zensar

Artificially Generating Structured Test Data

Using Generative AI for Maximizing Testing Coverage and Avoiding Data Privacy Challenges

An RPG Company

Executive Summary

This white paper explores the use of generative AI in creating additional structured test data to maximize testing coverage during migration from legacy systems to new advanced platforms. Generative AI is a specialized class of AI method that can generate new data similar to existing data which can be trained to generate synthetic data that appears realistic. Generative AI can be used across industries to generate synthetic samples in an adversary setting, allowing it to understand the interaction between different fields of data and create realistic synthetic samples. The paper also highlights Zensar's approach to testing new systems using generated synthetic data and comparing the results with legacy systems.

Maximizing test coverage with generative AI

Businesses are constantly upgrading to more advanced platforms to meet evolving operational needs. However, migrating data from one platform to another can be tedious and time-consuming and may involve several challenges. Some of the biggest challenges are gaining buy-in from stakeholders and ensuring data integrity, wherein data integrity also affects how you may impact the stakeholder buy-in. Multiple stakeholders from the businesses can be onboarded with the migration activity if we can ensure a seamless data migration solution that doesn't hamper the integrity of the data and is scalable and cost-effective, among other aspects.

This white paper explores the possibilities of generating additional structured test data, similar to historical data from legacy platforms or systems, to maximize testing coverage.

Massive test data that covers all possible scenarios and boundary conditions are required to compare the performance of the newer and more advanced systems to the legacy systems. The problem arises when the number of variables increase and the number of possible test cases grow exponentially.

Here generative AI comes to the rescue to create synthetic data based on the intrinsic patterns in the legacy system. With generative AI, those patterns can be maintained and retained in the synthetic test data, and all the possible test scenarios can be covered while testing the new advanced system.

What is Generative AI

Generative AI is a specialized class of method in AI designed to generate new data similar to existing data. It can learn the generation of synthetic data from random noise. Generative Adversarial Network (GAN) is a specific set of models that work on the principle of having an adversary setting defined when training the model. During model training, this setting assesses the quality of generated samples.

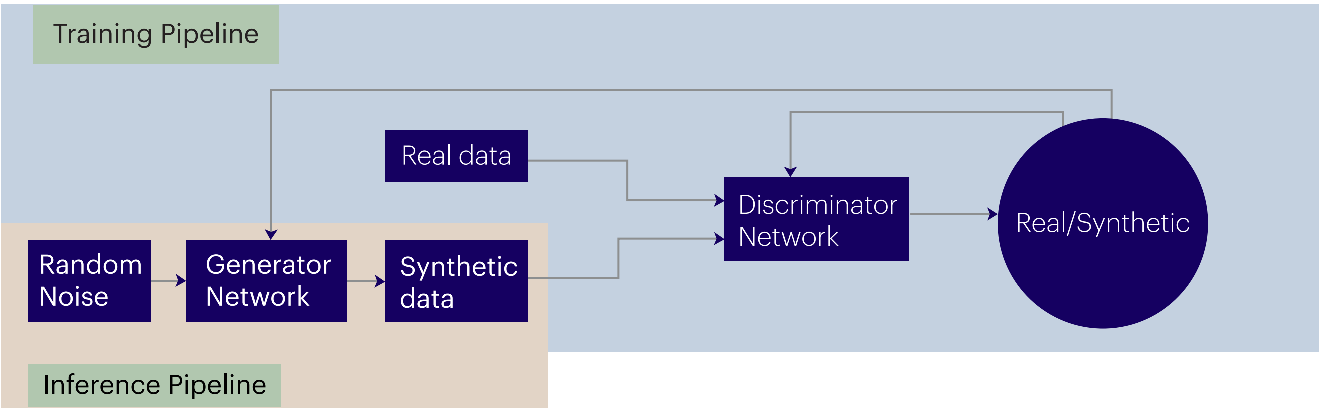


Figure 1: GAN overview proposed by GoodFellow

Figure 1 demonstrates the overview of GAN as proposed by GoodFellow et al.[1] Here, two networks, i.e., generator and discriminator, are trained with the help of actual data, including audio, image, text, or other forms of data.

The input random noise is transformed into synthetic data by the generator network. This synthetic data and real-world samples are input into the discriminator network to discriminate between real and synthetic samples. The errors made by the discriminator network in correctly identifying real and synthetic samples are used as feedback for both generator and discriminator models during the training stage.

