FACULTY OF SCIENCES

SYNTHETIC DATA GENERATION

Frie Van Bauwel, Marcin Jedrych, Xueting Li

GHENT UNIVERSITY

INTRODUCTION

- Why synthetic data?

  Data shortage, High acquisition costs,

  Privacy protection

- Benefits :

  More Data, Privacy Protection, Scalability,

  Cost-Effective

GHENT
UNIVERSITY

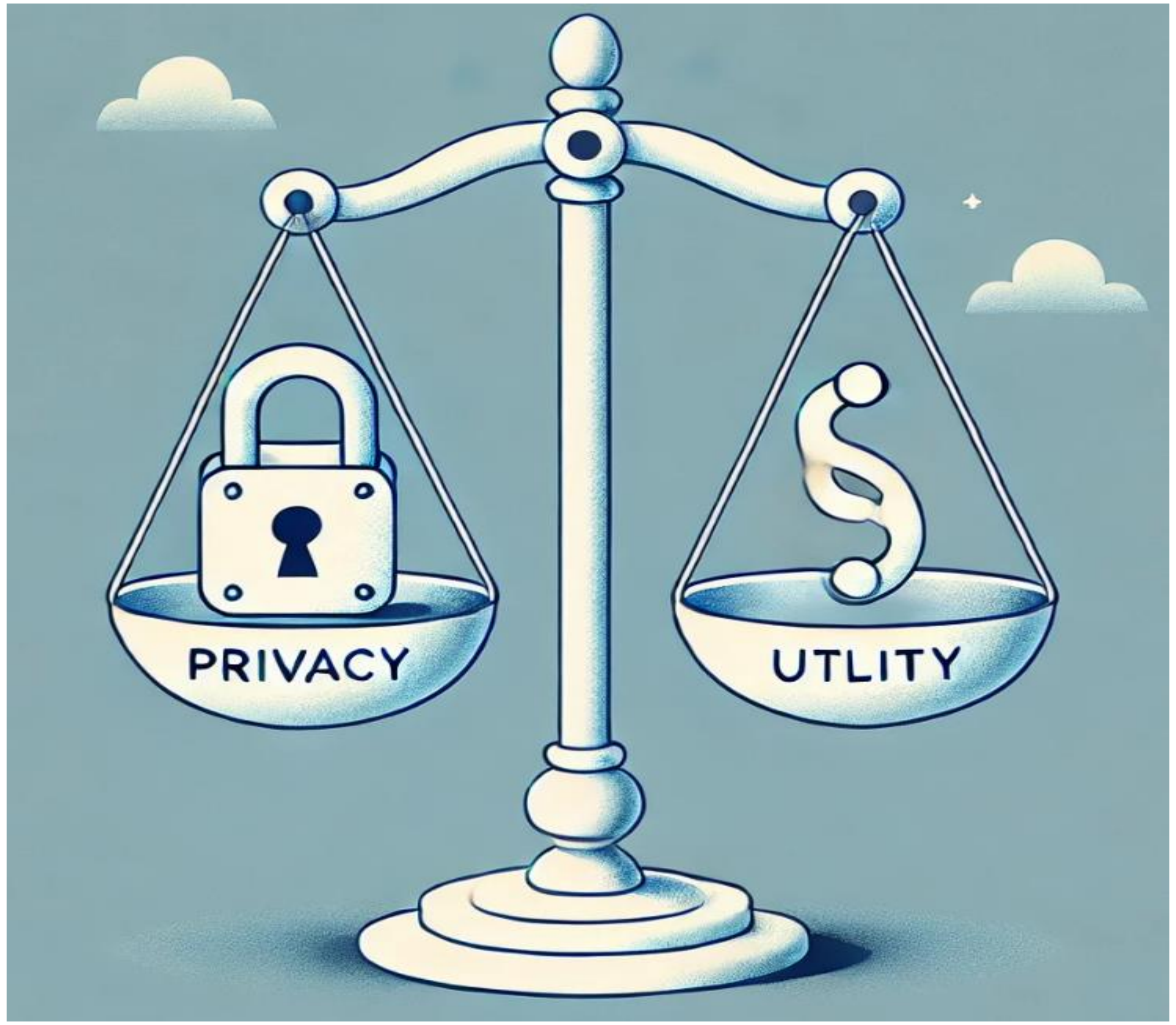
THE TRICK OF USEFUL SYNTHETIC DATA

Privacy-utility trade-off

- Synthetic data with high utility of comes at a cost in privacy

- Stronger privacy often reduces analytical benefits

- Makes it difficult to evaluate synthetic data



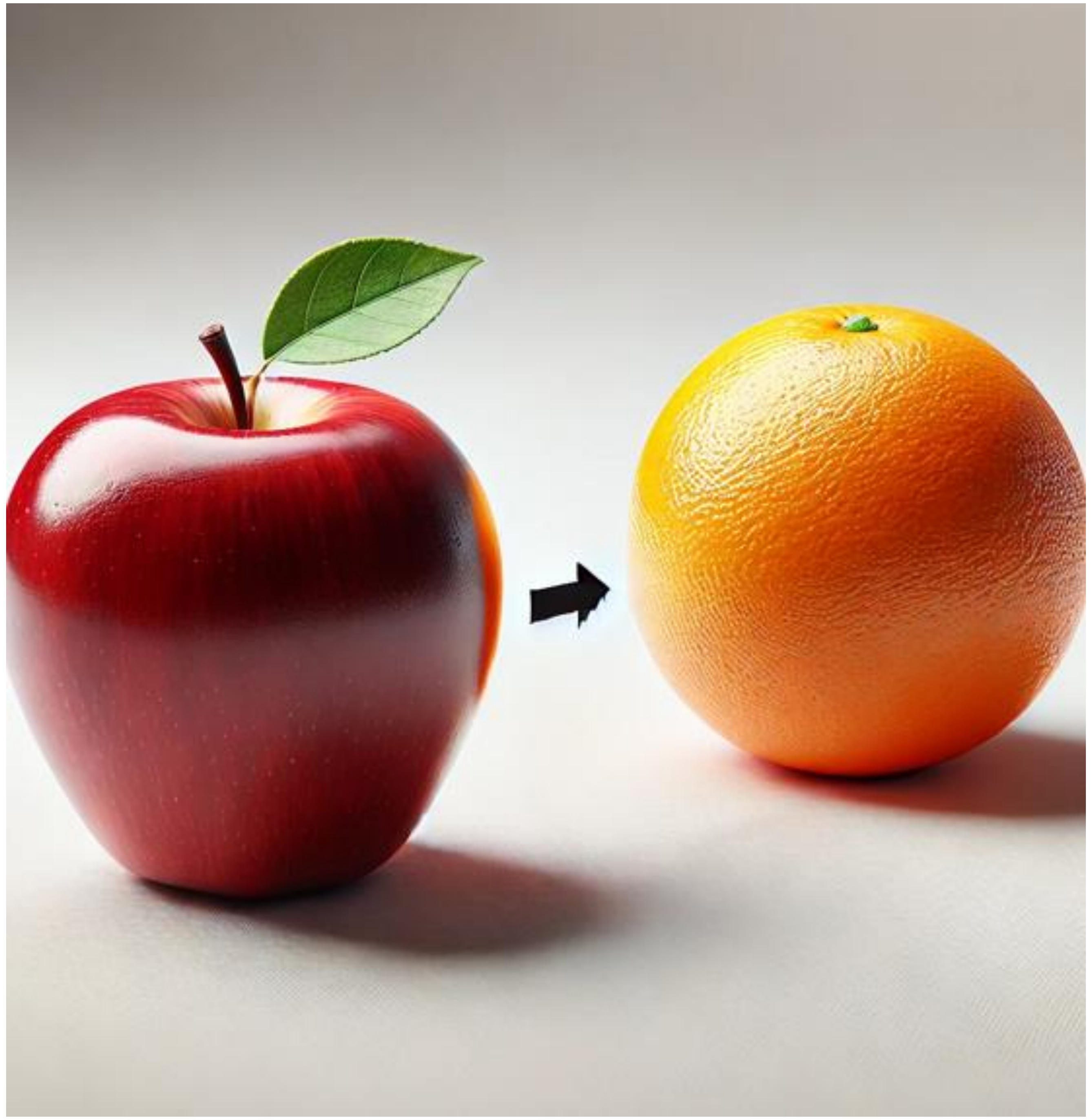
GENERATING SYNTHTETIC DATA

Multiple types of generation techniques

- **Traditional** e.g. Bootstrapping, Monte Carlo, Gaussian Mixture Models

- **Domain-specific** e.g. procedural generation for images, rule-based methods for tabulur data

