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

Timur Sattarov, Marco Schreyer

DEUTSCHE BUNDESBANK
EUROSYSTEM

University of St.Gallen
Institute of Computer Science

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

Agenda

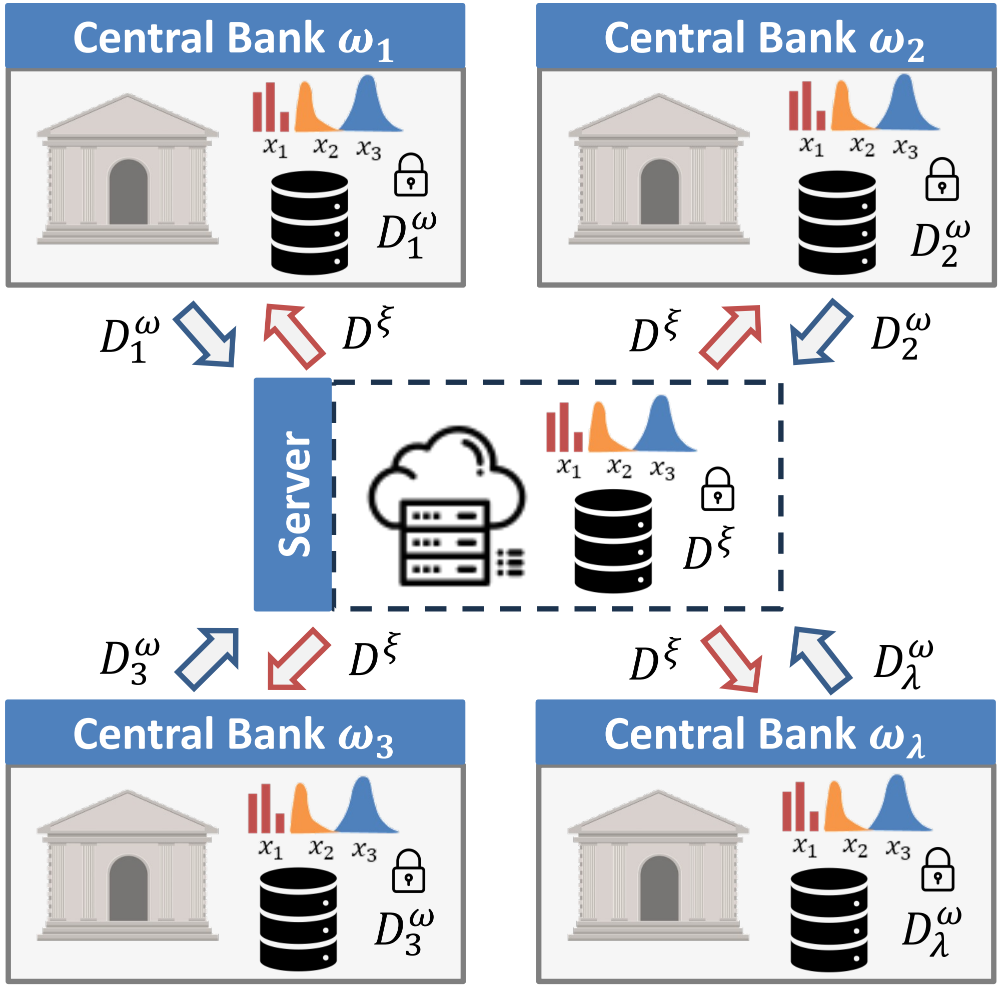
- Motivation
- Federated Learning
- Diffusion Models
- FedTabDiff
- Experimental Results
- Conclusion

