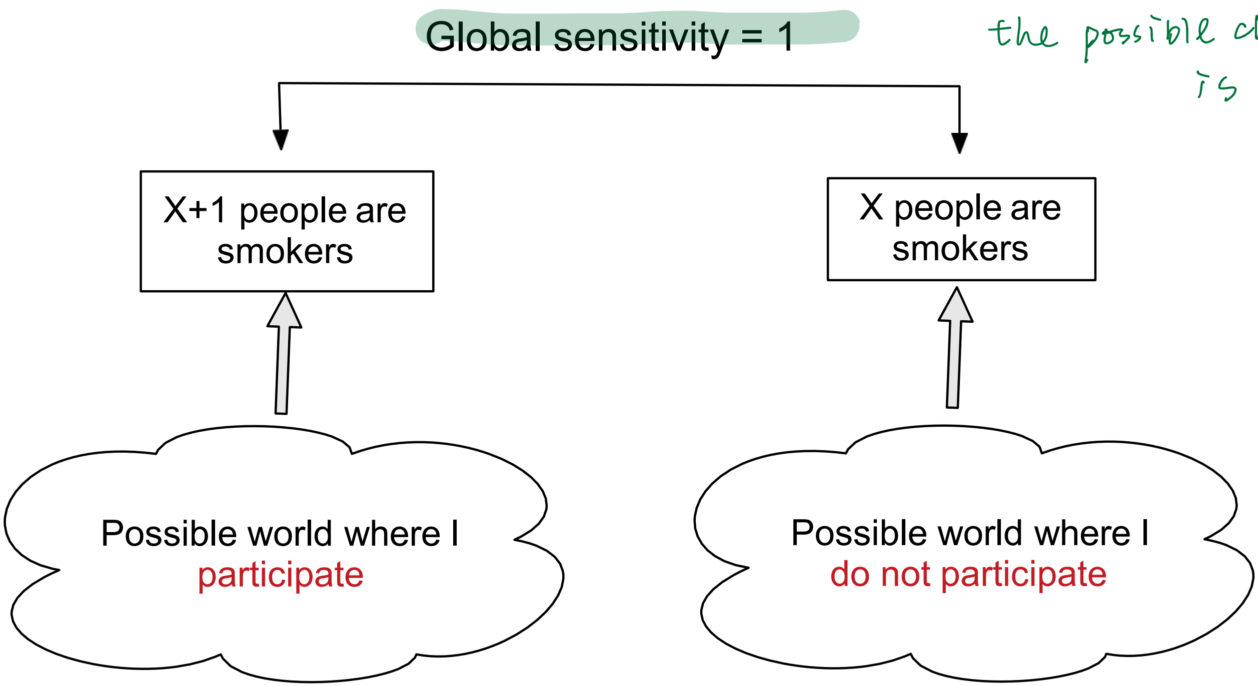
Global sensitivity

- QUERY: How many people in the dataset are female?



Global sensitivity

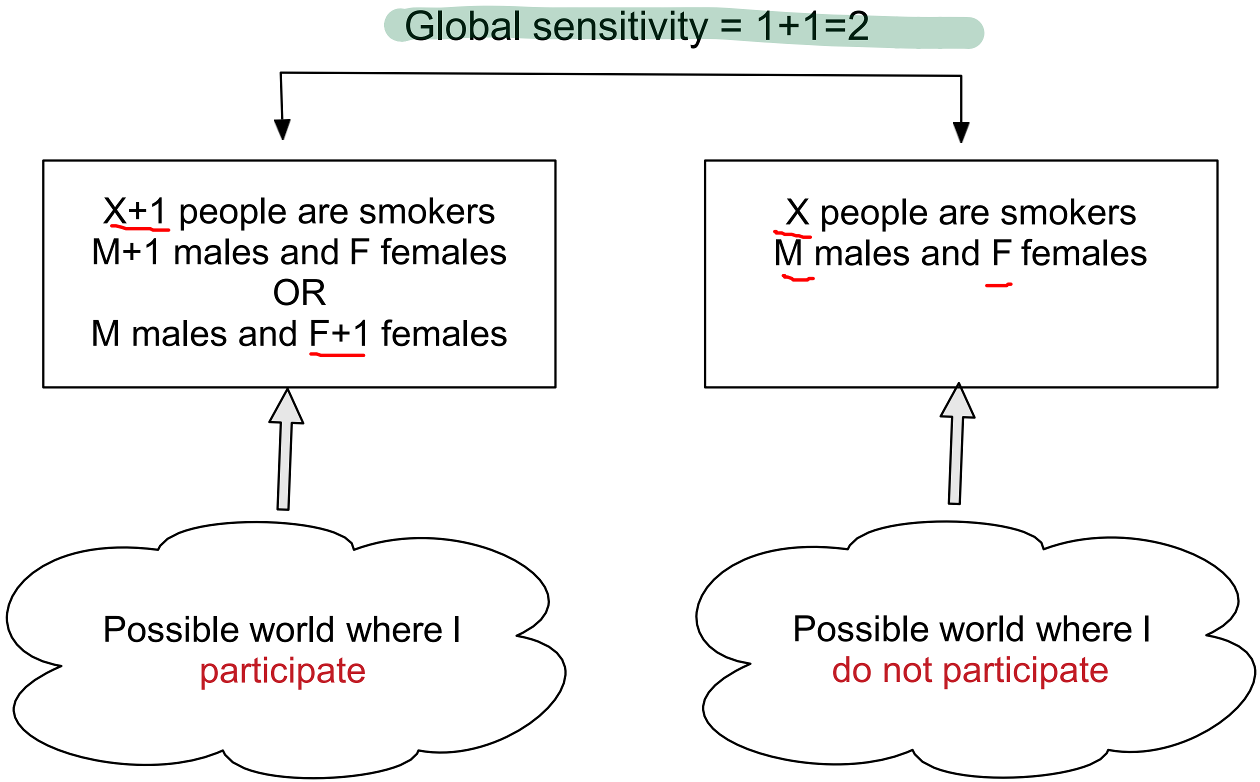
- QUERY: How many people in the dataset are smokers?