- **Deep Learning** e.g. GANs, VAEs



GHENT UNIVERSITY

GENERATIVE ADVERSARIAL NETWORKS (GAN)

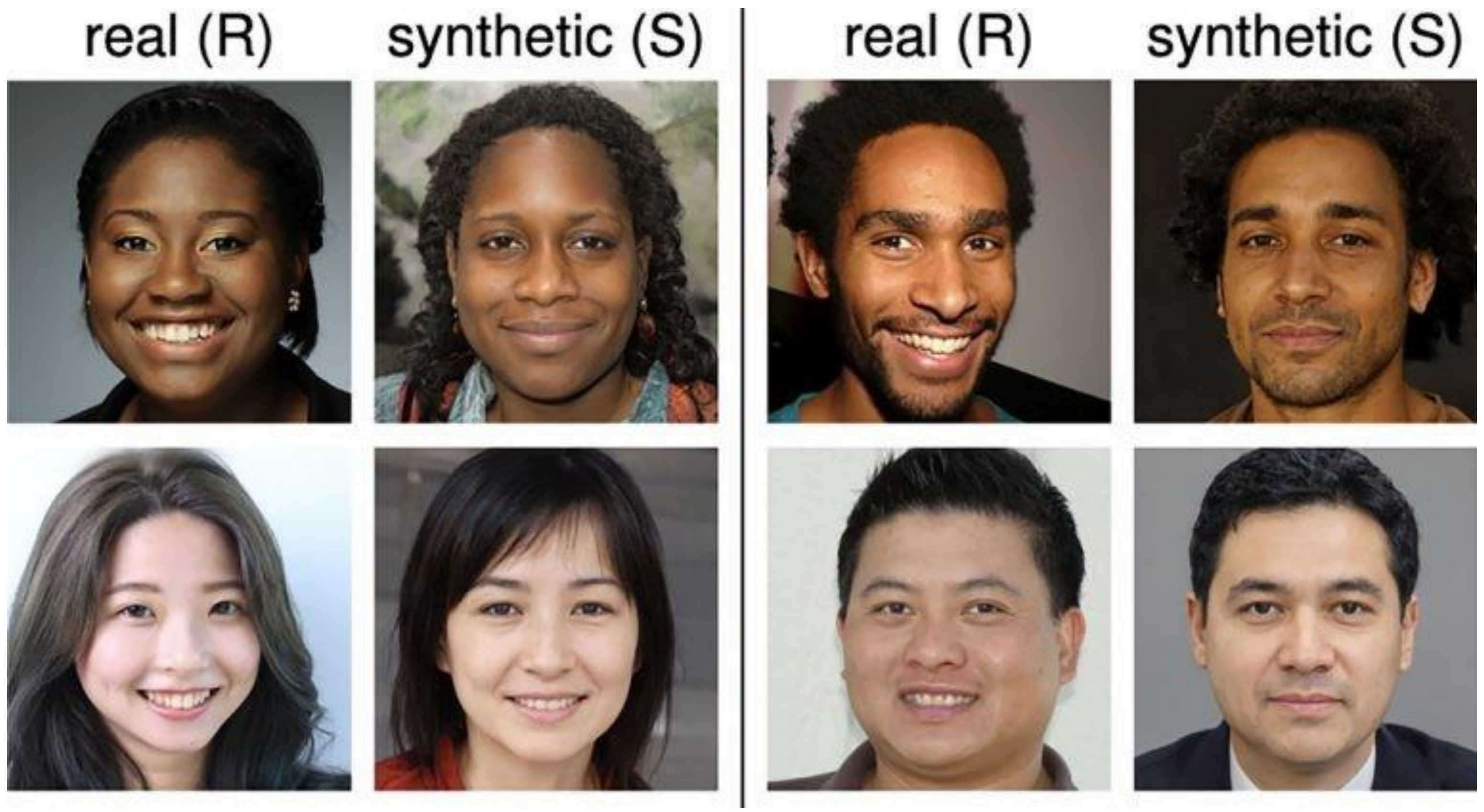
GHENT UNIVERSITY

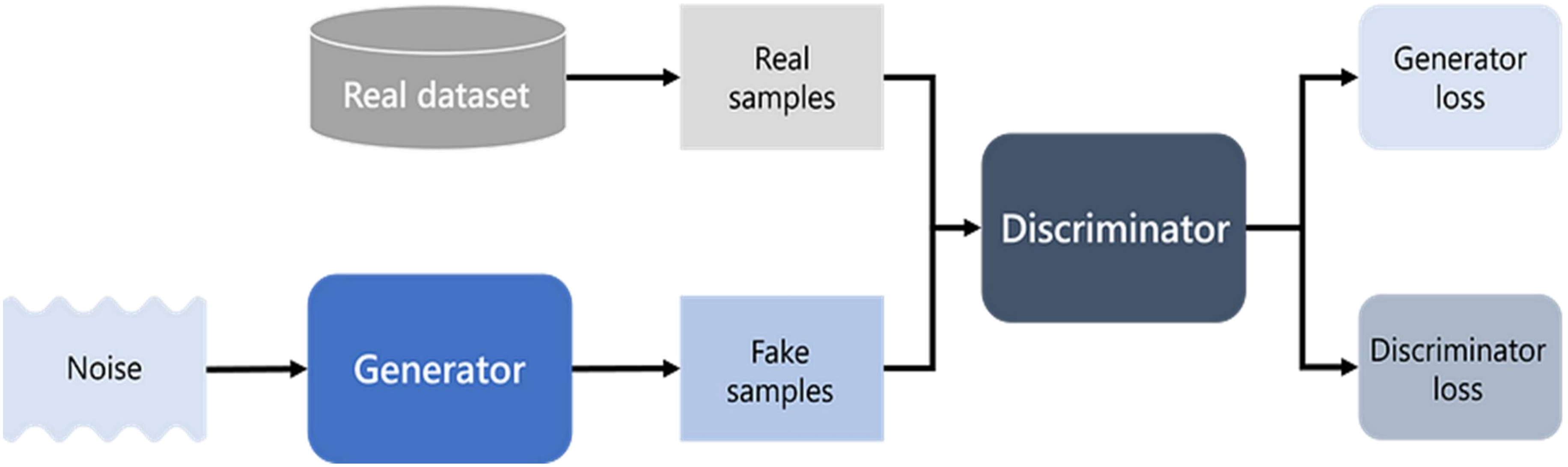
GENERATIVE ADVERSARIAL NETWORKS (GAN)

- Deep learning technique by Ian Goodfellow (2014)

