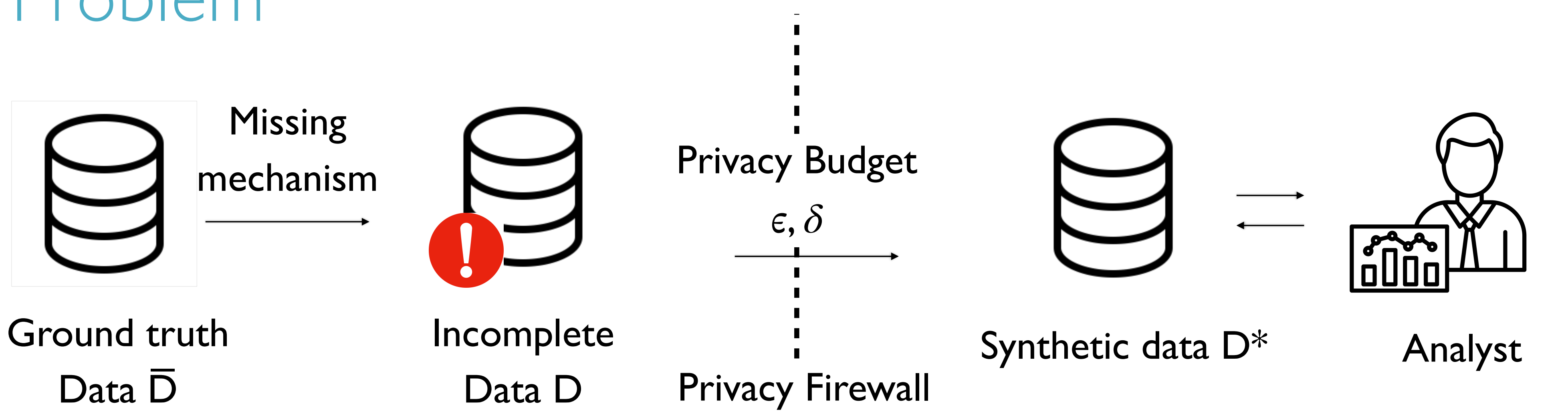
# Differentially Private Data Generation with Missing Data

Shubhankar Mohapatra, Florian Kerschbaum, Xi He
University of Waterloo

## Problem



Missing mechanism

Ground truth Data $\bar{D}$

Incomplete Data D

Privacy Budget $\epsilon, \delta$

Privacy Firewall

Synthetic data D*

Analyst

1. How does missing data affect private synthetic data generation?
2. How can we adapt existing methods to work on missing data?
3. How can we account privacy for the ground truth data?

Differential Privacy [1]: A randomized algorithm A: $\mathcal{D} \rightarrow \mathcal{R}$ satisfies $\epsilon, \delta$ -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 :

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

The definition changes according to the input incomplete data D or the ground truth data $\bar{D}$.

Missing mechanism[2]: Missing data is classified into different types using missing mechanisms:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

