Risk Evaluation of LDP scheme LoPub against Variational Autoencoder

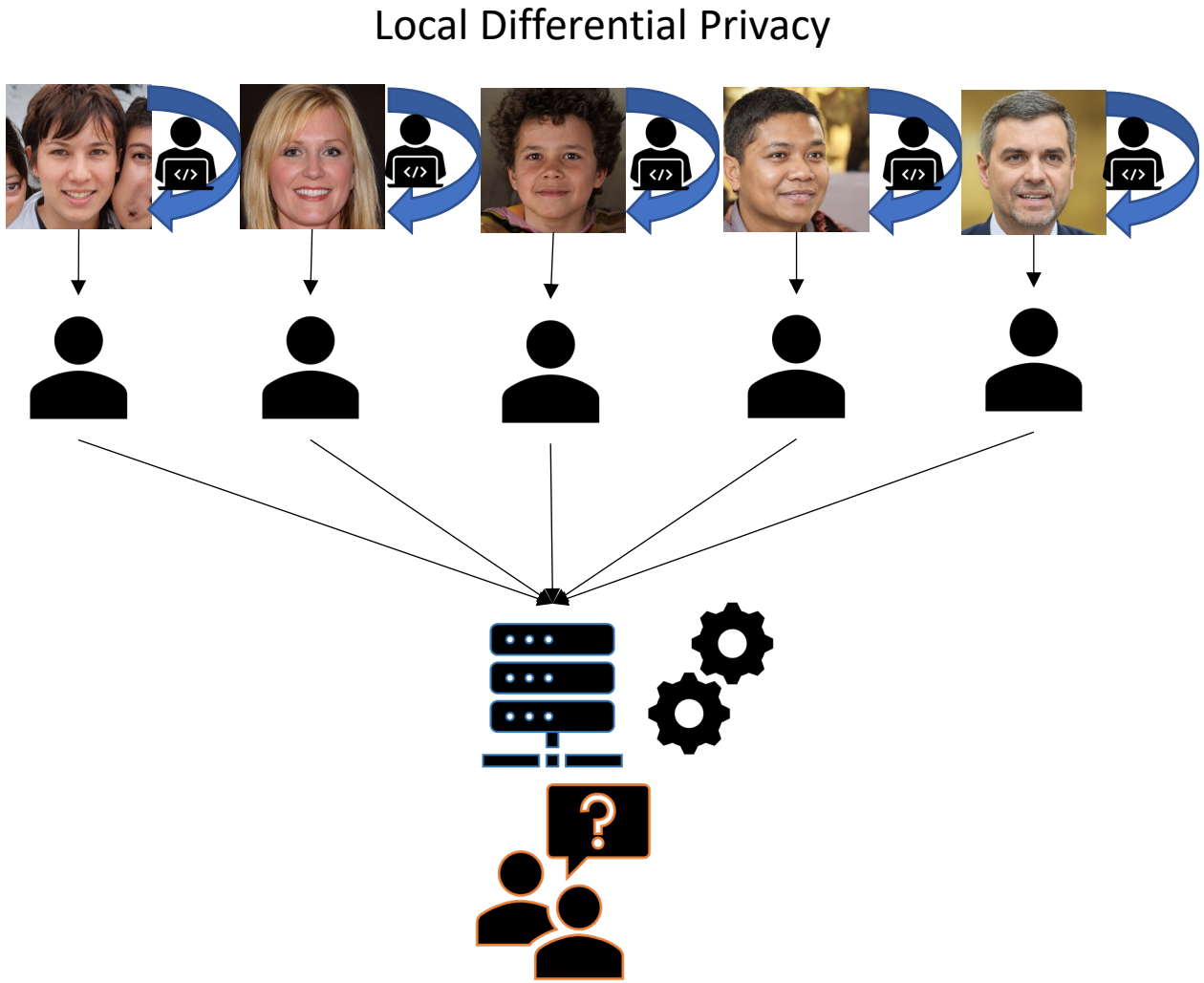
Andres Hernandez-Matamoros* and Hiroaki Kikuchi

Meiji University

Computer Security Symposium 2022
2022/10/24

*matamoros@meiji.ac.jp

What is LDP?



* Face images were taken from https://thispersondoesnotexist.com/

d(Attributes)

	A_1	A_2	⋯	A_d
	x_1^1	x_2^1	⋯	x_d^1
	x_1^2	x_2^2	⋯	x_d^2
	⋮	⋮	⋯	⋮
	x_1^{N-1}	x_2^{N-1}	⋯	x_d^{N-1}
	x_1^N	x_2^N	⋯	x_d^N

N

B.F.

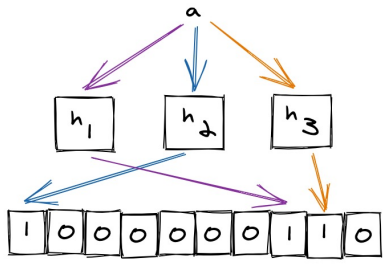
m_1		m_2			⋯	m_d		
0	1	0	1	1	⋯	0	⋯	1
1	0	1	1	0	⋯	1	⋯	0
⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋯	⋮
1	0	0	1	1	⋯	1	⋯	0
0	1	1	1	0	⋯	0	⋯	1

