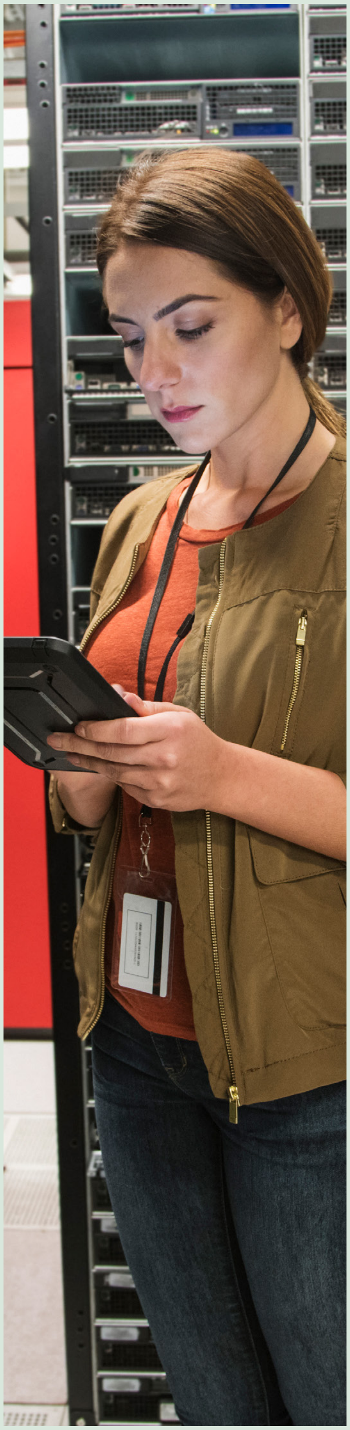
Infosys topaz

WHITE PAPER

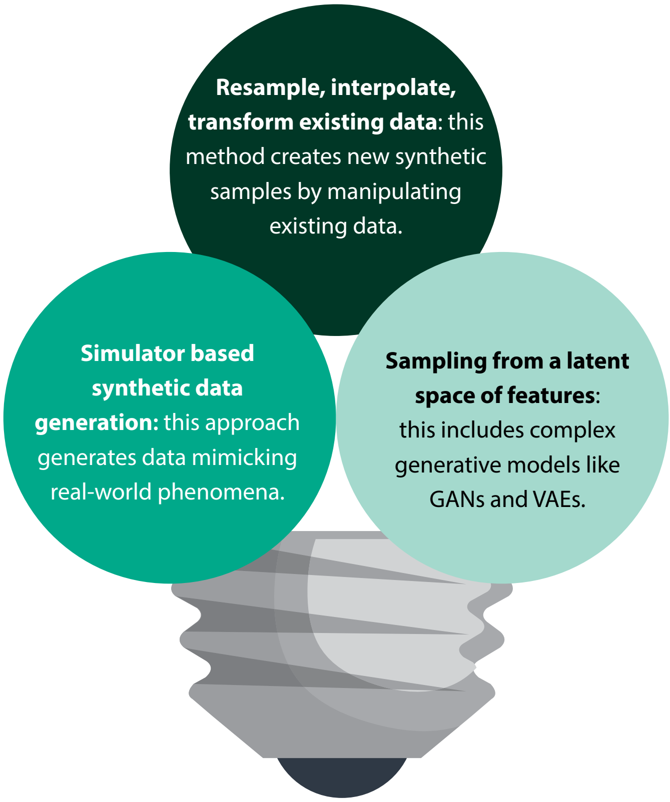


# SYNTHETIC DATA GENERATION: WHAT, WHERE AND HOW?

Infosys®
Navigate your next

In recent years, the importance of labeled data for training machine learning models has persisted despite advancements in self-supervised learning. Manual labeling is still labor-intensive and costly. Synthetic data generation [1][2] offers a promising solution by creating labeled data that reduces the need for manual efforts. This technique helps address the challenges associated with acquiring and labeling large datasets needed for specialized tasks.

There is, however, no single method for synthetic data generation that is universally applicable across all situations. Broadly, synthetic data generation can be categorized into three main types:



**Resample, interpolate, transform existing data**: this method creates new synthetic samples by manipulating existing data.

**Simulator based synthetic data generation:** this approach generates data mimicking real-world phenomena.

**Sampling from a latent space of features**: this includes complex generative models like GANs and VAEs.

In this white paper, we will discuss these three types of synthetic data generation techniques together with their suitability for applications as well as their shortcomings. Thereafter we will briefly look at the aspect of using LLMs (large language models) for purposes of training other smaller or specialized LLM models. Finally, we will delve into problems that often arise with synthetic data together with techniques for evaluation of related quality aspects of data generated.

## Simulator Based Synthetic Data Generation

Many machine learning environments require voluminous data that might be difficult or hazardous or simply too expensive to acquire. Let us look at some typical examples:

1. Training autonomous driving models requires data on rare dangerous events like car collisions and drifts.
2. Gas leaks in domestic scenarios are rare, making it hard to represent all potential variations.
3. Creating financial fraud instances from real examples yields only a small, specialized subset of possibilities.
4. Manually labeling data for object recognition in stores is prohibitively expensive and time-consuming.

