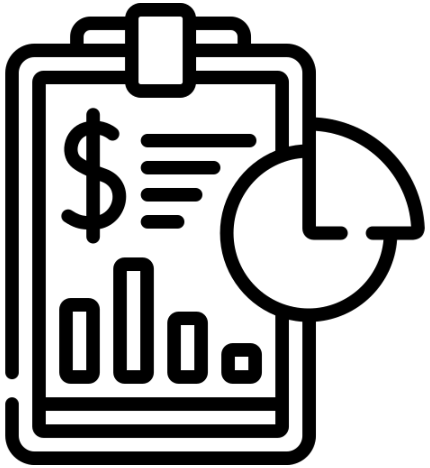
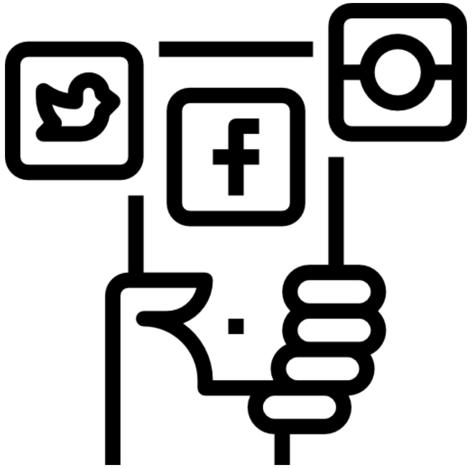
Differentially Private Synthetic Data Generation with Missing Data

Shubhankar Mohapatra, Jianqiao Zong, Florian Kerschbaum, Xi He

To be presented at VLDB 2024

UNIVERSITY OF WATERLOO
Data Systems Group





26 Billion Records Exposed in Data Breach – How To Check if You're Affected

The unprecedented leak has affected LinkedIn, Twitter/X, Dropbox, and many more popular sites.

Written by · Published on

Nova Scotia

Personal data of 50,000 N.S. health-care workers may have been leaked through pension plan

Names, birthdays, addresses, social insurance numbers among the information compromised

Taryn Grant · CBC News · Posted: Mar 06, 2021 1:11 PM EST | Last Updated: March 6, 2021

'Anonymised' data can never be totally anonymous, says study

Findings say it is impossible for researchers to fully protect real identities in datasets

Netflix's Impending (But Still Avoidable) Multi-Million Dollar Privacy Blunder

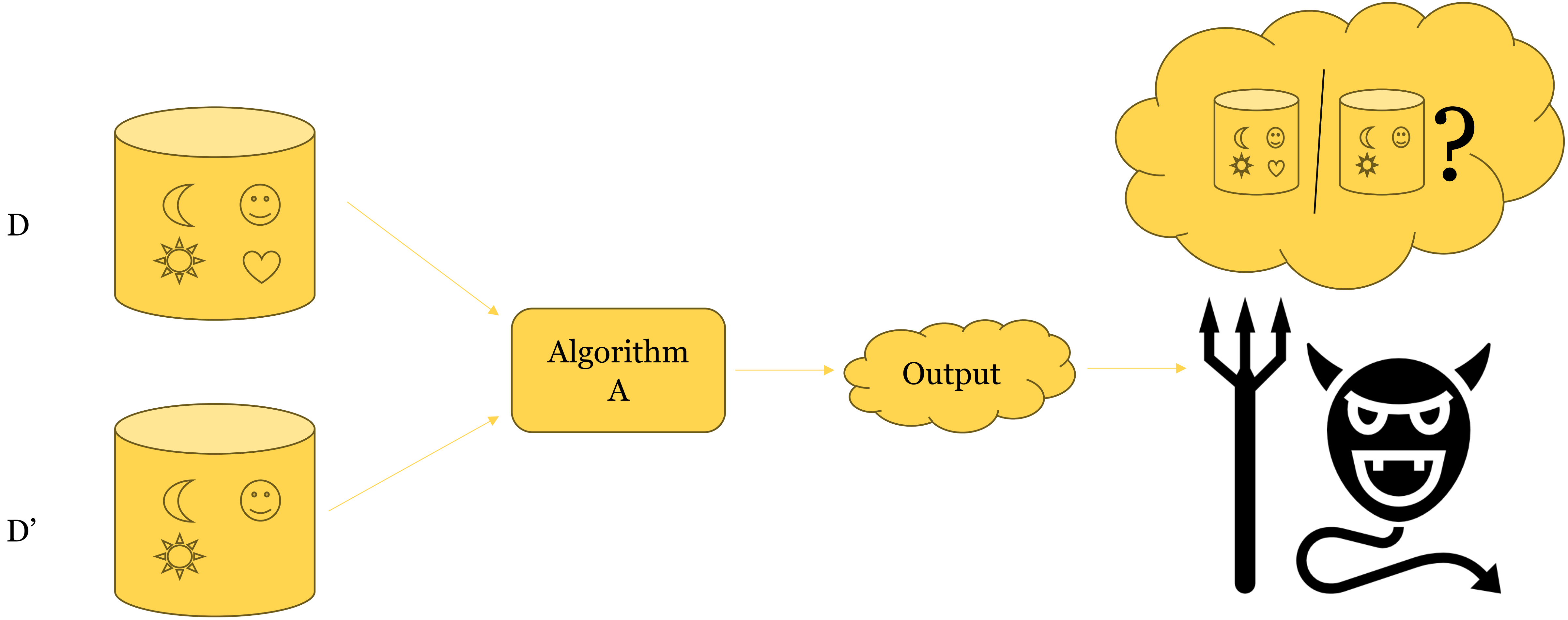
SEPTEMBER 21, 2009 BY PAUL OHM

In my **last post**, I had promised to say more about **my article on the limits of anonymization and the power of reidentification**. Although I haven't said anything for a few weeks, others have, and I especially appreciate posts by **Susannah Fox**, **Seth Schoen**, and **Nate Anderson**. Not only have these people summarized my article well, they have also added a lot of insightful commentary, and I commend these three posts to you.

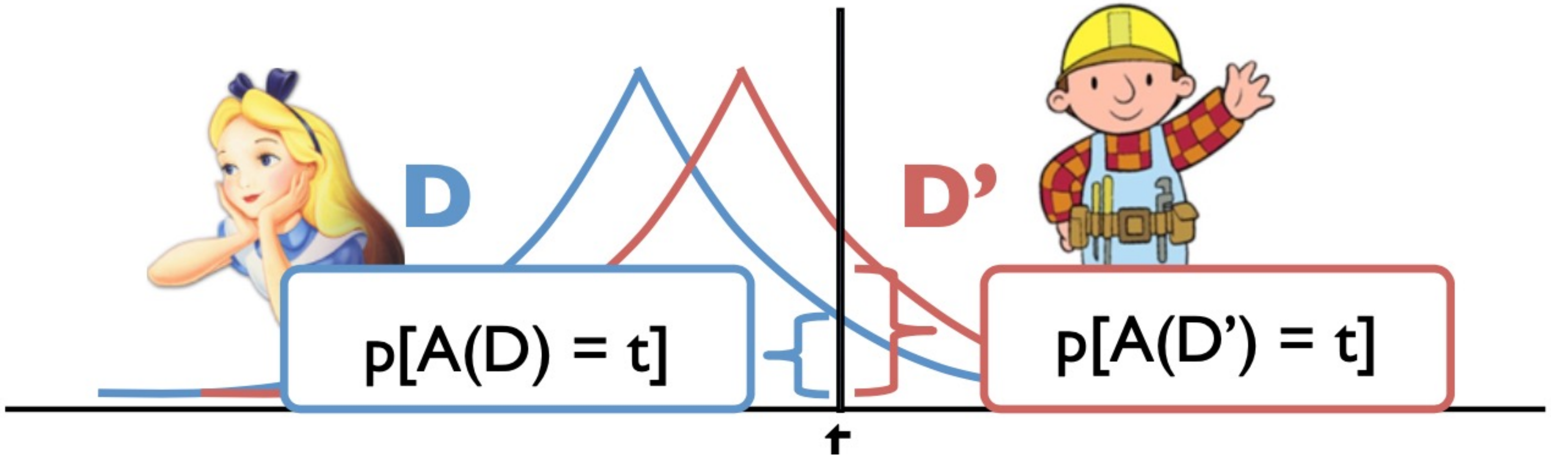
Today brings news relating to one of the central examples in my paper: Netflix has announced plans to commit a privacy blunder that could cost it millions of dollars in fines and civil damages.

