PERFORMANCE OF HEALTHCARE ANALYSIS UNDER LDP

ANDRES HERNANDEZ-MATAMOROS* AND HIROAKI KIKUCHI
*matamoros@meiji.ac.jp

AGENDA

- Objectives
- Motivation
- What is LDP?
- Castell Approach
  - Anonymization Algorithm
  - Estimating JPD
- Results

OBJECTIVES

- *Secure healthcare data* through anonymization techniques.

- *Estimate* Joint Probability Distributions (*JPD*) to ensure demographic information can be recovered *without compromising individual user privacy*.

MOTIVATION

Sensitive information such as:

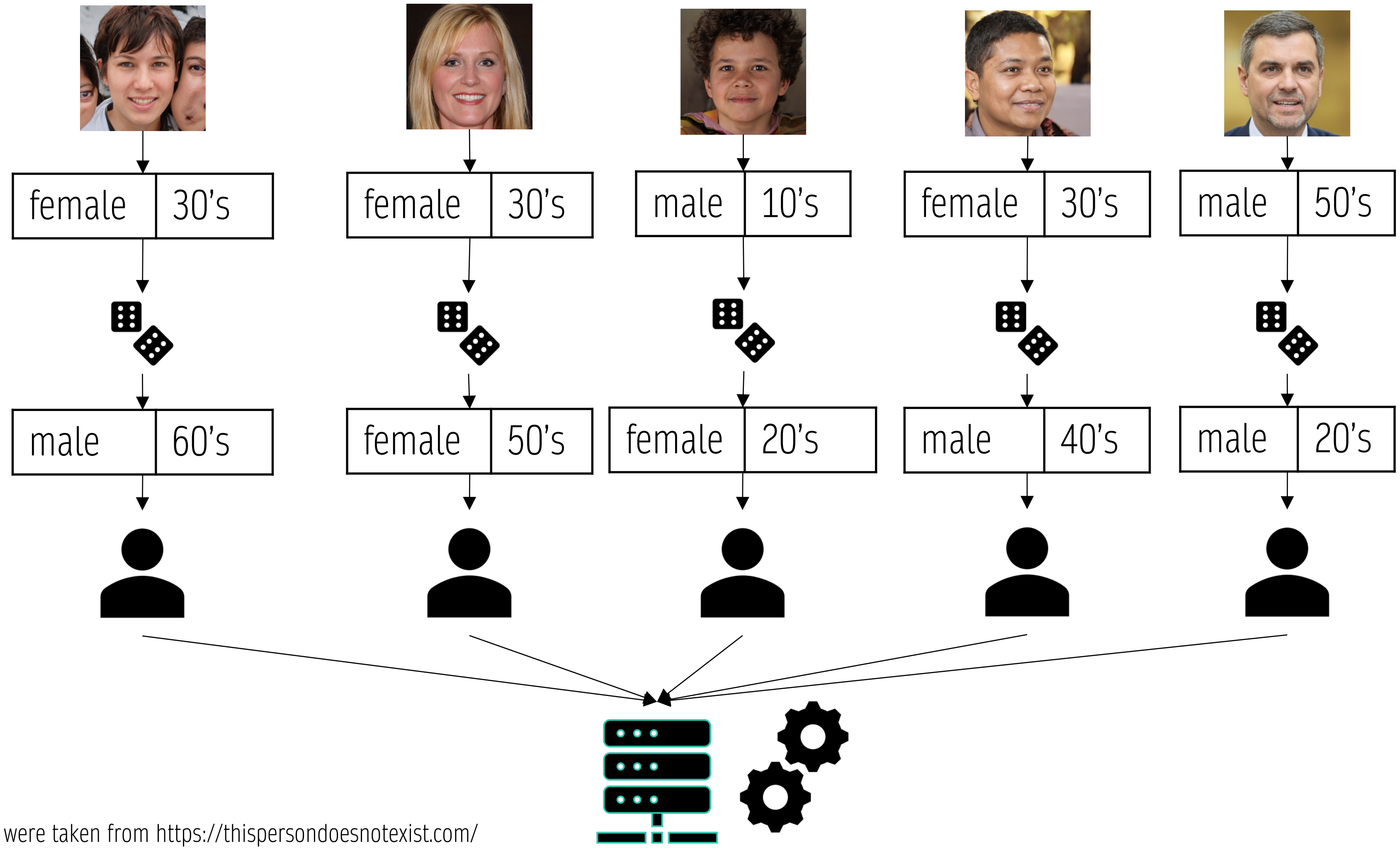
- Diagnoses
- Treatments
- Billing Records

Exposing this information:

- Ethical issues
- Financial issues
- Legal issues

WHAT IS LDP?

Local Differential Privacy



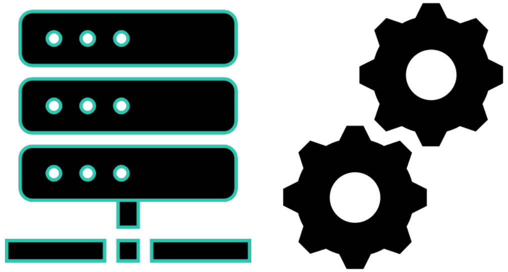
\# USERS
male,10’s ->1
male,20’s ->0
male,30’s ->0
male,40’s ->0
male,50’s ->1
male,60’s ->0
female,10’s ->0
female,20’s ->0
female,30’s ->3
female,40’s ->0
female,50’s ->0
female,60’s ->0

\# USERS
male,10’s ->0
male,20’s ->1
male,30’s ->0
male,40’s ->1
male,50’s ->0
male,60’s ->1
female,10’s ->0
female,20’s ->1
female,30’s ->0
female,40’s ->0
female,50’s ->1
female,60’s ->0

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

LDP'S GOAL?

