COMP20008 Elements of Data Processing

Differential Privacy

Announcements

- Phase 4 marks will be released today
- The exam study guide has been updated. It covers up to lecture 20
- A sample exam and sketch answers is also now provided

Plan today

- Recap of *k*-anonymity and *l*-diversity
  - Concept
  - Homogeneity and background attack
  - Location/trajectory privacy
- An introduction to differential privacy

k-anonymity recap

- Data owner determines quasi identifier(s)
- Data owner or individuals choose parameter *k*

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

l-Diversity: Privacy Beyond k-Anonymity. Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam, 2007

l-diversity

- To protect privacy against
  - Homogeneity attack
  - Background knowledge attack

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	$*$	Heart Disease
4	1305*	≤ 40	$*$	Viral Infection
9	1305*	≤ 40	$*$	Cancer
10	1305*	≤ 40	$*$	Cancer
5	1485*	> 40	$*$	Cancer
6	1485*	> 40	$*$	Heart Disease
7	1485*	> 40	$*$	Viral Infection
8	1485*	> 40	$*$	Viral Infection
2	1306*	≤ 40	$*$	Heart Disease
3	1306*	≤ 40	$*$	Viral Infection
11	1306*	≤ 40	$*$	Cancer
12	1306*	≤ 40	$*$	Cancer

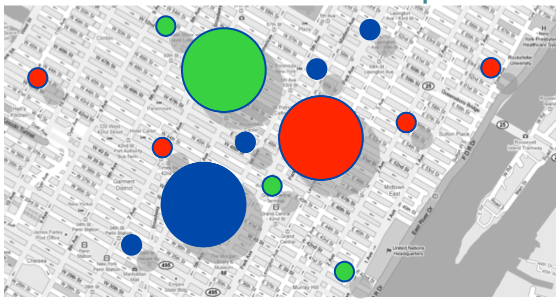
Overview of Location Privacy Models

- Location privacy
  - *k*-anonymity
    - if individuals' location information cannot be distinguished from $k-1$ other individuals
  - Obfuscation
    - The greater the imperfect knowledge about a user's location, the greater the user's privacy

Exact location points

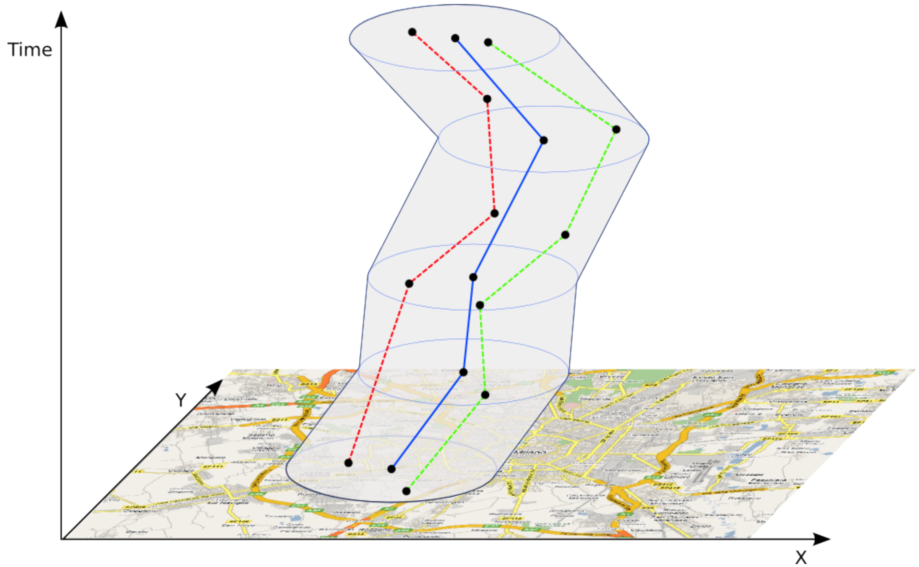
3-anonymized location points

Obfuscated location points



Trajectory Privacy

- Clustering k similar trajectories:
  - At each timestamp a point with the least distance to all trajectories is reported



- **Question:**
  - Shortcomings of trajectory cloaking?