In my article, I focus on Netflix's 2006 decision to release millions of records containing the

Differential Privacy



Differential Privacy



A randomized algorithm A: $\mathcal{D} \rightarrow \mathcal{R}$ satisfies (ε, δ)-differential privacy (DP) if for any two adjacent inputs $\mathcal{D}$, $\mathcal{D}' \in \mathcal{D}$ that differ in an entry and for any subset of outputs $t \subseteq R$ it holds that :

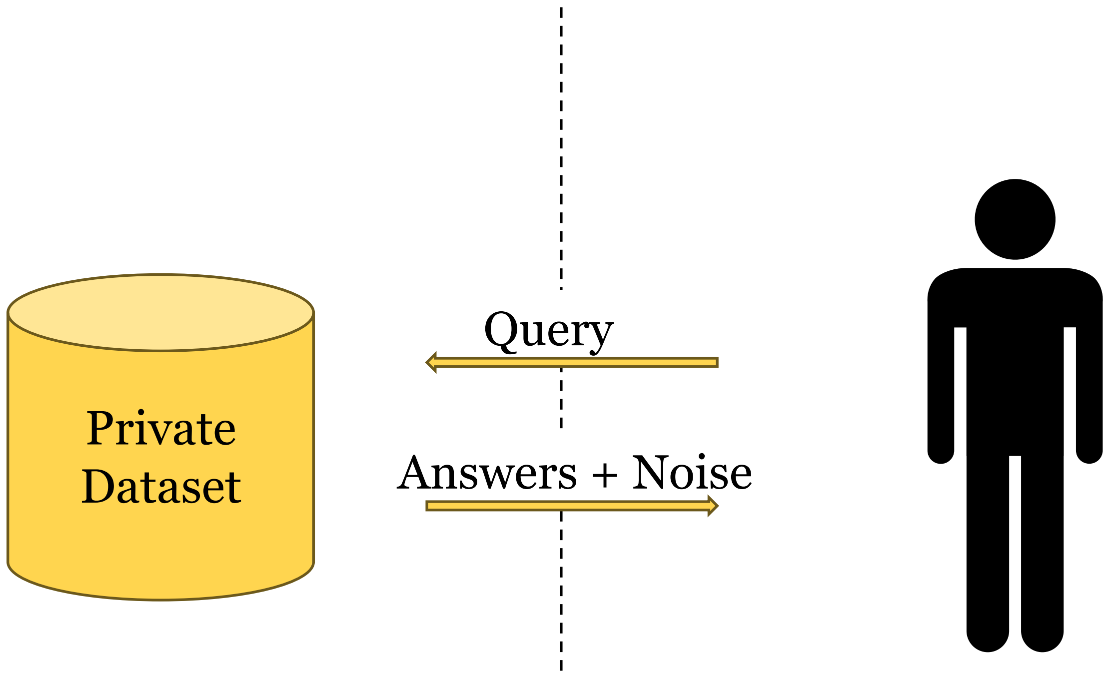
$$\Pr[A(\mathcal{D}) \in t] \leq e^{\varepsilon} \Pr[A(\mathcal{D}') \in t] + \delta$$

Quantifies information leakage

Allows for small probability of failure

(ε, δ) referred to as privacy budget

Problem Setup



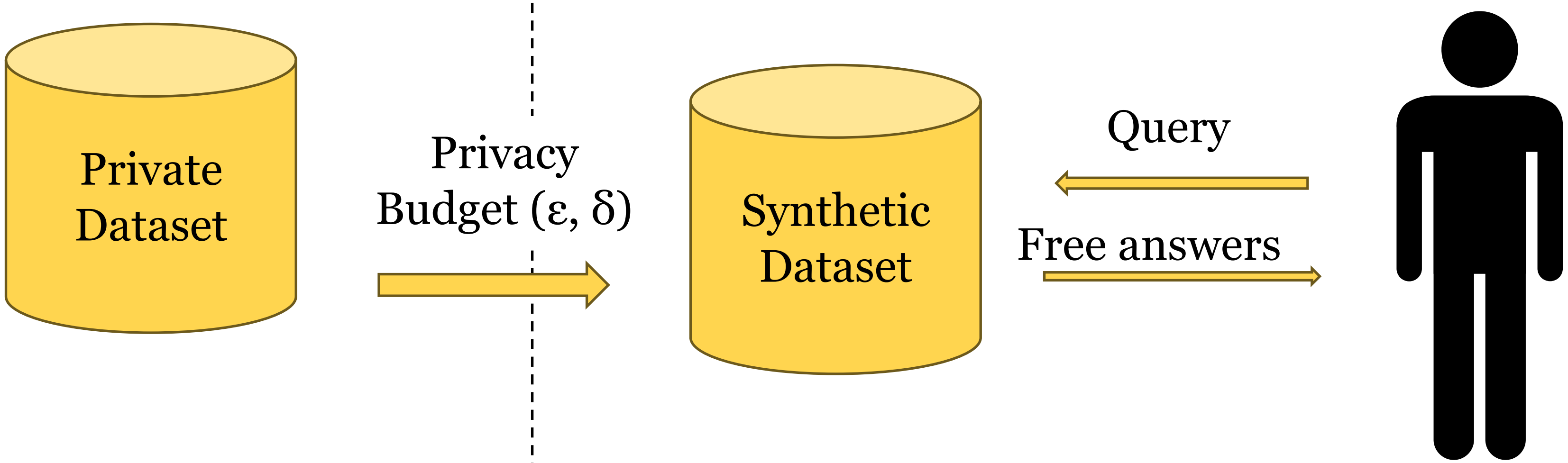
- Privacy budget (ε, δ) spent for each query
- Limited number of queries
- Only DP supported queries

Privacy
Firewall

Problem Setup

May mention how we are evaluating the synthetic dataset

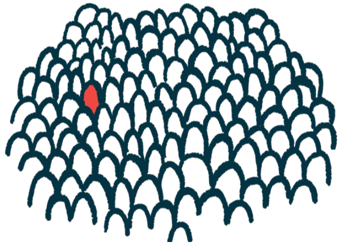
What if the private dataset has missing values?



