Softtek®

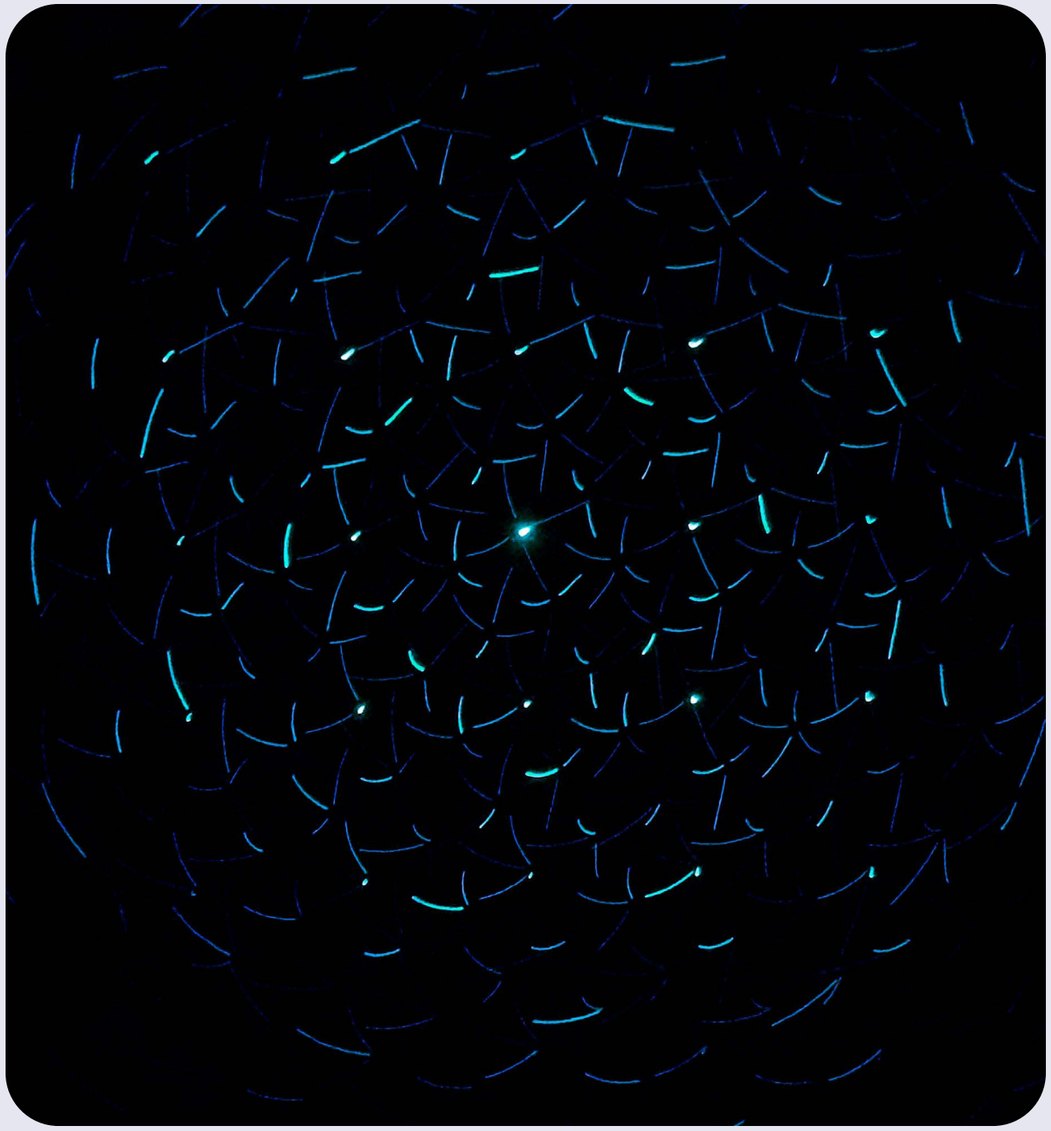
The rise of Synthetic Data: data without borders

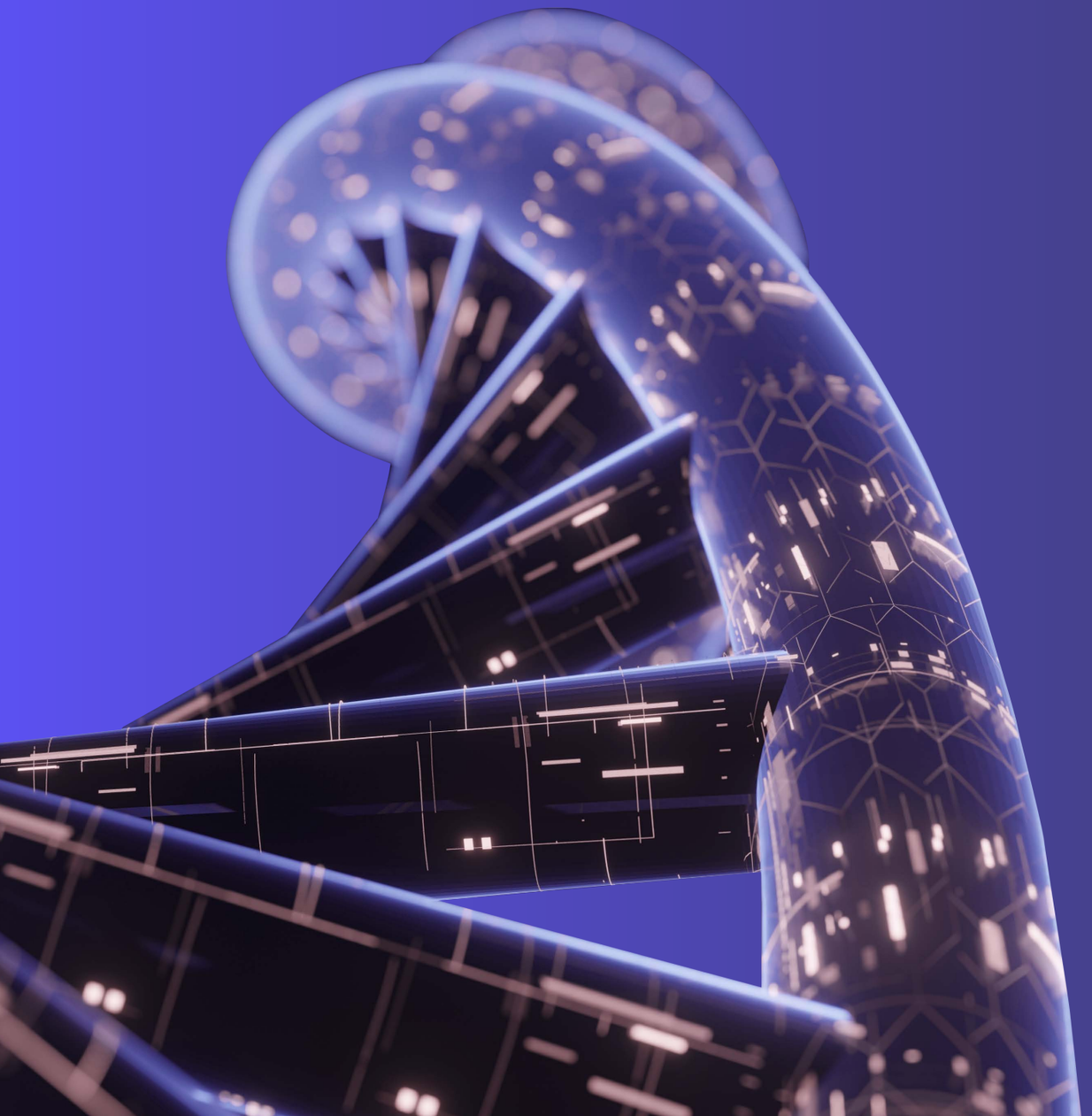
Synthetic data is the key to driving innovation and ensuring privacy in the new era

Synthetic data has emerged as a transformative solution in the tech industry, especially in a world where access to real data is increasingly limited and protected by regulations, laws, and privacy concerns. This technology enables the generation of artificial data that precisely mimics the statistical properties of real data, using advanced techniques such as deep learning and generative AI. From its beginnings in the 1930s with audio synthesis experiments, to its modern application in creating datasets to train AI models in sectors like healthcare, finance, autonomous vehicles, or entertainment, synthetic data has evolved into a key pillar of innovation. According to Gartner, **by 2026, 75% of companies are expected to use generative AI to create synthetic customer data.**

This advancement has direct implications for software development, the evolution of the new era of code, and advanced analytics. **Synthetic data helps overcome obstacles such as data scarcity, high collection costs, and privacy barriers**, allowing companies to build more robust, inclusive, and ethical AI models. As data generation technologies continue to improve, this type of data not only complements real data but also enables the creation of more accurate and efficient AI systems.

Thanks to synthetic data, **companies can feed their algorithms with oceans of data without drowning in the sea of legal restrictions**, thus reconciling AI's hunger for data with the expectations of customers, regulators, and society regarding privacy.