Summary

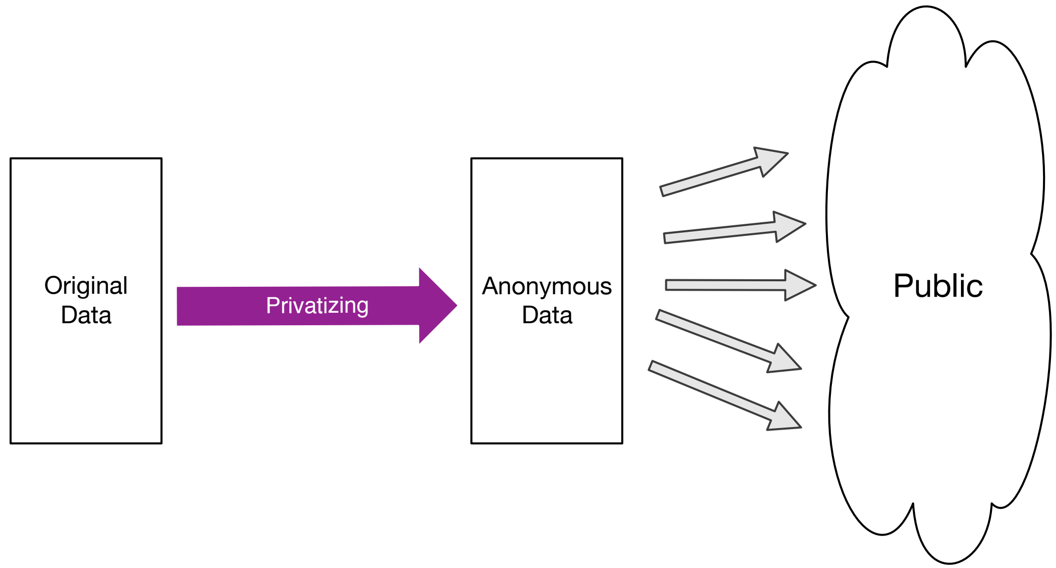
- To reduce risk of re-identification of individuals in released datasets
  - Choose value of *k*
  - Manipulate data to make it *k*-anonymous, either
    - Replace categories by broader categories
    - Suppress attributes with a * (limited utility)
  - Further manipulate data to make it *l*-diverse
    - Ensure there are at least *l* different values of the sensitive attribute in each group
- Privacy is difficult to maintain in high-dimensional datasets like trajectory datasets
  - Cloaking provides spatial *k*-anonymity
  - Obfuscation ensures location imprecision

Differential Privacy: why?

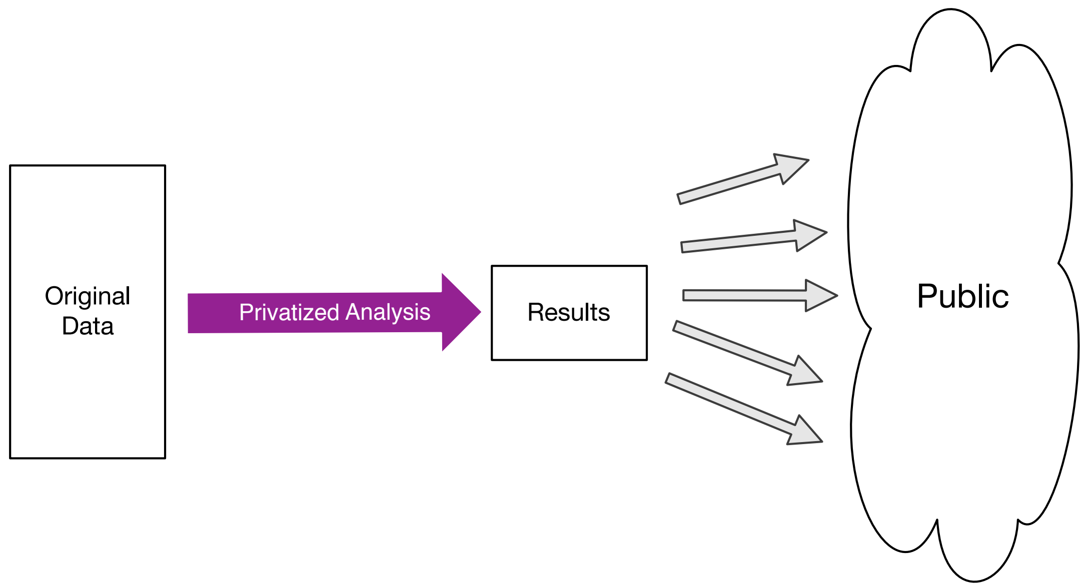
- Since its introduction in 2006:
  - US Census Bureau in 2012: *On The Map* project
  - Google in 2014 and 2015: private collection of telemetry and private release of snapshots of traffic
  - Apple in 2016: iOS 10

Differential Privacy: Our Focus

k-anonymity
l-diversity



Differential privacy



What is being protected?

- Imagine a survey is asking you:
  - How old are you?
  - What is your gender?
  - Are you a smoker?

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

- Would you take part in it?

What is being protected?

