IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Overcoming data-sharing challenges in central banking: federated learning of diffusion models for synthetic mixed-type tabular data generation[1]

Timur Sattarov,
Deutsche Bundesbank,

Marco Schreyer,
University of St Gallen

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Overcoming data-sharing challenges in central banking: Federated Learning of Diffusion Models for Synthetic Mixed-Type Tabular Data Generation

Timur Sattarov[1,2] Marco Schreyer[2]

Abstract

Realistic synthetic tabular data generation encounters significant challenges in preserving privacy, especially when dealing with sensitive information in domains like finance. In this paper, we introduce FedTabDiff, a novel federated learning framework for generating high-fidelity mixed-type tabular data without centralized access to the original datasets. Leveraging the strengths of Denoising Diffusion Probabilistic Models (DDPM), our approach addresses the inherent complexities in tabular data, such as mixed attribute types and implicit relationships. More critically, FedTabDiff realizes a decentralized learning scheme that permits multiple entities to collaboratively train a generative model while respecting data privacy and locality. We extend DDPM into the federated setting for tabular data generation, which includes a synchronous update scheme and weighted averaging for effective model aggregation. Experimental evaluations on real-world financial and medical datasets attest to the framework's capability to produce synthetic data that maintains high fidelity, utility, privacy, and coverage.

Keywords: federated learning, diffusion models, tabular data, data sharing

JEL classification: C81, C18

[1] Research Data and Service Centre, Deutsche Bundesbank, Frankfurt am Main, Germany (timur.sattarov@bundesbank.de). The views expressed in this article are those of the author and do not necessarily represent those of the Bundesbank or the Eurosystem.

[2] School of Computer Science, University of St. Gallen, St. Gallen, Switzerland (marco.schreyer@unisg.ch)

Contents

Overcoming data-sharing challenges in central banking: Federated Learning of Diffusion Models for Synthetic Mixed-Type Tabular Data Generation 1

1. Introduction 3

2. Related Work 4

3. Federated Learning of Diffusion Models 5

4. Experimental Setup 7

   Datasets and Data Preparation 7

   Model Architecture 8

   Evaluation Metrics 9

5. Experimental Results 10

6. Conclusion 12

References 12

1. Introduction

In the rapidly changing financial regulatory landscape, data analytics has become increasingly vital, with central banks collecting extensive microdata to guide policy, assess risks, and maintain stability globally. The detailed nature of this data introduces unique challenges, notably in data privacy. Real-world tabular data is crucial for developing complex, real-life dynamic models but often contains sensitive information, necessitating strict regulations like the EU's *General Data Protection Regulation* and the *California Privacy Rights Act*. Despite implementing such safeguards, institutions remain wary of deploying Artificial Intelligence (AI) models due to risks of data leakage (Carlini et al. 2020; Bender et al. 2021; Kairouz et al. 2019) and model attacks (Fredrikson, Jha, and Ristenpart 2015; Shokri et al. 2017; Salem et al. 2018), which could potentially reveal personally identifiable information or confidential training data (Carlini et al. 2020; Bender et al. 2021; Kairouz et al. 2019; Fredrikson, Jha, and Ristenpart 2015; Shokri et al. 2017; Salem et al. 2018). A promising avenue to address these risks originates from generating high-quality synthetic data. In this context, the term '*synthetic data*' denotes data of a generative process grounded in the inherent properties of real data. Modeling these processes provides a nuanced understanding of underlying patterns, unlocking insights, especially in high-stake domains (Assefa et al. 2020; Schreyer et al. 2019). Generating high-fidelity synthetic tabular data is crucial for regulatory-compliant data sharing, fostering collaboration among various entities, and enabling the modeling of rare but impactful events like fraud (Charitou, Dragicevic, and d'Avila Garcez 2021; Barse, Kvarnstrom, and Jonsson 2003). The Irving Fisher Committee (IFC) and the United Nations Economic Commission for Europe (UN-ECE) have both underscored the importance of data sharing and collaboration in the financial sector. The IFC's guidance note[3] discusses the challenges and benefits of cross-border cooperation, advocating for harmonized data practices and the use of advanced technologies. Similarly, the UNECE's guide[4] recommends that national statistical offices expand their data sources beyond national confines, incorporating international statistics to improve economic data. However, data sharing comes with a number of challenges such as data privacy or security concerns which can have severe consequences in case of any breaches of confidential information. In addition, the data transmission plays a big role as moving extensive datasets between central banks can become bandwidth-intensive and expensive. This approach is particularly pertinent as real-world tabular data is often characterized by inherent complexities, including:

1. **Mixed Attribute Types**: Tabular data comprises diverse attribute data types, including categorical, numerical, and ordinal data distributions. Modeling these complexities requires effectively integrating different data types into a cohesive model.

2. **Implicit Relationships**: Tabular data encompasses implicit relationships between individual records and attributes. Modeling these complexities necessitates

[3] https://www.bis.org/ifc/data_sharing_practices.pdf

[4] https://unece.org/sites/default/files/2021-02/Data%20sharing%20guide%20on%20web_1.pdf

disentangling the dependencies into a model representing the relationships between records and attributes.

3. **Distribution Imbalance**: The skewed distributions and imbalances in financial tabular data represent a significant hurdle. These factors demand advanced modeling techniques that precisely encapsulate and represent the nuanced patterns in financial data.

In recent years, deep generative models (Goodfellow, Shlens, and Szegedy 2014) have made impressive strides in the creation of diverse and realistic content, such as images (Brock, Donahue, and Simonyan 2018; Rombach et al. 2022), videos (Chan et al. 2019; Singer et al. 2022; Yu et al. 2023), audio (Wang et al. 2023; Le et al. 2023; Bai et al. 2022), code (Li et al. 2022; Chen et al. 2022; Le et al. 2022), and natural language (OpenAI 2023; Bubeck et al. 2023; Touvron et al. 2023). Lately, ***Denoising Diffusion Probabilistic Models (DDPM)*** (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) demonstrated the ability to generate synthetic images of exceptional quality and realism (Dhariwal and Nichol 2021; Rombach et al. 2022; Gao et al. 2023). To effectively train these models, characterized by extensive parameters, substantial compute resources, and extensive training data are crucial (de Goede 2023). However, challenges arise when sensitive data is distributed across various institutions, such as financial authorities, and cannot be shared due to privacy concerns (Assefa et al. 2020; Schreyer et al. 2022; Schreyer, Sattarov, and Borth 2022). Recently, the concept of ***Federated Learning (FL)*** (McMahan et al. 2017; McMahan and Ram-age 2017) was proposed in which multiple devices, such as smartphones or servers, collaboratively train AI models under the orchestration of a central server (Kairouz et al. 2019). The training data remains decentralized, providing a pathway to learn a shared model without direct data exchange.

To summarize, the main contributions of this study are:

- Introduction of *FedTabDiff*, a novel federated learning framework for generating synthetic mixed-type tabular data, merging denoising diffusion models with federated learning to enhance data privacy.
- Empirical evaluation of *FedTabDiff* using real-world financial and healthcare datasets, illustrating its efficacy in synthesizing high-quality, privacy-compliant data.

The reference implementation of the code can be accessed through: https://github.com/sattarov/FedTabDiff.

2. Related Work

Lately, diffusion models (Cao et al. 2022; Yang et al. 2022; Croitoru et al. 2023) and federated learning (Aledhari et al. 2020; Li et al. 2021; Zhang et al. 2021a) triggered a considerable amount of research. In the realm of this work, the following review of related literature focuses on the federated deep generative modeling of tabular data:

Deep Generative Models: Xu et al. (Xu et al. 2019) introduced CTGAN, a conditional generator for tabular data, addressing mixed data types to surpass previous models' limitations. Building on GANs for oversampling, Engelmann and Lessmann (Engelmann and Lessmann 2021) proposed a solution for class imbalances by integrating conditional Wasserstein GANs with auxiliary classifier loss. Jordon et al. (Jordon, Yoon, and Van Der Schaar 2018) formulated PATE-GAN to enhance data synthesis privacy, providing differential privacy guarantees by modifying the PATE framework. Torfi et al. (Torfi, Fox, and Reddy 2022) presented a differentially private framework focusing on preserving synthetic healthcare data characteristics. To handle diverse data types more efficiently, Zhao et al. (Zhao et al. 2021) developed CTAB-GAN, a conditional table GAN that efficiently addresses data imbalance and distributions. Kim et al. (Kim et al. 2021) enhanced synthetic tabular data utility using neural ODEs, and Wen et al. (Wen et al. 2022) introduced Causal-TGAN, leveraging inter-variable causal relationships to improve generated data quality. Zhang et al. (Zhang et al. 2021b) offered GANBLR for a deeper understanding of feature importance, and Noock and Guillame-Bert (Nock and Guillame-Bert 2022) proposed a tree-based approach as an interpretable alternative. Kotelnikov et al.(Kotelnikov et al. 2022) explored tabular data modeling using multinomial diffusion models (Hoogeboom et al. 2021) and one-hot encodings, while FinDiff (Sattarov, Schreyer, and Borth 2023), foundational for our framework, uses embeddings for encoding.

Federated Deep Generative Models: De Goede et al. in (de Goede 2023) devise a federated diffusion model framework utilizing Federated Averaging (McMahan et al. 2017) and a UNet (Ronneberger, Fischer, and Brox 2015) backbone algorithm to train DDPMs on the Fashion-MNIST and CelebA datasets. This approach reduces the parameter exchange during training without compromising image quality. Concurrently, Jothiraj and Mashhadi in (Jothiraj and Mashhadi 2023) introduce *Phoenix*, an unconditional diffusion model that employs a UNet (Ronneberger, Fischer, and Brox 2015) backbone to train DDPMs on the CIFAR-10 image database. Both studies underscore the pivotal role of federated learning techniques in advancing the domain.

To the best of our knowledge, this is the first attempt utilizing diffusion models in the federated learning setup for synthesizing mixed-type tabular financial data.

3. Federated Learning of Diffusion Models

This section outlines our proposed *FedTabDiff* model for integrating *Federated Learning* (FL) with *Denoising Diffusion Probabilistic Models* (DDPM), aiming to enhance the privacy of generated mixed-type tabular data.

