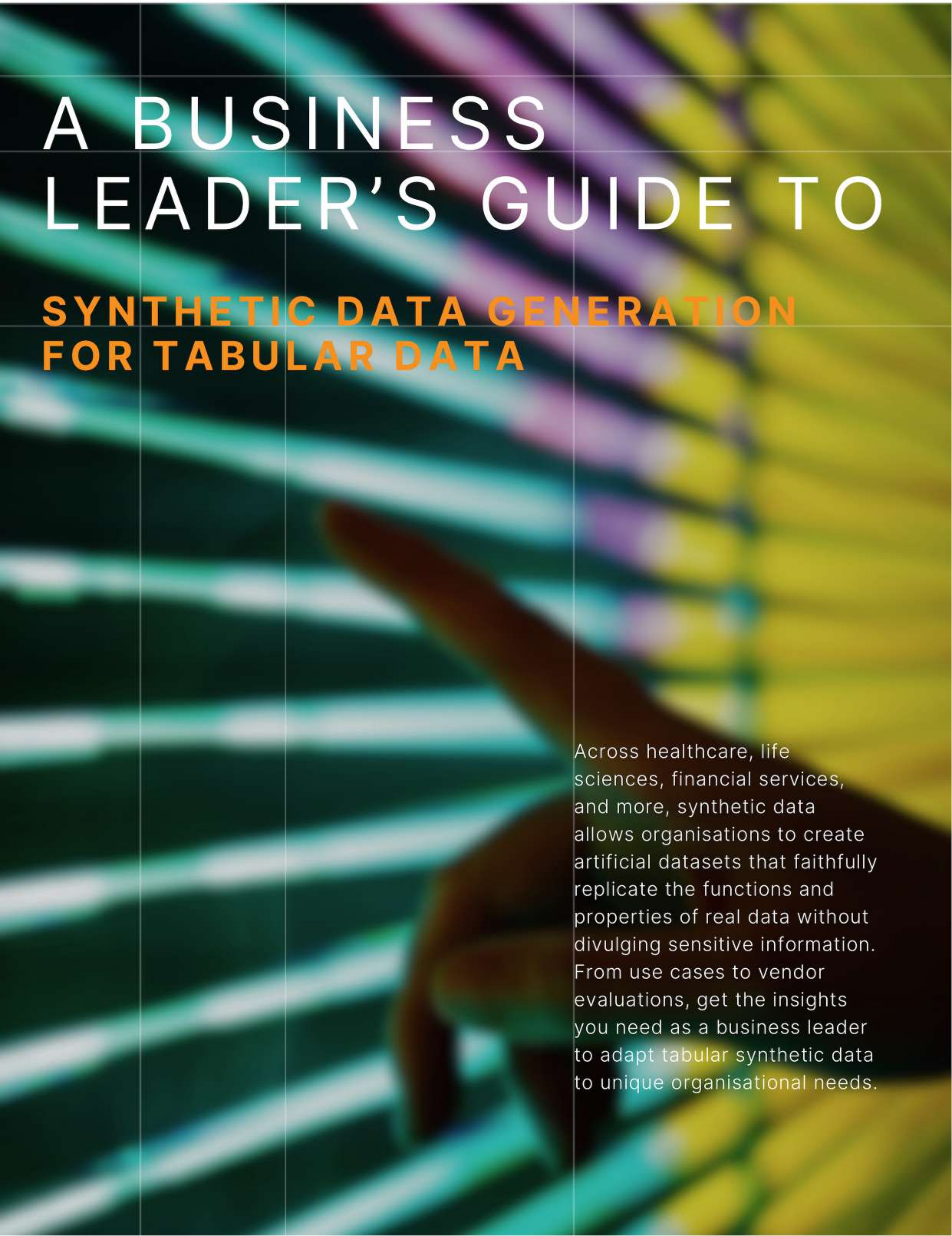
Tiger Analytics



A BUSINESS LEADER'S GUIDE TO

SYNTHETIC DATA GENERATION FOR TABULAR DATA

Across healthcare, life sciences, financial services, and more, synthetic data allows organisations to create artificial datasets that faithfully replicate the functions and properties of real data without divulging sensitive information. From use cases to vendor evaluations, get the insights you need as a business leader to adapt tabular synthetic data to unique organisational needs.

CONTENTS

WHY TIGER ANALYTICS?

Tiger Analytics, with vast experience in data analytics, AI, and GenAI, has been partnering with a roster of Fortune 500 clients across industries. Over the years, we have helped untangle several data issues – building and deploying real-world AI projects from the ground up. As a service company, we specialise in addressing gaps in data capabilities.

We have also conducted thorough research on synthetic tabular data generation - the tools, vendors, etc.

It's why we, at Tiger Analytics, are uniquely positioned to offer expert insights, backed by real-world wisdom, on synthetic data generation.

KEY TAKEAWAYS

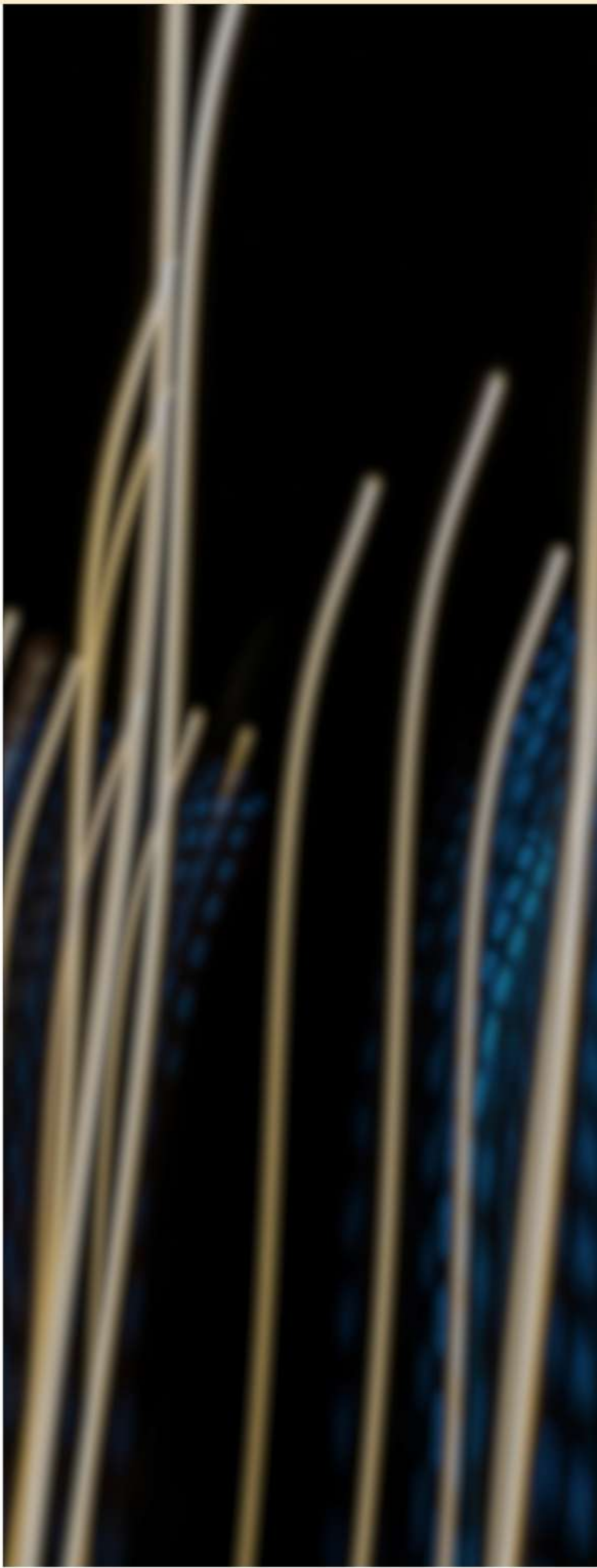
In this whitepaper, we explore how tabular synthetic data generation can address data scarcity and what organisations must keep in mind while choosing the right partners.