I would feel safe submitting the survey if:

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

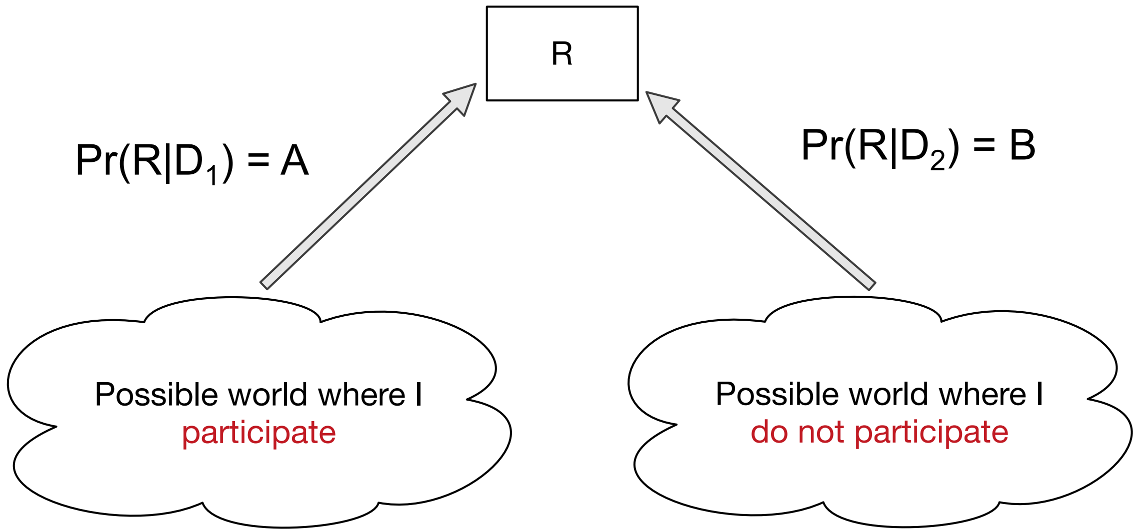
- Does this mean that an individual's answer has no impact on the release result?

Quick Reminder

- Conditional probability
  - Probability of an event given another event has happened
- Notation:
  - Pr(A|B)
  - For example Pr(rain) versus Pr(rain|winter)

The Promise of Differential Privacy

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



- If we guarantee A≅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?

- Two key concept:
  - Two datasets with or without an individual -> neighboring datasets and global sensitivity
  - Probability of having nearly the same result -> privacy budget

Global sensitivity

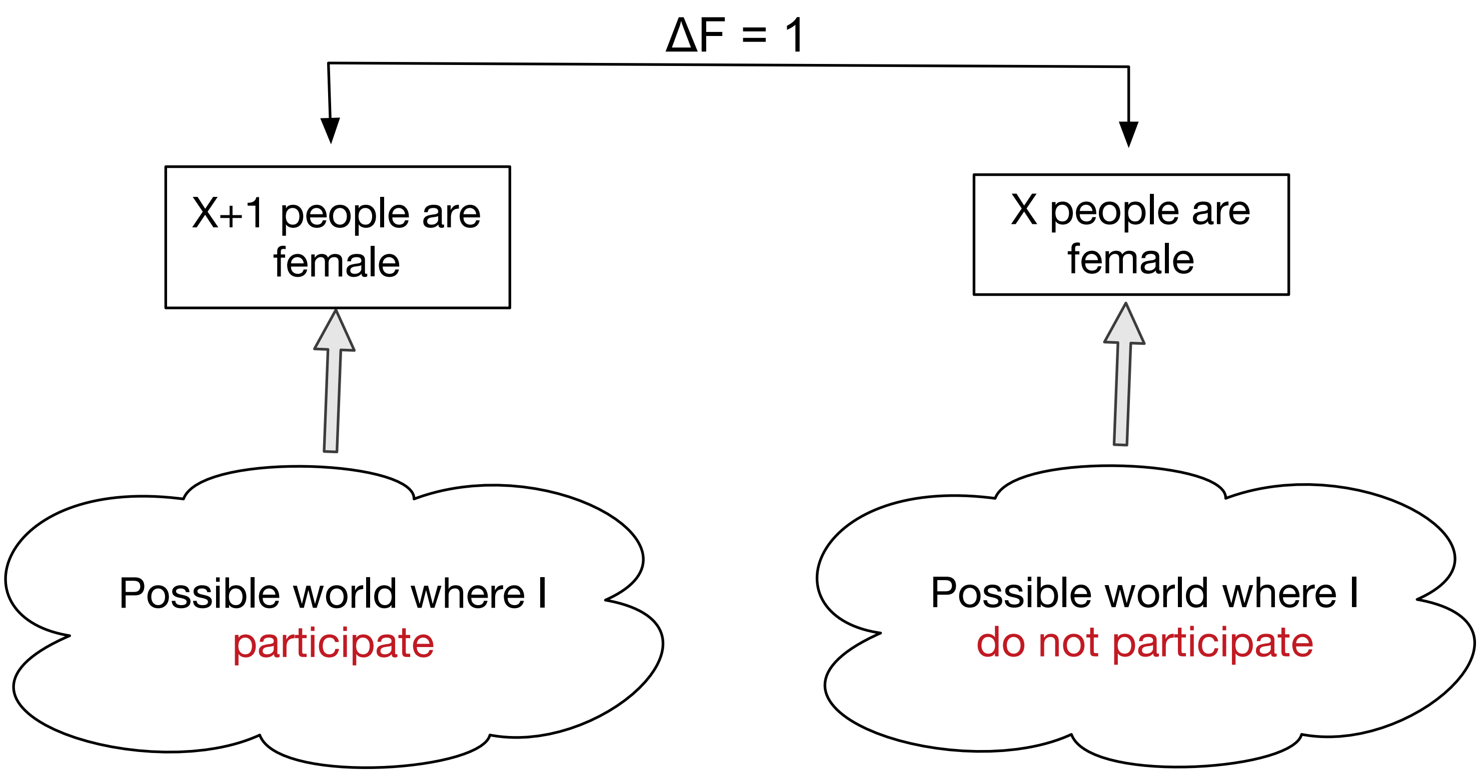
- Global sensitivity of a function (query) is the maximum difference in answers that adding or removing any individual from the dataset can cause

