

# A Gentle Introduction to Synthetic Data Generation

*GPA 2025, Seoul*

Khaled El Emam



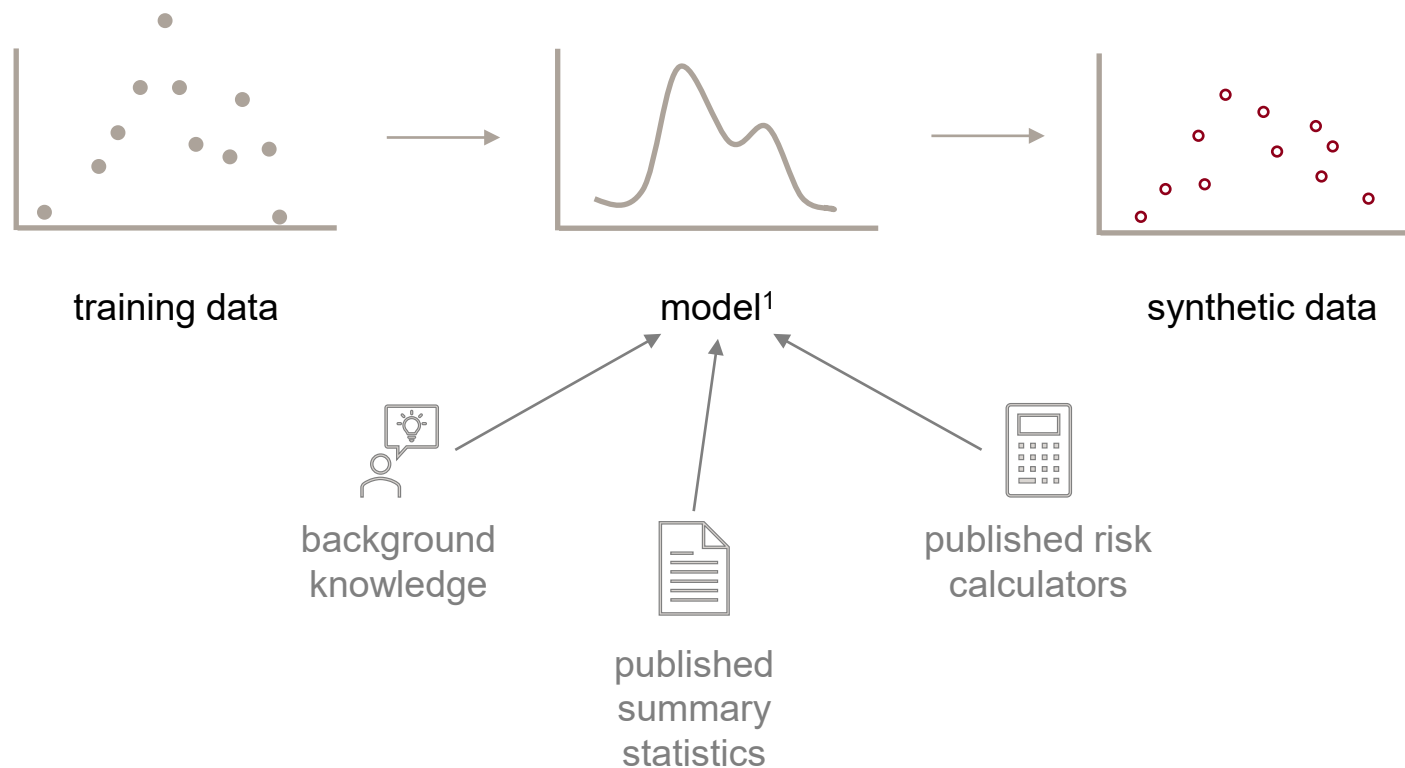
# What is synthetic data generation ?