Gaussian Diffusion Models. The Denoising Diffusion Probabilistic Model (Sohl-Dickstein et al. 2015), (Ho, Jain, and Abbeel 2020) is a latent variable model that uses a forward process to perturb data $x_0 \in \mathbb{R}^d$ incrementally with Gaussian noise ϵ, then restores it through a reverse process. Starting at x_0, latent variables $x_1, \ldots, x_T$ are derived via a Markov Chain, transforming them into Gaussian noise $x_T \sim \mathcal{N}(0,\mathrm{I})$, defined as: $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I})$

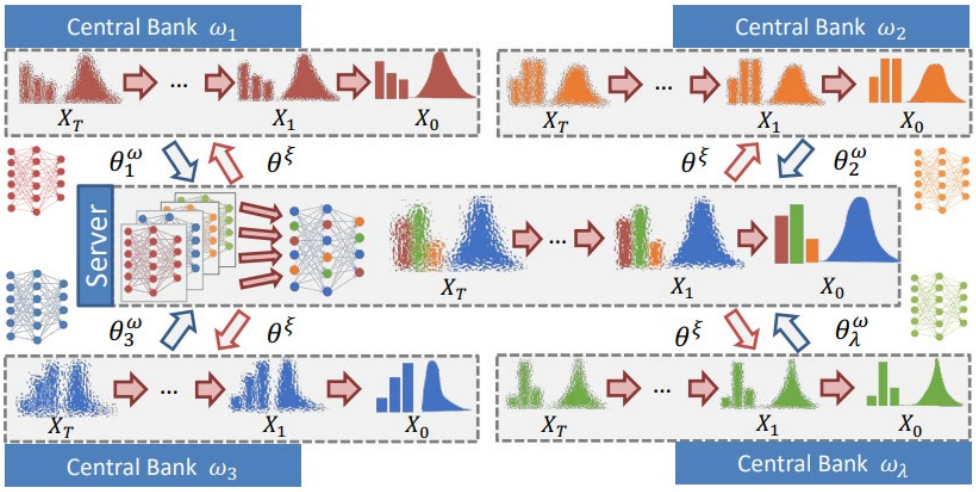


Figure 1. *Schematic illustration of the FedTabDiff model depicting each central bank ω_i training a local diffusion model 'FinDiff'. Each timestep $X_T, \dots, X_1, X_0$ reflects the latent data representations in the reverse diffusion. Model parameters θ_i^ω are aggregated on the server producing model θ^ξ, which is subsequently shared back after every training round r = 1,..., R.*

Here, β_t is the noise level at timestep *t*. Sampling x_t from x_0 for any *t* is expressed as

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{1-\widehat{\beta}_t}x_0, \widehat{\beta}_t \mathrm{I}),$$

where $\widehat{\beta}_t = 1 - \prod_{i=0}^{t}(1-\beta_i)$. In the reverse process, the model denoises x_t to recover x_0. To approximate, a neural network with parameters θ is trained, parameterizing each step as following:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$

where μ_θ and Σ_θ are the estimated mean and covariance. Following Ho et al. (Ho, Jain, and Abbeel 2020), with Σ_θ being diagonal, μ_θ is computed as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\widehat{\alpha_t}}}\epsilon_\theta(x_t, t))$$

Here, $\alpha_t \coloneqq 1 - \beta_t, \widehat{\alpha_t} \coloneqq \prod_{i=0}^{t}\alpha_i$ and $\epsilon_\theta(x_t, t)$ is the predicted noise component. Empirical evidence shows that using a simplified mean-squared error loss yields superior results compared to the variational lower bound $log\ p_{(\theta)}(x_0)$.

$$\mathcal{L}_t = \mathbb{E}_{x_0,\epsilon,t}\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2$$

Federated Learning. The DDPM learning process is enhanced by *Federated Learning* (McMahan et al. 2017). FL enables knowledge acquisition from data distributed across *C* clients, denoted as $\{\omega_i\}_{i=1}^{c}$. The training dataset is partitioned into *C* sub-datasets, $D = \{D_i\}_{i=1}^{c}$, each privately accessible only to a single client ω_i. Data subsets D_i and D_j, where $i \neq j$, can have different data distributions. We build on the FinDiff model for mixed-type tabular data generation (Sattarov, Schreyer, and Borth 2023), inspired by the work presented in (de Goede 2023; Jothiraj and Mashhadi 2023). We introduce a central FinDiff model f_θ^ξ with parameters θ^ξ, maintained by a trusted entity and collectively learned by clients $\{\omega_i\}_{i=1}^{c}$. Each client ω_i maintains a decentralized FinDiff

model $f_{\theta,i}^{\xi}$ and contributes to learning the central model. The federated optimization uses a synchronous update scheme over $r = 1, \ldots, R$ communication rounds (McMahan et al. 2017). In each round, a subset of clients $\omega_{i,r} \subseteq \{\omega_i\}_{i=1}^{c}$ is selected. These clients receive the current central model parameters $\theta_{i,r}^{\xi}$, perform optimization, and then send their updated model parameters back for aggregation. The schematic process with four clients is depicted on fig. 1. Following Federated Averaging (McMahan et al. 2017) is used to compute a weighted average over the decentralized model updates, defined as:

$$\theta_{r+1}^{\xi} \leftarrow \frac{1}{D} \sum_{i=1}^{\lambda} |D_i| \, \theta_{i,r+1}^{\omega}$$

where λ denotes the clients, r current communication round, $|D|$ is the total sample count, and $|D_i| \subseteq |D|$ is the number of samples accessible by the client ω_i.

4. Experimental Setup

This section describes the details of the conducted experiments, encompassing the data preparation steps, model architecture, and evaluation metrics.

Datasets and Data Preparation

Two real-world tabular datasets are used for the evaluation:

1. **City of Philadelphia Payments**[5] encompasses a total of 238,894 payments generated by 58 distinct city departments in 2017. Each payment includes 10 categorical and 1 numerical attribute. The feature *doc_ref_no_prefix* definition (document reference number prefix definition) was used for the non-iid split.

2. **Diabetes hospital data**[6] encompasses a total of 101,767 clinical care records collected by 130 US hospitals between the years 1999-2008. Each care record includes 40 categorical and 8 numerical attributes. The feature *age* (a patient's age group) was used for the non-iid split.

Non-iid partition. We aim to substantiate the model's efficacy and effectiveness, ensuring generalizability across non-iid data splits. The data is non-iid partitioned by segregating based on a categorical feature. The five dominant categories within this feature serve as a basis for data allocation to different clients, thereby introducing a realistic non-iid and unbalanced data environment for federated training. We present the descriptive statistics and details on the non-iid data partitioning scheme in the table 1.

[5] https://www.phila.gov/2019-03-29-philadelphias-initial-release-of-city-payments-data

[6] https://www.kaggle.com/datasets/brandao/diabetes

Table 1. *Non-iid and unbalanced data partitioning scheme. Only a subset $D_i \subset D$ is privately accessible to a client ω_i.*

	# samples in D_i	
Client	**Philadelphia**	**Diabetes**
ω_1	40,038	9,685
ω_2	28,521	17,256
ω_3	16,831	22,483
ω_4	93,119	26,068
ω_5	36,793	17,197
all	215,302	92,689

Numeric features were normalized using the Quantile Transformer[7] and categorical attributes were encoded using embeddings as per (Sattarov, Schreyer, and Borth 2023).

Model Architecture

The model architecture for both datasets involves the feed forward neural network with four layers of 1024 neurons each. Models are trained for up to 1000 communication rounds R = 1000 with a mini-batch size of 512 using Adam optimizer (Kingma and Ba 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$. PyTorch v2.0.1 (Paszke and Gross 2019) is used for optimization of the model parameters.

Diffusion Model Training Hyperparameters. We determined that the optimal number of diffusion steps for training the diffusion model is T = 500. We used a linear scheduler with $\beta_{start} = 0.0001$ and $\beta_{end} = 0.02$. Also, each embedding dimension of categorical attributes was assigned to 2.

Federated Learning Hyperparameters. At every communication round $r = 1, \dots, R$ each client ω_i performed 20 model optimization updates (client rounds) on its local model θ_i^{ω} before sharing the model parameters. All five clients $\lambda = 5$ were selected at every communication round for decentralized optimization iteration. The Flower framework v1.4.0 (Beutel et al. 2022) simulates the federated learning scenario and trains the model with multiple clients.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html

Evaluation Metrics

We employ a series of evaluation metrics, including fidelity, utility, coverage, and privacy, to evaluate the effectiveness of our model quantitatively. The metrics are selected to represent diverse aspects of data generation quality, ensuring a holistic view of model performance.

Fidelity. Fidelity evaluates the extent to which the synthetic data replicates the real data, a process executed at both column and row levels. The fidelity of columns is investigated by assessing the similarity between each column in the synthetic and real datasets. For numeric attributes, the Kolmogorov-Smirnov statistic (Massey 1951), is applied to quantify the utmost difference between empirical and theoretical cumulative distributions. The Total Variation Distance measures the disparity between the real and synthetic categorical columns. It is defined as $\sum_{c \in C} | p(x^{d_c}) - p(s^{d_c}) |$, with $p(s^{d_c})$ representing the proportion of existing categories c in attribute d. The row fidelity is measured using correlations between a pair of columns. For numeric attributes, a Pearson correlation (Benesty et al. 2009) is computed between a pair of columns. Next, the absolute difference between the correlations of the real and synthetic attribute pairs is used to estimate the coefficient. We used the contingency table for categorical attributes by computing the Total Variation Distance (TVD) on a pair of categories in attributes a and b.

Utility. The evaluation of synthetic data crucially depends on the measure of its utility, which involves quantifying its functional equivalence to real-world data. This metric involves training machine learning models on synthetic datasets and subsequent evaluations on original ones. This criterion measures the applicability and quality of synthesized data in various downstream tasks, like classification or anomaly detection. This study measures utility by training classifiers on synthetic data sharing the same dimensions with the training set, followed by evaluations on the actual test set. The average accuracy denotes the resultant utility, consolidated over all classifiers. The classifiers utilized in this study for a diversified approach include Random Forest, Decision Trees, Logistic Regression, Ada Boost, and Naive Bayes.

Coverage. The coverage metric quantifies how much a synthetic column encompasses all potential categories found within a real column. In dealing with categorical features, this metric initially calculates the number of unique categories inherent in the real column. Subsequently, it determines the presence of those categories within the synthetic column, and returns the proportion of real categories mirrored in the synthetic data.

Privacy. The privacy metric quantifies the degree to which the synthetic data inhibits the identification of original data entries. The Distance to Closest Records (DCR) is utilized to assess the privacy of the generated data. The DCR is determined as the nearest distance from a synthetic data point to the authentic data points. The final score is the median of the DCRs for all synthetic data points.

The implementation of the evaluation metrics originates from the Synthetic Data Vault (SDV) library v1.3.0 (Patki, Wedge, and Veeramachaneni 2016).

Evaluation Process. The effectiveness of the *FedTabDiff* federated approach is assessed by comparing it with a non-federated scenario. Here, each client trains a

FinDiff synthesizer, and the quality of generated data is evaluated against all data partitions, including the entire dataset. This comparison helps determine the added value of the federated setting and whether the global model effectively aggregates comprehensive data knowledge. Furthermore, it reveals whether clients (or central banks), especially those from underrepresented groups with limited data volume, benefit from the global model's capability to generate data reflecting broad characteristics.