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

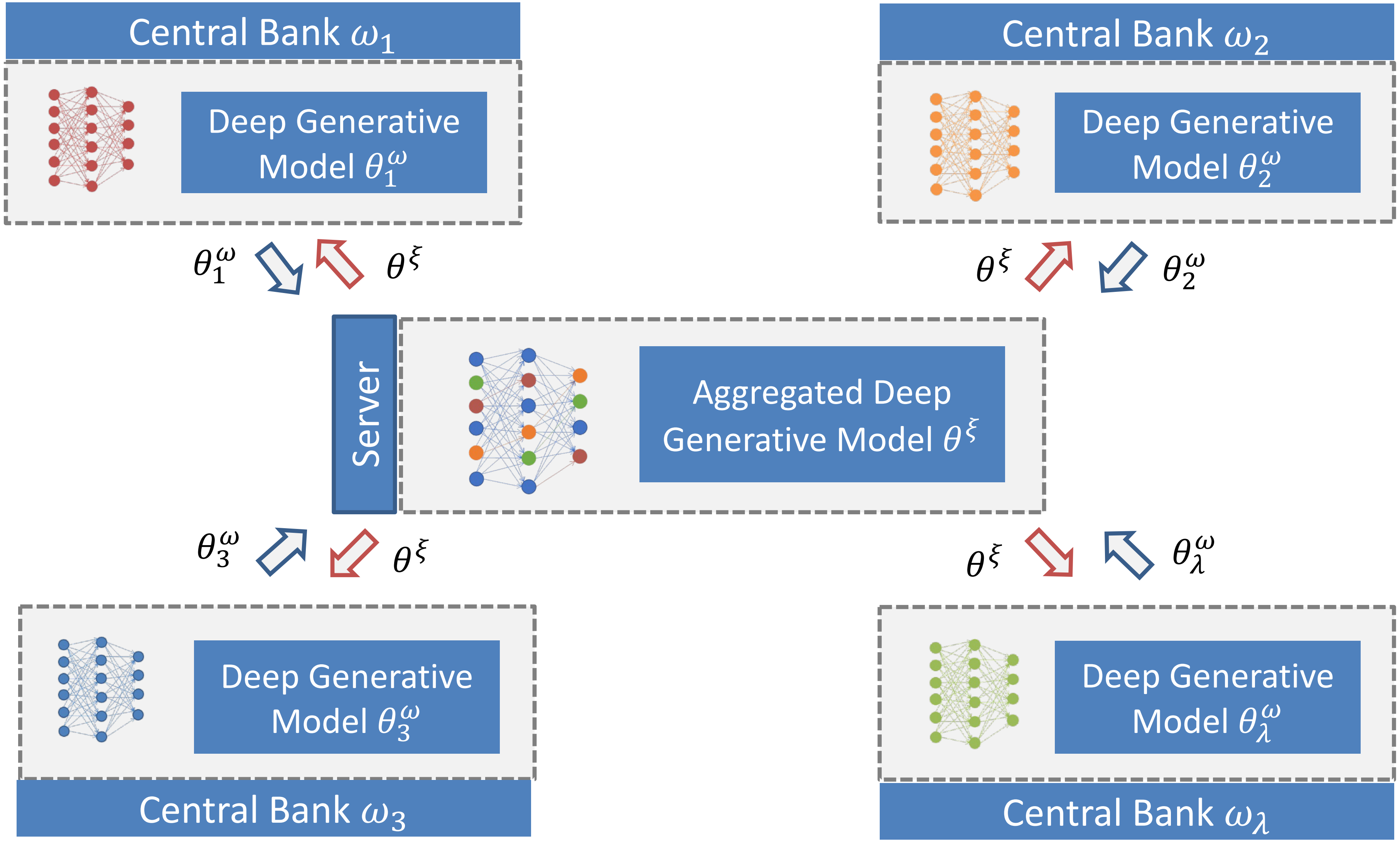
Challenges:

- **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

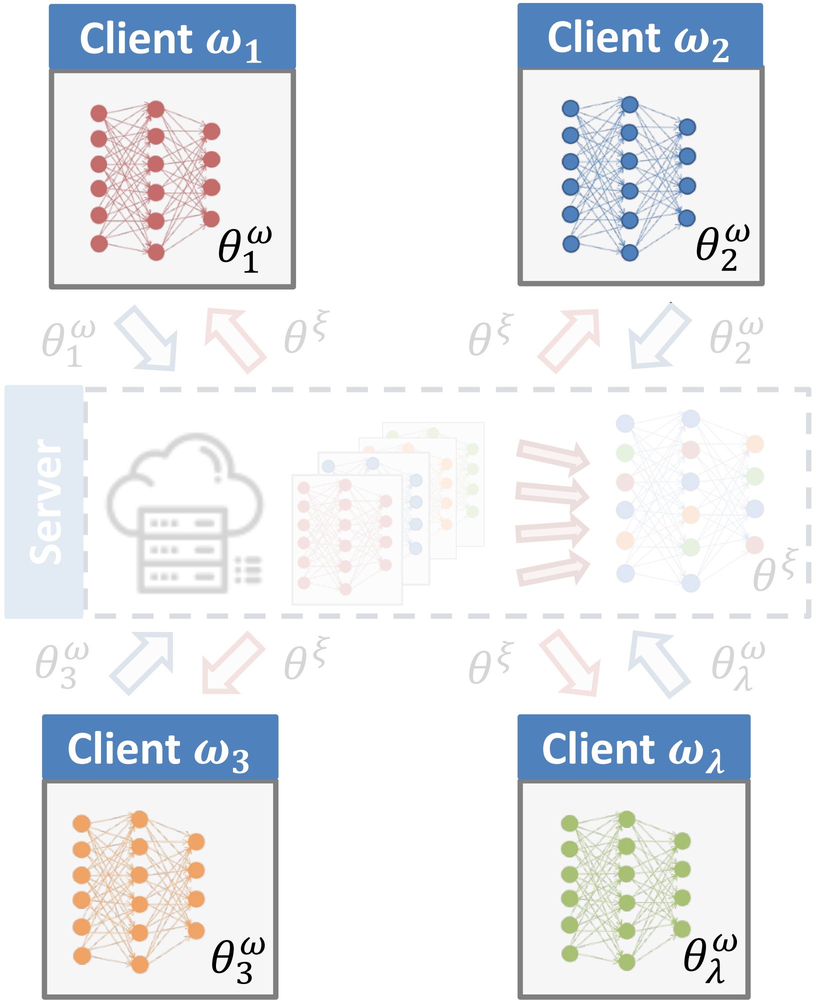


Agenda

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

Main ingredients:

