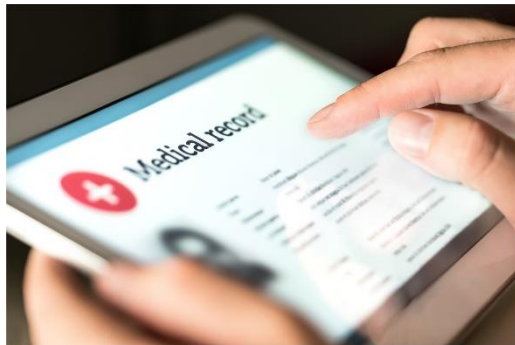




# Introduction to Quantitative Privacy

---


# Data privacy



# A toy example: Publishing medical data

	Non-Sensitive		Sensitive
Name	Zip	Age	Condition
Jack	13053	28	Flu
Peter	13068	29	Flu
Alice	13068	21	Diabetes
Robert	13053	23	Diabetes
James	14853	50	Cancer
Kevin	14853	55	Flu
Lucas	14850	47	Diabetes
Ella	14850	49	Diabetes
Jennifer	13053	31	Cancer
Emily	13053	37	Cancer
Emma	13068	36	Cancer
Jason	13068	35	Cancer

# A toy example: Publishing medical data

	Non-Sensitive		Sensitive
Name	Zip	Age	Condition
Jack	13053	28	Flu
Peter			Flu
Alice			Diabetes
Robert			Diabetes
James			Cancer
Kevin			Flu
Lucas			Diabetes
Ella			Diabetes
Jennifer			Cancer
Emily	13053	37	Cancer
Emma	13068	36	Cancer
Jason	13068	35	Cancer

# Give a try

	Non-Sensitive		Sensitive
Name	Zip	Age	Condition
Jack	13053	28	Flu
Peter	13068	29	Flu
Alice	13068	21	Diabetes
Robert	13053	23	Diabetes
James	14853	50	Cancer
Kevin	14853	55	Flu
Lucas	14850	47	Diabetes
Ella	14850	49	Diabetes
Jennifer	13053	31	Cancer
Emily	13053	37	Cancer
Emma	13068	36	Cancer
Jason	13068	35	Cancer

Remove  
names  
⇒

	Non-Sensitive		Sensitive
	Zip	Age	Condition
1	13053	28	Flu
2	13068	29	Flu
3	13068	21	Diabetes
4	13053	23	Diabetes
5	14853	50	Cancer
6	14853	55	Flu
7	14850	47	Diabetes
8	14850	49	Diabetes
9	13053	31	Cancer
10	13053	37	Cancer
11	13068	36	Cancer
12	13068	35	Cancer

# Not enough

	Non-Sensitive		Sensitive
	Zip	Age	Condition
1	13053	28	Flu
2	13068	29	Flu
3	13068	21	Diabetes
4	13053	23	Diabetes
5	14853	50	Cancer
6	14853	55	Flu
7	14850	47	Diabetes
8	14850	49	Diabetes
9	13053	31	Cancer
10	13053	37	Cancer
11	13068	36	Cancer
12	13068	35	Cancer

Linkage  
attack

Name	Zip	Age	Sport
Peter	13068	29	Tennis
John	13068	25	Soccer
Theo	14851	21	Soccer
Betty	14853	32	Tennis
Lucas	14850	47	Golf
Frank	13068	56	Snooker
Ben	13068	36	Golf

# $k$ -anonymity

---

- An equivalence class is a set of records having the same non-sensitive attributes.
- A table is  $k$ -anonymous if each equivalence class has at least  $k$  individuals.
- Suppression: replace part of a non-sensitive attribute by '\*'.
- Generalization: replace a non-sensitive attribute by a broader category.

# $k$ -anonymity

	Non-Sensitive		Sensitive
	Zip	Age	Condition
1	13053	28	Flu
2	13068	29	Flu
3	13068	21	Diabetes
4	13053	23	Diabetes
5	14853	50	Cancer
6	14853	55	Flu
7	14850	47	Diabetes
8	14850	49	Diabetes
9	13053	31	Cancer
10	13053	37	Cancer
11	13068	36	Cancer
12	13068	35	Cancer

4-  
anonymous  
⇒

	Non-Sensitive		Sensitive
	Zip	Age	Condition
1	130**	<30	Flu
2	130**	<30	Flu
3	130**	<30	Diabetes
4	130**	<30	Diabetes
5	1485*	>40	Cancer
6	1485*	>40	Flu
7	1485*	>40	Diabetes
8	1485*	>40	Diabetes
9	130**	3*	Cancer
10	130**	3*	Cancer
11	130**	3*	Cancer
12	130**	3*	Cancer