- Synthetic data generation (SDG) is a type of privacy enhancing technology (PET)
- One of its primary use cases is to create non-personal information from personal information
- This would then enable the use and disclosure of the synthetic data with reduced (minimal?) obligations



K. El Emam, L. Mosquera, and R. Hoptroff, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. Sebastopol, CA: O'Reilly Media, 2020.

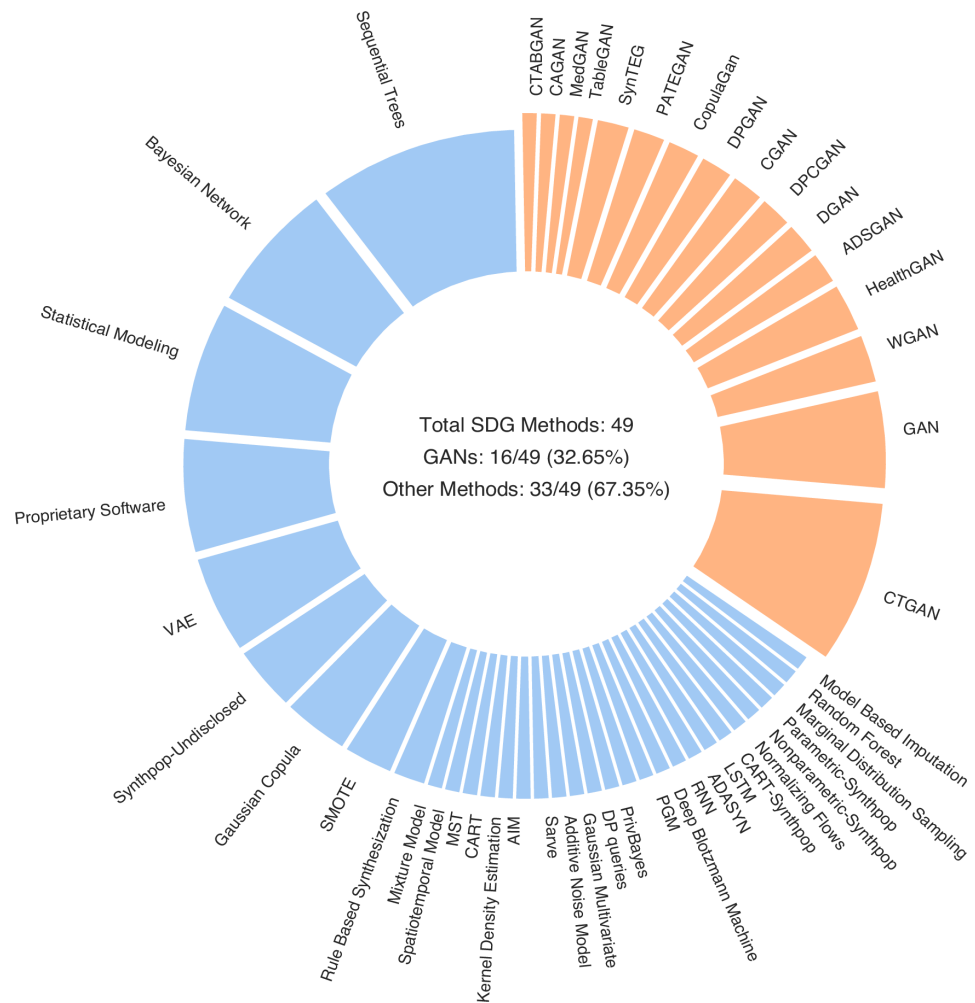
# Synthetic data generation typically involves a training dataset



<sup>1</sup> Our lab maintains the Python library **pysdg** with a unified interface to multiple SDG models: <https://github.com/CHEO-EHIL/pysdg-releases>