A new paradigm: privacy and innovation in the age of Artificial Data

Data is consolidating as key assets, while its use faces stricter regulations and increasingly rigorous protocols

Data is evolving as a key driver of business transformation

The data market has undergone a **remarkable transformation in recent decades**, driven by technological advances such as Cloud Computing and microservices or the implementation of artificial intelligence in many of its areas. In the past, the collection and use of data largely depended on physical infrastructure, such as leased lines for transmitting data between providers and consumers.

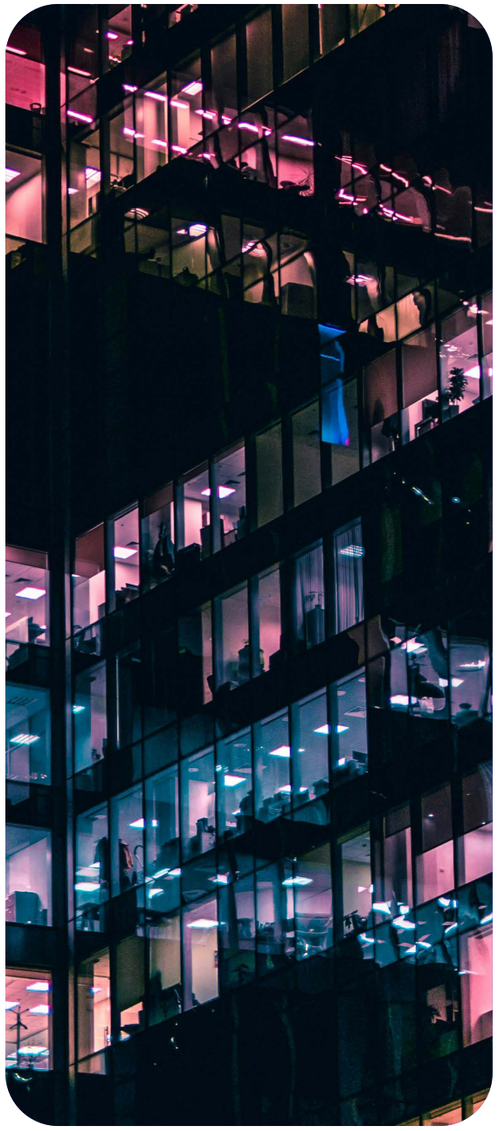
The strategic value of data has become more evident thanks to the fact that organizations have recognized that it is not simply a by-product of their operations, but an essential resource for decision-making, innovation, and the design of new business models.

Today, data is considered a corporate asset that can generate direct value or enable more sophisticated and personalized solutions.

Large corporations have shown that data can be the core of their revenue model, monetizing user information through services such as targeted advertising or personalized recommendations. Besides being a potential revenue source, **data drives innovation through the development of personalized products and the use of artificial intelligence and machine learning.**

180 zettabytes

is the volume of data generated, consumed, copied, and stored projected to be surpassed by 2025.



Source: Synthesis

But data scarcity and exclusivity limit innovation and business collaboration

However, even with this awareness of the importance of data, companies today face a **growing shortage of real data suitable for training AI models.**

Obtaining this data is not only complicated and costly, it also involves long and complex processes such as surveys, controlled experiments, or purchasing licenses to access private databases, which require significant investment. These costs are compounded by expenses on specialized teams to collect and process the information.

Even when organizations overcome these economic barriers, **data quality continues to be a critical problem.** Many datasets contain biases or are incomplete, which directly affects the precision and effectiveness of the models.

Moreover, **the current landscape is characterized by a competitive environment** in which organizations are becoming increasingly protective of their information, hindering collaboration and data sharing, which in turn obstructs the development of collaborative and innovative models. The exclusivity of data has become a limiting factor, closing access to key resources for many companies, especially smaller ones that do not have access to massive data repositories.

The evolution toward synthetic data reflects a strategic response to the growing demands for privacy and regulatory compliance

This landscape is joined by privacy regulations such as the GDPR in Europe and the CCPA in California, **which impose increasingly strict restrictions on the use of personal data.** These regulations require transparency, informed user consent, and the right to access and delete personal data, transforming how information is managed within companies.

Legislation is leading organizations to rethink their data collection and usage practices, forcing them to find alternatives that respect both user privacy and the need to innovate.

In fact, **85% of consumers are fully aware of the value and importance of their data**, and for this reason, protecting customer information is a priority for organizations, with 96% of them recognizing its relevance well beyond what regulations require.

Sources: Martech I Nymiz I Synthesis

How regulatory evolution drives innovation in privacy protection

This regulatory evolution is also driving innovation in privacy protection technologies, or Privacy-Enhancing Technologies (PET). These technologies allow the extraction of value from sensitive data without compromising individuals' privacy. **Methods such as advanced anonymization ensure that data remains protected during processing,** allowing its use for collaborative analysis while still complying with strict privacy laws.

Anonymization removes direct identifiers such as names, addresses, or ID numbers.

However, over time it has been shown that although useful, it is not always sufficient. In many cases, **combinations of seemingly anonymous data can be re-identified**, especially when crossed with other publicly available sources.

Faced with these limitations and the current hyper-protective regulatory pressures, **organizations have begun to adopt more robust approaches**, such as differential privacy and, more recently, synthetic data. These are datasets generated artificially through algorithms, simulations, or AI models.

82%

of companies admit putting themselves at risk when collecting real data

91%

state that they are making greater efforts to reassure their clients about how their data is used with AI

94%

agree that their customers would abandon them if their data is not well protected

Faced with legal and privacy risks, synthetic data emerges as the most robust and scalable option available

Comparative table: Real Data vs. Anonymized vs. Synthetic

Criterion	Real data	Anonymized data	Synthetic data
Privacy	**Low** Contains identifiable information	**Medium** Direct identifiers are removed	**High** Does not contain real data
Re-identification risk	**High**	**Medium** Risk if combined with external sources	**Very low** If well generated
Analytical value	**Very high**	**High** May lose precision	**Low** If original distribution is preserved
Regulatory compliance	**Low** Requires legal justification	**Medium** But subject to regulations	**High** If reversal to real data is not possible
Scalability	**Limited** Due to availability and regulation	**Limited** Due to quality and manual effort	**High** Can generate large volumes
Cost of acquisition/use	**High** Requires permissions, secure infrastructure	**Medium** Semi-automated or manual processing	**Low or medium** Depends on the technology used
Access time	**Low** Legal reviews and edits	**Medium** Anonymization validation required	**High** If demand is predictable

The synthetic data market will grow rapidly, driven by AI, machine learning, IoT, and emerging connected technologies

Global synthetic data growth: driven by AI, regulations, and regional expansion

In fact, the **global synthetic data generation market is projected to reach $1,788.1 million by 2030**, with a compound annual growth rate (CAGR) of 35.3% between 2024 and 2030, mainly driven by the growing adoption of emerging technologies such as AI, ML, and IoT, along with an increase in the use of connected device technologies. In fact, Forbes named it one of the “5 Biggest Data Science Trends in 2022”.

Although North America has dominated the market since 2023, **Asia-Pacific is the region with the fastest and most scalable growth.** In North America, growth drivers in this market focus on various factors: its advanced ecosystems for artificial intelligence research, stricter privacy regulations, and early adoption by sectors such as finance, technology, and services.

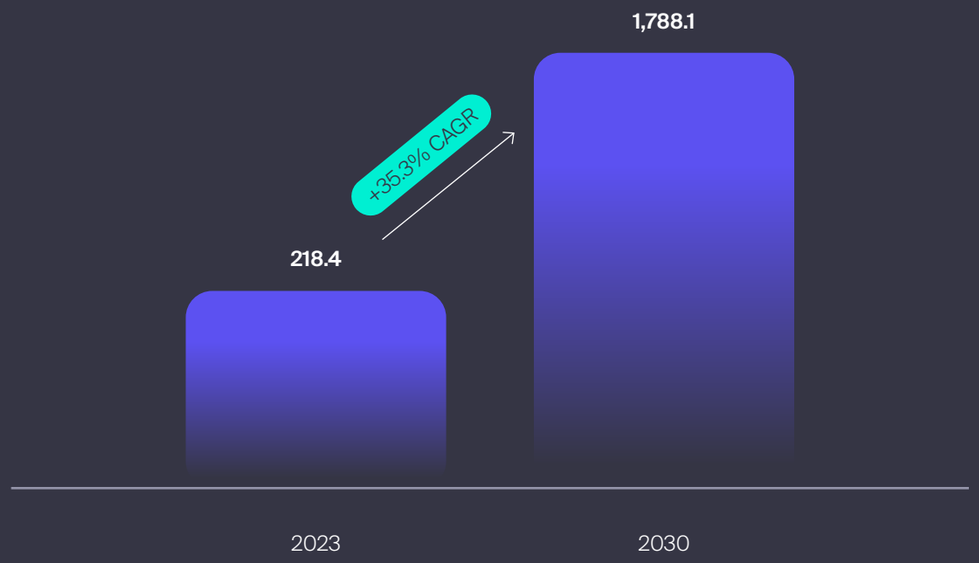
60%

of the data used for AI will be synthetic by the end of 2024, compared to 1% in 2021, highlighting its key role in simulation, modeling, and risk reduction.

