Private Data Synthesis: State of the Art and Challenges

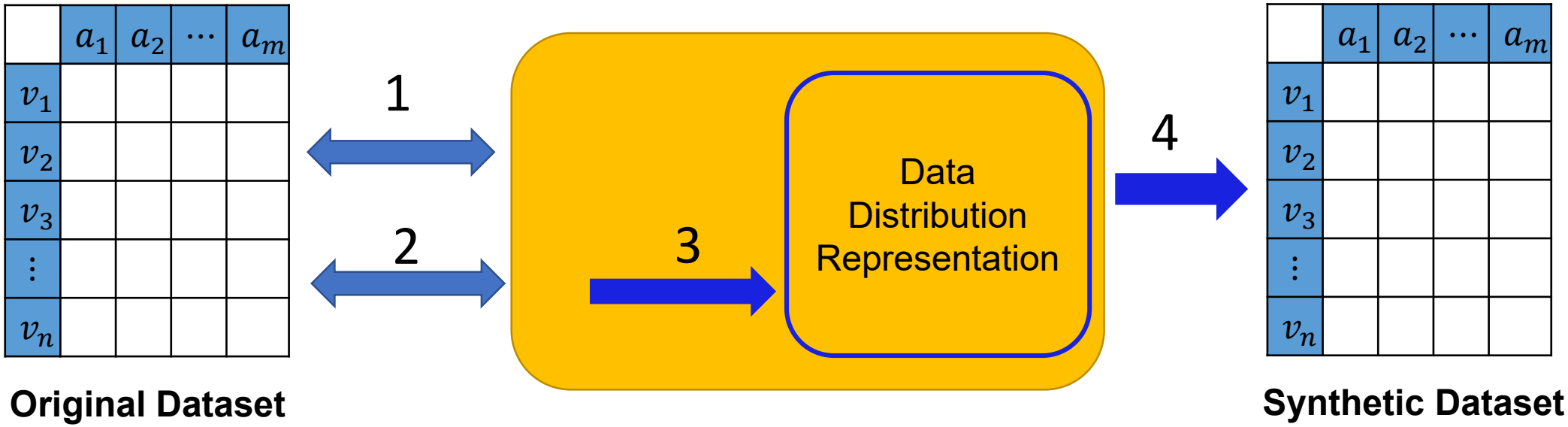
Presenter: Ninghui Li

Purdue University

11/14/2024 @ Synthetic Data and GenAI Workshop

Approaches to Private Data Synthesis

Private Synthesis Framework

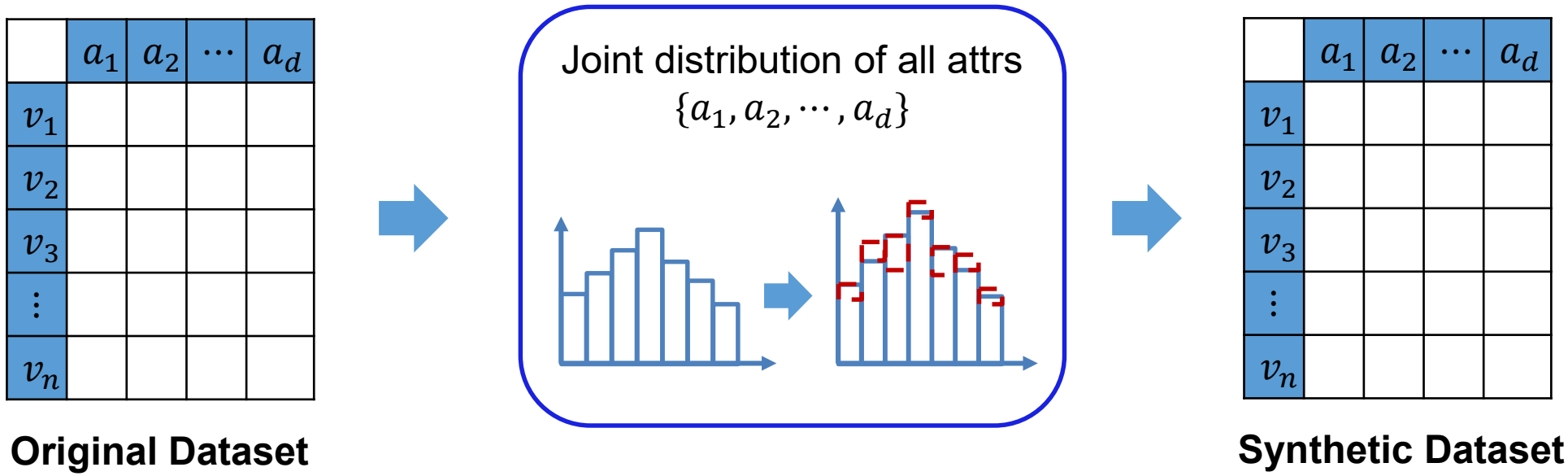


1. Choose appropriate queries (possibly depending on the dataset)
2. Obtain query answers (DP-based approaches satisfy DP)
3. Construct data representation from query answers
4. Synthesize dataset from the representation

Types of Data Distribution Representation

- Full domain probability density
- Probabilistic Graphical Model
- A set of consistent low-dimensional marginals
- Deep Neural Network Generative Model
  - Generative Adversarial Network