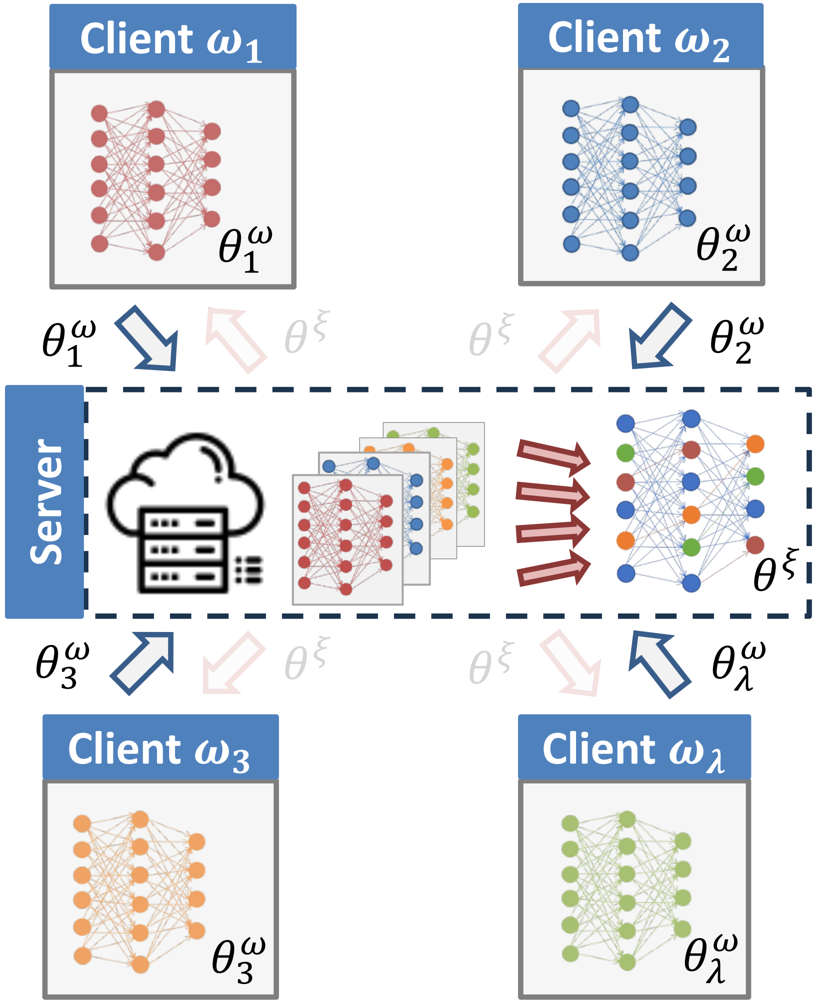
1. **Training on the Client**: the initial model θ_i^ω is trained locally on each client ω_i using the local data D_i, ensuring data privacy and security.



Federated Learning: the mechanics

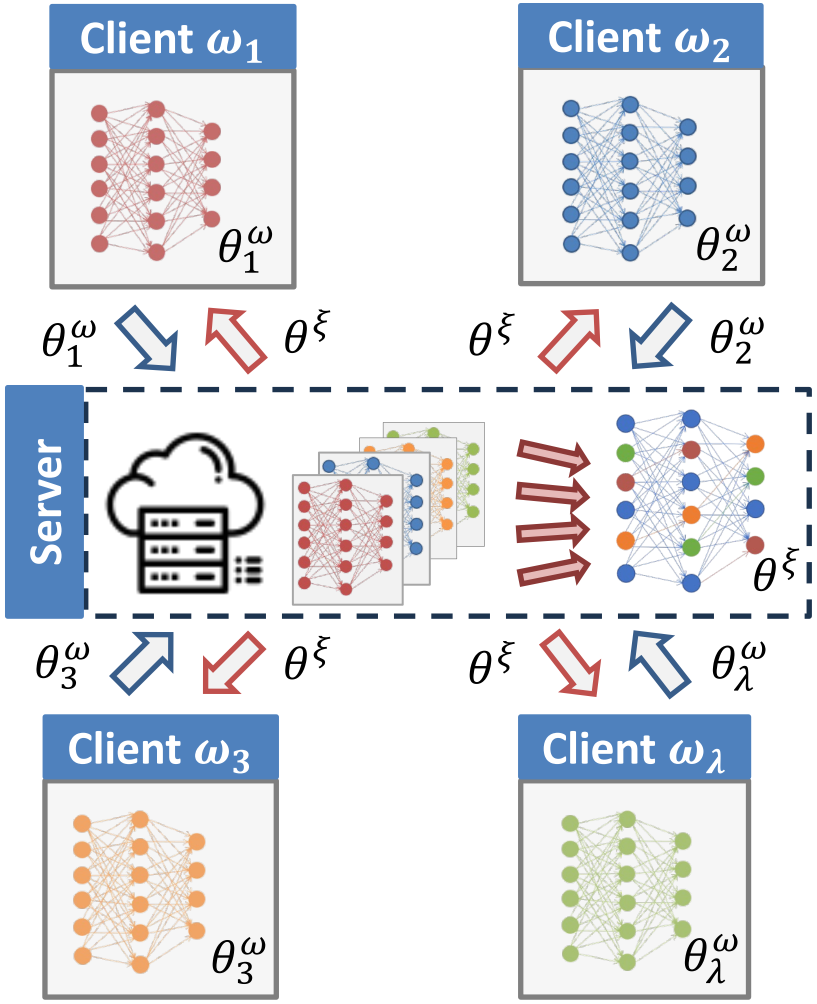
Main ingredients:

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 $r = 1, \ldots, R$ only the model parameters are sent to a centralized server, where they are aggregated into a "global" model θ^ξ.