5. Experimental Results

This section presents the results of the experiments, demonstrating the efficacy of the *FedTabDiff* model and providing quantitative analyses. The results, detailed in table 2 together with depicted heatmaps in fig. 2, compare the Federated *FedTabDiff* model with Non-Federated FinDiff models across fidelity, utility, coverage, and privacy.

***Table 2**. Comparative analysis of the Federated (FedTabDiff) versus Non-Federated (FinDiff) models, evaluated using the full dataset D. Non-Federated diffusion models are trained individually at each client ω_i with subset $D_i \subset D$ (reflected in column `Split') and evaluated against the entire dataset D.*

			Evaluation Measures			
Dataset	**Client**	**Split D_i**	**Fidelity ↑**	**Utility ↑**	**Coverage ↑**	**Privacy ↓**
Philadelphia	ω_1	19%	0.267 ± 0.03	0.263 ± 0.04	0.689 ± 0.03	3.162 ± 0.19
	ω_2	13%	0.264 ± 0.03	0.325 ± 0.06	0.681 ± 0.02	3.103 ± 0.13
	ω_3	8%	0.207 ± 0.03	0.118 ± 0.04	0.847 ± 0.04	3.178 ± 0.03
	ω_4	43%	0.394 ± 0.01	0.430 ± 0.01	0.863 ± 0.02	2.919 ± 0.14
	ω_5	17%	0.238 ± 0.03	0.197 ± 0.03	0.898 ± 0.01	3.359 ± 0.33
	FedTabDiff		**0.590 ± 0.01**	**0.837 ± 0.03**	**0.944 ± 0.03**	**2.607 ± 0.18**
Diabetes	ω_1	10%	0.217 ± 0.01	0.104 ± 0.03	0.944 ± 0.02	10.261 ± 0.25
	ω_2	18%	0.269 ± 0.01	0.186 ± 0.01	0.943 ± 0.03	10.091 ± 0.38
	ω_3	24%	0.314 ± 0.01	0.242 ± 0.01	**0.946 ± 0.01**	9.895 ± 0.31
	ω_4	28%	0.331 ± 0.01	**0.281 ± 0.01**	0.939 ± 0.01	9.941 ± 0.21
	ω_5	18%	0.269 ± 0.01	0.185 ± 0.01	0.943 ± 0.02	10.139 ± 0.19
	FedTabDiff		**0.720 ± 0.01**	0.265 ± 0.01	0.906 ± 0.01	**3.120 ± 0.09**

*Scores are derived from the averaged results and standard deviations of five experiments, each initiated with distinct random seeds

Fidelity. The *FedTabDiff* model demonstrates superior fidelity, outperforming non-federated models by up to 56% in the Philadelphia dataset and by approximately 118% in the Diabetes dataset. The model's adaptability across varying data distributions is underscored by examining fidelity across different data subsets, D_i, as presented in fig. 2a. The *FedTabDiff* showcases stable fidelity scores across all clients indicating its suitability for the proposed task. Conversely, non-Federated models achieve analogous scores solely on data subsets exposed to local training, a phenomenon visibly depicted through the emphasized diagonal scores in the associated heatmaps. This outcome aligns with expectations, considering that the model, trained only on the local subset D_i, acquaints itself exclusively with the proper-

ties therein, remaining uninformed about the structural intricacies of the comprehensive (global) dataset D.

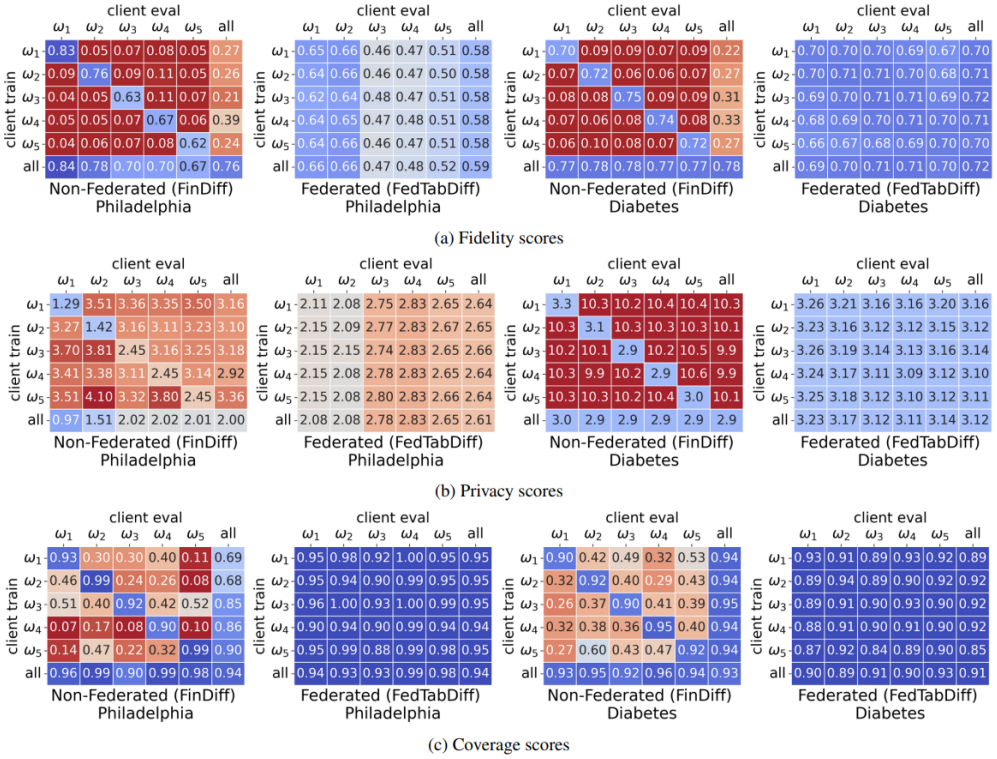


***Figure 2**: Comparison of evaluation metrics between Federated (FedTabDiff) and Non-Federated (FinDiff) diffusion models on individual client ω_i. In the Non-Federated model, each client is trained on its data subset D_i and evaluated across all subsets.*

Utility. The Utility scores from the table 2 indicate a substantial performance enhancement by *FedTabDiff* in the Philadelphia dataset, with an over 218% improvement compared to individual client models. With a score of 0.837, it outperforms all client models, highlighting its superior capacity to synthesize data with broader insights in a federated setting. For the Diabetes dataset, *FedTabDiff* ranks second, slightly behind client ω_4. This variation is attributed to the datasets' partitioning, where the Philadelphia dataset's imbalance is more pronounced, influencing the federated model's relative effectiveness.

Privacy. The privacy metric shows a substantial improvement with *FedTabDiff*, which scores 2.607 for Philadelphia and 3.120 for Diabetes. Although the scores for *FedTabDiff* seem promising, carefully examining the chosen privacy metric is imperative to ensure the validity of the synthetic data generated. Generating samples with substantial deviation from the original might introduce a risk of producing unrealistic data. Simultaneously, generating samples analogous to the original data may result in data replication, potentially breaching privacy safeguards. A balanced strategy, which moderately navigates between these two extremes, may offer a viable path forward.

Coverage. In the Diabetes dataset, *FedTabDiff* scores 0.906, slightly below the highest non-federated score of 0.946, while leading in Philadelphia with 0.944. These results highlight its performance and bring nuances in coverage metrics to light. For instance,

uniform random generation of categorical entries might inflate scores, as seen in Diabetes dataset results in table 2 and the last column of the non-Federated heatmaps in fig. 2c. Conversely, while the *FedTabDiff* gives attention to distribution shape, there is a chance to overlook minority entries, especially within skewed distributions.

Overall, the *FedTabDiff* model consistently surpasses non-federated models across all evaluated metrics. This superior performance of the federated learning model demonstrates its effective learning capabilities from the combined datasets of all clients. Consequently, it affirms that the federated learning setup allows diffusion models to effectively learn the underlying data structures across all clients, thereby generating data of higher quality.

6. Conclusion

This study introduces *FedTabDiff*, a novel federated diffusion-based generative model for the high-fidelity synthesis of mixed-type tabular data, enabling collaborative model training without sharing sensitive data. The model enhances performance metrics across two diverse datasets and maintains data privacy in a federated setting. The methodological strength of *FedTabDiff* is thoroughly examined through extensive experiments rooted in a realistic, non-iid data environment. The model's ability to prevent sharing sensitive information during the distributed training of generative models, specifically for synthesizing tabular data for downstream tasks, is apparent. Finally, through this federated approach, the positions of underrepresented entities with limited data are substantially enhanced, thereby promoting the practice of responsible AI in the financial sector. This phenomenon is particularly profound in financial regulation, where the official statistics are distributed across central banks in an unbalanced and non-iid manner. Models of this nature enable central banks to overcome these challenges by adopting collaborative training of generative models. Developed approach facilitates the sharing of models themselves rather than the direct sharing of sensitive data. Future directions for this preliminary research include more advanced federated learning methods and diffusion models.

References

Aledhari, M.; Razzak, R.; Parizi, R. M.; and Saeed, F. 2020. Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications. IEEE Access, 8: 140699–140725.

Assefa, S. A.; Dervovic, D.; Mahfouz, M.; Tillman, R. E.; Reddy, P.; and Veloso, M. 2020. Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls. In Pro-ceedings of the First ACM International Conference on AI in Finance, 1–8.

Bai, H.; Zheng, R.; Chen, J.; Ma, M.; Li, X.; and Huang, L. 2022. A3T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing. In International Confer-ence on Machine Learning, 1399–1411. PMLR.

Barse, E.; Kvarnstrom, H.; and Jonsson, E. 2003. Syn-thesizing test data for fraud detection systems. In 19th Annual Computer Security Applications Conference, 2003. Proceedings., 384–394.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Par-rots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. In Noise reduction in speech processing, 37–40. Springer.

Beutel, D. J.; Topal, T.; Mathur, A.; Qiu, X.; Fernandez-Marques, J.; Gao, Y.; Sani, L.; Li, K. H.; Parcollet, T.; de Gusmäo, P. P. B.; and Lane, N. D. 2022. Flower: A Friendly Federated Learning Research Framework. arXiv:2007.14390.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In International Conference on Learning Representations.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Cao, H.; Tan, C.; Gao, Z.; Xu, Y.; Chen, G.; Heng, P.-A.; and Li, S. Z. 2022. A Survey on Generative Diffusion Model. arXiv preprint arXiv:2209.02646.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2020. Extracting Training Data from Large Language Models. arXiv preprint arXiv:2012.07805.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody Dance Now. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 5933–5942.

Charitou, C.; Dragicevic, S.; and d'Avila Garcez, A. 2021. Synthetic Data Generation for Fraud Detection using GANs. arXiv:2109.12546.

Chen, B.; Zhang, F.; Nguyen, A.; Zan, D.; Lin, Z.; Lou, J.-G.; and Chen, W. 2022. CodeT: Code Generation with Gen-erated Tests. arXiv preprint arXiv:2207.10397.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion Models in Vision: A Survey. IEEE Trans-actions on Pattern Analysis and Machine Intelligence.

de Goede, M. 2023. Training Diffusion Models with Federated Learning: A Communication-Efficient Model for Cross-Silo Federated Image Generation.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems, 34: 8780–8794.