  - Diffusion Models
  - Fine-tuned Large Language Models

Using Full Domain Distribution



Original Dataset

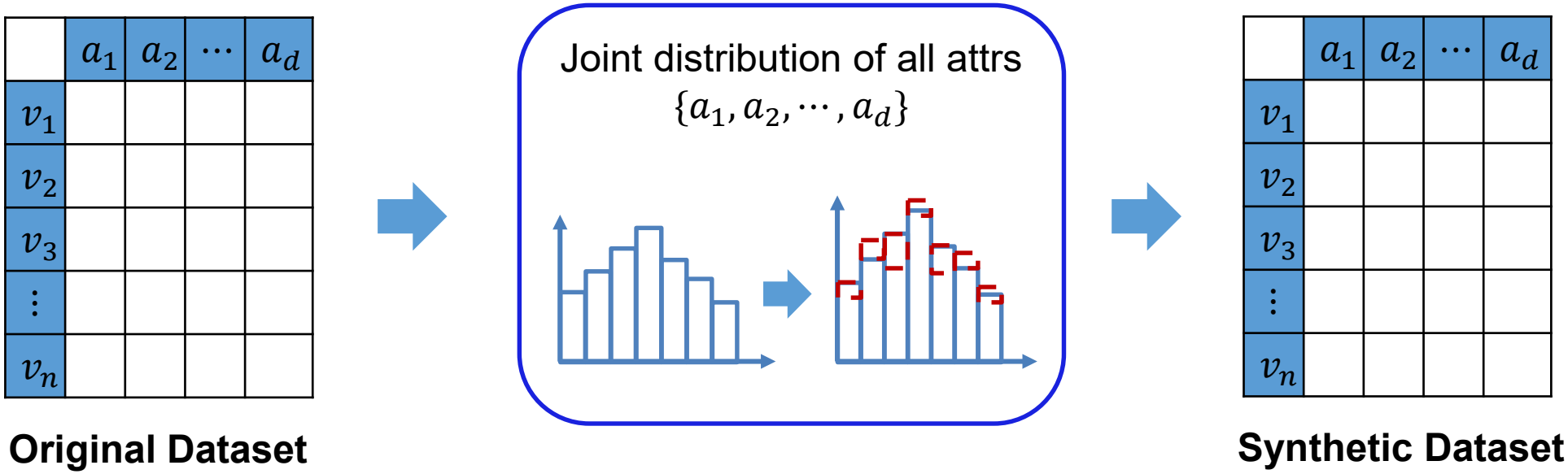
Synthetic Dataset

MWEM: An Example Using Full Domain Distribution (DP)

- Multiplicative Weights update rule with the Exponential Mechanism
  - Assumes as input a set of queries one wants to answer accurately
  - Maintains a probability distribution over the whole domain, initialized to be all uniform
  - Repeat a number of times
    - Uses Exponential Mechanism to choose a query that has the largest error given current distribution
    - Obtain a noisy answer to the chosen query
    - Update the distribution using the query answers answer (using the multiplicative update rule)

