An Efficient Local Differential Privacy Scheme Using Bayesian Ridge Regression

Andres Hernandez-Matamoros* and Hiroaki Kikuchi

明治大学
MEIJI UNIVERSITY

PST 2023
2023/08/22

*matamoros@meiji.ac.jp

Why is Privacy necessary?

Sensitive information such as:

- Diagnoses
- Treatments
- Billing Records

Exposing this information:

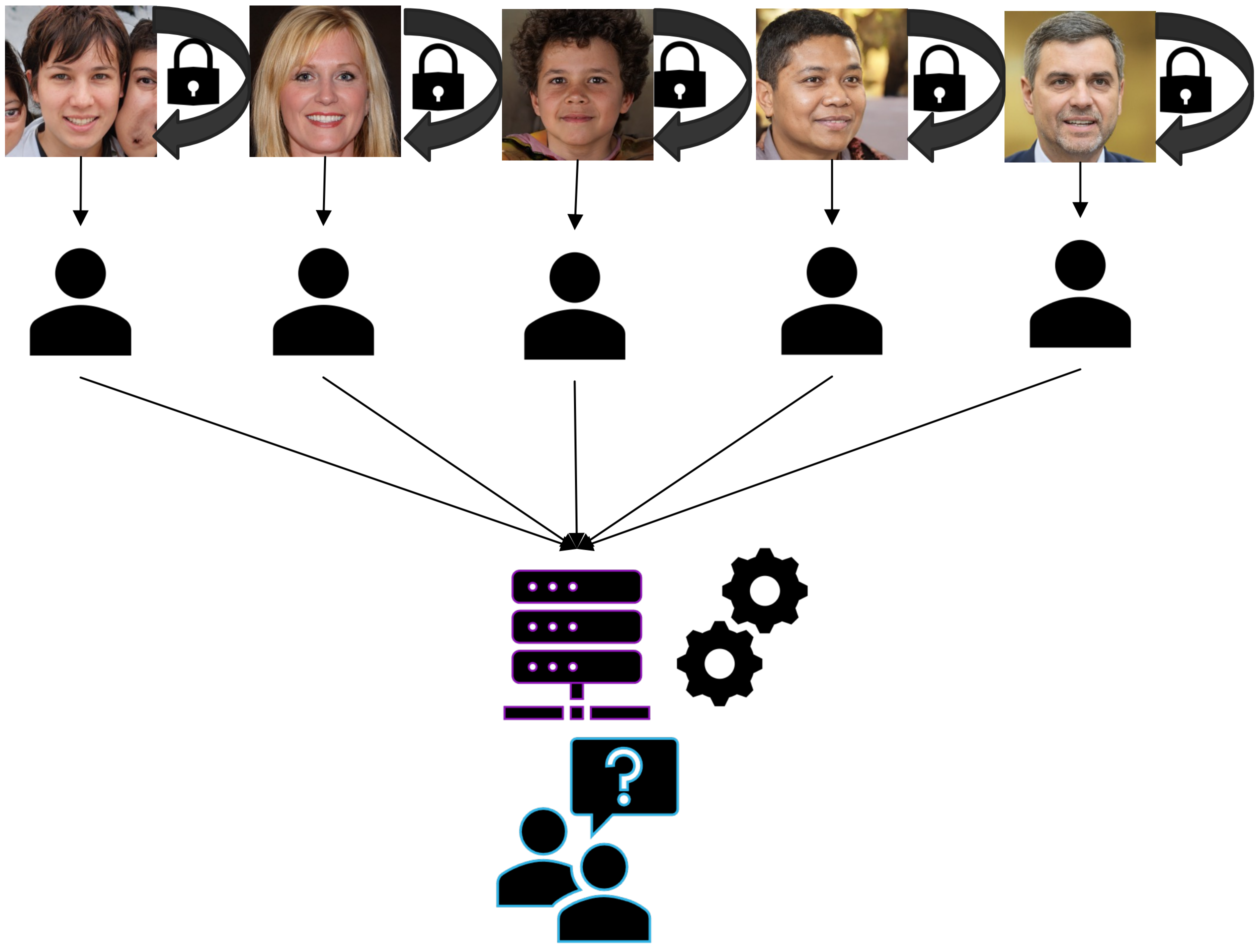
- Ethical issues
- Financial issues
- Legal issues

Public release of medical data is subject to restrictions due to stringent privacy regulations*.

