



Advancing Privacy-Preserving Statistical Innovation: **Bank Indonesia's Pilot on Synthetic Data for Secure Microdata Access**

Anggraini Widjanarti, Akhmad Zacky Nugraha,
Mohammad Khoirul Hidayat, Dwi Cahyo Ardianto, Larasati
Ratusinkaya (Bank Indonesia, Indonesia)



September 14th, 2026

The views expressed here are those of the authors and do not necessarily reflect the views of Bank Indonesia

Outline

01

Background

02

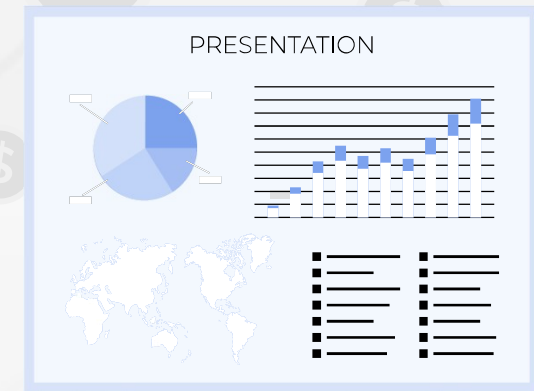
Synthetic Data Overview

03

Bank Indonesia's Pilot

04

Conclusion and Future Works

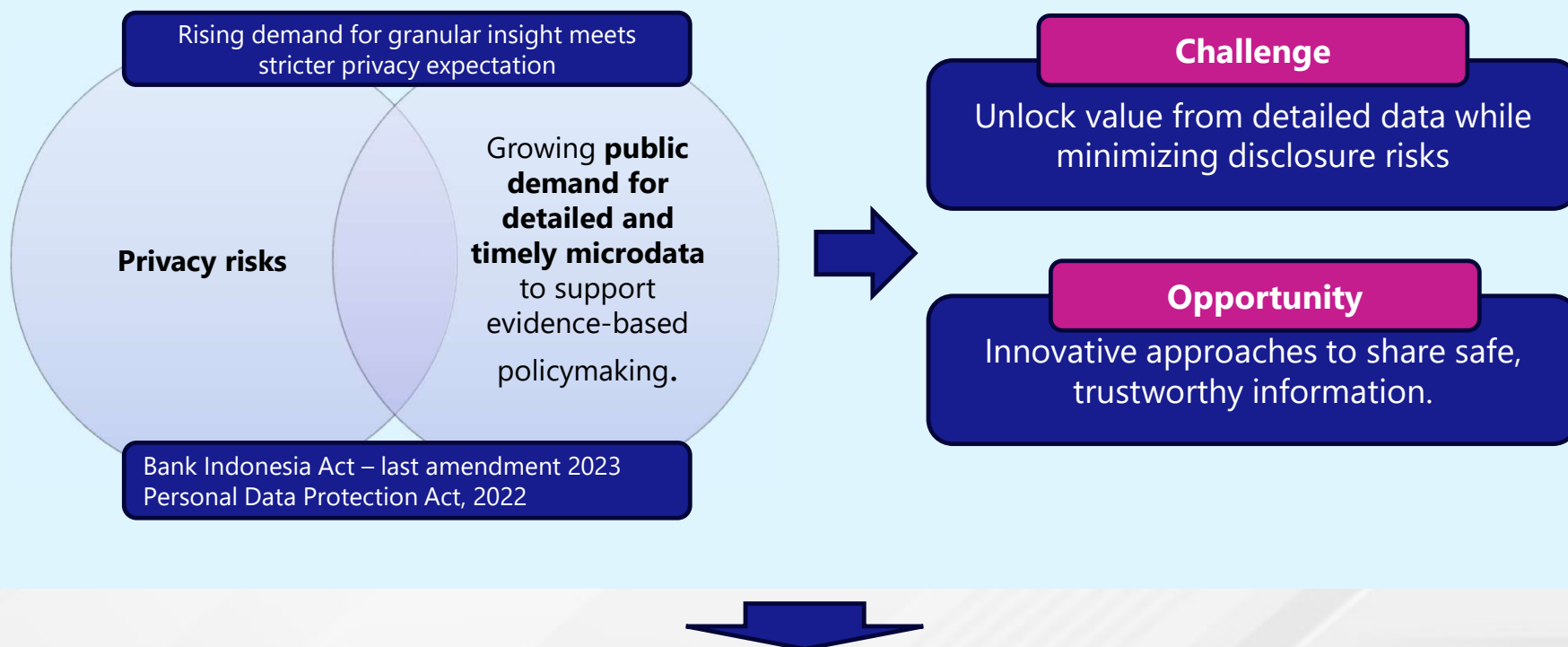




Background

////

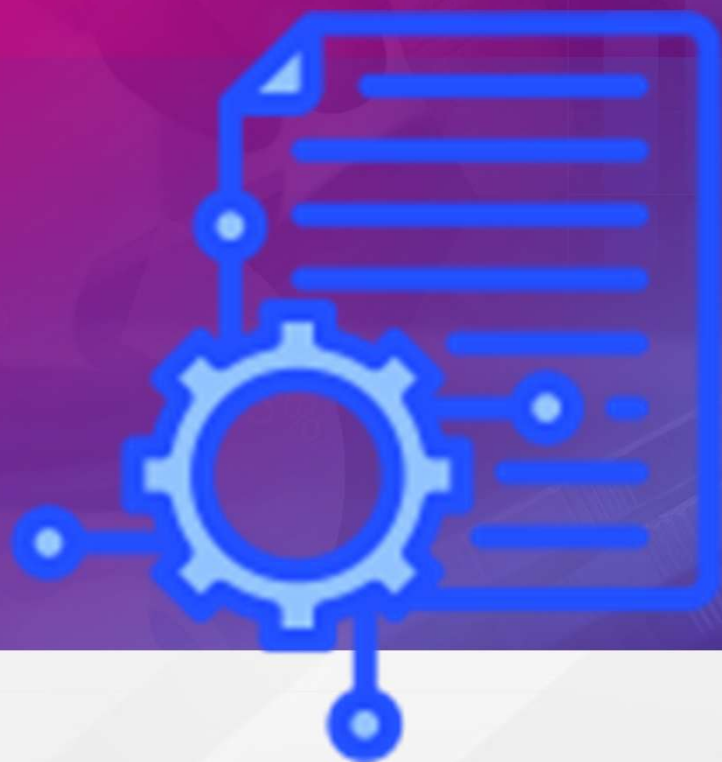
The Trust–Transparency Dilemma



Synthetic data is an emerging innovation increasingly explored by statistical institutions worldwide



How can synthetic data serve as a practical and responsible innovation for Bank Indonesia to make data dissemination safer, more transparent, and more trusted?



Synthetic Data ////

Overview Synthetic Data



Definition:

Artificially generated datasets that mimic the statistical properties of real data **without containing actual records** (Rubin, 1993)



Principle:

Synthetic data for secure microdata access should closely **replicate the statistical patterns, structure, and distribution of the original data while enabling in-depth analysis without exposing** the underlying sensitive microdata.



Advantages:

- Deeper, more granular access for authorized research & internal analytical use
- Lower re-identification and disclosure risks compared to raw microdata sharing
- Aligned with the concept of controlled and efficient access to sensitive statistical data



Purposes:

1. **Data augmentation:** enriching or balancing datasets to improve the performance of machine learning models
2. **Privacy Enhancing Technology (PETs):** synthetic data enables organizations to share granular datasets with external users while safeguarding confidentiality

Best Practices for Implementing Synthetic Data Use Cases

Effective synthetic data implementation depends on three pillars: (1) **robust technical evaluation**, (2) **alignment with legal and ethical safeguards**, and (3) **transparent communication with data users**.