- No limit on number of queries
- May ask unsupported queries

Human errors

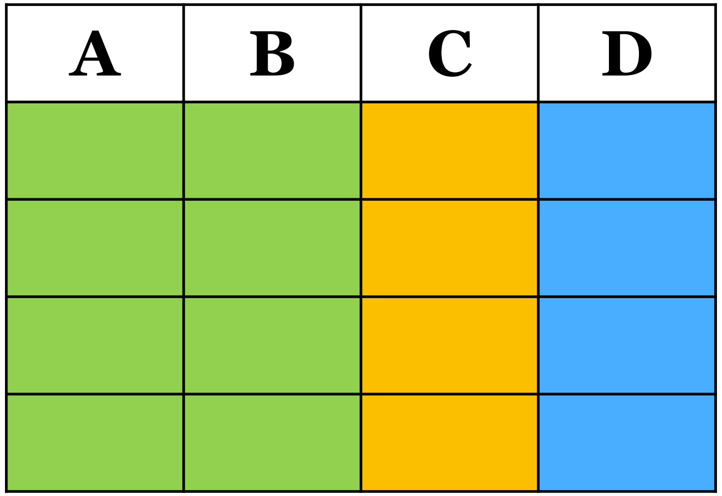
Privacy regulations (Right to forget)

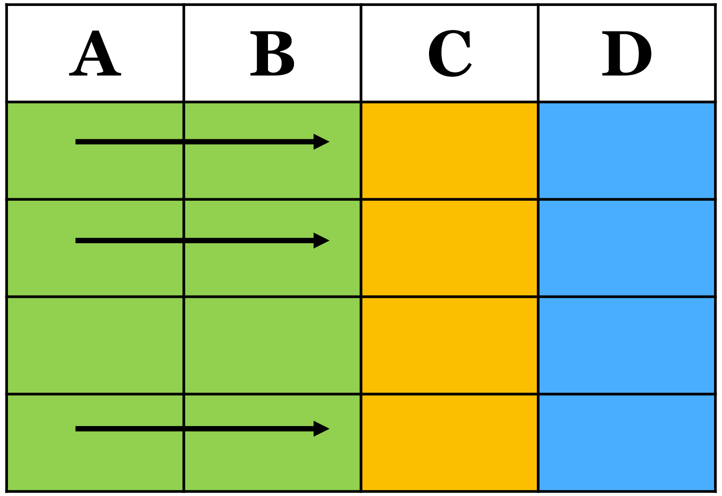


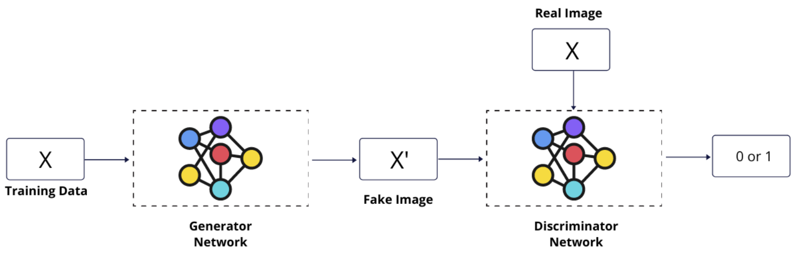
Rare conditions

Privacy Firewall

DP Synthetic data generation







Marginal based methods

Learn marginals over attributes
Add DP noise to the marginals
Sample synthetic data from noisy marginals

E.g. PrivBayes and AIM

Column based methods

Learn attributes in a sequence
Train intermediate models to predict next attribute
Use intermediate models to generate synthetic data

E.g. Kamino

GAN based methods

Generator generates fake data
Discriminator/Critic identifies fake vs real
Generator can be used to generate synthetic data

E.g DPautoGAN, DPCTGAN

What’s coming up next

- Effect of missing values on private synthetic data generation
- Types of missing data
- Preliminary solutions
- **Adaptive recourse for all three types of methods**
- Experimental setup
- Results
- **Privacy amplification due to missing data**
- Conclusion

Effect of missing values on private synthetic data generation

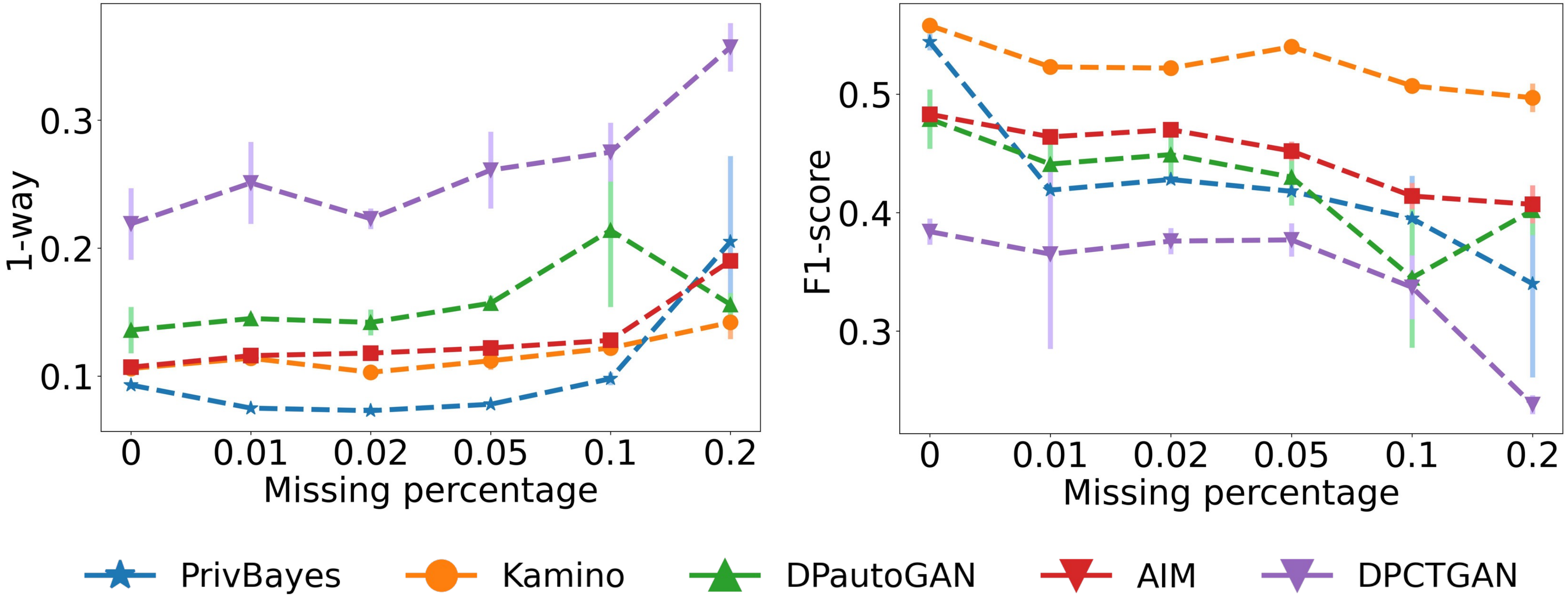


Fig: The effect of missing data on DP synthetic data generation algorithms on Adult