Sources: Gartner I Grand View Research I Nature

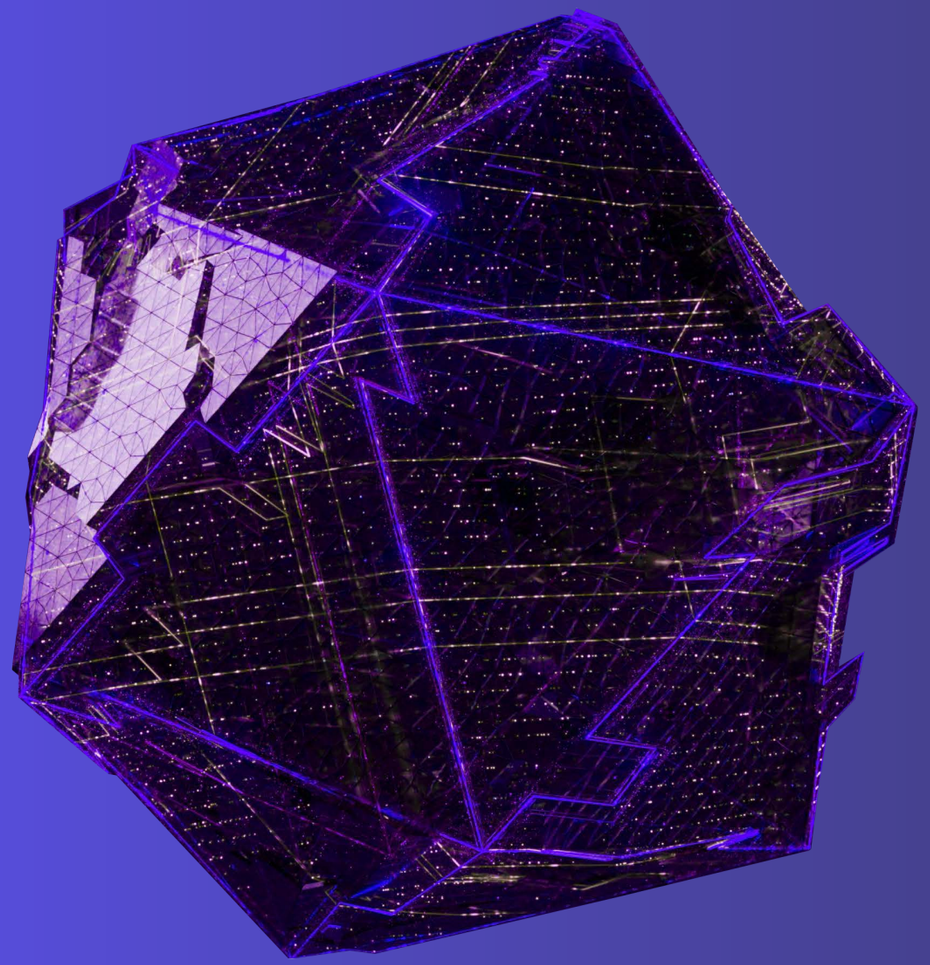
Global synthetic data generation market size

(in $ millions)



In **Europe**, compliance with the General Data Protection Regulation is a key driver. Additionally, key industrial sectors such as automotive and manufacturing are boosting the demand for synthetic data. Moreover, the existence of **regulatory sandboxes** in some European countries also facilitates experimentation with them.

In the **Asia-Pacific region, investment is being made in AI infrastructure,** and there is rapid adoption of digital technologies in countries like China, India, and Japan due to digital expansion and increased connected devices. Furthermore, governmental support for AI innovation is driving greater adoption of these technologies in the region.



Synthetic Data: reinventing digital reality

Synthetic data can be generated in various forms depending on the degree to which they are artificial and how they integrate with real data

Beyond real replication: a solution for model creation and validation

Synthetic data are not simply copies of real information, but **new creations based on the statistical distributions and relationships of the original data.** Today, they have the potential to cover a wide range of applications, including testing, new model development, algorithm training in Machine Learning, and predictive model validation.

The generation of synthetic data involves the use of algorithms and statistical models to produce data not collected from real-world sources. These data can take various forms.

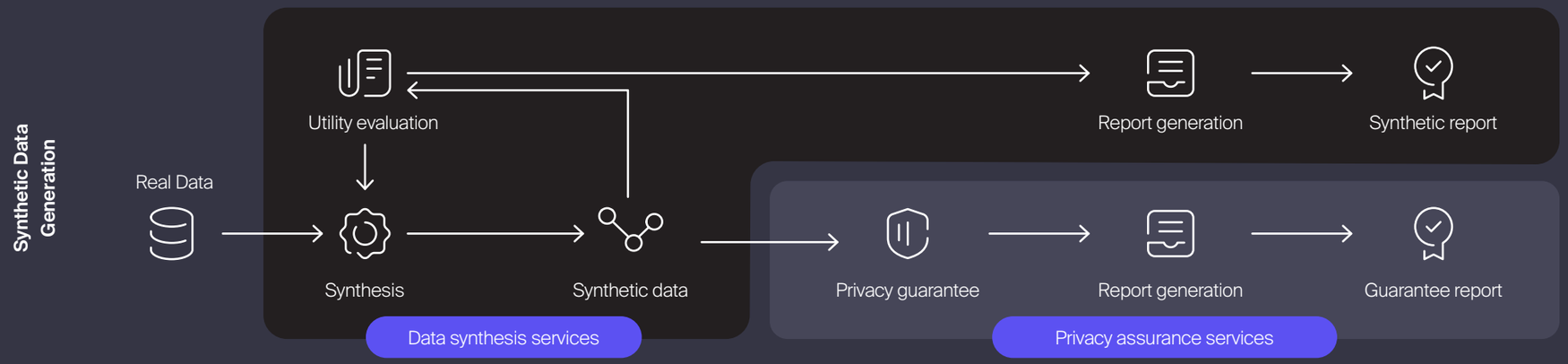
The techniques used for generating synthetic data are based on the analysis of underlying statistical distributions, machine learning models, and deep learning.

Their different forms are:

- **Partial synthetic data:** those that **replace only part of a real dataset with synthetic information.** They are useful for protecting sensitive information such as names or personal details. This technique preserves information privacy and helps protect personal data while maintaining the relevant characteristics of real data.

- **Hybrid synthetic data:** this type of **data combines real data with fully artificial data, meaning that a real dataset is randomly mixed with synthetic records.** This approach is very useful for simulating more complete scenarios without using direct data, reducing the risk of re-identification.

- **Fully synthetic data:** those that **do not contain any real data at all.** Entirely new data is generated that follows the same statistical relationships and properties of real data, but is not linked to any real person or event. This approach is particularly useful when there is not enough data available to train Machine Learning models or to simulate processes.



Source: LinkedIn

Synthetic Data offers greater security, eliminating re-identification risks that persist in traditional techniques

Anonymization: the solution of the past, synthetic data: the solution of the future

