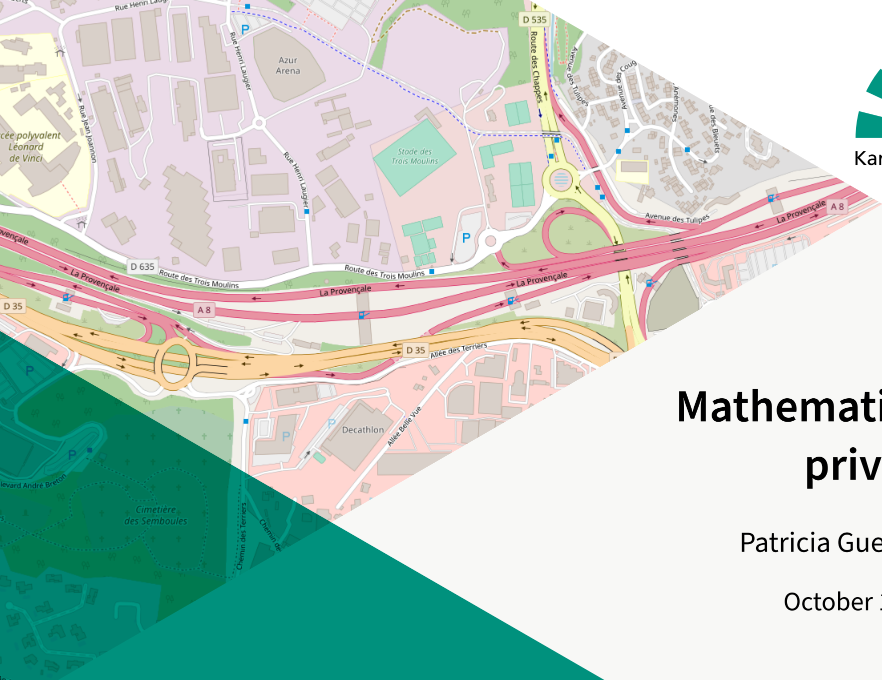


Karlsruhe Institute of Technology

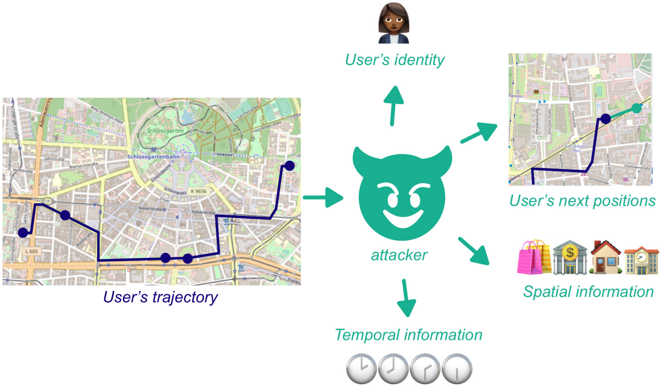
Mathematics behind privacy

Patricia Guerra-Balboa

October 14, 2025

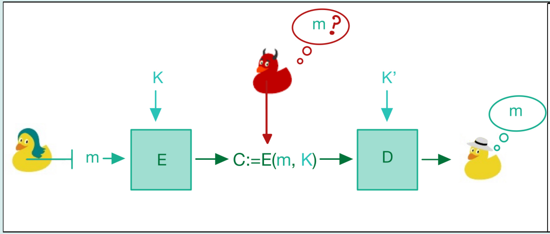
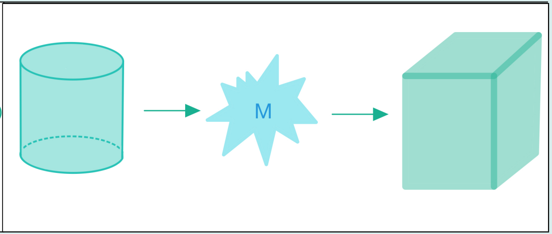
Motivation

“Privacy is one the biggest problems in this new electronic age”- Andy Grove (former INTEL Ceo)