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.



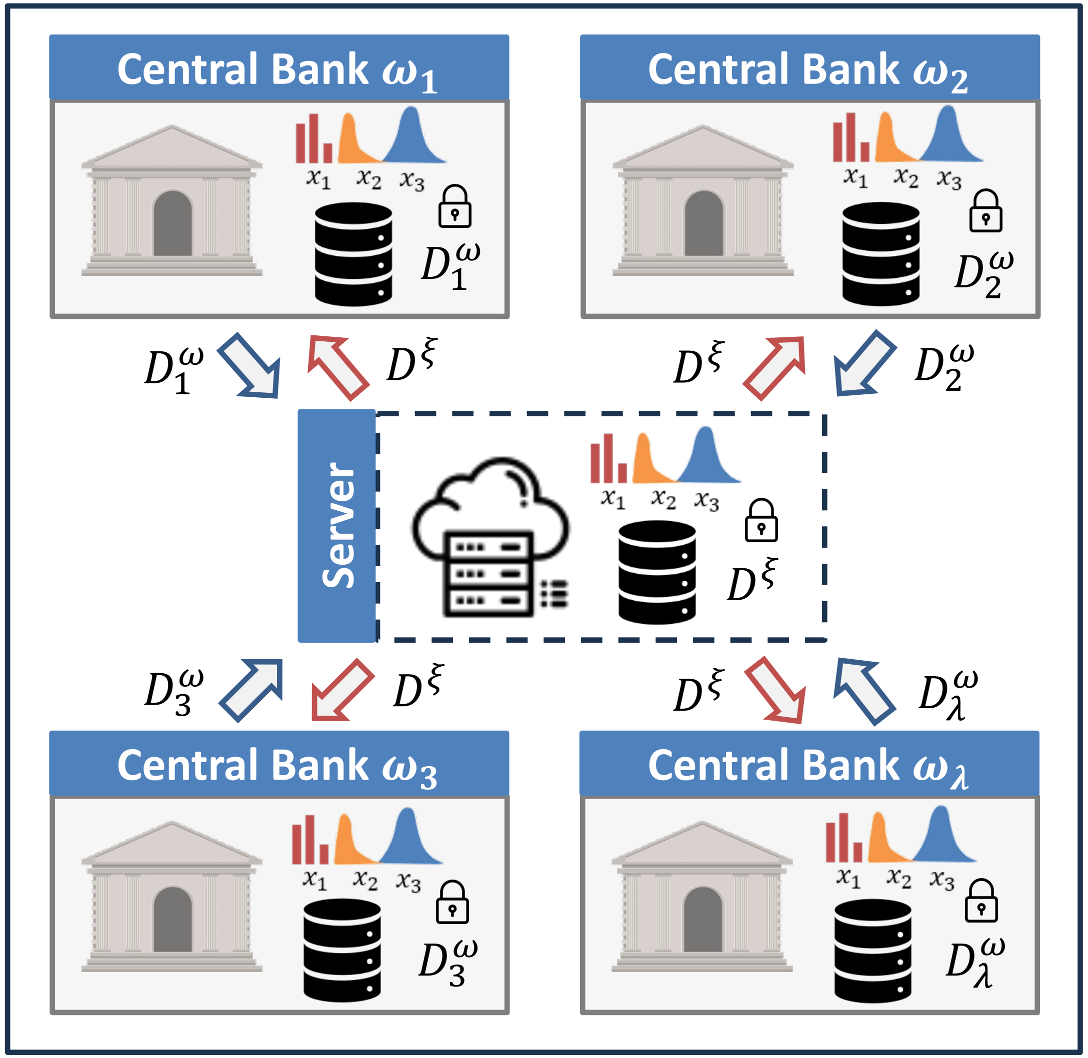
Federated Learning: Benefits

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.