In-depth understanding of the tabular synthetic data landscape

Strategic advantages and practical applications of tabular synthetic data generation

Glimpse into how tabular synthetic data solves practical problems and where it can be applied

Framework to evaluate vendors and solutions based on unique needs



INTRODUCTION

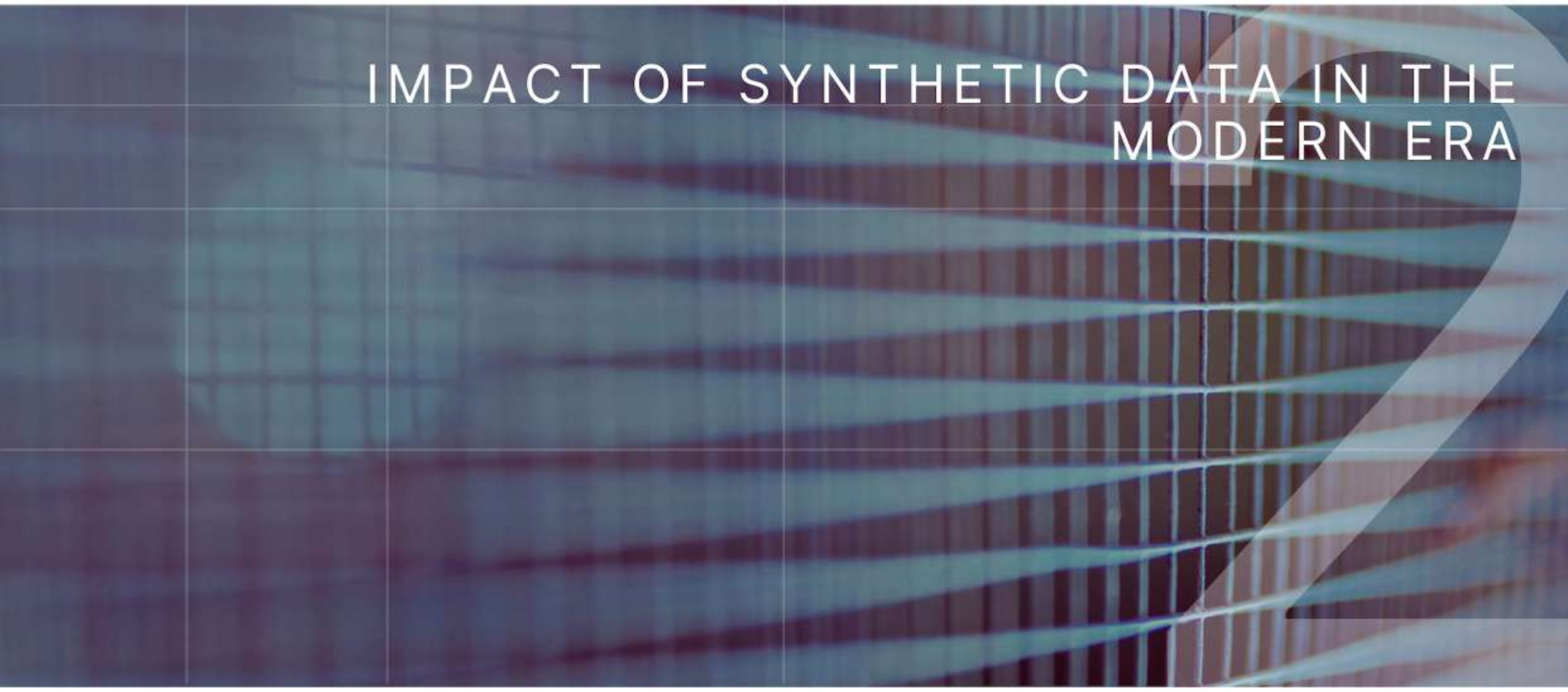
Tabular data is considered an extremely valuable asset across industries. However, the scarcity of good-quality data remains a major challenge. On the one hand, it can be attributed to stringent privacy concerns, which prevent organisations from sharing their data assets. On the other hand, it may simply be a shortfall in the volume of available data.

This is where tabular synthetic data comes into play.

Synthetic data generation techniques have been applied across various domains to replicate the characteristics and statistical properties of real-world datasets. These techniques cover a range of data types, such as images, natural language text, audio, and more.

Modern synthetic data generation methods offer a powerful solution. It allows organisations to create artificial datasets that faithfully replicate the functions and properties of real data without divulging sensitive information.

However, the success of this process can only be determined by the choice of method and its alignment with specific use cases.



IMPACT OF SYNTHETIC DATA IN THE MODERN ERA

Synthetic data plays a crucial role in overcoming the challenges faced by organisations that strive to harness the power of high-quality, real-world data. These typically comprise privacy concerns, data availability, and the cost of data acquisition.

But what exactly is synthetic data?

It refers to artificially generated data that imitates the characteristics and statistical properties of real-world data sets. It is created using generative models, simulations, and algorithms – mimicking the patterns, distributions, and correlations found in actual data.

While synthetic data is not derived directly from real observations, it retains similar statistical properties. Hence, it can be used for various analytical purposes without putting the privacy or security of sensitive information in jeopardy – ultimately ensuring data availability.

Synthetic data is also extremely beneficial in helping mitigate **critical business challenges.** These include:

DATA PRIVACY RISKS

Business Problem: Regulatory risks prevent essential data sharing for modeling.