## Effect of missing data and naive solutions

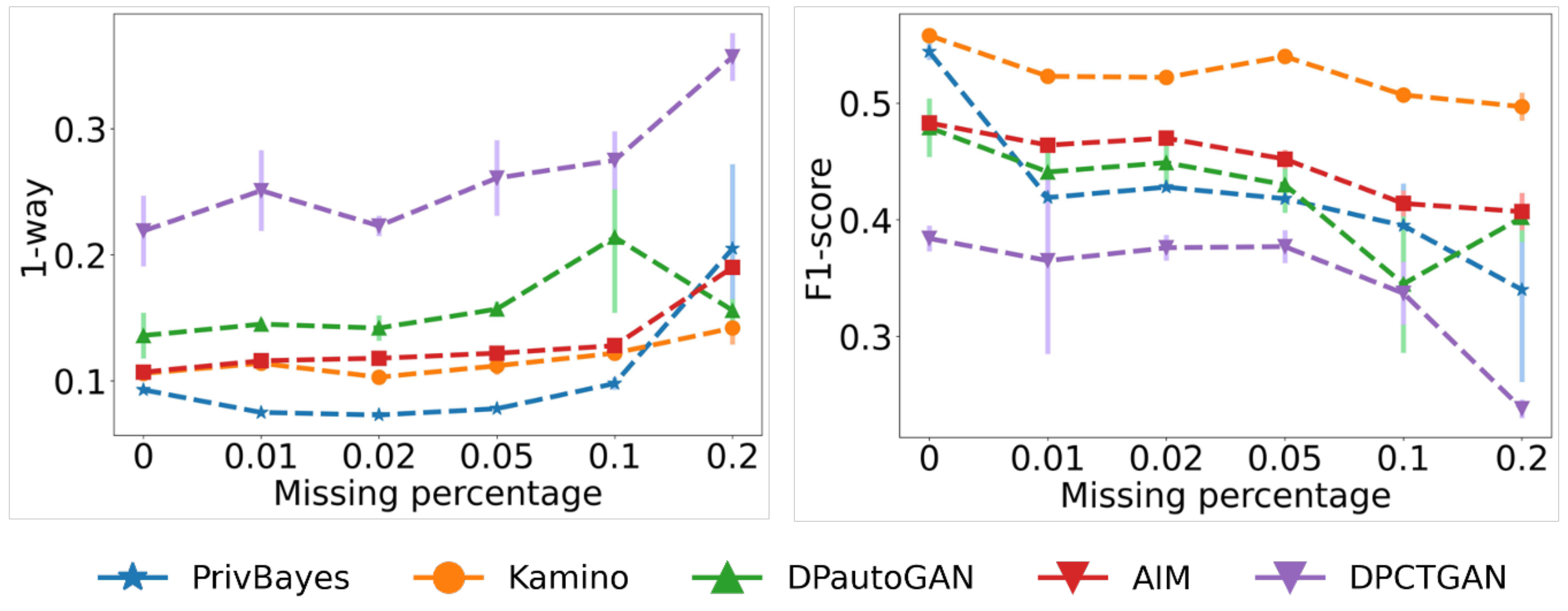


Figure: The performance of DP data generation algorithms decreases with increasing missing data

Naive approaches to tackle missing data include:

1. Complete row approach
2. Data imputation

Complete row results in bias of attributes

Data imputation is costly with DP

1. Requires splitting budget
2. Has sensitivity in $\mathcal{O}(n)$

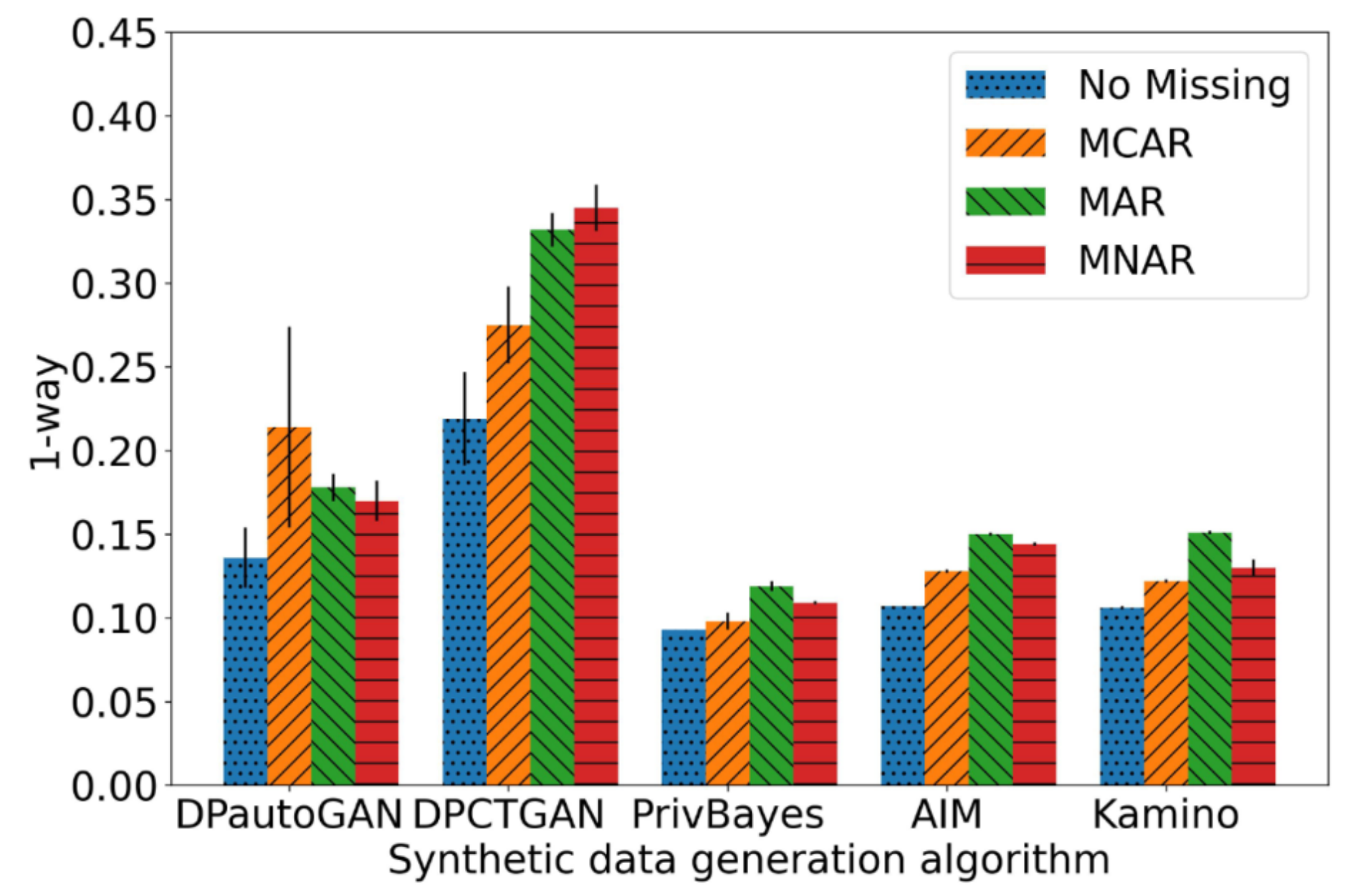
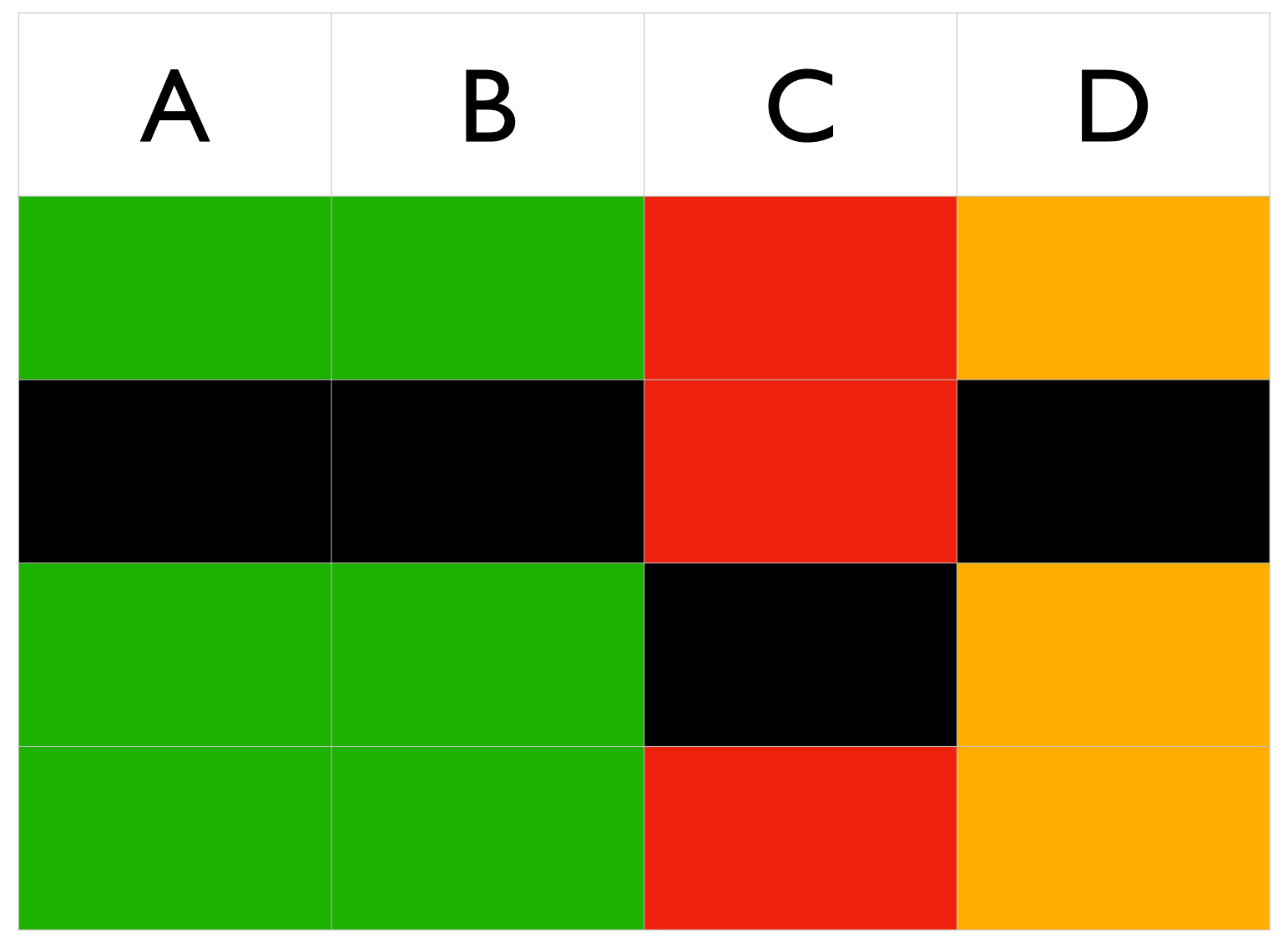