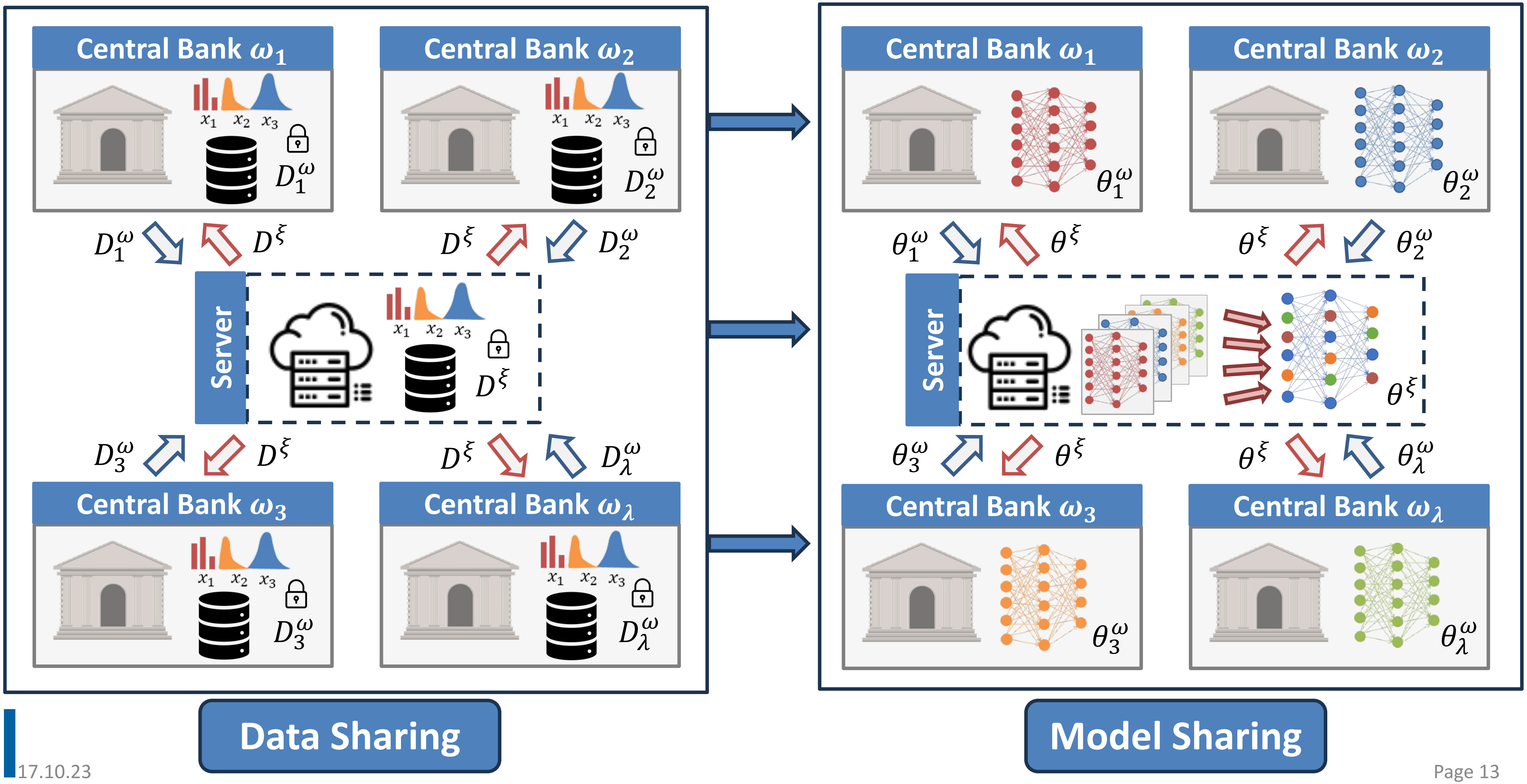
Global Insights: financial institutions can benefit from insights gathered globally across different markets and segments, but applied in a way that is tailored to local market conditions and regulations.

