

Advancing Trustworthy AI/ML: Profile- based Synthetic Data Generation

A Unissant Accelerator

Advancing Trustworthy AI/ML: Profile-based Synthetic Data Generation

A Unissant Accelerator

Table of Contents

Taking secure steps: building robust AI/ML models.....	1
Making confident strides: developing and sustaining AI/ML models.....	2
Avoiding rocks on the path: the role of profile-based synthetic data generation	3
Rock 1: Lack of data for training models.....	4
Rock 2: Incorrect, non-representative data.....	5
Rock 3: Trust in synthetic data	6
Rock 4: Perpetuating bias	7
Rock 5: One size does not fit all.....	8
Reaching your destination: secure, ethical models	9
About Unissant.....	10

The path to best-in-class AI and ML models requires three focal points: unwavering commitment to data privacy, ability to ensure model accuracy, and adherence to established data management procedures. Every step matters.

Taking secure steps: building robust AI/ML models

Demand for solutions leveraging artificial intelligence and machine learning continues to surge. Traditional data segregation methods and sanitized test datasets struggle to keep pace with the need for rapid development and robust security, especially during growth spurts or real-world deployments. These techniques also often fail to eliminate inherent biases within data.

The cloud revolution has amplified the importance of data privacy and security. Yet, acquiring enough data for effective AI/ML training remains an obstacle. With the growing adoption of AI/ML solutions by CTOs, CIOs, and CDOs, ensuring a balance between data protection and efficient AI/ML design, development, and maintenance has become a critical priority.

President Biden's October 30, 2023 Executive Order (EO) on AI emphasizes the need for secure, safe, and responsible development and use. It highlights the importance of establishing new standards for responsible AI applications, particularly regarding safety, security, and data privacy. The Executive Order encourages the development of trustworthy and accountable AI models.

Similarly, the Federal Data Strategy Framework outlines a 10-year vision for the government's use of data. This Framework underscores the importance of balancing security, privacy, and confidentiality with data-driven decision-making, AI/ML training and development, and data interoperability across agencies and external partners.

The imperative is clear: agencies must identify the most secure and ethical pathways to implementing AI/ML models.

In late 2023, the Government Accountability Office (GAO) released a report on the use of AI/ML across federal agencies. GAO noted that most agencies have data gaps and inaccuracies in the data used for training and development of AI/ML models. The report observed that the government's management of its use of AI is hindered by incomplete and inaccurate data.

Making confident strides: developing and sustaining AI/ML models

Agencies must secure private and confidential data during AI/ML model development, training, and testing. Production data that includes personally identifiable information, public health information, or confidential data are not appropriate in the model lifecycle. Rather, synthetic data can improve data privacy and security, enhance model performance, and accelerate AI development cycles by providing readily available training data.

The idea of creating synthetic data is not new. However, traditional rule-based or statistical approaches have their limitations:

- Limited control: these approaches rely on predetermined rules that can be inflexible, limiting the variety and complexity of data they can generate.
- Lack of realism: while data may appear statistically similar, often that data lacks the nuances associated with real production data.
- Bias perpetuation: Synthetic data can inherit biases that exist in the production data.

Instead, we suggest a **Synthetic Data Generation Framework**, informed by real-world experience and industry best practices. A key accelerator within our Framework is Unissant's **Profile-based Synthetic Generator**. When implemented correctly, this accelerator delivers realistic results while preserving confidential information.

Profile-based synthetic data generation allows developers to construct and test AI/ML models without sacrificing precision. It provides datasets with fully controlled class distribution, randomization methods, and programmatic constraints, all customized to specific use cases.

The result, the ability to leap across common obstacles, or rocks, on the path to secure AI models that protect civil liberties and sensitive data.

Avoiding rocks on the path: the role of profile-based synthetic data generation



Incorporating profile-based synthetic data generation at the early stages of model development establishes a secure and efficient system for AI/ML deployment. By creating artificial datasets drawn from real-world data profiles, organizations can generate extensive, dependable datasets. These datasets provide realistic insights sufficient for effective AI/ML model training and development while preserving anonymity and ensuring sensitive information remains protected. Moreover, by adjusting the parameters of data generation, we can significantly reduce the impact of both implicit and explicit bias.