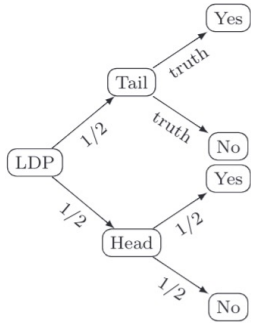
N

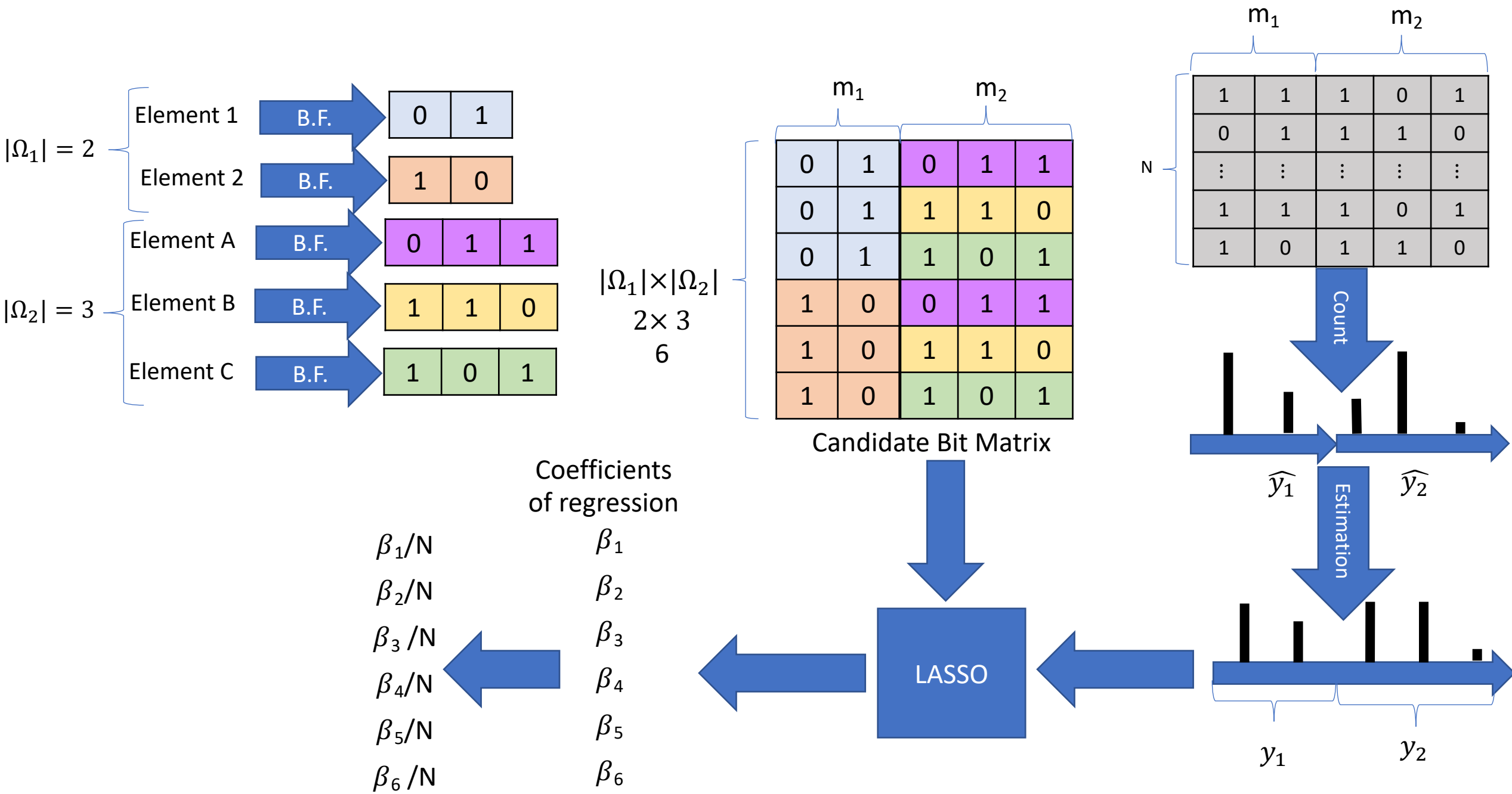
RR

m_1			m_d		
1	1	⋯	1	⋯	1
0	1	⋯	1	⋯	0
⋮	⋮	⋯	⋮	⋯	⋮
1	1	⋯	1	⋯	1
1	0	⋯	1	⋯	0

N






$|\Omega_1| = 2$
Element 1
B.F.
0 1
Element 2
B.F.
1 0
$|\Omega_2| = 3$
Element A
B.F.
0 1 1
Element B
B.F.
1 1 0
Element C
B.F.
1 0 1
m_1 m_2
$|\Omega_1| \times |\Omega_2|$
2× 3
6
0 1 0 1 1
0 1 1 1 0
0 1 1 0 1
1 0 0 1 1
1 0 1 1 0
1 0 1 0 1
Candidate Bit Matrix
m_1 m_2
N
1 1 1 0 1
0 1 1 1 0
⋮ ⋮ ⋮ ⋮ ⋮
1 1 1 0 1
1 0 1 1 0
Count
$\widehat{y_1}$ $\widehat{y_2}$
Estimation
y_1 y_2
LASSO
Coefficients
of regression
β_1
β_2
β_3
β_4
β_5
β_6
β_1/N
β_2/N
β_3 /N
β_4/N
β_5/N
β_6 /N

Pro vs Cons

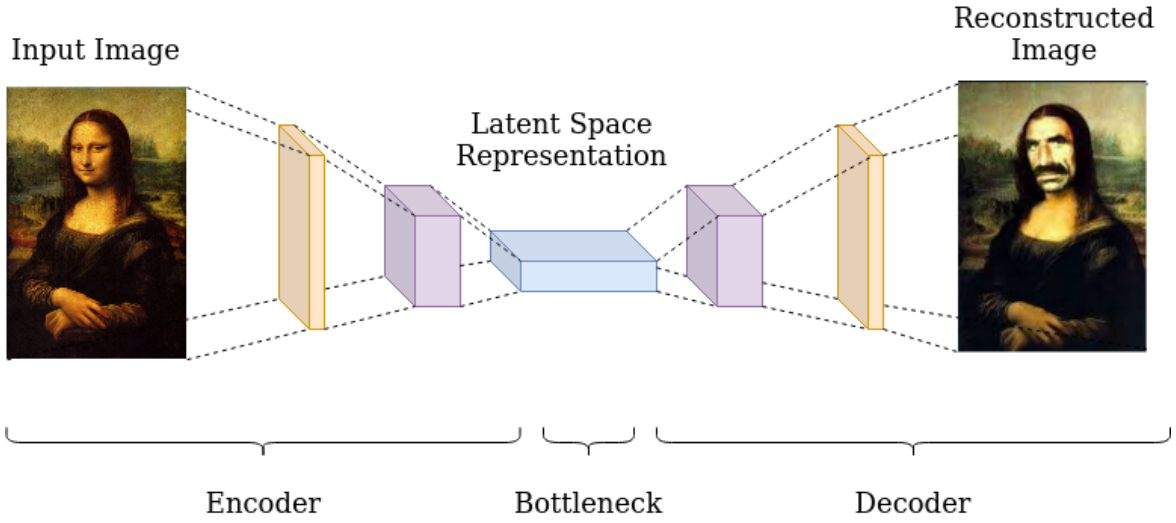
One/two-dimensional probability distributions can be efficiently estimated through the Lasso regression-based algorithm.

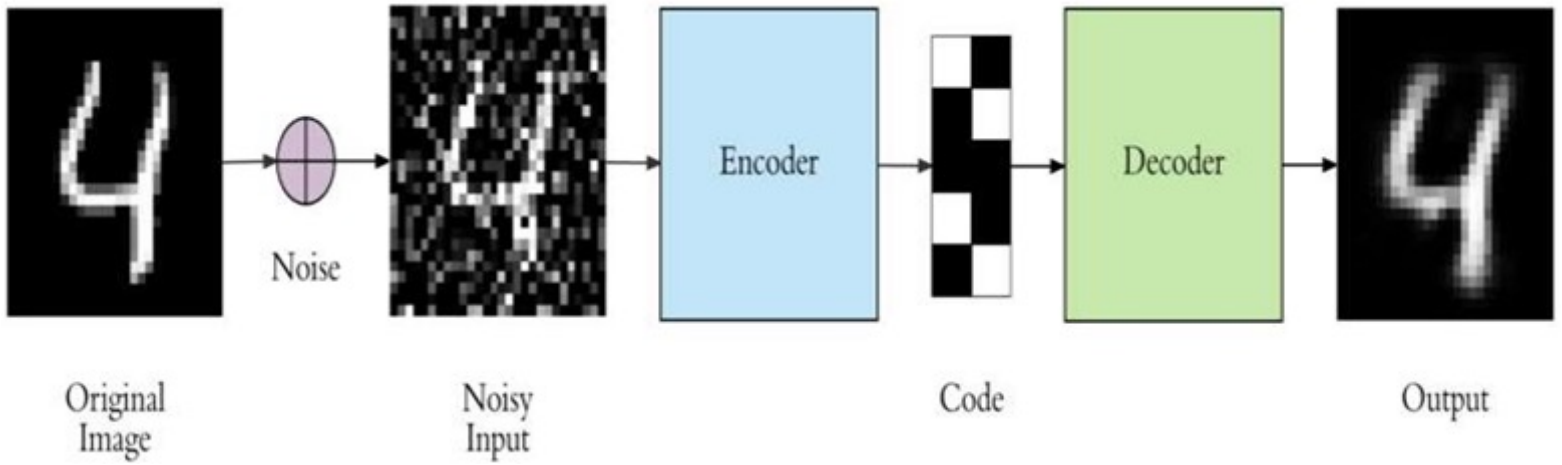
The k-dimensional distribution estimations in LoPub still suffer from the low data utility when k is large.