From Data Sharing to Model Sharing

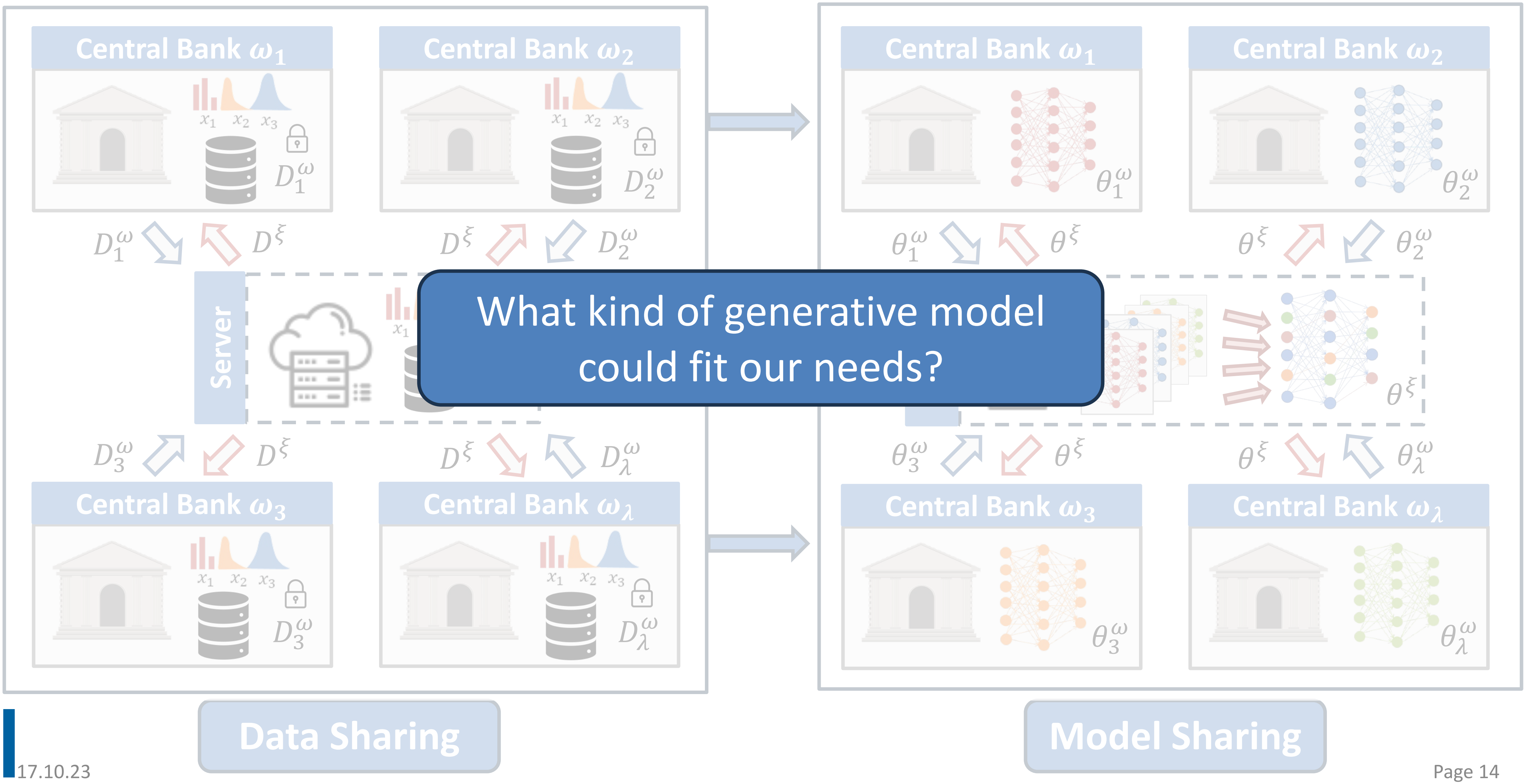


Data Sharing

From Data Sharing to Model Sharing



From Data Sharing to Model Sharing



What kind of generative model could fit our needs?

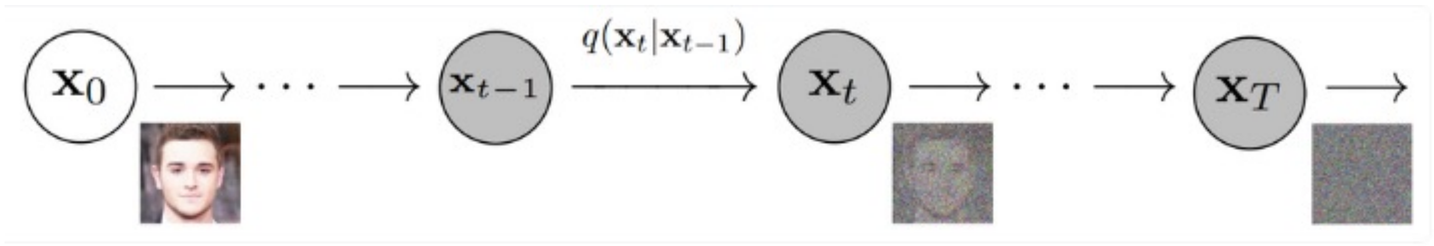
Agenda

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

Diffusion Models

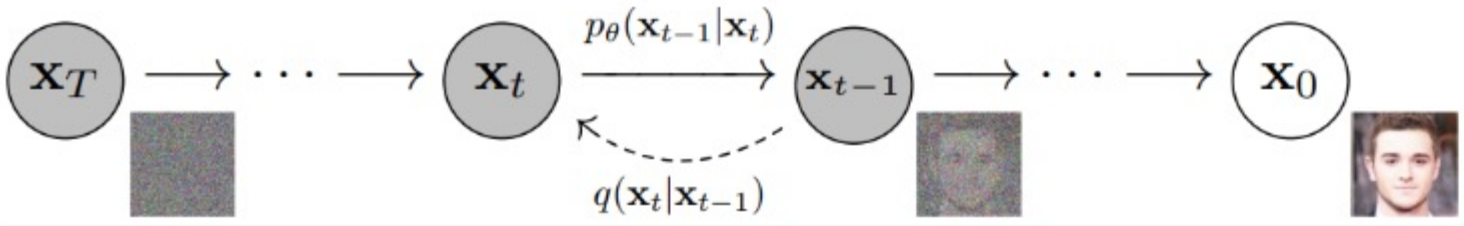
- Diffusion Models – are **generative models** trained with the objective of noise removal and subsequently constructing the desired **data samples from pure noise**.
- First introduced by Jascha Sohl-Dickstein et al. in 2015 with motivation from non-equilibrium thermodynamics. They can be thought as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