Decrease of 5 - 23% in utility for <5% and 10 - 190% for <20% missing values

Types of missing mechanisms

Occupation	Age	Income
	35	
Academic	40	60k
Business	30	200k

Missing completely at random (MCAR)

No correlation between missing values and observed values

Occupation	Age	Income
Academic	35	60k
Business	40	
Business	30	

Missing at random (MAR)

Missing values correlated to observed values

Occupation	Age	Smokes
Academic	35	No
Business	40	
Business	30	

Missing not at random (MNAR)

Missing values correlated to other missing values

Solution 1 : Complete row approach

- Number of complete rows remaining can be very small
  - 32k rows reduces to ≈ 5k complete rows with 20% MAR
  - ≈ 1k complete rows with 20% MCAR/MNAR
- Complete row also introduces bias in the distribution of attributes

MAR and MNAR are most affected by complete row approach

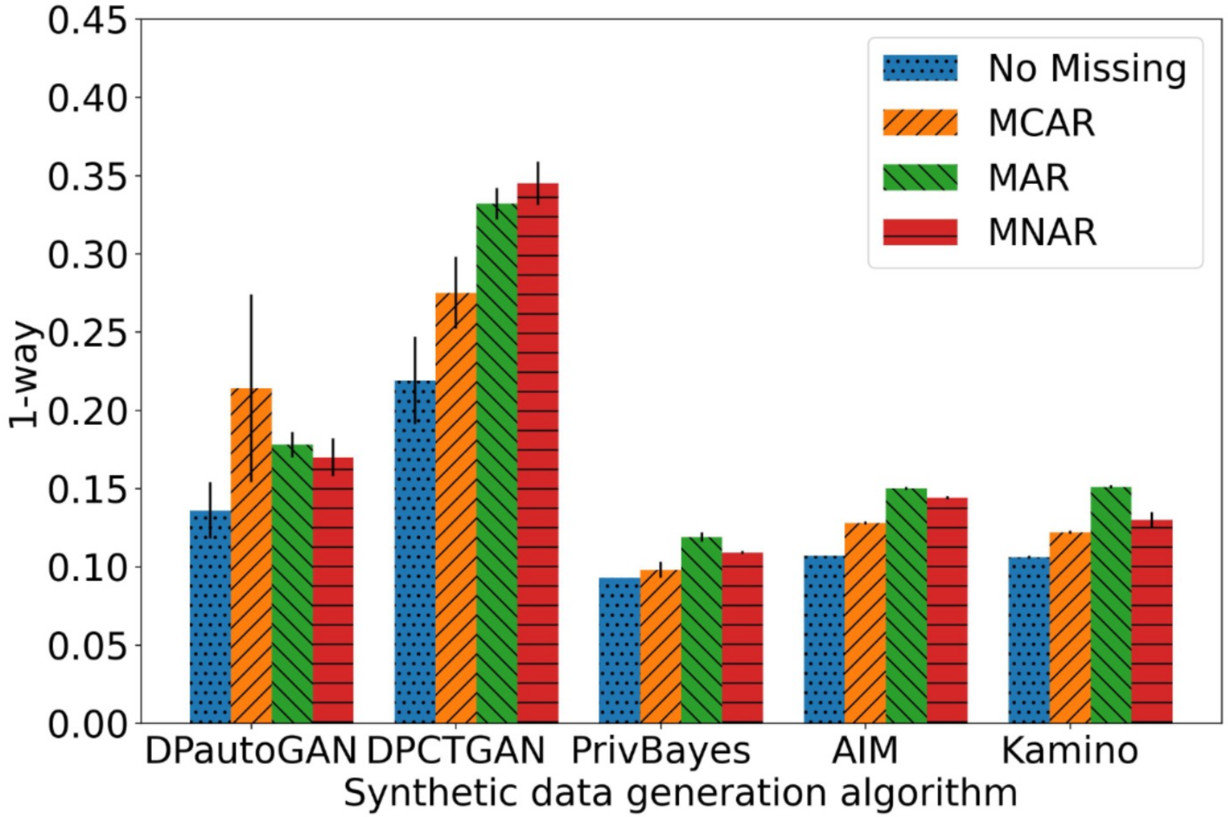


Fig: Complete row approach on various missing mechanisms

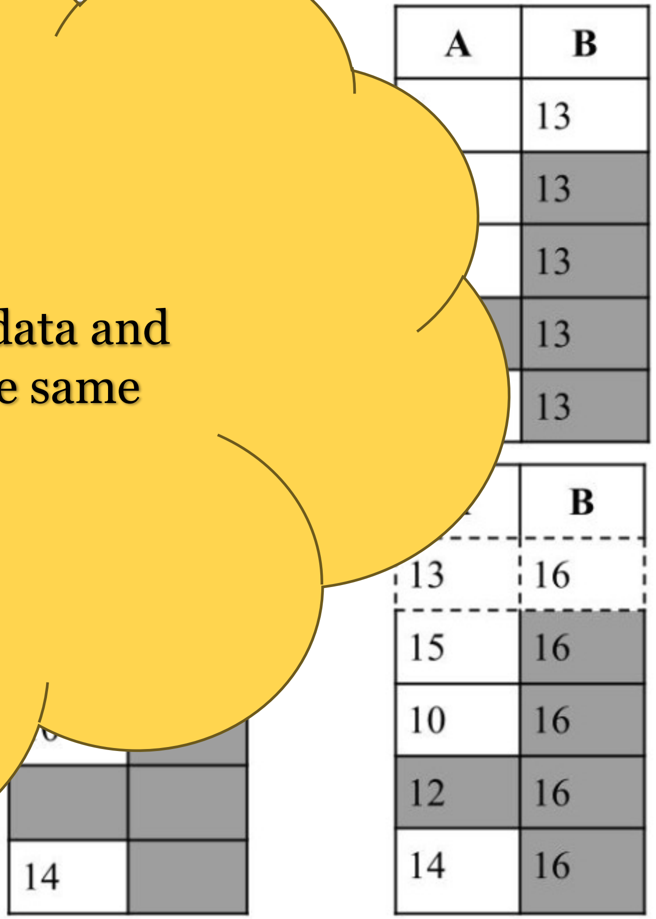
Solution 2 : Imputation

Private imputation can be do
ways:
1. Split privacy budget
2. Formulate imputati
   transformation fu