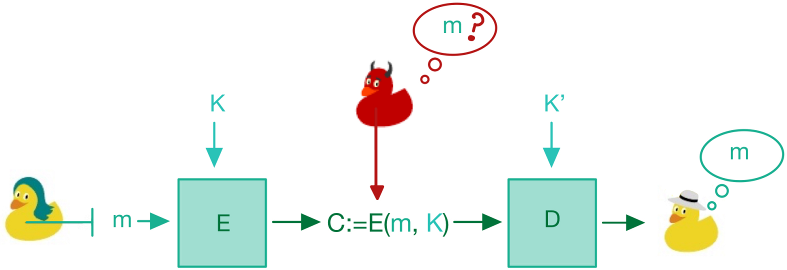


Data Privacy

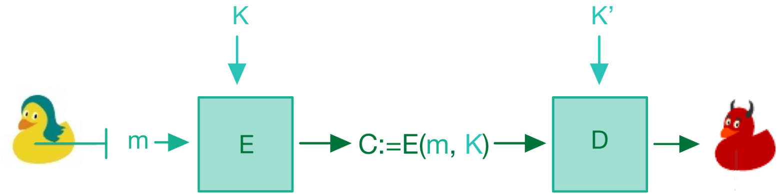
Privacy frameworks

Cryptography	SDC
	
secure message delivery	information sharing with privacy

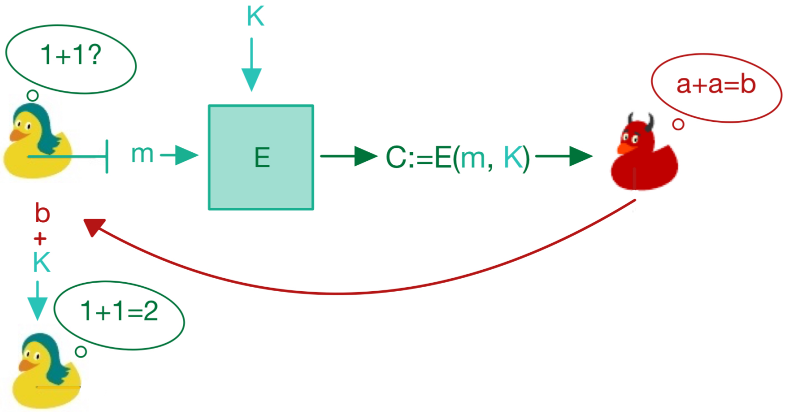
Cryptography



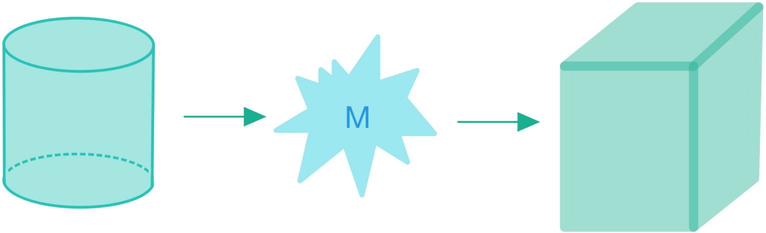
Cryptography



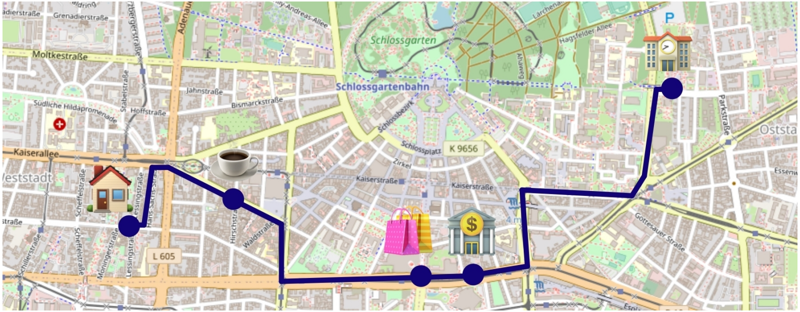
Cryptography (FHE)