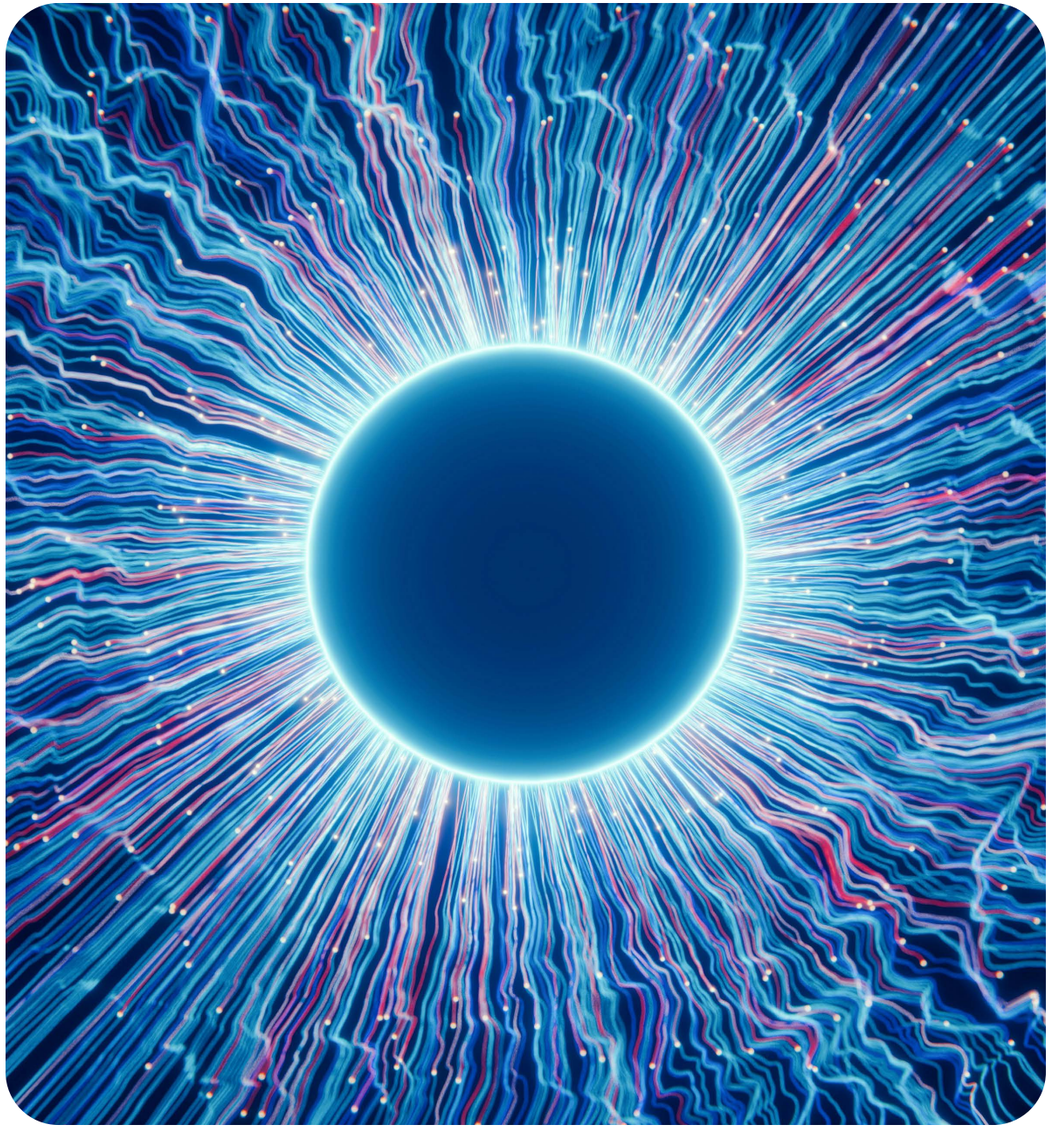
Traditionally, **anonymization, pseudonymization, and masking techniques have been used to protect personal data.** However, with increasing data complexity and advanced analytics, these techniques reveal clear limitations, leading to synthetic data as a more secure and scalable alternative.

80%

reduction in potential fines, since synthetic data minimizes the risk of data leakage compared to real data.

25%

faster creation of AI prototypes, offering competitive advantages in sectors such as finance and healthcare.



Sources: LinkedIn I Syntho I Geeks for Geeks I Better Data

Different privacy protection techniques

	Anonymization	Pseudonymization	Masking	Synthetic data
	Transforms personal data so it can no longer be linked to an individual, and cannot be reversed.	It is a process in which **direct personal identifiers are replaced with pseudonyms.** Although it is an improvement over anonymization, it does not completely eliminate the risk of re-identification, since the pseudonyms can be reversed	**This process alters the values of the original data so they are unrecognizable.** For example, names and addresses can be replaced with fictitious values, while maintaining the structure and format of the data.	**This technique artificially generates data using AI algorithms that mimic the characteristics and patterns of real data**, but without containing any personal or identifiable information. Unlike other techniques, it does not depend on original data and eliminates the risk of re-identification.
Disadvantage	• Since these techniques are static, **they may not be sufficient against new re-identification methods.** • In many cases, **it reduces the granularity of the information,** eliminating details that may be crucial for advanced analytics.	• **Its inherently reversible nature makes it vulnerable** and a significant risk in terms of privacy. • If the keys can be accessed or if the data is combined with other databases, **it is possible to re-identify** individuals, which must be specially controlled.	• If the keys are accessed, or if its implementation is weak**, the data can be restored**, resulting in vulnerabilities if the protection mechanisms are not strong. • Even if the structure is preserved, **the quality of masked data is often compromised**, which may limit its usefulness in AI models and predictive analysis.	• **Generating high-quality data can be a computationally expensive process,** especially when advanced models are used. • **Its creation requires the use of advanced AI models and experience** in data management.

Each organization must evaluate its case, although synthetic data stands out as a superior solution in privacy and utility

Anonymization methods

Method	Real data	Anonymized data
Masking	EFGH3456	EFGH**** | ****3456
Pseudonymization	David	a0001
Generalization	Age: 24	20-25 | 20-30
Data Swapping	David Lorenzo John Carlos	David Carlos John Lorenzo

Synthetic data

	Name	Email	ID number	Age
Real data	David	david@gmail.com	1234	36
Fully synthetic	James	james@gmail.com	8456	36
Partially synthetic	David	David@outlook.com	4312	36

Even techniques of anonymization, pseudonymization y masking have been useful in the past, **limitations in terms of data utility and re-identification risks** have led to the adoption of **synthetic data** as the most robust solution.

They offer greater privacy, maintain data utility, and are easily scalable, making them the preferred choice in modern applications—especially when dealing with large data volumes or rare scenarios that traditional methods cannot effectively address.

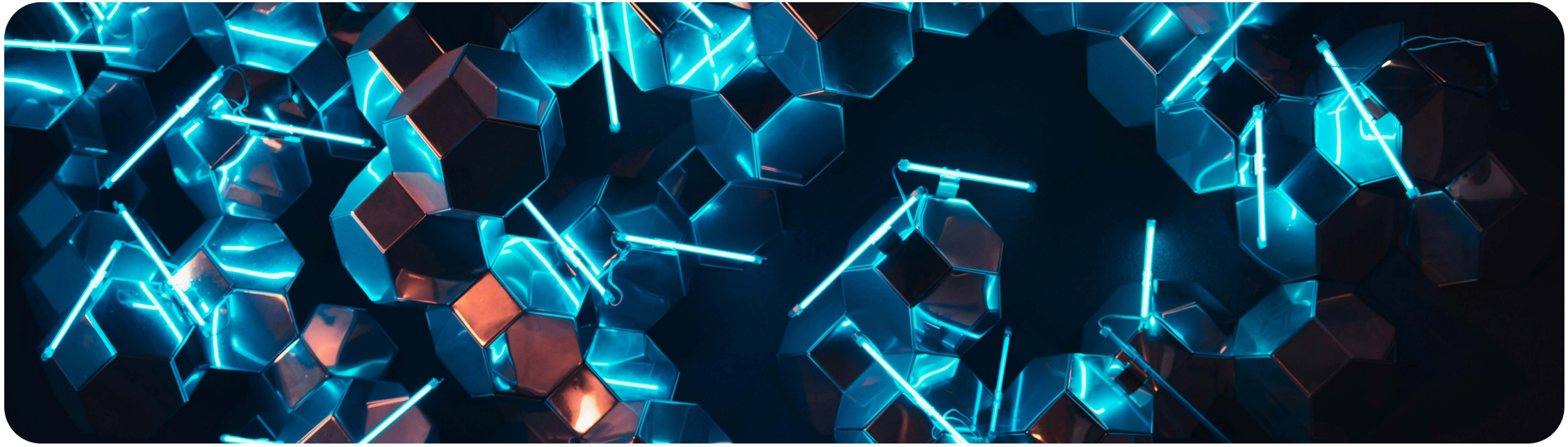
However, the selection among these methods must depend on **each organization's specific needs**, the type of data they handle, and the privacy regulations they must comply with.

Yet, synthetic data is increasingly being consolidated as the best alternative to protect privacy without sacrificing utility, becoming an indispensable tool in the modern world of Big Data and artificial intelligence.

Managing data implies constant trade-offs, but synthetic data enables progress without compromising privacy, utility, or cost

Trade-offs between privacy, utility and cost

Any decision involving data implies a trade-off between maximizing privacy, maximizing utility, and factoring in cost or effort. Synthetic data is emerging as a solution that balances all three axes in a novel way—though not a silver bullet.



Privacy vs. utility

In traditional methods, this trade-off was severe—removing or masking fields reduced exploitable information and added too much noise, ruining fine-grained analytics. With synthetic data, the goal is to achieve **high privacy with high utility.** However, to guarantee privacy, slight inaccuracies are sometimes introduced to avoid matches with real data. The trick is finding the right balance. For example, techniques like differential privacy allow tuning a parameter epsilon (a smaller epsilon means more privacy, but less utility—and vice versa).

Moreover, **synthetic data can even add utility** in some cases by generating larger volumes of data for algorithms (e.g., better generalization), or by eliminating real-world data errors (e.g., typos or measurement noise) that synthetic methods can avoid.

Therefore, the privacy/utility trade-off is significantly narrower with synthetic data compared to traditional methods.