After the training is complete, the generator network is expected to convert the noise to realistic appearing synthetic samples, especially in cases that the discriminator network finds challenging to distinguish from the real samples.

When the trained GAN model is put into real-world usage, it uses only the trained generator network and takes the generated synthetic samples for the downstream task under consideration. No further training of the model can occur beyond this point. The existing research supports that the GAN models display state-of-the-art performance for generating a variety of modalities of data, including text, audio, images, and videos. [2][3][4]

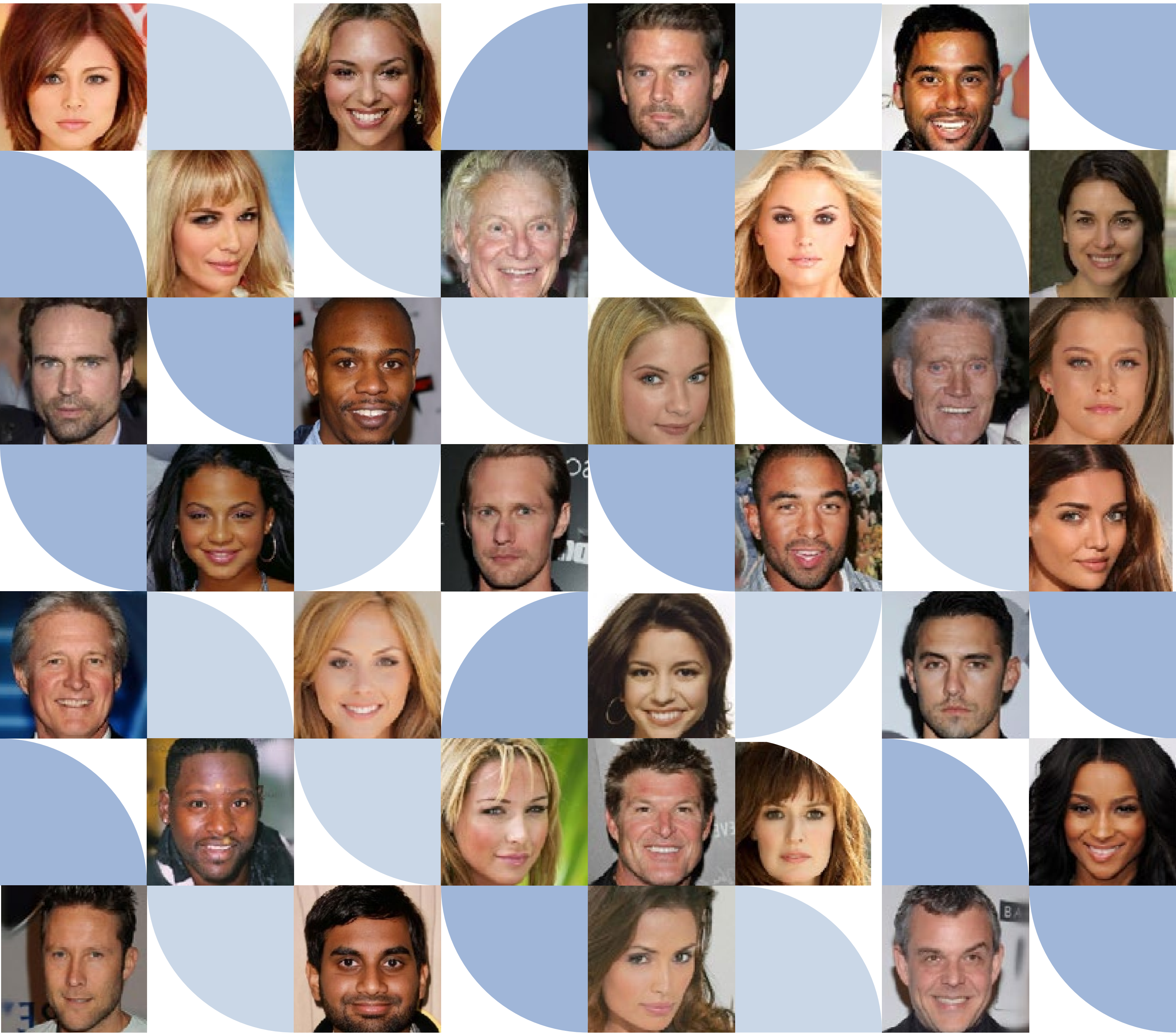


Figure 2: Faces generated by GAN[5]

The faces in Figure 2 belong to people who actually don't exist but have been generated by GAN. They appear realistic and of real people. Such is the power of GAN.