Solution: Synthetic datasets serve as a privacy-preserving alternative for testing, validation, and benchmarking, ensuring that an organisation's data handling practices comply with legal demands. It allows rigorous assessment, driving adherence to data privacy regulations such as GDPR or HIPAA and building trust among stakeholders.

LARGE SCALE DATA COLLECTION IS CHALLENGING

Business Problem: Data collection challenge at a large scale impedes data availability in certain environments.

Solution: Enabling data collection, especially in industrial environments, can be costly and risky. Synthetic data generation can be applied to create virtual simulations of these processes, providing a safe and cost-effective way to analyse and improve industrial operations. It is critical to develop accurate models of processes like chemical reactions, manufacturing lines, or power generation that capture the physics, chemistry, and dynamics of the system.

BIASED DECISION MAKING

Business Problem: Imbalanced data that leads to poor decision-making.

Solution: Synthetic data combats imbalanced data and bias simultaneously. It rebalances datasets by creating artificial instances of underrepresented classes and mitigates bias, ensuring fairness and informed decision-making in one stride.

LACK OF DATA FOR LOAD TESTING

Business Problem: Limited data that hampers load stress testing effectiveness. Due to data restrictions in non-production environments, there is insufficient data for functional, non-functional and load testing of systems.

Solution: Synthetic data enables organisations to simulate high-demand scenarios and assess system performance under extreme conditions. Organisations can optimise their infrastructure while achieving resilient performance during peak loads. Synthetic data offers a controlled and adaptable environment for functional load testing. Organisations can create diverse test cases to cover several scenarios, from routine operations to edge cases.

WHAT INDUSTRIES CAN BENEFIT FROM SYNTHETIC DATA?

Let’s look at the functional Use Cases of GenAI-based synthetic data generation for tabular data.

CROSS-INDUSTRY RELEVANCE

- Drives testing and developing algorithms, models, or software applications
- Enables the creation of large datasets with known properties to evaluate system performance, assess scalability, or test new features without relying on real-world data.

INSURANCE

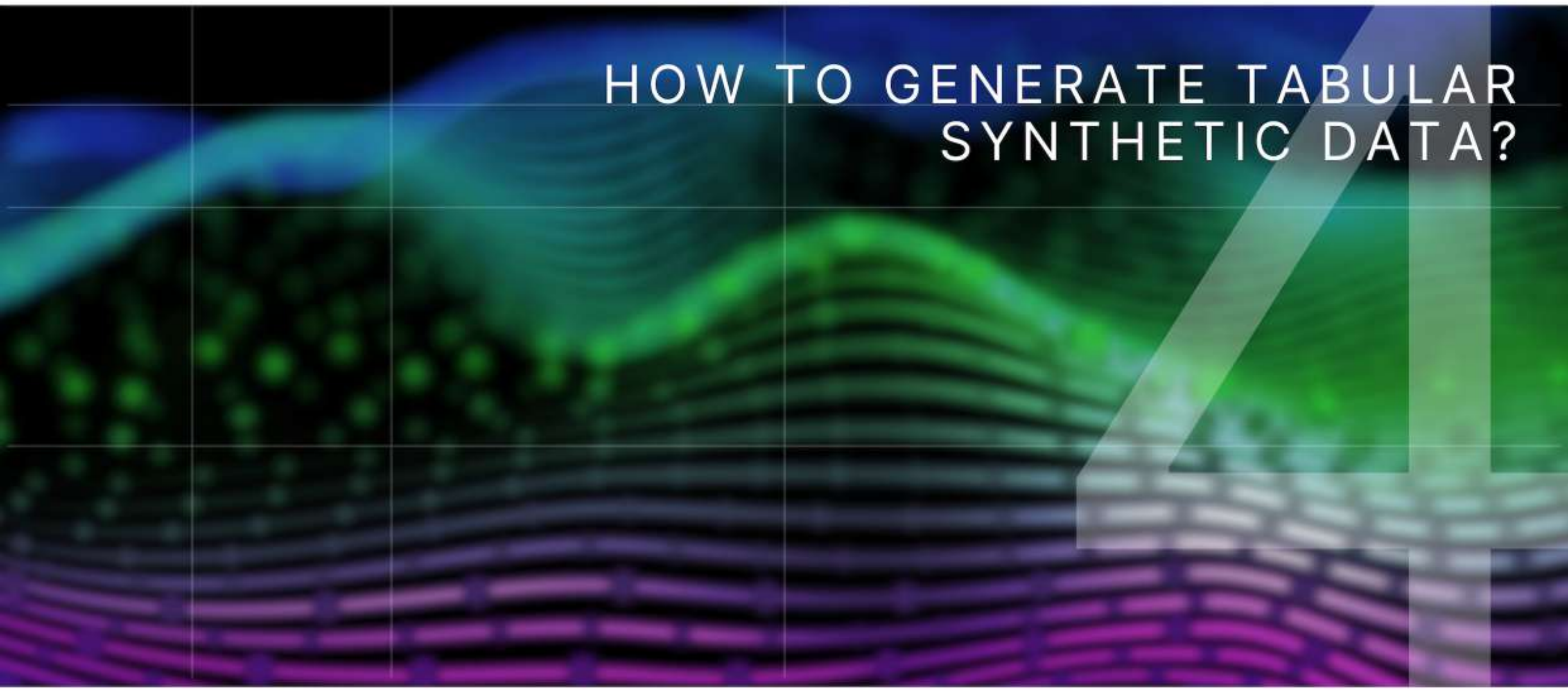
- Analyses historical claims data, policy documents, and customer information to generate synthetic claim scenarios, enabling efficient risk assessment and fraud detection.
- Helps automate claims processing and improves efficiency.

HEALTHCARE

- Generates synthetic patient data that closely mimics real patient profiles while ensuring privacy and data security.
- Enables training AI algorithms, improves research collaborations, and addresses data sharing limitations or regulatory constraints.

