

Whitepaper



Synthetic Data Generation

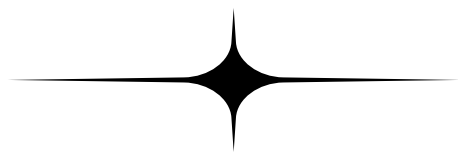


dataxid.com | hello@dataxid.com

Executive Summary

AI represents a strategic priority for organizations pursuing digital transformation across industries. Despite sustained investment, many AI initiatives fail to deliver measurable business value. The primary constraint is not the maturity of models or algorithms, but restricted access to data that is usable, compliant and suitable for enterprise-scale deployment. Organizations face a data bottleneck when there is not enough high-quality data available, or when existing data cannot be safely used or shared.

Synthetic data addresses this challenge by enabling the use of statistically meaningful, high-volume datasets without exposing real and sensitive data. It allows organizations to share data securely, accelerate AI development, improve model reliability and scale innovation with confidence.



DataXID delivers this transformation through a secure, high-fidelity synthetic data engine powered by advanced generative models. The platform enables organizations to unlock the full economic value of their data assets, reduce development cycles and operational costs, and accelerate the deployment of AI solutions. By enabling compliant data sharing and faster innovation, DataXID helps organizations improve return on AI investments, strengthen competitive positioning and build responsible AI systems aligned with evolving regulatory frameworks, including the AI Act.

CONTENT

dataxid.com

- 1 INTRODUCTION
- 2 WHAT IS SYNTHETIC DATA?
- 3 BENEFITS OF SYNTHETIC DATA
- 4 USE CASES OF SYNTHETIC DATA
- 5 BEST PRACTICES FOR GENERATING SYNTHETIC DATA
- 6 SYNTHETIC DATA GENERATION METHODS
- 7 TYPES OF SYNTHETIC DATA
- 8 GENERATIVE MODELS FOR SPECIFIC DATA TYPES
- 9 SYNTHETIC DATA VS REAL DATA VS MOCK DATA

Introduction

In recent years, organizations have increased their investments in AI to improve operational efficiency, reduce costs and enable new business models. Despite major advances in model architectures and computational capacity, the impact of these investments remains constrained by a persistent and critical challenge: access to high quality, safe and usable data.

Data has become simultaneously the most valuable and the most restricted organizational asset. Highly sensitive sectors such as finance, healthcare, telecommunications and public services generate large volumes of data, yet privacy and regulatory requirements, and fragmented data silos limit its use into software and AI pipelines. Also, the global regulatory landscape is evolving toward significantly stricter oversight of data and AI practices. Regulatory frameworks such as GDPR, KVKK, HIPAA and EU AI Act establish explicit expectations for transparency, accountability, risk management and robust data governance across the AI lifecycle. This context requires a fundamental shift in how data is accessed and used. Sustainable AI innovation can no longer depend solely on real world data; it must rely on privacy preserving, ethically aligned and regulation compliant data foundations by design.

Synthetic data provides a practical response to this shift. It enables the creation of statistically representative datasets without exposing sensitive or identifiable information. As a result, organizations can address data scarcity, reduce compliance risk, enable secure data sharing and support fairness across AI systems. Synthetic data transforms otherwise restricted datasets into reusable and shareable assets while maintaining strong protections for personal, confidential and operationally sensitive information.

DataXID is positioned as a key enabler of responsible and scalable AI ecosystems. By combining advanced generative models with a privacy first architecture, DataXID helps organizations to accelerate AI development, meet evolving regulatory expectations and unlock the full value of their data assets with confidence. The next phase of AI adoption will be defined not only by algorithms, but by the quality, safety and availability of the data that supports them. Synthetic data plays a central role in this transition, and DataXID is designed to enable it.

What is Synthetic Data?

Synthetic data refers to artificially generated data that preserves the statistical patterns, correlations and structural properties of real datasets without containing information related to real individuals or events. Rather than masking, anonymizing or modifying sensitive data, synthetic data is produced through generative models that learn the underlying behavior of the original data and generate entirely new, artificial samples.

Modern generative techniques, including GANs, VAEs, and LMs-based architectures, enable the generation of high-fidelity synthetic datasets for tabular data, time-series, text, images and multimodal formats. Crucially, synthetic data does not replicate or expose any real entries; it captures the underlying probability distribution and structural relationships of the dataset while eliminating the privacy risks associated with personal data. Core properties of synthetic data:

- **Privacy-Safe:** No one-to-one mapping to real individuals; supports compliance with GDPR, KVKK and AI Act compliance.
- **Statistically Accurate:** Maintains distributions, correlations, variances, rare patterns and edge cases.
- **Machine-Ready:** Designed for direct use in AI/ML workflows, including model training, simulation, testing, LLM fine-tuning and secure data sharing.

Synthetic data is a high-utility, low-risk and fully shareable data asset that enables organizations to extract value from data without exposing sensitive information.

What Synthetic Data Is NOT?

Synthetic data is not anonymized or masked or partially modified real data. It is not produced by removing identifiers, perturbing sensitive attributes, or transforming real samples. These methods still carry re-identification risks and rely on real individuals' data. Synthetic data is also not mock data. Mock data is manually crafted and fails to preserve the real statistical behavior needed for AI and analytics.

Synthetic data represents a fundamentally different paradigm: the value of real data without exposing real data.

Benefits of Synthetic Data



Privacy, Compliance and Risk Reduction

Eliminates re-identification risks and supports regulatory compliance.



Cost Efficiency

Cuts data acquisition, anonymization and governance costs significantly.



Operational Simplicity

Removes data bottlenecks, accelerates approvals and development.



Scalability and Flexibility

Generates unlimited, scenario-specific data for large-scale AI workflows.



Improved AI Performance

Strengthens model accuracy by filling gaps and simulating rare cases.



Fairness & Bias Mitigation

Enables targeted balancing to support fair and bias-resistant AI models.



Enhanced Collaboration and Innovation

Enables data sharing across teams and partners without privacy risks.



Cloud-Scale Experimentation & AI Development

Supports large-scale cloud experimentation while maintaining privacy.

Benefits of Synthetic Data

Synthetic data provides a rare combination of high analytical value and minimal privacy risk, which empowers organizations to develop next-generation data-driven systems without compromising sensitive information. The following seven benefits represent the core strategic advantages of synthetic data.

1. Privacy, Compliance and Risk Reduction

Synthetic data is privacy-preserving by design and contains no information that can be linked to real individuals. Thus, it eliminates the re-identification risk that can remain in anonymized or masked datasets. This fundamental property supports organizations to achieve demonstrable compliance with regulatory frameworks, including GDPR, KVKK, and EU AI Act.

2. Cost Efficiency and Time-to-Value

Time is often the largest cost in data-driven projects. Traditional data collection, preparation and labeling can take months or years, while synthetic data enables high-quality datasets to be generated within minutes or hours. In addition, lengthy approval and compliance processes frequently delay projects and increase opportunity costs. This reduction in time-to-data lowers overall costs, accelerates AI deployment and allows organizations to gain competitive advantage through faster innovation.