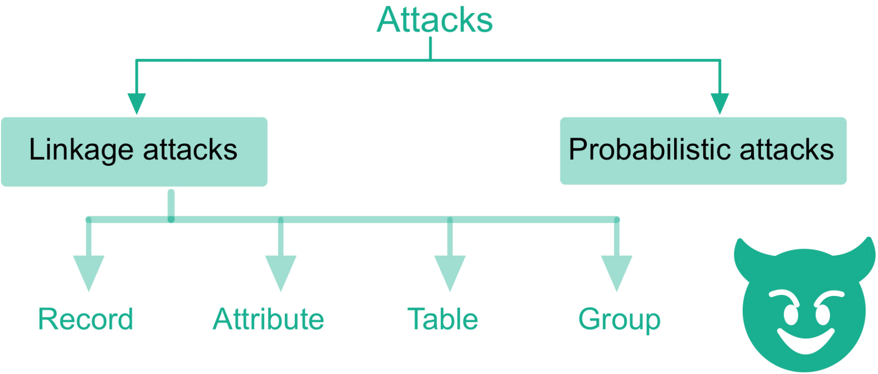
Statistical Disclosure Control



Identifiers Vs quasi-identifiers



Attacks



Attacks
Linkage attacks
Probabilistic attacks
Record
Attribute
Table
Group

Real examples

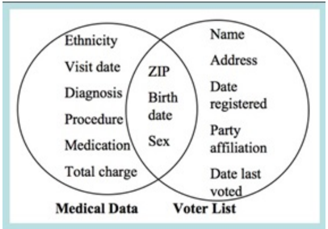


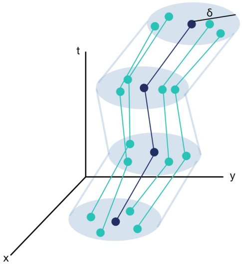
Figure 1: Zip code, gender, and birth date were likely sufficient in 1990 to identify 87% of individuals in the U.S.

Figure 2: 8 movie ratings and dates were enough to uniquely identify 99% of viewers in the Netflix Prize dataset

Privacy Notions in SDC

Syntactic Notions

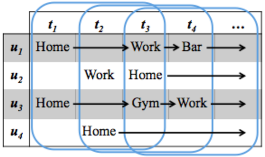
Database properties



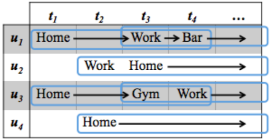
- k-anonymity
- l-diversity
- t-closeness
- Attribute Privacy

Semantic Notions

ϵ-differential privacy



	t_1	t_2	t_3	t_4	...
u_1	Home		Work	Bar	
u_2		Work	Home		
u_3	Home		Gym	Work	
u_4		Home			



	t_1	t_2	t_3	t_4	...
u_1	Home		Work	Bar	
u_2		Work	Home		
u_3	Home		Gym	Work	
u_4		Home			

Syntactic Notions

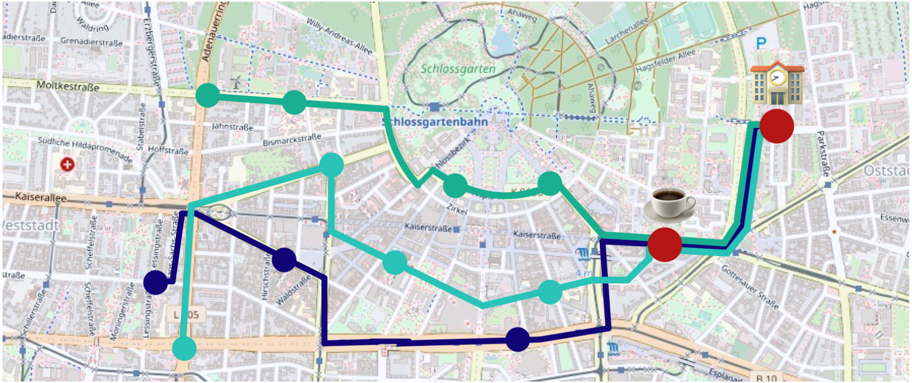
k-anonymity

We say that a dataset D satisfies k-Anonymity for a given value $k \in \mathbb{Z}$ if: For each row $r_1 \in D$, there exist at least $k - 1$ other rows $r_2 \ldots r_k \in D$ such that

$$\Pi_{qi(D)} r_1 = \Pi_{qi(D)} r_2, \ldots, \Pi_{qi(D)} r_1 = \Pi_{qi(D)} r_k$$

where $q_i(D)$ is the quasi-identifiers of D and $\Pi q_i(D) r$ represents the columns of r containing quasi-identifiers (i.e. the projection of the quasi-identifiers).