*General data protection regulation (GDPR) – official legal text, general data protection regulation (GDPR), 2021, https://gdpr-info.eu/ (accessed May 20, 2021).

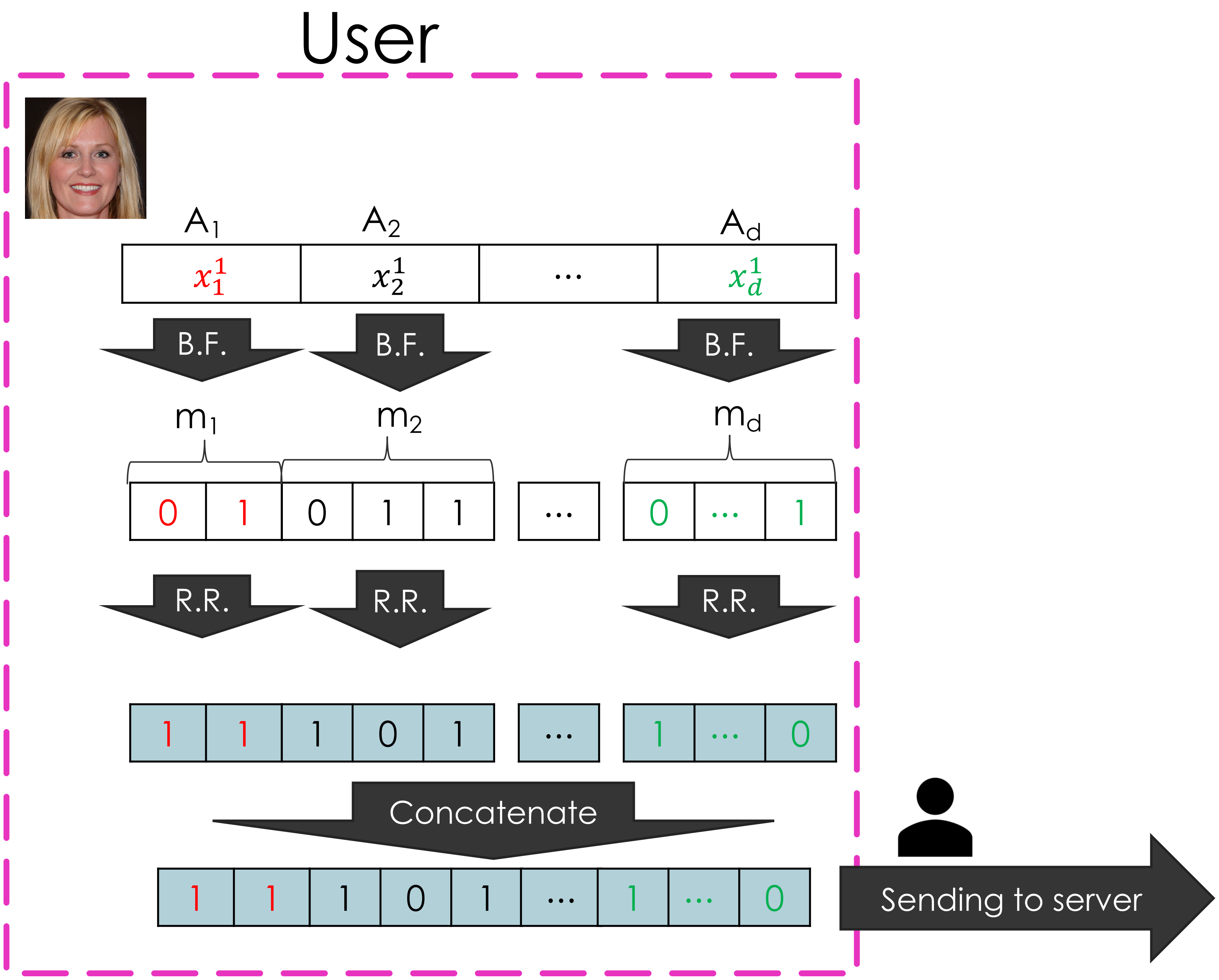
What is LDP?

