# FLAIM: AIM-based Synthetic Data Generation in the Federated Setting

Samuel Maddock[1], Graham Cormode[1 2], Carsten Maple[1]

*[1]University of Warwick, [2]Meta AI*


KDD2024
BARCELONA, SPAIN

# Synthetic Data Generation (SDG)

- **Goal:** Produce “fake” data with the properties of real data

- Synthetic data attractive for many reasons
  - **Key reason:** Privacy
  - Allow general release for downstream tasks e.g., training models, analytics

- Lots of solutions when data is centralised in one place
  - GANs, LLMs, Statistical models, etc.

- Methods prone to “memorisation”
  - Can produce verbatim copies of real data
  - Prevention via **Differential Privacy (DP)**

KDD2024
BARCELONA, SPAIN

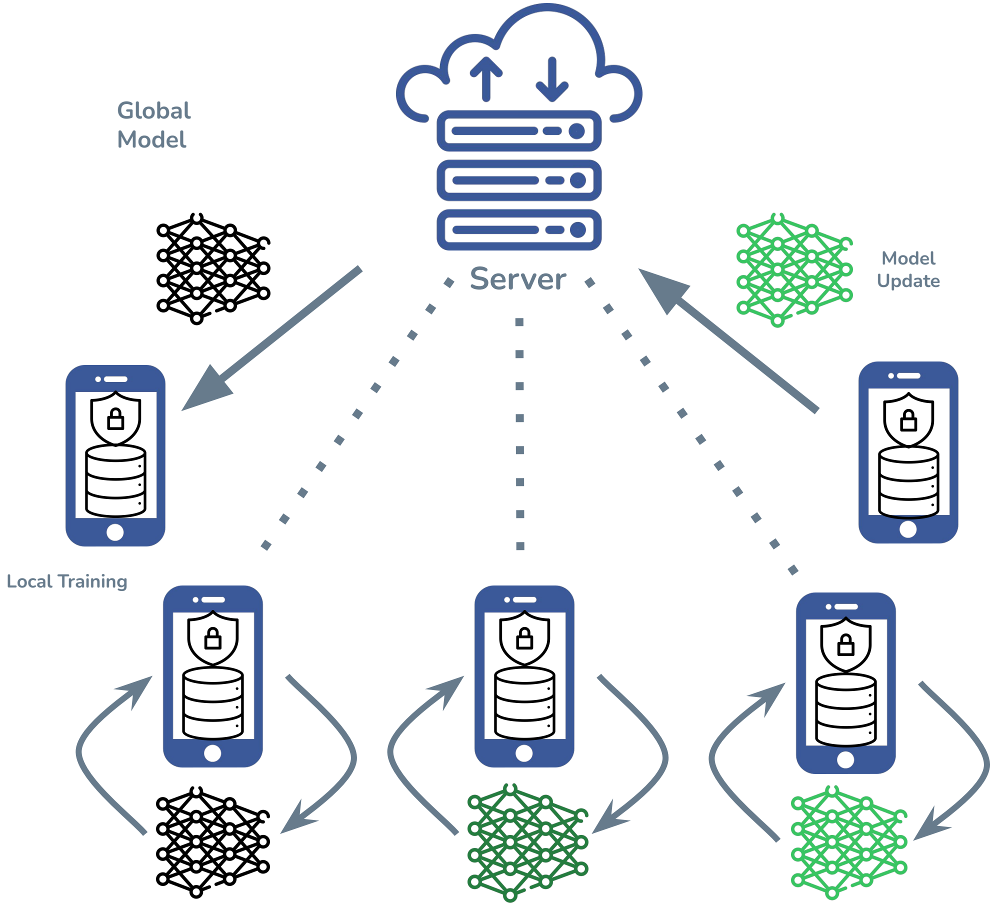
# Federated Learning (FL)

- **Federated Setting**
  - Millions of clients, holding local data
  - Wish to participate in model training
  - Perform local work and send to server

- "Realistic" scenario for large organisations

- Synthetic data not well studied in FL
  - Generic image/language generation (e.g. GANs)

- **Our focus:** Federated Synthetic Tabular Data




KDD2024
BARCELONA, SPAIN

# Differential Privacy (DP)

- Parameterized by ($\varepsilon$ , $\delta$ ):
  - $\varepsilon$ - Privacy budget, larger implies less privacy (noise)
  - $\delta$ - Small probability of failure, set “cryptographically” small