MANUFACTURING, SUPPLY CHAIN & OPERATIONS

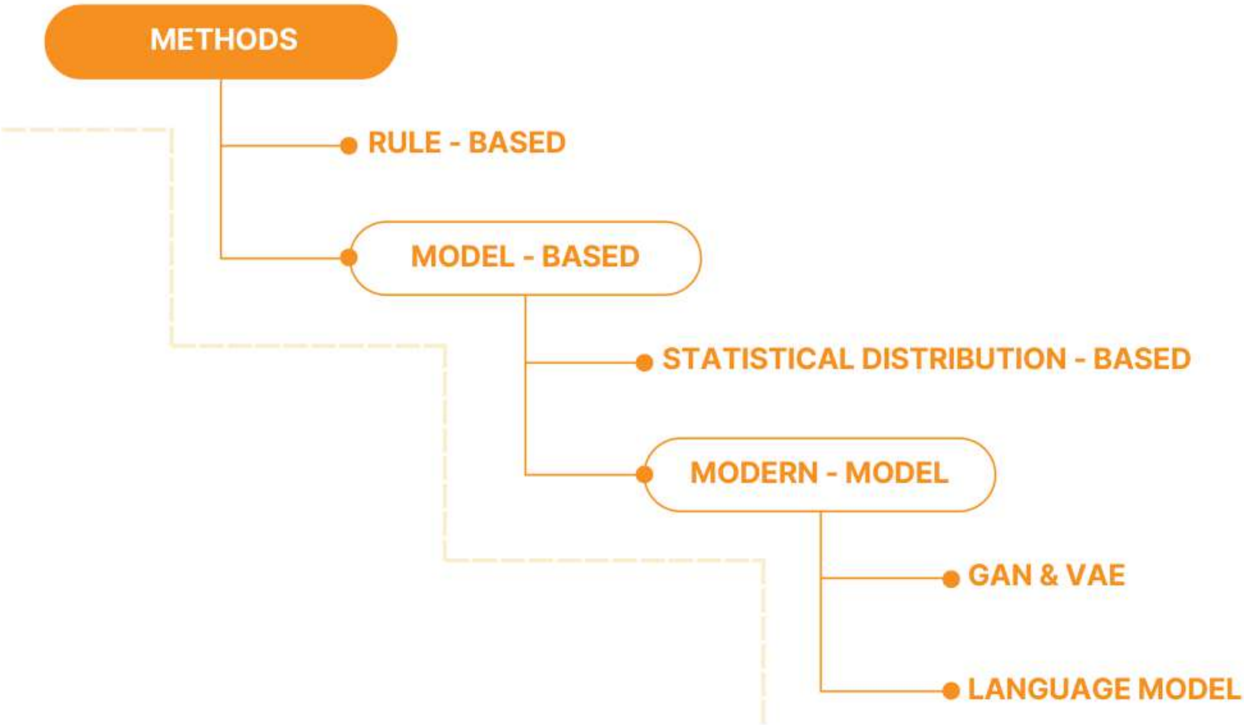
- Creates simulated scenarios or generates artificial datasets that match certain characteristics.
- Performs predictive modeling, scenario planning, or risk assessment without relying solely on existing data.



HOW TO GENERATE TABULAR SYNTHETIC DATA?

The available methods can be divided into two broad categories:

Model-based and Rule-based.
Both offer distinct attributes and trade-offs for meeting specific business needs.



METHODS

RULE - BASED

MODEL - BASED

STATISTICAL DISTRIBUTION - BASED

MODERN - MODEL

GAN & VAE

LANGUAGE MODEL

Methods of generating synthetic data:

RULE - BASED METHODS

With rule-based synthetic data generation, organizations can explicitly define rules, constraints, or algorithms that mirror original data patterns. It provides granular control and speed in quickly producing high volumes of data.

However, the accuracy depends on capturing data patterns manually, often requiring the expertise of a subject matter specialist. It means that the rule-based method introduces potential biases and limits based on the knowledge of experts.

When to use it?

- When control over data generation is crucial.
- When rapid generation of large volumes of data is needed.
- When data patterns are well understood and can be explicitly defined by rules and constraints.

Who can use it?

- Subject matter experts who can define data patterns and constraints.
- Organisations needing quick and controlled synthetic data for specific scenarios.
- Projects where the speed of data generation and adherence to known rules are prioritised over the discovery of hidden patterns.

MODEL - BASED METHODS

Model-based methods can be further classified into traditional and modern approaches.

TRADITIONAL METHODS

Add random noise: A common practice is to introduce random noise to real data, obscuring its initial structure and content. This approach gives rise to apprehensions regarding the potential for reverse engineering the data, which could lead to privacy vulnerabilities.

Replicate statistical distributions: Alternatively, traditional methods seek to replicate the statistical distributions of the columns of data. Univariate approaches focus on individual data columns, while multivariate methods aim to capture dependencies among multiple columns. But, they may struggle with capturing correlations between columns, when dealing with data that deviates from standard distributions.

MODERN METHODS

Generative Adversarial Networks (GAN) & Variational Autoencoders (VAE) models:

These cutting-edge approaches are designed to learn the combined distribution of data columns.

GANs employ a two-network setup consisting of a generator and a discriminator. The generator generates synthetic data samples by learning from real data, while the discriminator tries to distinguish between real and synthetic samples. This adversarial training encourages the generator to produce data that closely resembles the real data distribution.

VAEs work by encoding real data into a lower-dimensional latent space and then decoding it to generate synthetic data. They offer a probabilistic framework, allowing for the modeling of uncertainty in data generation.

Both models observe interactions and correlations between data columns, identify underlying patterns, and replicate these patterns in the generated data.

While GANs and VAEs are powerful tools for synthetic data generation, they can present challenges like training complexity, lack of control over generated data, and heavy computational resource requirements. They also could be less effective when applied to datasets with categorical data.

Language model-based methods:

This new paradigm embraces language models such as Long-Short-Term Memory (LSTM) and transformer-based models. These models excel in replicating patterns within categorical data because they can leverage the contextual relationships and hierarchies present in text-based data.

However, they may have limitations when dealing with numeric data, as their architecture is primarily tailored for processing and understanding textual information. After all, Large Language Models (LLMs) treat numbers as text, which can lead to challenges when attempting to capture the characteristics of numerical data.

In summary, while the traditional method is an unstructured (hack-like) approach, the modern methods are more structured. Hence, while the former offers ease of implementation and simplicity, the latter may be more effective for "near real" data generation.

Hence, while the traditional method offers simplicity and ease of implementation, the modern approach leverages advanced Deep Learning techniques and language models for more effective synthetic data generation.

OVERVIEW

TRADITIONAL METHODS

- Adds random noise, which obscures the initial structure/content
- Replicates statistical distributions
- Focuses on individual data columns
- Captures dependencies among multiple columns
- May cause reverse data engineering issues

MODERN METHODS

- Leverages GAN models to learn the joint distribution of data columns
- Employ a two-network setup (generator and discriminator)
- Generates synthetic data samples by learning from real data
- Leverages VAE models to encode real data into a lower-dimensional latent space
- Decodes it to generate synthetic data
- Leverages language models like Term Memory (LSTM) and transformer-based models

When to use Modern Methods?