Splitting results in less
synthetic data generatio

DP imputation requires adding n
proportional to number of rows

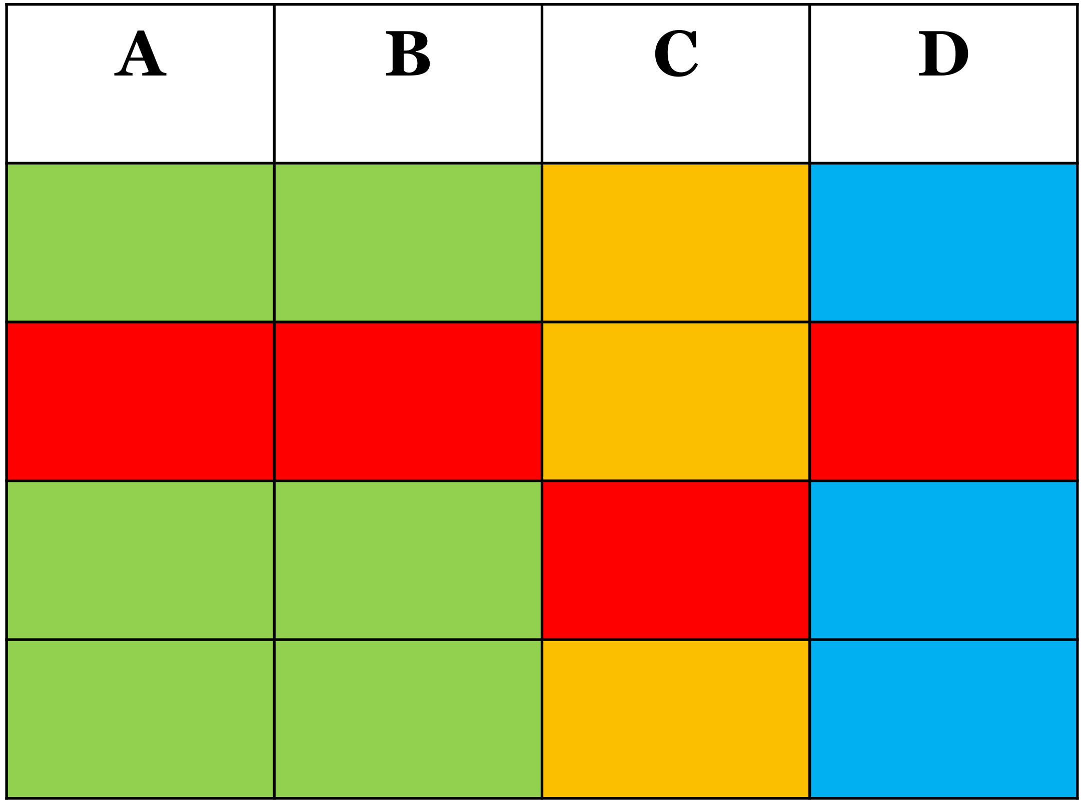
What if we deal with missing data and generate synthetic data at the same time?



UNIVERSITY OF WATERLOO | DSG Data Systems Group

Solution 3 : Adaptive Recourse

Partial marginal observation-based

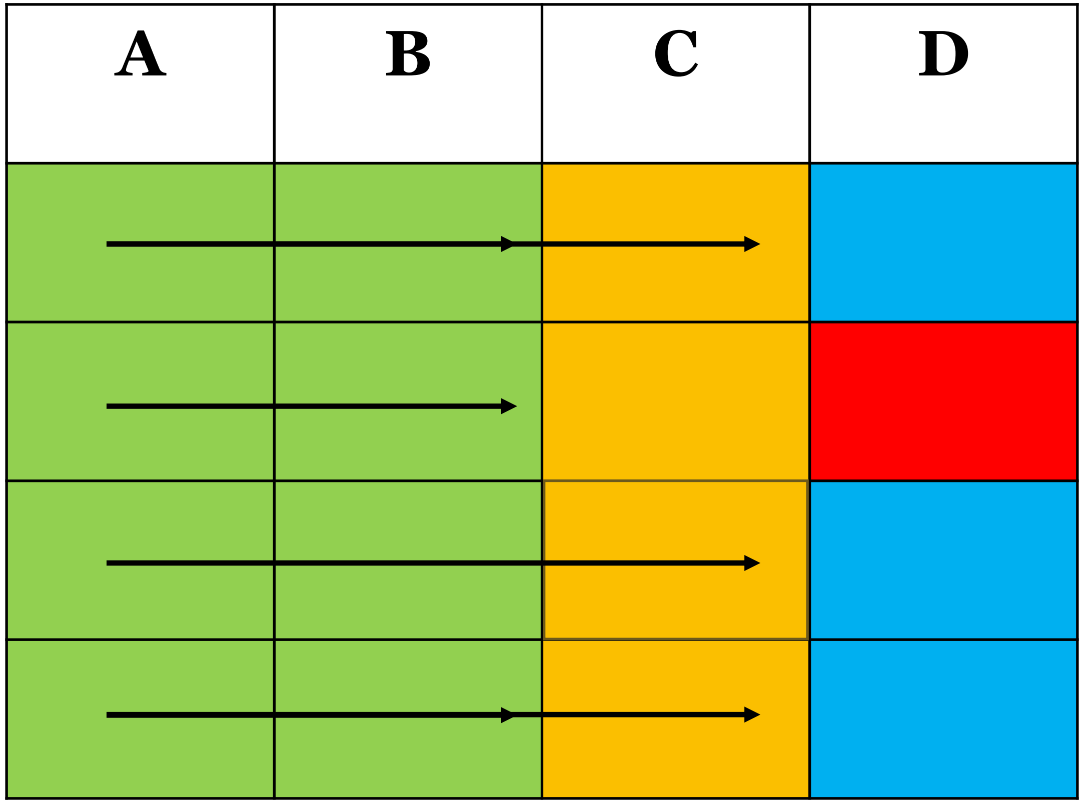


- Works for marginal based approaches
- The marginals are computed over subset of attributes : AB, C, D
- Learn from complete values in queried marginals
- Save incomplete rows over queried subset

AB and D can save information from this row

Solution 3 : Adaptive Recourse

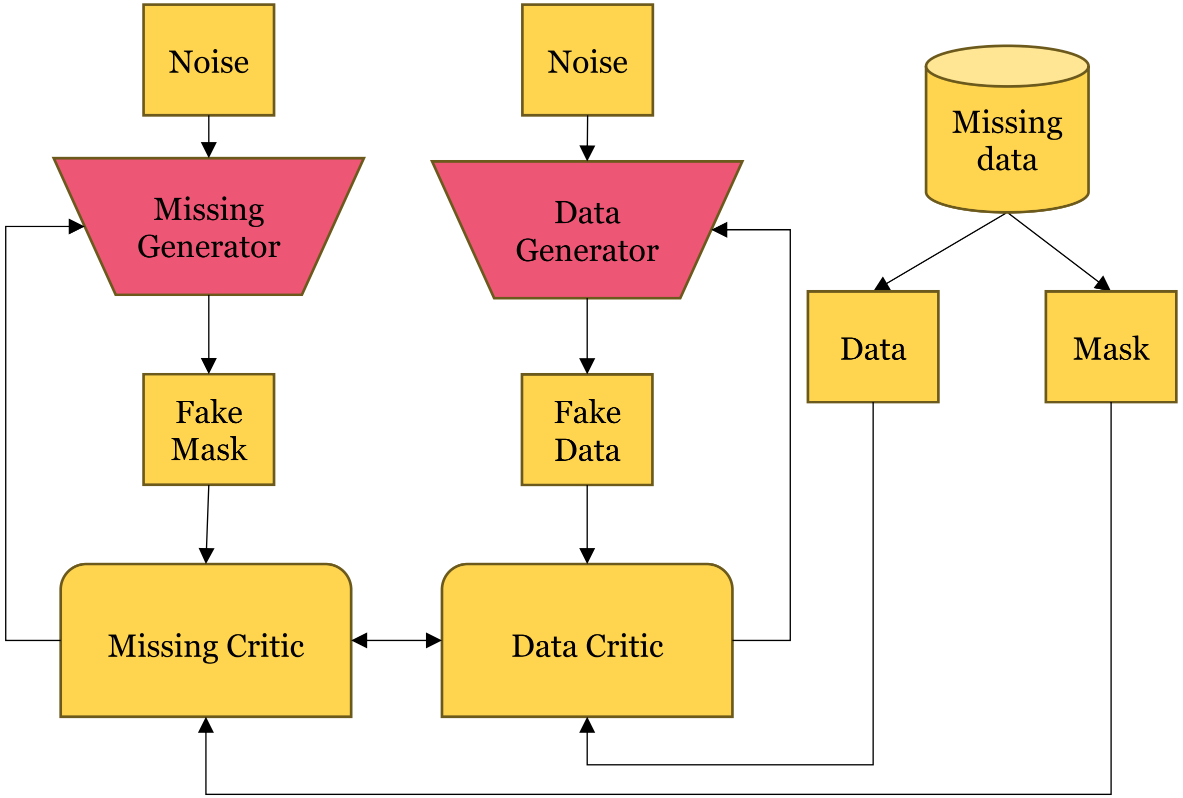
Column-wise data generation-based



- Attributes are learnt in a sequence
- Synthetic data is generated by leveraging intermediate model
- Impute missing values using learnt models