- Used for realistic data generation

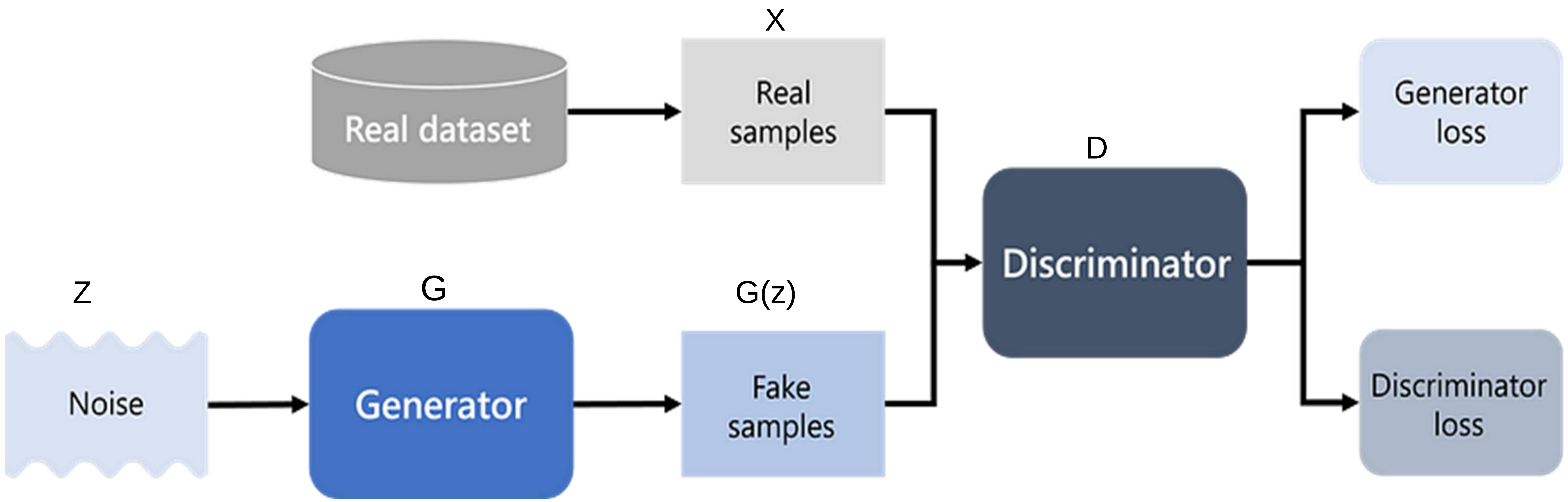
- Applications: deepfakes, data augmentation, etc.





GHENT UNIVERSITY

GENERATIVE ADVERSARIAL NETWORKS (GAN)



$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

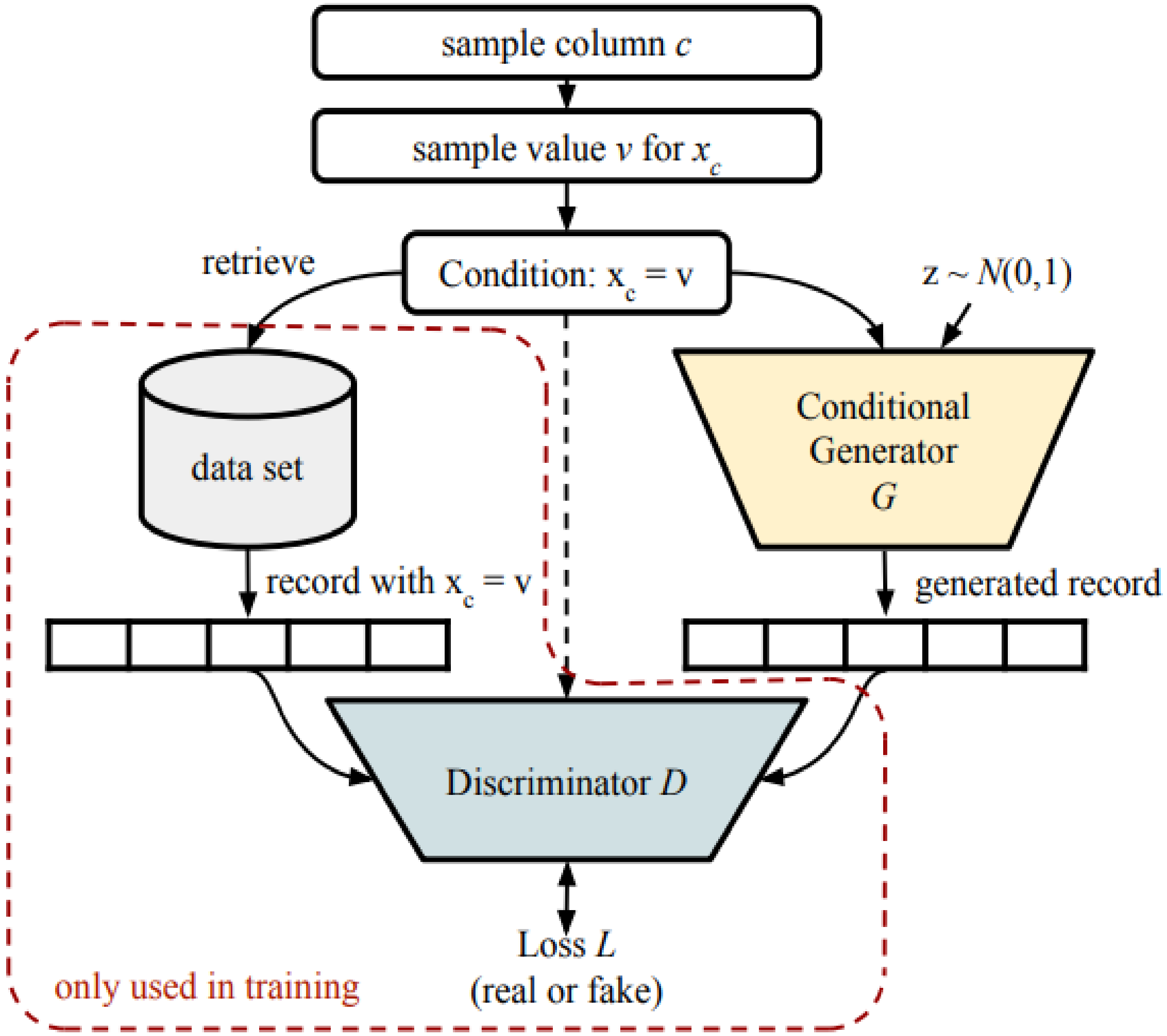
GHENT UNIVERSITY

CTGAN

- CTGAN is optimized for **tabular data** (often a mix of categorical and continuous variables)

- CTGAN uses a **conditional generator** which makes it better in capturing dependencies between features.