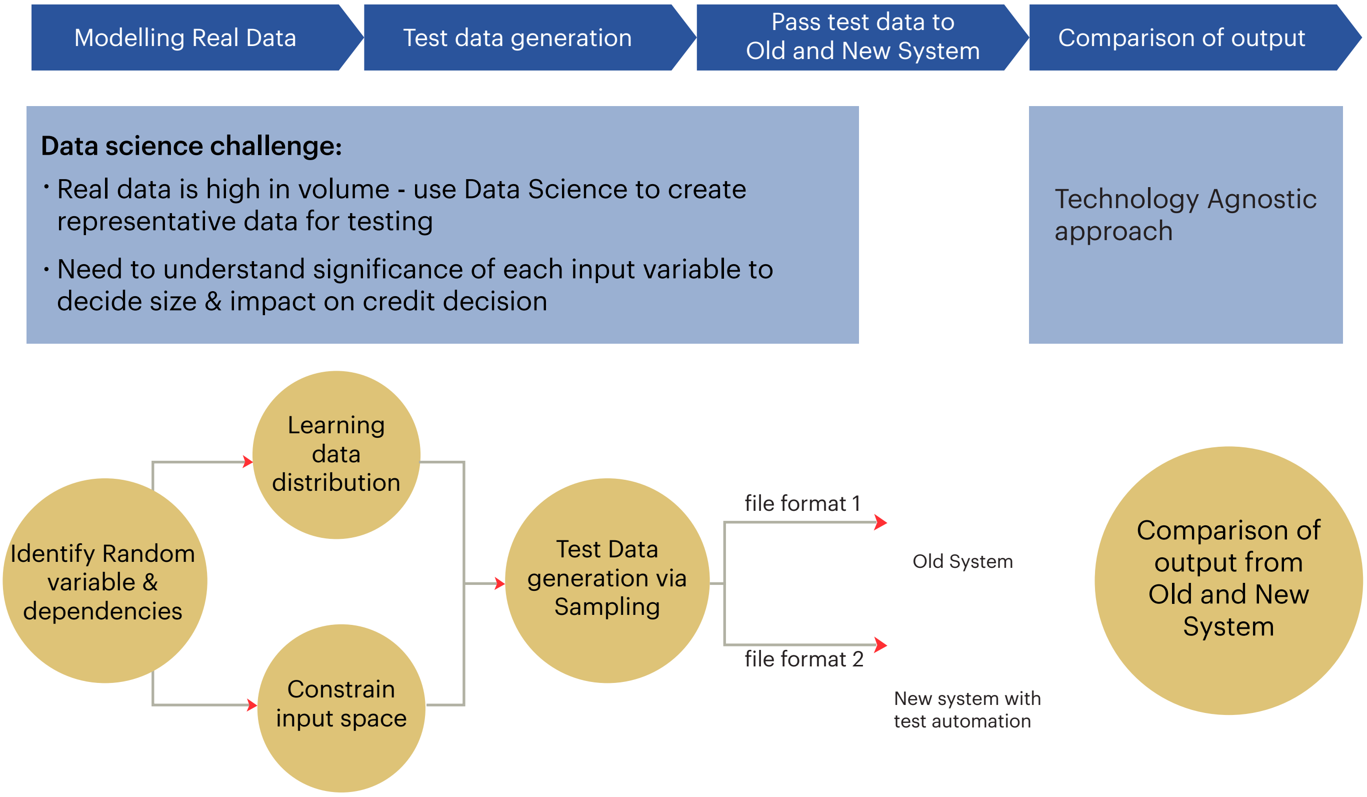
How is generative AI relevant here?

Addressing the problem at hand, the use case of test data generation for comparing the two systems, i.e., the existing legacy system and the planned advanced system, requires comparing the output of the two systems for the same input data. For a faster deployment of the new advanced system, it must be tested for wider and specific test cases. However, the testing is applicable only in test cases close to the real-world data samples and representative of new upcoming samples. Also, the test data should have both volume and variety of samples to save time in testing duplicates.