Engelmann, J.; and Lessmann, S. 2021. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications, 174: 114582.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 1322–1333.

Gao, S.; Zhou, P.; Cheng, M.-M.; and Yan, S. 2023. Masked Diffusion Transformer is a Strong Image Synthesizer. arXiv preprint arXiv:2303.14389.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. arXiv preprint arXiv:1412.6572.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. Advances in Neural Information Pro-cessing Systems, 33: 6840–6851.

Hoogeboom, E.; Nielsen, D.; Jaini, P.; Forre´, P.; and Welling, M. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. Advances in Neural Information Processing Systems, 34: 12454–12465.

Jordon, J.; Yoon, J.; and Van Der Schaar, M. 2018. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In International conference on learning representations.

Jothiraj, F. V. S.; and Mashhadi, A. 2023. Phoenix: A Federated Generative Diffusion Model. arXiv preprint arXiv:2306.04098.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2019. Advances and Open Problems in Federated Learning. arXiv preprint arXiv:1912.04977.

Kim, J.; Jeon, J.; Lee, J.; Hyeong, J.; and Park, N. 2021. OCT-GAN: Neural ODE-Based Conditional Tabular GANs. In Proceedings of the Web Conference 2021, 1506–1515.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; and Babenko, A. 2022. TabDDPM: Modelling Tabular Data with Diffusion Models. arXiv:2209.15421.

Le, H.; Wang, Y.; Gotmare, A. D.; Savarese, S.; and Hoi, S. C. H. 2022. CoderL: Mastering Code Generation Through Pretrained Models and Deep Reinforcement Learning. Advances in Neural Information Processing Systems, 35: 21314–21328.

Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Manohar, V.; Adi, Y.; Mahadeokar, J.; et al. 2023. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. arXiv preprint arXiv:2306.15687.

Li, Q.; Wen, Z.; Wu, Z.; Hu, S.; Wang, N.; Li, Y.; Liu, X.; and He, B. 2021. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering.

Li, Y.; Choi, D.; Chung, J.; Kushman, N.; Schrittwieser, J.; Leblond, R.; Eccles, T.; Keeling, J.; Gimeno, F.; Dal Lago, A.; et al. 2022. Competition-Level Code Generation with AlphaCode. Science, 378(6624): 1092–1097.

Massey, F. J. 1951. The Kolmogorov-Smirnov test for good-ness of fit. Journal of the American Statistical Association, 46(253): 68–78.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Artificial Intelligence and Statistics, 1273–1282. PMLR.

McMahan, B.; and Ramage, D. 2017. Federated Learning: Collaborative Machine Learning Without Centralized Training Data. Google Research Blog, 3.

Nock, R.; and Guillame-Bert, M. 2022. Generative Trees: Adversarial and Copycat. arXiv preprint arXiv:2201.11205.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Paszke, A.; and Gross, S. e. a. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d' Alche´Buc, F.; Fox, E.; and Garnett, R., eds., Advances in Neural Information Processing Systems 32, 8024–8035.

Patki, N.; Wedge, R.; and Veeramachaneni, K. 2016. The Synthetic data vault. In IEEE International Conference on Data Science and Advanced Analytics (DSAA), 399–410.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In IEEE/CVF conference on computer vision and pattern recognition, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 234–241. Springer.

Salem, A.; Zhang, Y.; Humbert, M.; Berrang, P.; Fritz, M.; and Backes, M. 2018. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Ma-chine Learning Models. arXiv preprint arXiv:1806.01246.

Sattarov, T.; Schreyer, M.; and Borth, D. 2023. FinDiff: Diffusion Models for Financial Tabular Data Generation. arXiv:2309.01472.

Schreyer, M.; Hemati, H.; Borth, D.; and Vasarhelyi, M. A. 2022. Federated Continual Learning to Detect Accounting Anomalies in Financial Auditing. In Workshop on Federated Learning: Recent Advances and New Challenges (in Con-junction with NeurIPS 2022).

Schreyer, M.; Sattarov, T.; and Borth, D. 2022. Federated and Privacy-Preserving Learning of Accounting Data in Financial Statement Audits. In Proceedings of the Third ACM International Conference on AI in Finance, 105–113.

Schreyer, M.; Sattarov, T.; Reimer, B.; and Borth, D. 2019. Adversarial Learning of Deepfakes in Accounting. NeurIPS'19 Workshop on Robust AI in Financial Services.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership Inference Attacks Against Machine Learning Models. In 2017 IEEE Symposium on Security and Privacy (SP), 3–18. IEEE.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-Video: Text-to-Video Generation Without Text-Video Data. arXiv preprint arXiv:2209.14792.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In International confer-ence on machine learning, 2256–2265. PMLR.

Torfi, A.; Fox, E. A.; and Reddy, C. K. 2022. Differentially Private Synthetic Medical Data Generation Using Convolutional GANs. Information Sciences, 586: 485–500.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288.

Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv preprint arXiv:2301.02111.

Wen, B.; Cao, Y.; Yang, F.; Subbalakshmi, K.; and Chandramouli, R. 2022. Causal-TGAN: Modeling Tabular Data Using Causally-Aware GAN. In ICLR Workshop on Deep Generative Models for Highly Structured Data.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling tabular data using conditional gan. NeurIPS, 32.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Shao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2022. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv preprint arXiv:2209.00796.

Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A. G.; Yang, M.-H.; Hao, Y.; Essa, I.; et al. 2023. Magvit: Masked Generative Video Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10459–10469.

Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; and Gao, Y. 2021a. A Survey on Federated Learning. Knowledge-Based Systems, 216: 106775.

Zhang, Y.; Zaidi, N. A.; Zhou, J.; and Li, G. 2021b. GAN-BLR: A Tabular Data Generation Model. In International Conference on Data Mining (ICDM), 181–190. IEEE.

Zhao, Z.; Kunar, A.; Birke, R.; and Chen, L. Y. 2021. CTAB-GAN: Effective Table Data Synthesizing. In Asian Confer-ence on Machine Learning, 97–112. PMLR.

DEUTSCHE BUNDESBANK EUROSYSTEM

University of St.Gallen
Institute of Computer Science

Overcoming Data-Sharing Challenges in Central Banking: Federated Learning of Diffusion Models for Synthetic Data Generation

Timur Sattarov, Marco Schreyer

3rd IFC Workshop on Data Science in Central Banking
„Data Sharing and Data Access“
17-19 October 2023, Rome, Bank of Italy

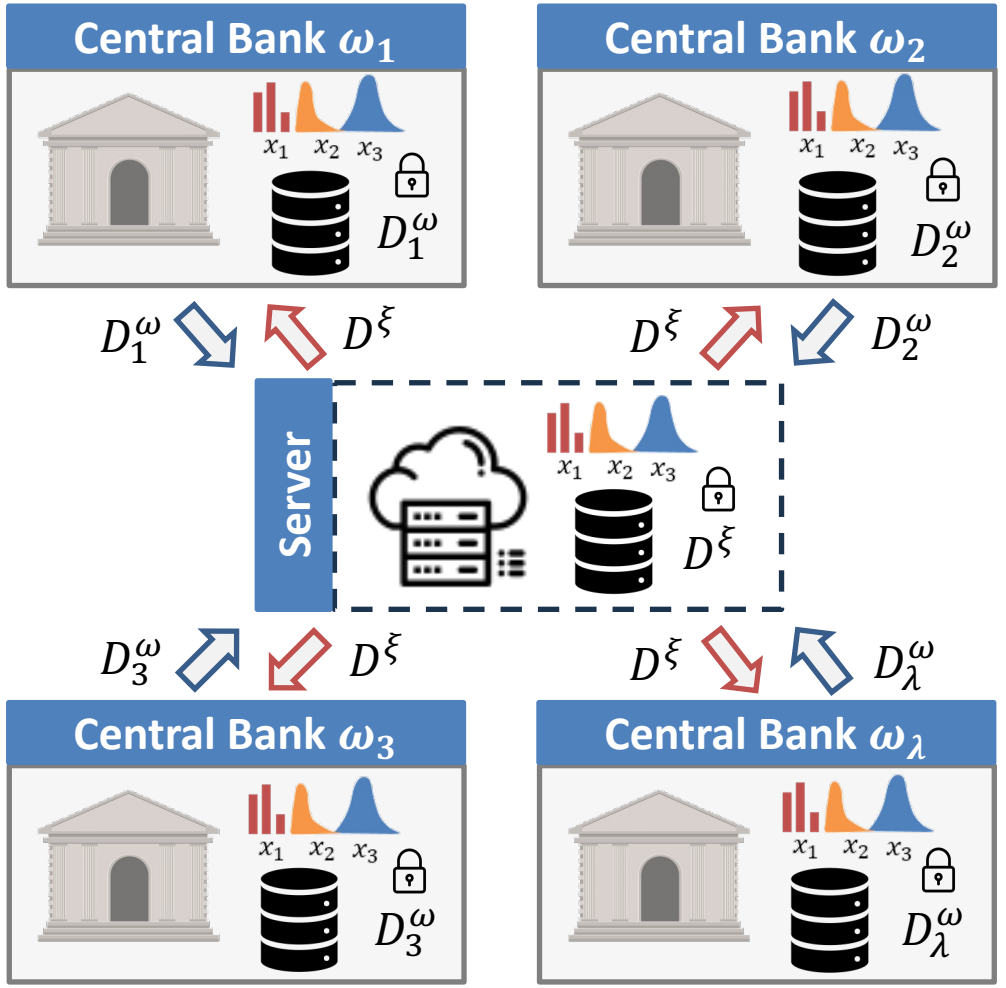
Agenda

- Motivation
- Federated Learning
- Diffusion Models
- FedTabDiff
- Experimental Results
- Conclusion

Motivation

The sharing of financial data between central banks is crucial for managing economic policy and financial sector supervision.
Key advantages:

- **Improved economic policy**: Informed responses to global trends and risks.
- **Boosted financial stability**: Identifies risks and strengthens systems proactively.
- **Better cross-border regulation**: Enables consistent standards and detects misconduct.



Motivation

Challenges:

- **Legal and Regulator Constraints**: varying laws and data protection regulations in different countries can be cumbersome for data sharing.
- **Data Privacy and Security Concerns**: any breaches of confidential information can have severe consequences.
- **Data Transmission**: moving extensive datasets between central banks can become bandwidth-intensive and expensive.