- In Python: CTGANSythesizer from **SDV package**



GHENT UNIVERSITY

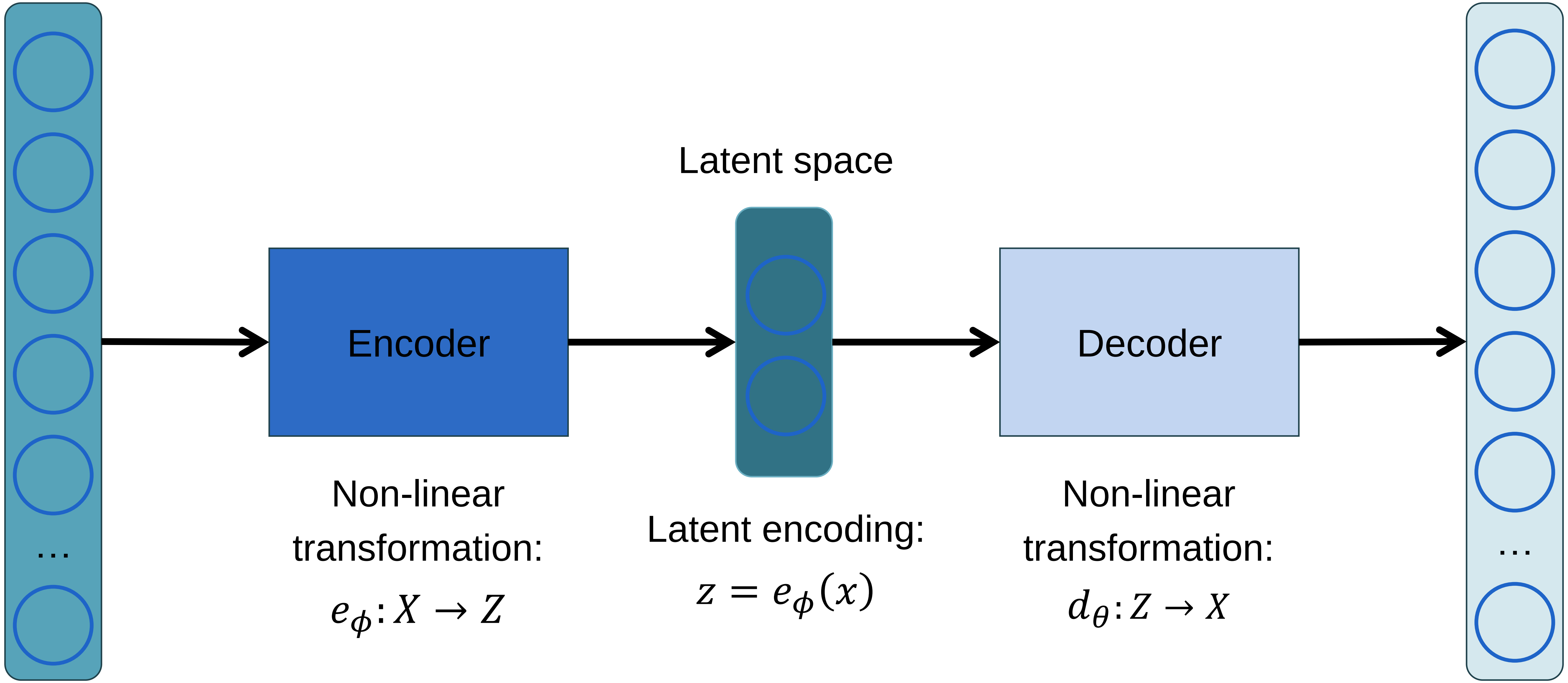
VARIATIONAL AUTOENCODERS (VAE)

First proposed by Kingma & Welling, 2013

GHENT UNIVERSITY

AUTO- ENCODER

$$f(x) = d_\theta(e_\phi(x))$$



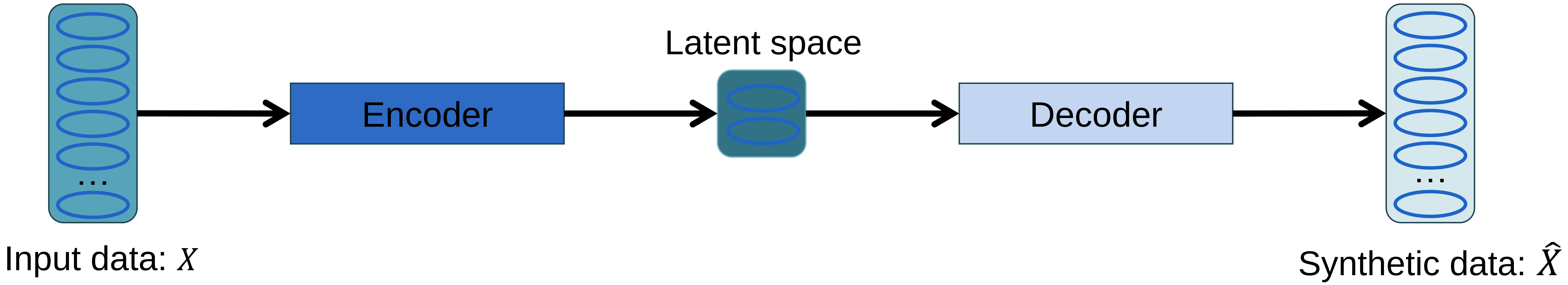
Input data: X

Synthetic data: $\hat{X}$

$\hat{x} = d_\theta(z)$

GHENT UNIVERSITY

AUTO-ENCODER



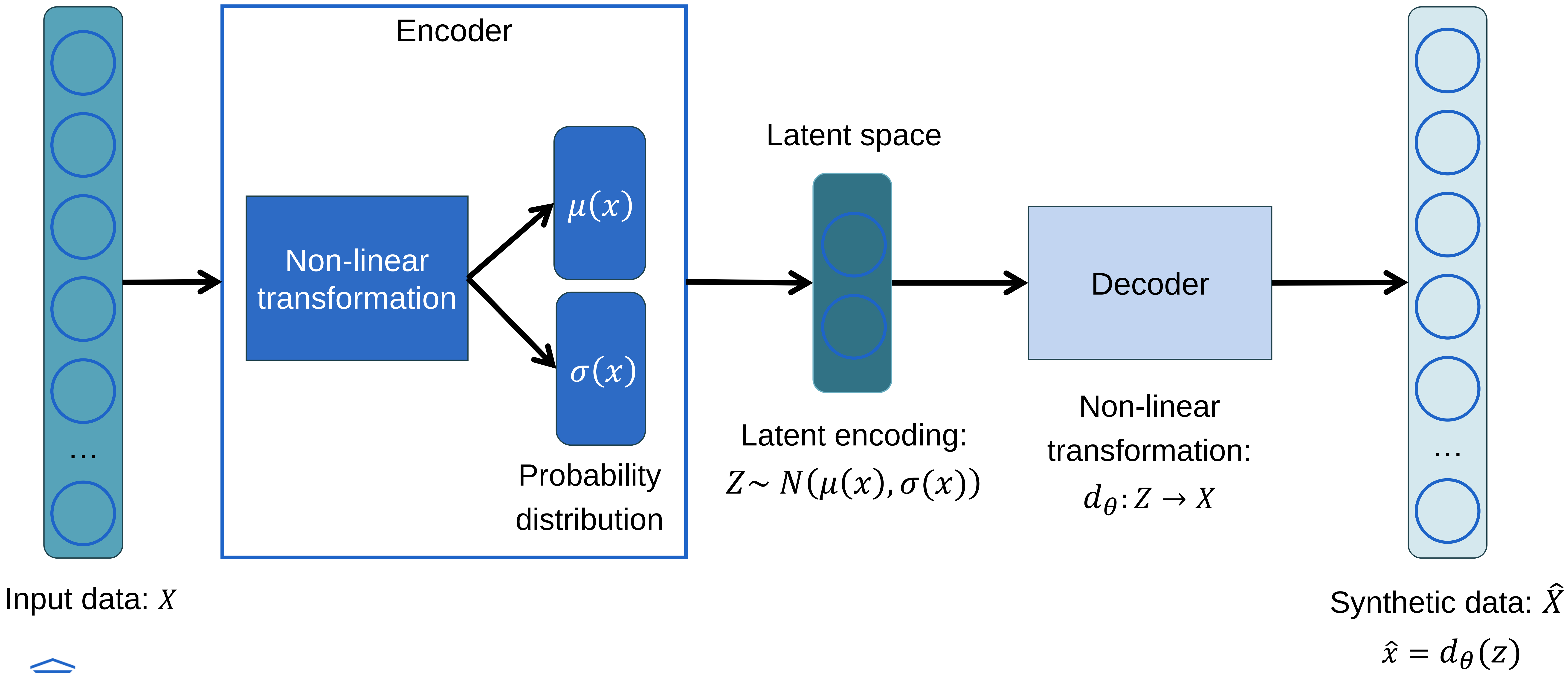
Challenges auto-encoder:

- Similar samples not necessarily close to each other in the latent space
- Realistic outcomes not guaranteed