- When autonomous pattern recognition is important
- When the focus is on replicating complex data distributions and relationships
- When aiming to leverage advanced ML techniques like GANs and VAEs

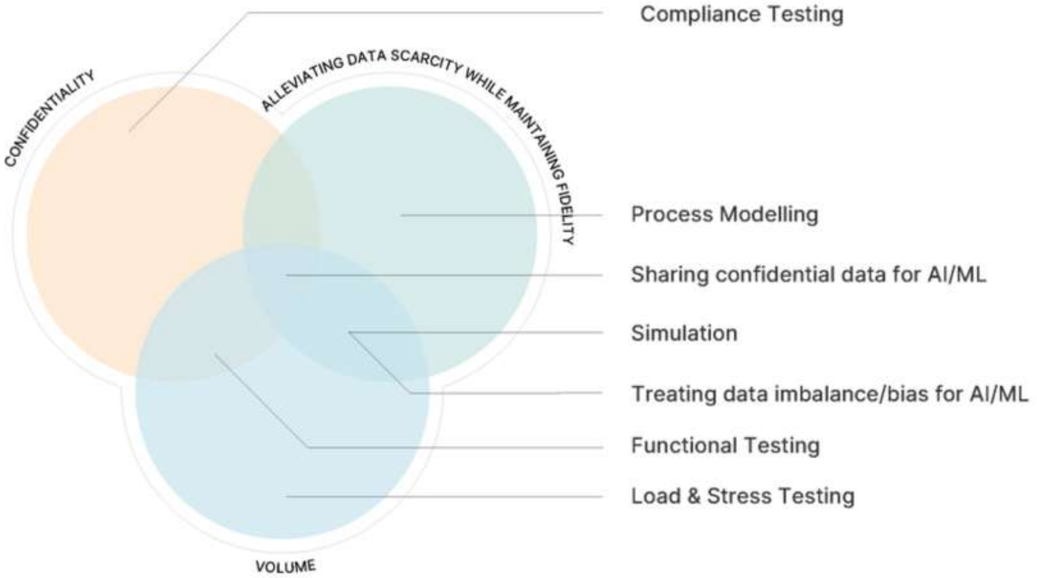
Who can use it?

- Data specialists and ML engineers familiar with complex models
- Organisations that have computational resources to train and deploy such models
- Projects requiring high fidelity in the synthetic data, mimicking real data patterns

THREE MAIN CLASSES OF USE CASES FOR TABULAR SYNTHETIC DATA

Our experts at Tiger Analytics help illustrate scenarios where synthetic data fulfils varied requirements across the intersections of three classes of Use Case:

Three Intersections of Use Cases: Confidentiality, Mitigating bias, and High-volume of Data.


CONFIDENTIALITY
ALLEVIATING DATA SCARCITY WHILE MAINTAINING FIDELITY
VOLUME
Compliance Testing
Process Modelling
Sharing confidential data for AI/ML
Simulation
Treating data imbalance/bias for AI/ML
Functional Testing
Load & Stress Testing

CONFIDENTIAL DATA SHARING

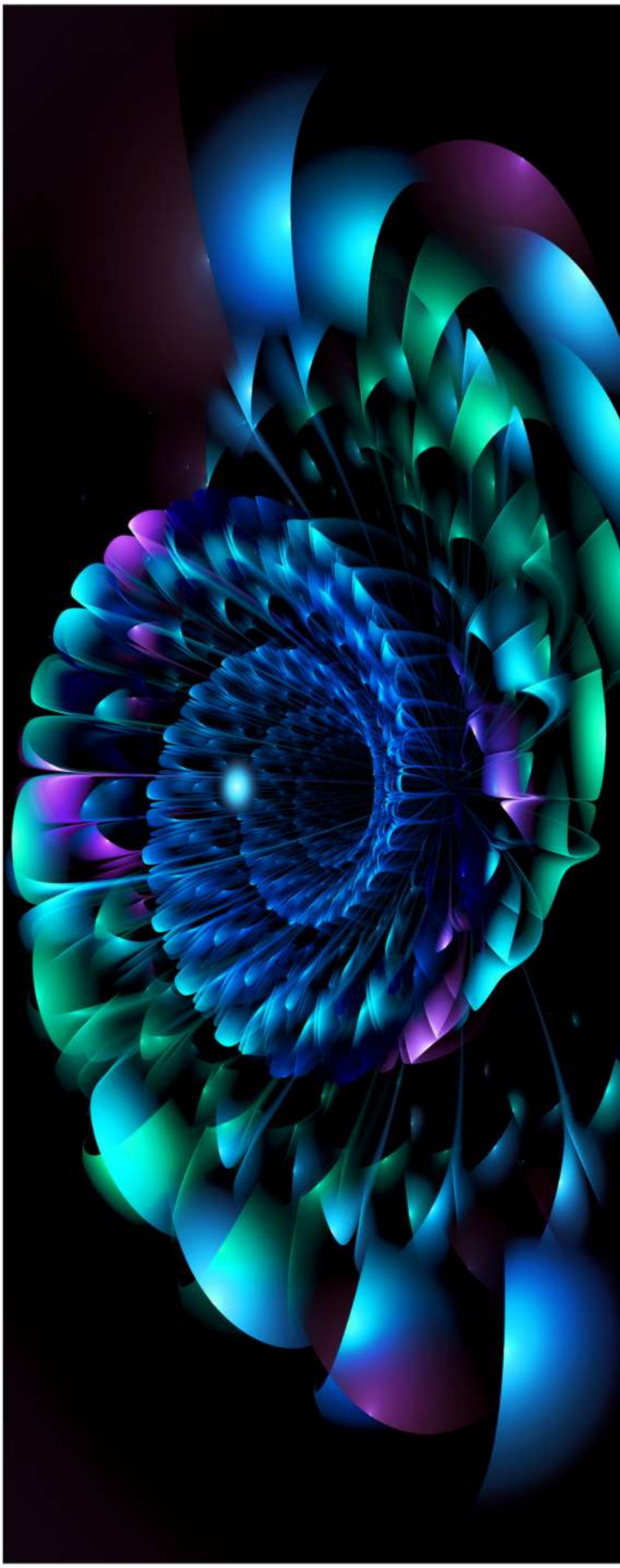
Organisations grapple with the dilemma of sharing proprietary data while safeguarding user privacy and PII. Confidential data sharing is the need of the hour, especially for model building and analysis. Synthetic data bridges the gap by enabling the training of models to understand data distributions, patterns, and characteristics. It results in a synthetic dataset that mirrors real data without impacting privacy.

MITIGATING BIAS

Synthetic data addresses scarcity and biases when data is insufficient, imbalanced, or skewed (for example, a particular gender), leading to biased models. Creating synthetic samples ensures that models have access to the requisite data diversity and balance for better decision-making.