Motivation

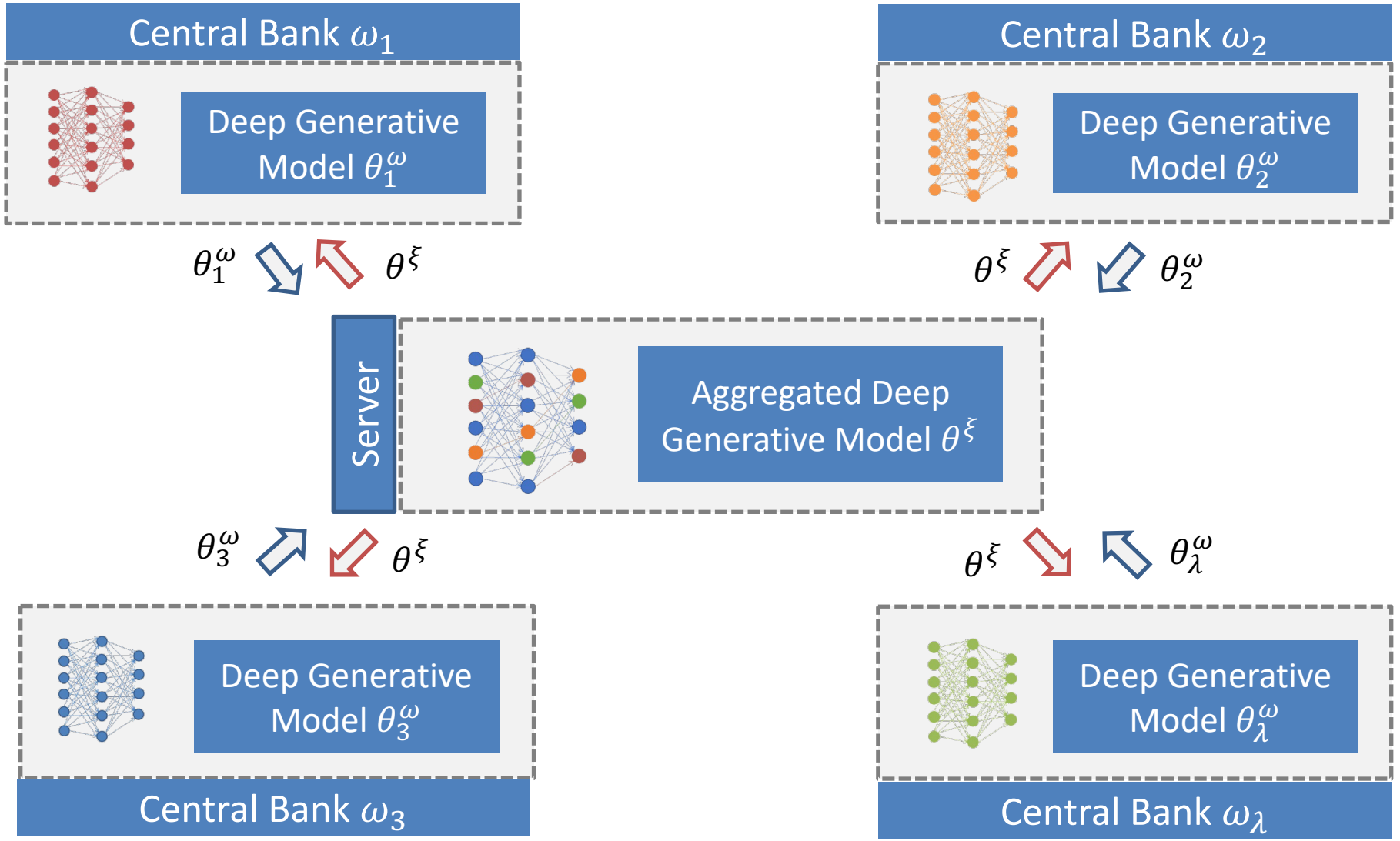
Challenges:

- **Legal and Regulator Constraints**: varying laws and data protection regulations in different countries can be cumbersome for data sharing.
- **Data Privacy and Security Concerns**: any breaches of confidential information can have severe consequences.
- **Data Transmission**: moving extensive datasets between central banks can become bandwidth-intensive and expensive.

Idea: “Federated Learning + Diffusion Models”

- Use **Federated Learning** for decentralized node training without data exchange.
- Utilize **Diffusion Models** to synthesize central bank’s local data; then share the trained synthesizer model as part of the federated training.
- The **global model aggregates** knowledge from central banks to produce high-quality synthetic data.

Data Sharing with Federated Learning



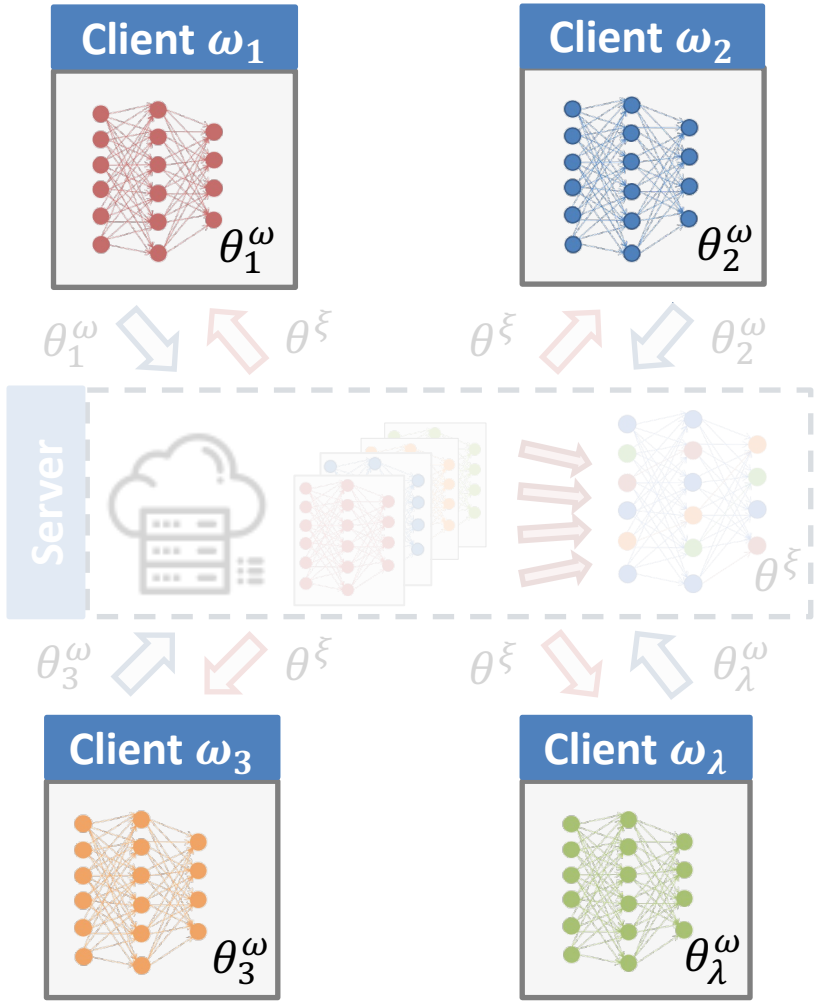
Agenda

- Motivation
- **Federated Learning**
- Diffusion Models
- FedTabDiff
- Experimental Results
- Conclusion

Federated Learning: the mechanics

Main ingredients:

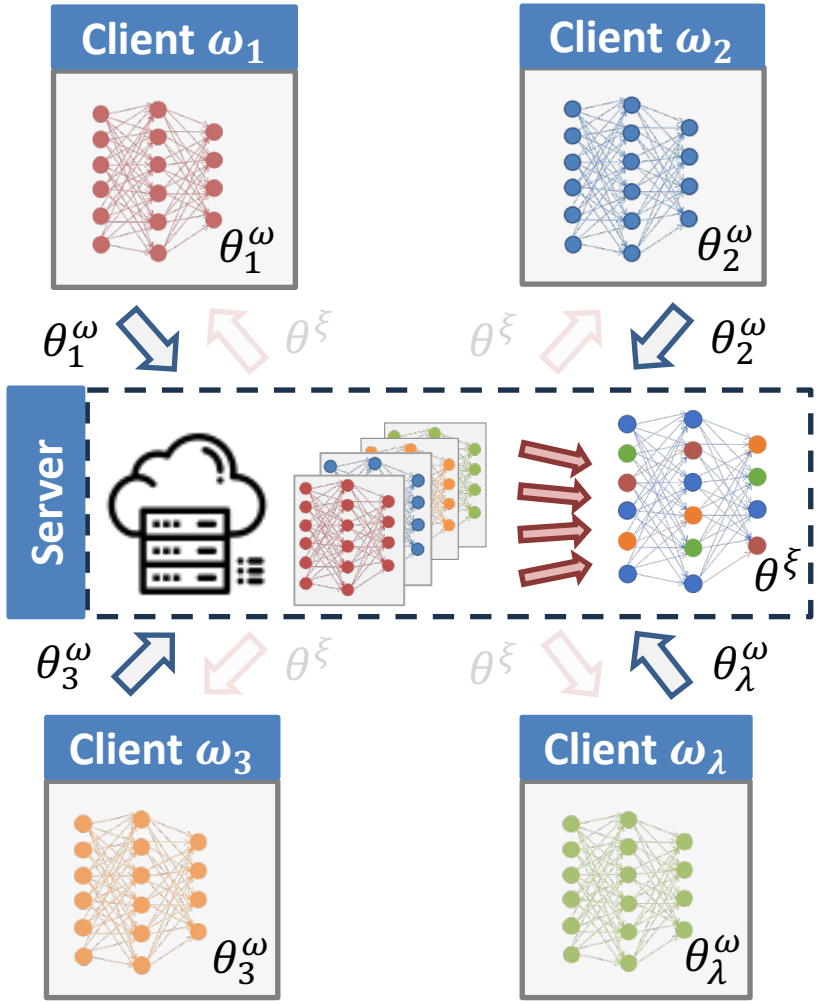
1. **Training on the Client**: the initial model θ_i^{ω} is trained locally on each client ω_i using the local data D_i, ensuring data privacy and security.