--> Variational auto-encoder

GHENT UNIVERSITY

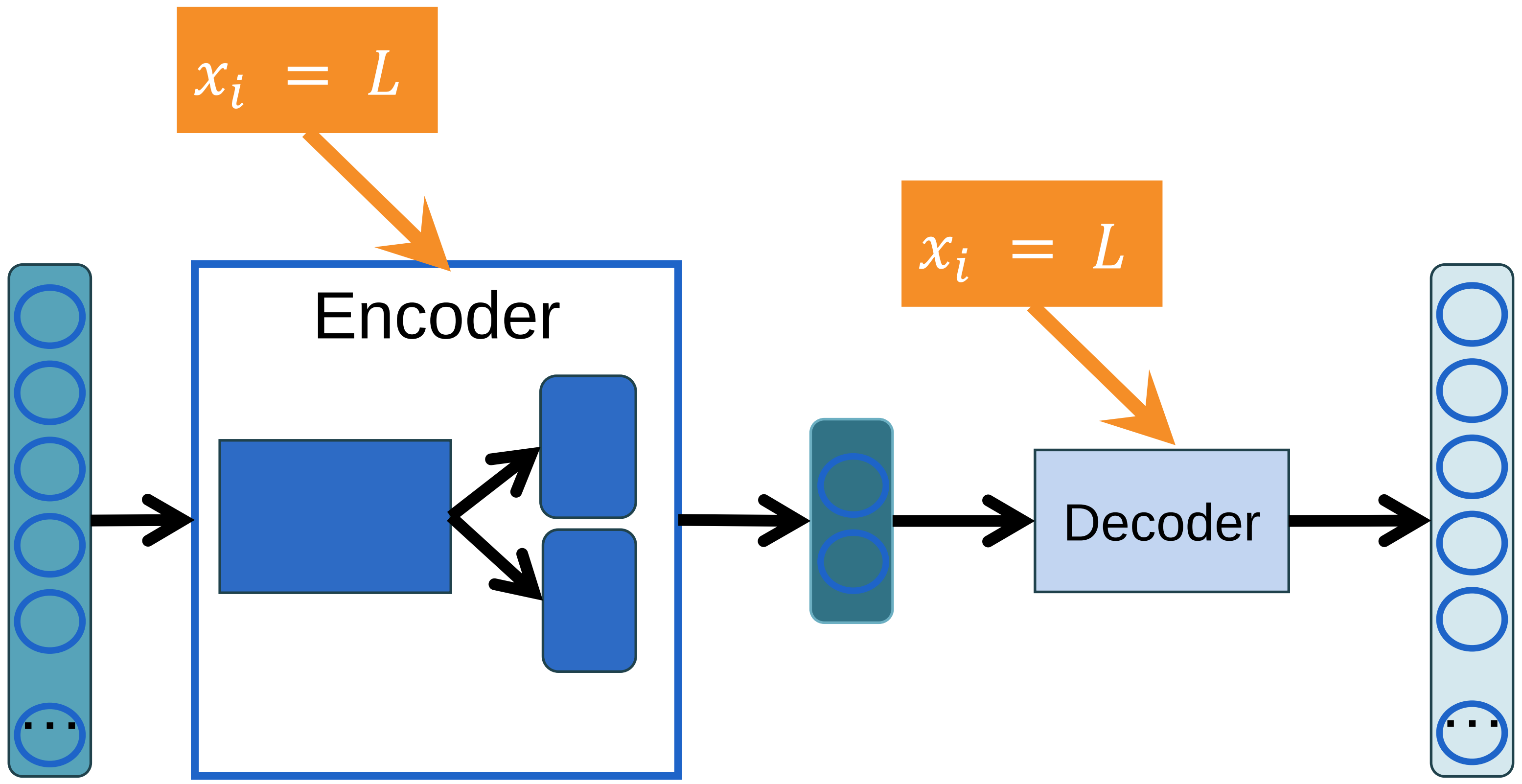
VARIATIONAL AUTO-ENCODER



GHENT UNIVERSITY

POSSIBLE EXTENSIONS

– CVAE *(Doersch, 2021)*
  - Fill gaps in existing entries
  - Condition the model on input