- To guarantee DP → add noise into training process

- Smaller the privacy budget = more noise needed

- Has many useful properties
  - Post-processing
  - Composition

# Differentially Private Synthetic Data Generators (DP-SDG)

- Define workload of queries **Q**

- **Goal:** Produce synthetic data with accurate answers over workload **Q**
- **Example:** Marginal Query e.g., “How many rows have Sex=”M” and Employed=True?”
- **Want to learn:** Model producing synthetic data with low error over **Q**

- Data can still be used for any number of downstream tasks
  - e.g., training ML models
  - No guarantees outside defined workload **Q**

KDD2024 BARCELONA, SPAIN

# DP-SDG: "Select-Measure-Generate"

- Private tabular SDG methods follow **"Select-Measure-Generate"**

- For t = 1, … ,T
  1. **Select:** query q ∈ Q with highest error (privately)
     a. *Exponential mechanism with utility scores u(q)*

  2. **Measure:** Measure chosen marginal q under calibrated noise
     a. *Gaussian mechanism*

  3. **Generate:** Update model to learn noisy marginal

# Adaptive Iterative Mechanism (AIM)

*McKenna et al. (VLDB 24)*

- Follows "*Select-Measure-Generate*" paradigm
  - ("Generate") - uses Private-PGM → Markov Random Field (MRF)

- Modifications to improve utility:
  - **Augmented utility scores -** "Select" step performed in smarter way
  - **Budget annealing -** Rounds (T) do not need to be set in advance
  - **zCDP accounting** - Add less noise for same privacy guarantees

- Translating AIM to the federated setting is the core focus of our work

# Federated DP-SDG

- **Key Question:** How do we federate AIM?
  - = how to federate "Select-Measure-Generate" paradigm

- **Distributed setting**
  - All clients participate over a single (or few) rounds
  - Typically assume all participants are available

- **Federated setting**
  - Client participation is subset of true population (e.g., dropout, availability)
  - Client data exhibits strong heterogeneity (e.g., distribution skew)

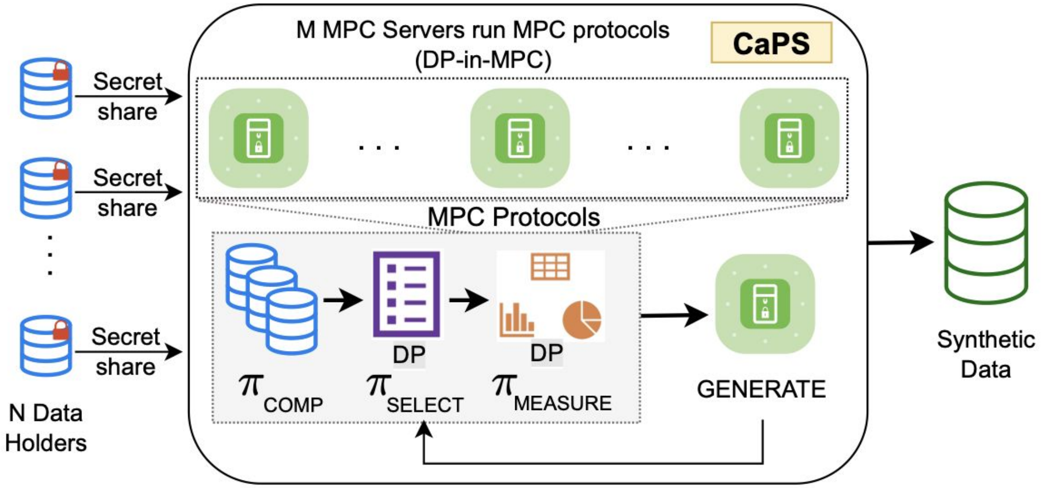
KDD2024
BARCELONA, SPAIN

# **Prior Work:** Pereira et al. 2022

- **Distributed setting**
  - Secure Multi-party Computation (MPC)
  - 2/3-party settings, all clients available

- All clients secret-share workload answers to computing server(s)
- Servers work to emulate central algorithm
  - Distributed **Select** + **Measure** steps

- **Drawbacks**
  - Focus on MWEM - poor data representation
  - “Fully-MPC” solution has overheads




KDD2024 BARCELONA, SPAIN

# **Our Work:** Distributed AIM

- Pereira et al., 2022 distribute MWEM using MPC

- **Our Work:** DistAIM
    - Plug AIM into their framework replacing MWEM
    - Gain utility boost due to AIM over prior work