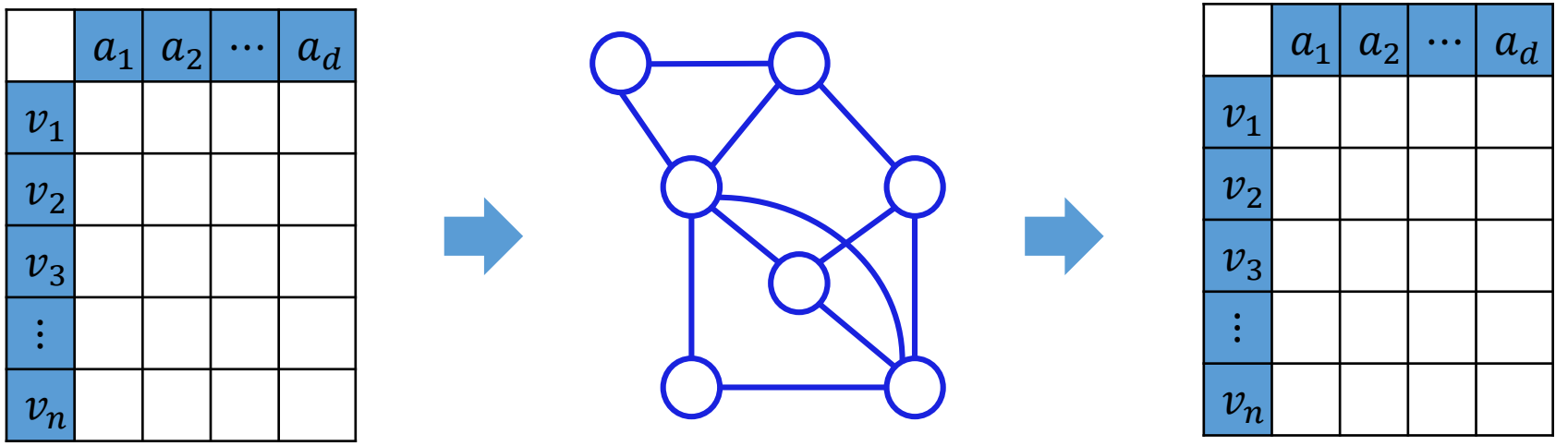
Hardt, Ligett, McSherry: A Simple and Practical Algorithm for Differentially Private Data Release. NIPS 2012.

Using Full Domain Distribution: Limitations



❖ **When the number of attributes is large, the domain of joint distribution is large, leading to prohibitive computational cost.**

Graphical Model



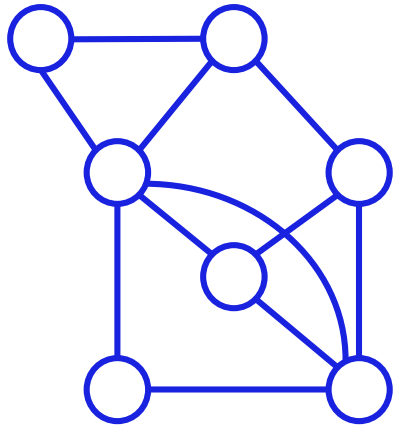
- Uses a Probabilistic Graphical Model (e.g., Bayesian Network or Markov Random Fields) as representation.
- Learn the PGM structure and parameters (through marginal tables)
- Sample the model to generate synthetic data

Zhang et al.: PrivBayes: Private data release via Bayesian networks. SIGMOD 2014.
McKenna, Sheldon, and Miklau: Graphical model based estimation and inference for differential privacy. ICML 2019.