Under these kinds of circumstances, a **physical model based on a set of rules or dynamics** that adequately represent the environment or the system for which the machine learning model is constructed can provide for a simulator that matches real world data [3][4]. In the recent past we have successfully generated a gas leak model based on the physical dynamics of a leaking gas pipe while 3D models generated from a few images of individual products have been used to generate large volumes of synthetic data for object recognition.

A simulator can generate vast, varied synthetic datasets once built, covering many cases. However, they need to be custom-built; furthermore, designing a simulator requires deep understanding of system dynamics and may not always capture the full complexity of real-world scenarios.

This implies that simulator-based data generation is ideal when we can fully model the simulated system. Also, before using a simulator to create synthetic data, it must be validated to ensure the data aligns with physical principles and boundary conditions.

## Resample, Interpolate, Transform Existing Data

In situations where the training of a machine learning model requires way more data than is available, we can generate a larger data set through oversampling or transformation [5] of the original samples. Some typical situations are:

1. Customer churn detection or anomaly detection where the training data is usually imbalanced -- that is one of the classes has very few representative samples (maybe 5% of the overall data set).
2. Healthcare data sets, where privacy concerns prevail in addition to insufficiency of data for certain categories.
3. Applications based on speech, text or image data lacking enough variations required to train robust models.

One prominent technique frequently used to address imbalanced datasets is the Synthetic Minority Over-sampling Technique **(SMOTE)** [6][7] applicable to multivariate numerical data. SMOTE generates synthetic samples for the minority class by interpolating between existing samples, effectively balancing the dataset. Its variant **DeepSMOTE** [8] extends the interpolation functionality with a deep learning model to generate image data for medical imaging applications.

SMOTE and its variants like DeepSMOTE are robust, computationally efficient and can work with a small initial data set. However, biased or low-quality input data can result in generation of similarly flawed synthetic data. Additionally, SMOTE can occasionally generate out-of-distribution synthetic samples, thereby requiring filtering of generated data.

**Data augmentation** techniques, widely used in both computer vision and natural language processing, enhance diversity in existing data by introducing variations such as rotations, scaling, and cropping for images, synonym replacement and back-translation for text, occasionally adding a controlled amount of noise for robust modeling or source obfuscation (as with healthcare data).

Data augmentation can significantly enhance model performance and generalization capabilities in a cost-effective manner. However, the transformations in data augmentation need to be appropriate, otherwise distorted output data can result. Augmented data therefore requires validity checking, for example, to ensure that outputs are grammatically correct and semantically similar for paraphrased or translated sentences.

Prompting **diffusion model-based image generators** to generate similar images to a given image can also be regarded as an augmentation technique. However, unlike mainstream data augmentation techniques, validity checking, lacking objective criteria, becomes difficult if not impossible to automate.

## Sampling From a Latent Space of Features

Sometimes there might be sufficient data for an application to train a generative model, whereafter the same may be used to synthesize voluminous and varied data on lines of a simulator, e.g. synthesizing health records, data for self-driving cars, even tabular data.

The method here typically involves sampling from a probability distribution followed by transformation using neural networks, as in Generative Adversarial Networks (**GAN**s) and Variational Autoencoders (**VAE**s) [9][10]. Both methods in an abstract sense involves sampling from a latent representation space of features. GAN consists of a generator that synthesizes data and a discriminator that performs a reality check; this adversarial process allows GANs to create high-quality outputs that closely resemble real data. Conversely, VAEs encode input data into a latent space and then decode it, ensuring that the generated samples maintain the statistical properties of the original dataset.

While these approaches can generate diverse and high-fidelity datasets, they also present challenges. GANs can be complex to train and prone to issues like mode collapse, where the generator produces limited variations. VAEs, while generally more stable, may yield noisier outputs compared to GANs. Both need a fair amount of quality training data as well, because of which these techniques are typically limited to synthesizing data for areas where this applies.

## LLM Based Synthetic Data Generation

Large Language Models (**LLM**s) revolutionize synthetic data generation by producing text that mirrors real-world patterns [11][12][13]. Their instruction-following capabilities allow for the generation of tailored datasets, enhancing controllability and minimizing human effort. This automation facilitates scalability and efficiency, exemplified by over 300 synthetic datasets available on platforms like Hugging Face as of June 2024.

LLMs utilize two main strategies for synthetic data generation: **prompt engineering** and **multi-step generation**. Prompt engineering involves crafting effective prompts that include task specifications, generation conditions, and in-context demonstrations. In contrast, multi-step generation breaks complex tasks into simpler sub-tasks, allowing for iterative data production.

LLM based data synthesis suffers from challenges such as noisy or toxic samples. On one hand, this necessitates robust **data curation** strategies, including sample filtering using heuristic metrics to remove low-quality samples and label enhancement to correct erroneous annotations. On the other hand, **evaluating the generated data** in terms of direct assessments of data quality and indirect evaluations based on downstream model performance is crucial.