- **Problem:** not designed with FL in mind - inherits issues of Pereira et al.
    1. Assumes all clients available to secret-share answers
    2. Overhead for clients sharing all workload answers
    3. Overhead for server due to MPC operations for exponential mechanism

KDD2024 BARCELONA, SPAIN

# **Our Work:** Naive FLAIM

- DistAIM obtains good utility but w/ overheads not compatible with typical FL
- Can we design an analog to traditional FL training?
  - Offload work to clients (make local steps)
  - Client(s) distill work into update => server aggregates and updates global model

- FLAIM
  - **“Select”:** have each (available) client perform a number of local steps
    - Under LDP
  - **“Measure”**: server performs under lightweight cryptography i.e., *secure-aggregation*
    - Distributed DP
  - **“Generate”:** update graphical model => post-processing
- Avoids (heavy) MPC → secure exponential mechanism

KDD2024 BARCELONA, SPAIN

# **Our Work:** AugFLAIM (Non-private)

- **Problem:** clients w/ strong heterogeneity more likely to choose skewed marginals

$$u(q; D_k) \propto ||M_q(D_k) - M_q(\hat{D}^{(t)})||_1$$

- **Solution:** correct local skew by penalising q with strong heterogeneity

- **How to define heterogeneity?** Deviation of clients marginal from global

$$\tau_k(q) := ||M_q(D_k) - M_q(D)||_1$$

- **Problem:** $M_q(D)$ is *exactly* what we are trying to learn (privately) via AIM !

KDD2024
BARCELONA, SPAIN

# **Our Work:** AugFLAIM (Private)

- **Problem:** Can't ever learn "true" heterogeneity of clients local marginals

- **Private Proxy:** have clients submit 1-way marginals every round
    - Pay privacy cost in the number of features
    - Obtain subsequently more accurate 1-way answers

$$\tilde{\tau}_k(q) := \frac{1}{|q|} \sum_{j \in q} \|M_{\{j\}}(D_k) - \tilde{M}_{\{j\}}(D)\|_1$$

KDD2024
BARCELONA, SPAIN

# Methods

1. Naive FLAIM
    - Translation of AIM to FL with no modifications
    - “SecAgg + noise”

2. AugFLAIM (Oracle)
    - Assumes knowledge of heterogeneity skew
    - Modify select step for local clients taking this into account

3. AugFLAIM (Private)
    - Private proxy of heterogeneity
    - Estimates all 1-way marginals and query from “select” step at each round

KDD2024 BARCELONA, SPAIN

# **Experiment:** Comparison with Baselines

- Popular deep learning alternative
  - **DP-CTGAN**

- FLAIM baselines
  - **NaiveBayes** - 1-way marginals only
  - **FLAIM (Random)** - random decisions
  - **NaiveFLAIM** - no modification to utility score