Proposal

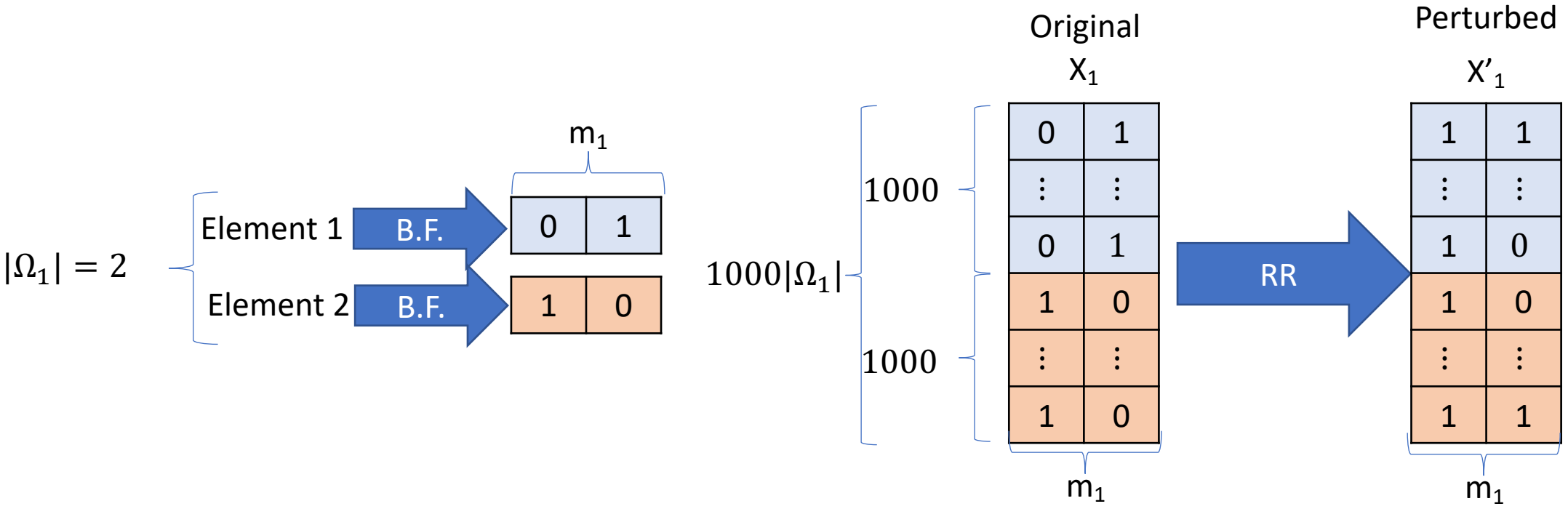
	LoPub	Ours
Randomization	BF+RR	
Estimation	LASSO	VAE