Noisy # users
male,10's ->0
male,20's ->1
male,30's ->0
male,40's ->1
male,50's ->0
male,60's ->1
female,10's ->0
female,20's ->1
female,30's ->0
female,40's ->0
female,50's ->1
female,60's ->0



How can we recover???

Original # USERS
male,10's ->1
male,20's ->0
male,30's ->0
male,40's ->0
male,50's ->1
male,60's ->0
female,10's ->0
female,20's ->0
female,30's ->3
female,40's ->0
female,50's ->0
female,60's ->0

LDP try to recover demographic information

Never try to recover information of only one user

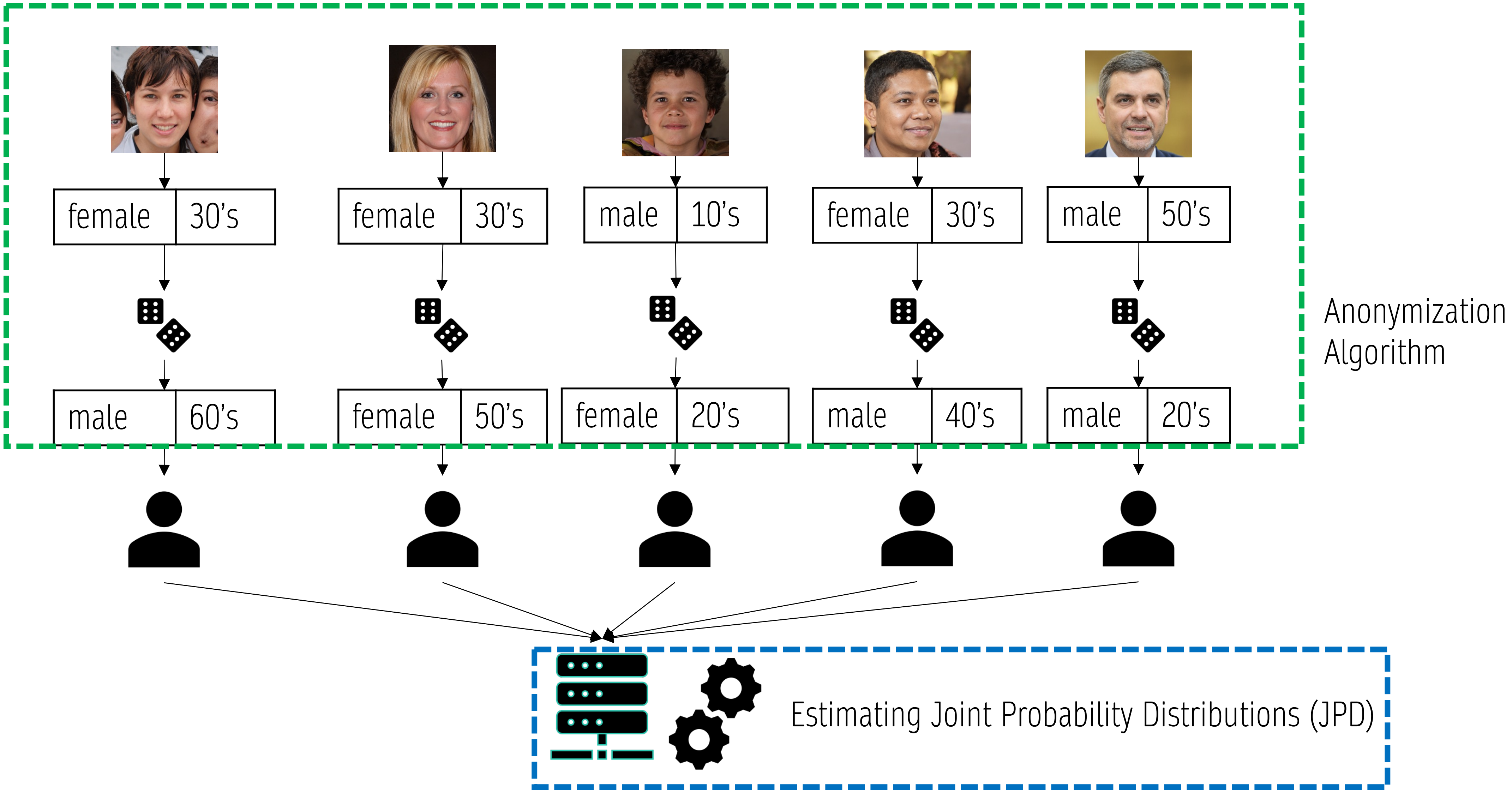
LDP APPROACHES COMPARISON

	Lopub	Locop	Br	Castell
Anonymization Algorithm	Bloom Filters Randomize Response	Bloom Filters Randomize Response	Bloom Filters Randomize Response	Randomize Response
JPD estimation Algorithm	LASSO	LASSO Gaussian Copula	Bayesian Ridge Regression	Inverse of Probability matrix multiplication

One/two-dimensional probability distributions
can be efficiently estimated

Lopub) Ren, X.; Yu, C.M., Yu, W., Yang, S., Yang, X., McCann, J.A. and Philip, S.Y. LoPub: High-Dimensional Crowdsourced Data Publication with Local Differential Privacy. IEEE Trans. Inf. Forensics Secur. 2018, 13, 2151–2166. https://doi.org/10.1109/TIFS.2018.2812146
Locop) Wang, T.; Yang, X.; Ren, X.; Yu, W.; Yang, S. Locally Private High-Dimensional Crowdsourced Data Release Based on Copula Functions. IEEE Trans. Serv. Comput. 2022, 15, 778–792. https://doi.org/10.1109/TSC.2019.2961092
Br) Hernandez-Matamoros, Andres, and Hiroaki Kikuchi. 2024. "Comparative Analysis of Local Differential Privacy Schemes in Healthcare Datasets" Applied Sciences 14, no. 7: 2864. https://doi.org/10.3390/app14072864
Castell) Hiroaki Kikuchi, Castell: Scalable Joint Probability Estimation of Multi-dimensional Data Randomized with Local Differential Privacy. 2022, arXiv preprint, https://arxiv.org/abs/2212.01627.