Graphical Model: Some Limitations

	a_1	a_2	$\cdots$	a_d
v_1				
v_2				
v_3				
$\vdots$				
v_n				

Original Dataset

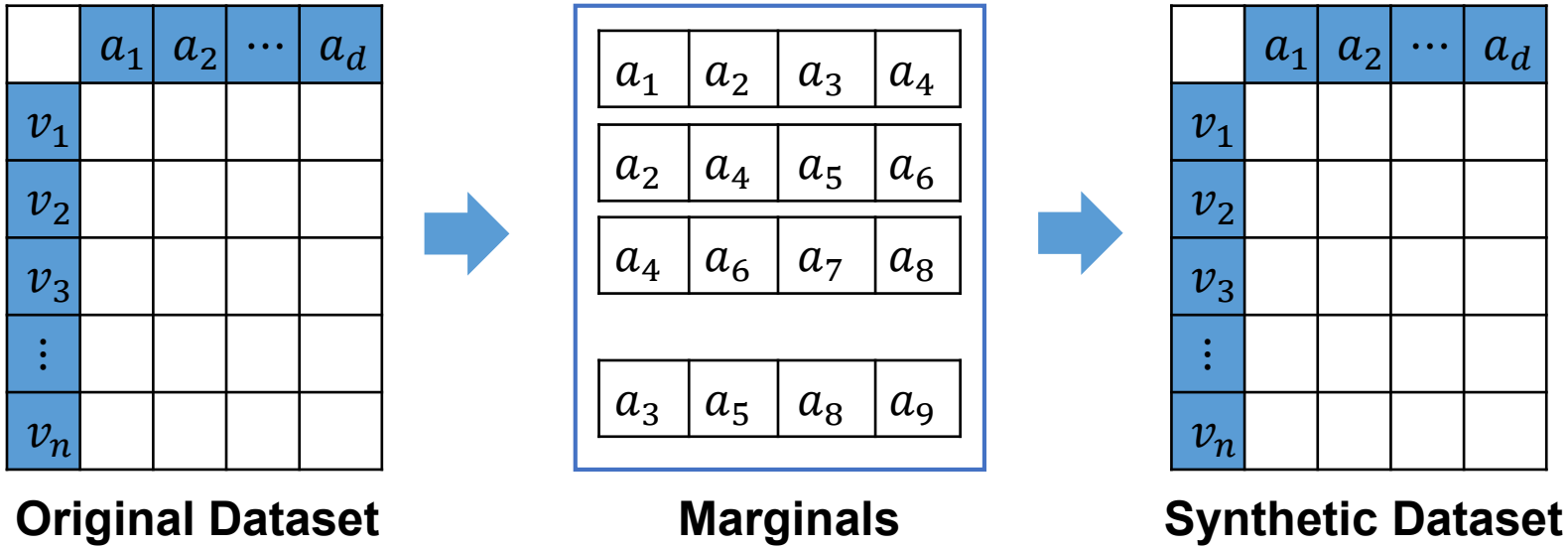


	a_1	a_2	$\cdots$	a_d
v_1				
v_2				
v_3				
$\vdots$				
v_n				

Synthetic Dataset

- **Bayesian Network [1]: Can only exploit d-1 marginals, losing many correlation information.**
- **Markov Random Field [2]: Some cliques can be very large when the number of marginals is large, leading to high storage cost.**