Solution 3 : Adaptive Recourse

GAN-based



- Missing data is split into data and missing mask
- Two GAN models on missing mask and complete are learnt

After convergence, the data generator is used to sample synthetic data

Experimental Setup

Methods
Marginal based approaches: PrivBayes and AIM
Column wise data generation: Kamino
GAN based: DPCTGAN, DPAutoGAN

Datasets
Adult, BR2000, Bank and National

Dataset	Cardinality	#Numerical Attr	#Categorical Attr
Adult	32561	5	10
Bank	45211	3	14
BR2000	38000	3	11
National	15012	6	14

Metrics
1-way and 2-way metric (smaller values are better) [Mention how exactly its being calculated]
F1-score of 9 ML models (higher values are better)

GAN based methods are trained with ϵ=3 and others with ϵ=1
All results are averaged over 3 runs.

Results

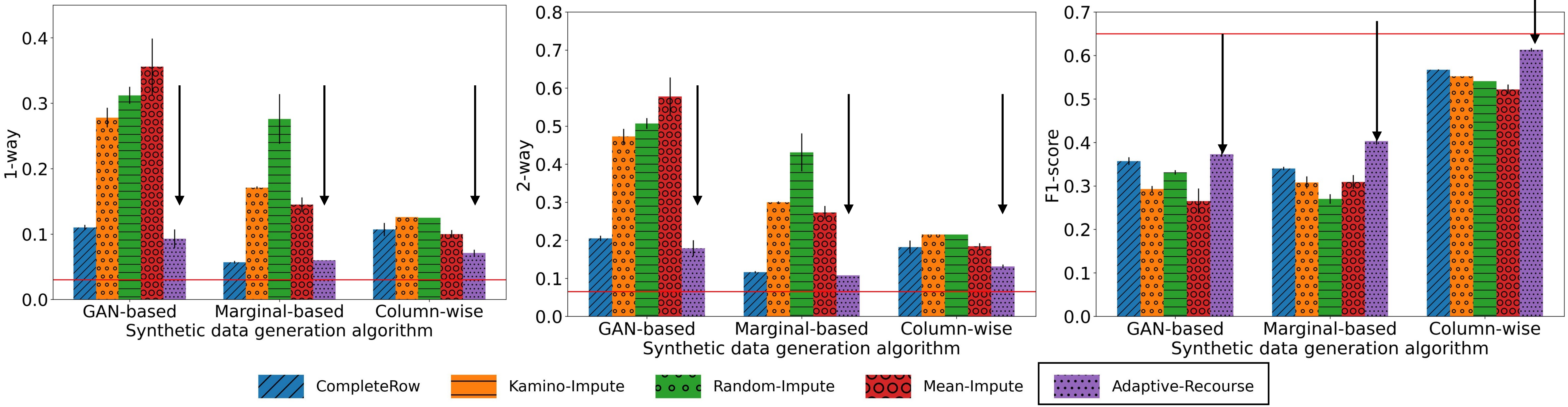


Fig: Comparison of all solutions on Adult dataset with 10% MCAR data

- Adaptive recourse methods perform better than the complete row approach or imputation
- Complete row approach performs second best

Results

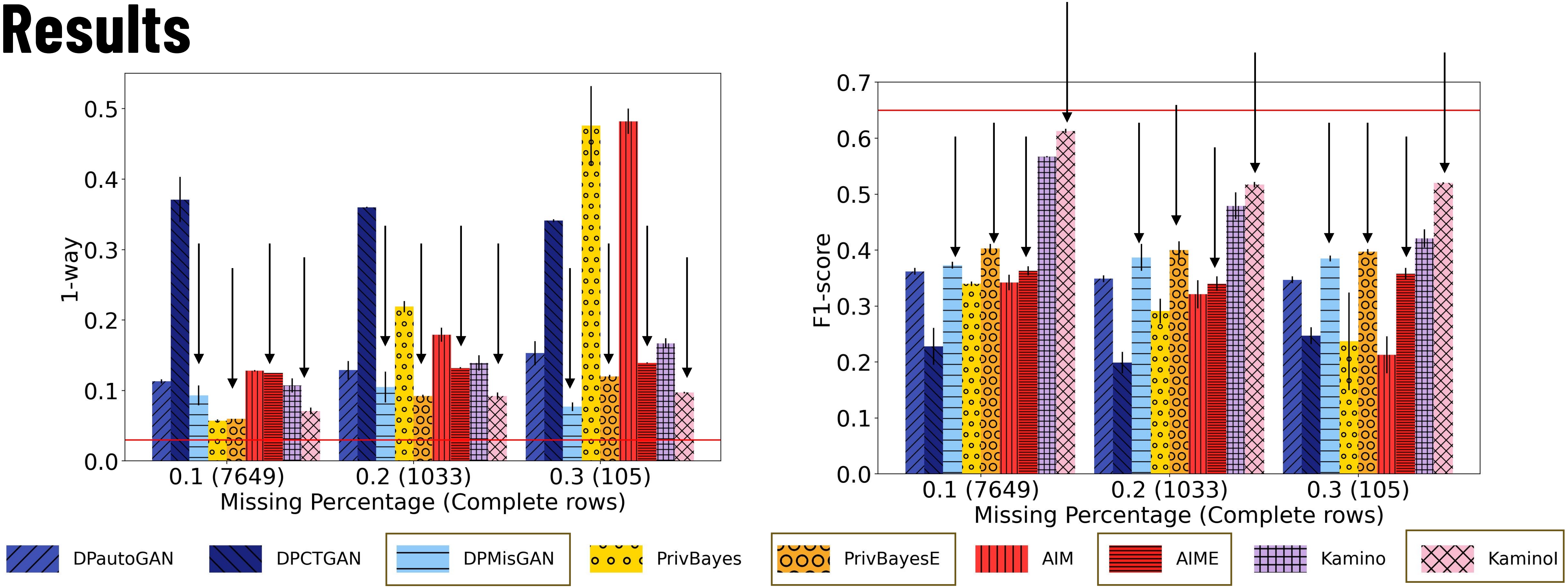


Fig: Comparison of adaptive strategies vs complete row approach for all baselines on Bank