Local Differential Privacy

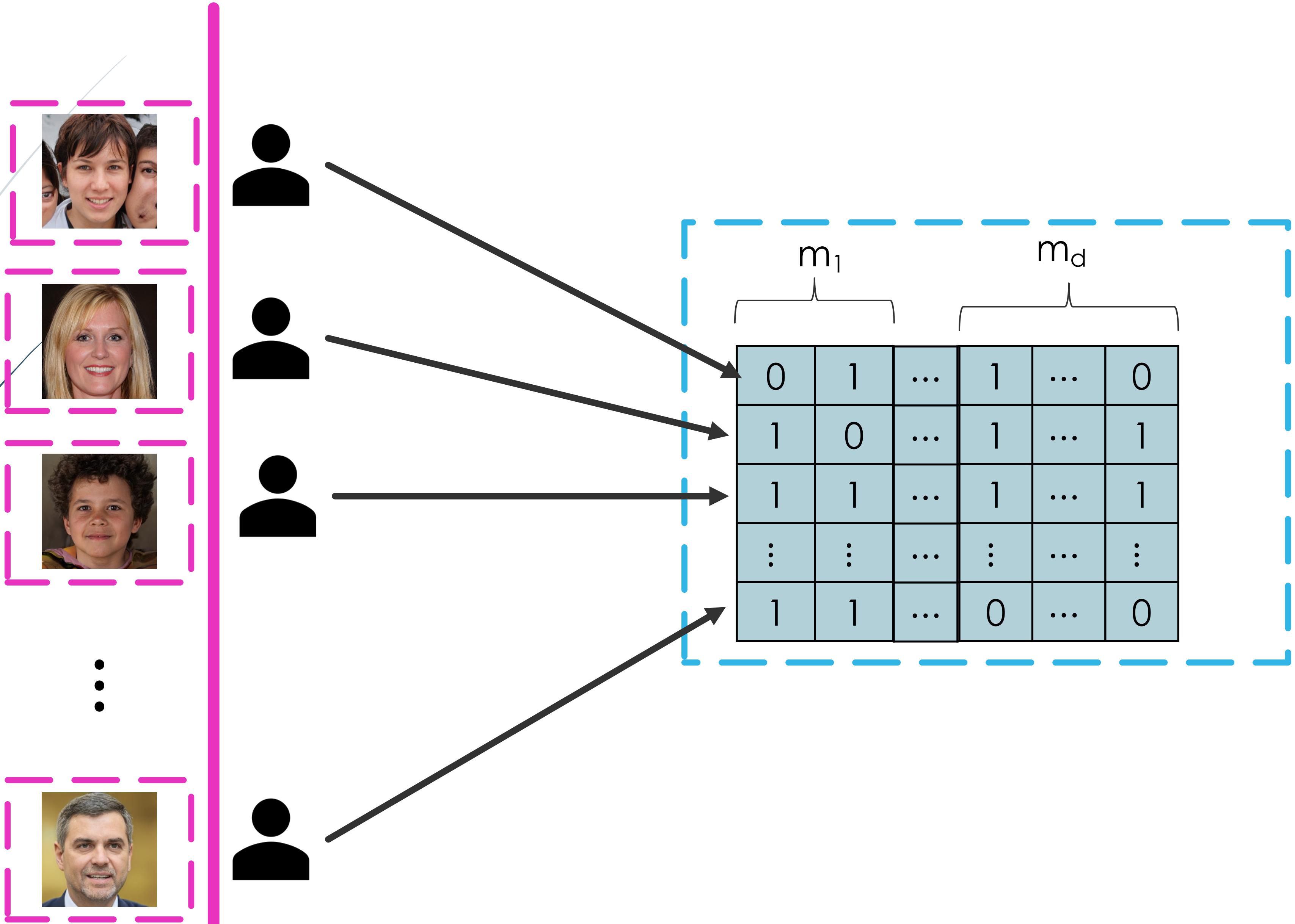


* Face images were taken from https://thispersondoesnotexist.com/

$$m_j = \frac{\ln(\frac{1}{p})}{(\ln 2)^2} |\Omega_j|$$



Sending to the Central Server