PrivSyn: Using Marginals

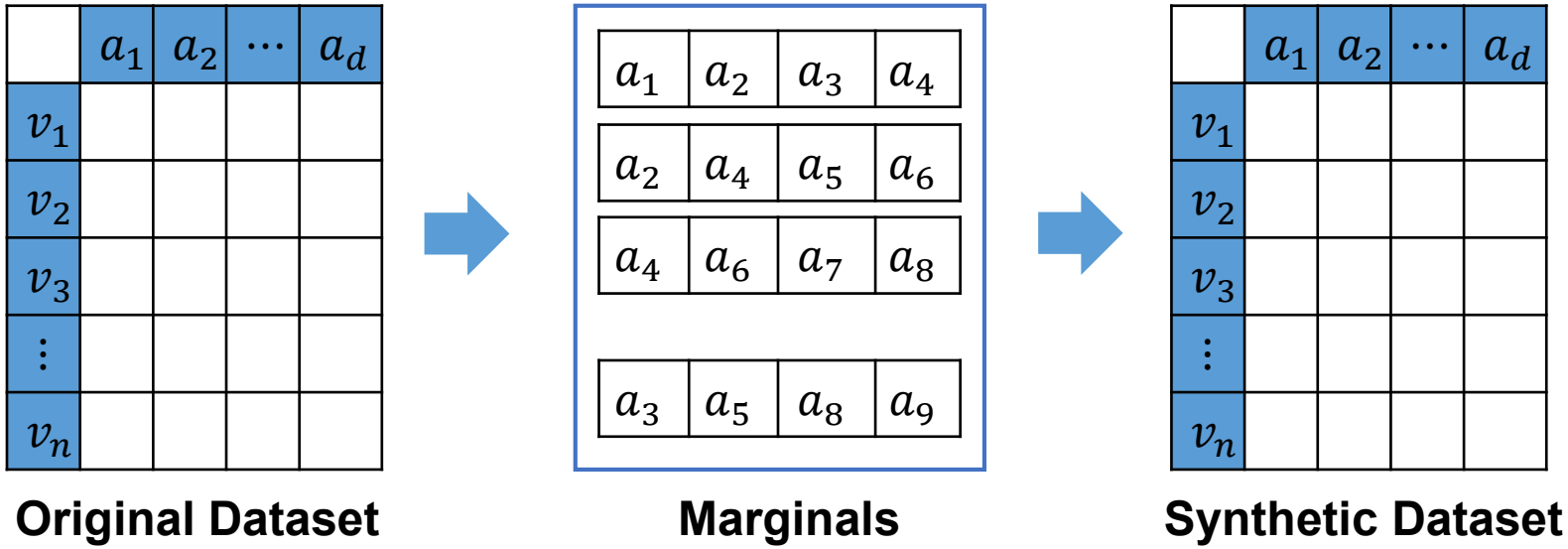


Original Dataset **Marginals** **Synthetic Dataset**

- **Use a set of consistent low-dimensional marginals as representation**
- **Synthesize data directly from marginals**
- **Can be viewed as a non-parametric method compared to PGM**

Qardaji, Yang, and Li: PriView: practical differentially private release of marginal contingency tables. SIGMOD 2014.
Zhang et al.: PrivSyn: Differentially Private Data Synthesis. USENIX Security 2021.

PrivSyn: Our Approach



Original Dataset **Marginals** **Synthetic Dataset**

- **Challenge 1: How to choose a set of marginals that capture as much as correlation information and avoid excessive noise.**
- **Challenge 2: How to generate a synthetic dataset from the selected marginals.**

Marginal Selection: DenseMarg

- Example
  - Attributes: a b c d
  - All two-way marginals: (a, b), (a, c), (a, d), (b, c), (b, d), (c, d)

Noise Error

- If a two-way marginal is chosen
  - Add **Gaussian noise** to obtain the marginals
  - Proportional to the number of cells