HIGH-VOLUME DATA NEEDS

Extensive data quantities are important for load testing applications, which prioritise data volume over accurate data patterns. For instance, organisations may need a lot of data for stress-testing GenAI applications or conducting software load tests. Synthetic data acts as a scalable solution, alleviating the need to share sensitive data with third-party service providers.



CURRENT MARKET SCENARIO: COMMERCIAL & OPEN SOURCE OFFERINGS

Commercial offerings

Several vendors offer synthetic data generation as a service. Many of them operate in the cloud and require users to upload their data to their servers. Some of these providers also offer on-premises installations through docker containers.

> *These off-the-shelf offerings can expedite the execution of synthetic data.*

INABILITY TO MEET UNIQUE NEEDS

Standardised (universal) solutions may not perfectly match the complex requirements of dynamic business use cases. Therefore, directly inputting the data into a model will not capture all its subtleties. Frequently, multiple data preprocessing steps are necessary before training a model and after generating synthetic data, as well, to achieve realistic results.

LACK OF TRANSPARENCY

Certain commercial vendors do not openly share their methodologies or the fundamental processes behind data generation. It can lead to concerns about the data generation practices.

Open-source tools

In the open-source world, there are several data generation tools that provide flexibility, transparency, and performance comparable to their commercial counterparts. Let’s look at three of the most popular open-source tools:

SYNTHCITY

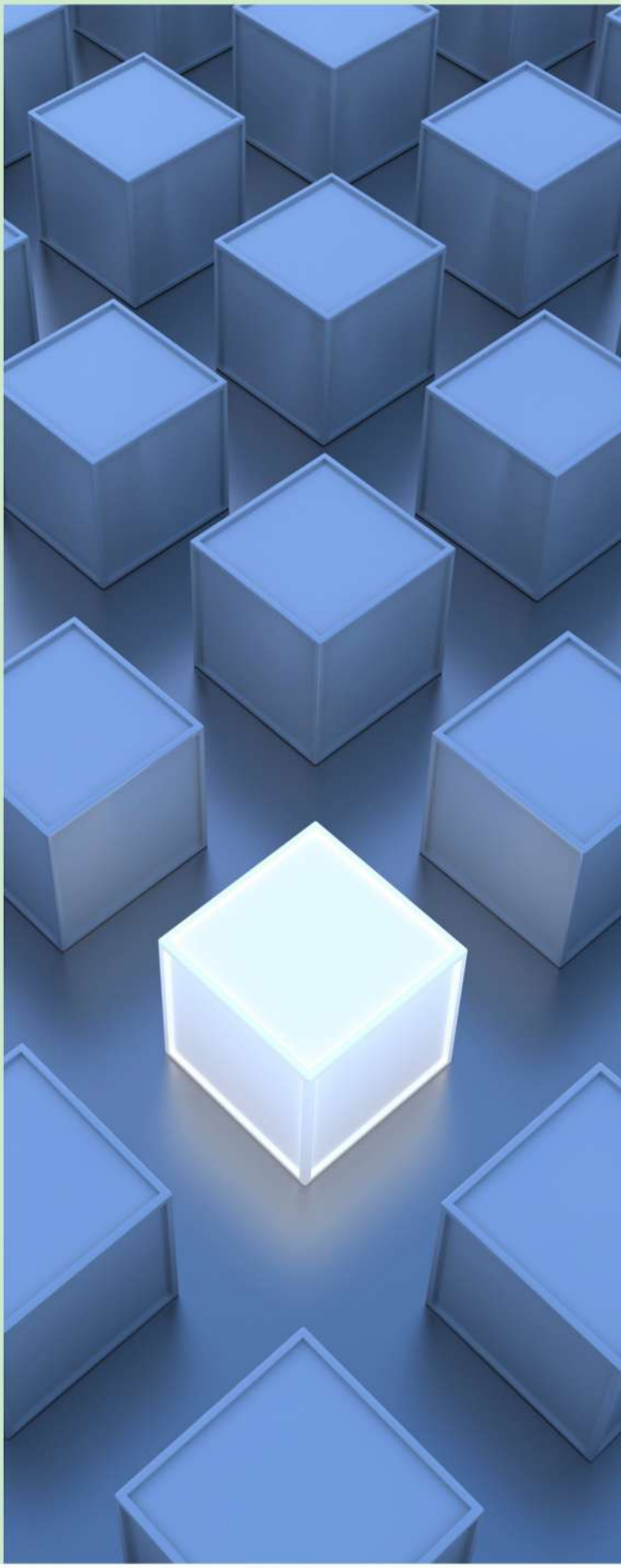
Synthcity offers a variety of GAN and VAE-based models, customised for tabular data generation.

BE-GREAT

be-great, an open-source Python package, leverages the capabilities of LLMs available through Hugging Face. It fine-tunes LLMs using Parameter-Efficient Fine-Tuning (PEFT) techniques, streamlining the accurate synthesis of data.

DBLDATAGEN

For rule-based synthetic data solutions, dbldatagen is a versatile open-source tool. Users can define rules, constraints, and algorithms, granting them the power to create synthetic data that aligns with their requirements.



The open-source nature of these tools provides transparency and adaptability. It enables the development of customised, top-tier solutions, which perfectly match the distinct business use case needs.

It's why, when choosing between enterprise and open-source solutions, it is important to consider how quickly you require results, how complex and customisable the data generation is, and how much you value transparency and data privacy.

Pros and cons of commercial and open-source offerings

COMMERCIAL OFFERINGS

PROS

- Ready-to-use solutions that accelerate the process of generating synthetic data
- Professional support and maintenance to quickly resolve any issues

CONS

- Inability to cater to specific business requirements, leading to suboptimal performance
- Lack of transparency, raising doubts about data integrity and suitability

OPEN-SOURCE OFFERINGS

PROS

- Flexibility to modify or extend the codebase to meet specific needs
- Transparency for users to fully understand and trust data generation processes

CONS

- Significant in-house resource requirements, which may not be feasible
- Lack of highly available professional support

SYNTHETIC DATA GENERATION: THE TIGER POV

In today's dynamic business landscape, the choice of an appropriate data generation method holds significant sway over the success of any data-driven initiative. We, at Tiger Analytics, have in-depth know-how in delivering custom commercial offerings and follow a customised methodology for synthetic data generation based on the use case requirements.

We carefully analyze the current requirements of the client, along with their future needs, to deploy the right offering that meets their exact specifications.

It's why our selection process is truly unique as we have the ability to provide a custom toolkit of synthetic data generation tools, which vary as per organisational needs and industry demands.