Central Server

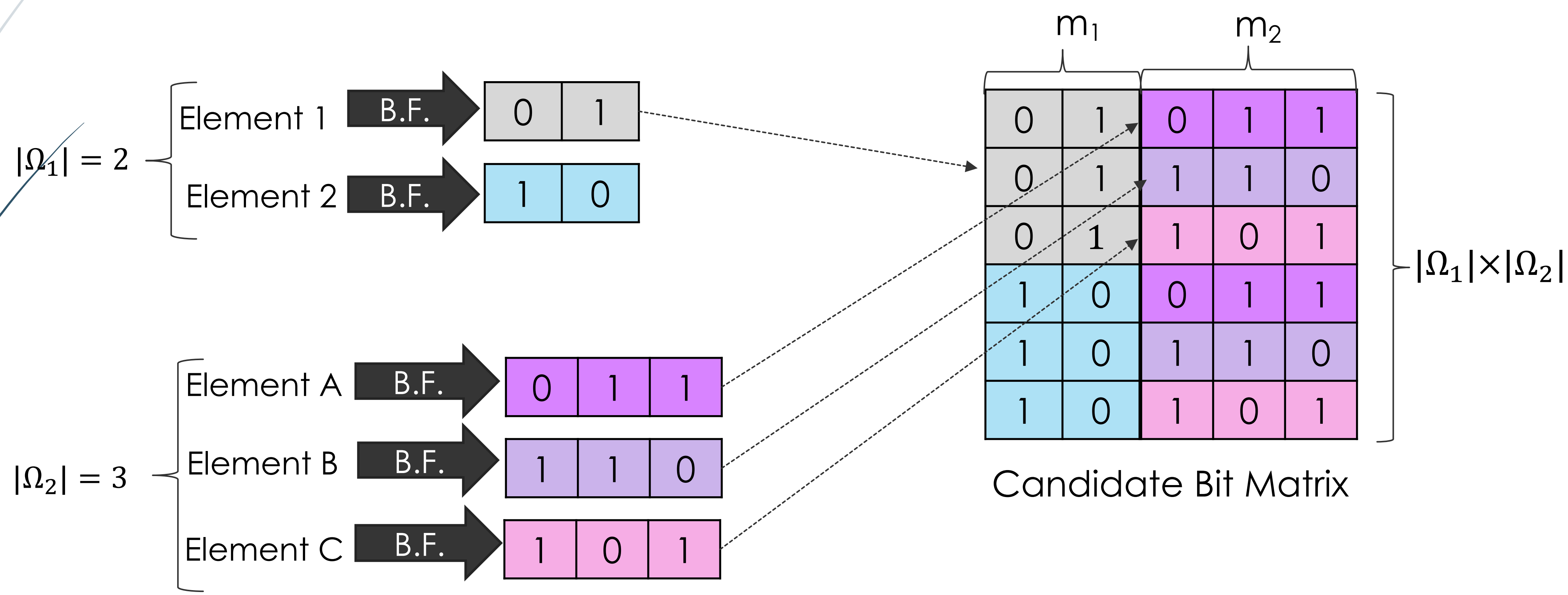
C.S. creates
the candidate Bit Matrix

B.R.R

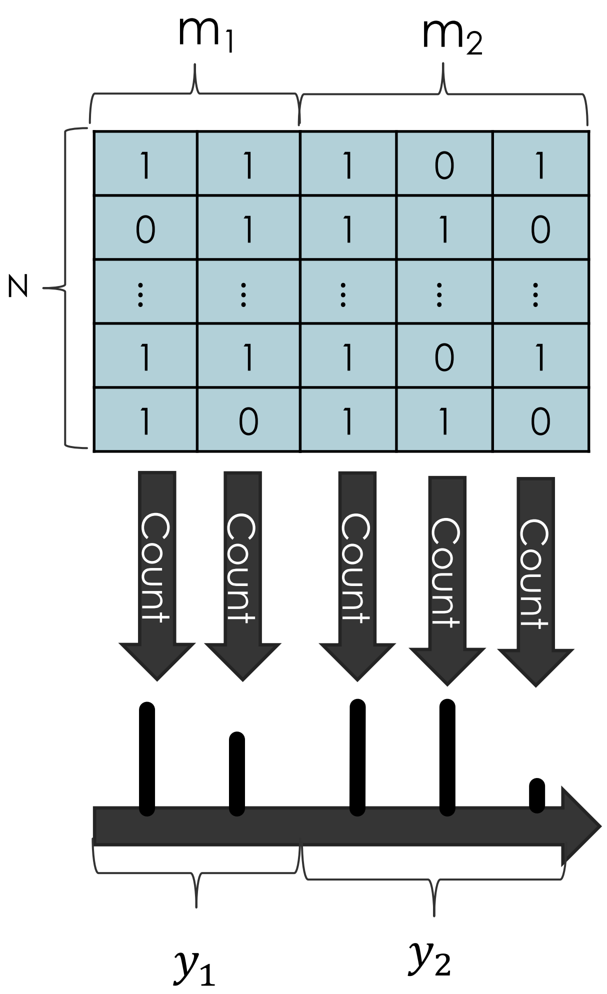
C.S Counts the number of
frequencies
of the perturbed value