Utility vs. cost

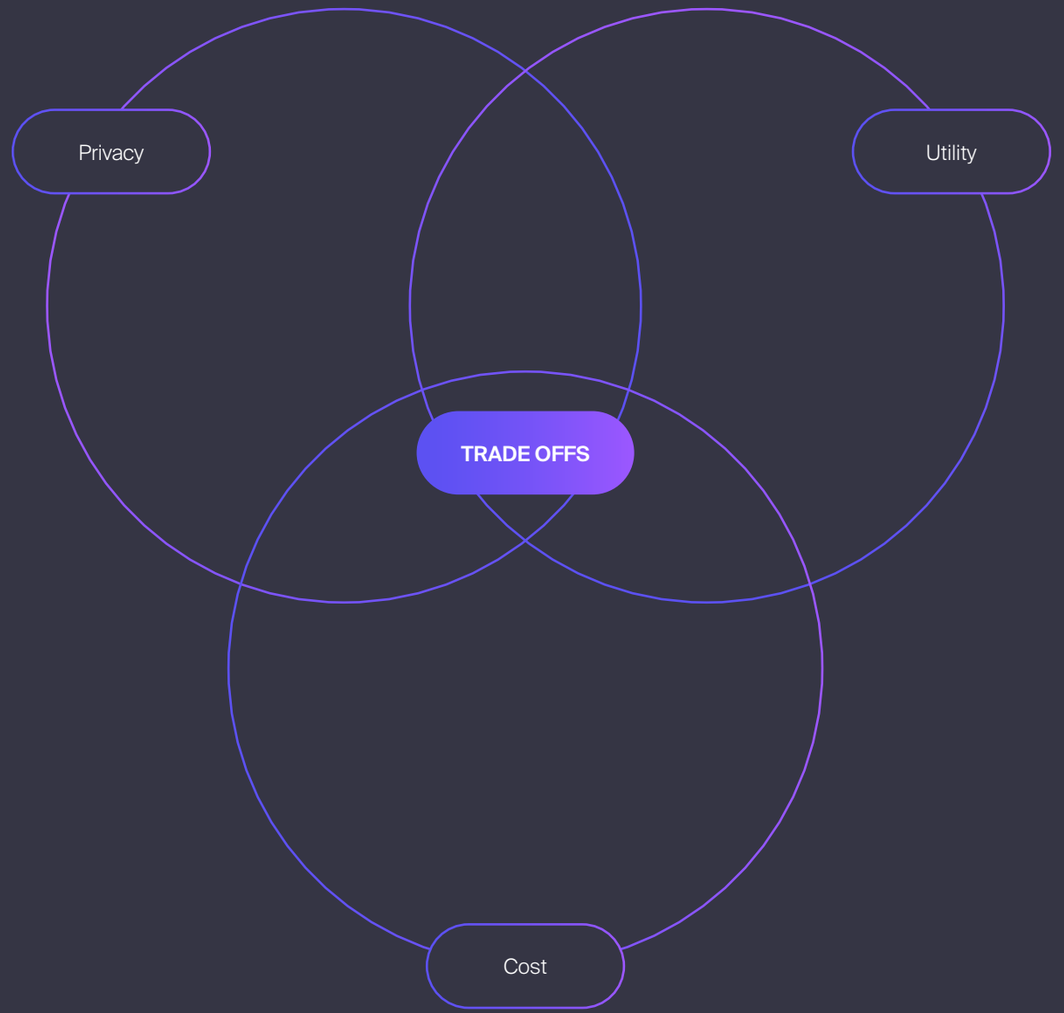
Traditionally, collecting large volumes of high-quality real data is expensive (in time, money, and human effort). Additionally, cleaning and labeling data delays AI project execution. **Synthetic data can reduce these costs**, and it's confirmed that generating synthetic data is often cheaper than collecting and annotating real-world data.

It's also **potentially scalable at low cost.** In general, synthetic data tends to improve the utility available within a given budget.

Privacy vs. cost

In classical mechanisms, adding privacy often means extra cost (consultants for anonymization, time to prepare safe data, etc.). With synthetic data, once the process is implemented, privacy is embedded in the generated data itself. This can simplify workflows and reduce operational compliance costs. From a regulatory perspective, companies report that **projects requiring complex legal agreements (time-consuming and costly) now move faster with synthetic data, reducing opportunity costs.**

Sources: LinkedIn I Aetion



Moreover, achieving the right balance between fidelity and generalization in synthetic data is key to avoiding re-identification or loss of value

The fidelity dilemma: false positives or useful generalization?

One of the biggest challenges is **finding the right balance between fidelity and generalization.** This dilemma involves deciding to what extent generated data should faithfully replicate the patterns and relationships present in real data.

Too much fidelity can introduce false positives or re-identification risks, whereas too little fidelity may lead to excessive generalization, causing the model to lose critical and useful information.

High fidelity:
risk of false positives

When synthetic data tries to replicate real data almost identically, **there is a risk of reproducing irrelevant patterns or statistical noise** present in the original data—commonly known as overfitting.

Models trained with such high-fidelity synthetic data may learn **spurious or non-causal correlations.**

If synthetic data replicates reality too closely, **it could compromise individuals' privacy** by memorizing confidential information (like outliers).

Moreover, **excessive fidelity limits the model's ability to generalize**, making it less robust to unseen scenarios.

Low fidelity:
useful generalization, but at what cost

On the other hand, too little fidelity means that synthetic data is **overly generalized.**

While this avoids overfitting, it comes at the cost of omitting important details that could be crucial for building robust models. A generative model that over-smooths distributions may **fail to capture key behaviors in certain subgroups or patterns of interest.**

In statistical terms, this refers to the classic bias–variance trade-off: **too much smoothing introduces bias, while too little smoothing leads to high variance.** The key is finding a balance where synthetic data does not omit useful signals.

Key technologies like GANs, Transformers, and simulators drive synthetic data, balancing realism, control, and practical applicability

Generative models can be divided into explicit and implicit types

Explicit models are those that directly and transparently model the underlying distribution of the data. This means you can observe and understand how the results are generated, as the model uses well-defined mathematical functions to create the data. They are easier to interpret and apply when high precision and traceability are required in the generated data. These models are more focused on replicating data with high fidelity to original statistical distributions.

Implicit models, on the other hand, do not directly model the data distribution. Instead, they learn to approximate this distribution through training with real data. These models are more powerful in terms of generating realistic data, but harder to interpret and control, as there is no explicit mapping between the data distribution and the generated outcomes.

Explicit models

Variational Autoencoders (VAE):

Learn latent representations of data through a process of compression and reconstruction. An encoder compresses data into a lower-dimensional space, and a decoder reconstructs synthetic data from that representation. Through a probabilistic process, **new instances of data are generated that are diverse and realistic**, without replicating specific points from the original data.

Probabilistic Modeling:

Involves the use of statistical distributions and simulations (like Monte Carlo) to generate synthetic data. This approach is based on observing patterns in real data and creating synthetic samples that respect those distributions.

Simulators:

Use physical, mathematical, or logical models to generate synthetic data via **simulated systems or process models.** These technologies allow experimentation with hypothetical or controlled scenarios, making them especially useful in testing environments.

Implicit models

Generative Adversarial Networks (GANs):

Deep learning models composed of two neural networks: a generator that creates synthetic data, and a discriminator that evaluates their quality by comparing them to real data. Both networks compete with each other, improving their performance. Over time, the generator improves until the discriminator can no longer distinguish real from synthetic data.

Transformer Models (GPT, BERT):

Powerful deep learning models, especially effective in sequence processing like natural language. Models like GPT or BERT learn structure and linguistic relationships from large volumes of data. Encoders transform input data into numerical representations (called embeddings), while decoders generate output sequences based on those embeddings. This enables the model to focus on relevant tokens while generating coherent synthetic outputs.

Diffusion Models:

Generative models that create data gradually, especially images, through progressive noise diffusion. In this process, random noise is added and then reversed until realistic and usable data emerges.

Moreover, choosing between generative models or simulations depends on the balance between realism, privacy, explainability, and usage context

Common applications of generative models

Create Realistic Data

Image and video generation

Text generation

AI model training

Common applications of simulation

Explore Different Scenarios

Industrial process optimization

Financial systems simulation

Healthcare system simulation

Generative models vs. simulation

When working with synthetic data generation, it's important to understand the distinction between two different approaches: **generative models—including GANs, VAEs, Transformers—and simulation-based models.**

Generative models are designed to create synthetic data that mimics the **statistical characteristics of real data.** Using machine learning algorithms, generative models learn the distributions of original data and then generate new samples that follow those same distributions. This generation technique is well-suited for realistic replication of both the structure and statistical properties of the data, making it ideal for reproducing unstructured data such as images, text, and audio.