Unissant's **Synthetic Data Generation Framework** embraces a blend of tactical approaches and methodologies for data generation. Controlled randomness aids in the production of varied and distinct data points that align with the original data distributions. Use of histograms preserves statistical attributes of input datasets; this ensures that synthetic data accurately mirrors the original dataset, all while safeguarding individual privacy.

Undoubtedly, agencies face various hurdles when developing and sustaining AI/ML. Let us explore the "rocks" threatening a smooth path and how Unissant's Synthetic Data Generation Framework and enabling accelerator, Unissant's Profile-Based Synthetic Data Generator, help agencies stay safely and securely on the road to success.

Rock 1: Lack of data for training models

Data privacy, sensitivity, and confidentiality is a major concern when considering how to make data available to train AI/ML models. Data that is sensitive in nature is not appropriate for developing and training AI/ML models in non-production environments.

	
<p>The federal medical research community may use x-rays or pathology images to predict the likelihood of certain types of cancer. As Protected Health Information, these data would be difficult to incorporate into non-production environments for the training and testing of models.</p>	<p>Satellite imagery, drone data, and facility security footage provide critical insights to national security missions. Failure to sanitize and anonymize this data before use in models could introduce risks to missions and lives.</p>

Unissant’s approach

As a flexible accelerator, Unissant’s Profile-based Synthetic Data Generator can be configured to generate any volume of data. Additionally, we can daisy chain different data profiles for different use cases and generate use case-specific synthetic data as needed.

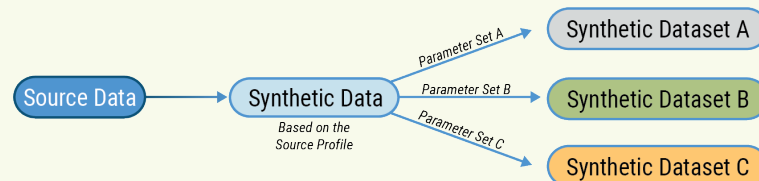


Figure 1: Configurable data generation flow

We also configure algorithms to generate realistic PII data for names, addresses, or phone numbers and specify the distribution curve to either 1) match the real data or 2) have a custom distribution. For example, for a dataset with city and state as attributes for address, we can configure the Synthetic Data Generator to match the city and state distributions observed in the source data or be balanced at, say, 100 rows per city or state, or any combination thereof.

Rock 2: Incorrect, non-representative data

AI/ML model development and training are data-hungry operations requiring large volumes of data that cover all aspects of the use case. The principle of Garbage-In-Garbage-Out (GIGO) holds true for the AI/ML domain, as models trained with incorrect or non-representative datasets produce predictions that are incorrect or biased. The volume of data used to train models invariably determines the accuracy of the predictions.

Data must be both correct and comprehensive, covering each aspect of the inference use case. For example, for a classification model, training data must contain a balanced representation of the classes for the predictions that the model is expected to make. Training data imbalanced toward one possible classification will produce a biased outcome.

Unissant's approach

Using profile-based synthetic data generation, we follow a similar distribution of attribute values as observed in the source data. This allows us to build a true and realistic representation of the source data for the secure training and development of models without the risk of production data disclosure.

Rock 3: Trust in synthetic data

When using synthetic data, it is important to establish its authenticity and ensure its effectiveness in training models. Given the importance of data privacy, agencies must understand the intricacies of synthetic data and its role in data security.

Use of legacy approaches such as random synthetic data generation create inaccurate data, leading to incorrect training of models. For example, a national security component analyzing potential threats from foreign travelers may wish to use synthetic data generation to create profiles of suspicious individuals. Random data generation could create entities with illogical combinations, like an 80-year-old athlete or a toddler experienced in international espionage.

Unissant's approach

Profile-based synthetic data generation publishes data profiles for the source data and the generated synthetic data, allowing for thorough comparative analysis of source and generated data. Experts can fine-tune attributes that digress from the source data and feed that back into the model as part of the reinforcement learning loop. In the national security example, adopting profile-based synthetic data generation avoids the creation of unrealistic profiles, allowing analysts to focus on potentially risky individuals.

Publishing the data profiles used to train the AI/ML model helps data scientists validate the predictions made by the model. This brings accountability into AI/ML predictions, supporting core tenets of the Executive Order for building transparent and trustable AI/ML use cases.

Agencies find that datasets produced using Unissant's Profile-based Synthetic Data Generator are statistically similar to real-world data and preserve its distributions as seen in the histograms for each data attribute.