Adaptive methods outperform other baselines at various amounts of missing data

Results

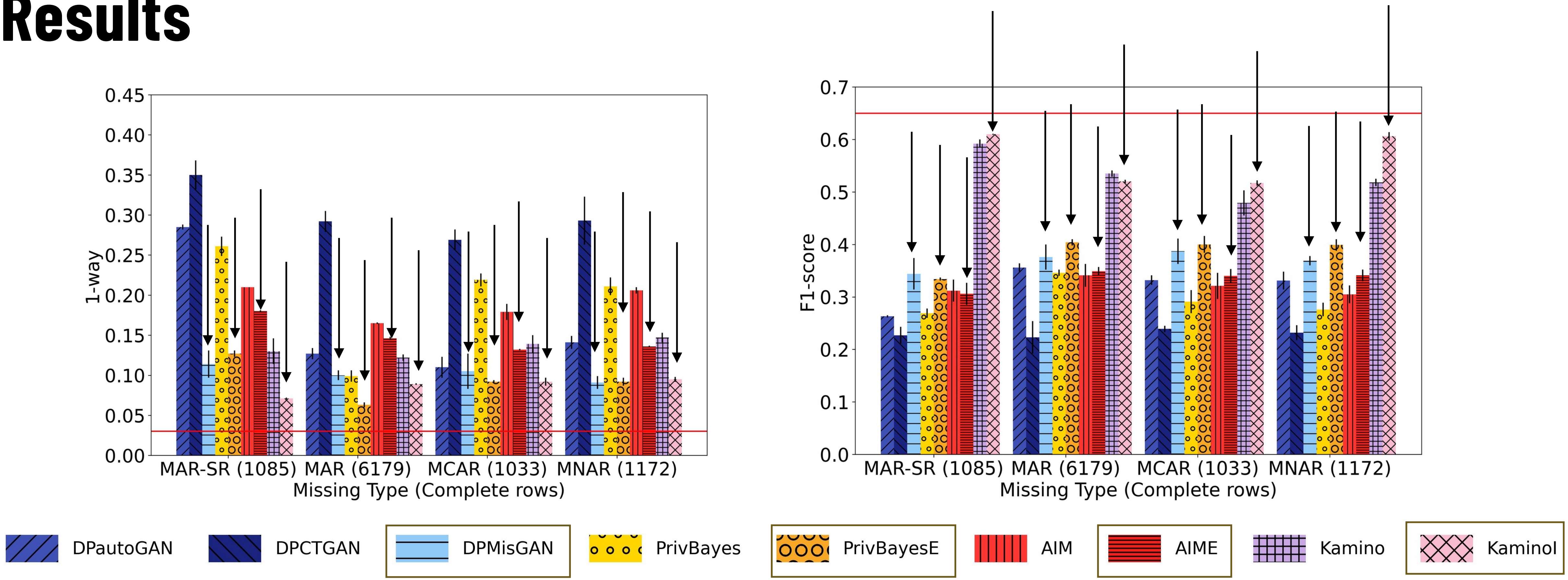
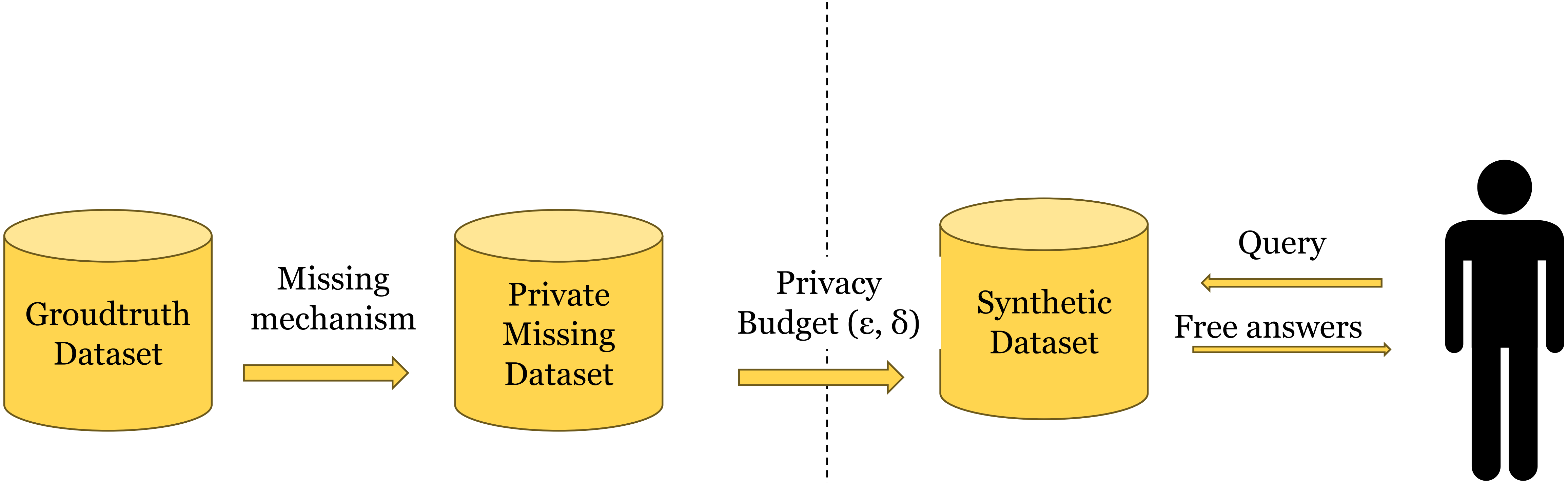


Fig: Comparison of adaptive strategies vs complete row approach for all baselines with different missing mechanisms on Bank 20%