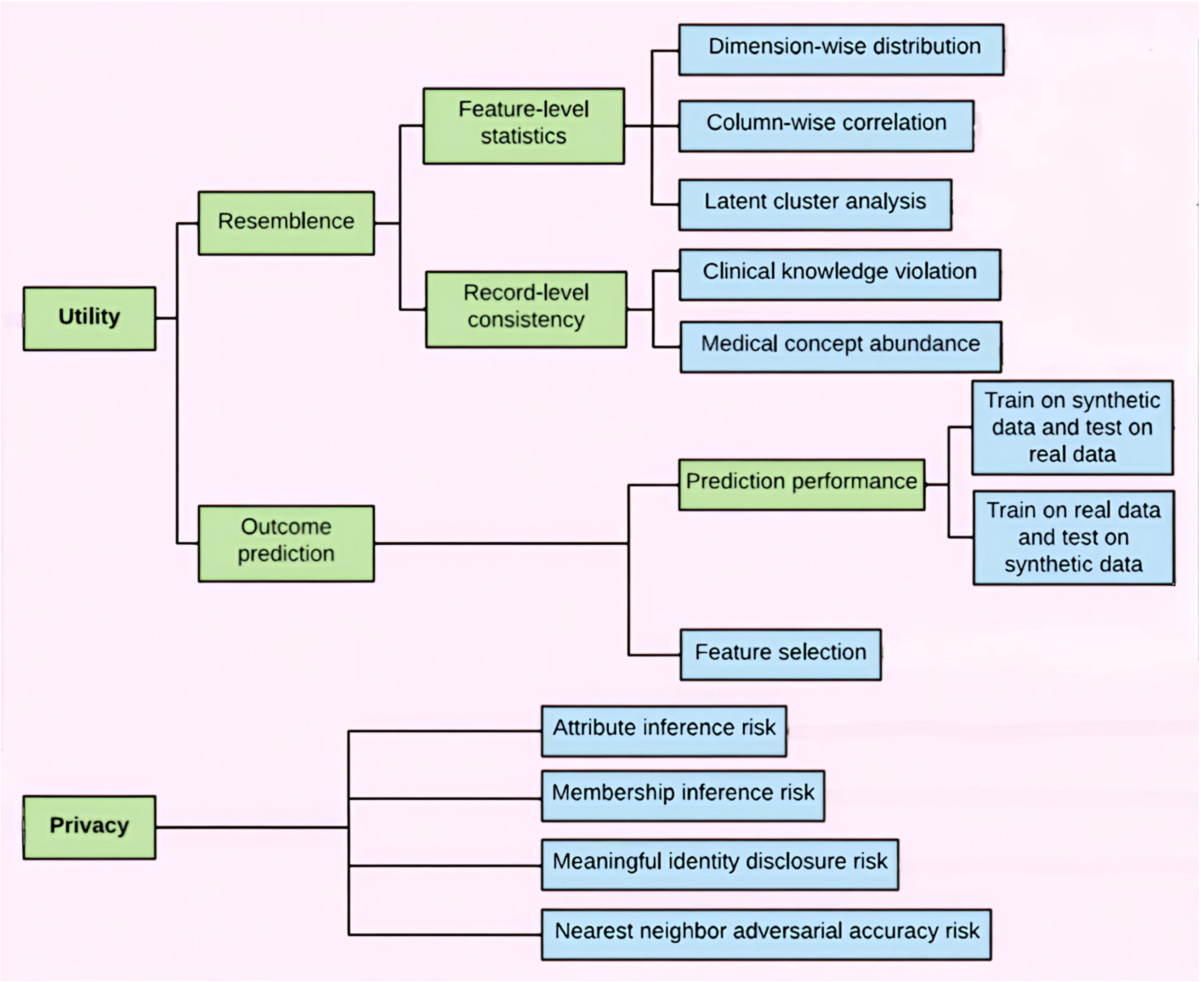


– TVAE *(Xu et al. 2019)*
  - Tabular data
  - SDV python package

EVALUATION OF SYNTHETIC DATA

EVALUATION OF SYNTHETIC DATA

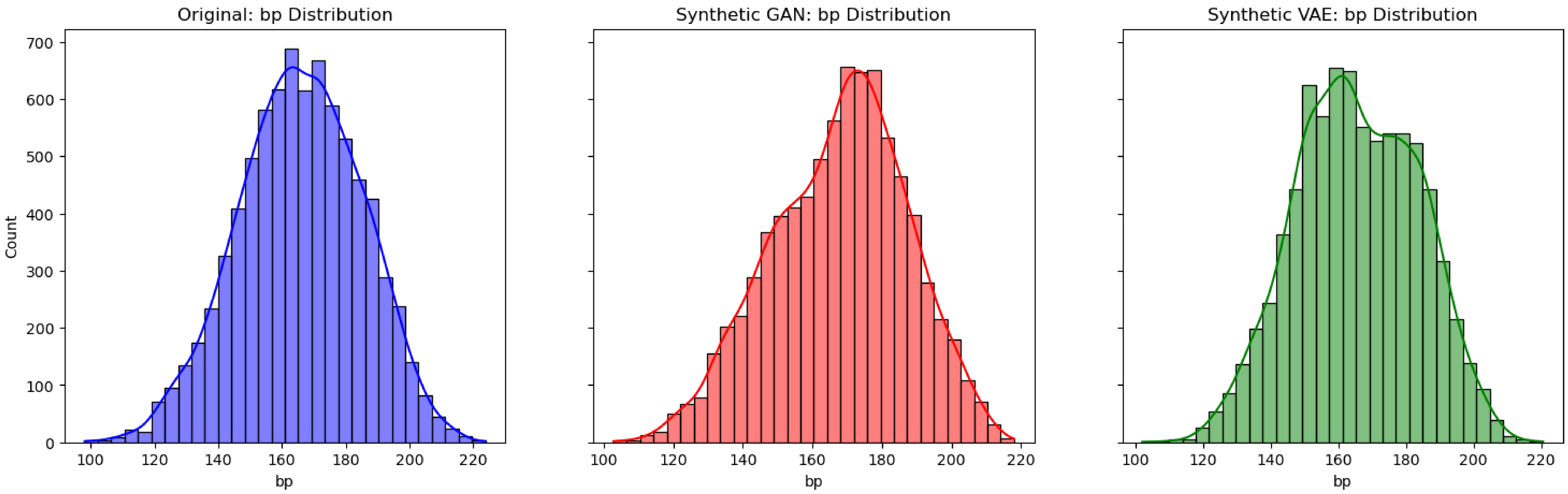
- Univariate / Bivariate
- Utility (e.g. prediction performance)
- Privacy (e.g. MIA)



GHENT UNIVERSITY

UNIVARIATE

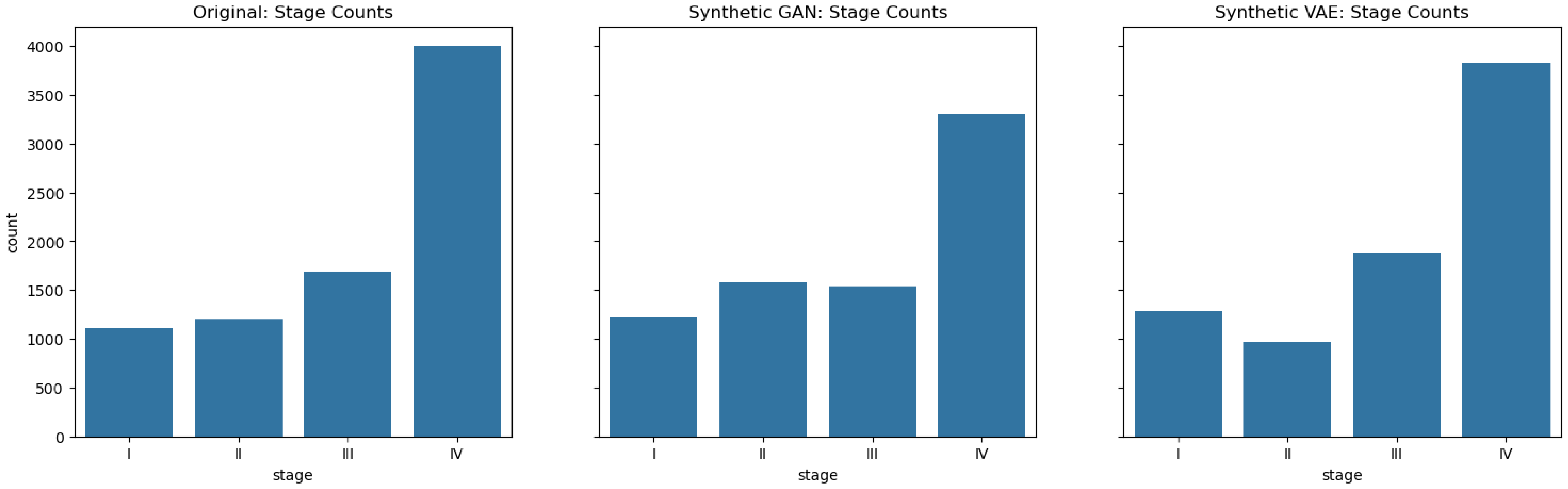
How similar are distributions?



GHENT UNIVERSITY

UNIVARIATE

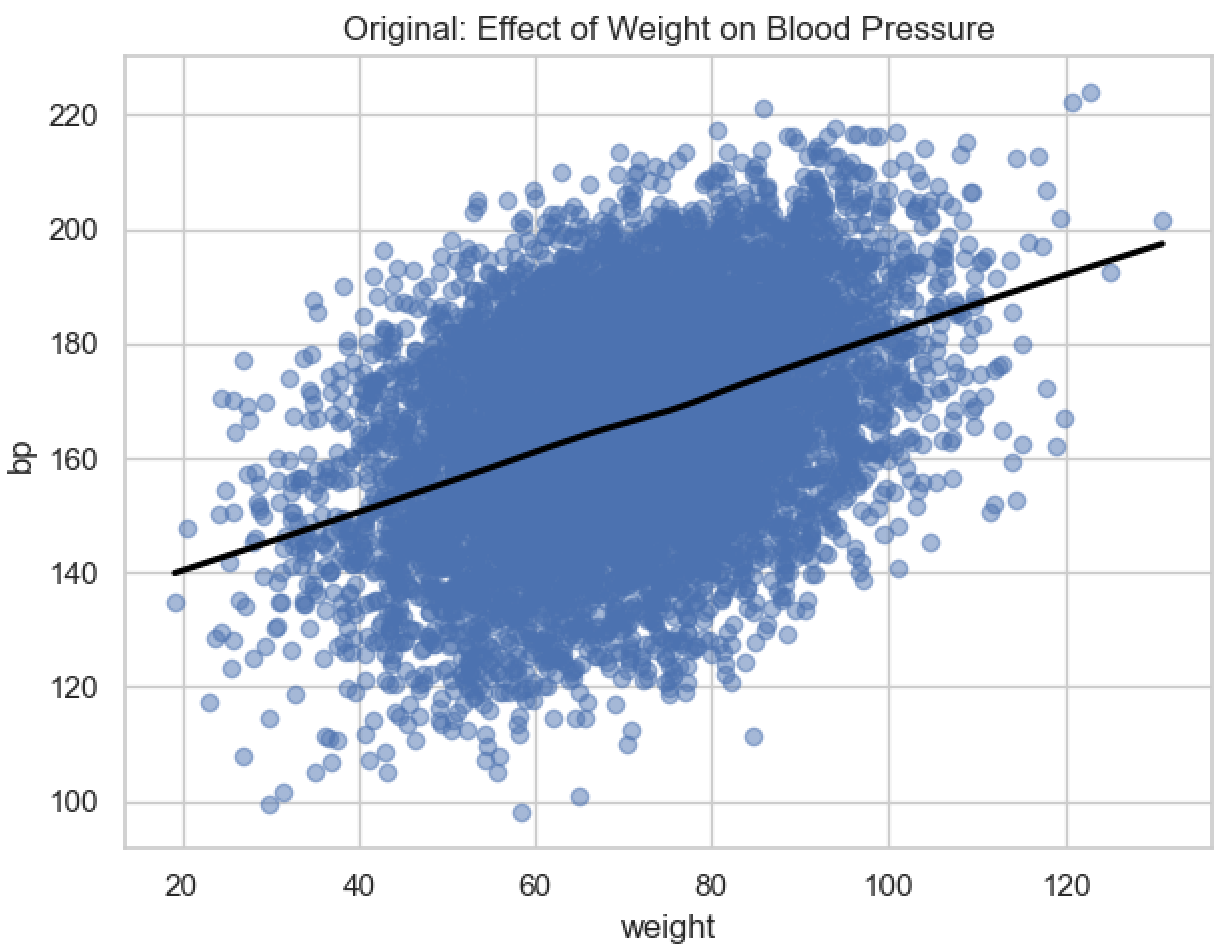
How similar are distributions?