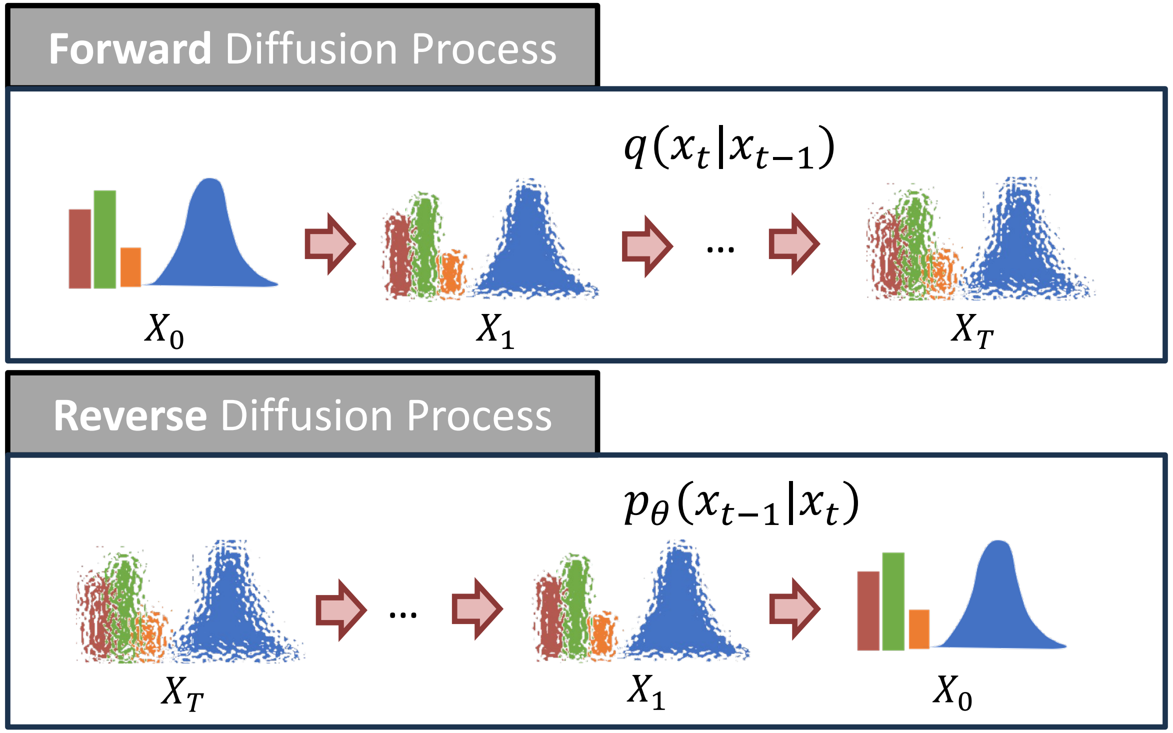
Figure is taken from „Denoising Diffusion Probabilistic Models " by Ho et al. https://arxiv.org/pdf/2006.11239.pdf

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

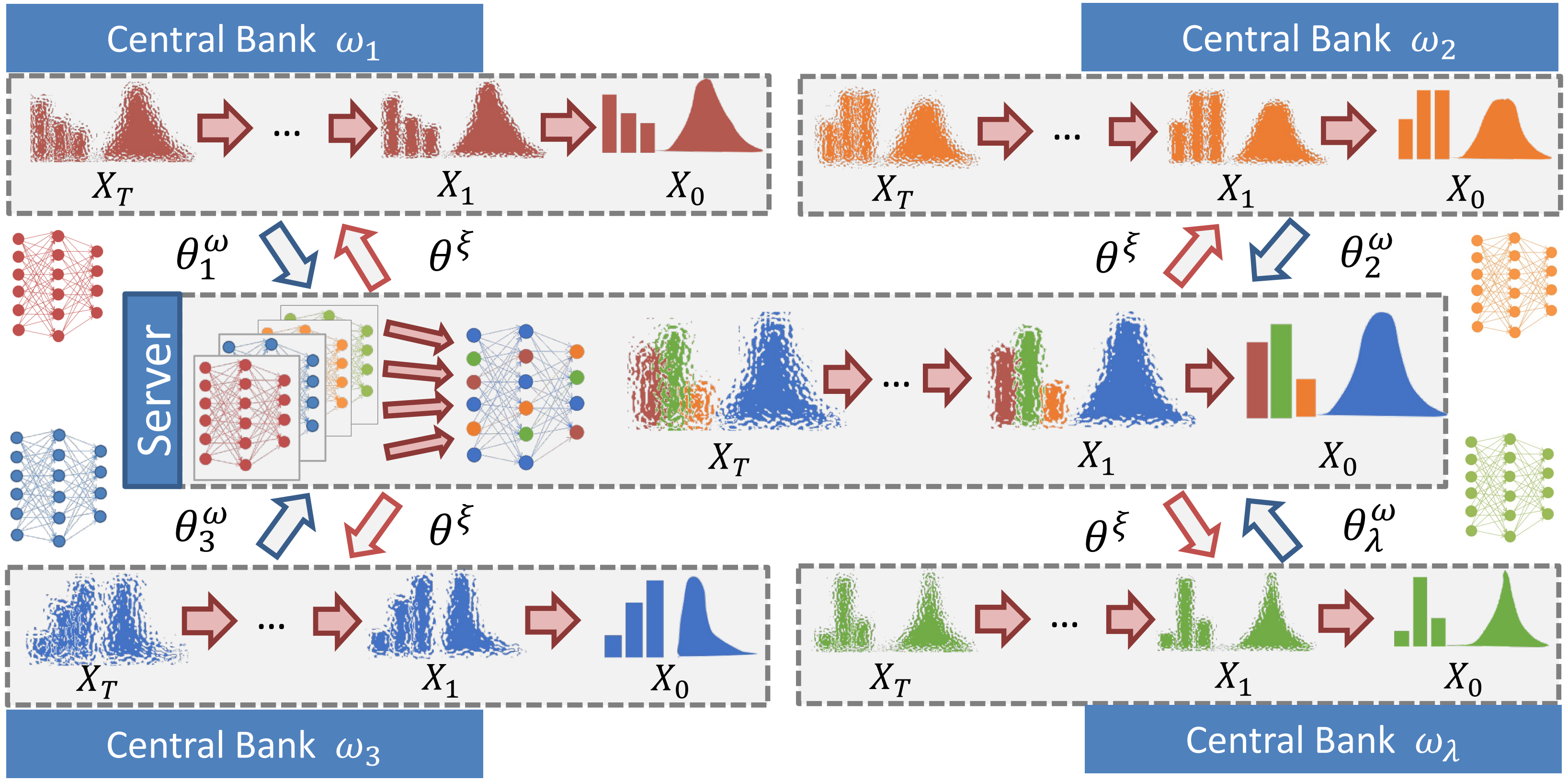
- FinDiff is a diffusion based generative model, that synthesizes financial tabular data for **regulatory downstream tasks**.
- It uses **embeddings** for mixed modality financial data, comprising both **categorical and numeric** attributes.
- Empirical results demonstrate **high fidelity, privacy, and 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 $\boldsymbol{\theta_i^\omega}$.**
2. The Server aggregates all models into a **global generative model $\boldsymbol{\theta^\xi}$** 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