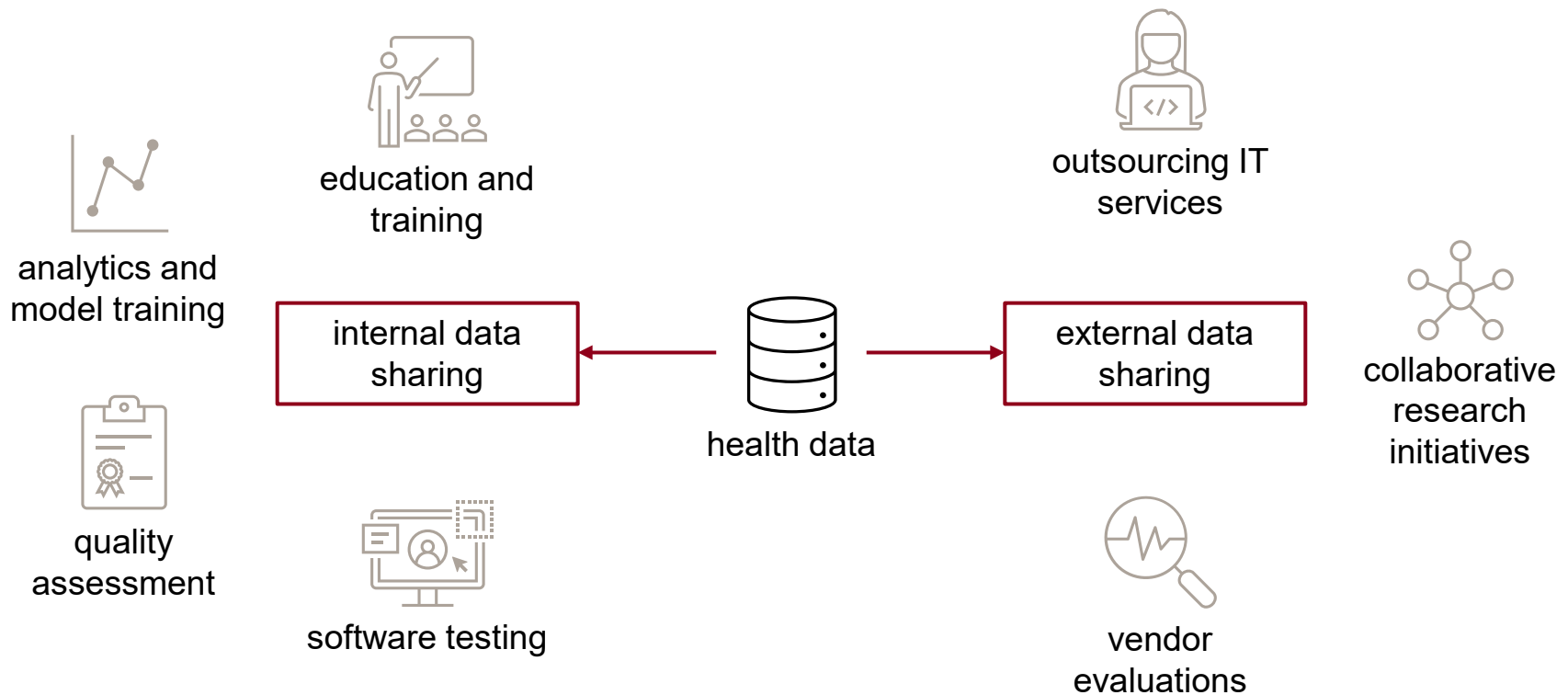
# There are multiple AI/ML models for SDG

- Generative Adversarial Networks (GANs) are among the most used SDG models
- More “traditional” statistical approaches include Sequential Trees or Bayesian Networks (BN)
- Recent approaches include Adversarial Random Forests (ARF) or Diffusion Models