$\beta/sum(\beta)$

C.S. creates the candidate Bit Matrix



Counting perturbed values

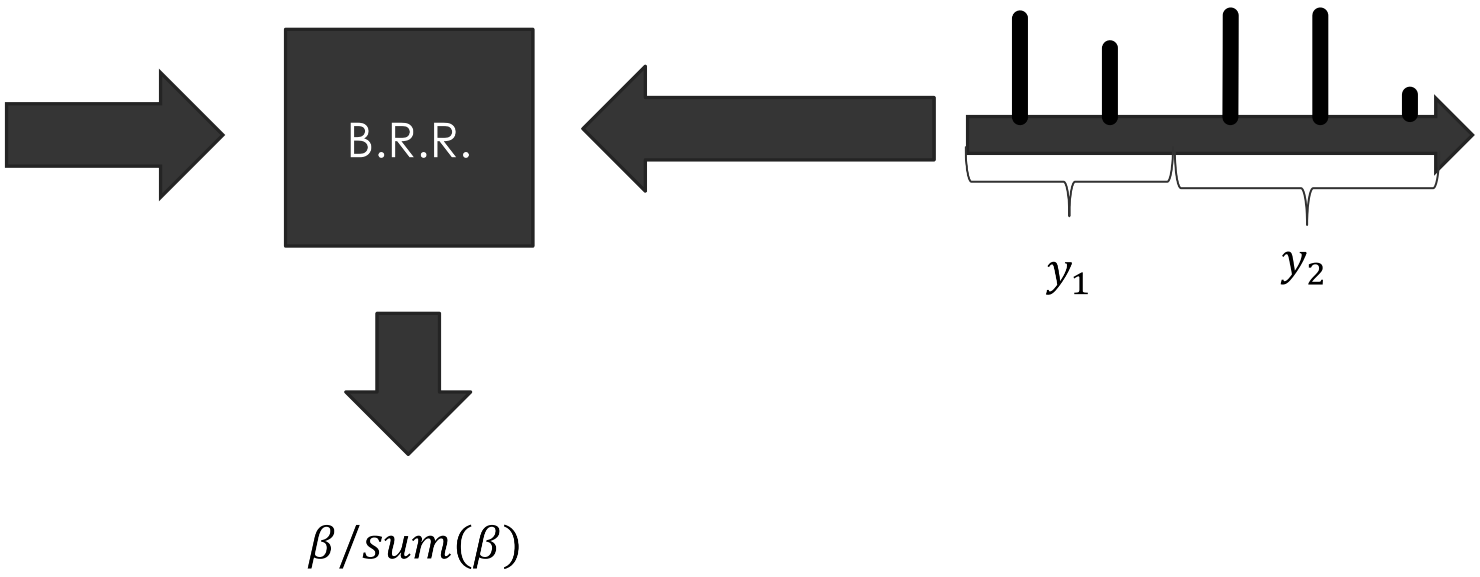


Central Server

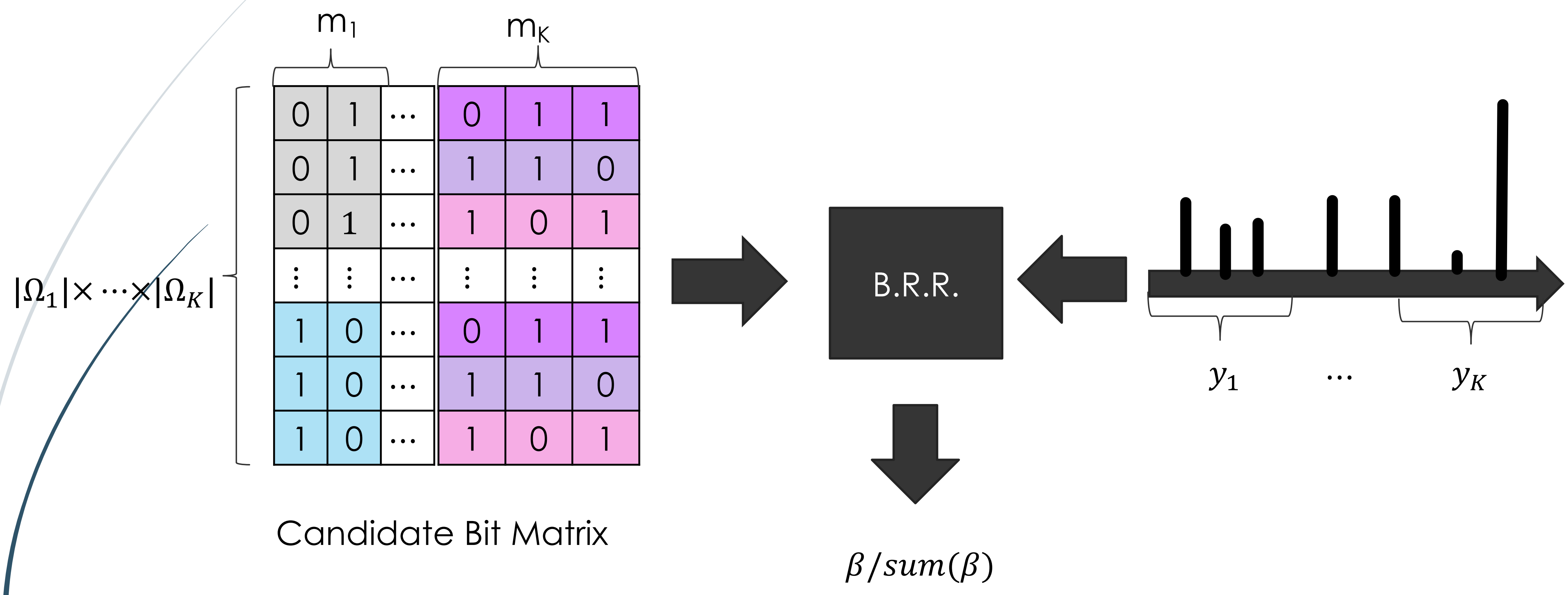
$|\Omega_1| \times |\Omega_2|$

m_1		m_2		
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



Central Server



Proposal vs LoPub vs LoCop

	LoPub[1]	LoCop[2]	Ours
User	Bloom Filters Randomize Response	Bloom Filters Randomize Response	Bloom Filters Randomize Response
Central Server	LASSO	LASSO Gaussian Copula	Bayesian Ridge Regression

One/two-dimensional probability distributions can be efficiently estimated

1) Ren, Xuebin and Yu, Chia-Mu and Yu, Weiren and Yang, Shusen and Yang, Xinyu and McCann, Julie A. and Yu, Philip S., IEEE Transactions on Information Forensics and Security, LoPub: High-Dimensional Crowdsourced Data Publication With Local Differential Privacy, 2018, doi=10.1109/TIFS.2018.2812146.

2) Wang, Teng and Yang, Xinyu and Ren, Xuebin and Yu, Wei and Yang, Shusen, Locally Private High-Dimensional Crowdsourced Data Release Based on Copula Functions, IEEE Transactions on Services Computing, 2022, 15, 2, 778-792.