Dependency Error

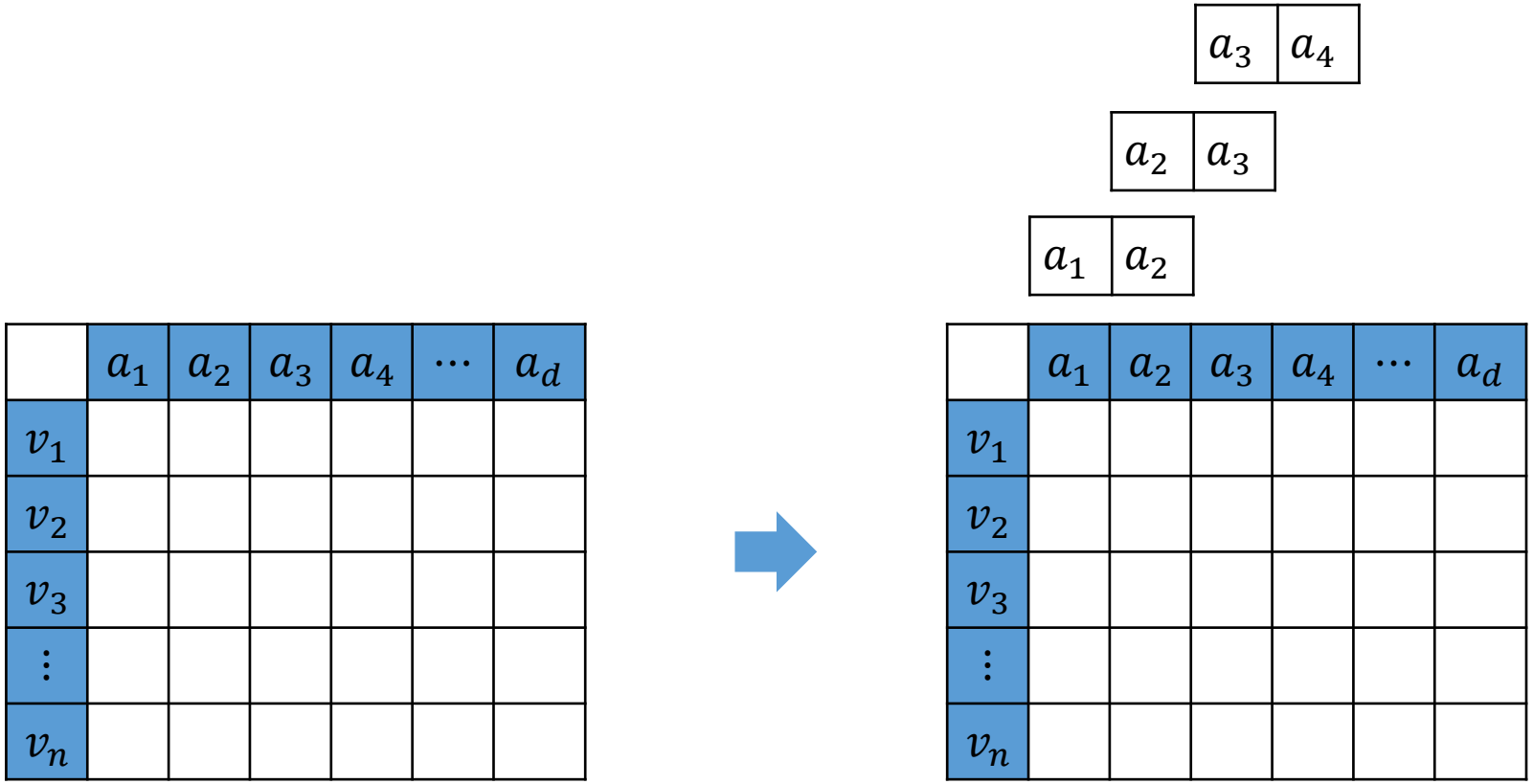
- If a two-way marginal is **not** chosen
  - Mutual information (high sensitivity)
  - **InDif** (low sensitivity)

Optimization Problem Formulation:

$$\text{minimize} \sum_{i=1}^{m} [\psi_i x_i + \phi_i (1 - x_i)]$$

$$\text{subject to } x_i \in \{0, 1\}$$

Dataset Generation: GUM



Step I: Randomly generate dataset.

Step II: Iteratively using all marginals to update the dataset.