United States Census Bureau & Office for National Statistics: Synthetic datasets must be accompanied by **transparent documentation, guidance on appropriate use, and explicit communication** about their limitations (Abowd, 2018; ONS, 2022)

ONS Data Science Campus: Developed the SynthGauge framework to measure **fidelity, utility, and privacy risk in a balanced way** (2022)

UNECE: Recommended that synthetic data be evaluated **not only on statistical performance**, but also on governance principles such as **accountability, proportionality, and transparency** (2021)

Synthetic Data in Practice



Healthcare

Evaluation frameworks emphasize the need to measure disclosure risks quantitatively before synthetic datasets are shared with researchers (El Emam et al., 2020)



Transportation

Synthetic trip data for car-sharing programs requires validation against real-world distributions to ensure credibility for decision-making (Albrecht et al, 2024)



Financial

Synthetic data can enable scenario testing for payment systems, but also stressed that governance frameworks must clearly delineate experimental analysis from official statistics (Docherty and Wang, 2010)



Bank Indonesia's Pilot ////

Synthetic Data for Secure Microdata Access: International Transaction Reporting System (ITRS)

ITRS Data Overview

ITRS reporting is mandated through Law No.4 of 1999 on FX Flows Act and BI regulation no.21 of 2019 as amended lastly by BI regulation no.9 of 2024 on FX Flows reporting, to report flows of foreign assets and liabilities of Indonesian residents through banks.

Scope:



Reporters: banks



Threshold: $\geq$ USD10,000



Granularity: transaction level, with identified counterparties and banks involved



Transaction reported: current, capital, financial account, others



Reporting purposes: external statistics compilation, fx flows monitoring, regulatory and supervisory needs

Why This Dataset for the Pilot?

ITRS combines the highest statistical complexity with the highest information sensitivity — the most demanding conditions for testing synthetic data:

1. High transaction granularity
2. Numerical and categorical variables
3. Sensitive microdata
4. Relatively high disclosure risk
5. High analytical value

PILOT FOCUS: generating synthetic data for individual payer and payee information after unnecessary direct identifiers are removed.

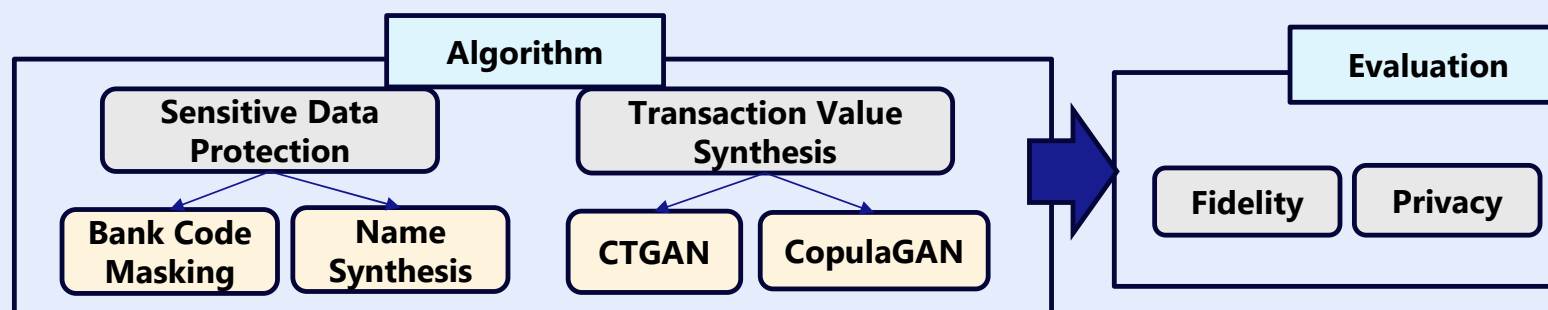
Synthetic Data for Secure Microdata Access: ITRS Data

Developing synthetic data to enable the utilization and data sharing (ITRS data) without exposing sensitive information from the original dataset.