Federated Learning: the mechanics

Main ingredients:

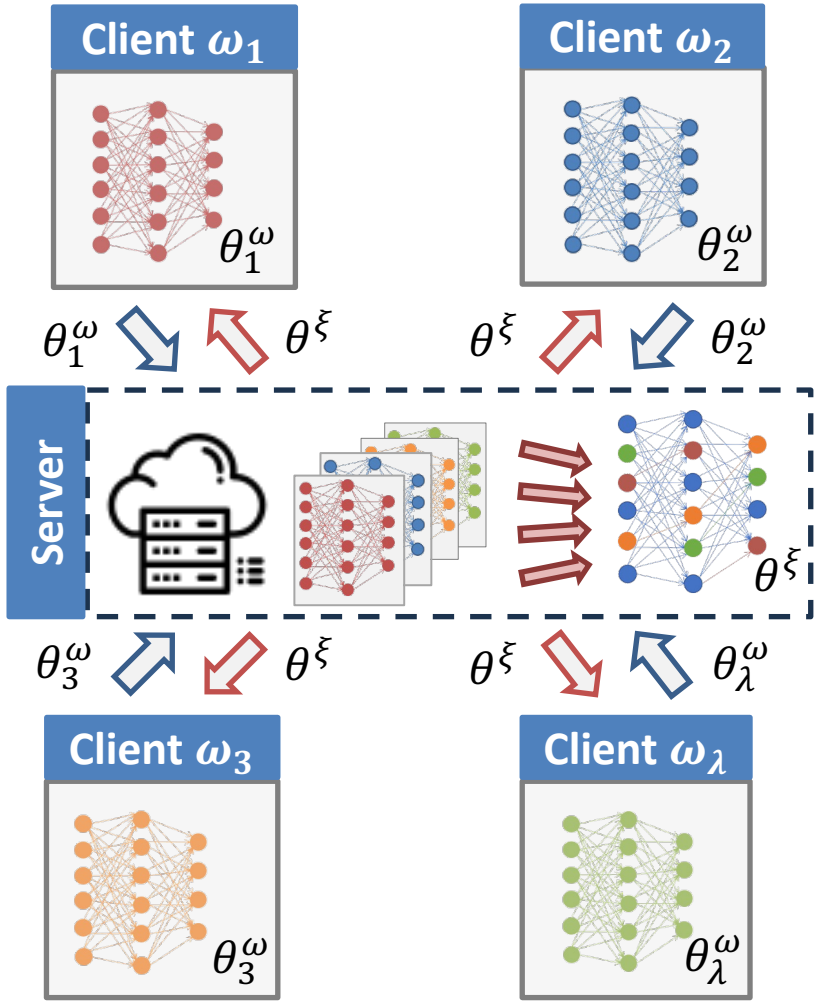
1. **Training on the Client**: the initial model θ_i^{ω} is trained locally on each client ω_i using the local data D_i, ensuring data privacy and security.
2. **Model Aggregation at the Server**: after each communication round $r = 1, \dots, R$ only the model parameters are sent to a centralized server, where they are aggregated into a "global" model θ^{ξ}.



Federated Learning: the mechanics

Main ingredients:

1. **Training on the Client**: the initial model θ_i^{ω} is trained locally on each client ω_i using the local data D_i, ensuring data privacy and security.
2. **Model Aggregation at the Server**: after each communication round $r = 1, \dots, R$ only the model parameters are sent to a centralized server, where they are aggregated into a "global" model θ^{ξ}.
3. **Continuous Learning Cycle**: The updated global model is sent back to the clients for further local training, creating a continuous cycle of learning and improvement.



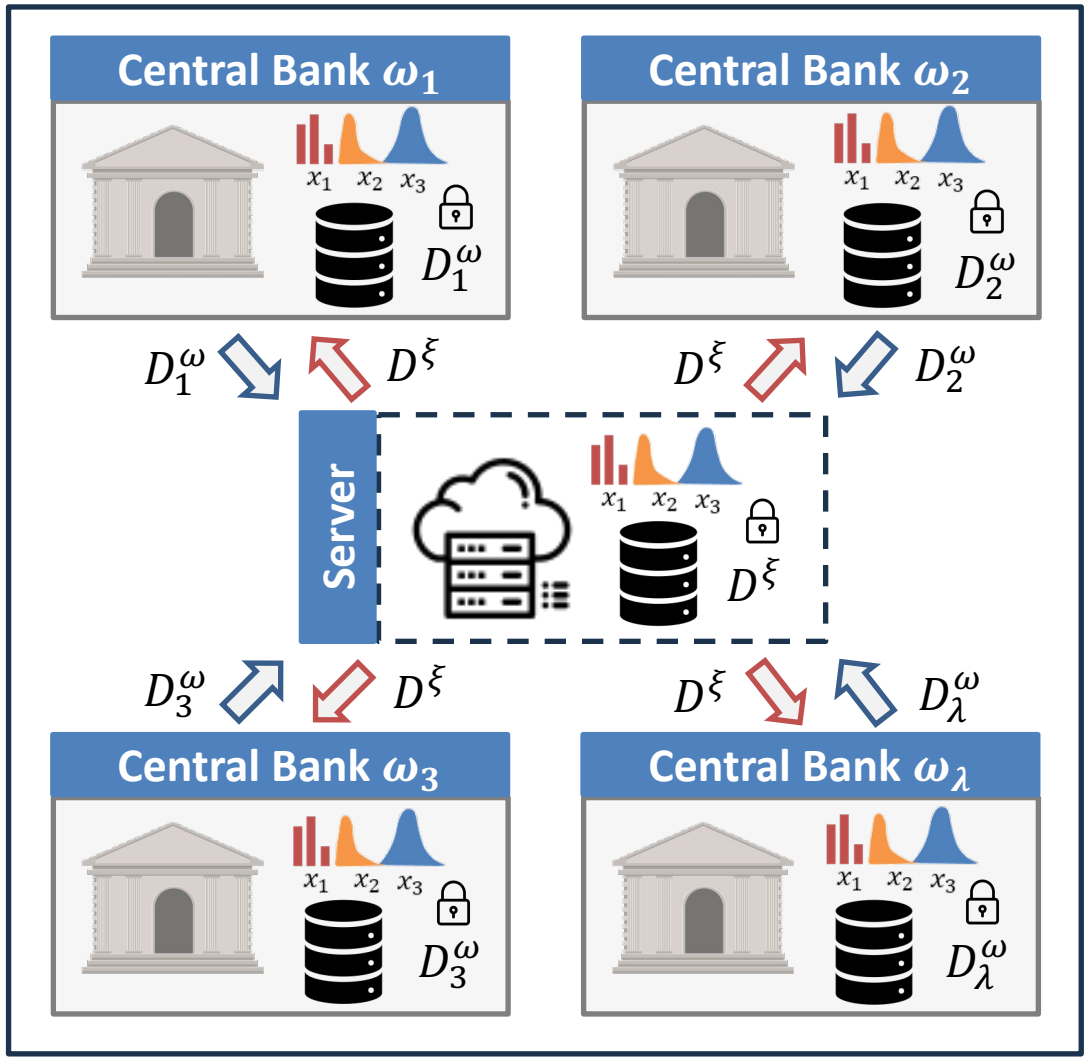
Federated Learning: Benefits

Privacy Preserving: Federated Learning enhances user privacy by keeping all the sensitive data on the local device, never sending raw data to the central server.

Reduced Data Transfer Costs: in the Federated Learning setup only the model parameters are being exchanged therefore avoiding transmission of large data volumes.

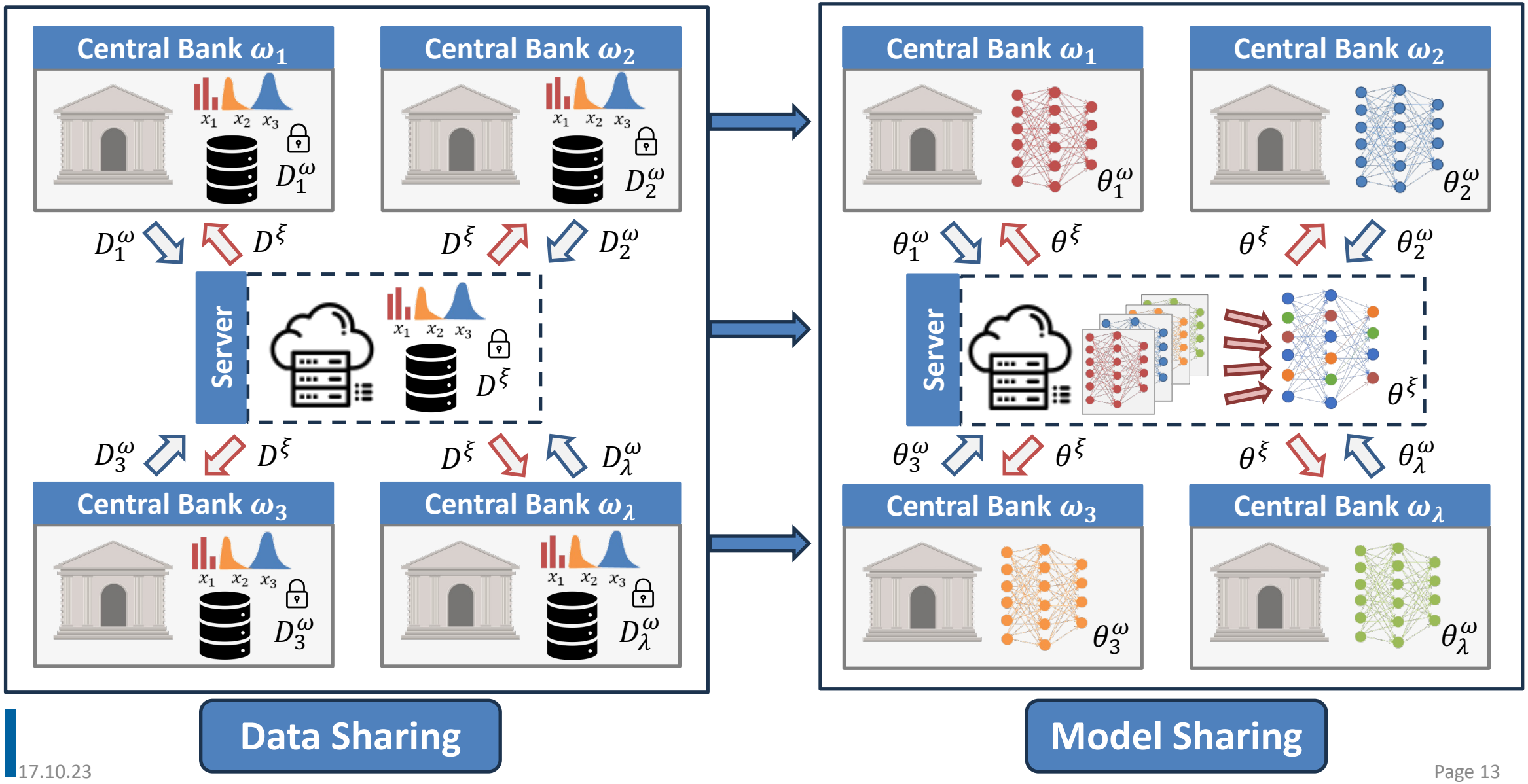
Global Insights: financial institutions can benefit from insights gathered globally across different markets and segments, but applied in a way that is tailored to local market conditions and regulations.