3. Operational Simplicity

Data workflows become significantly more streamlined when workflows are no longer dependent on real data. Synthetic data accelerates internal approvals, shortens development cycles and eliminates the bottlenecks caused by restricted or siloed datasets. This leads to faster experimentation, lower coordination costs, easier deployment and more agile project execution.

4. Scalability and Flexibility

Synthetic data can be generated in virtually unlimited volumes and tailored to a wide range of complex scenarios, from rare events to demographic balancing.

Because development no longer depends on the use of real data, synthetic datasets can be produced and used efficiently in cloud environments. Thus, it allows scalable and flexible AI workflows while reducing infrastructure and hardware costs. Also, it improves the effectiveness of AI development, testing and simulation, while shortening data availability timelines from months or years to weeks. This combination of scalability and flexibility delivers clear cost advantages and strengthens competitive positioning.

5. Improved AI Performance

The performance of machine learning models is fundamentally driven by the quality, diversity and availability of the data used for training, while real-world data collection is often time-intensive and costly. Synthetic data strengthens model robustness and accuracy by addressing data coverage gaps and enabling systematic simulation of edge cases without the delays associated with prolonged data acquisition. Models trained on high-quality synthetic datasets demonstrate improved generalization, greater stability and stronger predictive performance, particularly in domains where real data is limited. Higher model quality reduces retraining costs, lowers failure rates in production and increases the overall return on AI investments.

6. Fairness and Bias Mitigation

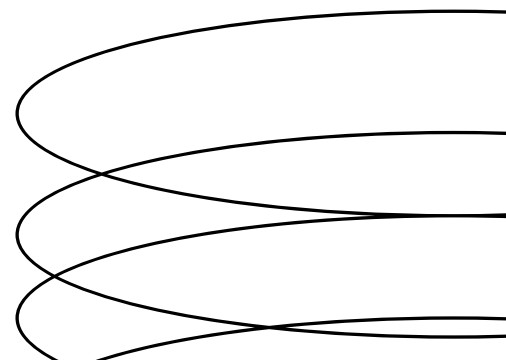
Synthetic data enables the generation of sufficient and statistically representative samples for underrepresented classes without introducing privacy risk. As a result, predictive performance for these groups improves, while bias-related remediation efforts decrease and long-term model sustainability is strengthened. It also supports the development of fairer and more robust models aligned with ethical principles and regulatory expectations, including the fairness requirements of the EU AI Act.

7. Enhanced Collaboration and Innovation

Synthetic data removes privacy and confidentiality barriers that limit data sharing across organizational boundaries. It enables secure collaboration among internal teams, external partners, vendors and research institutions without exposing sensitive information. In cases where real data cannot be shared externally or processed in cloud environments, organizations are often constrained by costly and complex on-premise development models. Synthetic data overcomes these limitations by allowing safe data usage beyond internal systems and by reducing infrastructure overhead and operational complexity. This capability accelerates innovation cycles, supports data exchange and monetization models, and allows organizations to unlock value from data assets that were previously inaccessible.

8. Cloud-Scale Experimentation & Flexible AI Development

Synthetic data helps organizations, particularly in highly regulated sectors, to leverage the flexibility and scalability of cloud computing without exposing sensitive information. Privacy-preserving synthetic datasets allow teams to conduct large-scale experimentation, perform model optimization, execute extensive simulations and utilize high-performance computing resources in public cloud environments with confidence. This model supports a hybrid development approach in which cloud-based experimentation and validation are performed on synthetic data, while real data remains restricted to controlled environments. Such an approach combines scalable computational capacity with regulatory alignment and enables significantly shorter innovation cycles.



Use Cases of Synthetic Data

AI/ML Model
Development and
Training

Internal Data Sharing
and Enterprise
Analytics

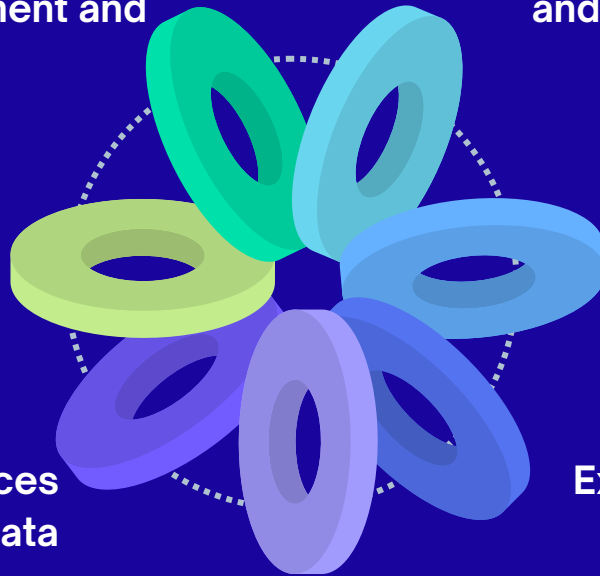
Software
Development, Quality
Assurance and Testing

Cloud-Scale Simulation,
Digital Twins and
Infrastructure Testing

Data Marketplaces
and Synthetic Data
Exchanges

External Collaboration
and Ecosystem
Partnerships

SLMs/LLMs
Fine Tuning



Use Cases of Synthetic Data

Synthetic data is not only limited to privacy protection it also plays a strategic role in driving innovation across the full data and AI lifecycle. It addresses constraints related to data availability, privacy, sharing and scalability and supports a broad range of enterprise use cases. The following use cases demonstrate how organizations across industries apply synthetic data to accelerate innovation, reduce risk and create measurable business value.

1. AI/ML Model Development and Training

A major constraint in AI/ML initiatives is the long time and high cost required for data collection and labeling, which often delays project initiation. Synthetic data removes this barrier by allowing model development to begin with limited real data, without waiting for extended data accumulation cycles.

When sensitive data cannot be used due to privacy constraints, organizations often turn to alternatives such as mock data, anonymized or masked datasets, or generic open-source data. However, these alternatives typically fail to capture the specific statistical properties and behavioral patterns of real-world data, which limits model performance in production environments. In contrast, models trained on high-fidelity synthetic data are able to achieve results closely aligned with outcomes obtained from real data.

Furthermore, in domains such as fraud detection, anomaly detection, risk assessment and healthcare analytics, where data distributions are highly imbalanced, synthetic data increases representation of rare cases and enables simulation of previously unseen scenarios. This capability leads to measurable improvements in predictive outputs and model robustness, enhances operational reliability and strengthens the overall return on AI investments.

2. Software Development, Quality Assurance and Testing

In practice, teams often rely on mock or anonymized datasets in situations where sensitive production data cannot be used; however, these alternatives rarely reflect real system behavior and frequently fail to expose performance bottlenecks, edge conditions or data related failures.

Synthetic data provides realistic and privacy safe test scenarios, stress conditions and boundary cases that help identify system weaknesses prior to release. With the ability to refresh test data on demand, continuous integration and automated test pipelines operate with greater speed and consistency. This approach improves software reliability, reduces production incidents and shortens release cycles, leading to lower operational costs and stronger returns from software and data-driven investments.