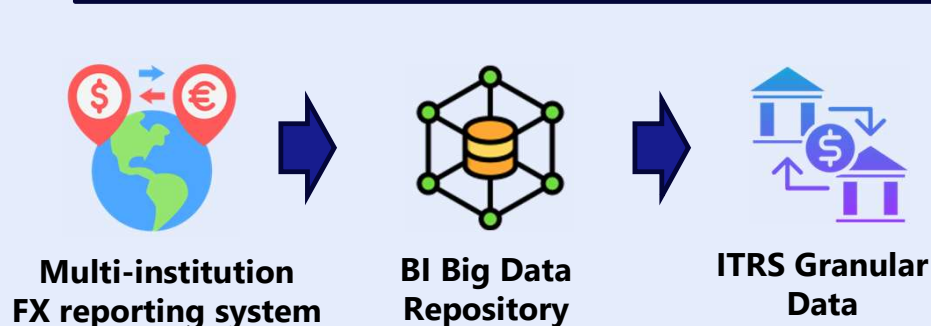
Objective

1. Assess whether synthetic data **can replicate the key statistical properties** of the original dataset without disclosing **confidential records**.
2. **Evaluate synthetic data** based on fidelity and privacy.
3. Assess the **potential of synthetic data** for **secure data sharing** and **analysis**.