This requires an understanding of various fields of data. In the financial services use case we examined, the user-specified queries in the form of multiple fields. These fields were either categorical or numerical. The queries were the samples that we refer to as real-world samples, and they acted as references for the generation of synthetic samples. This kind of data is usually referred to as tabular data and is encountered in other domains as well. To generate synthetic samples that appear realistic, the model should be able to understand the interaction between different fields and generate the value of each field with respect to the other fields. The existing simple sampling-based systems can generate exhaustive samples based on the ranges and values each field takes. However, they do not model the relationship between the different fields. This problem can be solved with the help of GAN.

Generative AI learns to generate synthetic samples in an adversary setting and can understand the interaction of different fields in the data to create realistic synthetic samples. Thus, the generative AI solution can be applied in various settings and modalities. We can also use synthetic data generation to create data sets in cases where using real data might raise legal or other compliance issues. In this paper, we will focus the discussion on tabular data generation.

Zensar's approach to the problem



Modelling Real Data → Test data generation → Pass test data to Old and New System → Comparison of output

Data science challenge:

- Real data is high in volume - use Data Science to create representative data for testing
- Need to understand significance of each input variable to decide size & impact on credit decision

Technology Agnostic approach

Identify Random variable & dependencies
Learning data distribution
Constrain input space
Test Data generation via Sampling
file format 1
file format 2
Old System
New system with test automation
Comparison of output from Old and New System

Figure 3: Representation of Zensar's approach in testing a new system

Figure 3 highlights our approach for testing the new system with generated synthetic data and compares the results with the legacy system. Here, the first step is to understand various constraints in the input data for multiple fields like ranges, values, and valid combinations of different fields. These inputs can also come from user/subject matter experts and must be input into the system during training. Once trained, the model can be used in real-world settings to generate high volumes of data that follow the constraints (as in input data) in the generated samples. These synthetically generated samples can then be sent to the legacy and the new system. The legacy system's output can then be compared to analyze the new system's performance against the legacy system.

The testing can further be enriched by comparing the performance of the two systems against edge/corner cases of testing samples. These corner case samples are rare occurring samples and hence need to be explicitly generated by conditional sampling/reject sampling by a trained GAN model.

Key challenges and the way around it

Any advanced AI algorithm, including GAN, is widely limited by the amount and variety of data required to train it. Furthermore, the trained model is expected to generate the same quality as the real samples. Another challenge is that the real sample distribution may have some cases that are not that frequently occurring but are important for testing the system. This issue can be targeted by applying a training strategy that works on conditional distributions of column/set of columns for specific values or ranges of values. There could be cases where both categorical and numerical data are present.

Business-related challenges:

Data unavailability

Obtaining representative and comprehensive data for training synthetic data generation is a challenge. Edge cases may be rare in real-world datasets, but are essential for testing. This can be addressed by continuously improving the model with more data and developing edge case-specific models using generative AI.

Data privacy

Data privacy laws introduce technical challenges for data collection, storage, and sharing, making it difficult to develop systems. This includes difficulties in modeling user interaction and behavior. This can be mitigated by designing the system, training the model on the same server as the legacy system, and sharing observations and results with the new system for test case generation. Generative AI can also be used to generate diverse samples and share the trained model across testing teams while maintaining data privacy.

Data drift analysis

Data drift occurs when a model is trained on a given data and fields change due to environmental and seasonal changes, causing the AI system to lose predictive power. It is essential to understand the nature of change in the data and fine-tune the generative AI system with recent data. A system can be trained in multiple iterations for periodic drifts to model the analyzed behavior.

Obsolete test data

Testing a system with an independent dataset that mimics real-world situations is challenging, especially if some fields' ranges and dependencies change over time. Testing data must be updated regularly to reflect changing user/query behavior and trends.

Key benefits of Zensar's approach

A data-centric approach based on generative AI can be leveraged for:

Maintained data integrity

The existing input test data can be used to effectively learn the distribution of different fields. This can help generate more test samples (volume) following the distribution of the existing real data. This would enable the integrity of the synthetic data to be maintained.