HOW WE NAVIGATE DATA GENERATION COMPLEXITIES:

- **CUSTOMIZED SOLUTIONS**
- **TAILORED TOOL SELECTION**
- **MODEL SELECTION GUIDELINES**
- **HYBRID DATA COMPOSITION**
- **SYNTHETIC DATA SUCCESS STORIES**
- **ONGOING RISKS AND CONSIDERATIONS**

CUSTOMIZED SOLUTIONS

Choosing the right method is important. We, at Tiger Analytics, help organisations to select the right data generation method, which aligns with their objectives, considering factors such as control, speed, and accuracy. Rule-based methods excel in volume-based scenarios or when capturing patterns in not too important, while model-based methods autonomously uncover patterns, catering to unique business needs.

TAILORED TOOL SELECTION

There is no one-size-fits-all approach. We ensure that the choice of tool should be driven by the unique characteristics of the use case.

MODEL SELECTION GUIDELINES

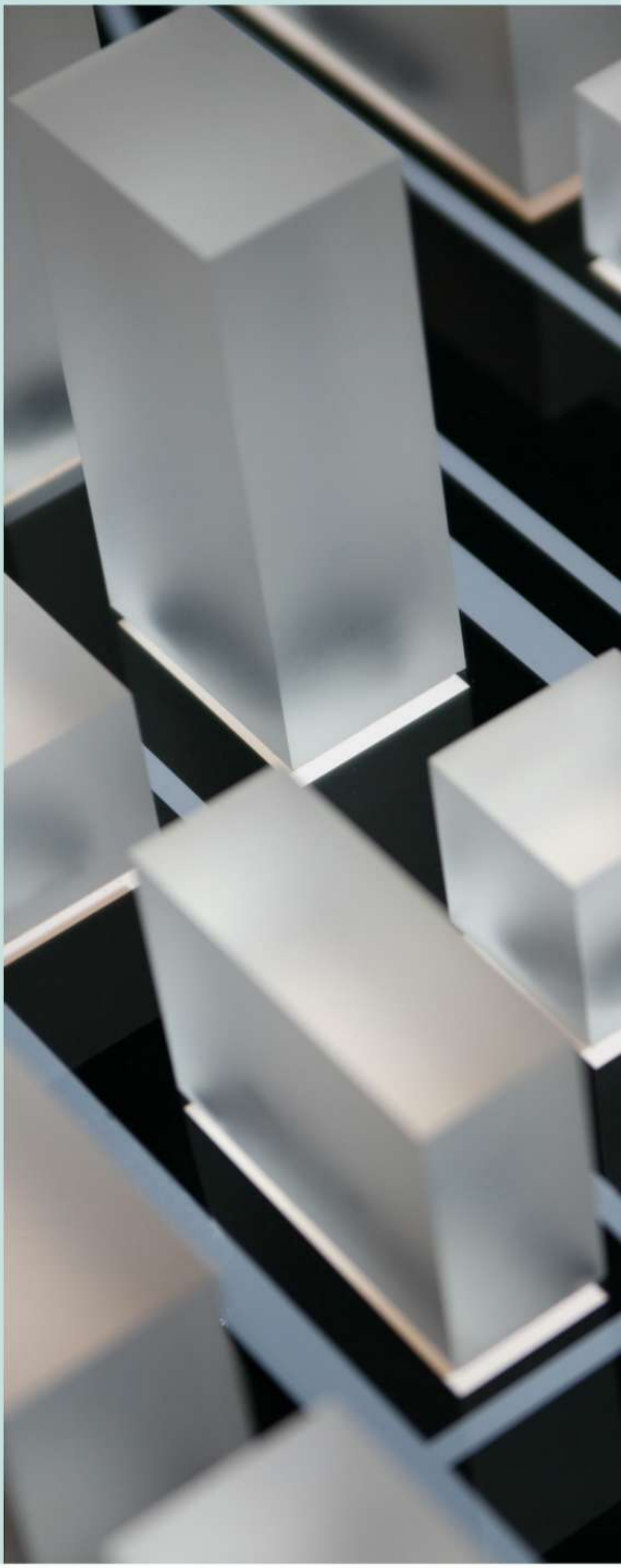
No single model fits all scenarios. While experimentation is essential, a general rule of thumb can be applied based on the following inputs by Tiger Analytics:

- For numerical data, consider starting with TVAE
- When dealing with categorical data, begin with an LLM-based approach
- For datasets with a mix of both, start with GAN-based models

HYBRID DATA COMPOSITION

The practice of blending real and synthetic data is applicable in situations, where the goal is to address imbalanced data scenarios and ensure the development of fair AI models. Replacing real data entirely with synthetic data for modeling purposes may not yield optimal results, especially when working with imbalanced datasets.

In such cases, we, at Tiger Analytics, recommend a balanced blend of actual and synthetic data often outperforms using real data alone. This approach helps mitigate data imbalance issues and contributes to building fair AI models. However, determining the ideal blend ratio remains a critical hyper-parameter that demands careful consideration.