Methodology



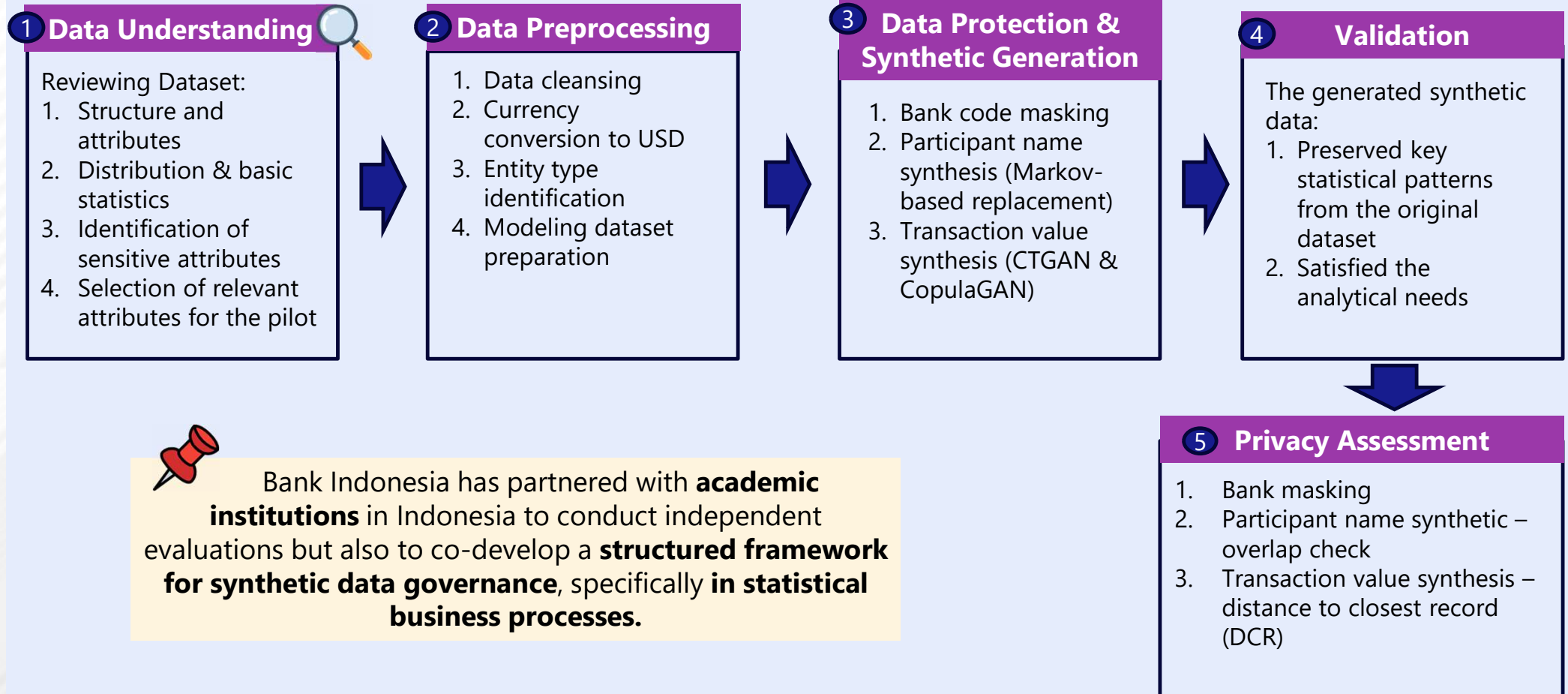
Data Sources



The data are highly granular as they contain individual-level information with a large sample size

Synthetic Data Pilot Framework and Methodology (PET Approach)

The pilot follows the principle that synthetic data should preserve the key statistical patterns, structure, and distribution of the original ITRS data, while enabling analytical use without exposing sensitive microdata.



3 Synthetic Data Method

Methods Used for Synthetic Name & Transaction Value Generation

Faker Library & Markov-chain

- **Hybrid approach:** rule-based library (**Faker**) + statistical model (**Markov Chain**)
- **Faker** – generates realistic names from a locale-based library
- **Markov Chain** – learns transition patterns from real names to generate new, statistically consistent surrogate names
- Balances speed and statistical realism, lowering re-identification risk

Reference: Faker Developers. (2024). *Faker: Python package that generates fake data*.

<https://faker.readthedocs.io/>

A Markov Chain Replacement Strategy for Surrogate Identifiers: Minimizing Re-Identification Risk While Preserving Text Reuse. *Electronics*, 2025. DOI: <https://doi.org/10.3390/electronics14193945>

Conditional Tabular GAN (CTGAN)

- **A GAN-based generative model adapted for tabular data** with mixed numerical and categorical variables
- Works through **two conditional** components:
 - **Generator** – produces synthetic data conditioned on specific categories/feature values
 - **Discriminator** – learns to distinguish real vs. synthetic samples
- As the generator and discriminator compete during training, the **generator steadily improves at replicating the real data's statistical patterns** and category distributions.

Reference: A. Nigam and S. Srivastava, "Generating Realistic Synthetic Traffic Data using Conditional Tabular Generative Adversarial Networks for Intelligent Transportation Systems," *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, Bilbao, Spain, 2023, pp. 2881-2886, doi: 10.1109/ITSC57777.2023.10422234

CopulaGAN

- **A hybrid model combining** statistical (**Gaussian Copula**) transformation with **CTGAN's** deep learning architecture.
- **Applies a copula-based** (Cumulative Distribution Function) **transformation to the data first**, making it easier for the underlying GAN to learn patterns.
- Designed to **bridge the gap between purely statistical models and purely GAN-based models**.

Reference: Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 399–410). <https://doi.org/10.1109/DSAA.2016.49>

Descriptive Statistics on Transaction Value (Real Dataset)