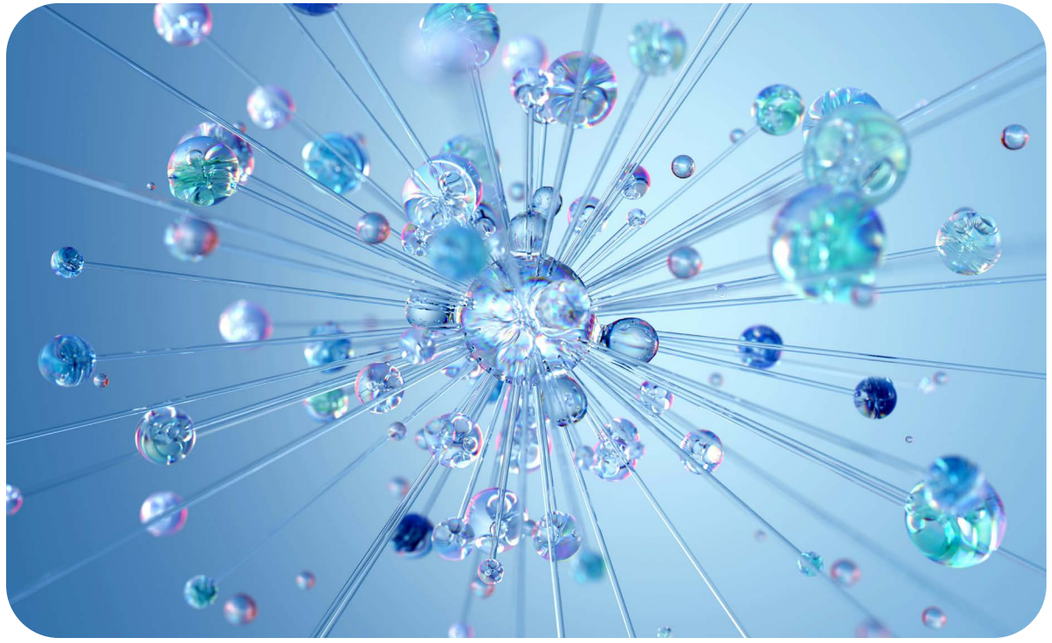
Moreover, to function properly, **these models must be trained on real data**, allowing them to replicate behaviors present in real datasets. Such models are used across various sectors—from content creation (images, music, or text) to AI training and sensitive data protection, since they can generate high-quality data without compromising individual privacy. In contrast, simulation models do not focus on creating new data from existing datasets.

Instead, they aim to **replicate the behavior of real-world systems under controlled conditions.** These models use mathematical, physical, or logical equations to simulate phenomena and real-world events. They are especially effective for analyzing dynamic and complex systems such as economic, healthcare, or industrial processes, where interactions between variables are not always directly observable.

They are also ideal when the goal is to study a system's behavior under different conditions, without needing real data. Simulations allow for modeling of workflows, agent interactions, or physical processes. Through simulation, it is possible to **explore a variety of "what-if"** scenarios, adjusting variables and observing results without risk in real-life situations.

Unlike generative models, simulation models allow full control over the variables involved, making them well-suited for hypothesis testing and forecasting in controlled environments.

The use of synthetic data revolutionizes internal sharing and external collaboration, ensuring privacy, agility, and total trust



When and how to use synthetic data: from internal sharing to external collaboration

Their flexibility and ability to represent diverse scenarios make them a key tool across different stages of the **artificial intelligence and machine learning** model lifecycle. To maximize their value, it is essential to understand in which contexts and processes they are most suitable, **approaching their implementation with a strategic analysis that considers objectives, needs, and environment.**

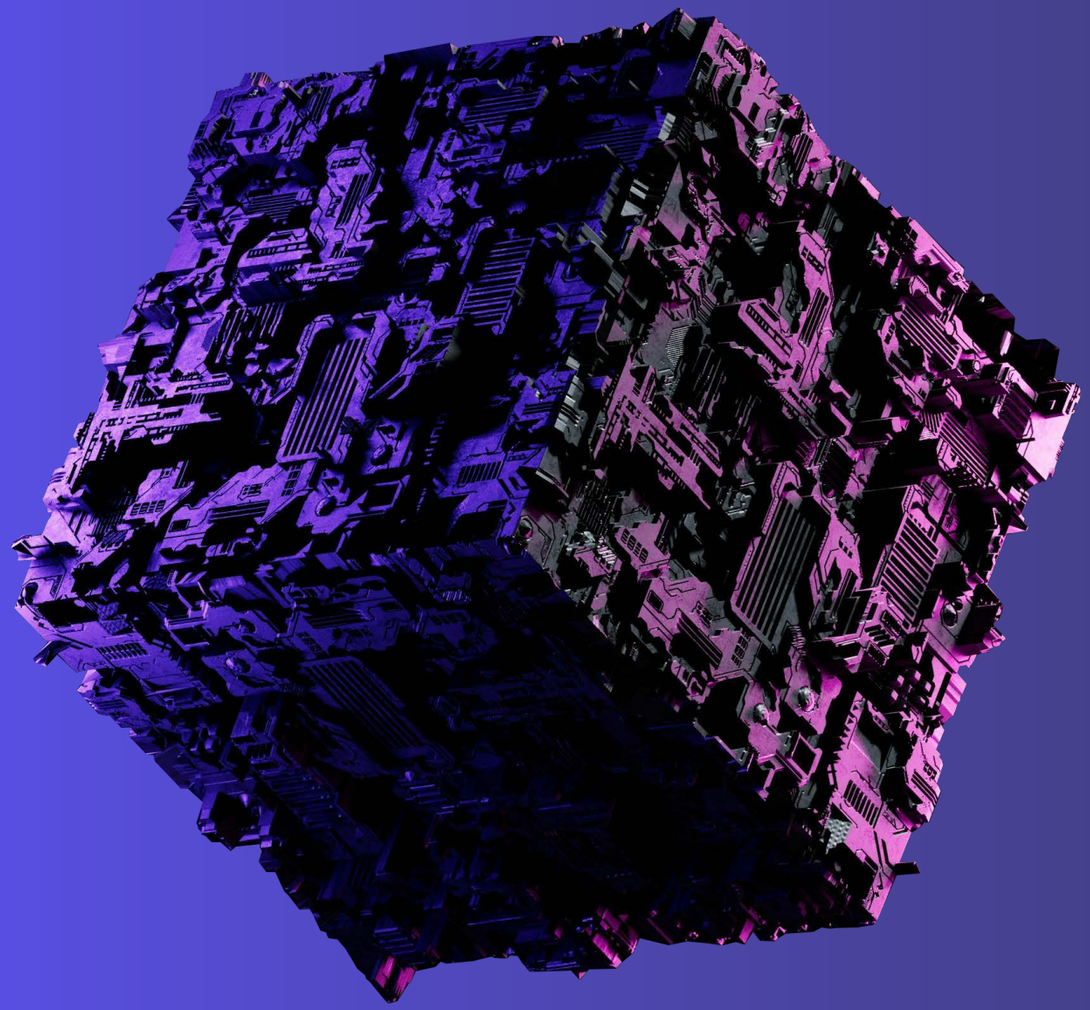
Before deciding on their use, it is necessary to conduct a **study of organizational needs and the existing limitations in handling real data**—such as access restrictions, acquisition costs, or data quality. This analysis will help determine whether synthetic data offers a viable alternative.

Once these aspects are evaluated, it is important **to define how and when to apply them across the various phases of the AI model lifecycle** and in business operations. Some key strategic factors include:

- **Product or service development stage:** allows testing and validation without compromising real data.
- **AI and ML model training:** useful when real data is insufficient, incomplete, or biased.
- **Simulations and hyper-realistic scenarios:** enable the recreation of risk situations or extreme conditions that are difficult to obtain in reality.
- **Internal and external collaboration:** facilitate the exchange of representative datasets without compromising privacy or security.

By identifying organizational needs and barriers, it becomes possible to effectively determine when and how to implement them. However, their successful adoption **requires careful planning that aligns technological goals with the company's overall strategy.**

Synthetic data strengthens internal and external collaboration in research, innovation, and training



From Lab to Market: opportunities and challenges

With synthetic data, companies turn restrictions into advantages, drive rapid innovation, and collaborate without risk

Unlocking new opportunities with synthetic data

Synthetic data is emerging as a **new technique to solve a wide range of challenges** in data management and analysis, particularly in sectors facing limitations due to privacy regulations, data scarcity, or high costs. However, its implementation brings both opportunities and challenges that must be carefully understood and managed.

Privacy protection without limiting innovation

Proactively integrates privacy by eliminating personally identifiable information, reducing risks and easing audits, **turning it into a strategic advantage that unlocks access to data and builds trust among regulators and clients.**

Operational efficiency and speed

Eliminates inefficiencies in real data management by providing ready-to-use datasets without the need for legal approvals, forms, or cleanups. This accelerates **AI development, testing, and product design, resulting in faster experimentation cycles and reduced development costs.**

Secure collaboration and ecosystem growth

Turns the problem of data scarcity into an opportunity by allowing legal and secure data sharing. **This enables innovation scaling and high-definition (HD) modeling through external collaborations, joint ventures, or data-sharing partnerships.**

Differentiation and future innovation

Allows modeling of rare or unusual events, simulating complex real-world systems, all without using real data. It also supports the creation of **new experimental models for emerging technologies**, boosting sustainability, traceability, and regulatory readiness.