Figure: Complete row approach fails drastically for MAR and MNAR missing values

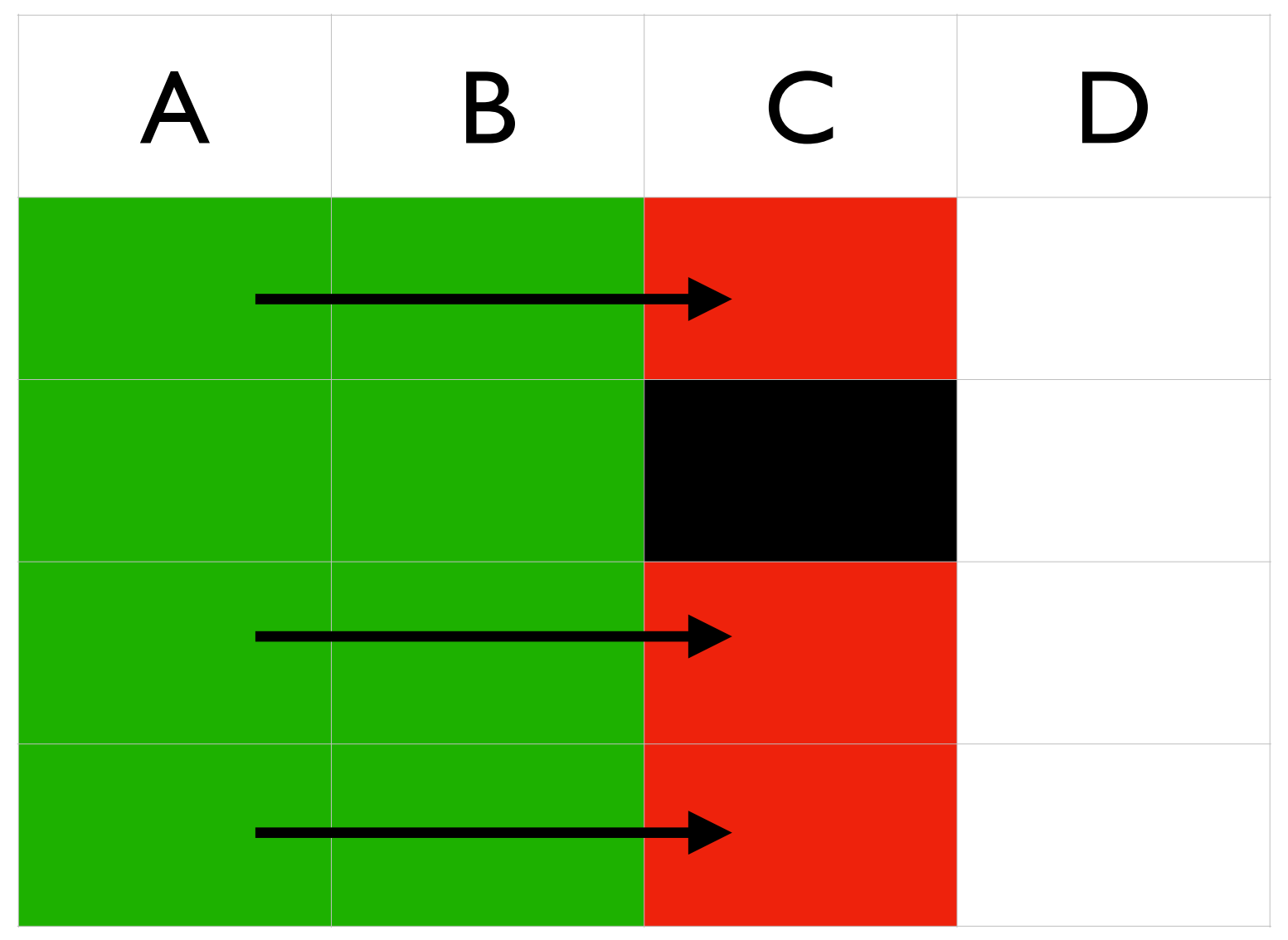
## Adaptive recourse



Partial marginal observation

Works for marginal based approaches

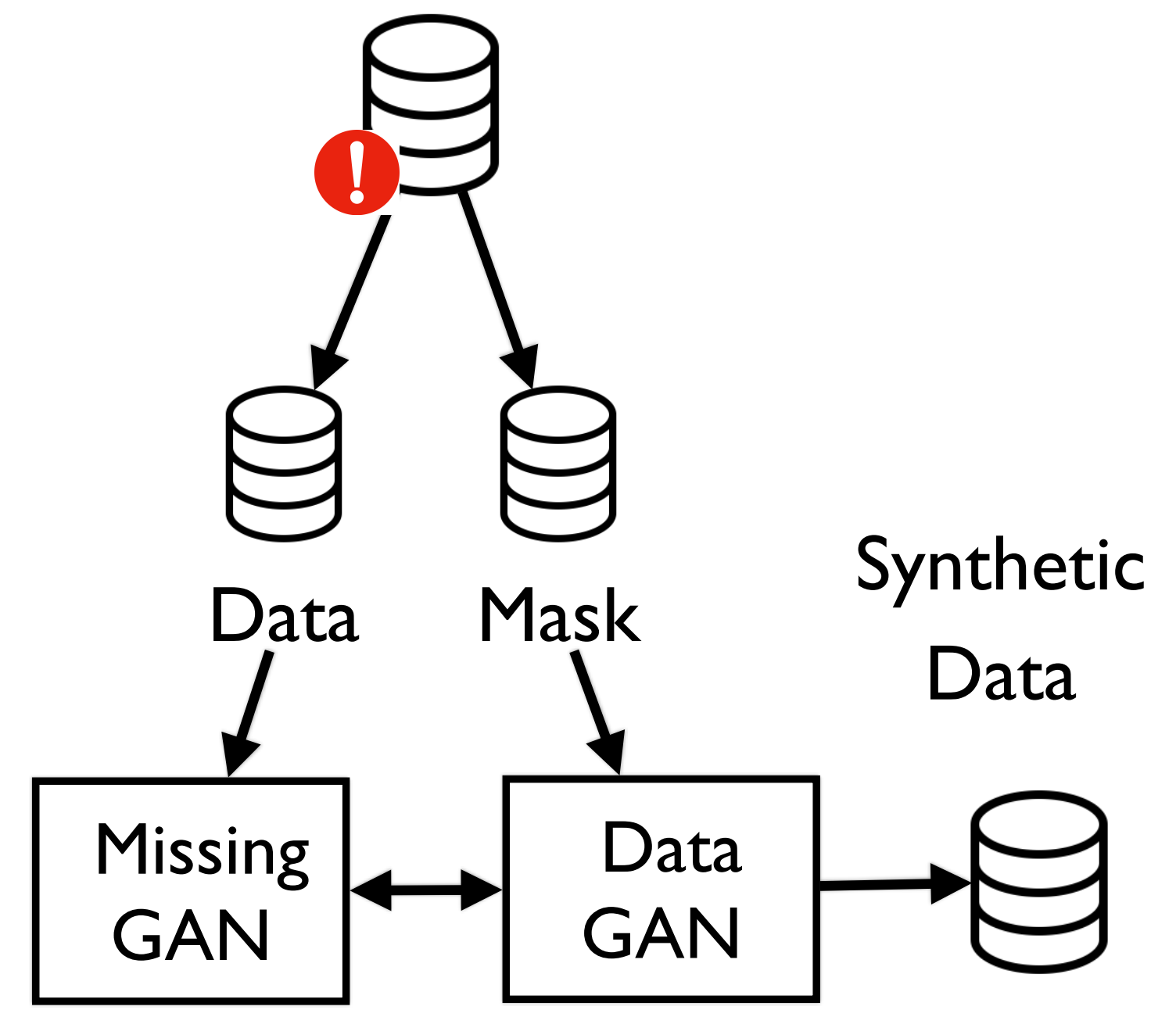
Learns from complete values in queried marginals