The ITRS data used in this pilot covers the period of January–December 2024

Month	Count	Mean	Std	Min	25%	Median	75%	Max
Jan	497,949	1,969,532	17,562,581	0	21,553	60,293	366,113	1,208,000,000
Feb	431,998	2,030,921	25,294,083	0	22,275	64,592	406,030	6,035,000,000
Mar	461,341	2,075,537	24,745,118	0	23,313	66,600	415,157	6,035,000,000
April	429,815	1,890,721	13,383,584	0	23,816	67,029	400,000	1,350,000,000
May	471,623	1,868,933	21,944,877	0	23,380	64,411	364,656	6,035,000,000
Jun	452,029	2,050,962	15,873,685	0	23,155	64,993	401,157	1,482,199,334
Jul	540,661	1,932,165	13,879,513	0	22,319	63,051	400,000	1,000,000,000
Aug	544,818	2,182,741	23,118,596	0	23,721	68,241	480,704	6,035,000,000
Sep	526,672	2,230,310	17,174,151	0	23,713	67,649	437,535	900,130,750
Oct	550,642	2,153,238	16,845,673	0	22,938	65,324	407,664	1,099,588,759
Nov	500,760	1,931,757	22,836,969	0	22,227	62,488	369,924	6,035,000,000
Dec	524,995	1,969,047	22,003,108	0	21,822	60,118	330,439	6,035,000,000

- **Transaction values are highly right-skewed**, the mean (~\$1.9 – 2.2 M) consistently exceeds the median (~\$60k–68k), indicating a few very large transactions pulling the average upward.
- **High variance relative to the mean** (Std: \$13M–25M) confirms the presence of extreme outliers in the dataset.
- **Recurring identical maximum values** (\$6,035,000,000) across several months suggest a possible reporting threshold or recurring high-value transaction pattern.
- **Minimum values of 0 appear across all months**; further **lookup via the transaction purpose** field is needed to determine the cause (e.g., non-monetary reporting)
- These skewed and extreme-value patterns represent the **rare events and high-sensitivity attributes** that challenge generative models (CTGAN, CopulaGAN) in replicating statistical fidelity.

Evaluation: Fidelity

CTGAN

Month	KS Score	Histogram Similarity
Jan	0.8807	0.9872
Feb	0.9212	0.9985
Mar	0.8918	0.9992
April	0.9034	0.9964
May	0.8674	0.9988
Jun	0.9389	0.9945
Jul	0.9163	0.9968
Aug	0.9138	0.9978
Sep	0.8446	0.9936
Oct	0.9212	0.9969
Nov	0.9197	0.9983
Dec	0.8474	0.9988
Average	0.897	0.9964

CopulaGAN

Month	KS Score	Histogram Similarity
Jan	0.9251	0.9927
Feb	0.8997	0.9985
Mar	0.9569	0.9984
April	0.9304	0.9941
May	0.9384	0.9988
Jun	0.9486	0.9936
Jul	0.9413	0.9951
Aug	0.9476	0.9978
Sep	0.8796	0.9900
Oct	0.9244	0.9960
Nov	0.8939	0.9983
Dec	0.9028	0.9981
Average	0.924	0.9960

- **CopulaGAN outperforms CTGAN in KS Score**, scoring higher in 10 of 12 months, indicating stronger marginal distribution fidelity
- **Histogram similarity is comparably high for both models**
- **CTGAN shows greater month-to-month variability** in KS Score (range: 0.8446–0.9389) compared to CopulaGAN's more stable performance (range: 0.8796–0.9569)
- Overall, **CopulaGAN demonstrates more consistent statistical fidelity**, while both models maintain strong histogram-level similarity to the original data

Evaluation: Privacy Score (DCR)

Month	Mean DCR	Median DCR	Min DCR	Max DCR	Exact Match (%)
Jan	0.000000878462	0.000000000392	0	0.00041174	2.61041593
Feb	0.000000121530	0.000000000073	0	0.00007578	2.28319815
Mar	0.000000352224	0.000000000079	0	0.00040892	0.95769355
April	0.000000504127	0.000000000347	0	0.00034917	0.89108105
May	0.000000077278	0.000000000057	0	0.00009204	7.83676793
Jun	0.000001368653	0.000000000304	0	0.00130912	3.89908612
Jul	0.000000586035	0.000000000403	0	0.00045604	1.82091921
Aug	0.000000090720	0.000000000067	0	0.00007891	5.05288004
Sep	0.000000420677	0.000000000551	0	0.00036739	2.02915667
Oct	0.000000913172	0.000000000311	0	0.00186036	4.13644437
Nov	0.000000047421	0.000000000070	0	0.00004864	3.11646298
Dec	0.000000114088	0.000000000066	0	0.00020869	0.73905466