3. Data Marketplaces and Synthetic Data Exchanges

Synthetic data plays a central role in the development of data marketplaces and exchange platforms as organizations seek new sources of value creation from data assets. It provides organizations with a structured approach for the commercial distribution of privacy safe datasets. This approach creates new revenue opportunities from internal data that cannot be shared in its original form. Industry specific synthetic datasets can be distributed to partners, vendors and developers with no regulatory exposure, lower compliance costs and faster commercialization cycles. This approach increases data utilization, strengthens monetization strategies and contributes directly to improved returns on data and AI investments, while supporting a collaborative data economy.

4. LLM/SLM Fine Tuning

Large and small language models typically require high volumes, high quality and well structured data to achieve strong performance. In practice, this requirement is constrained by several factors, including limited data availability, concerns around privacy and data misuse, and the significant time and computational costs associated with large scale model training. Hardware limitations, GPU availability and infrastructure expenses further increase the barrier, particularly when training must remain on premise rather than in cloud environments. These constraints collectively slow down model development and increase the overall cost of LLM and SLM projects.

Also, organizations increasingly seek to develop domain specific or organization specific language models tailored to their own data and use cases. However, many are unable to apply data efficient fine tuning in a way that meaningfully reflects the impact of their data on model behavior. Insufficient data volume, privacy restrictions, high infrastructure costs and limited access to scalable cloud resources often prevent effective fine tuning efforts.

Synthetic data addresses these challenges by enabling the generation of privacy safe, scalable and statistically representative datasets derived from existing data. This approach allows organizations to expand training datasets to cloud scale, reduce dependency on specialized hardware and perform fine tuning without exposing sensitive information. Thus, it significantly lowers cost and time barriers and accelerates the adoption of customized language models.

5. Internal Data Sharing and Enterprise Analytics

Data silos continue to limit innovation in large organizations. Privacy constraints, governance requirements and uncertainty around sensitive data elements often prevent teams from accessing shared datasets. In many cases, these limitations create substantial hidden costs, as analytical work slows down, efforts are duplicated and coordination costs increase across business units. Lengthy approval processes for data access and sharing further delay analysis and reduce overall productivity and time to insight.

Synthetic data addresses these challenges by providing privacy safe and high fidelity representations of production datasets that can be accessed and analyzed across departments without exposing personal information. Faster experimentation and insight generation improve scalability, while data evolves into a strategic enterprise asset that strengthens decision making and return on data and AI investments.

Breaking Internal Silos with Synthetic Data

Several organizations adopting synthetic data have reported dramatic improvements in internal collaboration. By generating synthetic copies of their core operational datasets, hundreds of analysts across R&D, risk, operations and marketing were able to run queries without waiting for lengthy compliance reviews. Development cycles shortened from months to days, and synthetic data became the strategic bridge connecting previously isolated teams. This model unlocks latent value in organizations and accelerates decision-making across the entire enterprise.

6. External Collaboration and Ecosystem Partnerships

Synthetic data supports collaboration with external researchers, universities, technology partners, startups and international consortia while preserving confidentiality and data protection requirements. Sensitive datasets can be shared in synthetic form without disclosure of personal or proprietary information, which enables joint research, benchmarking activities and coordinated development efforts across multiple organizations. As a result, organizations can participate in collaborative innovation ecosystems without the legal, compliance and reputational risks associated with real data sharing.

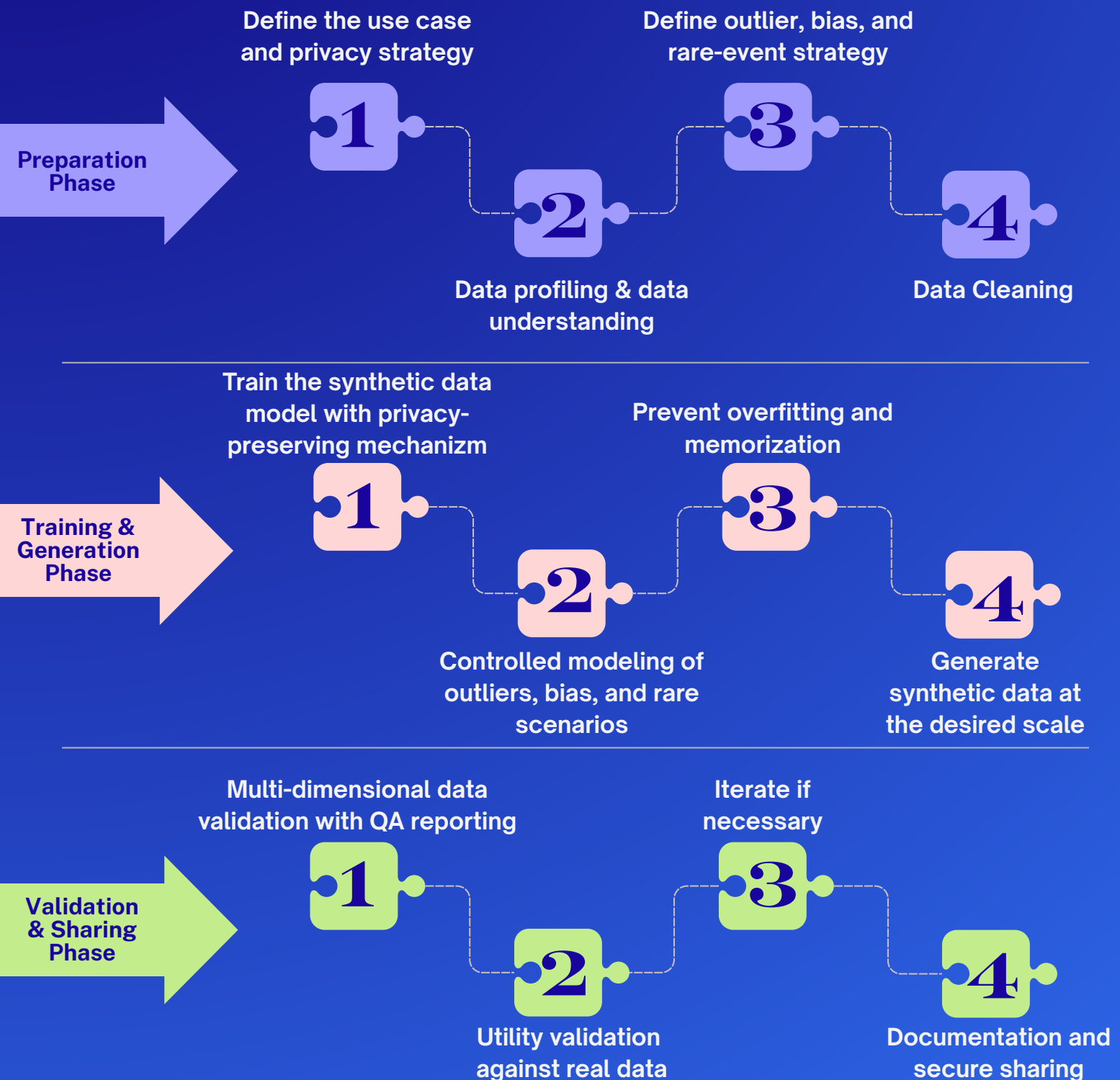
Access to privacy safe and shareable data reduces barriers to partnership formation and shortens project initiation timelines. Collaborative projects can progress with lower legal overhead, reduced negotiation effort and faster alignment among stakeholders. This capability is particularly valuable for AI initiatives that depend on diverse, multi party data inputs across a wide range of industries with complex data requirements. Over time, improved collaboration efficiency leads to faster innovation cycles, lower project costs and stronger returns on data-driven investments.