From Data Sharing to Model Sharing

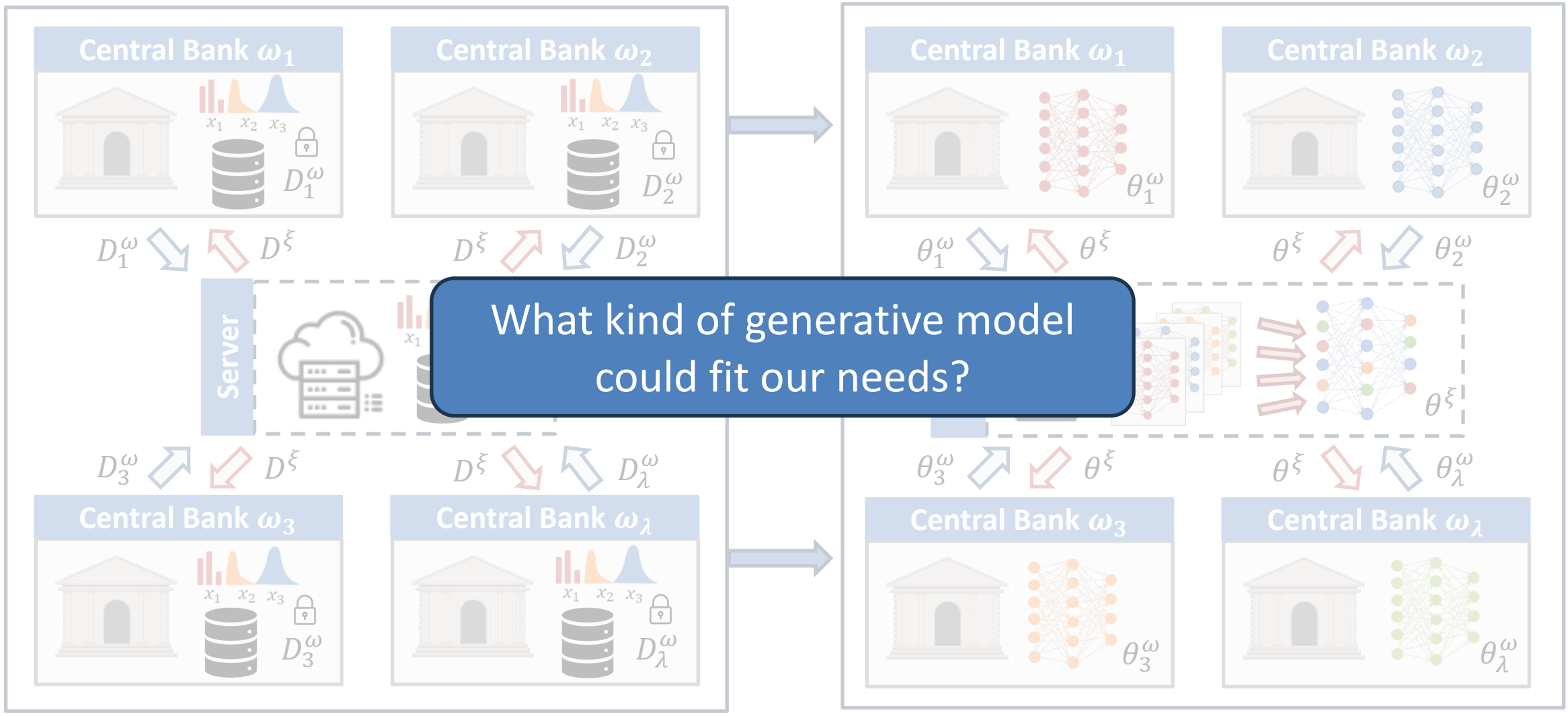


Data Sharing

From Data Sharing to Model Sharing



From Data Sharing to Model Sharing



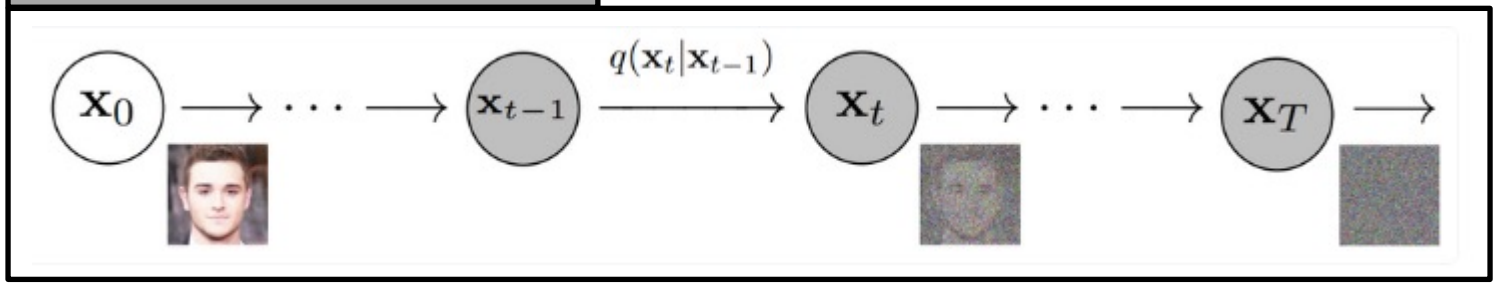
Agenda

- Motivation
- Federated Learning
- **Diffusion Models**
- FedTabDiff
- Experimental Results
- Conclusion

Diffusion Models

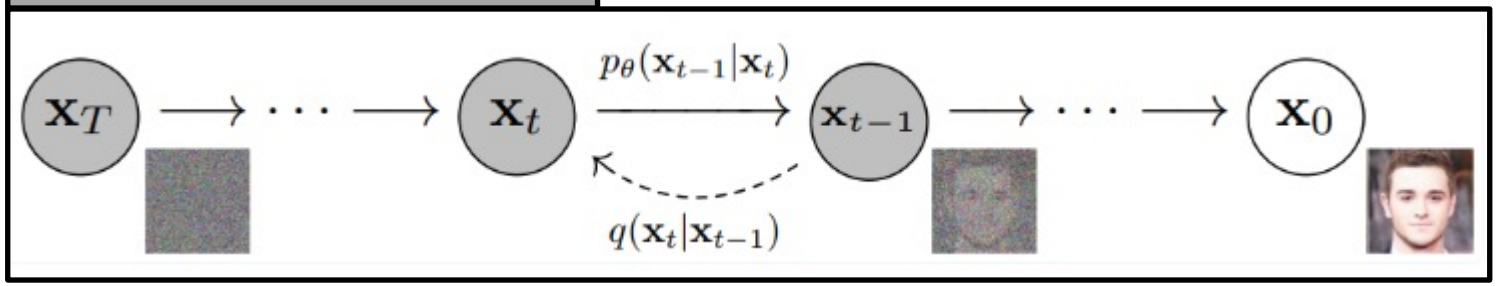
- Diffusion Models – are **generative models** trained with the objective of noise removal and subsequently constructing the desired **data samples from pure noise**.
- First introduced by Jascha Sohl-Dickstein et al. in 2015 with motivation from non-equilibrium thermodynamics. They can be thought as a **sequence of denoising autoencoders**.

Forward Diffusion Process



- Markov chain of diffusion steps.
- Add Gaussian noise in *T* steps.
- When $T \to \infty$, x_t is equivalent to an isotropic Gaussian.
- No learning at this step.

Reverse Diffusion Process



- Reverse process of *T* steps.
- Data generation from Isotropic Gaussian noise.
- Usually is called sampling.
- Learning is required.

Figure is taken from „Denoising Diffusion Probabilistic Models “ by Ho et al. https://arxiv.org/pdf/2006.11239.pdf

Diffusion Models for Financial Tabular Data

FinDiff: Diffusion Models for Financial Tabular Data Generation

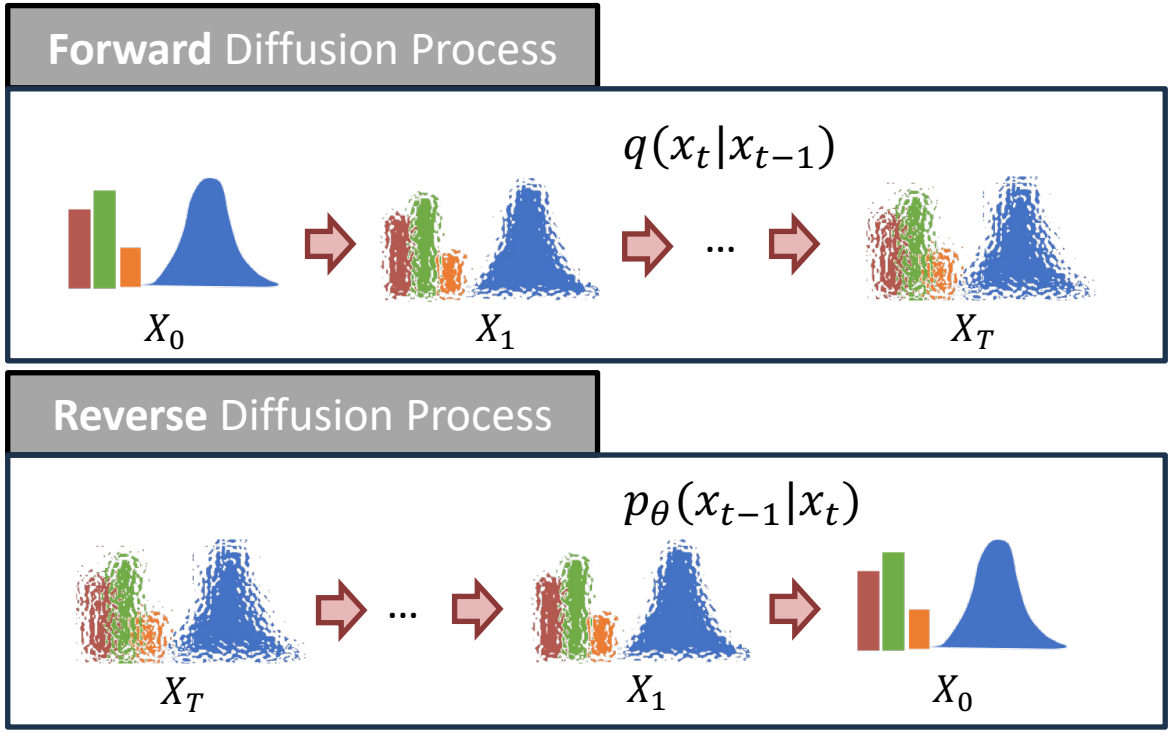
Timur Sattarov[1,2]
timur.sattarov@bundesbank.de