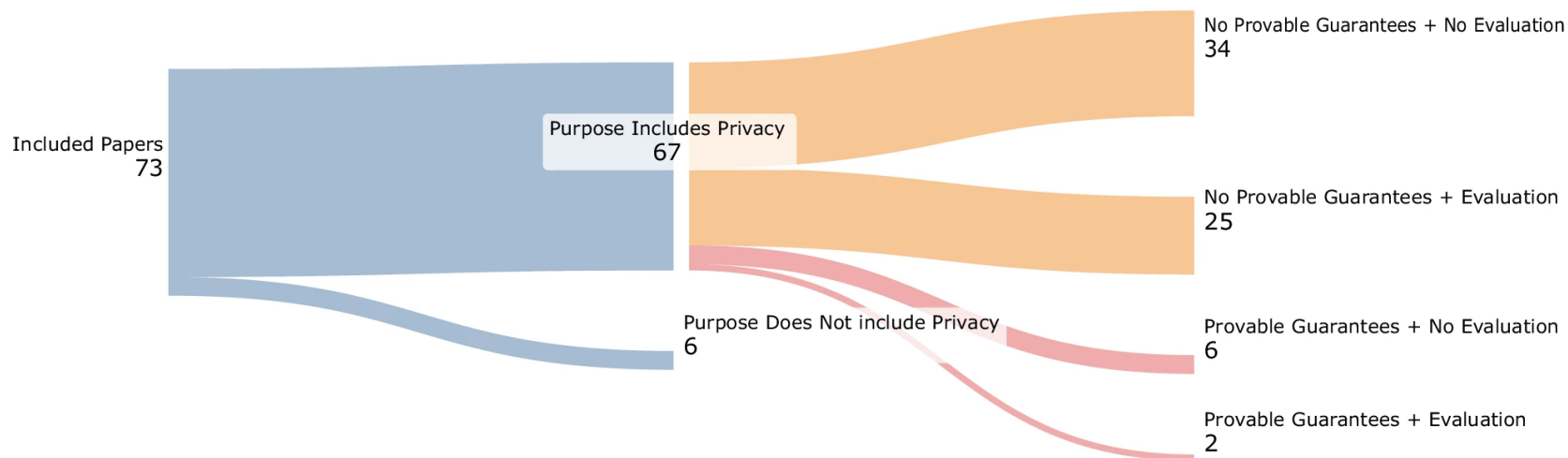


Kaabachi, B. et al. (2025). A scoping review of privacy and utility metrics in medical synthetic data. NPJ Digit Med 8, 60. <https://doi.org/10.1038/s41746-024-01359-3>.

# Non-personal data is essential for multiple internal and external applications



# Privacy is not always assessed in synthetic data studies



Kaabachi, B. et al. (2025). A scoping review of privacy and utility metrics in medical synthetic data. NPJ Digit Med 8, 60. <https://doi.org/10.1038/s41746-024-01359-3>.

Until now there has been no consensus on how to measure the residual privacy vulnerability in synthetic data and that uncertainty may be one reason privacy often isn't evaluated at all.

# How is synthetic data regulated ?

- Very few laws mention synthetic data explicitly
- Some regulatory guidance, with analogy to de-identification / anonymization
- Perspectives of regulators

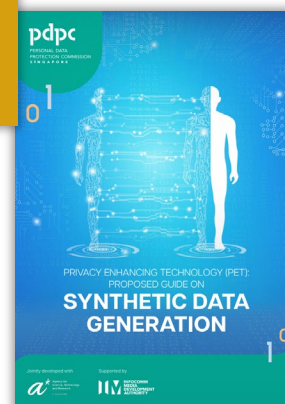


# Regulatory guidance on synthetic data

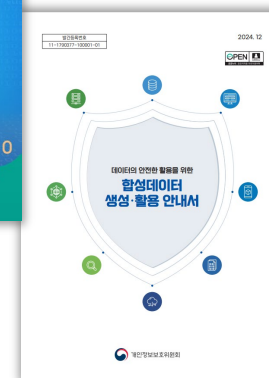
- Synthetic data is not inherently devoid of risks
- Privacy evaluation is emphasized as an integral part of good practices
- Utility and fairness are relevant considerations for managing risks from synthetic data



United  
Kingdom  
2023



Singapore  
2024



South Korea  
2024

Pilgram L, Ko H., Tung A. and El Emam K. Is Synthetic Data Protective of Individual Patient Privacy ? A Regulatory Perspective. Under Review (2025).