PROPOSED APPROACH



ANONYMIZATION ALGORITHM

	U_1	U_2	U_3
U^1	Male	Chinese	Standard
U^2	Female	Mexican	Underweight
⋮	⋮	⋮	⋮
U^{N-1}	Male	Korean	Overweight
U^N	Female	Mexican	Standard

RR

$\widehat{U_1}$	$\widehat{U_2}$	$\widehat{U_3}$	
Female	Korean	Overweight	To the server
Male	Mexican	Standard	To the server
⋮	⋮	⋮	⋮
Female	Chinese	Underweight	To the server
Female	Korean	Overweight	To the server

Privacy budget ε

$$\Omega = \{10\text{'}s, 20\text{'}s, 30\text{'}s, 40\text{'}s, 50\text{'}s, 60\text{'}s\}$$

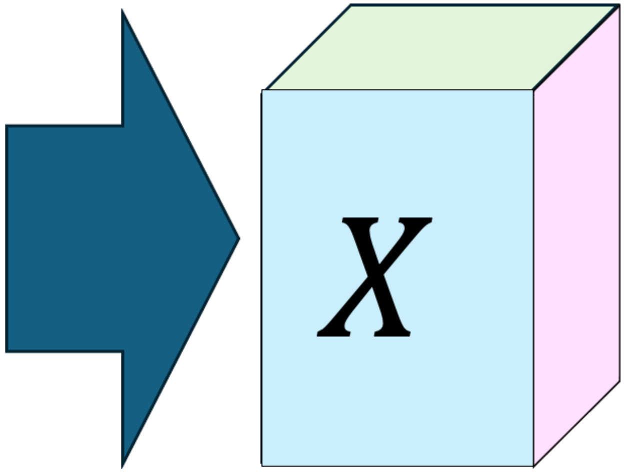
$$|\Omega| = 6$$

$$p = \frac{e^{\varepsilon}}{e^{\varepsilon} + |\Omega| - 1} \qquad q = \frac{1}{e^{\varepsilon} + |\Omega| - 1}$$

```
Randomize Response
sample= random(0, 1)
computes p,q
if sample > p - q:
        out = random (Ω)
else:
        out = original value
```

ESTIMATING JOINT PROBABILITY DISTRIBUTIONS (JPD) *CASTELL

	$\widehat{U_1}$	$\widehat{U_2}$	$\widehat{U_3}$
$\widehat{U^1}$	Female	Chinese	Overweight
$\widehat{U^2}$	Male	Mexican	Underweight
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\widehat{U^{N-1}}$	Female	Korean	Standard
$\widehat{U^N}$	Female	Mexican	Overweight



$$\begin{array}{cc} & |\Omega_1| \\ |\Omega_1| & \boxed{P_1^{-1}} \end{array}$$

$$P_{v,l} = \begin{cases} \dfrac{1-p}{|\Omega|-1} & if\ v \neq l, \\ p & if\ v = l, \end{cases} \quad \text{where} \quad p = \frac{e^{\varepsilon}}{e^{\varepsilon} + |\Omega| - 1}$$

*Hiroaki Kikuchi, Castell: Scalable Joint Probability Estimation of Multi-dimensional Data Randomized with Local Differential Privacy. 2022, arXiv preprint,https://arxiv.org/abs/2212.01627.

ESTIMATING JOINT PROBABILITY DISTRIBUTIONS (JPD)

$\underbrace{P_1^{-1}}_{|\Omega_1| \times |\Omega_1|} \times$

	Overweight	Underweight	Standard
	Korean(K)		
Male(M)	$X_{M,K,O}$	$X_{M,K,U}$	$X_{M,K,S}$
Female(F)	$X_{F,K,O}$	$X_{F,K,U}$	$X_{F,K,S}$

$\underbrace{P_1^{-1}}_{|\Omega_1| \times |\Omega_1|} \times$

	Overweight	Underweight	Standard
	Mexican(MX)		
Male(M)	$X_{M,MX,O}$	$X_{M,MX,U}$	$X_{M,MX,S}$
Female(F)	$X_{F,MX,O}$	$X_{F,MX,U}$	$X_{F,MX,S}$

$\underbrace{P_1^{-1}}_{|\Omega_1| \times |\Omega_1|} \times$

	Overweight(O)	Underweight(U)	Standard(S)
	Chinese(C)		
Male(M)	$X_{M,C,O}$	$X_{M,C,U}$	$X_{M,C,S}$
Female(F)	$X_{F,C,O}$	$X_{F,C,U}$	$X_{F,C,S}$

$\Rightarrow$

	O	U	S
	C		
M	$\tau^1_{M,C,O}$	$\tau^1_{M,C,U}$	$\tau^1_{M,C,S}$
F	$\tau^1_{F,C,O}$	$\tau^1_{F,C,U}$	$\tau^1_{F,C,S}$

	O	U	S
	MX		
M	$\tau^1_{M,MX,O}$	$\tau^1_{M,MX,U}$	$\tau^1_{M,MX,S}$
F	$\tau^1_{F,MX,O}$	$\tau^1_{F,MX,U}$	$\tau^1_{F,MX,S}$