LASSO VS Bayesian Ridge Regression

- ❌ LASSO often selects only one attribute from a group of highly correlated attributes[3]
- ✔ BRR[4,5] solves the problem of the evaluation of highly correlated attributes.
- ✔ BRR has the ability to incorporate prior information about the parameters and to construct good prior distributions[6].
- ✔ Sambasivan[7] applied BRR in the fields of sparse modeling and machine learning.
- ✔ Assat[8] shown that this approach can be effective in constructing good prior distributions.

3) Konstantin Posch, Maximilian Arbeiter, Juergen Pilz, A novel Bayesian approach for variable selection in linear regression models, Computational Statistics & Data Analysis,Volume 144,2020,106881,ISSN 0167-9473, https://doi.org/10.1016/j.csda.2019.106881.
4) Michimae, H., Emura, T. Bayesian ridge estimators based on copula-based joint prior distributions for regression coefficients. Comput Stat 37, 2741–2769 (2022). https://doi.org/10.1007/s00180-022-01213-8
5) Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Techno- metrics 12:55–67
6) Van Wieringen WN (2021) Lecture notes on ridge regression. arXiv preprint https://arxiv.org/pdf/1509.09169
7) Sambasivan R, Das S, Sahu SK (2020) A Bayesian perspective of statistical machine learning for big data. Comput Stat 35:893–930
8) Assaf AG, Tsionas M, Tasiopoulos A (2019) Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression. Tour Manag 71:1–8

Datasets

Dataset	Users	Attributes
Adult[9]	45,223	8
Ms Fimu[10]	88,936	5
Nursery[11]	12960	9

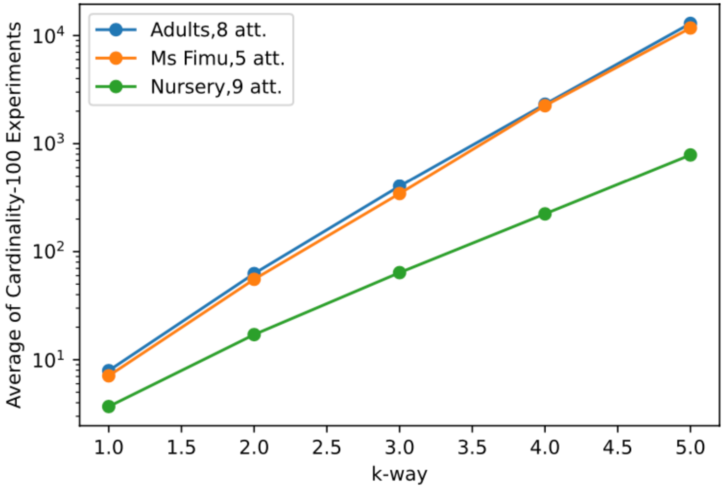
9)Adult, 1996, UCI Machine Learning Repository.
10)Arcolezi HH, Couchot JF, Al Bouna B, Xiao X (2021a) Random sampling plus fake data: multidimensional frequency estimates with local differential privacy. Int Conf Inf Knowl Manag Proc. https://doi.org/10.1145/3459637.3482467
11)Rajkovic,Vladislav. (1997). Nursery. UCI Machine Learning Repository