SYNTHETIC DATA SUCCESS STORIES

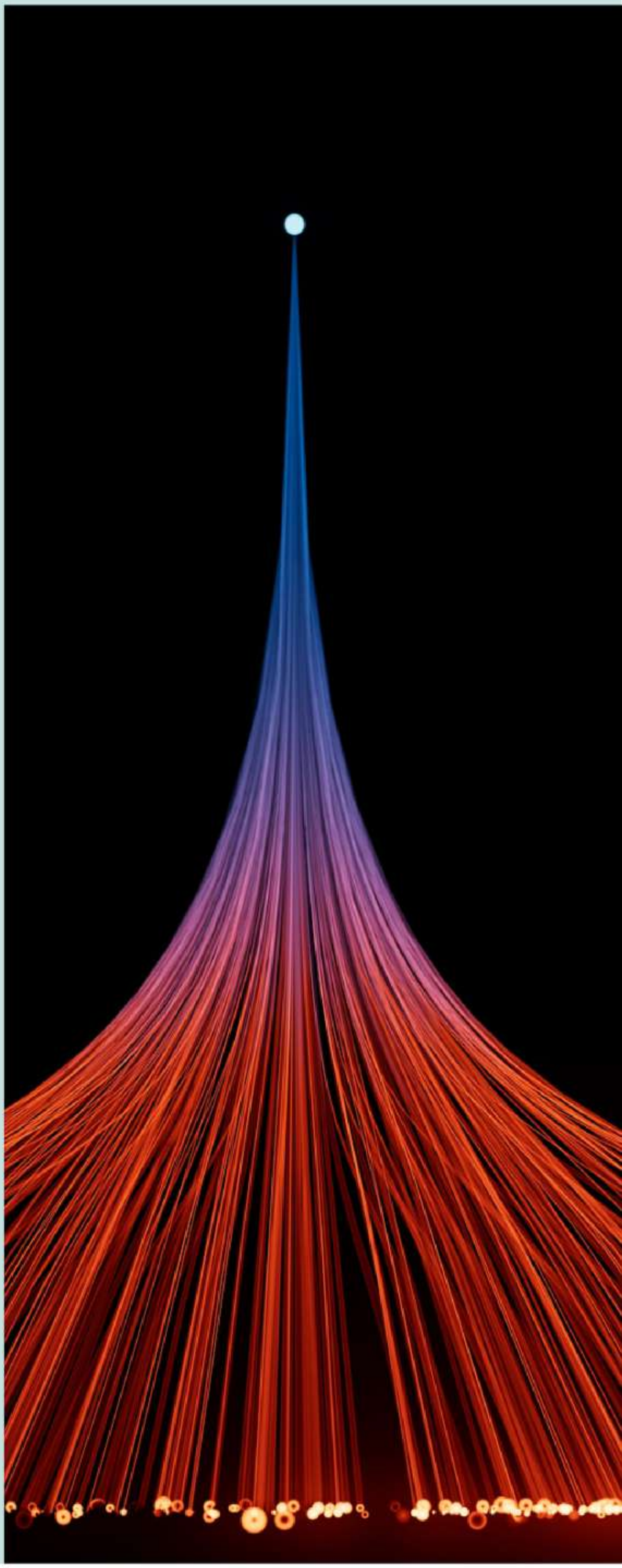
Synthetic data excels in preserving privacy and capturing essential statistical characteristics, including min, max, mean, and variable correlations. We have demonstrated proven expertise in helping organisations leverage these to achieve success.

ONGOING RISKS AND CONSIDERATIONS

While highlighting the potential of data generation, we recommend recognising the ongoing risks associated with training models and preserving complex, business-specific patterns.

FIVE CRITICAL CONSIDERATIONS

- Adequate datasets are required to effectively train models
- Models may struggle to capture rare category levels
- Complex, non-linear relationships between columns may elude certain models
- Certain business-specific patterns may not be fully preserved, affecting data authenticity
- No perfect solution exists currently since this is a rapidly evolving stream of work



8 MEASURING QUALITY OF GENERATED DATA

The measures we need to use to judge the quality of generated data depends on the Use Case. For a few popular Use Cases described below, we have a checklist.

USE CASES	CHECKLIST
DATA INTEGRITY CHECKS	□ Are all the levels in the categorical columns present in actual data? □ Are the various combinations of columns preserved?
MODELING REQUIREMENTS	□ Column distributions are similar in real and synthetic data. □ Correlations between different columns are maintained □ ML efficacy metrics calculate the success of using synthetic data to perform an ML prediction task
PRIVACY PROTECTION	□ ML Detection metric calculates how difficult it is for an ML model to differentiate the real data from the synthetic data □ Distance to Closest Record (DCR) measures the proximity of synthetic records to their closest original records

TIGER ANALYTICS' 3-STEP FRAMEWORK TO SELECT THE RIGHT METHOD

Step 1:

HOW TO CHOOSE BETWEEN SYNTHETIC DATA GENERATION MODELS OR METHODS IN CHOOSING THE CORRECT PROVIDER?

- Choose rule-based methods for scenarios requiring high degree of control on the output and huge volumes of data to be produced.
- Choose model-based methods when the goal is to learn patterns autonomously
- Assess the amount of data available (model based methods need a certain amount of data to train on. Rule base methods can work with little or no data but might require an SME to guide on the rules to apply)
- Understand the specific needs of the use case so that synthetic data generation tools can meet them

Step 2:

WHAT SHOULD YOU FOCUS ON WHILE MAKING BUDGETING DECISIONS?

- Evaluate the cost implications of synthetic data generation tools based on factors like licensing fees
- Avoid overspending on advanced models if a simpler method can achieve the desired results
- Budget for resources like hardware, software, and skilled personnel to manage the data generation method

Step 3:

ARE THERE ANY TECH/INFRA DEPENDENCIES THAT WILL SWAY THE DECISION?

- Ensure adequate computational power is available (Model-based methods, especially those using GANs or LLMs, require a lot of processing power)
- Assess the compatibility of the chosen synthetic data generation tools with existing systems and workflows
- Choose solutions that can scale with business growth
- Evaluate the technical expertise needed to implement and maintain the synthetic data generation methods

FINAL THOUGHTS

Undoubtedly, synthetic data generation plays an all-encompassing role in building an Insight-rich future. The need for data-driven decision-making will soar in the future. Still, challenges like data scarcity, privacy concerns, and biased datasets will continue to be a concern for organisations of all sizes.

Data-driven decision-making will soar in the future.

Hence, it's important to focus on building a future where data scarcity is a thing of the past, where every business, regardless of its size or resources, can harness the power of data to drive meaningful change in how they operate.

The time for synthetic data is now, and with the right expert overseeing the generation process - the future is brighter than ever.

About the Authors:

VASUDEVA MAIYA - DATA SCIENTIST

He is a seasoned Data Science professional with over 16 years of experience in Data Science and Consulting. Throughout his career, he has worked across diverse domains, including Consumer Packaged Goods (CPG), Pharmaceuticals, Media, Transportation, Insurance, and Real Estate.

SEEMANT SINGH - LEAD DATA SCIENTIST

With over 7 years of experience in Data Science across diverse domains including Insurance, Retail, Supply Chain, and Real Estate, Seemant has led numerous projects on predictive modeling, customer segmentation, and generative AI. Currently, his work focuses on developing advanced machine learning models to optimise business processes and drive data-driven decision-making.

AARTI KAPUR - PARTNER, DATA SCIENCE

With 16 years of experience in data analytics and machine learning, she has worked across diverse domains, including CPG, retail, transportation, and manufacturing. Aarti excels at solving complex business problems using mathematical and statistical constructs and technology. In recent years, she has been focusing on challenges in supply chain, pricing, and generative AI.

ABOUT TIGER ANALYTICS

Tiger Analytics is a global leader in AI and analytics, helping Fortune 1000 companies solve their toughest challenges. We offer full-stack AI and analytics services and solutions to help businesses achieve real outcomes and value at scale. We are on a mission to push the boundaries of what AI and analytics can do to help enterprises navigate uncertainty and move forward decisively. Our purpose is to provide certainty to shape a better tomorrow

Being a recipient of multiple industry awards and recognitions, we have 4000 technologists and consultants, working from multiple cities in 5 continents.

www.tigeranalytics.com

US | UK | CANADA | INDIA | SINGAPORE | AUSTRALIA