VAE-Applications

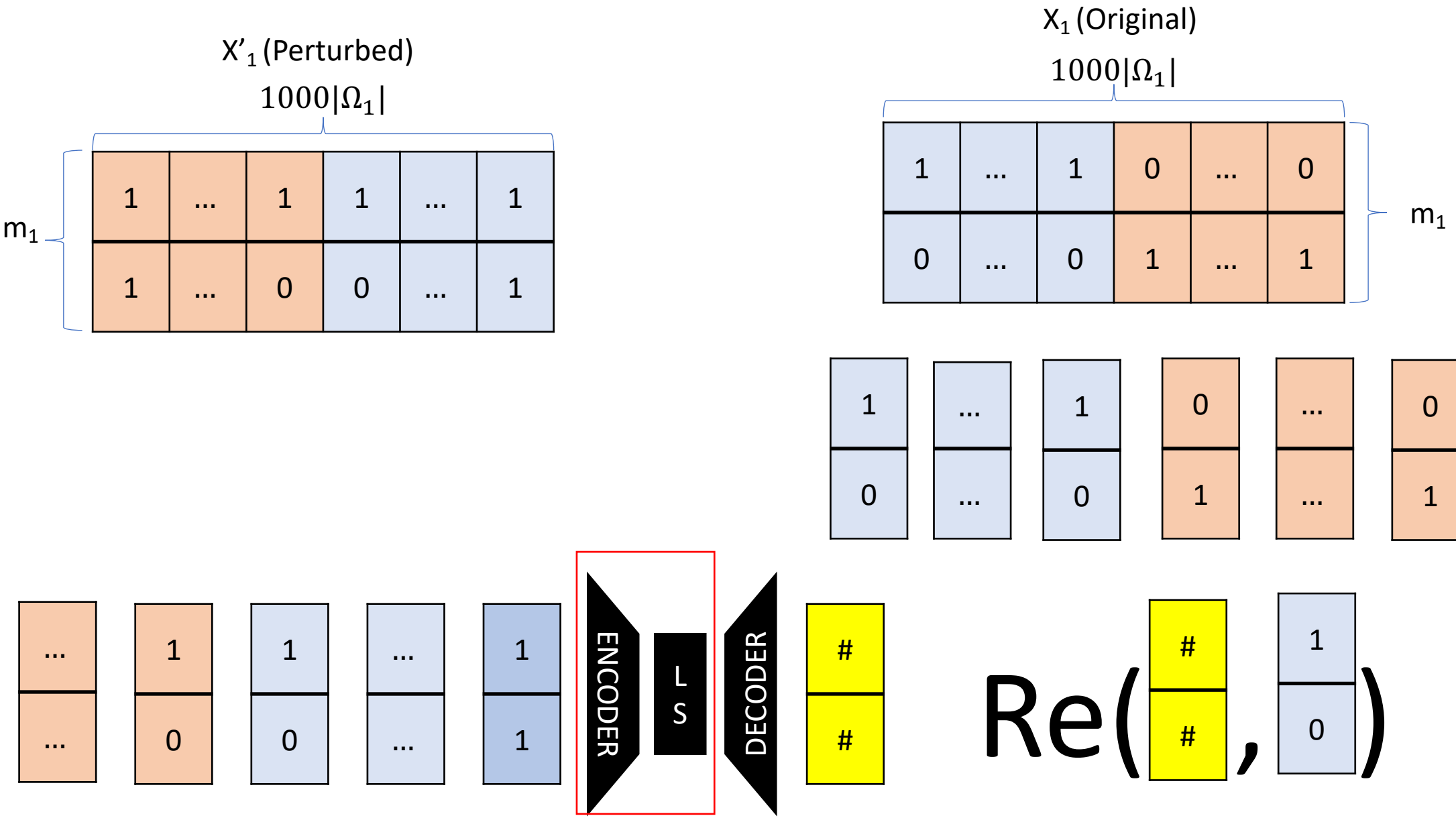




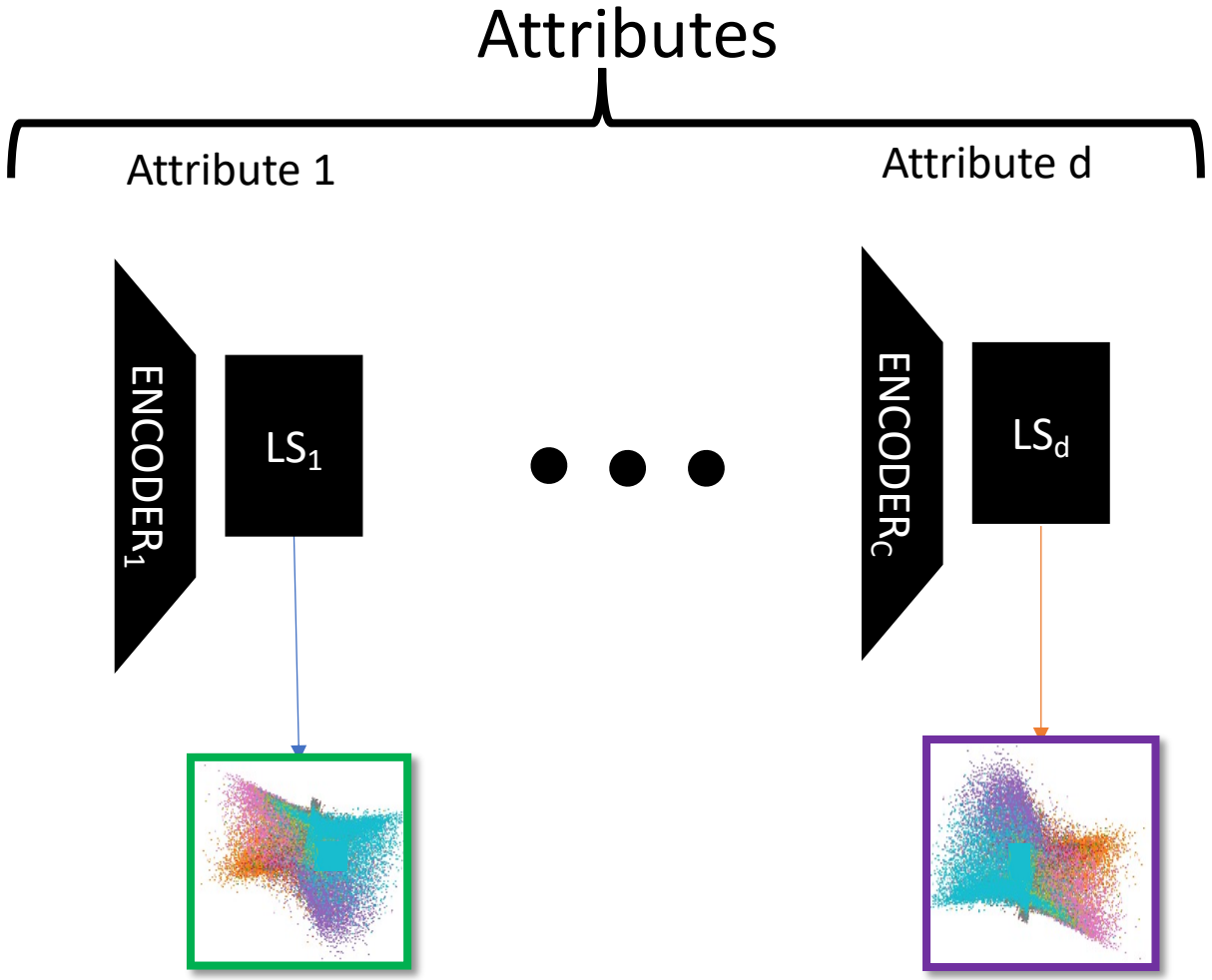
VAE Training-Dataset



VAE Training