$$\Delta F = \max_{D_1, D2} ||F(D_1) - F(D_2)||$$

  - Intuitively, we want to consider the worst case scenario
  - If asking multiple queries, global sensitivity is equal to the sum of the differences

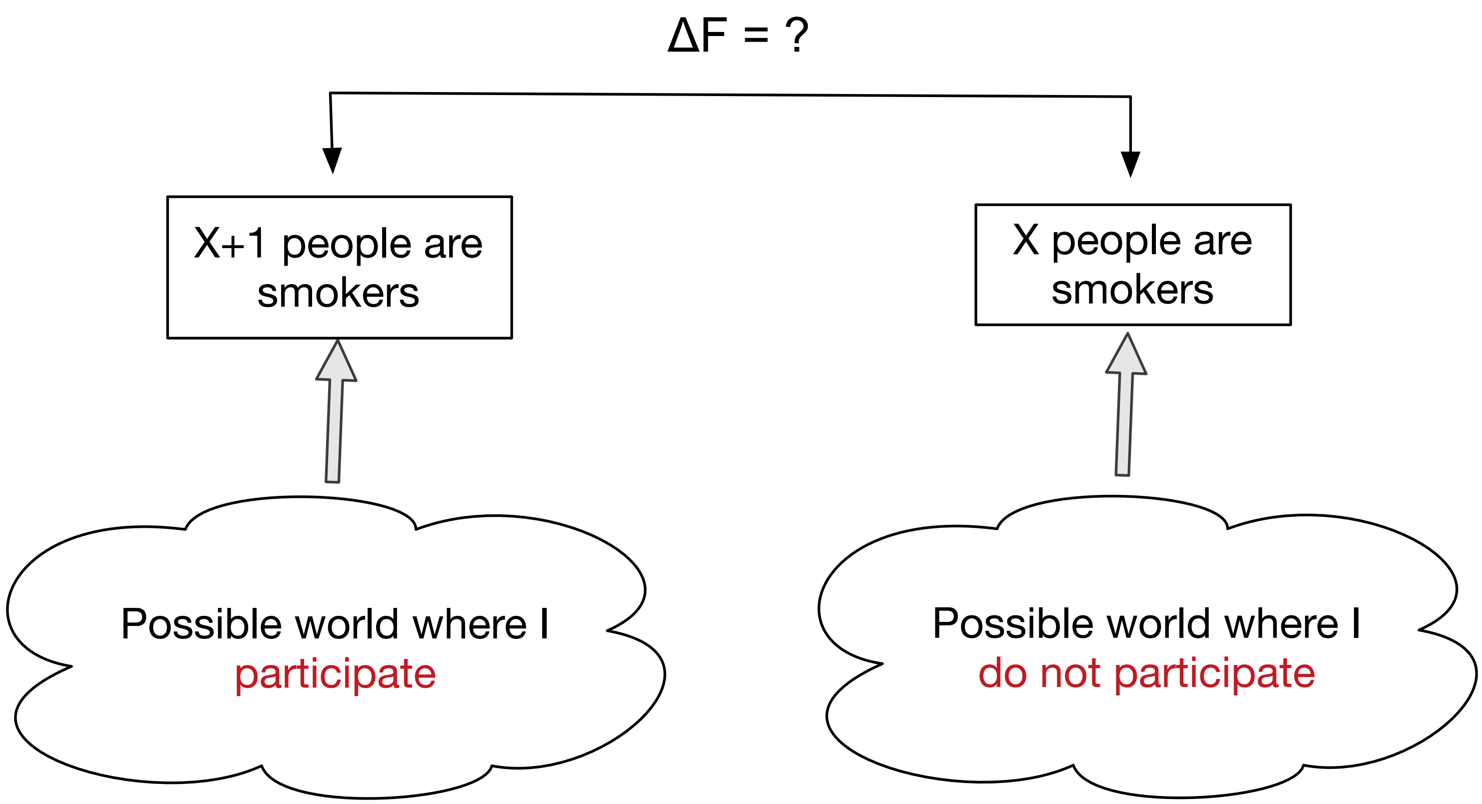
Global Sensitivity

- How many people in the dataset are female?



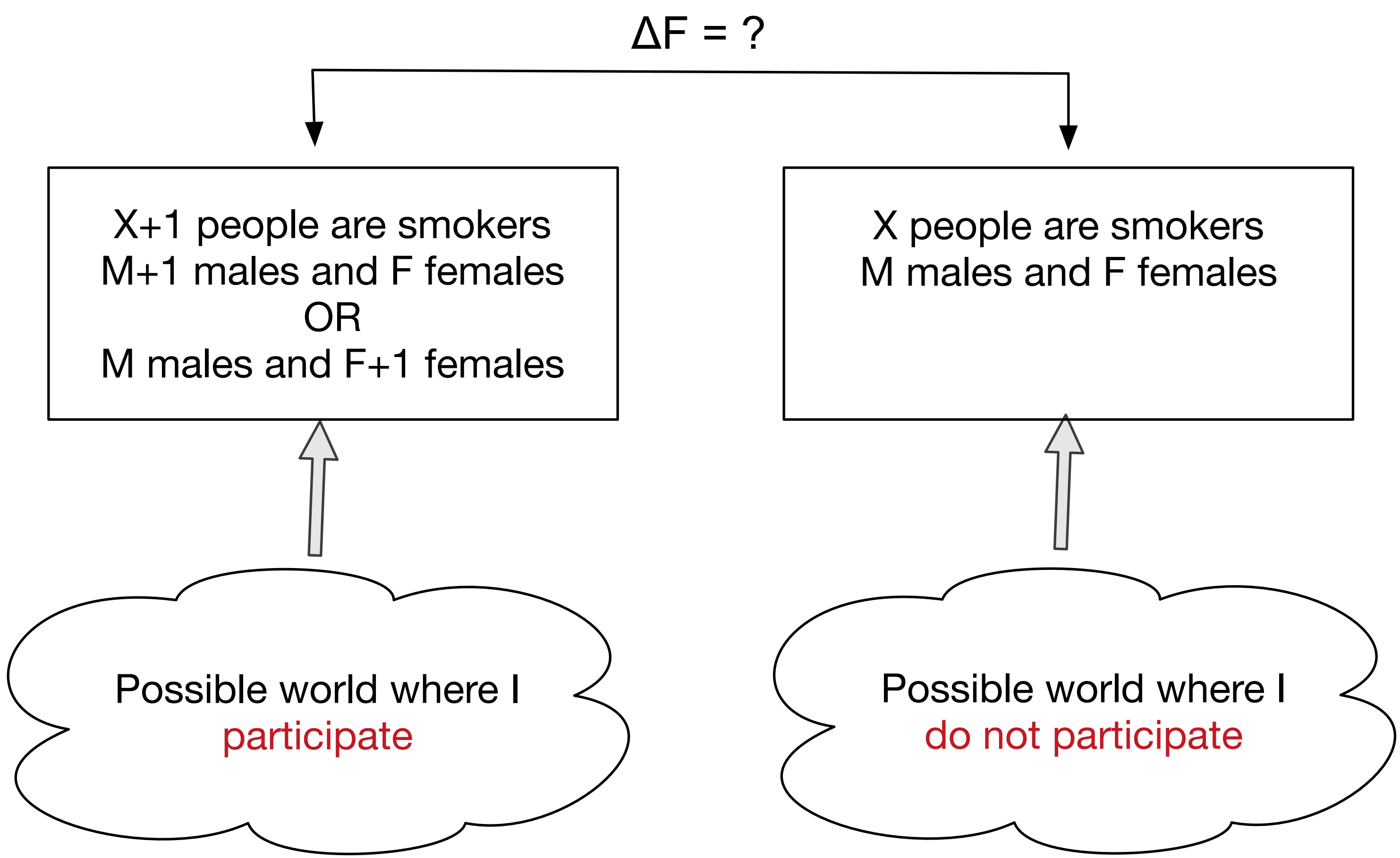
Global Sensitivity

- How many people in the dataset are smokers?