Rock 4: Perpetuating bias

Randomized and statistical approaches to synthetic data generation rely on existing datasets that may contain inherent biases. Data generated through profile-based approaches can inherit these same biases.

To handle bias, we must manage the datasets used for generating synthetic data and the data used in training the AI/ML models. CDOs and data providers must ensure that the datasets used for training represent a broad spectrum of attributes and attribute values that cover all possible ranges and use cases.

Unissant's approach

Applying Unissant's Profile-based Synthetic Data Generator, experts can define specific characteristics of synthetic data, allowing for profiles that minimize bias. Well-documented profiles make it easier to identify and mitigate potential biases within the profiles themselves.

Adjusting attribute values supports the generation of multiple data profiles with adjusted attribute value ranges to produce multiple training datasets. Data scientists and AI/ML experts can then evaluate the synthetic data to identify and address potential sources of biases. Production data remains untouched and can be profiled again for comparative analysis.

Data trends change, making the ongoing evaluation and audit of AI/ML systems essential. Good model hygiene helps identify potential biases for mitigation. Regular audits help ensure AI/ML systems conform to the organization's ethical standards and are in line with regulations. These processes can be automated using data profiles and version control, allowing teams to proactively monitor drift and flag models for retraining when the drift is beyond an acceptable threshold.

Rock 5: One size does not fit all

Different use cases may need different distributions for data attributes. Data sharing may require greater considerations for anonymization while profile sharing may demand a more nuanced approach to data generation and anonymization. Tailoring the approach to the use case helps ensure that the synthetic data is fit for purpose and positioned to provide reliable insights for data-driven initiatives.

Unissant's approach

We configure our Profile-based Synthetic Data Generator to adjust individual attribute distributions to match a specific use case as required. Generated data can cater to different scenarios, including scenarios not covered by the source data. We consider potential changes in data and prepare for the future by capturing those use cases during the AI/ML model training process.

For example, an agency focused on public health plans to develop AI/ML models to predict and prepare for potential disease outbreaks in a given region. Accurately predicting outbreaks is crucial, but relying solely on historical data has limitations: new diseases emerge, population demographics shift. In lieu of real-time disease or demographics data, synthetic data generation allows public health agencies to define profiles representing different population segments (e.g., age, vaccination rates). Algorithms can create synthetic datasets that consider variables such as change in demographics, emergence of new disease strains, or increases in vaccination coverage due to targeted outreach campaigns.

Reaching your destination: secure, ethical models

Unissant's Profile-based Synthetic Data Generator offers a compelling solution for the secure creation of large datasets, generating values that are realistic and synthesized. Adopting a Synthetic Data Generation Framework minimizes exposure risk while enabling agencies to confidently comply with regulations for the handling of sensitive data.

To maximize benefits, agencies should tailor their synthetic data generation strategy to their unique use cases and objectives. Common elements include best practices such as encryption, anonymization, masking, data attribute concealment, and user authorization when necessary. Monitoring generator performance ensures outputs meet expectations while mitigating risks.

Collaborating with our clients, we help agencies mature their data strategies and achieve mission results. Unissant's Synthetic Data Generation Framework helps organizations take control of AI/ML data flows while safeguarding confidential information. Incorporating our Framework into your data management strategy supports compliance with the Presidential EO and the Federal Data Strategy. Our Profile-based Synthetic Data Generator tracks data drifts in the source data over time for proactive tracking of AI/ML model performance.

A mature approach to synthetic data generation lays the groundwork for a future where AI/ML technologies work responsibly and effectively for the benefit of all. Together, we can navigate the wave of AI/ML innovation with due diligence, consideration, and respect for the principles of justice, fairness, and privacy.

About Unissant

Mission-focused, data-driven—Unissant Inc. (Unissant) delivers for the agencies that keep our nation healthy and safe. Keeping people and mission at the forefront, we apply our domain expertise, data acumen, and technology know-how to achieve breakthrough results. Agencies turn to Unissant for our expertise in AI, advanced analytics, digital excellence, and cybersecurity solutions. Our proven frameworks drive successful execution of complex projects at enterprise scale. With an unwavering commitment to advancing mission outcomes, our teams engineer human-centered, innovative solutions that accelerate time to value. We bring honesty, integrity, and dependability to every interaction with our employees, clients, and partners.

For more information, visit us at www.unissant.com.