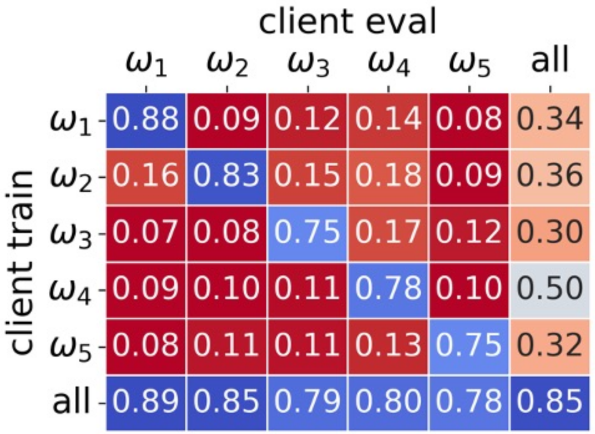
- Comparative analysis of the **Federated** (FedTabDiff) versus **Non-Federated** (FinDiff) models, evaluated using the full dataset D.
- Non-Federated diffusion models are trained individually at each client ω_i with subset $D_i \subseteq D$ (column "Split") and evaluated against the entire dataset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

			Evaluation Measures			
Dataset	**Client**	**Split $\mathcal{D}_i$**	**Fidelity [5, 32] ↑**	**Utility [50] ↑**	**Coverage [7] ↑**	**Privacy [55] ↓**
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	$\mathbf{3.120 \pm 0.09}$

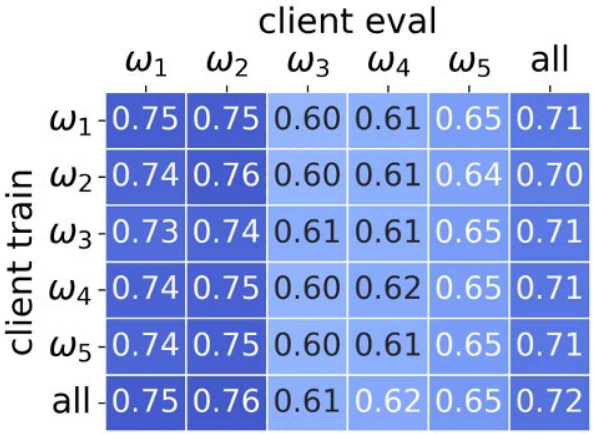
*Scores are derived from the averaged results and standard deviations of five experiments, each initiated with distinct random seeds

Experimental Results

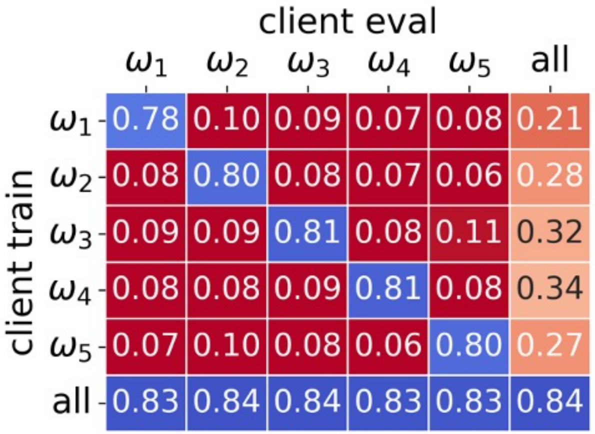
- Fidelity - **similarity of every column** in the synthetic dataset against the real dataset.
- Fidelity score comparison between **Federated** (FedTabDiff) and **Non-Federated** (FinDiff).
- In the Non-Federated model, each client is trained on its data subset $D_i \subseteq D$ and evaluated across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



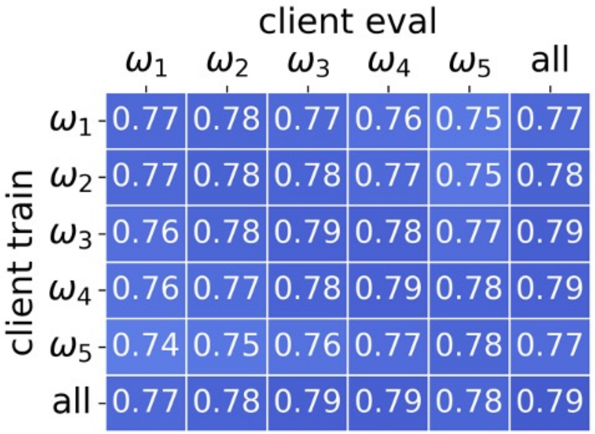
Non-Federated (FinDiff)
Philadelphia



Federated (FedTabDiff)
Philadelphia



Non-Federated (FinDiff)
Diabetes



Federated (FedTabDiff)
Diabetes

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