the possible change is 1

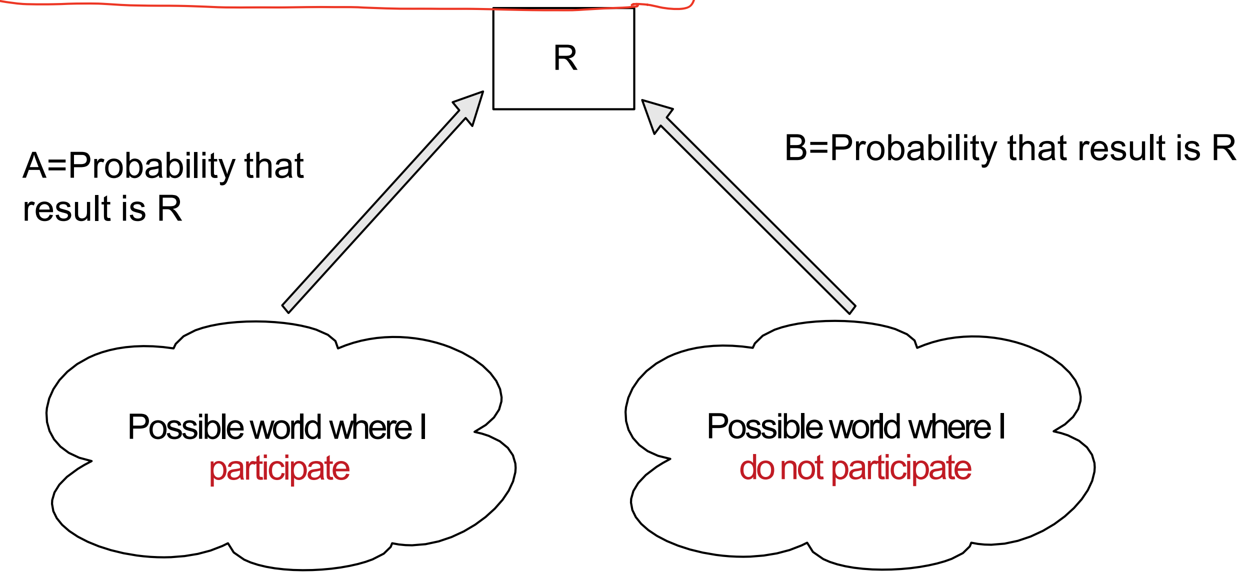
Global sensitivity

- QUERY: How many people in the dataset are female? And how many people are smokers?