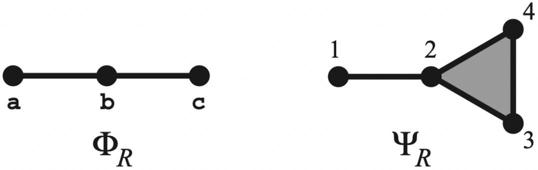
Syntactic Notions



Attribute Privacy

$$\Phi_R := \{\gamma \subseteq Y \quad | \quad \exists x \in X : (x,y) \in R \quad \forall y \in \gamma\}$$
$$\Psi_R := \{\sigma \subseteq x \quad | \quad \exists y \in Y : (x,y) \in R \quad \forall x \in \sigma\}$$

R	a	b	c
1	•	•	
2		•	•
3			•
4			•



Attribute Privacy

$$\phi_R : \quad \Psi_R \longrightarrow \Phi_R \qquad \psi_R : \quad \Phi_R \longrightarrow \Psi_R$$

$$\sigma \rightsquigarrow \cap_{x\in\sigma} Y_x \qquad \gamma \rightsquigarrow \cap_{y\in\gamma} X_y$$

Atribute Privacy

Let D be a database. X, Y sets of users and atttributes of D resp. We say that D has attribute privacy if the relation R drawn from D veryfies:

$$\phi_R \circ \psi_R = Id_{\Phi_R}$$

Attribute Privacy

Theorem

Let R relation. X, Y non empty sets, then:

$$\Phi_R \text{ has not free faces } \Rightarrow \phi_R \circ \psi_R = Id_{\Phi_R}(A.P)$$

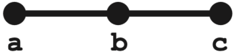
Theorem

Let R relation. X, Y non empty sets, then:

$$\begin{array}{c} \phi_R \circ \psi_R = Id_{\Phi_R}(A.P) \\ \wedge \\ \psi_R(Y_x) = \{x\}(U.I) \end{array} \quad \Rightarrow \quad \Phi_R \text{ has not free faces}$$

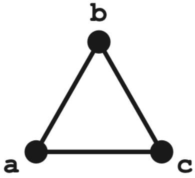
Attribute Privacy

R	a	b	c
1	•	•	
2		•	•
3			•
4			•



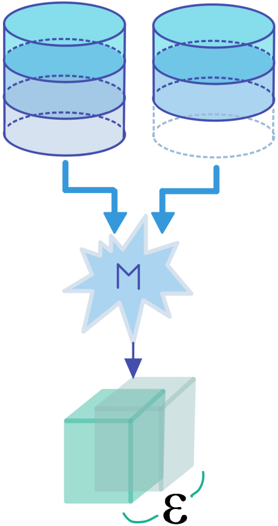
Φ_R

R'	a	b	c
1	•	•	
2		•	•
3	•		•
4			•



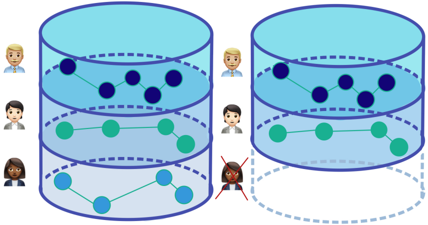
$\Phi_{R'}$

Differential Privacy

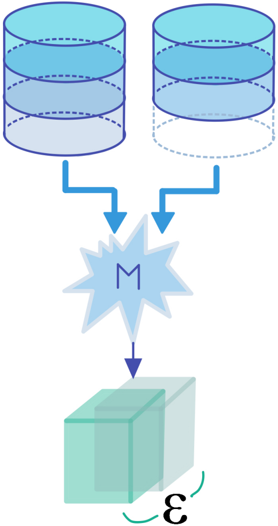


ϵ-Differential Privacy

$$\mathbb{P}(M(D) = r) \leq e^{\epsilon} \cdot \mathbb{P}(M(D') = r)$$

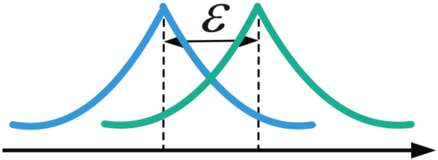


Differential Privacy



Privacy Loss (by observing r)

$$\mathcal{L}^{r}_{M(D)||M(D')} = ln\left(\frac{\mathbb{P}(M(D) = r)}{\mathbb{P}(M(D') = r)}\right)$$