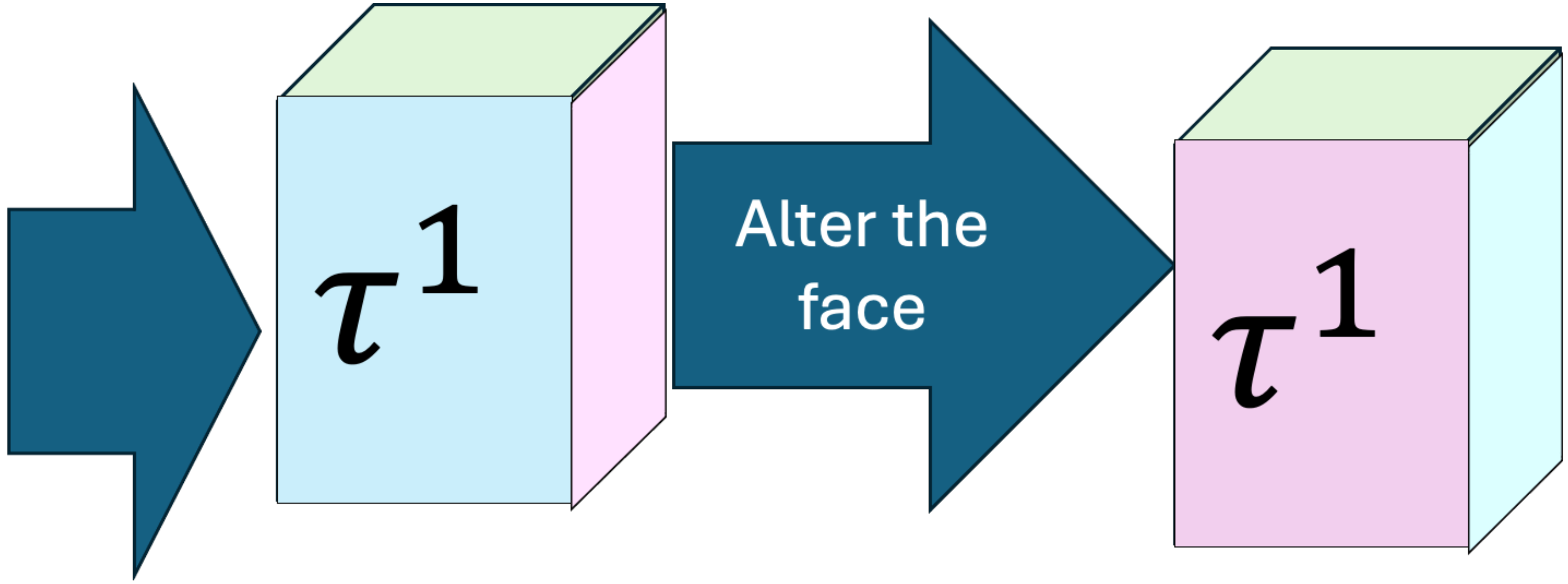
	O	U	S
	K		
M	$\tau^1_{M,K,O}$	$\tau^1_{M,K,U}$	$\tau^1_{M,K,S}$
F	$\tau^1_{F,K,O}$	$\tau^1_{F,K,U}$	$\tau^1_{F,K,S}$

ESTIMATING JOINT PROBABILITY DISTRIBUTIONS (JPD)

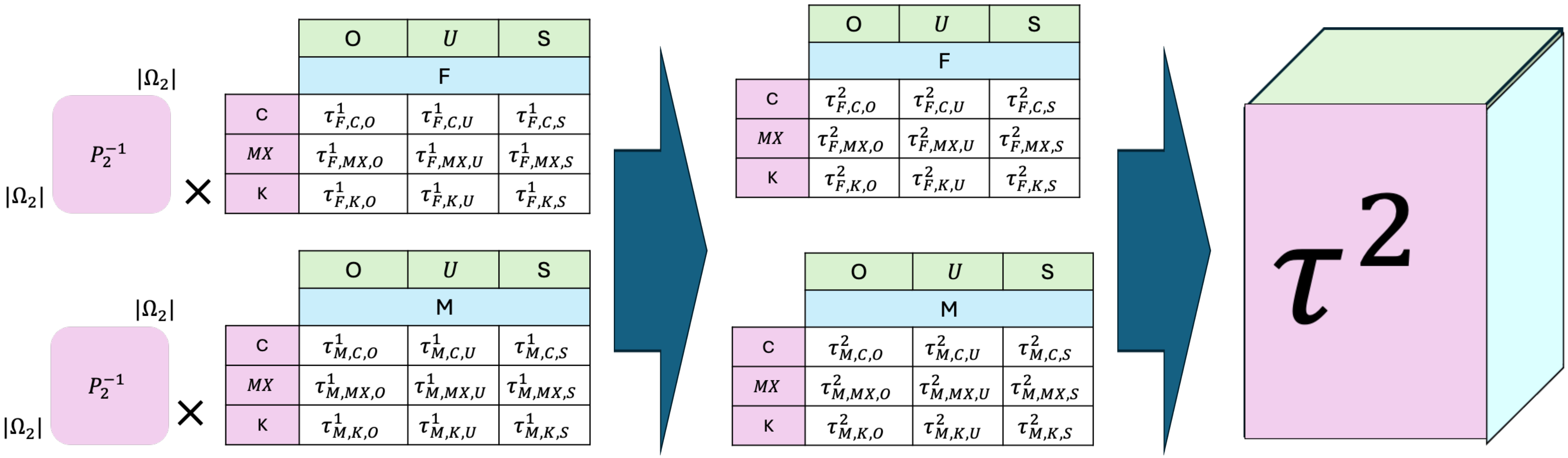
	O	U	S
	C		
M	$\tau^1_{M,C,O}$	$\tau^1_{M,C,U}$	$\tau^1_{M,C,S}$
F	$\tau^1_{F,C,O}$	$\tau^1_{F,C,U}$	$\tau^1_{F,C,S}$