7. Cloud-Scale Simulation, Digital Twins and Infrastructure Testing

Synthetic data supports simulation and digital twin environments by providing datasets at the required volume for scenario analysis, system validation and performance assessment. Synthetic datasets can be adapted to different operational conditions and test requirements, which allows organizations to evaluate complex system behavior, test extreme conditions and validate infrastructure designs. This approach reduces preparation effort, lowers testing costs and shortens validation timelines for complex systems.

In cloud based architectures, synthetic data supports simulation and testing workflows that scale with available compute resources, without long term infrastructure commitments. This approach is particularly valuable in complex and cost sensitive domains such as smart cities, manufacturing, logistics and mobility. By accelerating validation cycles and reducing infrastructure dependency, synthetic data improves infrastructure resilience, lowers operational risk and enables faster returns on investments in digital twin and AI systems.

Best Practices for Generating Synthetic Data



Best Practices for Generating Synthetic Data

Effective synthetic data generation requires a structured, end to end methodology that balances data utility, privacy protection, regulatory compliance and operational robustness. Synthetic data generation follows a multi phase process that spans strategic planning, model training and controlled data release, rather than a single modeling activity. Each phase addresses distinct technical and governance challenges by ensuring that synthetic datasets are not only statistically valid, but also safe to use, aligned with regulatory requirements and fit for real world analytical and AI workloads. The following best practices outline a three phase framework that organizations can adopt to systematically design, generate, validate and share high quality synthetic data.

1. Preparation Phase

The preparation phase defines what type of synthetic data will be generated, for which purpose, and under which privacy, regulatory and risk constraints the process will operate. During this phase, business objectives are aligned with data characteristics and relevant regulatory and compliance requirements. This alignment ensures that model training and data generation steps are technically robust and clearly justified from both governance and regulatory perspectives. A well designed preparation phase establishes the foundation for high data utility, structural consistency and responsible synthetic data generation that supports analytical and AI use cases.

- **Define the Use Case and Privacy Strategy:** In this step, the purpose and scope of the synthetic data generation process are clearly defined. The target use case, such as machine learning model training, software testing, data sharing or LLM fine tuning, determines which statistical properties must be preserved and which variations are acceptable. At the same time, relevant regulatory frameworks such as GDPR, KVKK, EU AI Act and sector specific requirements are identified. Based on the defined use case, an appropriate privacy strategy is selected. The selected privacy approach directly shapes the complexity of the generative model and influences the level of training effort required in subsequent phases.

- **Data Profiling and Data Understanding:** A thorough profiling of the source data is essential before any synthetic modeling begins. This step involves analyzing variable distributions, valid ranges, data types and schema consistency, as well as identifying correlations and dependencies across features. Early signals of bias, sparsity and outliers are examined to determine whether they reflect meaningful real world patterns or underlying data quality issues. In addition, attributes that require additional privacy consideration are identified to guide modeling choices and the selection of suitable privacy preservation techniques.
- **Data Cleaning:** Once the structure and behavior of the data are understood, the dataset is prepared for modeling through systematic cleaning. During this step, missing values are handled using appropriate strategies, invalid or inconsistent instances are corrected or removed. Categorical values are standardized to ensure internal consistency across the dataset. In addition, logical and business constraints that must hold in both real and synthetic data, such as valid value ranges or cross-field dependencies, are enforced at this stage. The objective of data cleaning is to ensure that the generative model learns meaningful data patterns and relationships rather than artifacts, noise, or errors inherent in the raw source data.
- **Define Outlier, Bias, and Rare-Event Strategy:** Not all extreme or rare patterns should be evaluated uniformly in synthetic data generation. In this step, explicit decisions are made regarding how outliers, biases and rare scenarios will be handled. Outliers may be preserved when they are operationally meaningful, generalized when they increase privacy exposure, or excluded when they originate from data quality issues. Bias and fairness considerations are evaluated to determine whether real world distributions should be maintained or adjusted to support fairness and performance objectives. For rare events, the need for conditional generation or scenario specific synthesis is assessed to ensure sufficient representation without compromising privacy.

At the conclusion of the preparation phase, a clean and model ready dataset, a clearly defined use case and privacy framework, and well structured modeling and risk management decisions are established.

2. Training and Generation Phase

The training and generation phase focuses on learning the statistical structure of real data and producing synthetic datasets that preserve analytical utility while preventing the disclosure of any real data. The aim is to generate synthetic data that reflects the statistical characteristics of real data, supports analytical and AI tasks, and prevents exposure of original data. At this stage, decisions defined during the preparation phase guide both modeling and privacy controls.

- **Train the Synthetic Data Model and Integrate Privacy-Preserving Mechanisms:** Model training is performed using a generative approach selected based on the data type, structure and defined use case. During training, the model learns univariate distributions as well as multivariate relationships and dependencies across features, with an emphasis on capturing the data structure. Privacy protection can be embedded directly into the training process rather than applied as a post-processing step. Based on the selected privacy strategy, techniques such as differential privacy and defined privacy constraints are integrated into the model training pipeline.
- **Controlled Modeling of Outliers, Bias, and Rare Scenarios:** Outliers, biases and rare scenarios are modeled in a deliberate and controlled manner to preserve statistical fidelity and help reduce potential risks. Conditional generation techniques may be applied to increase the representation of rare scenarios or to address distributional imbalances, based on the requirements of the defined use case. Extreme values can be generalized to maintain their statistical relevance while avoiding the replication of unique or identifiable individuals.
- **Prevent Overfitting and Memorization:** To prevent overfitting and memorization, synthetic data models must be designed to learn general statistical patterns rather than replicate real samples. This is achieved by controlling model capacity, applying regularization techniques and integrating privacy-preserving constraints such as differential privacy directly into the training process. Continuous validation methods, including similarity and disclosure risk checks, are used to detect and avoid sample-level replication and support that synthetic data remains both privacy-safe and analytically reliable.

- **Generate Synthetic Data at Scale:** Once the model demonstrates stable, generalized and privacy-preserving behavior, synthetic data is generated at the required scale. Scalable generation allows synthetic data to support multiple use cases, including model training, testing, data sharing and system validation, without introducing privacy risk or compromising data quality. This capability is particularly important for data intensive use cases that require large volumes of high quality data.