# $k$ -anonymity

	Non-Sensitive		Sensitive
	Zip	Age	Condition
1	130**	<30	Flu
2	130**	<30	Flu
3	130**	<30	Diabetes
4	130**	<30	Diabetes
5	1485*	>40	Cancer
6	1485*	>40	Flu
7	1485*	>40	Diabetes
8	1485*	>40	Diabetes
9	130**	3*	Cancer
10	130**	3*	Cancer
11	130**	3*	Cancer
12	130**	3*	Cancer

May leak sensitive attributes:

$(130 **, 3 *) \Rightarrow \text{Cancer}$



# $\ell$ -diversity

---

- A table is  $\ell$ -diverse if each equivalence class has at least  $\ell$  different values for all sensitive attributes.

# $\ell$ -diversity

	Non-Sensitive		Sensitive
	Zip	Age	Condition
1	13053	28	Flu
2	13068	29	Flu
3	13068	21	Diabetes
4	13053	23	Diabetes
5	14853	50	Cancer
6	14853	55	Flu
7	14850	47	Diabetes
8	14850	49	Diabetes
9	13053	31	Cancer
10	13053	37	Cancer
11	13068	36	Cancer
12	13068	35	Cancer

3-diverse  
⇒

	Non-Sensitive		Sensitive
	Zip	Age	Condition
1	1305*	<40	Flu
4	1305*	<40	Diabetes
9	1305*	<40	Cancer
10	1305*	<40	Cancer
5	1485*	>40	Cancer
6	1485*	>40	Flu
7	1485*	>40	Diabetes
8	1485*	>40	Diabetes
2	1306*	<40	Flu
3	1306*	<40	Diabetes
11	1306*	<40	Cancer
12	1306*	<40	Cancer

# $\ell$ -diversity

	Non-Sensitive		Sensitive
	Zip	Age	Condition
1	1305*	<40	Flu
4	1305*	<40	Diabetes
9	1305*	<40	Cancer
10	1305*	<40	Cancer
5	1485*	>40	Cancer
6	1485*	>40	Flu
7	1485*	>40	Diabetes
8	1485*	>40	Diabetes
2	1306*	<40	Flu
3	1306*	<40	Diabetes
11	1306*	<40	Cancer
12	1306*	<40	Cancer

May not be enough if one has *a priori* skewed distribution of sensitive attributes, e.g., 1% of the overall population have cancer.

First equivalence class  $\Rightarrow$  50% cancer  
 $\gg$  1%

# Another example

	Non-Sensitive		Sensitive	
	Zip	Age	Salary	Condition
1	47677	29	3K	Gastric Ulcer
2	47602	22	4K	Gastritis
3	47678	27	5K	Stomach Cancer
4	47905	43	6K	Gastritis
5	47909	52	11K	Flu
6	47906	47	8K	Bronchitis
7	47605	30	7K	Bronchitis
8	47673	36	9K	Pneumonia
9	47607	32	10K	Stomach Cancer

3-diverse  
⇒

	Non-Sensitive		Sensitive	
	Zip	Age	Salary	Condition
1	476**	2*	3K	Gastric Ulcer
2	476**	2*	4K	Gastritis
3	476**	2*	5K	Stomach Cancer
4	4790*	>40	6K	Gastritis
5	4790*	>40	11K	Flu
6	4790*	>40	8K	Bronchitis
7	476**	3*	7K	Bronchitis
8	476**	3*	9K	Pneumonia
9	476**	3*	10K	Stomach Cancer

# Another example

	Non-Sensitive		Sensitive	
	Zip	Age	Salary	Condition
1	476**	2*	3K	Gastric Ulcer
2	476**	2*	4K	Gastritis
3	476**	2*	5K	Stomach Cancer
4	4790*	>40	6K	Gastritis
5	4790*	>40	11K	Flu
6	4790*	>40	8K	Bronchitis
7	476**	3*	7K	Bronchitis
8	476**	3*	9K	Pneumonia
9	476**	3*	10K	Stomach Cancer

Does not take into account semantical closeness of sensitive attributes in an equivalence class.