	O	U	S
	MX		
M	$\tau^1_{M,MX,O}$	$\tau^1_{M,MX,U}$	$\tau^1_{M,MX,S}$
F	$\tau^1_{F,MX,O}$	$\tau^1_{F,MX,U}$	$\tau^1_{F,MX,S}$

	O	U	S
	K		
M	$\tau^1_{M,K,O}$	$\tau^1_{M,K,U}$	$\tau^1_{M,K,S}$
F	$\tau^1_{F,K,O}$	$\tau^1_{F,K,U}$	$\tau^1_{F,K,S}$



ESTIMATING JOINT PROBABILITY DISTRIBUTIONS (JPD)



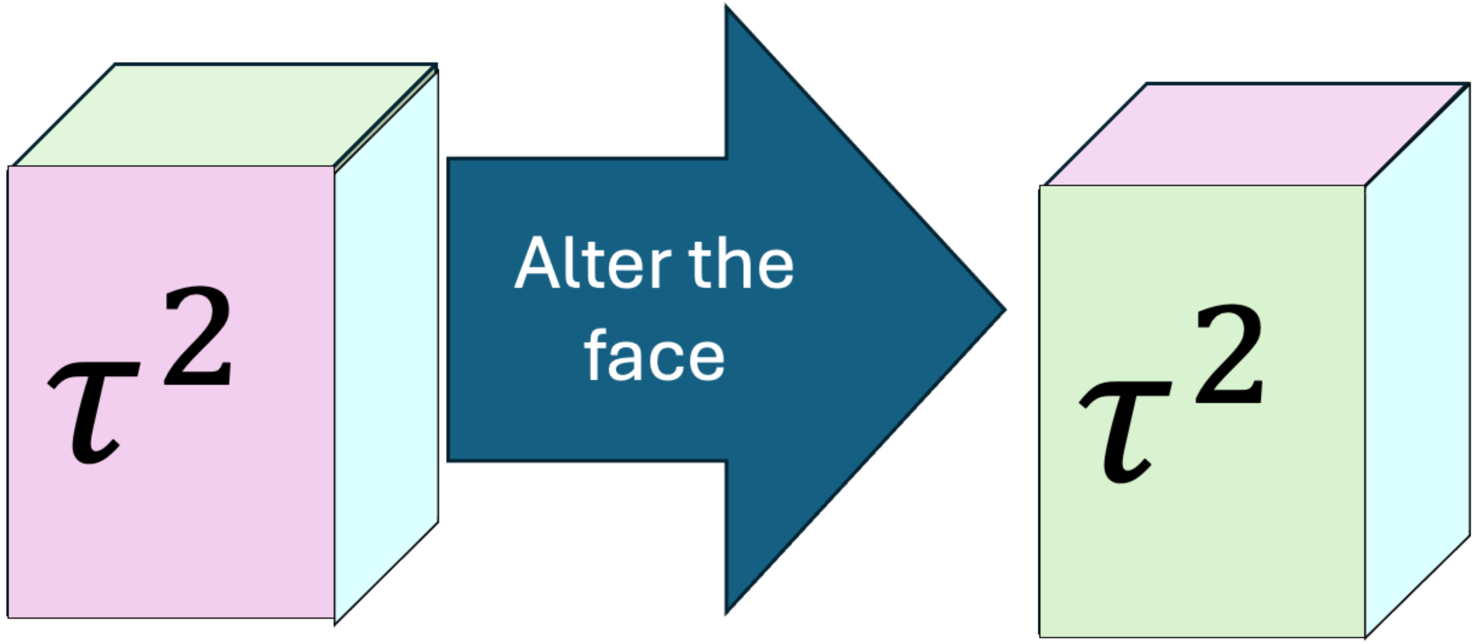
	O	U	S
	F		
C	$\tau^1_{F,C,O}$	$\tau^1_{F,C,U}$	$\tau^1_{F,C,S}$
MX	$\tau^1_{F,MX,O}$	$\tau^1_{F,MX,U}$	$\tau^1_{F,MX,S}$
K	$\tau^1_{F,K,O}$	$\tau^1_{F,K,U}$	$\tau^1_{F,K,S}$

	O	U	S
	M		
C	$\tau^1_{M,C,O}$	$\tau^1_{M,C,U}$	$\tau^1_{M,C,S}$
MX	$\tau^1_{M,MX,O}$	$\tau^1_{M,MX,U}$	$\tau^1_{M,MX,S}$
K	$\tau^1_{M,K,O}$	$\tau^1_{M,K,U}$	$\tau^1_{M,K,S}$

	O	U	S
	F		
C	$\tau^2_{F,C,O}$	$\tau^2_{F,C,U}$	$\tau^2_{F,C,S}$
MX	$\tau^2_{F,MX,O}$	$\tau^2_{F,MX,U}$	$\tau^2_{F,MX,S}$
K	$\tau^2_{F,K,O}$	$\tau^2_{F,K,U}$	$\tau^2_{F,K,S}$

	O	U	S
	M		
C	$\tau^2_{M,C,O}$	$\tau^2_{M,C,U}$	$\tau^2_{M,C,S}$
MX	$\tau^2_{M,MX,O}$	$\tau^2_{M,MX,U}$	$\tau^2_{M,MX,S}$
K	$\tau^2_{M,K,O}$	$\tau^2_{M,K,U}$	$\tau^2_{M,K,S}$