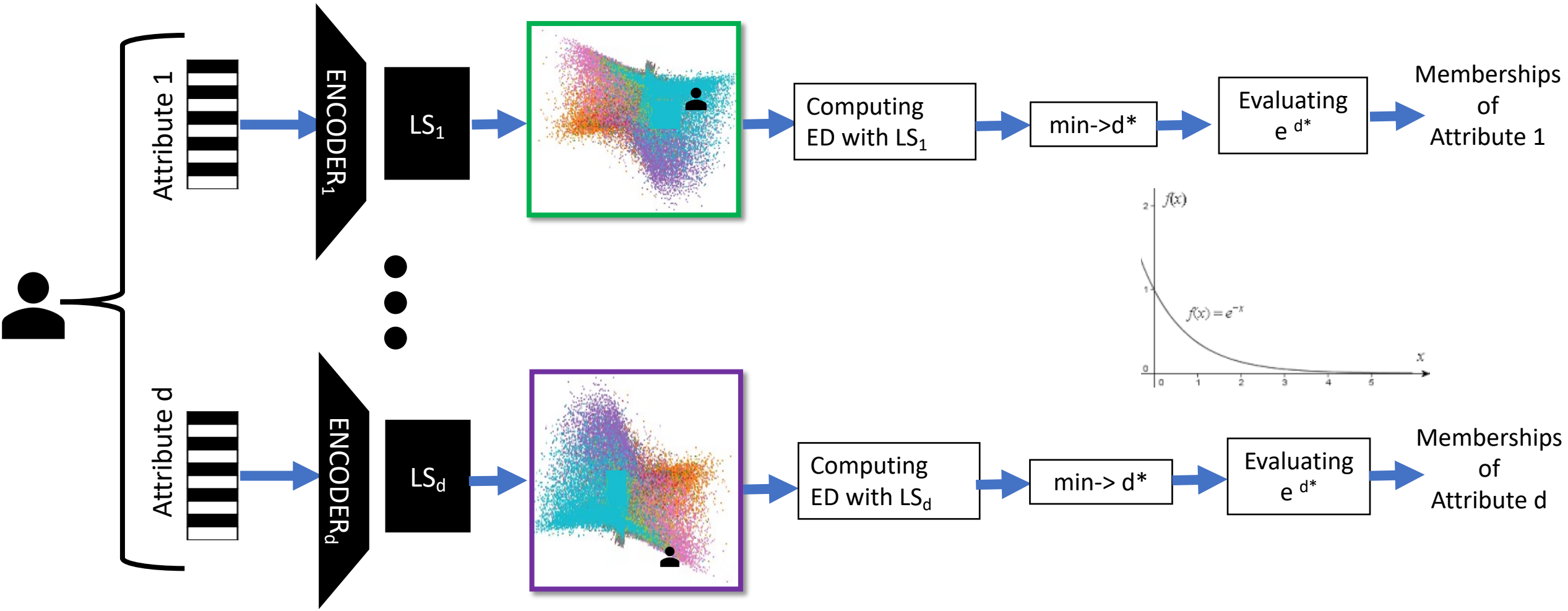


Latent Space's Attributes



Latent Space Examples

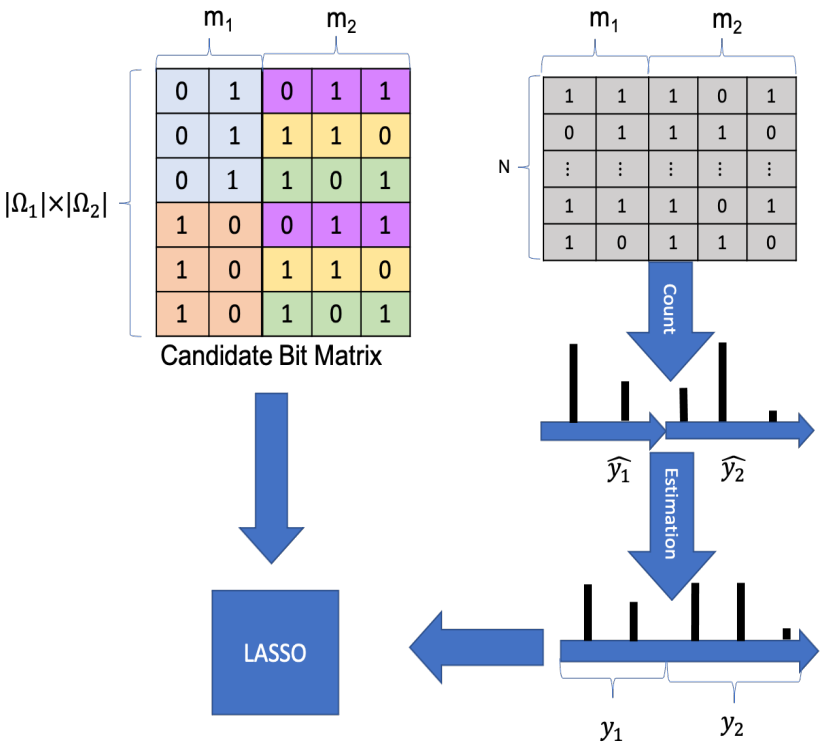
Evaluation



Euclidean Distance (ED)