# Canadian perspectives

- Uncertainty about whether specific consent is required for the act of generating synthetic data
- As a practical matter most of the time consent is not sought
- Regulators said that developing codes of practices or standards would be valuable for regulating synthetic data



- K. El Emam, A. Fineberg, E. Jonker, and L. Pilgram, "Perspectives of Canadian privacy regulators on anonymization practices and anonymized information: a qualitative study," *Int. Data Priv. Law*, p. ipae017, Dec. 2024, doi: 10.1093/idpl/ipae017.
- L. Pilgram, A. Fineberg, E. Jonker, and K. El Emam, "An assessment of synthetic data generation, use and disclosure under Canadian privacy regulations," *AI Ethics*, Aug. 2025, doi: 10.1007/s43681-025-00819-0.



# Emerging international frameworks and standards around synthetic data



Synthetic Data Guidelines

**APPROVED WORK ITEM**

## ISO/IEC AWI TR 42103

Information technology — Artificial intelligence — Overview of synthetic data in the context of AI systems

---

**Under development**  
A working group has prepared a draft.

**IEEE SA** STANDARDS ASSOCIATION

**Synthetic Data**  
Industry Connections Activity Initiation Document (ICAID)  
Version: 2.0, 13 November 2023  
**IC21-013-02 Approved by the CAG 22 December 2023**

International Standards

# Consensus framework for privacy evaluation in synthetic data

## Membership Disclosure Vulnerability



Formulate an explicit threat model and check if it aligns with the implicit assumptions in the metric



Model the member prevalence in the attack dataset: consider the sampling fraction



Factor in the adversary's background knowledge: match on QIs (and their possible subsets)

$$F_{\beta}$$

Inform performance measurement by the threat model: pick proper weights for the  $F_{\beta}$  score

$$F_{rel} = \frac{F_{\beta} - F_{naive}}{1 - F_{naive}}$$

Account for a naïve membership guess: calculate the relative  $F_{\beta}$  score

## Attribute Disclosure Vulnerability



Consider knowledge generation: use a holdout dataset as proxy for the non-member baseline



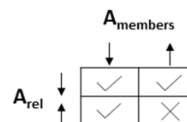
Avoid pre-selection bias: calculate vulnerability for all potential targets



Inform performance measurement by the threat model: use e.g. AUROC for binary classification

$$A_{rel} = A_m - A_{nm}$$

Incorporate the non-member baseline: calculate the relative  $A_{rel}$



Account for a random guess: use both,  $A_m$  and  $A_{rel}$  to guide decision-making

# Consensus framework for privacy evaluation in synthetic data

## Membership Disclosure Vulnerability



Formulate an explicit threat model and check if it aligns with the implicit assumptions in the metric



Model the member dataset: consider the s



Factor in the adversarial match on QIs (and their

$$F_{\beta}$$

Inform performance r model: pick proper weights for the  $F_{\beta}$  score

$$F_{rel} = \frac{F_{\beta} - F_{naive}}{1 - F_{naive}}$$

Account for a naïve membership guess: calculate the relative  $F_{\beta}$  score

## Attribute Disclosure Vulnerability



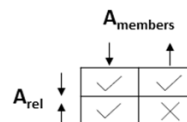
Consider knowledge generation: use a holdout dataset as proxy for the non-member baseline



pre-selection bias: calculate vulnerability for potential targets

performance measurement by the threat use e.g. AUROC for binary classification

incorporate the non-member baseline: calculate the  $A_{rel}$



Account for a random guess: use both,  $A_m$  and  $A_{rel}$  to guide decision-making



**QUESTIONS**