ESTIMATING JOINT PROBABILITY DISTRIBUTIONS (JPD)



ESTIMATING JOINT PROBABILITY DISTRIBUTIONS (JPD)

$|\Omega_3|$

$|\Omega_3|$ P_3^{-1} $\times$

	C	MX	K
	F		
O	$\tau^2_{F,C,O}$	$\tau^2_{F,MX,O}$	$\tau^2_{F,K,O}$
U	$\tau^2_{F,C,U}$	$\tau^2_{F,MX,U}$	$\tau^2_{F,K,U}$
S	$\tau^2_{F,C,S}$	$\tau^2_{F,MX,S}$	$\tau^2_{F,K,S}$

$|\Omega_3|$

$|\Omega_3|$ P_3^{-1} $\times$

	C	MX	K
	M		
O	$\tau^2_{M,C,O}$	$\tau^2_{M,MX,O}$	$\tau^2_{M,K,O}$
U	$\tau^2_{M,C,U}$	$\tau^2_{M,MX,U}$	$\tau^2_{M,K,U}$
S	$\tau^2_{M,C,S}$	$\tau^2_{M,MX,S}$	$\tau^2_{M,K,S}$

	C	MX	K
	F		
O	$\tau^3_{F,C,O}$	$\tau^3_{F,MX,O}$	$\tau^3_{F,K,O}$
U	$\tau^3_{F,C,U}$	$\tau^3_{F,MX,U}$	$\tau^3_{F,K,U}$
S	$\tau^3_{F,C,S}$	$\tau^3_{F,MX,S}$	$\tau^3_{F,K,S}$

	C	MX	K
	M		
O	$\tau^3_{M,C,O}$	$\tau^3_{M,MX,O}$	$\tau^3_{M,K,O}$
U	$\tau^3_{M,C,U}$	$\tau^3_{M,MX,U}$	$\tau^3_{M,K,U}$
S	$\tau^3_{M,C,S}$	$\tau^3_{M,MX,S}$	$\tau^3_{M,K,S}$

ESTIMATING JOINT PROBABILITY DISTRIBUTIONS (JPD)

	C	MX	K
	F		
O	$\tau^3_{F,C,O}$	$\tau^3_{F,MX,O}$	$\tau^3_{F,K,O}$
U	$\tau^3_{F,C,U}$	$\tau^3_{F,MX,U}$	$\tau^3_{F,K,U}$
S	$\tau^3_{F,C,S}$	$\tau^3_{F,MX,S}$	$\tau^3_{F,K,S}$

	C	MX	K
	M		
O	$\tau^3_{M,C,O}$	$\tau^3_{M,MX,O}$	$\tau^3_{M,K,O}$
U	$\tau^3_{M,C,U}$	$\tau^3_{M,MX,U}$	$\tau^3_{M,K,U}$
S	$\tau^3_{M,C,S}$	$\tau^3_{M,MX,S}$	$\tau^3_{M,K,S}$

JPD

DATASETS

Dataset	#Patients (N)	Attributes
Skin Cancer	10,015	5
Nursery	12,960	9
Diabetes	70,592	18

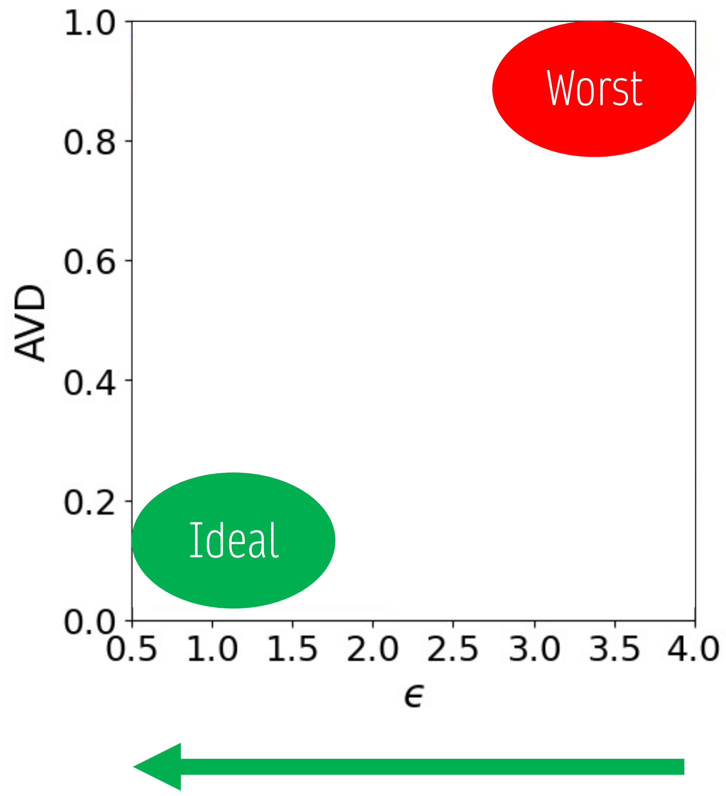
RESULTS

A subset of k attributes is randomly chosen from each dataset, and their JPD of k attributes is estimated, repeating this process one hundred times.

To assess the accuracy of the JPD estimation, the average variant distance (AVD) metric is employed to quantify the difference between the true and estimated JPD.