At the conclusion of this phase, one or more privacy preserving synthetic datasets are produced through a controlled and scalable generation pipeline. These outputs are designed to support a wide range of analytical, development, and data-sharing use cases while maintaining strong privacy protections and consistent data quality.

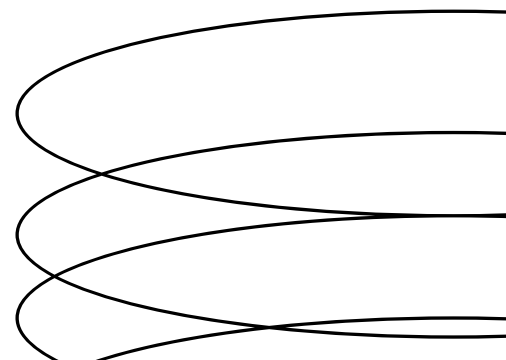
3. Validation and Sharing Phase

The validation and sharing phase provides assurance in the generated synthetic data by demonstrating its data quality, privacy protection and utility for analytical and AI use cases. This phase confirms that synthetic datasets accurately reflect the statistical properties of real data and deliver reliable results across analytics and machine learning applications. In addition, validation ensures that synthetic data can be shared across internal teams and external stakeholders without introducing privacy or compliance concerns.

- **Multi-Dimensional Data Validation with QA Reporting:** In this step, the quality and structural integrity of the synthetic data are evaluated across multiple dimensions. Univariate, bivariate and multivariate distributions are compared between real and synthetic datasets to assess similarity in value ranges and frequency patterns. Multivariate correlations and dependencies are analyzed to verify that relationships across features are preserved. In addition, structural rules and business constraints are validated to ensure that the synthetic data remains consistent with the logical requirements of the real data. Privacy-related indicators are also evaluated as part of the overall QA process.

- **Utility Validation Against Real Data:** Beyond statistical similarity, the practical utility of the synthetic data is assessed by comparing its performance against real data in representative workflows. Train test compare experiments are conducted to evaluate whether models trained on synthetic data achieve results comparable to models trained on real data. Machine learning performance metrics, including prediction accuracy, robustness and fairness impact, are analyzed to confirm that the synthetic data supports the defined analytical and operational use cases.
- **Iteration and Improvement:** Validation results are reviewed to assess whether further improvement is necessary. When predefined quality, privacy, or utility requirements are not satisfied, generative models are retrained or fine-tuned. Accordingly, synthetic datasets are regenerated. The iterative process enables systematic convergence toward a balanced outcome that meets data quality, utility, and privacy objectives.
- **Documentation and Secure Sharing:** The final step focuses on transparency, traceability, and controlled access. Methods, assumptions and known limitations of the synthetic data generation process are documented in a transparent and structured format. Quality assurance, privacy, and utility assessment reports are provided to support internal review or external sharing. Validated synthetic datasets are then made available through controlled access mechanisms in line with defined regulatory and access policies to allow both internal use and external collaboration.

At the conclusion of validation and sharing phase, approved and validated synthetic datasets are produced along with comprehensive quality, privacy, and utility assessment reports. These outputs form shareable synthetic data assets that are ready for responsible use across development, analytics and collaboration scenarios.



Synthetic Data Generation Methods

Synthetic data can be generated using many techniques with different degrees of complexity, statistical fidelity and scalability. Method selection is determined by the data type, use case, required fidelity and privacy constraints. Synthetic data generation approaches can be categorized into rule based and statistical methods, agent based methods and machine learning based generative methods.

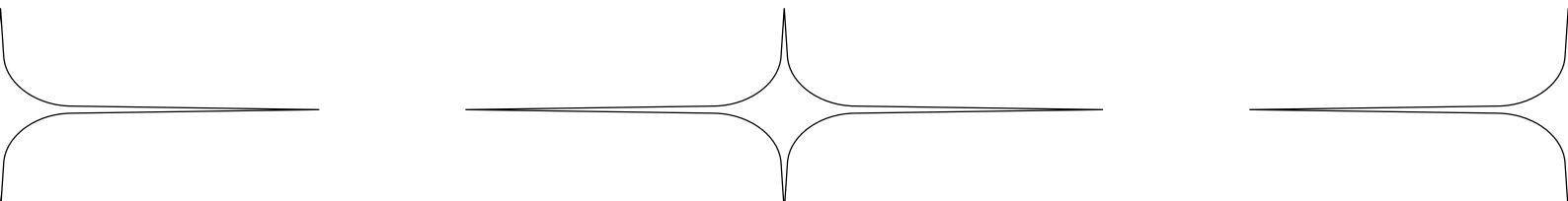
1. Rule-Based and Statistical Methods

Rule based and statistical methods represent the traditional approach to synthetic data generation. These methods are based on predefined rules, probability distributions and statistical assumptions derived from domain knowledge or statistical properties of real data. Common methods include:

- **Marginal Distribution Models (Simple Sampling):** Marginal distribution models generate synthetic data by analyzing the univariate distributions of individual variables (e.g., normal, exponential, uniform) and sampling independently from these distributions. The approach preserves some statistical properties, including mean and variance at the variable level, but it assumes independence across variables and does not capture multivariate relationships.
- **Rule-Based Synthesis:** Rule-based synthesis generates synthetic data according to predefined business rules or logical constraints (e.g., "If Age is greater than 18, then driver's license status must be 'valid'"). The method ensures consistency and compliance with domain logic, but it does not scale effectively for complex datasets with statistical dependencies.
- **Multivariate Parametric Models:** Multivariate parametric models assume that the data follows a predefined joint distribution, such as a multivariate normal distribution. They preserve linear relationships; however, their ability to represent non linear behavior and complex data distributions is limited.

- **Simple Sequential Models:** Simple sequential models, such as Markov chains, generate synthetic sequences based on probabilistic state transitions. These models assume dependence limited to the current state, which constrains their ability to capture long term temporal dependencies.
- **Interpolation-Based Models:** Data space interpolation techniques create synthetic instances by interpolating between existing data points. Although these methods capture surface level structure, they are insufficient for modeling complex, high dimensional data patterns.

Rule-based and statistical methods are easy to implement and interpret, however they offer limited capacity to represent the complex, high dimensional and non linear relationships present in real data. Consequently, these methods are suited for testing, simulation and analytical use cases with simple and well understood data structures, but are insufficient for advanced analytical and machine learning workflows that require realistic multivariate behavior.



2. Agent-Based Methods

Agent-based methods generate synthetic data by representing systems as collections of autonomous entities, referred to as agents, that interact with each other within a defined environment. Each agent follows a set of explicit behavioral rules that determine its actions and interactions. Interactions among agents give rise to emergent system-level patterns. This characteristic makes the approach suitable for modeling dynamic processes that are difficult, costly or impractical to observe directly in real-world environments. Therefore, agent-based methods are applied in domains that involve complex system interactions, including traffic and mobility systems, epidemiological spread, financial markets, supply chains and operational processes.