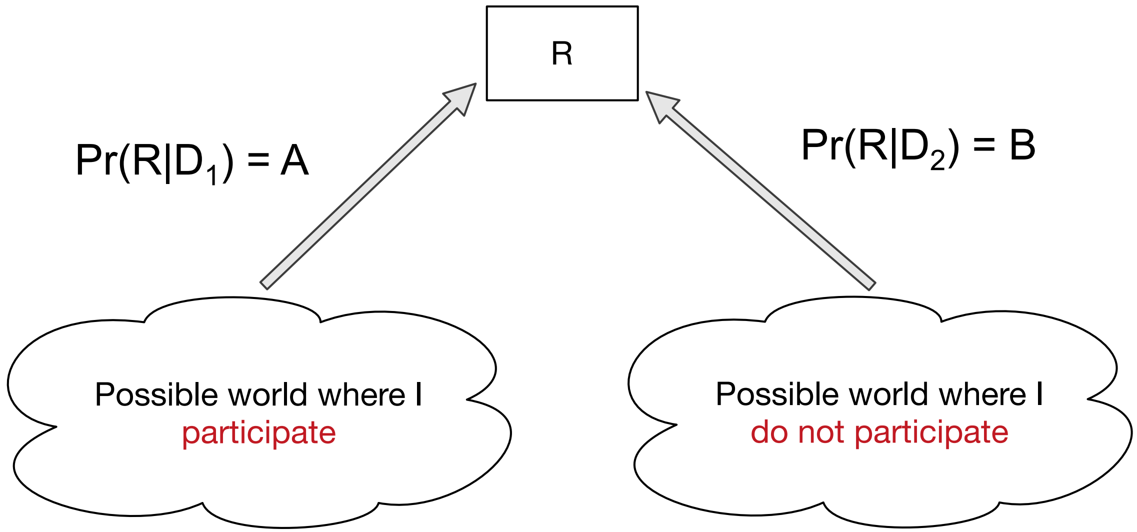
Global Sensitivity

- How many people in the dataset are male or female? And how many people are smokers?



Privacy Budget

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



- Privacy budget, denoted as ε determines how close the chance of having R is:

$$Pr(R|D_1) \leq e^{\epsilon} Pr(R|D_2)$$

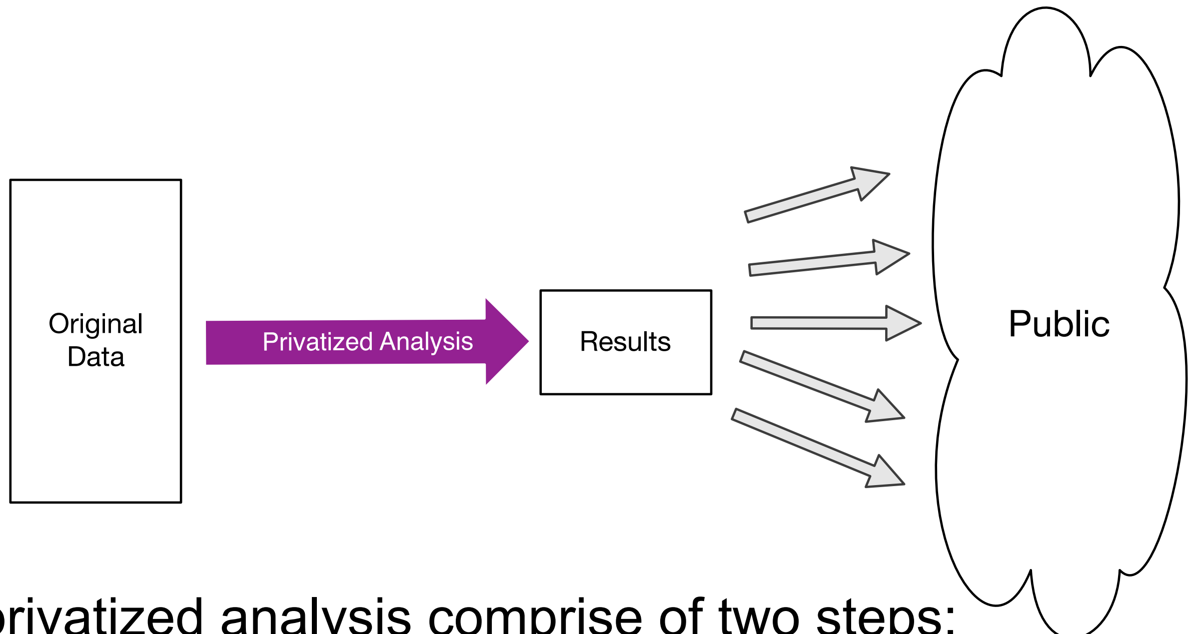
Privacy Budget

- Intuitively, the privacy budget determines how strict we are

$$Pr(R|D_1) \leq e^{\epsilon} Pr(R|D_2)$$

- What does a privacy budget of ε = 0 imply?