LoPub vs Proposal

LoPub

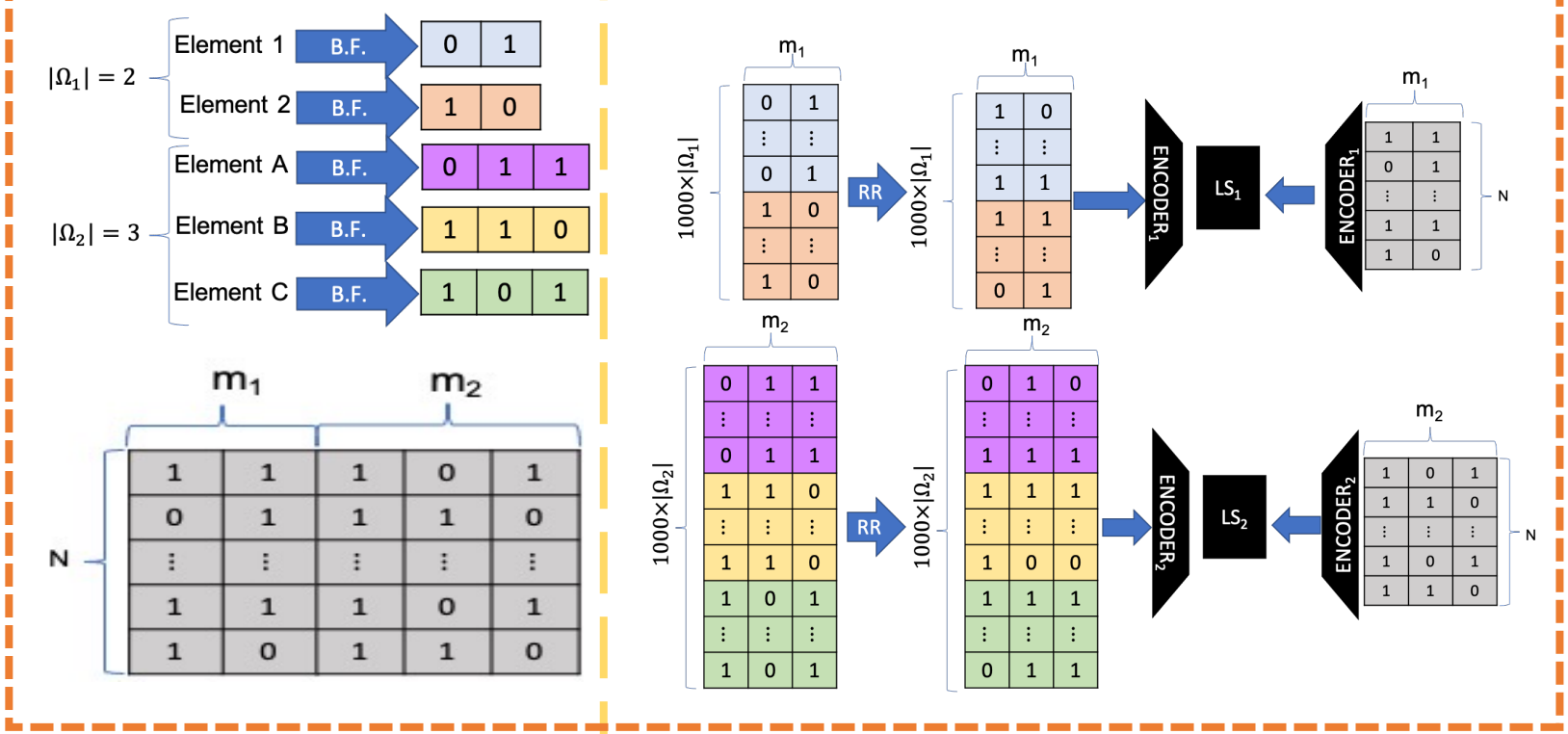


User

Encode

Perturb

Proposal

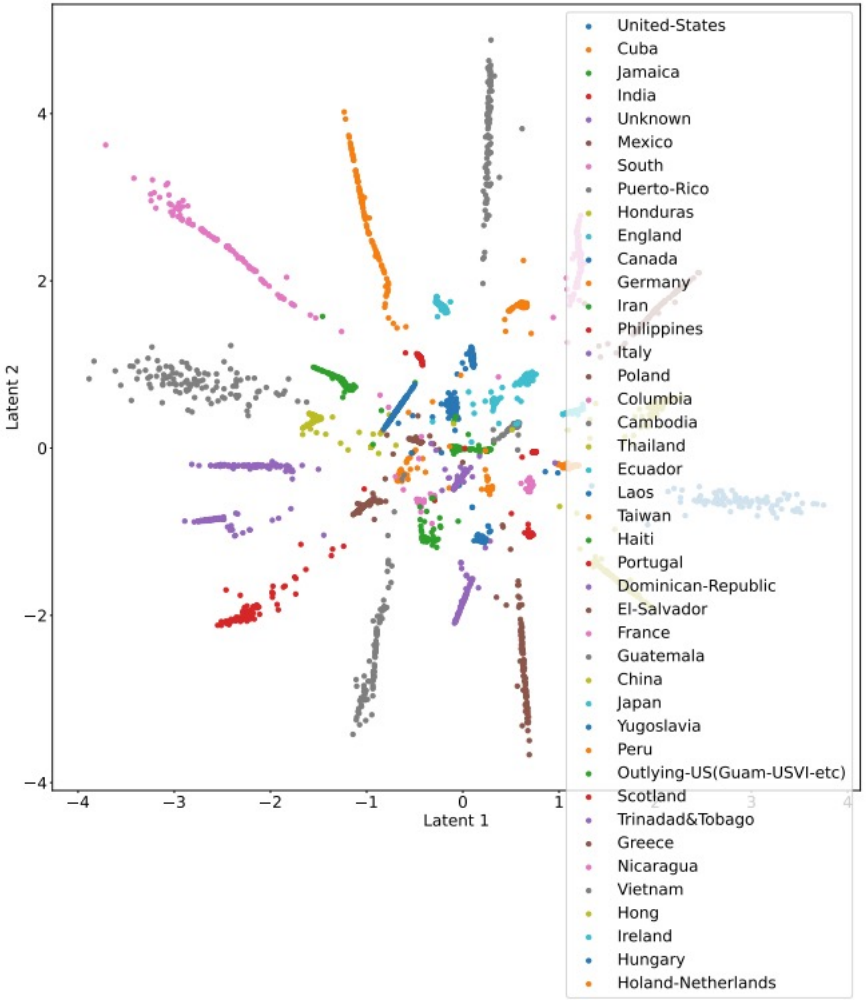


Datasets

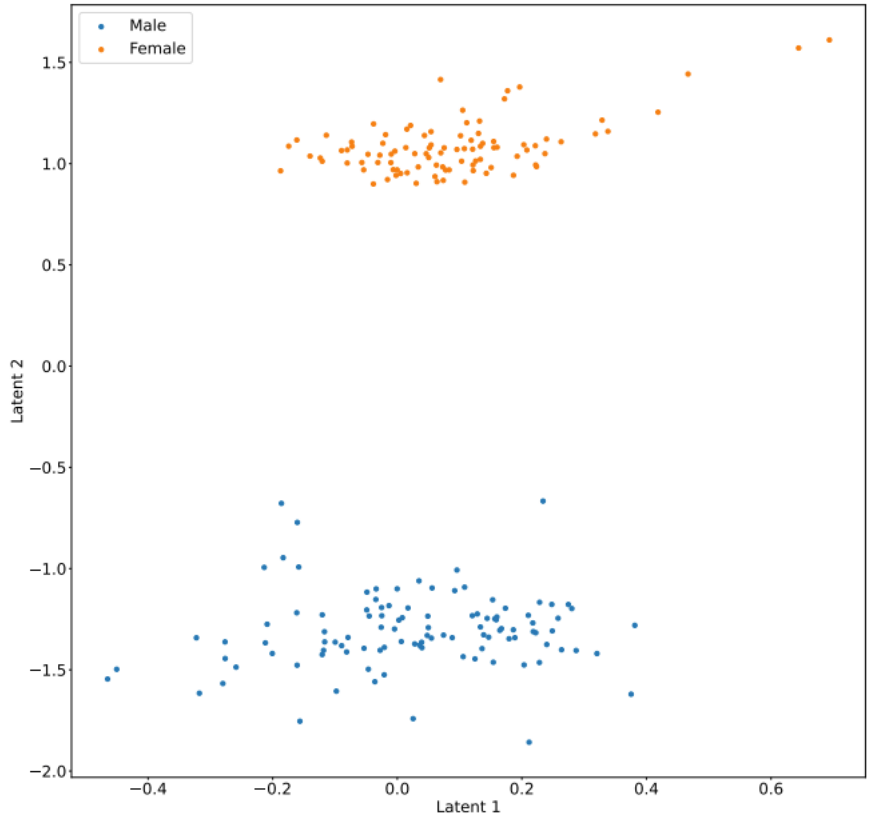
Dataset	Users	Attributes
NHANES	4189	$\lvert\omega_{Genre}\rvert = 2$ $\lvert\omega_{Race}\rvert = 5$ $\lvert\omega_{Education}\rvert = 5$ $\lvert\omega_{Marital}\rvert = 6$ $\lvert\omega_{Q_m}\rvert = 4$
Adult	32561	$\lvert\omega_{Workclass}\rvert = 9$ $\lvert\omega_{Education}\rvert = 16$ $\lvert\omega_{MaritalStatus}\rvert = 7$ $\lvert\omega_{Occupation}\rvert = 15$ $\lvert\omega_{Relationship}\rvert = 6$ $\lvert\omega_{Race}\rvert = 5$ $\lvert\omega_{Sex}\rvert = 2$ $\lvert\omega_{Country}\rvert = 42$ $\lvert\omega_{Target}\rvert = 2$
Bank	45211	$\lvert\omega_{Job}\rvert = 12$ $\lvert\omega_{Marital}\rvert = 3$ $\lvert\omega_{Education}\rvert = 4$ $\lvert\omega_{Default}\rvert = 2$ $\lvert\omega_{Housing}\rvert = 2$ $\lvert\omega_{Loan}\rvert = 2$ $\lvert\omega_{Contact}\rvert = 3$ $\lvert\omega_{Month}\rvert = 12$ $\lvert\omega_{Outcome}\rvert = 4$ $\lvert\omega_{Y}\rvert = 2$