Ethical advancements in AI

By generating data under similar conditions to real-world data, synthetic data makes it possible to train and test AI models more fairly and inclusively, **helping to detect and mitigate bias.** It reinforces ethical development by reducing exposure to sensitive or discriminatory data.

New sources of revenue

This generation technique opens up **new data monetization opportunities**, such as paid data sharing, R&D partnerships, synthetic data marketplaces, and other data commercialization strategies.

Overcoming the technical challenges of synthetic data is key to ensuring its value, utility, and widespread adoption

These challenges are mainly related to the quality, realism, and representativeness of the generated data — critical factors to ensure that AI models perform correctly when trained on them.

1. **Quality and realism**

   One of the most critical challenges of synthetic data is to ensure high quality and realism. **AI models largely depend on data quality to learn correct patterns and make accurate predictions.** If the synthetic data does not reflect real-world complexity, models may struggle to generalize or even produce incorrect results. While synthetic data is statistically similar to real data, the models must be able to capture subtle variations and unusual behaviors to avoid biased or inaccurate predictions. Although generation methods have improved significantly, models still risk producing biased or non-representative data if not properly trained.

2. **Representativeness**

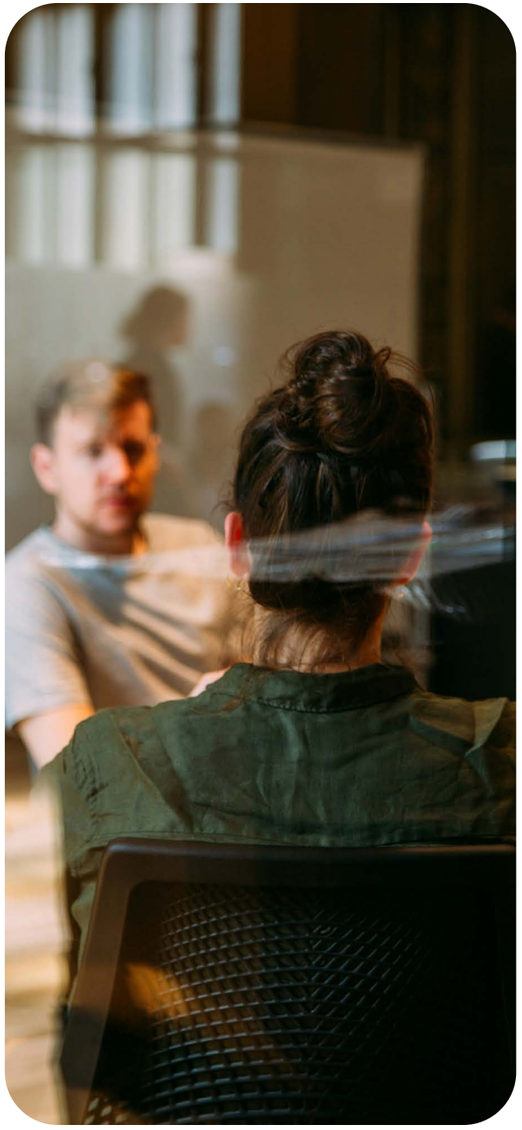
   Another challenge is representativeness. To be useful, **synthetic data must simulate not only common scenarios, but also rare or extreme cases**, which are critical in decision-making. These cases are crucial, for example, in fraud detection (finance) or rare diagnosis modeling (healthcare). As AI models train on synthetic data, the algorithms must accurately capture relationships and correlations between variables — something difficult when the real data is complex or non-linear.

3. **Balance**

   Combining real and synthetic data is key to building robust and accurate models. Synthetic data can help fill data gaps — especially when real data is scarce, restricted, or private. However, overreliance on synthetic data can compromise model accuracy and limit applicability in real-world scenarios.

4. **Validation**

   The lack of clear and standardized metrics to evaluate synthetic data quality has been a major barrier to widespread adoption. Currently, there is no unified framework to measure the quality of synthetic data, making it difficult to compare and validate across use cases and models. This creates **uncertainty that can hinder adoption and prevent confidence in their representativeness.**



Synthetic data can amplify biases and privacy risks if their generative processes are not carefully controlled

Risks and biases: can you audit what never happened?

A series of ethical and technical risks must be managed carefully. These include **biases inherent in original data, auditability and traceability challenges, and legal and regulatory concerns** related to the use of synthetic data. Furthermore, a fundamental dilemma arises: how can we audit a dataset that, by definition, does not come from real-world events?

1. If the original data contains biases, such as in the case of racially biased credit decisions, **synthetic data can reproduce or even worsen these biases**, affecting the fairness of the AI models trained with them. This problem becomes even more serious if the generative model does not adequately handle the data distributions, which could lead to the generation of biased or unbalanced synthetic data.

2. In addition, the auditing of synthetic data becomes a challenge because, as it does not come from real events, **it cannot be verified in the same way as traditional data.** Instead of verifying against reality, auditors should focus on the statistical properties of the generative model and the procedures used, which requires a new validation framework. Instead of seeking direct matches with real events, it must be ensured that synthetic data maintains statistical coherence and does not introduce unintended biases.

3. Another risk is re-identification, although synthetic data is designed to protect privacy, **if the generative model is not well regulated, it could memorize details of real data and replicate them in the synthetic data**, which would compromise the same privacy. This problem highlights the importance of applying differential privacy protection techniques, which add controlled noise to protect individuals' identities.

4. The feedback loop is another problem, occurring when a model is trained with synthetic data generated by another synthetic model. In this cycle, **errors or biases from the original model can be amplified in subsequent models, which can drift the generated data away from real patterns.** To mitigate this risk, real data should be periodically reintroduced or synthetic and real data combined in the process.

5. Finally, **ethical dilemmas about responsibility** arise when models trained with synthetic data fail or cause harm. In such cases, the question arises as to who is responsible—the model's creator or the user who implemented the model. This is especially relevant in areas such as automated decision-making, where models can directly impact people. The solution requires more technical and transparent auditing, focused on validating the processes and the statistical properties of the synthetic data, rather than verifying it against real events.

The impact and value of synthetic data lies in redefining privacy, boosting innovation, and democratizing access to data

Impact of synthetic data: driving a more efficient and democratic digital economy

The adoption of synthetic data by multiple actors in the sector represents a **complete transformation in the way different stakeholders approach privacy,** sustainability, and innovation.

1. **Agility, efficiency and new revenue**

   They allow companies to move faster without compromising security or compliance. By removing restrictions on personal data, waiting times are reduced, product development is accelerated, and operational costs are decreased. Additionally, they open access to new markets and previously inaccessible segments due to regulatory barriers, resulting in greater speed, less friction, and a higher capacity for data monetization.

2. **Innovation without compromising privacy**

   Their ability to generate unlimited data **allows simulating risks, validating ideas, and training algorithms in situations not covered by real data.** In an environment where innovation, ethics, and regulation must go hand in hand, synthetic data represents a responsible growth strategy that balances technological agility and privacy protection.

3. **Scalability and resilience**

   Their adaptation not only complies with the law but goes beyond it. This technology allows operating and scaling products in various jurisdictions without the legal risks of real data, becoming a lever for expansion and reduction of reputational risks.

4. **Sustainability and ethics**

   They allow reducing the footprint of real data, by generating artificial data instead of relying on personal information, relieving pressure on digital infrastructures, reducing storage and processing needs. Furthermore, companies adopt a more ethical and responsible model, integrating privacy by design and strengthening user trust.

5. **Democratization of access**

   Their **fast access and retrieval allow small and medium-sized companies to access valuable datasets without the associated risks**, leveling the playing field with technological giants. This opens new opportunities for collaboration and co-creation of value.

6. **Competitive differentiation and positioning**

   Their use demonstrates a commitment to privacy, algorithmic fairness, and responsible innovation. Integrating them streamlines processes and strengthens positioning as a digital leader. **User trust and transparency are more critical than ever.**