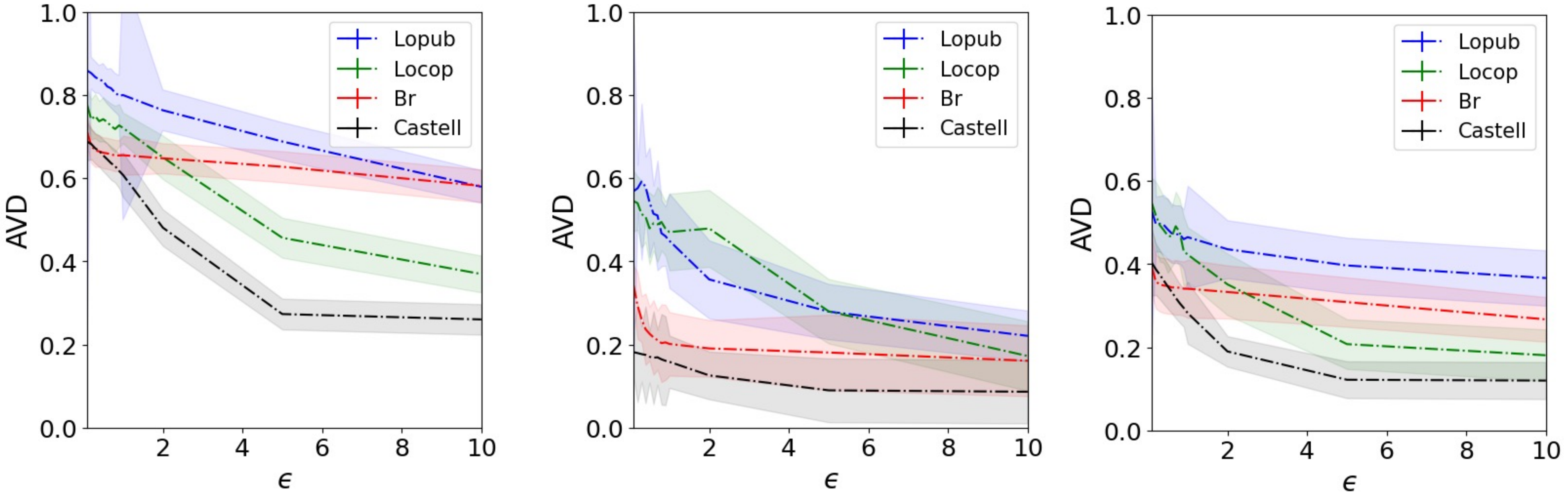
$$AVD = \frac{1}{2}\sum_{w \in \Theta} |O(\Theta) - S(\Theta)|$$

TRADE OFF DATA UTILITY VS PRIVACY



Strong privacy

RESULTS



Skin Cancer | Nursery | Diabetes

AVD vs Privacy Budget (ε), estimating JPD of 3 attributes

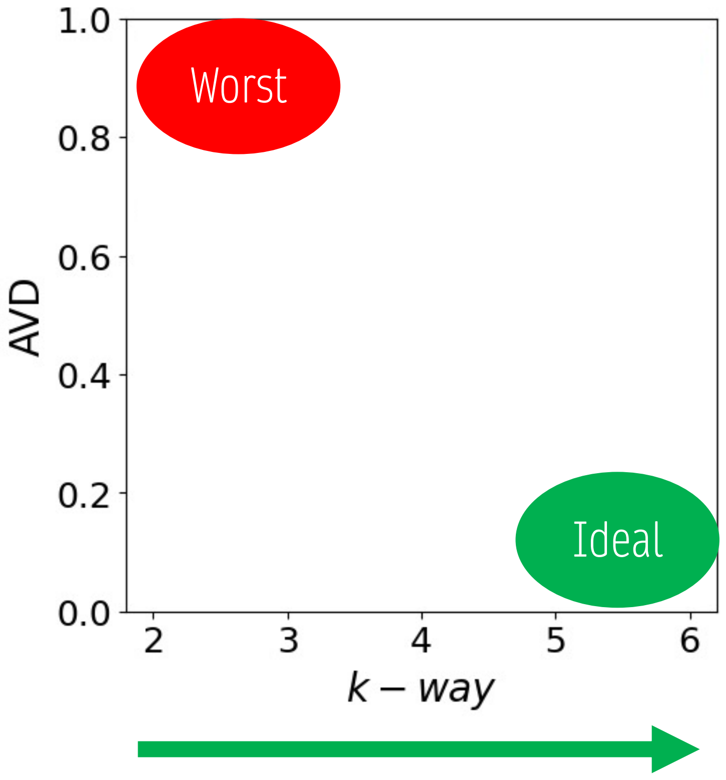
Lopub) Ren, X.; Yu, C.M., Yu, W., Yang, S., Yang, X., McCann, J.A. and Philip, S.Y. LoPub: High-Dimensional Crowdsourced Data Publication with Local Differential Privacy. IEEE Trans. Inf. Forensics Secur. 2018, 13, 2151–2166. https://doi.org/10.1109/TIFS.2018.2812146

Locop) Wang, T.; Yang, X.; Ren, X.; Yu, W.; Yang, S. Locally Private High-Dimensional Crowdsourced Data Release Based on Copula Functions. IEEE Trans. Serv. Comput. 2022, 15, 778–792. https://doi.org/10.1109/TSC.2019.2961092

Br) Hernandez-Matamoros, Andres, and Hiroaki Kikuchi. 2024. "Comparative Analysis of Local Differential Privacy Schemes in Healthcare Datasets" Applied Sciences 14, no. 7: 2864. https://doi.org/10.3390/app14072864