Maximized test coverage

The corner/edge cases during testing are highly important to cover all possible scenarios. The conditional GANs can generate corner cases by specifying the boundary conditions for the required fields. This will ensure the optimal representation of testing data.

Faster testing

The overall testing time would reduce. This is because more coverage of different scenarios, including the corner cases, is ensured in a lesser amount of data.

Reusability of model

If the business is spread across multiple geographies, the same test automation utility could be used by fine-tuning the base AI model for different ranges of fields, e.g., currency. This would lead to reduced costs for developing new models every time and ensure standardization in modeling across all geographies.

Avoiding data privacy challenges

Synthetically generated data can be used in scenarios where data privacy is of prime concern, like financial services, and real-world data cannot be shared with different testing teams outside the production environment.

Other use cases of generative AI

Although generative AI is comparatively a newer technology, it will be the future of creating new artificial data in varied scenarios. Some of the critical areas where it is in the talk include:

Exploring unexplored geographic locations:

The satellite images can be converted into map images and help to venture into the unexplored.

Creating new patient records:

GAN can revolutionize the medical industry by creating new medical images based on patterns in various patient records. These images can be used to provide better insights into diseases and understand rare cases, thereby improving intuitive patient care and ensuring patient data privacy.

Venturing into the entertainment industry:

GAN can be used to create new genres of music – by amalgamating two or more genres. It can also be used to develop unique videos by using multiple videos. Another application is to create human-like voices or faces – especially for dubbing. It can also be used for film restoration by enhancing the quality of old pictures and movies.

Revamping search:

Often, we can describe what we want to see but cannot find images to match that description. GAN can be used for text-to-image translation and to produce images of a textual description, taking search engine capabilities to a higher level.

GAN can also be used in various business case scenarios like better risk management, fraud detection, trading prediction, synthetic data generation for multiple scenarios, and risk factor modeling. [6] [7] [8]

Conclusion

Generative AI can be used to generate structured test data for businesses to improve testing coverage, avoid data privacy challenges, and ensure data integrity. Additionally, Generative AI can be applied in various industries to generate synthetic data for training models, simulate scenarios, and make predictions.

The ability of generative AI to learn patterns in existing data and generate new data similar to it opens up a wide range of possibilities for organizations to improve their operations and decision-making processes. It is recommended that organizations should explore the use of generative AI and understand how it can benefit their specific use cases and industries. Generative AI has the potential to improve the way businesses operate, and organizations should consider how they can leverage this technology to gain a competitive edge.

References:

1. Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets. 27." In Proceedings of conference on neural information processing systems (NeurIPS). 2014.

2. Nistal, Javier, Stefan Lattner, and Gael Richard. "Comparing representations for audio synthesis using generative adversarial networks." 2020 28th European Signal Processing Conference (EUSIPCO). IEEE, 2021.

3. de Rosa, Gustavo H., and João P. Papa. "A survey on text generation using generative adversarial networks." Pattern Recognition 119 (2021): 108098.

4. A. Karnewar and O. Wang, "MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7796-7805, doi: 10.1109/CVPR42600.2020.00782.

5. Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation." In International Conference on Learning Representations. 2018.

6. https://www.walkme.com/glossary/generative-ai/

7. https://10xds.com/blog/age-of-generative-ai/

8. https://research.aimultiple.com/generative-ai/

Dr. Sumant Kulkarni
AVP and Head, Artificial Intelligence and Machine Learning
sumant.kulkarni@zensar.com

Dr. Annapurna Sharma
Lead Scientist, Artificial Intelligence and Machine Learning
annapurna.sharma@zensar.com

Rahul Nimje
Solution Architect, Artificial Intelligence and Machine Learning
r.nimje@zensar.com

Ankita Vashisht
Business Consultant, Artificial Intelligence and Machine Learning
ankita.vashisht@zensar.com

zensar
An RPG Company

We conceptualize, build, and manage digital products through experience design, data engineering, and advanced analytics for over 145 leading companies. Our solutions leverage industry-leading platforms to help our clients be competitive, agile, and disruptive while moving with velocity through change and opportunity.

With headquarters in Pune, India, our 11,500+ associates work across 30+ locations, including Milpitas, Seattle, Princeton, Cape Town, London, Singapore, and Mexico City.

For more information please contact: velocity@zensar.com | www.zensar.com