Agent-based methods offer strong interpretability and allow domain experts to encode causal assumptions directly into the model. However, defining rules that accurately reflect real-world complexity requires significant domain expertise and manual effort.

Model outcomes are highly sensitive to initial conditions and rule definitions. Also, large-scale simulations can be computationally expensive. Therefore, agent-based methods are used for scenario analysis and system simulation rather than large-scale, general-purpose synthetic data generation.

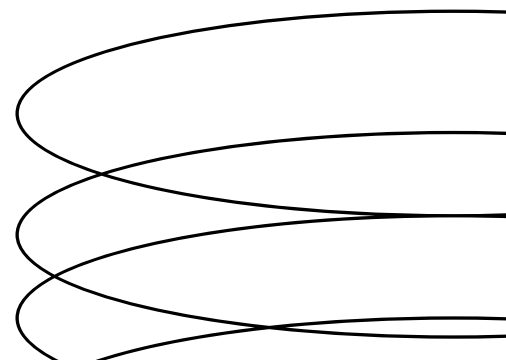
3. Machine Learning Based Generative Methods

Machine learning-based generative methods represent a paradigmatic shift in synthetic data generation by enabling models to learn complex, multi-dimensional and non-linear statistical structures from real data. The methods leverage deep learning architectures to generate synthetic datasets with high statistical fidelity, scalability and privacy-preserving properties. Their key advantage is the ability to model complex dependencies, multi-modal distributions and rare patterns without the need for manual rule definition. Some commonly used approaches include:

- **Generative Adversarial Networks (GANs):** GANs consist of two neural networks trained simultaneously; a generator that produces synthetic samples and a discriminator that attempts to discriminate between real and synthetic data. Training proceeds through an adversarial process between these networks until a stable convergence point is reached, at which the discriminator can no longer reliably differentiate synthetic samples from real ones. GANs are particularly effective at capturing complex correlations and generating high-fidelity, privacy-preserving synthetic data for tabular, image and time-series domains.
- **Variational Autoencoders (VAEs):** VAEs generate synthetic data by learning a probabilistic latent representation of the input data. During training, the model learns to encode input data into a latent space, from which new samples are decoded for data generation. The interpretable latent space of VAEs supports advanced applications such as fairness-focused improvements, conditional generation, and domain-specific data augmentation. VAEs are particularly suitable for scenarios where both statistical consistency and model interpretability are critical.

- **Diffusion-Based Models:** Diffusion-based models generate synthetic data through a probabilistic, multi-step process that gradually transforms structured data into noise and learns to reverse this transformation. During training, noise is incrementally added to real data, and the model learns to reconstruct the data by reversing this process step by step. Once trained, new synthetic samples are generated by starting from noise and iteratively refining it toward realistic data. Diffusion-based models are effective at capturing complex, high-dimensional and non-linear data distributions, including high-resolution images, video and complex tabular data. They provide stable training behavior, although their multi-step sampling process introduces higher computational cost.
- **Language Models (LMs-Based Generators):** Language Models (LMs), ranging from traditional Recurrent Neural Networks (RNNs) to modern transformer architectures, aim to learn the underlying probability distribution of the training data, such as a sequence of words or tokens. This capability allows the models to generate new samples from the learned distribution and to predict subsequent tokens or words in a sequence with high contextual coherence. LMs-based generators are particularly effective for producing synthetic text, including documents, customer interactions, medical notes, source code, and other semi-structured data types. Their advanced natural language processing capabilities enable the preservation of semantic consistency and domain-specific context.

Each synthetic data generation method offers several advantages and trade-offs. In practice, DataXID employs multiple approaches to balance data fidelity, privacy protection and operational efficiency. DataXID automatically selects the optimal generation model according to data characteristics and the target use case, and enhances it with additional privacy, quality and control layers to strengthen overall performance, reliability and governance.



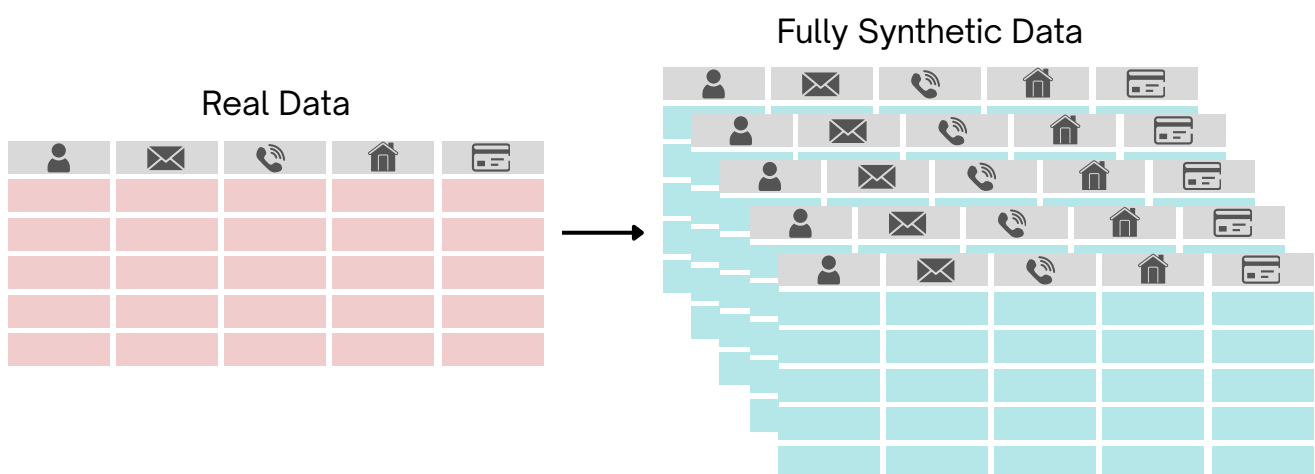
Types of Synthetic Data

Synthetic data can be classified into three types based on how it is generated and applied within data and AI workflows. These types help organizations evaluate the trade-offs between data utility, statistical fidelity and privacy protection across different use cases. The categories outlined below address different requirements related to analytics, model development and data sharing.

1. Fully Synthetic Data

Fully synthetic data consists entirely of artificially generated samples and contains no representations of real data. The generative model learns the statistical distributions, structural patterns and cross-variable dependencies present in the source dataset, and produces a synthetic dataset that preserves these properties without reproducing individual data points.