Privacy loss budget = k

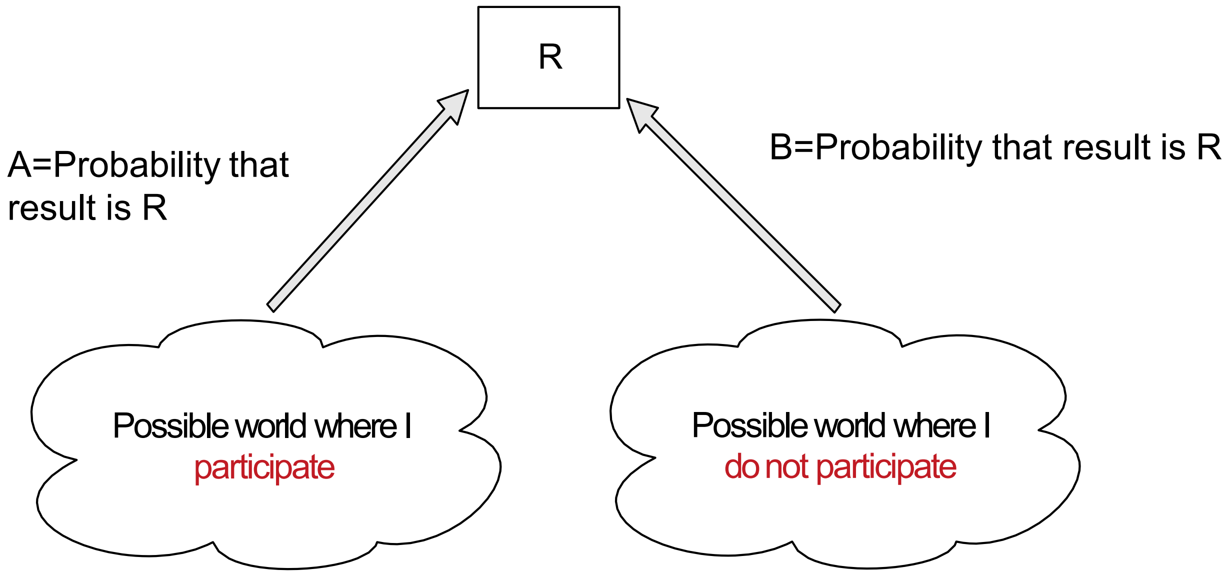
- We want that the presence or absence of a user in the dataset does not have a *considerable effect* on the released result



Privacy loss budget = k ($k \geq 0$)

Choose k to guarantee that $A \leq 2^k \times B$

Privacy loss budget = k



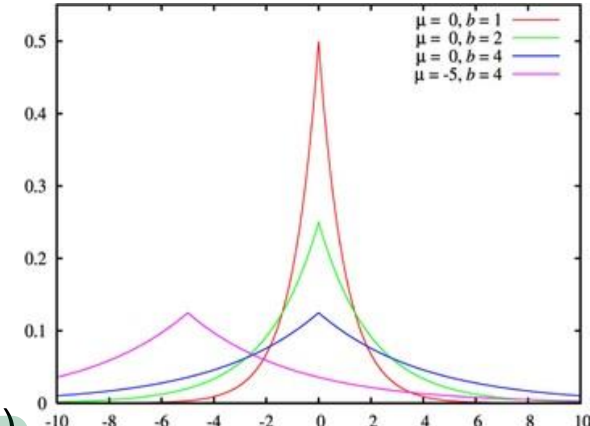
Privacy loss budget=k (k ≥0)

Choose k to guarantee that $A \leq 2^k \times B$

- k=0: No privacy loss (A=B), low utility no much information, useless
- k=high: Larger privacy loss, higher utility
- k=low: Low privacy loss, lower utility

Differential privacy: How?

- How much noise should we add to the result? This depends on
  - **Privacy loss budget (k):** How private we want the result to be (how hard for the attacker to guess the true result)
  - **Global sensitivity (G):** How much difference the presence of absence of an individual could make to the result.