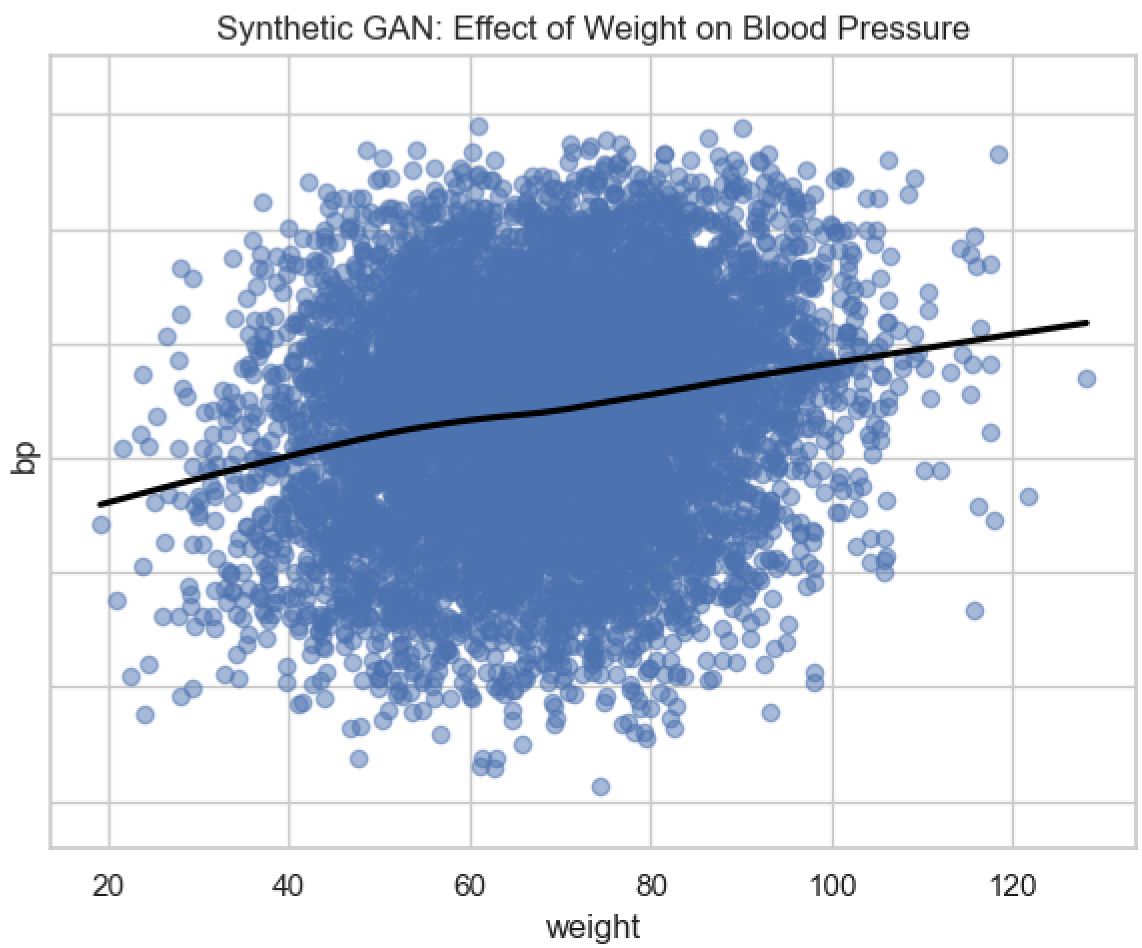
GHENT UNIVERSITY

BIVARIATE

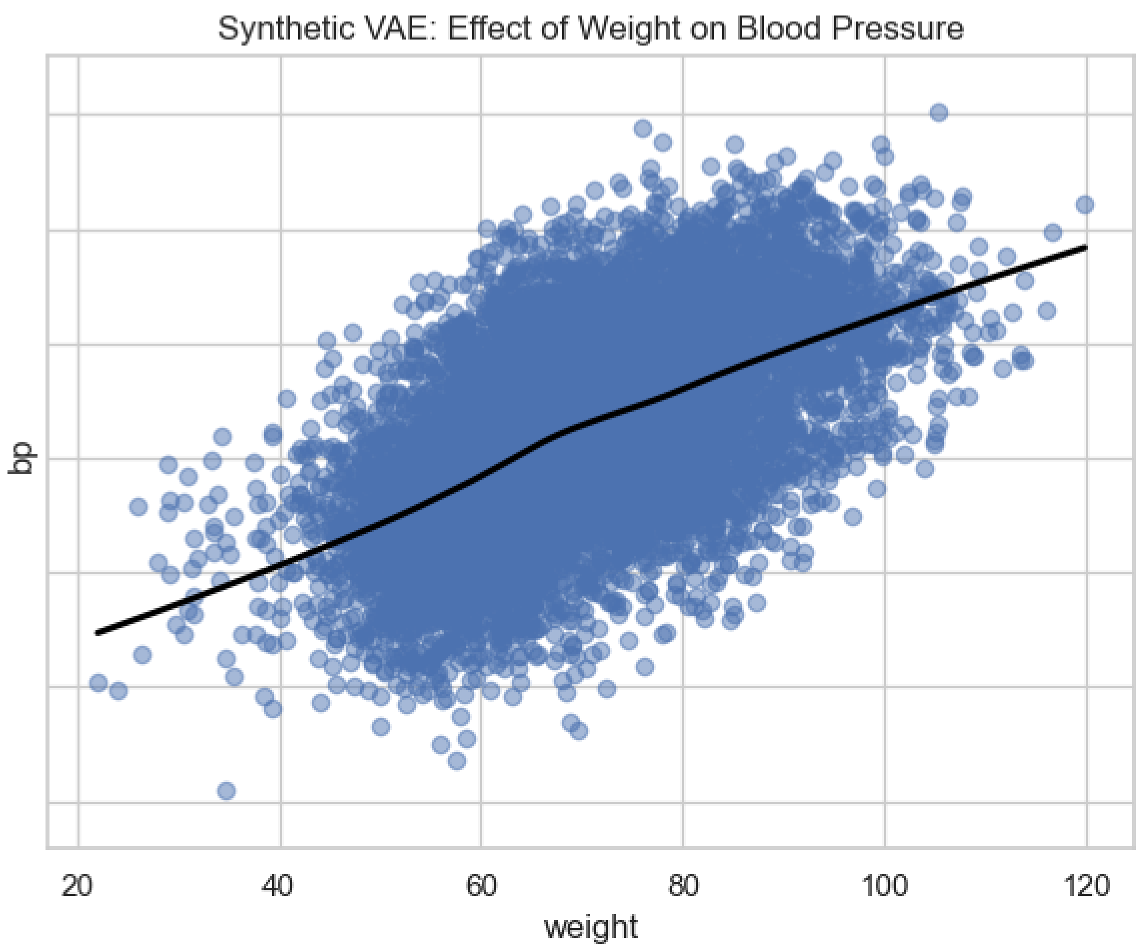
How similar are dependencies?