Putting it All Together



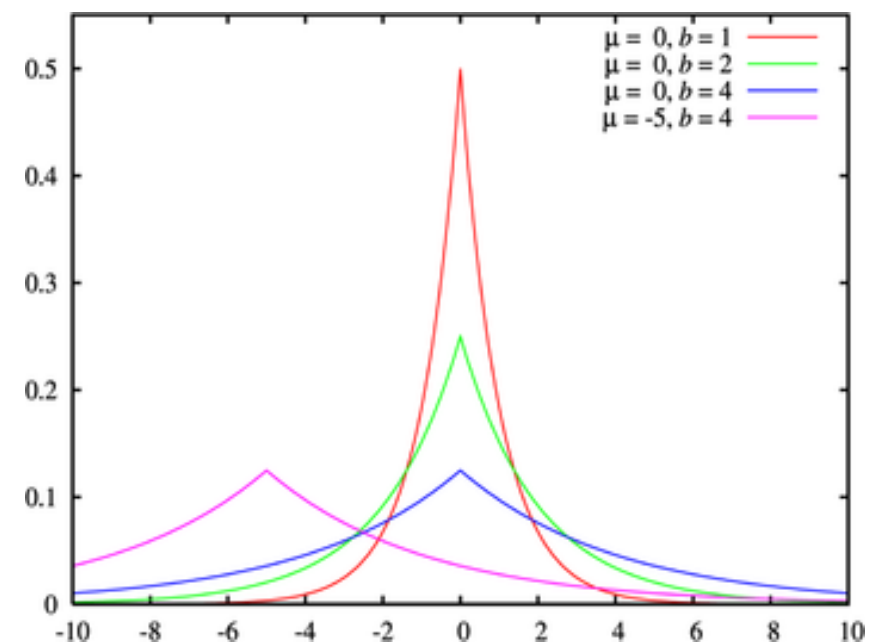
- The privatized analysis comprise of two steps:
  - Query the data and obtain the real result, e.g., how many female students are in the survey?
  - Add noise to hide the presence/absence of any individual

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

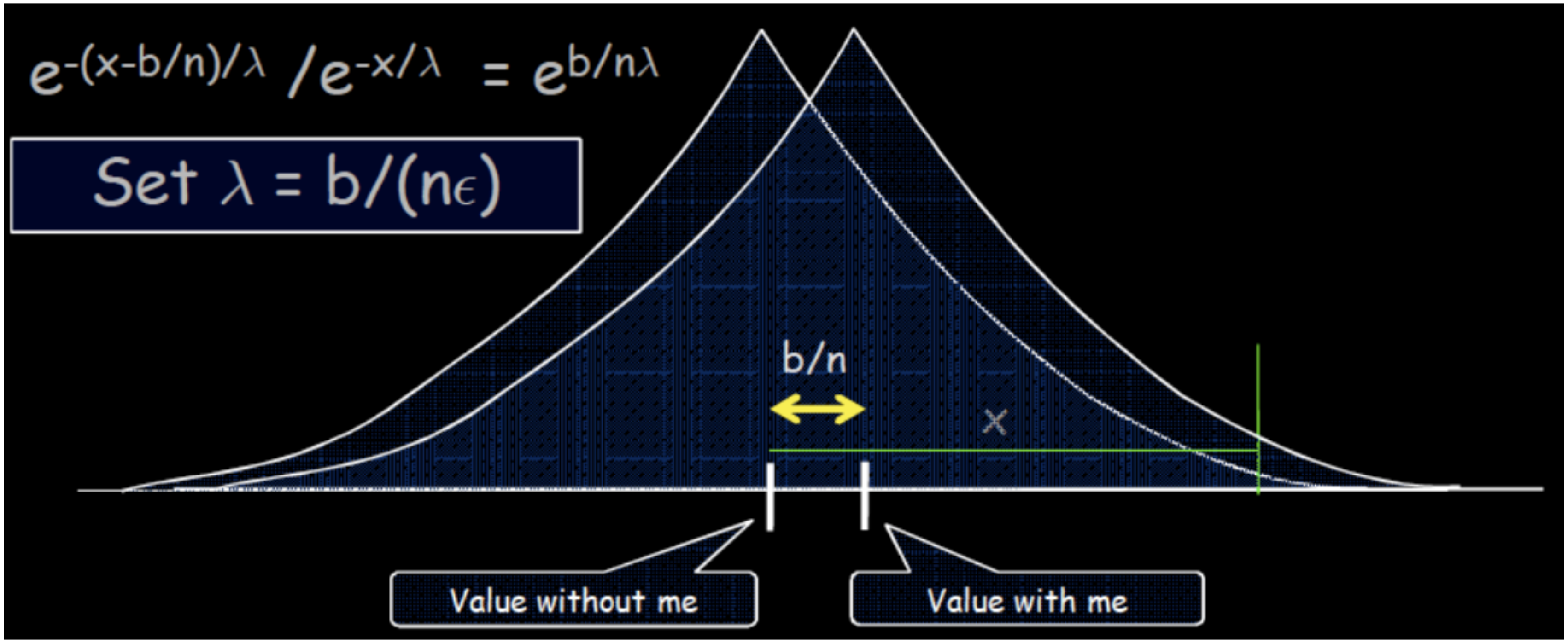
How much noise?