Marco Schreyer[1]
marco.schreyer@unisg.ch

Damian Borth[1]
damian.borth@unisg.ch

[1]University of St. Gallen, St. Gallen, Switzerland

[2]Deutsche Bundesbank, Frankfurt am Main, Germany



https://arxiv.org/abs/2309.01472

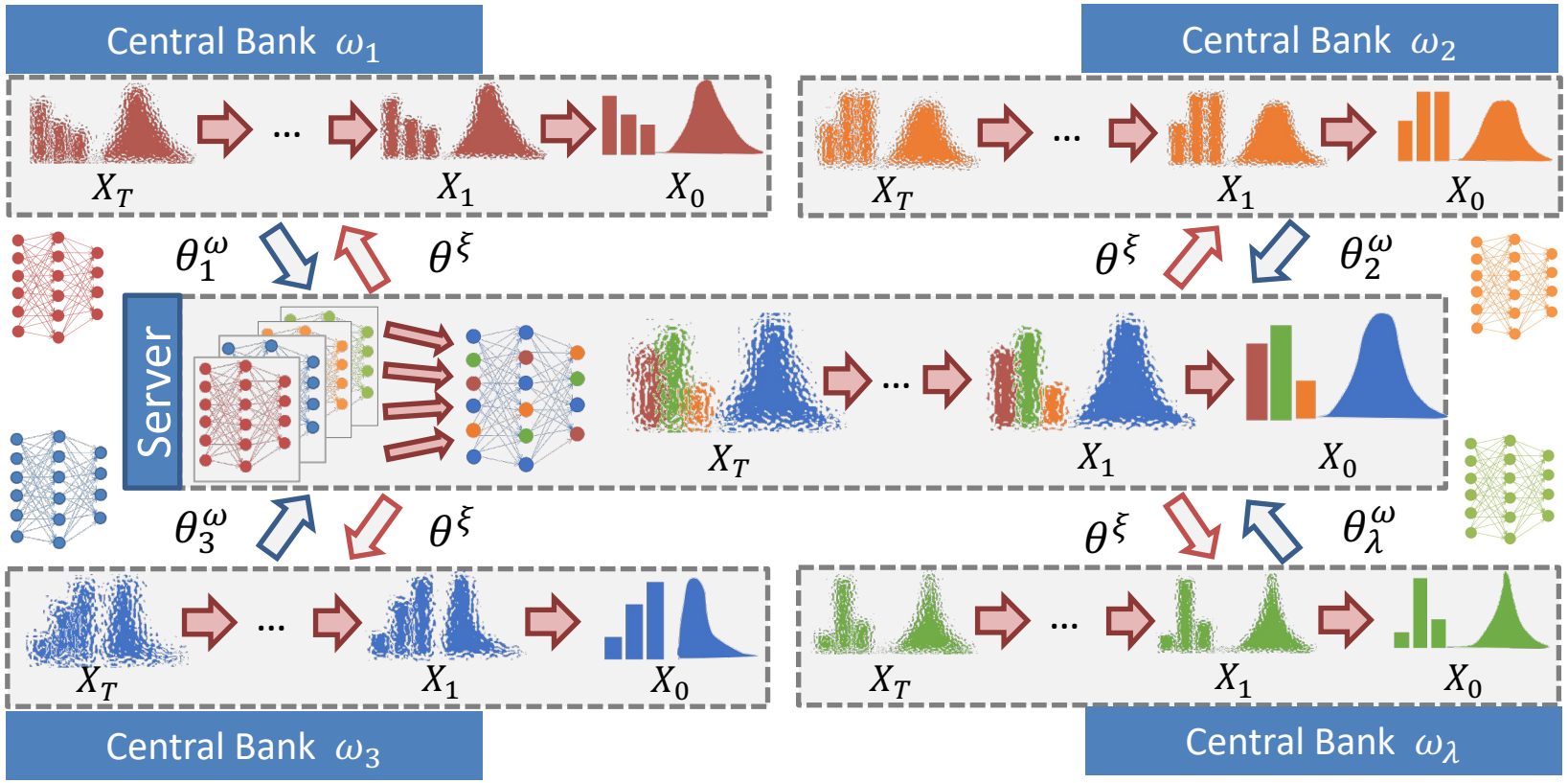
- FinDiff is a diffusion based generative model, that synthesizes financial tabular data for **regulatory downstream tasks**.
- It uses **embeddings** for mixed modality financial data, comprising both **categorical and numeric** attributes.
- Empirical results demonstrate **high fidelity, privacy, and utility** using FinDiff.

Agenda

- Motivation
- Federated Learning
- Diffusion Models
- **FedTabDiff**
- Experimental Results
- Conclusion

FedTabDiff schematic overview

1. Each central bank trains a **local generative model θ_i^{ω}.**
2. The Server aggregates all models into a **global generative model θ^{ξ}** accumulating knowledge from local datasets D_i^{ω}.
3. The global model θ^{ξ} is used to **generate high quality tabular data** without sharing the actual data.



Experimental Setup

- Mixed-type tabular datasets:
  City of Philadelphia Payments – 215,302 payments generated by 58 city departments in 2017. Contains 10 categorical and 1 numeric attributes.
  Diabetes Hospital Data – 92,689 clinical care records collected by 130 US hospitals between 1999-2008. Each record has 40 categorical and 8 numeric attributes.
- The dataset was **non-iid partitioned** $D_i \subseteq D$ across 5 clients ω_i.
- Evaluation **metrics**: fidelity, utility, privacy, and coverage.
- **Federated learning hyperparameters**
  total communication rounds: $R = 1000$
  client model θ_i^{ω} updates: $R_c = 20$
- **Diffusion model hyperparameters:**
  MLP layers: 1024 -> 1024 -> 1024 -> 1024
  activation: leakyRelu
  total diffusion steps: $T = 500$

Dataset	Client	# samples $\mathcal{D}_i$
Philadelphia	ω_1	40,038
	ω_2	28,521
	ω_3	16,831
	ω_4	93,119
	ω_5	36,793
	all	215,302
Diabetes	ω_1	9,685
	ω_2	17,256
	ω_3	22,483
	ω_4	26,068
	ω_5	17,197
	all	92,689

Experimental Results

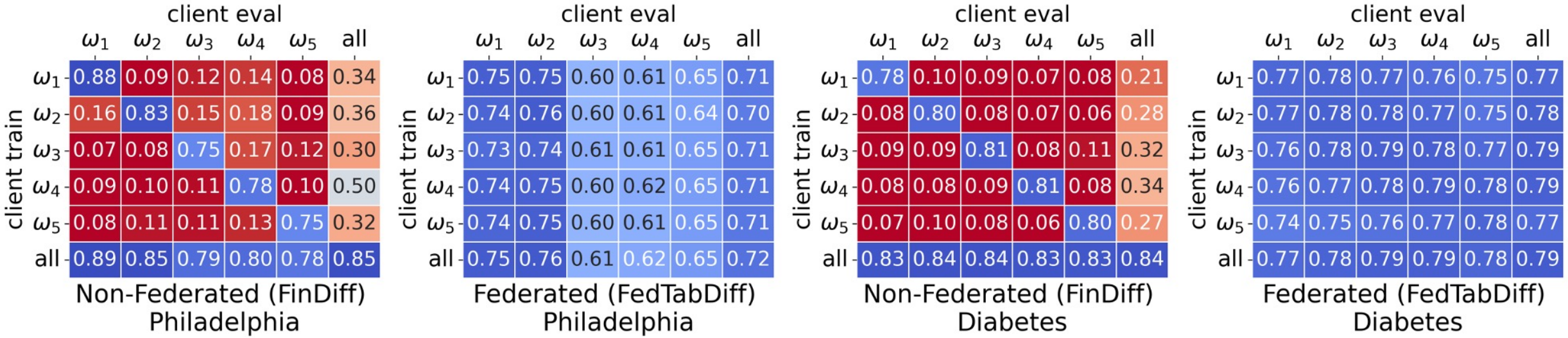
- Comparative analysis of the **Federated** (FedTabDiff) versus **Non-Federated** (FinDiff) models, evaluated using the full dataset D.
- Non-Federated diffusion models are trained individually at each client ω_i with subset $D_i \subseteq D$ (column "Split") and evaluated against the entire dataset D.

			Evaluation Measures			
Dataset	**Client**	**Split $\mathcal{D}_i$**	**Fidelity [8, 32] ↑**	**Utility [50] ↑**	**Coverage [7] ↑**	**Privacy [53] ↓**
Philadelphia	ω_1	19%	0.267 ± 0.03	0.263 ± 0.04	0.689 ± 0.03	3.162 ± 0.19
	ω_2	13%	0.264 ± 0.03	0.325 ± 0.06	0.681 ± 0.02	3.103 ± 0.13
	ω_3	8%	0.207 ± 0.03	0.118 ± 0.04	0.847 ± 0.04	3.178 ± 0.03
	ω_4	43%	0.394 ± 0.01	0.430 ± 0.01	0.863 ± 0.02	2.919 ± 0.14
	ω_5	17%	0.238 ± 0.03	0.197 ± 0.03	0.898 ± 0.01	3.359 ± 0.33
	FedTabDiff		$\mathbf{0.590 \pm 0.01}$	$\mathbf{0.837 \pm 0.03}$	$\mathbf{0.944 \pm 0.03}$	$\mathbf{2.607 \pm 0.18}$
Diabetes	ω_1	10%	0.217 ± 0.01	0.104 ± 0.03	0.944 ± 0.02	10.261 ± 0.25
	ω_2	18%	0.269 ± 0.01	0.186 ± 0.01	0.943 ± 0.03	10.091 ± 0.38
	ω_3	24%	0.314 ± 0.01	0.242 ± 0.01	$\mathbf{0.946 \pm 0.01}$	9.895 ± 0.31
	ω_4	28%	0.331 ± 0.01	0.281 ± 0.01	0.939 ± 0.01	9.941 ± 0.21
	ω_5	18%	0.269 ± 0.01	0.185 ± 0.01	0.943 ± 0.02	10.139 ± 0.19
	FedTabDiff		$\mathbf{0.720 \pm 0.01}$	$\mathbf{0.265 \pm 0.01}$	0.906 ± 0.01	$\mathbf{3.120 \pm 0.09}$

*Scores are derived from the averaged results and standard deviations of five experiments, each initiated with distinct random seeds

Experimental Results

- Fidelity - **similarity of every column** in the synthetic dataset against the real dataset.
- Fidelity score comparison between **Federated** (FedTabDiff) and **Non-Federated** (FinDiff).
- In the Non-Federated model, each client is trained on its data subset $D_i \subseteq D$ and evaluated across all subsets.



Conclusion and Future Work

- Through the adoption of federated learning methodologies central banks may transition **from data sharing to model sharing.**
- FedTabDiff is a **federated diffusion-based generative model** for high-fidelity synthesis of mixed-type tabular data.
- The model **avoids sharing of sensitive information** by training a generative model and sharing it across different authorities without distributing the underlying data.
- Generated tabular data can be used for a variety of **downstream tasks,** such as regulatory compliance, anti-money laundering, fraud detection, risk management and many others.
- Future trajectories: advancement of **privacy-preserving** techniques, mitigation of **information dissemination** risks and evaluation on the **proprietary regulatory financial statistics**.

Thank you

Timur Sattarov
Data Service Centre
Deutsche Bundesbank
Frankfurt am Main, Germany
timur.sattarov@bundesbank.de

Marco Schreyer
School of Computer Science
University of St.Gallen
St.Gallen, Switzerland
marco.schreyer@unisg.ch