Adaptive methods outperform other baselines at various missing mechanisms

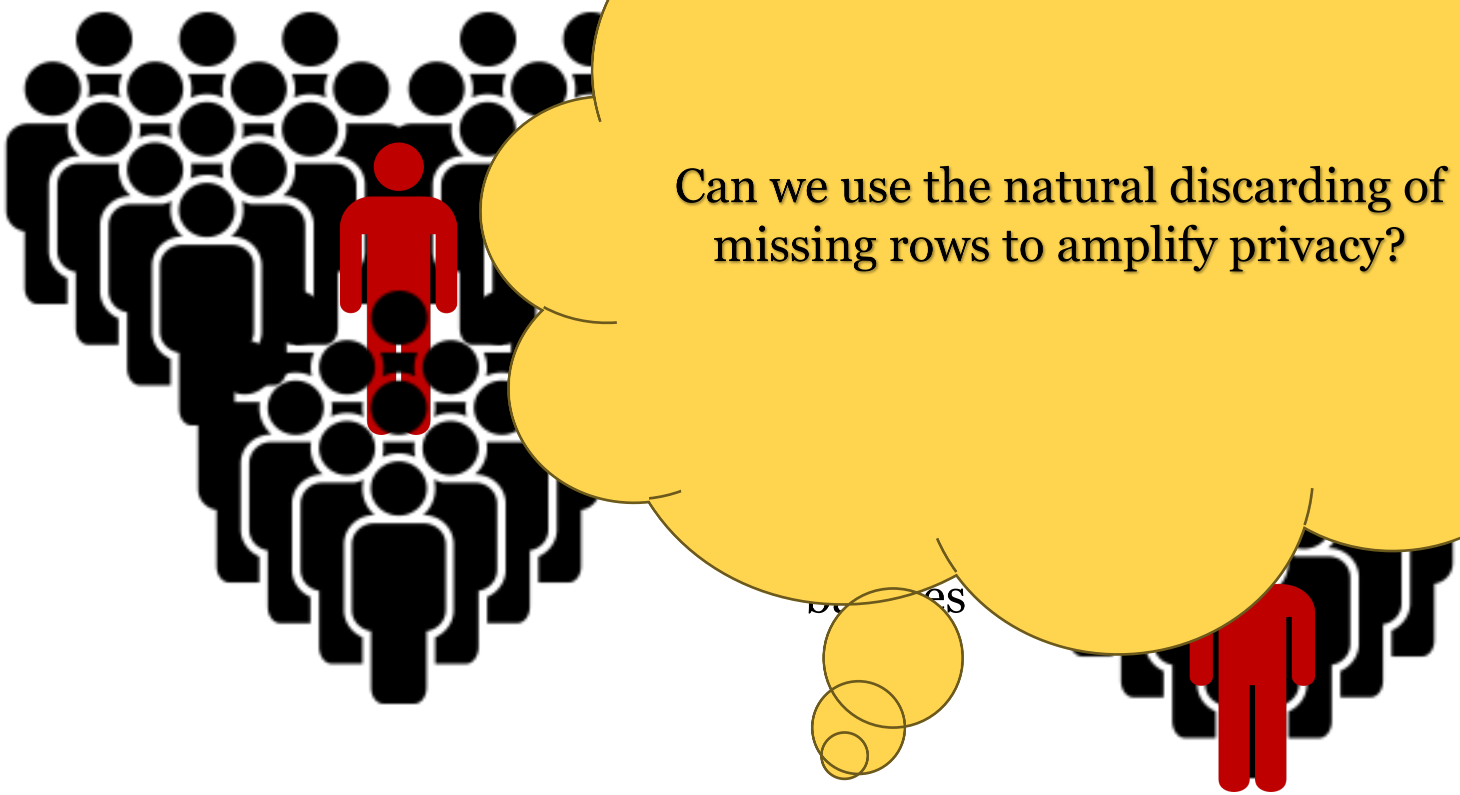
PRIVACY AMPLIFICATION DUE TO MISSING DATA

Recap Problem Setup



How does the privacy guarantee change for the groundtruth data?

Amplification due to subsampling



Can we use the natural discarding of missing rows to amplify privacy?

nsider an algorithm A: $\mathcal{D} \rightarrow \mathcal{R}$ sfies (ε, δ)-differential privacy if a pling mechanism S (D) that ples a random subset U from set D of n samples with p bability for each sample. Then, the omposite mechanism A(S (D)) offers (pϵ, pδ)-DP for small values of ϵ.

Is amplification always possible?

$\bar{D}$

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist	M	80k
BC	Artist	F	25k
AB	Business	F	100k

→ D

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist		80k
	Artist	F	25k
AB	Business	F	

$\bar{D}'$

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist	M	80k
BC	Artist	F	25k
AB	Business	F	80k

→ D'

State	Occupation	Gender	Income
ON	Business	M	
BC	Artist		
	Artist	F	25k
AB	Business	F	

State and Gender are MCAR and Income is MNAR

Amplification is possible only when each row has independent probability of having missing values.

Amplification for MCAR

$\bar{D}$

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist	M	80k
BC	Artist	F	25k
AB	Business	F	100k

→ D

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist		80k
	Artist	F	25k
AB	Business	F	

Assuming MCAR for all rows.
$\Phi_{state} = 0.25$, $\Phi_{occupation} = 0$, $\Phi_{gender} = 0.25$, $\Phi_{income} = 0.25$

For complete row approach,

$$\Pi_i(1-\phi_i) = 0.421\epsilon \ \ (saves \downarrow 0.579\epsilon\ )$$

Attributes $\{A1, \ldots Ak\}$
Missing probabilities $\{\phi1, \ldots, \phi k\}$

The probability of a row not having missing value = $\Pi_i(1-\phi_i)$

Amplification for partial marginal based approach

$\bar{D}$

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist	M	80k
BC	Artist	F	25k
AB	Business	F	100k

→ D

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist		80k
	Artist	F	25k
AB	Business	F	

Assuming MCAR for all rows.
$\Phi_{state} = 0.25$, $\Phi_{occupation} = 0$, $\Phi_{gender} = 0.25$, $\Phi_{income} = 0.25$

Marginals:
M1 <State> : $1 - \Phi_{state} = 0.75$
M2 <Occupation> : $1 - \Phi_{occupation} = 1$
M4 <Gender, Income>: $(1 - \Phi_{gender}) * (1 - \Phi_{income}) = 0.5625$

Assuming $\epsilon/3$ budget for each, total is $0.77\,\epsilon$

Amplification for partial marginal based approach

$\bar{D}$

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist	M	80k
BC	Artist	F	25k
AB	Business	F	100k

→ D

State	Occupation	Gender	Income
ON	Business	M	80k
BC	Artist		80k
	Artist	F	25k
AB	Business	F	

Assuming MCAR for all rows.
$\Phi_{state} = 0.25$, $\Phi_{occupation} = 0$, $\Phi_{gender} = 0.25$, $\Phi_{income} = 0.25$

Marginals:
M1 <State> : 1 - $\Phi_{state} = 0.75$
M2 <Occupation> : 1- $\Phi_{occupation} = 1$
M3 <Gender> : 1- $\Phi_{gender} = 0.75$
M4 <Gender, Income>: $(1 - \Phi_{gender}) * (1 - \Phi_{income}) = 0.5625$

Assuming equal budget $\epsilon/4$ budget for each
If we choose the marginal with most amplification M4 over M3,
M1, M2, M4 = 0.83ϵ ($\downarrow$ *saves* $0.17\ \epsilon$)

A better bound is possible if both M3 and M4 are amplified using intersecting attribute

M3 and M4 both amplified using Gender
Therefore,

$$0.75 * \frac{\epsilon}{4} + 1 * \frac{\epsilon}{4} + 0.75\left(2 * \frac{\epsilon}{4}\right) = 0.81\epsilon \ (\downarrow saves\ 0.19\ \epsilon)$$

Results

The problem of finding the best partition is exponential hard

We show optimizations to prune some partitions

Algorithm:

1. Find all possible disjoint partitions of the attribute set
2. Iterate all partitions
    1. Initialize cost for partition
    2. Choose best amplification factor in partition for every marginal
    3. Calculate total cost
3. Return partition with lowest cost

Dataset	MCAR missing %				
	0.1	0.2	0.3	0.4	0.5
Adult	0.88	0.77	0.65	0.47	0.44
BR2000	0.83	0.68	0.55	0.41	0.31

Table: Amplified privacy costs

Conclusion

- We show that missing data can drastically affect the performance of DP synthetic data generation methods
- Complete row approach and imputation are not effective solutions. They either discard too many rows or are costly in terms of privacy.
- Simple adaptive methods significantly improve the quality of DP synthetic data generations methods without spending extra privacy budget
- Missing data can be used to amplify privacy using subsampling techniques

Thank you! Questions? :)

Add your name to lunch list --->

UNIVERSITY OF WATERLOO | DSG Data Systems Group