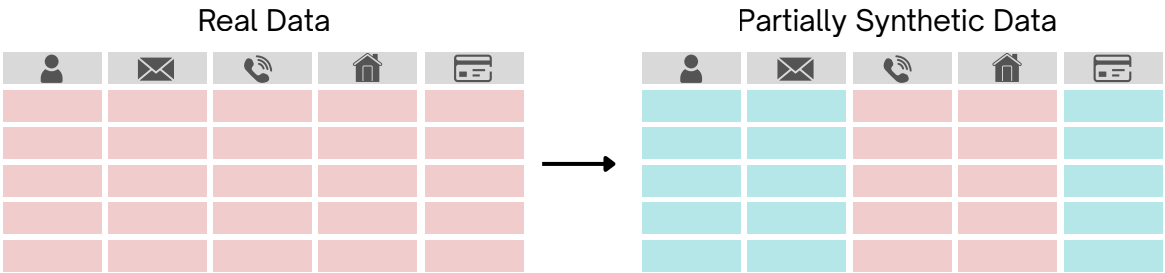
Fully synthetic data offers the highest level of privacy protection and is well suited for secure data sharing, external collaboration and use cases subject to strict regulatory and compliance requirements. Once the model has been trained on a limited volume of real data, it can generate synthetic datasets at virtually unlimited scale, enabling large-scale data availability without requiring continued access to sensitive real data.



2. Partially Synthetic Data

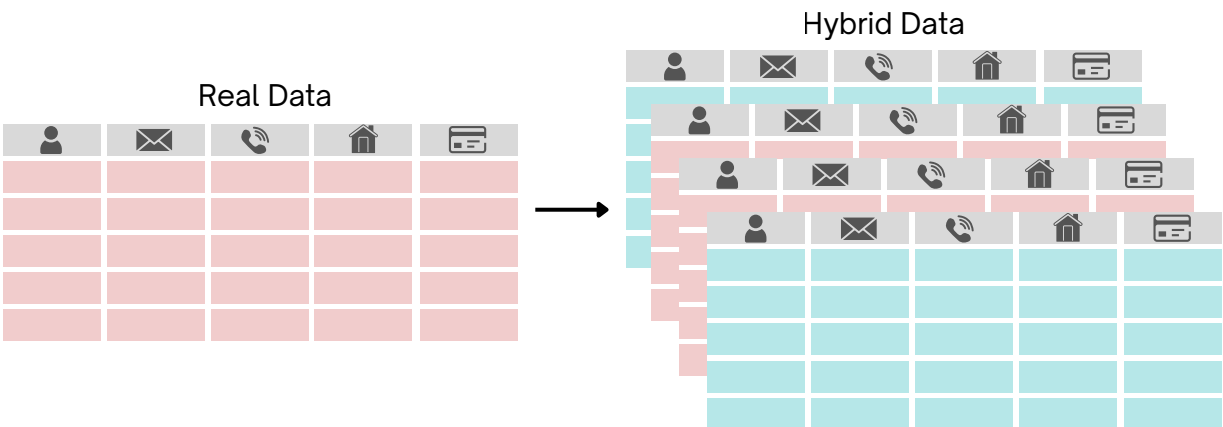
Partially synthetic data is obtained by combining real data with synthetically generated data, in which only selected or sensitive fields are replaced with artificial values while the remaining dataset is left unchanged. This approach is commonly applied to protect specific sensitive information while maintaining a high degree of fidelity to the original data structure and distributions.

Only a subset of fields is artificially generated, while the remaining data fields are retained from the real dataset. Therefore, the size of a partially synthetic dataset is typically limited to that of the real dataset and does not scale beyond the existing number of samples. Partially synthetic data is most often used for internal analytics, controlled testing environments and scenarios where full data replacement is unnecessary but targeted privacy protection is required.



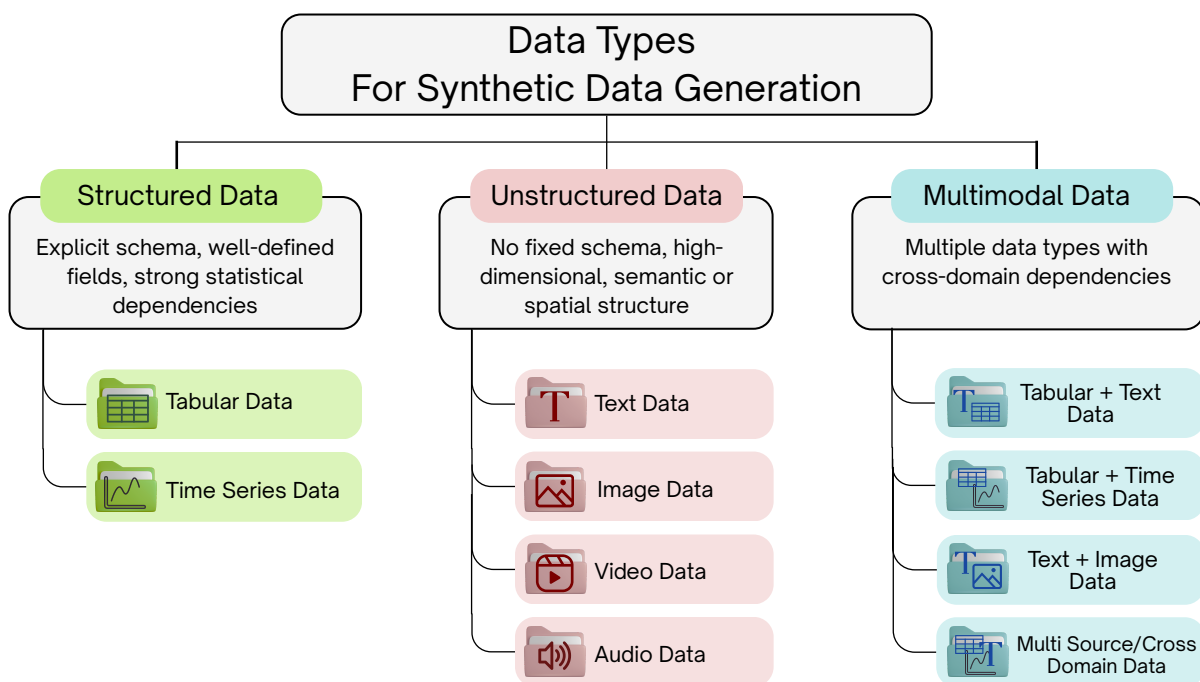
3. Hybrid Data

Hybrid synthetic data is obtained by integrating real and synthetic data at the dataset level. Artificially generated samples are combined with real samples to increase dataset scale, improve class balance or strengthen the representation of rare scenarios. Hybrid datasets are used for model training, stress testing and performance benchmarking when real data is available but insufficient, and additional data diversity is needed.



Data Types for Synthetic Data Generation

In synthetic data generation, data types are classified according to structural characteristics and dependency patterns, rather than file formats or storage representations. Such a classification establishes a clear and methodologically robust basis for selecting generative models and validation approaches aligned with defined use case requirements.



1. Structured Data

Structured data refers to datasets with a defined schema and explicit relationships between variables, including tabular and time-series data. Such datasets are commonly used in analytics and predictive modeling. They require synthetic generation methods that preserve multivariate dependencies, logical constraints. Machine learning-based generative models, such as GANs, VAEs and diffusion-based approaches, are commonly used to learn these joint and non-linear relationships. Tabular data consists of rows and columns with heterogeneous data types and complex inter-column dependencies. Synthetic generation must preserve individual column distributions as well as row-level consistency and semantic validity.