Key concerns about using LLM-based data generation for training smaller or domain-specific models include the potential for low-quality or biased data, lack of domain expertise, and challenges in evaluating generated content. Additionally, issues related to data privacy and insufficient diversity sometimes occur.

## Quality Issues and Evaluation Methods

The quality and validity of synthetic data are paramount, especially in sectors like healthcare and banking. Evaluating synthetic data [9][14][15] involves three main aspects: fidelity, utility, and privacy.

**Fidelity** assesses how well synthetic data mirrors real data through statistical comparisons and visual analyses. Techniques such as histograms, correlation plots, and metrics like KL Divergence and SSIM are employed. In addition, discriminators (as in GANs) wherever available, can also be used to assess data fidelity.

**Utility** evaluation involves comparing the performance of machine learning models trained on synthetic versus real data. Methods like Training on Real Data and Testing on Real Data (TRTR) and Training on Synthetic Data (TSTR) provide insights. Metrics such as GAN-train and GAN-test are also useful. For text data, ROUGE and BLEU scores are relevant for ensuring content invariance, while privacy-utility trade-off metrics balance utility with privacy concerns.

**Privacy** risks are significant especially in areas like healthcare with data potentially containing Personally Identifiable Information (PII). Re-identification attacks, where adversaries match synthetic to original data, highlight risks. Differential privacy methods, which add noise to the data generation process, offer strong privacy guarantees. K-anonymity ensures records remain indistinguishable among a group, though it might not be sufficient alone.

Summarily, generating high-quality synthetic data requires rigorous evaluation to maintain fidelity, utility, and privacy. Robust evaluation methods and metrics ensure synthetic data is useful and secure, protecting individual confidentiality while maintaining data usability.

## Conclusion

In conclusion, it is crucial to recognize that synthetic data generation does not adhere to a "one size fits all" paradigm. The choice of generation techniques and evaluation metrics must be customized to address the unique challenges of each domain, ensuring the production of reliable and effective synthetic datasets. Methods have their respective pros and cons and cost-benefit trade-offs. In all cases, the generated data should be rigorously tested for fidelity, utility, and privacy prior to use.

## References

1. Chang, Hsin-Yu, Pei-Yu Chen, Tun-Hsiang Chou, Chang-Sheng Kao, Hsuan-Yun Yu, Yen-Ting Lin, and Yun-Nung Chen. "A Survey of Data Synthesis Approaches." arXiv, 2024.
2. Wang, Ke, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. "A Survey on Data Synthesis and Augmentation for Large Language Models." arXiv, 2024.
3. De Melo, Celso, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. "Next-generation deep learning based on simulators and synthetic data." Trends in Cognitive Sciences 26, no. 2 (2022): 174-187.
4. Dankar, F.K., and M. Ibrahim. "Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation." Applied Sciences 11, no. 5 (2021): 2158.
5. Gandhi, Saumya, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. "Better Synthetic Data by Retrieving and Transforming Existing Datasets." In Findings of the Association for Computational Linguistics: ACL 2024, 6453–6466. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024.
6. Fernández, Alberto, Salvador García, Francisco Herrera, and Nitesh V. Chawla. "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary." Journal of Artificial Intelligence Research 61 (2018): 863-905.
7. Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of Artificial Intelligence Research 16 (2002): 321-357.
8. Dablain, Damien, Bartosz Krawczyk, and Nitesh V. Chawla. "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data." arXiv, 2021.
9. Figueira, Alvaro, and Bruno Vaz. "Survey on synthetic data generation, evaluation methods and GANs." Mathematics 10, no. 15 (2022): 2733.
10. Lu, Yingzhou, Meng Shen, Huazheng Wang, Xiaoyang Wang, Capucine van Rechem, Tian Fu, and Wenqi Wei. "Machine learning for synthetic data generation: a review." arXiv preprint arXiv:2302.04062 (2023).
11. Veselovsky, Veniamin, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. "Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science." arXiv, 2023.
12. Tan, Zhen, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. "Large Language Models for Data Annotation: A Survey." arXiv, 2024.
13. Long, Lin, Ruixuan Xiao, Junbo Zhao, Xiaodong Ding, Guojun Chen, and Haobo Wang. "On LLMs-driven synthetic data generation, curation, and evaluation: A survey." arXiv preprint arXiv:2406.15126 (2024).
14. Osorio-Marulanda, Paula Andrea, Gorka Epelde, Mikel Hernandez, Imanol Isasa, Natalia Maria Reyes, and Wenqi Wei. "Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review." IEEE Access (2024).
15. Livieris, Ioannis E., Nikos Alimpertis, George Domalis, and Dimitris Tsakalidis. "An Evaluation Framework for Synthetic Data Generation Models." arXiv, 2024.

## Authors

**Avina Datta**, Infosys,

**Dr. Puranjoy Bhattacharya**, Infosys

**Sudarshan Gopalan**, Infosys

For more information, contact askus@infosys.com

Infosys®
Navigate your next