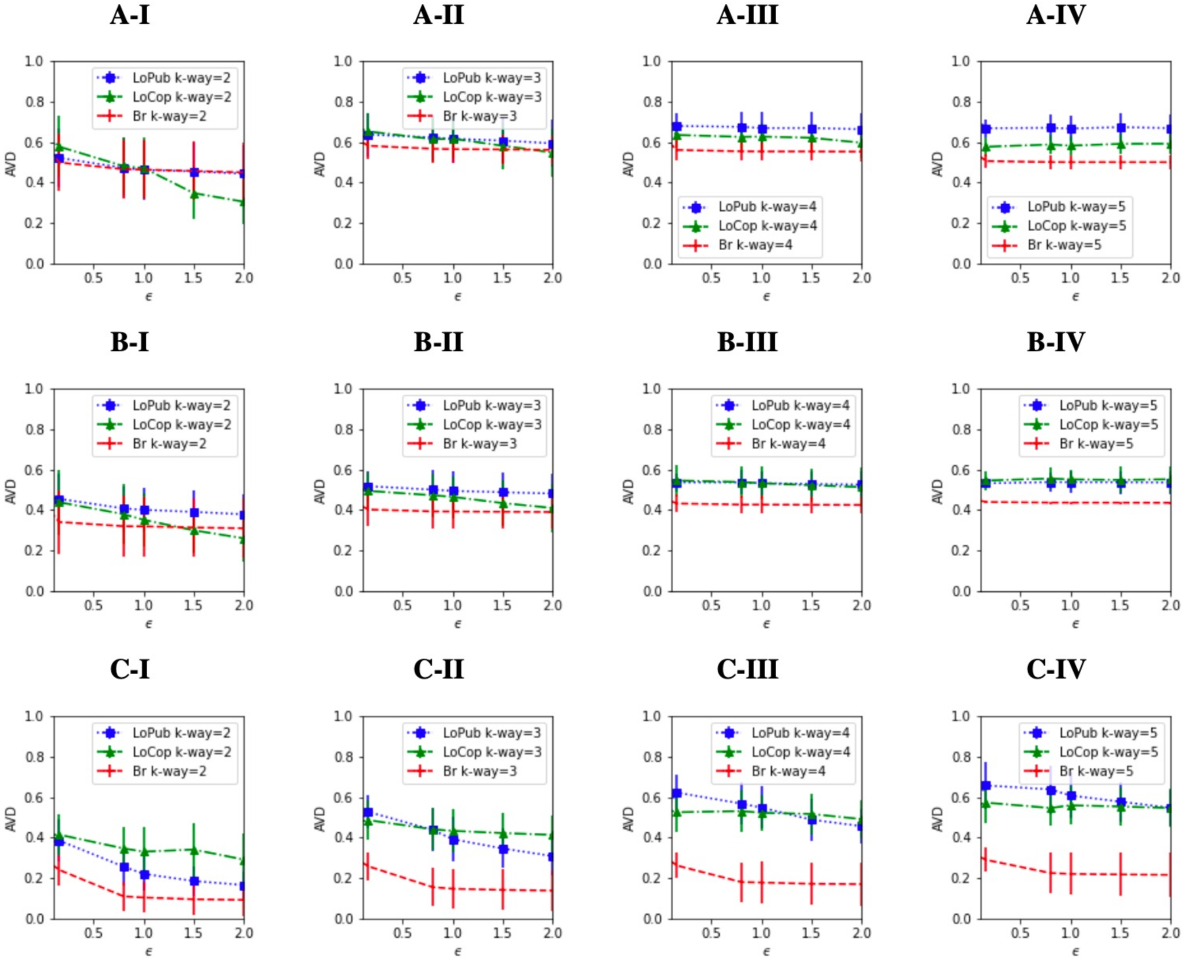
K-way evaluation

We randomly selected k-way joint probabilities of attributes one hundred times. To measure accuracy, we used the distance metric AVD (average variant distance), to quantify the closeness between the probability distributions $P(\omega)$ and $Q(\omega)$.

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

Accuracy K-way





Adult Dataset

MS FIMU Dataset

Nursery Dataset

Conclusions

- This work presents a Bayesian ridge regression approach of an LDP scheme for estimating joint probability.
- The results demonstrate that as the number of attributes k-way increases, the BRR outperforms LoPub and LoCop in terms of the AVD.
- In addition, the performance of the Bayesian ridge algorithm is less impacted by the increase in noise resulting from an increase in the number of users and attributes.
- These findings suggest the BRR can be an effective tool for privacy preservation in data publication
- Future work will involve creating synthetic datasets with varying user quantities, distributions, and cardinalities to evaluate how different element distributions affect the LDP scheme's performance.

An Efficient Local Differential Privacy Scheme Using Bayesian Ridge Regression

Andres Hernandez-Matamoros* and Hiroaki Kikuchi

Thank You for Your Attention!

明治大学
MEIJI UNIVERSITY

PST 2023
2023/08/22

*matamoros@meiji.ac.jp