Differential Privacy Properties

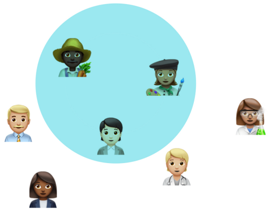
Group Privacy

Given M a ϵ-DP mechanism, for all $\|D - D'\|_1 \leq k$ and all $r \in Range(M)$

$$\mathbb{P}(M(D) = r) \leq e^{k\epsilon} \cdot \mathbb{P}(M(D') = r)$$

Post-procesing

Let $M \colon \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}$ be a randomized algorithm that is ϵ-DP. Let $f \colon \mathcal{R} \to \mathcal{R}'$ be an arbitrary map. Then $f \circ M \colon \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}'$ is ϵ-DP.



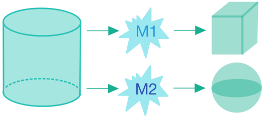
Differential Privacy Properties

Sequential Composition

Let $M_1 : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_1$ be an ϵ_1-DP algorithm, and let $M_2 : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_2$ be an ϵ_2-DP algorithm. Then their combination is $(\epsilon_1 + \epsilon_2)$-DP :

$$M_{1,2} : \quad \mathbb{N}^{|\mathcal{X}|} \longrightarrow \mathcal{R}_1 \times \mathcal{R}_2$$

$$D \rightsquigarrow (M_1(D), M(D_2))$$

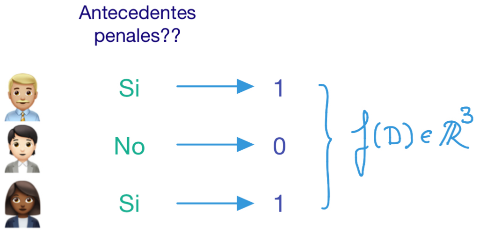


Algorithms Achieving Differential Privacy

ℓ_1-sensitivity

The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^n$ is:

$$\Delta(f) := \max_{\|D, D'\|_1 = 1} \|f(D) - f(D')\|_1$$

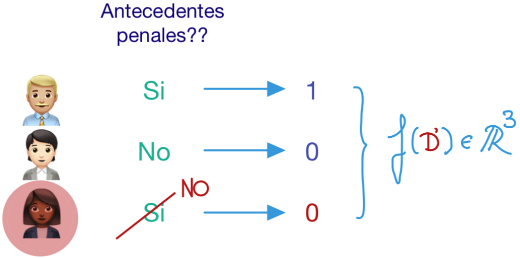


Algorithms Achieving Differential Privacy

ℓ_1-sensitivity

The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^n$ is:

$$\Delta(f) := \max_{\|D, D'\|_1 = 1} \|f(D) - f(D')\|_1$$

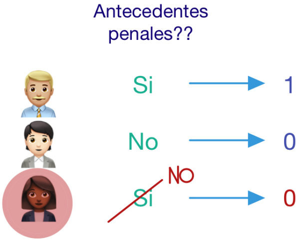


Algorithms Achieving Differential Privacy

ℓ_1-sensitivity

The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^n$ is:

$$\Delta(f) := \max_{\|D,D'\|_1=1} \|f(D) - f(D')\|_1$$



$\Delta f = 1$

Algorithms Achieving Differential Privacy

ℓ_1-sensitivity

The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^n$ is:

$$\Delta(f) := \max_{\|D, D'\|_1 = 1} \|f(D) - f(D')\|_1$$

UNBOUNDED SENSITIVITIES!!

outliers and huge noise

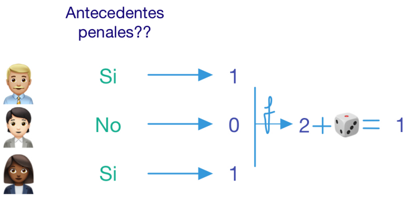
Algorithms Achieving Differential Privacy
Laplace Mechanism