First equivalence class  $\Rightarrow$  low salary & stomach-related condition

## $t$ -closeness

---

- Earth Mover's distance (EMD): the minimal amount of work required to transform one distribution to another via moving probability mass between each other.
- An equivalence class is said to have  $t$ -closeness if the EMD between the distribution of any sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold  $t$ .
- A table is said to have  $t$ -closeness if all equivalence classes have  $t$ -closeness.

# $t$ -closeness

	Non-Sensitive		Sensitive	
	Zip	Age	Salary	Condition
1	47677	29	3K	Gastric Ulcer
2	47602	22	4K	Gastritis
3	47678	27	5K	Stomach Cancer
4	47905	43	6K	Gastritis
5	47909	52	11K	Flu
6	47906	47	8K	Bronchitis
7	47605	30	7K	Bronchitis
8	47673	36	9K	Pneumonia
9	47607	32	10K	Stomach Cancer

0.167 – closeness  
w.r.t. Salary  
0.278 – closeness  
w.r.t. Condition

	Non-Sensitive		Sensitive	
	Zip	Age	Salary	Condition
1	4767*	<40	3K	Gastric Ulcer
3	4767*	<40	5K	Stomach Cancer
8	4767*	<40	9K	Pneumonia
4	4790*	>40	6K	Gastritis
5	4790*	>40	11K	Flu
6	4790*	>40	8K	Bronchitis
2	4760*	<40	4K	Gastritis
7	4760*	<40	7K	Bronchitis
9	4760*	<40	10K	Stomach Cancer



# Differential privacy

---

- Interactive query model instead of releasing data
- Protect privacy by adding random noises to true answers
- Can resist arbitrary auxiliary information

# An example

Name	Flu ( $X$ )
Ross	1
Monica	1
Joey	0
Phoebe	0
Chandler	1
Rachel	0

An adversary wants to find whether Chandler has flu or not. He knows in which row Chandler is. Suppose he is only allowed to use a particular form of query  $Q_i$  that returns the partial sum of the first  $i$  rows of the second column in the database.

- Queries: What is  $Q_4$ ? What is  $Q_5$ ?
- True answers:  $Q_4 = 2, Q_5 = 3 \Rightarrow X(\text{Chandler}) = Q_5 - Q_4 = 1$
- DP answers:  $Q_4 = 2 + \text{noise}, Q_5 = 3 + \text{noise}'$
- E.g.,  $Q_4 = 2 + 0.4 = 2.4, Q_5 = 3 - 0.2 = 2.8 \Rightarrow Q_5 - Q_4 = 0.4$
- Cannot tell whether Chandler has flu or not

# Formal definition

- Two databases  $D = \{d_1, \dots, d_n\}$  and  $D' = \{d'_1, \dots, d'_n\}$  are adjacent if there exists  $i \in \{1, \dots, n\}$  such that  $d_j = d'_j$  for all  $j \neq i$ .

Name	Flu ( $X$ )
Ross	1
Monica	1
Joey	0
Phoebe	0
Chandler	1
Rachel	0

Name	Flu ( $X$ )
Ross	1
Monica	1
Joey	0
Phoebe	0
Chandler	0
Rachel	0

- Given  $\varepsilon, \delta \geq 0$ , a randomized mechanism  $\mathcal{M}$  with domain  $\mathcal{D}$  is  $(\varepsilon, \delta)$ -differentially private if for all  $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$  and all adjacent databases  $D, D' \in \mathcal{D}$ , it holds that  $P[\mathcal{M}(D) \in \mathcal{S}] \leq e^\varepsilon P[\mathcal{M}(D') \in \mathcal{S}] + \delta$ .

# Laplace mechanism

- Laplace distribution  $x \sim \text{Lap}(b)$ :  $p(x) = \frac{1}{2b} \exp(-\frac{|x|}{b})$
- Sensitivity of a function  $f$ :  $\Delta f = \max_{D, D' \in \mathcal{D}: D, D' \text{ adjacent}} |f(D) - f(D')|$
- Given a function  $f: \mathcal{D} \rightarrow \mathbb{R}$  with sensitivity  $\Delta f$ . The Laplace mechanism  $\mathcal{M}(D) = f(D) + x$  with  $x \sim \text{Lap}(\Delta f/\epsilon)$  is  $\epsilon$ -differentially private.

Name	Flu ( $X$ )
Ross	1
Monica	1
Joey	0
Phoebe	0
Chandler	1
Rachel	0

$$\Delta f = 1$$

$$\mathcal{M}(D) = Q_i(D) + x, \text{ with } x \sim \text{Lap}(1/\epsilon)$$



# Questions?

---