Latent Space

Examples of Latent Spaces in 2D

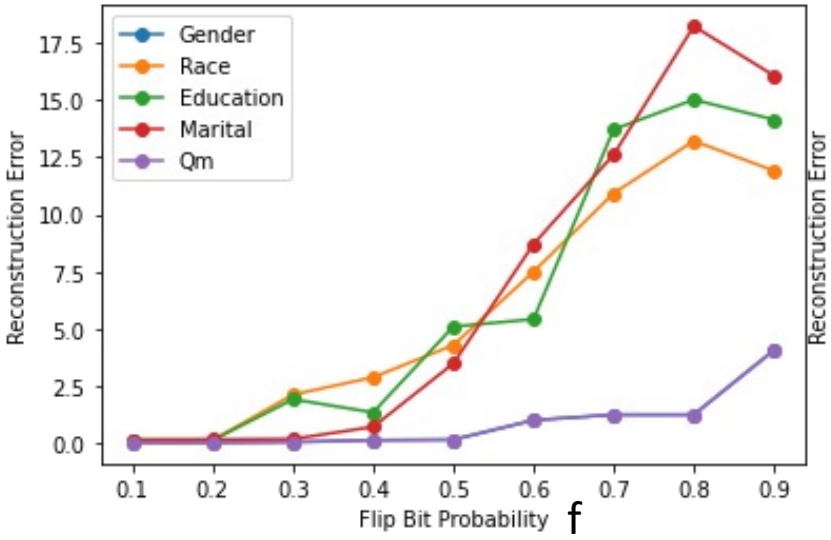


Country, Adults Dataset

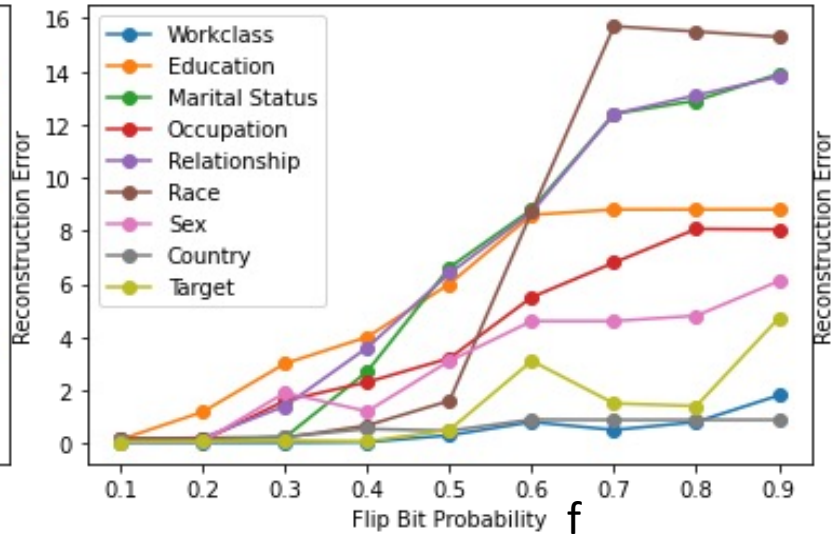


Gender, Adults Dataset

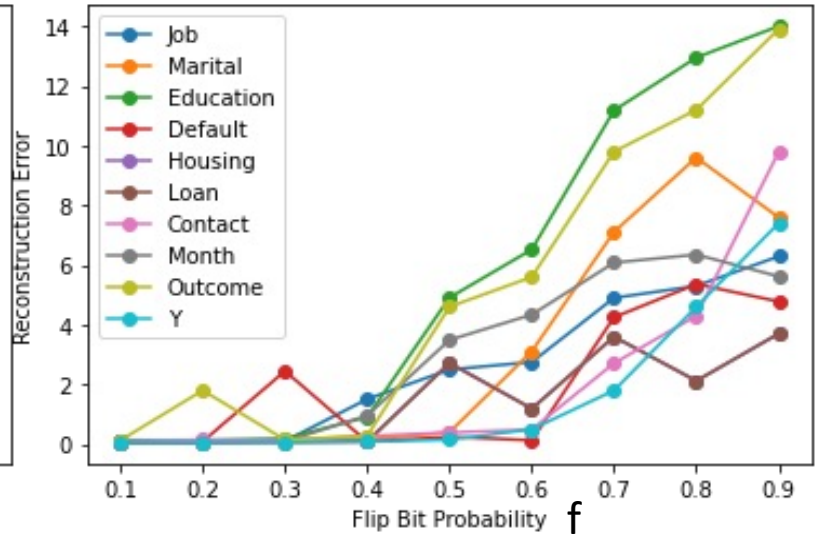
Reconstrucion Error



NHANES Dataset



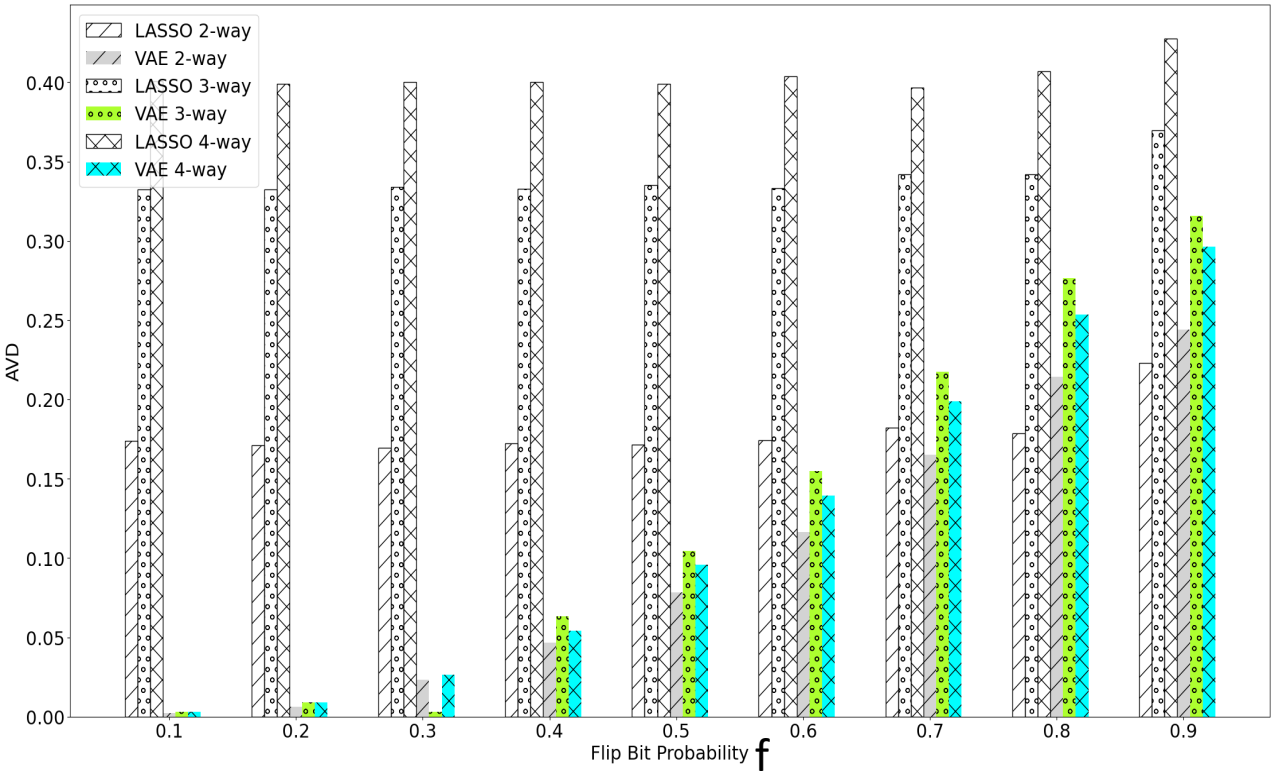
Adult Dataset



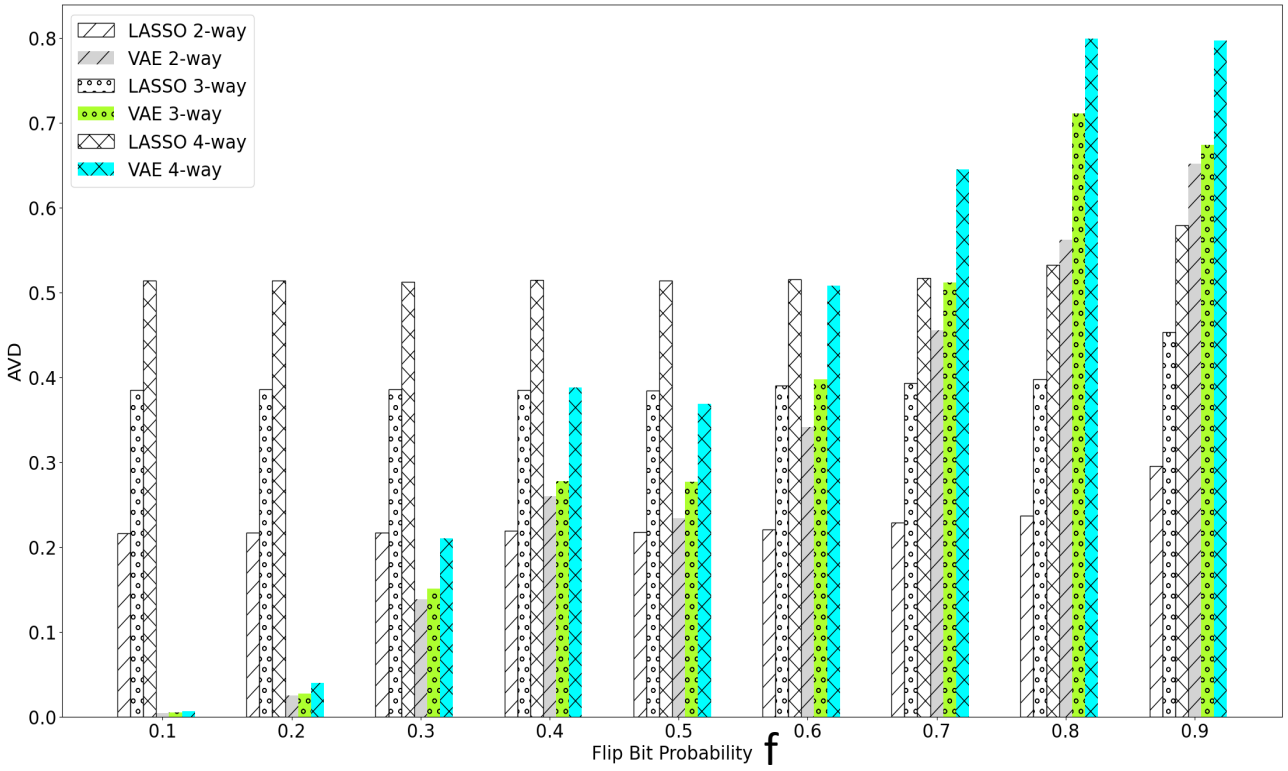
Bank Dataset

Accuracy K-way

NHANES Dataset



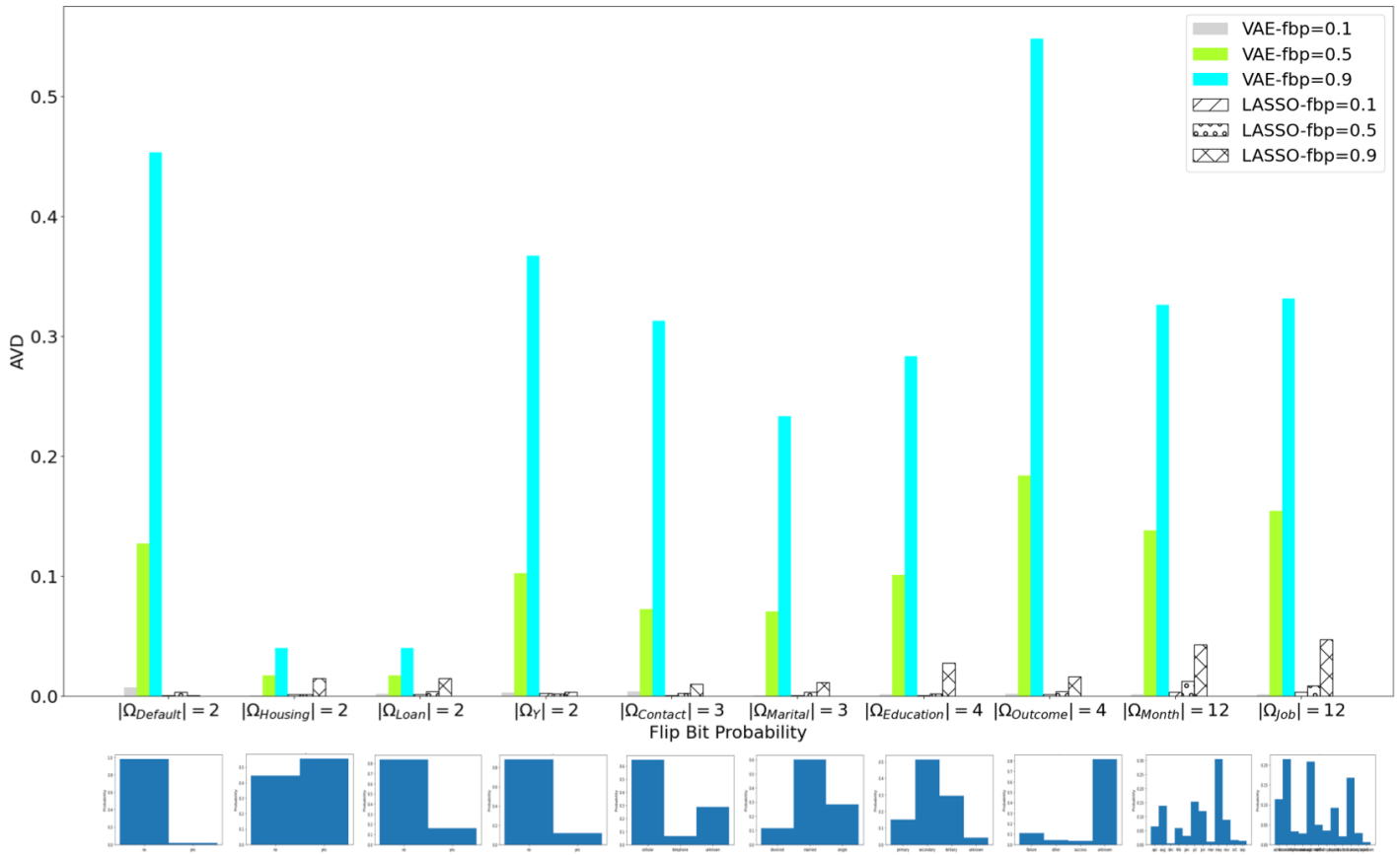
Adult Dataset



$$Dist_{\mathrm{AVD}}(P,Q) = \frac{1}{2}\sum_{\omega\in\Omega}|P(\omega) - Q(\omega)|$$

Bank cardinality

Accuracy vs Cardinality



Cardinality

Distribution of elements

Bank Dataset

Conclusions

- VAE performs better in small dataset. (less than 5k)
- Cardinality and the distribution of the attribute impact the performance.
- Future study variating quantities of users, distributions, and cardinalities to quantify the performance of an LDP schemes.
- Try specific VAE for attributes with different cardinality.