Time-series data represents observations indexed over time and requires modeling temporal dependencies such as autocorrelation, trends and seasonality. Therefore, sequence-aware generative architectures and temporal extensions of deep generative models are preferred for advanced analytical applications.

2. Unstructured Data

Unstructured data does not follow a fixed schema. It is characterized by high dimensionality, semantic richness, or spatial and temporal structure. This category includes text data, image data, video data, and audio data. Synthetic data generation for unstructured data focuses on preserving semantic coherence, contextual consistency, or perceptual patterns specific to each data type. Language models based on transformer architectures are the dominant approach for generating synthetic text and other sequential data, as they learn contextual dependencies at the token or sequence level. For visual data such as images and video, diffusion models and adversarial architectures are adopted due to their ability to produce high-fidelity and coherent outputs. Audio data similarly benefits from sequence-aware or diffusion-based generative techniques.

3. Multimodal Data

Multimodal data combines two or more distinct data types and requires preserving cross-domain relationships between modalities. Common examples include tabular data combined with text, tabular data combined with time-series data, text combined with images, and more complex multi-source or cross-domain datasets. In these settings, synthetic generation must ensure that relationships across modalities remain consistent and meaningful. Multimodal synthetic data generation typically relies on conditional or joint generative architectures that learn aligned representations across data types. Thus, multimodal synthetic data generation depends on conditional or joint generative architectures that learn aligned representations across data types. These architectures support coordinated generation across modalities, either by conditioning one modality on another or by jointly modeling multiple modalities within a shared latent representation. In practice, hybrid strategies that incorporate both real and synthetic data may also be applied to improve coverage or represent rare cross-modal scenarios.

Real vs Synthetic vs Mock

In data-driven world, organizations often must choose between different sources of data depending on availability, privacy requirements and analytical goals. The choice between real data, synthetic data and mock data can significantly impact project outcomes, model performance and the overall success of analytics and AI initiatives.

Real data is derived directly from real-world events and interactions. It represents the ground truth of a given process or domain. Although it offers maximum fidelity by definition, it introduces privacy, security, and compliance risks, particularly when it contains personally identifiable information or sensitive operational details. Therefore, real data is often subject to strict regulatory constraints, access controls, and sharing limitations, which can slow down innovation and hinder collaboration.

Synthetic data addresses many of these limitations by generating artificial datasets that preserve the statistical properties, structural patterns and relationships of real data. Through machine learning-based generation techniques, synthetic data delivers high analytical fidelity while significantly reducing privacy risk. This makes it well suited for AI training, testing, and data sharing scenarios, particularly in regulated environments. Moreover, synthetic data can be generated at scale and adapted to specific scenarios, capabilities that are difficult or impractical to achieve with real data alone.

Mock data is created through manual processes or simple rule-based definitions. It is useful for early-stage development activities such as interface testing, or schema validation, but it does not reflect the statistical distributions or dependencies in real data. Consequently, it provides limited value for advanced analytics or machine learning workflows and does not support regulatory, compliance or representativeness requirements.

Choosing the right data strategy is a critical decision. While real data offers maximum fidelity and mock data supports early development, synthetic data provides the optimal balance of fidelity, scalability and privacy for enterprise-grade AI. DataXID enables organizations to make this shift with confidence, transforming data constraints into a competitive advantage and accelerating the path from experimentation to measurable business value.

Real vs Synthetic vs Mock

Feature	Real Data	Synthetic Data	Mock Data
Definition	Data collected from real-world systems or users	Artificially generated data that reflects real data distributions	Simple, rule-based or manually created data
Data Source	Production systems or actual user/system interactions.	Machine Learning (GANs, VAEs, etc.) or advanced statistical models.	Manual input, basic programming logic, or rule-based generators.
Privacy/Risk	High risk. Contains PII. Subject to privacy regulations.	Low risk. No real data. Can include privacy preservation mechanisms.	Low risk. Usually contains random, meaningless, or generic values.
Fidelity (Realism)	100% (By definition).	High. High statistical fidelity	Low. Only meets format and structural rules.
Scalability	Limited to observed data	Unlimited and scenario-driven	Limited and manual
AI Training	Excellent.	Excellent. In some cases, improves AI performance through better balance and coverage.	Poor for model training
Sharing	Restricted by regulations	Fully shareable. Does not include real data.	Limited. Varies based on data structure.
Compliance	Difficult. Requires effort to comply with regulations (GDPR, HIPAA).	EU AI Act and Regulations-aligned by design	No compliance guarantees

Conclusion

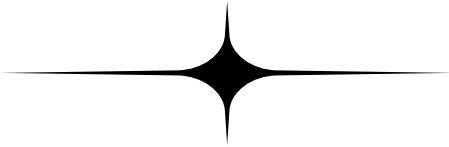
Data-driven technologies have reached a point where progress depends not only on computational power or advanced algorithms, but also on access to high-quality, safe and usable data. Across industries, organizations struggle with limited data accessibility due to privacy constraints, regulatory requirements, high acquisition costs and operational complexity. A paradigm shift is underway that elevates data from a back-office technical resource to a core strategic asset shaping competitiveness, innovation speed and organizational resilience. In this environment, strategies built only on real data increasingly limit growth, collaboration and the ability to scale AI initiatives with confidence.

Synthetic data provides a systematic approach to addressing these structural constraints. By enabling the generation of statistically representative and privacy-preserving datasets at scale, it allows organizations to derive analytical value without exposing sensitive information or introducing governance-related delays. As outlined in this document, synthetic data contributes to core organizational objectives such as shortening development cycles, reducing operational costs, improving model performance, ensuring regulatory compliance and facilitating collaboration across internal and external ecosystems. When applied through a structured, end-to-end methodology, synthetic data functions as a reliable, verifiable and operationally manageable data asset.

Sustainable data-driven growth depends not on having more data capacities, but on the ability to create, manage and trust it at scale. DataXID helps organizations to operationalize synthetic data as a controlled, AI-ready capability for innovation.

About DataXID

DataXID provides an end-to-end data platform delivering synthetic data, data sharing and AI-ready workflows through integrated and modular solutions. DataXID Synthetic Data Platform generates high-quality synthetic datasets using advanced AI models. DataXID helps organizations to develop software and AI systems without relying on sensitive real data. While collecting and using real data is costly, slow, limited and subject to strict regulations such as GDPR and EU AI Act, DataXID offers AI-ready data or test data within minutes with minimal effort. Since synthetic data contains no real samples, it is safe to use and share by design. DataXID reduces time-to-market by up to 50%, improves AI/ML project metric rates by up to 20%, and enables faster, compliant innovation for delivering measurable ROI through lower compute cost and robust governance.



More Information on DataXID: dataxid.com

Contact us: hello@dataxid.com

1773 İTÜ Teknopark, İstanbul, 34467



No data bottlenecks.
Faster innovation.

Book a Demo

A white hand cursor icon with a black outline, pointing towards the "Book a Demo" button. The hand is positioned as if about to click the button.

 dataxid.com | hello@dataxid.com