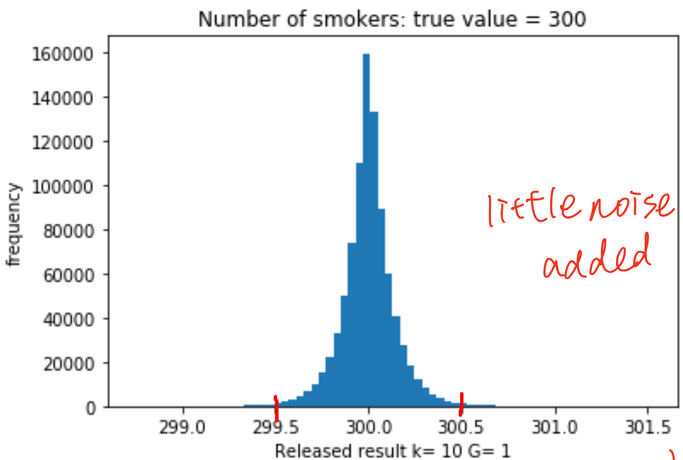


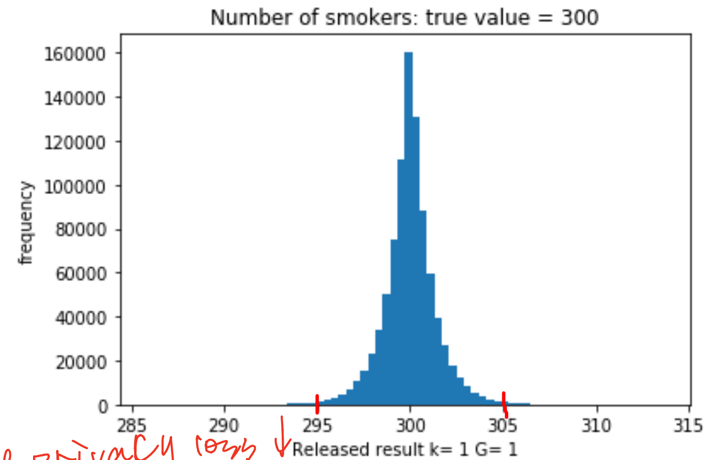
- Strategy: Add noise to the result according to
  - Released result = True result + noise
    - Where noise is a number randomly sampled from a distribution having
      remain same mean
      - average value = 0 (μ)
      - standard deviation (spread)= G/k (b)
    - Details about the distribution are beyond the scope of our study (it is called the Laplace distribution)

Example Code

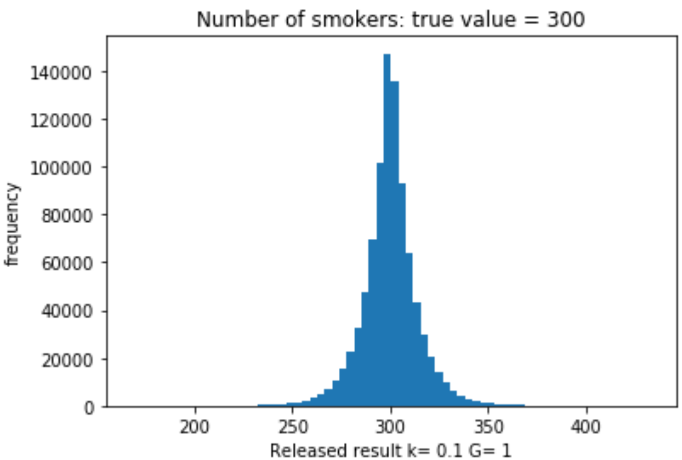
```
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline

G = 1
k = 10

deviation = G/k
loc,scale = 300., deviation

s = np.random.laplace(loc, scale, 900000)
plt.hist(s,70)
plt.ticklabel_format(useOffset=False)

plt.xlabel('Released result k= ' + str(k) + ' G= ' + str(G))
plt.ylabel('frequency')
plt.title('Number of smokers: true value = 300')
plt.show()
```

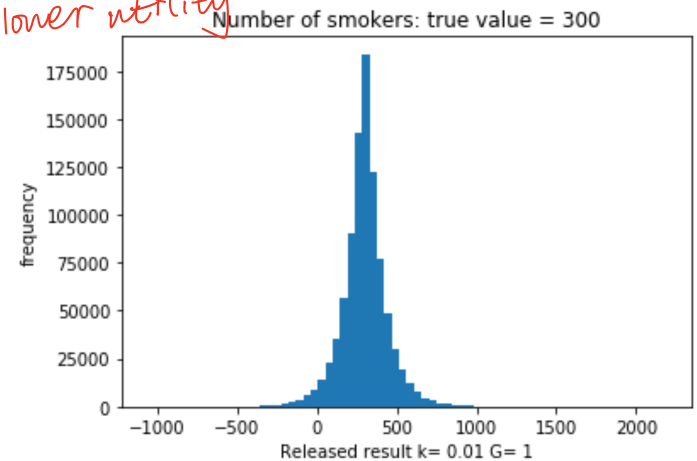
Example





k↓, the privacy loss ↓
but lower utility





Summary

- Differential privacy guarantees that the presence or absence of a user cannot be revealed after releasing the query result
  - It does not prevent attackers from drawing conclusions about individuals from the aggregate results over the population
- We need to determine the budget and global sensitivity to know what is the scale of the noise to be added

Summary

- Differential privacy guarantees that the presence or absence of a user cannot be revealed after releasing the query result
  - It does not prevent attackers from drawing conclusions about individuals from the aggregate results over the population

- We need to determine the budget and global sensitivity to know what is the scale of the noise to be added