Laplace Mechanism

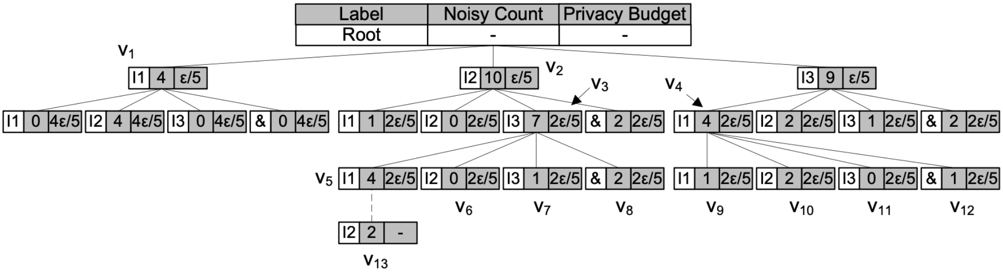
Given any function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^n$ the Laplace mechanism is defined as:

$$ML(D, f(\cdot), \epsilon) = f(D) + (Y_1, \dots, Y_n)$$

where Y_i are i.i.d. random variables drawn from $Lap(\frac{\Delta f}{\epsilon})$.



Algorithms Achieving Differential Privacy
Laplace Mechanism



Algorithms Achieving Differential Privacy
Exponential Mechanism

H	SMOKES	HAS_CANCER	DRINKS_SODA
	●	●	
		●	●
			●
			●

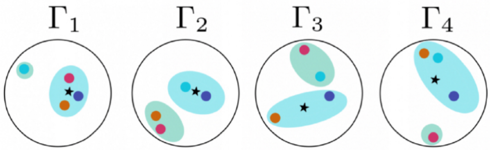
Algorithms Achieving Differential Privacy

Exponential Mechanism

Abstract range

$u: \quad \mathbb{N}^{|\mathcal{X}|} \times \mathcal{R} \longrightarrow \mathbb{R}$

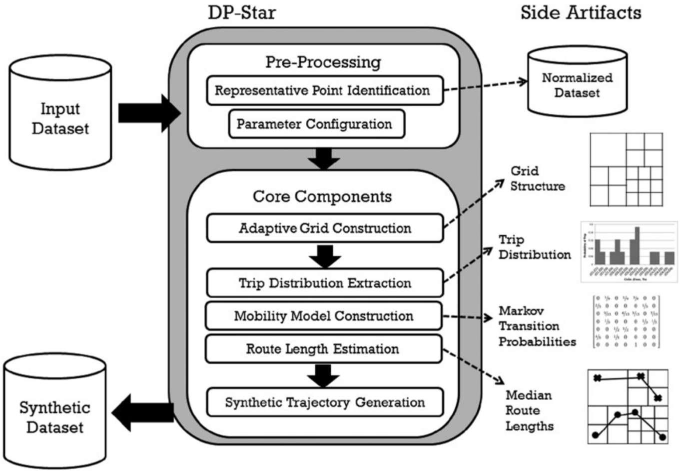
Score function



$$\Delta(u) := \max_{r \in \mathcal{R} \| D, D' \|_1 = 1} \| u(D,r) - u(D',r) \|_1$$

$M_E(D, u, \mathcal{R})$ selects and outputs an element $r \in \mathcal{R}$ with probability proportional to $exp(\frac{\epsilon u(D,r)}{2\Delta(u)})$. 2u

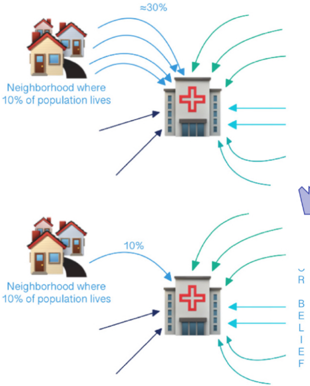
Mechanism Achieving Differential Privacy
Synthetic Data

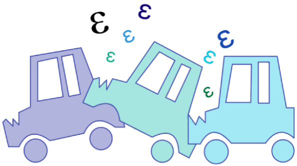


Limitations
on Differential Privacy

Correlation

Bayesian inference





Infinity streaming

Conclusions and Future Research