- Our proposal: **AugFLAIM (Private)**
- Table shows NLL compared to test set

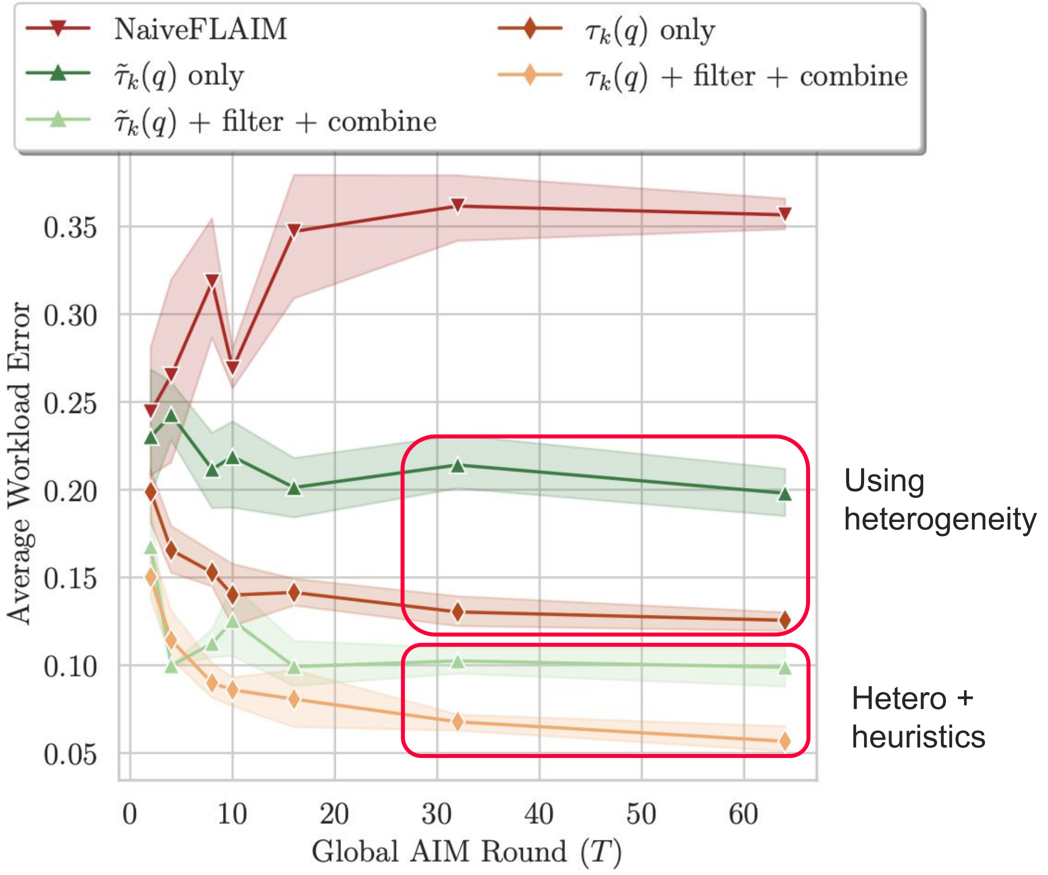
**Table 1: Comparison of FLAIM approaches against baselines for negative log-likelihood (NLL), $\varepsilon = 5$. Smaller NLL is better.**

| **Method / Dataset** | Adult | Credit | Covtype |
|---|---|---|---|
| Fed DP-CTGAN | 37.1 | 83.8 | 62.7 |
| FedNaiveBayes | 25.33 | 18.02 | 44.9 |
| FLAIM (Random) | 83.9 | 47.7 | 58.4 |
| NaiveFLAIM | 29.4 | 18 | 45.4 |
| AugFLAIM (Private) | **20.87** | **16.2** | **41.6** |
| DP-CTGAN | 28.6 | 27.6 | 45.9 |
| AIM | **19.2** | **15.57** | **40.92** |

KDD2024
BARCELONA, SPAIN

# Experiment: Ablation

- Why does AugFLAIM (Private) perform so well?
- NaiveFLAIM
    - No utility score modification
- AugFLAIM (Oracle)
    - Access to true heterogeneity
- AugFLAIM (Private)
    - Private proxy for heterogeneity




(b) Credit

KDD2024 BARCELONA, SPAIN

# **Experiment:** Overheads

- **If T is small**
  - Utility of AugFLAIM >= DistAIM
- **If T is large**
  - DistAIM favorable performance
- **Bandwidth** = Average client sent & receivec
- On Adult, DistAIM requires
  - 2x more rounds
  - 1300x increase in bandwidth
  - to reduce workload error by ~½

Table 2: Overhead of DistAIM vs. FLAIM at optimal $T$

| | $T(\uparrow)$ | Bandwidth $(\uparrow)$ | Err $(\downarrow)$ | NLL $(\downarrow)$ |
|---|---|---|---|---|
| Adult | $2\times$ | $1300\times$ | $0.58\times$ | $0.1\times$ |
| Magic | $3.2\times$ | $1643\times$ | $0.19\times$ | $0.15\times$ |
| Mushroom | $7\times$ | $7.5\times$ | $0.79x$ | $0.4\times$ |
| Nursery | $20\times$ | $3.4\times$ | $0.89\times$ | $0.17\times$ |

KDD2024
BARCELONA, SPAIN

# Conclusion

- FLAIM provides a way to
  - obtain comparable utility to DistAIM in practical FL
  - whilst reducing client overheads via lightweight MPC

- Limitations
  - Example-level DP
  - Inherits limitations of “select-measure-generate”
    - Continuous features
    - Specifying a workload Q
    - High-dimensional datasets

**Poster Number 91**
**Today, 6:30pm**

***arXiv:2310.03447***

KDD2024
BARCELONA, SPAIN