- We add noise to the real result of the query to span the sensitivity gap
- What noise?
  – Random value drawn from a Laplacian distribution
  – Mean zero to be close to the real result
  – Standard deviation large enough to cover the gap: ΔF/ε

μ : mean
b : standard deviation



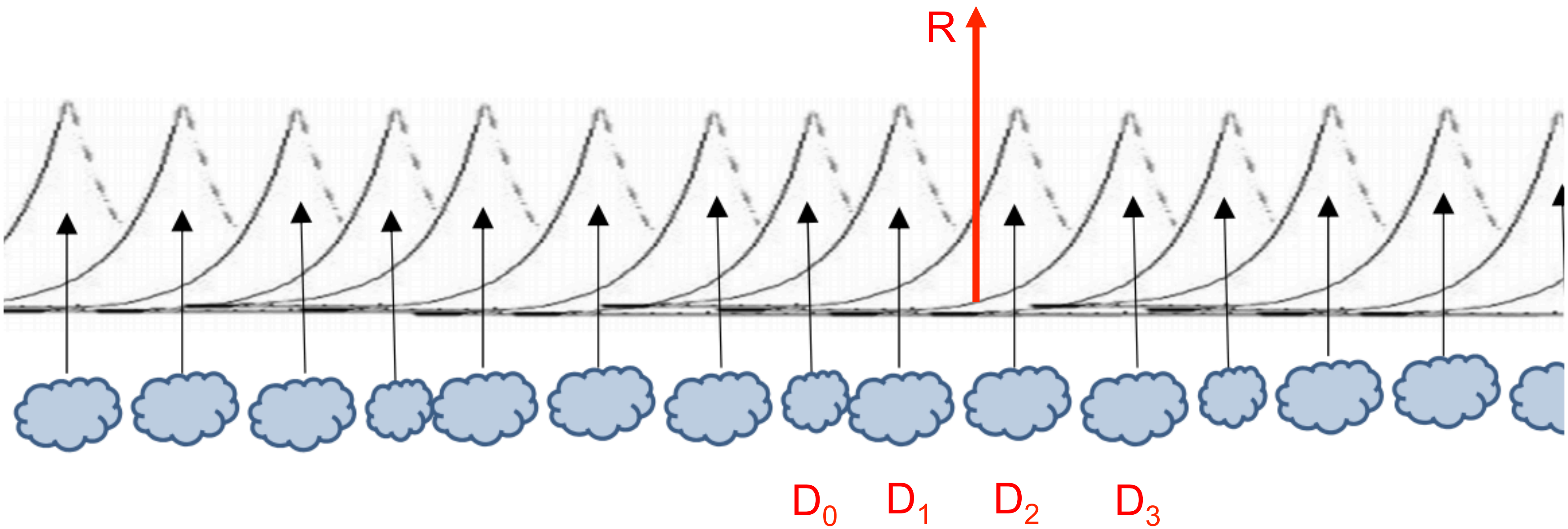
How the noise works?



https://www.cs.unb.ca/~rlu1/slide/20150120guomingwang.pdf

How the noise works?

- By looking at the result R
  - Very difficult to guess which world it came from and who was exactly in the dataset
  - General neighborhood of the actual answer is preserved for 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

THE UNIVERSITY OF MELBOURNE

Acknowledgements

This lecture was prepared using some material adapted from:

- Masachusette story
  - https://epic.org/privacy/reidentification/ohm_article.pdf
- From a social science perspective
  - http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006
- *l*-diversity
  - https://www.cs.cornell.edu/~vmuthu/research/ldiversity.pdf
- A Practical Beginner's Guide to Differential Privacy – Christine Task
- Location and Trajectory Privacy – Lars Kulik