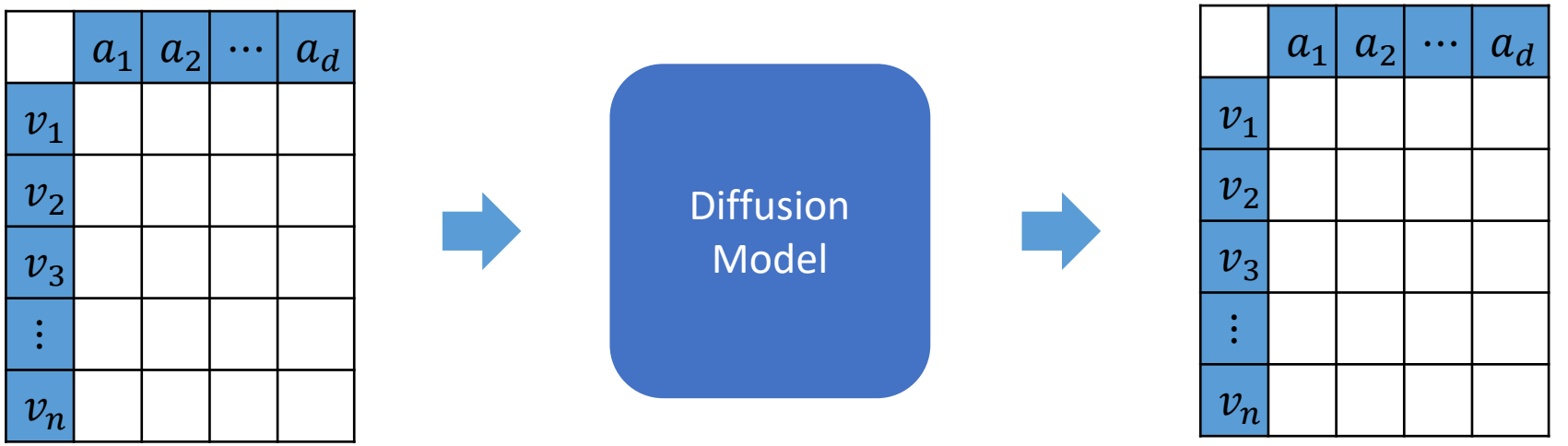
GAN-based Generative Model

Original Dataset **Synthetic Dataset**

- Using Neural Network to encode data distribution, and GAN to train the network
- CTGAN combines one-dimensional marginals with GAN
- When satisfying DP, does not perform well empirically for tabular data
  - Queries do not use privacy budget efficiently.
  - Cannot directly select what information is preserved.

Xu et al.: Modeling tabular data using conditional GAN. NeurIPS 2019.

Diffusion Models

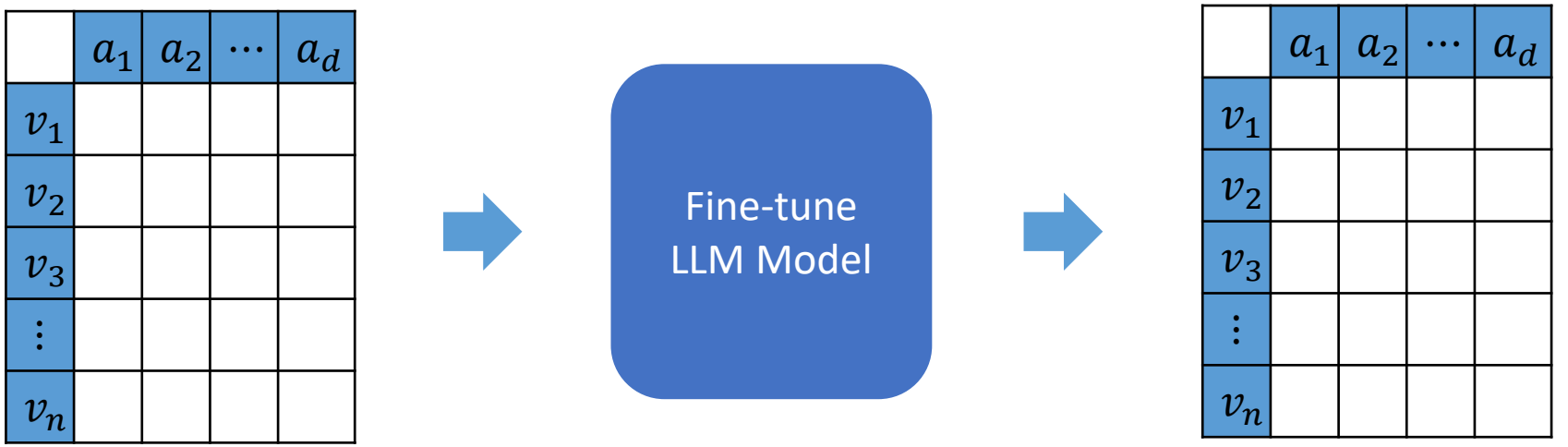


Original Dataset **Synthetic Dataset**

- Define separate forward diffusion processes
  - Gaussian diffusion models for numerical
  - Multinormial diffusion models for categorical features
- Optimization with ELBO

Kotelnikov et al.: TabDDPM: Modelling Tabular Data with Diffusion Models. ICML 2023.

Using LLM



Original Dataset **Synthetic Dataset**

- Convert input tabular data to text, with random feature order permutation
- Using the text to fine-tune an LLM model
- Can use partial records to prompt the LLM to generate new records

Borisov et al.: Language Models are Realistic Tabular Data Generators. ICLR 2023.