Cor(x,y) = 0.40



Cor(x,y) = 0.18

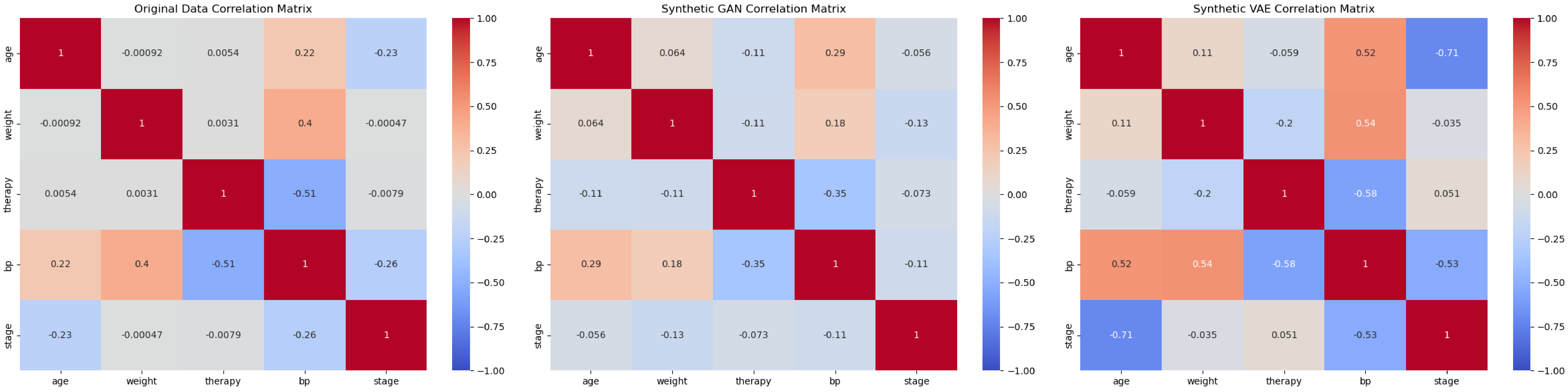


Cor(x,y) = 0.54

GHENT UNIVERSITY

BIVARIATE

How similar are dependencies?



GHENT UNIVERSITY

UTILITY

Predictive performance

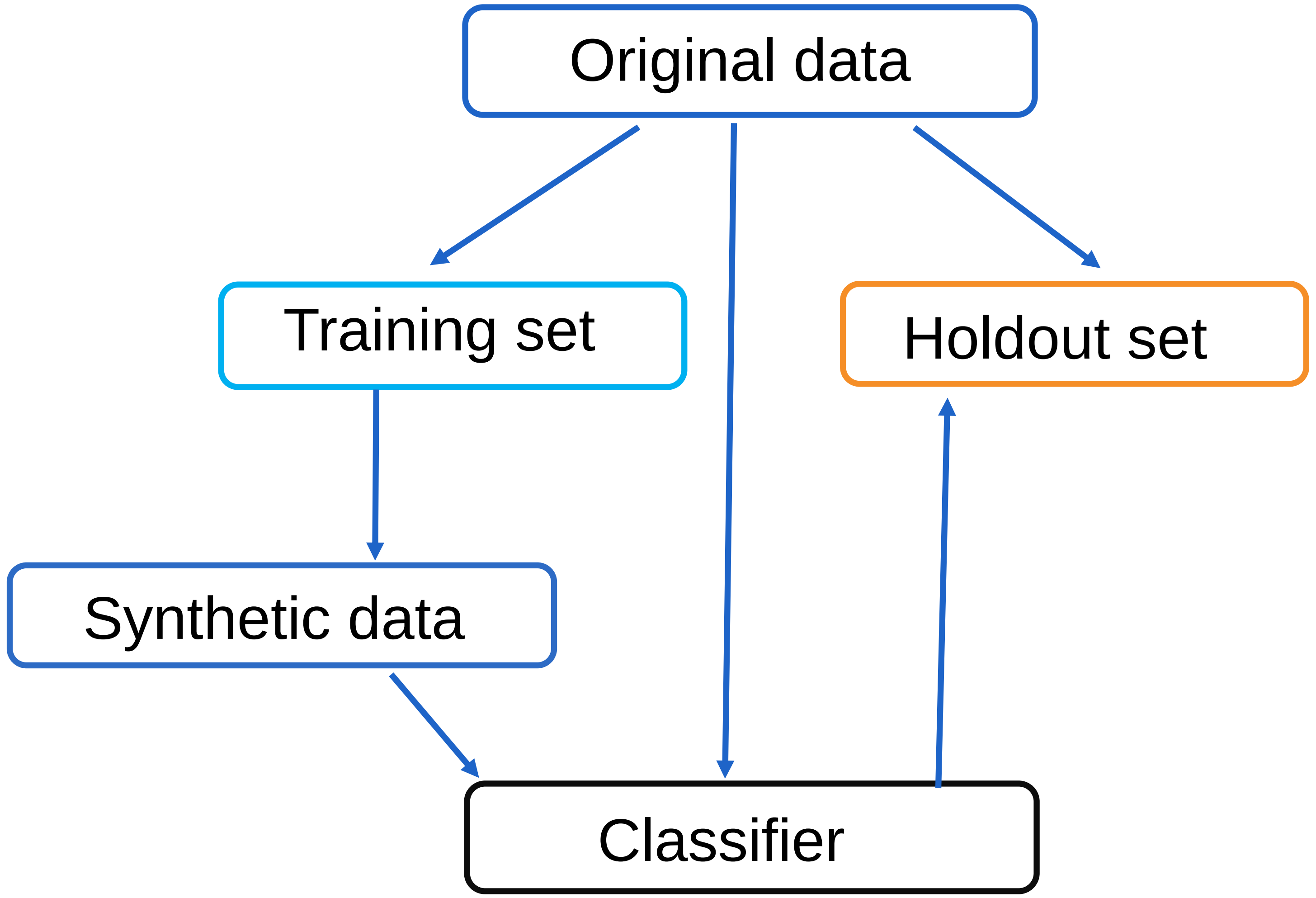
	Original	Synthetic CTGAN	Synthetic TVAE
MSE	95.66	203.19	119.03
Adjusted R^2	0.74	0.45	0.68

Multiple linear regression to predict bloodpressure

PRIVACY

Membership Inference Attack

How well an attacker can determine if a specific individual's data was used to train the synthetic data generator.



	Original	Synthetic CTGAN	Synthetic TVAE
Accuracy	0.50	0.72	0.62

GHENT UNIVERSITY

CONCLUSION

- Synthetic data is promising for applications with sensitive data
- Multiple techniques exist (e.g. GAN, VAE)
- Difficult to evaluate
  - Trade-off
  - No standardized method

FACULTY OF SCIENCES

Frie Van Bauwel, Marcin Jedrych, Xueting Li

Github: https://github.com/marcinjedrych/Project-BDA.git

Sources:

- Doersch, C. (2016). *Tutorial on Variational Autoencoders*. http://arxiv.org/abs/1606.05908
- Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *Npj Digital Medicine*, *6*(1). https://doi.org/10.1038/s41746-023-00927-3
- Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. In *Neurocomputing* (Vol. 493, pp. 28–45). Elsevier B.V. https://doi.org/10.1016/j.neucom.2022.04.053
- Jamotton, C., & Hainaut, D. (2024). Variational AutoEncoder for synthetic insurance data. *Intelligent Systems with Applications*, *24*. https://doi.org/10.1016/j.iswa.2024.200455
- Kingma, D. P., & Welling, M. (2013). *Auto-Encoding Variational Bayes*. http://arxiv.org/abs/1312.6114
- Mohammadi, M. (2021, June 15). *Synthetic data generation using Generative Adversarial Networks (GANs): Part 2.* from Medium: https://medium.com
- Patki, N., Wedge, R., & Veeramachaneni, K. (2016b). The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410. https://doi.org/10.1109/DSAA.2016.49
- Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling Tabular data using Conditional GAN*. http://arxiv.org/abs/1907.00503
- Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., Mooney, S. D., & Malin, B. A. (2022). A Multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-35295-1/doi.org/10.1109/DSAA.2016.49

GHENT UNIVERSITY