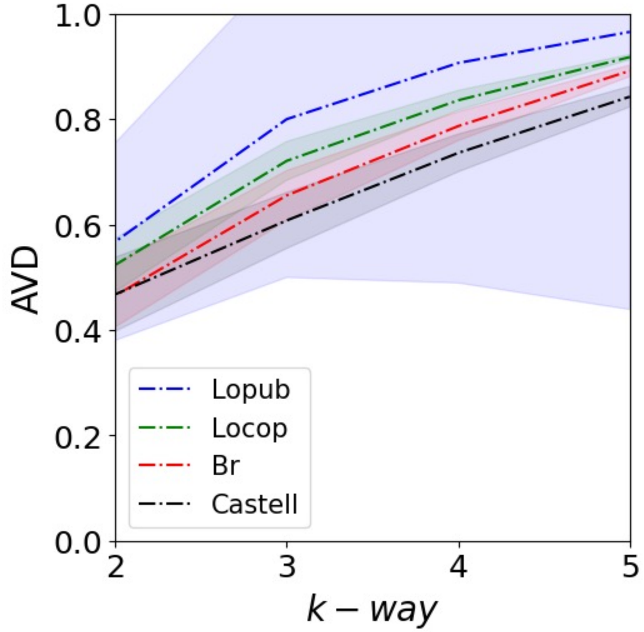
Castell) Hiroaki Kikuchi, Castell: Scalable Joint Probability Estimation of Multi-dimensional Data Randomized with Local Differential Privacy. 2022, arXiv preprint, https://arxiv.org/abs/2212.01627.

TRADE OFF DATA UTILITY VS K-WAY

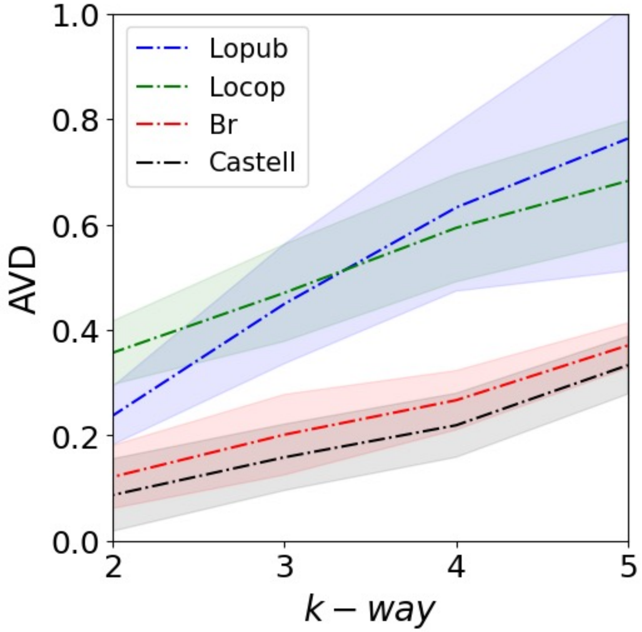


Attributes

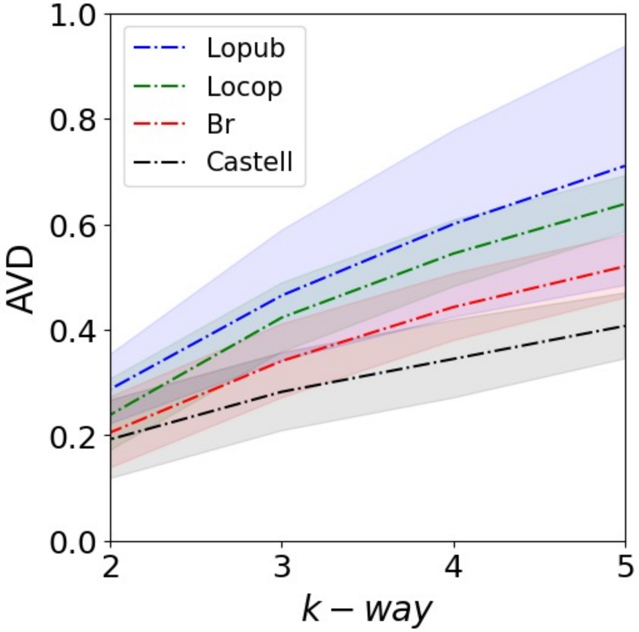
RESULTS



Skin Cancer



Nursery



Diabetes

AVD vs k-way with Privacy Budget (ε) set as 1

Lopub) Ren, X.; Yu, C.M., Yu, W., Yang, S., Yang, X., McCann, J.A. and Philip, S.Y. LoPub: High-Dimensional Crowdsourced Data Publication with Local Differential Privacy. IEEE Trans. Inf. Forensics Secur. 2018, 13, 2151–2166. https://doi.org/10.1109/TIFS.2018.2812146

Locop) Wang, T.; Yang, X.; Ren, X.; Yu, W.; Yang, S. Locally Private High-Dimensional Crowdsourced Data Release Based on Copula Functions. IEEE Trans. Serv. Comput. 2022, 15, 778–792. https://doi.org/10.1109/TSC.2019.2961092

Br) Hernandez-Matamoros, Andres, and Hiroaki Kikuchi. 2024. "Comparative Analysis of Local Differential Privacy Schemes in Healthcare Datasets" Applied Sciences 14, no. 7: 2864. https://doi.org/10.3390/app14072864

Castell) Hiroaki Kikuchi, Castell: Scalable Joint Probability Estimation of Multi-dimensional Data Randomized with Local Differential Privacy. 2022, arXiv preprint, https://arxiv.org/abs/2212.01627.

CONCLUSIONS

Four LDP approaches was tested.

Castell stood out for its ability to maintain a balance between privacy and accuracy

Future work, uses JPD to train Machine Learning models

THANK YOU

ANDRES HERNANDEZ-MATAMOROS
MATAMOROS@MEIJI.AC.JP
HTTPS://PHDMATAMOROS.GITHUB.IO/AGHM-CV/