Assessment of Existing Methods

- Without considering privacy, Diffusion models perform the best in terms of
  - fidelity (synthesized data distribution similar to original dataset)
  - and utility (downstream ML tasks have similar performance)
- If aiming to satisfy Differential Privacy
  - PGM performs the best
- There appears to be significant memorization/leakage in diffusion-based models
- Making diffusion models satisfying DP using current techniques results in significant fidelity/utility loss.

Du & Li:Systematic Assessment of Tabular Data Synthesis Algorithms. arXiv.

Overview of Differential Privacy

Differential Privacy [Dwork et al. 2006]

- Definition: Mechanism *A* satisfies ε-Differential Privacy if and only if
  - for any neighboring datasets D and D'
  - and any possible transcript $t \in$ Range(A),

$$\Pr[A(D) = t] \leq e^{\epsilon} \Pr[A(D') = t]$$

  - For relational datasets, typically, datasets are said to be neighboring if they differ by a single record.

Why Does Differential Privacy Make Sense?

- "Privacy as Secrecy" (i.e., hiding information) is impossible if one wants to share data at all.
  - Consider the following example: Assume that smoking causes lung cancer is not yet known, and an organization conducted a study that demonstrates this connection.
  - A smoker Carl was not involved in the study, but complains that publishing the result of this study affects his privacy, because others would learn new information about him, namely he has a higher chance of getting lung cancer, and as a result he may suffer damages.
- Differential Privacy attempts to simulate "privacy as opting out": The most one can do to protect one's privacy is to take one's data out of the dataset.

Differential Privacy: Basic Mechanisms

- Laplace Mechanism:

$$A_f(D) = f(D) + Lap(\frac{\Delta_f}{\varepsilon})$$

- Exponential Mechanism: Sample an answer with probability determined by the quality of the answer.

Answering Queries Under DP

- Easy queries: their answers do not change much when one record changes
  - E.g., counting number of records that satisfy certain conditions.
- Hard queries: their answers are easily affected by one record change
  - E.g., max / min
- Usually easy to answer, but worst-case exist: their answers typically change very little when one record changes, but could change a lot in pathological cases
  - E.g., median
  - One can come up with mechanisms that work well in practice without meaningful proven bound of utility

From Relational Data to Contingency Table

Gender	Age	Income
Female	31	150k
Male	28	100k
Male	30	110k
Male	45	200k
Female	19	50k
Male	24	40k

Gender Male: 1 Female: 0	Age Larger than 25: 1 Otherwise : 0	Income More than 100k: 1 Otherwise: 0	Count
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	3

From Contingency Table to Marginal Tables

Gen	Age	Inc	Cnt
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	3

Gen	Age	Cnt
0	0	1
0	1	1
1	0	1
1	1	3

Gen	Cnt
0	2
1	4

Marginal table queries can be answered with the same privacy cost as each counting query. They are arguably the most privacy-efficient way to extract information from input dataset.

Hueristic Privacy Metrics: Distance to Closest Records (DCR)

- Given a synthetic dataset and the original dataset
  - Compute the distances from each synthetic data point to its nearest real one
  - Uses the 5th percentile (or the mean) as the privacy score
  - A small score indicates that the synthetic dataset is too close (similar) to real data, signaling a high risk of information leakage.
- Weaknesses
  - Underestimates privacy risk in some cases because it does not consider the worst-case synthesized record
  - Overestimates privacy risk in other cases when there are close clusters of data points.

Du & Li:Systematic Assessment of Tabular Data Synthesis Algorithms. arXiv.

Proposed Metric: Membership Disclosure Score (MDS)

- First, given an input dataset D and synthesizer A, we quantify the disclosure risk of one record $x \in$ D, DS(x, A, D), as:
  - Expected difference between
    - Distance of closest record to x in synthetic dataset when x is included in input to A
    - And such distance when x is not included in input to A

- MDS(A,D) is defined to be $\max_{x \in D}$ DS (x, A, D)

- Open question: Is MDS a sufficiently strong privacy metric?

Du & Li:Systematic Assessment of Tabular Data Synthesis Algorithms. arXiv.

Thank you for your listening

Q & A

Email: ninghui@purdue.edu