Column wise data generation

Attributes are learnt in sequence

Imputes values using intermediate learnt models



Data

Mask

Synthetic Data

Missing GAN

Data GAN

GAN based

Missing data is split into data and missing mask

Two GAN models are learnt

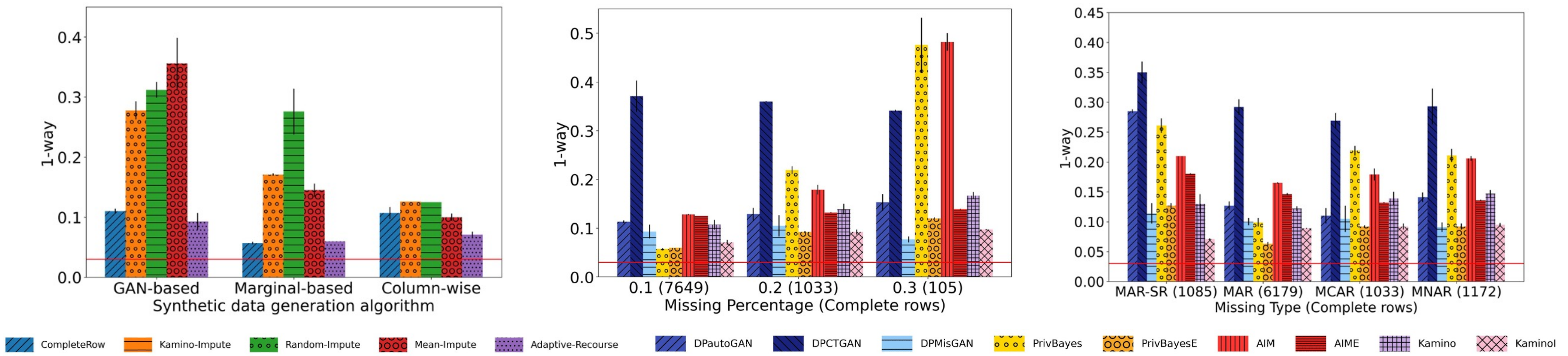


Figure: Adaptive recourse methods perform better than naive solutions. Left image shows adaptive recourse (purple) is better than naive approaches. Right image shows adaptive methods (DPMisGAN, PrivBayesE, AIME and KaminoI) are better than their counterparts at different amounts and types of missing data.

## Privacy accounting and amplification for ground truth data

Natural discarding of missing values can be leveraged to account for privacy of ground truth data $\bar{D}$ when missing data is MCAR.

Privacy amplification due to subsampling [3]: DP Mechanism on random subsample of data improves privacy. If each row is sampled with $p$ probability, amplified mechanism is $(p\epsilon, p\delta)$- DP.

- Every marginal can be amplified using missing probability of its participating attributes
- Generate all possible valid partitions of attributes
- Prune suboptimal partitions
- Return maximum amplification of remaining partitions

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

| Marginal | Suboptimal | Optimal |
|---|---|---|
| $M_1$ <State> | 0.75 | 0.75 |
| $M_2$ <Occupation> | 1 | 1 |
| $M_3$ <Gender> | ✗ | 0.75 |
| $M_4$ <Gender, Income> | 0.5625 | 0.75 |
| Amplification factor | 0.83 | **0.81** |

Figure: Optimal results are often achieved by amplifying multiple marginals ($M_3$ and $M_4$) using common attribute (Gender).

References:

1.Dwork, Cynthia, et al. "Calibrating noise to sensitivity in private data analysis." *TCC 2006*

2.Rubin, Donald "Inference and Missing Data". *Biometrika 1976*

3.Balle, Borja, et al. "Privacy amplification by subsampling: Tight analyses via couplings and divergences." *NeurIPS 2018*