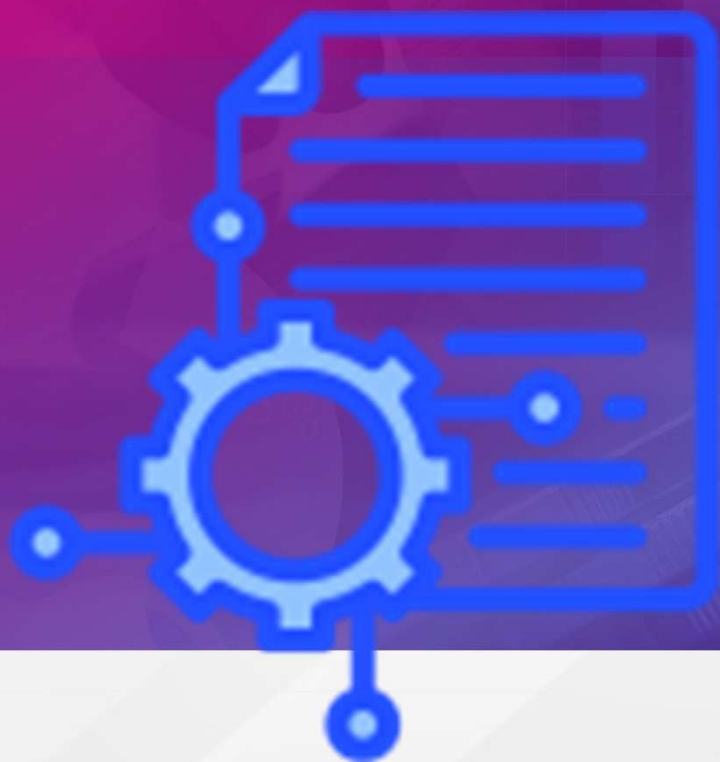
Methodological Note:

DCR in this evaluation **was calculated using only a single synthesized variable (transaction value)**, which limits the scope of the assessment and may result in higher exact match rates than those observed in a full multi-variable evaluation.

Implication:

These findings highlight the need for well-calibrated risk thresholds and a comprehensive multi-variable DCR assessment in future iterations of the pilot.

- **Mean and median DCR are near zero** across most months, indicating many synthetic records fall extremely close to their nearest real record.
- **Exact Match rate ranges from 0.74% to 7.84%** across months, with the highest occurring in May, meaning a non-trivial share of synthetic records are identical to real records.
- Elevated exact match rates likely correspond to **months with more skewed, concentrated transaction value distributions** (consistent with earlier descriptive statistics findings)



Conclusion and Future Works ////

Conclusion and Future Works

Conclusion

Synthetic data is best positioned as a **Privacy Enhancing Technology (PET)**, widening access to granular statistics while keeping disclosure risk within acceptable limits, consistent with the Personal Data Protection Act (2022) and international guidance (UNECE, 2021). Bank Indonesia's ITRS pilot supports this principle: **both models replicated the distribution of transaction values with high fidelity** (average KS 0.924 for CopulaGAN, 0.897 for CTGAN), **while exact-match rates of 0.74–7.84% show that fidelity alone does not guarantee privacy**. Synthetic ITRS data is therefore fit for research and exploratory analysis, not yet for regulatory reporting or policy calibration.

Future Works

The pilot also identified several areas for further improvement before the approach can be scaled beyond the pilot stage. Future work will **focus on strengthening the evaluation of disclosure risk**, refining the governance and release criteria for synthetic ITRS data, and improving documentation and user engagement to support responsible use and wider dissemination.



THANK YOU