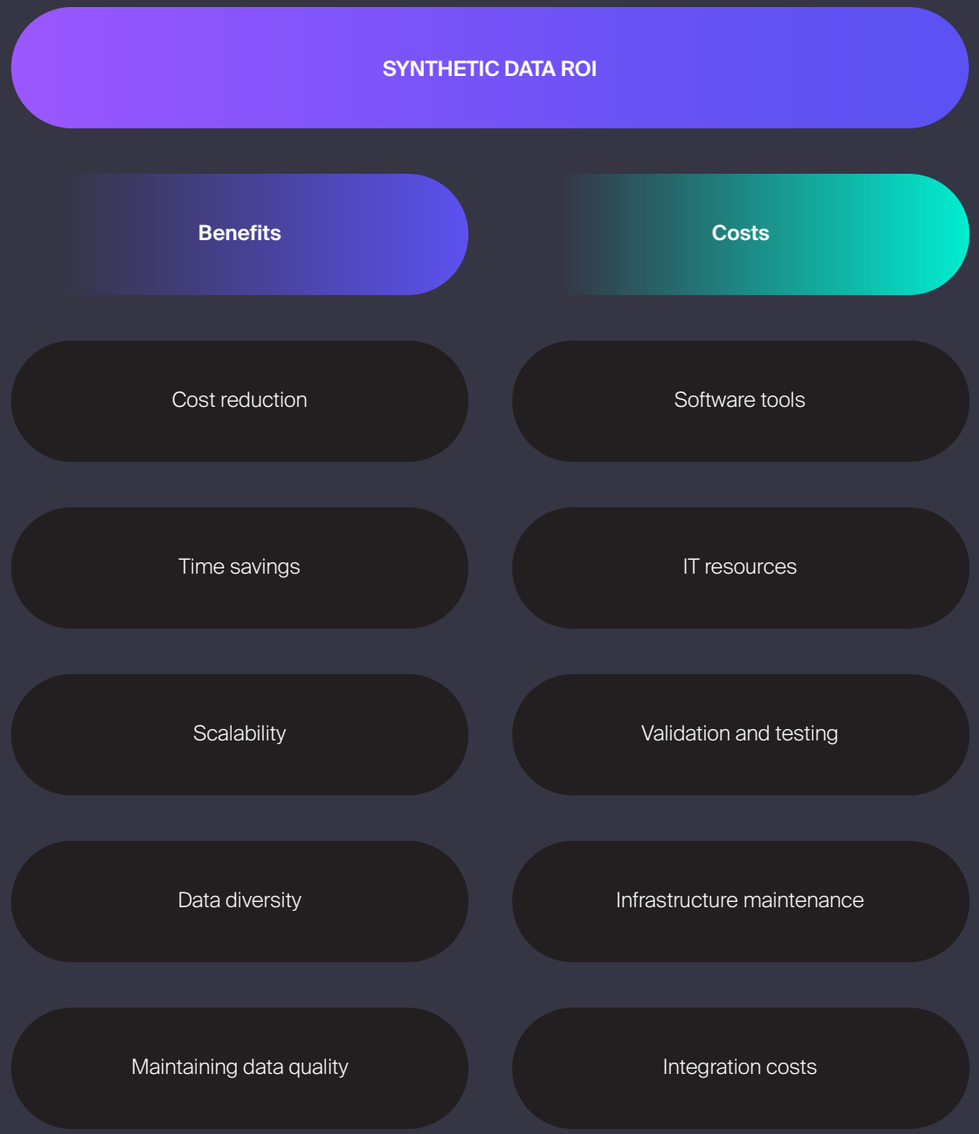
Calculating the ROI of synthetic data helps demonstrate immediate benefits in agility, savings, compliance, and competitive advantage

How to quickly calculate and visualize the immediate return on investment when implementing synthetic data

Alongside the impact this trend has in the sector, it also stands out for the ROI it can generate through its use. The key to the ROI of synthetic data **lies in cost reduction and the mitigation of previously mentioned risks.**

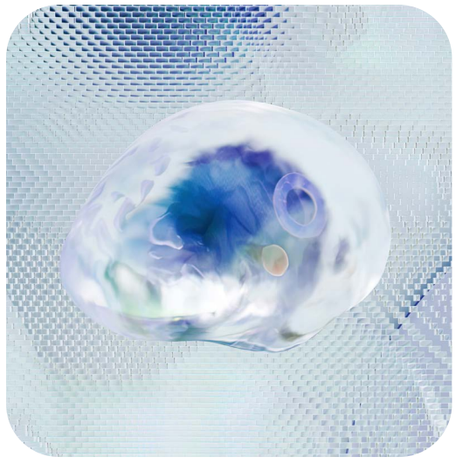
Calculating the ROI of synthetic data is based on three steps:

1. **Identification of tangible benefits** such as reduced compliance costs, time savings in data processing, and improved operational efficiency.
2. **Quantification of benefits by assigning a monetary value,** such as development time reduction or savings.
3. **Evaluation of costs** of the initial investment in software tools, IT resources, and validation.



In the end, not adopting synthetic data exposes companies to risks, slows innovation, and leaves them out of future competition

Why is the time now? Factors demanding the adoption of synthetic data



The growing adoption of synthetic data is being driven by a convergence of **technological, regulatory, and ethical factors** that reflect both the urgent demand for high-quality data and the need to comply with strict privacy regulations. Delaying the adoption of synthetic data can have significant consequences across various fronts, from regulatory costs to competitive disadvantages and the loss of agility in product development. Companies that do not rapidly integrate synthetic data into their operations face risks such as:

Higher regulatory risk and legal exposure

As privacy regulations tighten, relying exclusively on real data exposes companies to sanctions, security breaches, and demands. Without synthetic data, it's harder to comply with principles like data minimization or explicit consent. The consequences include **million-dollar fines, loss of customer trust, and operational shutdowns.**

Limited access to critical data for innovation

AI models and advanced analytics require large volumes of diverse data, while legal restrictions reduce the availability of real data. Without synthetic data, companies face data shortages and biases in their models. This affects **incomplete models, slowed innovation, and lack of capacity for early scenario simulation or risk anticipation.**

Slower times and higher operational costs

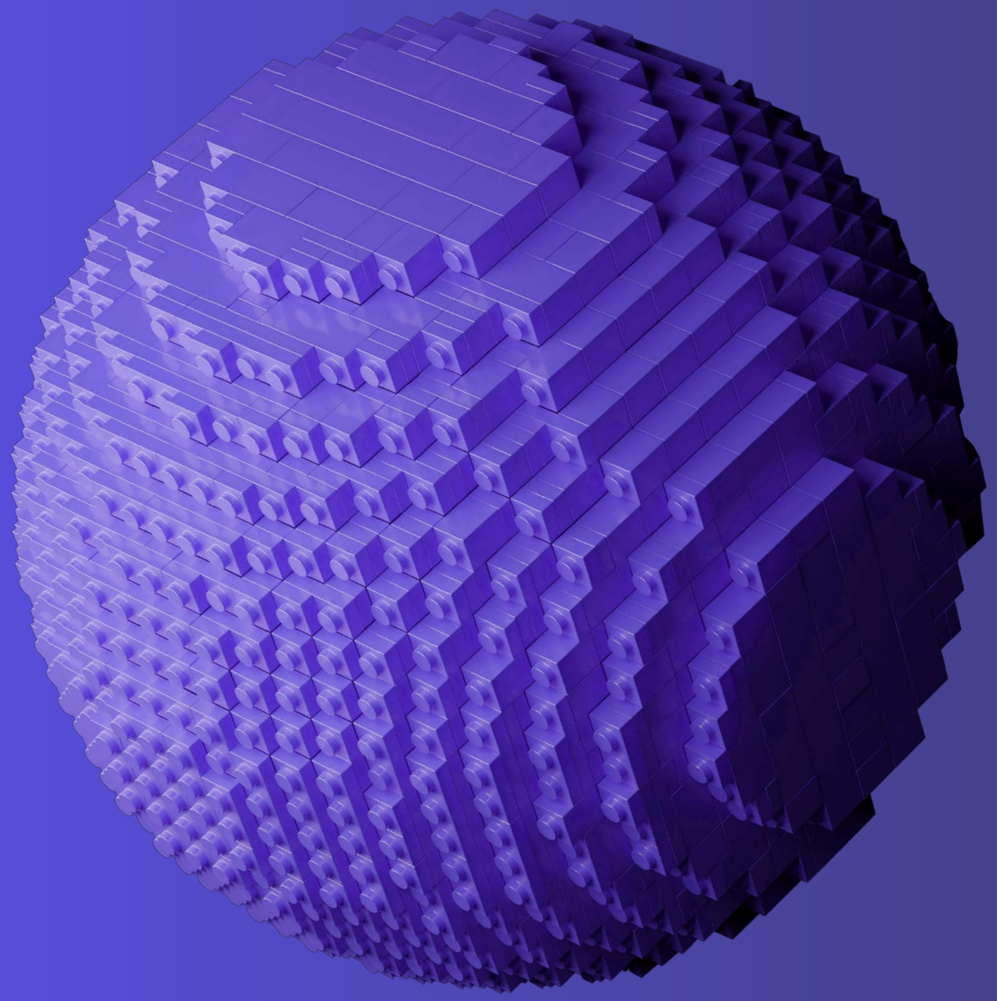
Using real data implies long approval, anonymization, and legal review cycles. Synthetic data allows faster prototyping, testing, and scaling without legal friction. Not using them results in **longer time-to-market, high operational costs, and lagging behind more agile competitors.**

Competitive disadvantage in AI and data strategy

Leaders in health, banking, insurance, and tech are already using synthetic data to train better models and reduce risk. Not adopting them implies falling behind technologically versus companies that prioritize privacy and innovation. The impact is reflected in **loss of positioning, lower personalization capacity, and less competitiveness in regulated markets.**

Difficulties collaborating and scaling data

Sharing real data between teams or external partners is increasingly difficult and risky. Synthetic data enables secure collaboration, interoperability, and integration in complex environments. Not using it results in **data silos, project roadblocks, and slowed digital transformation.**

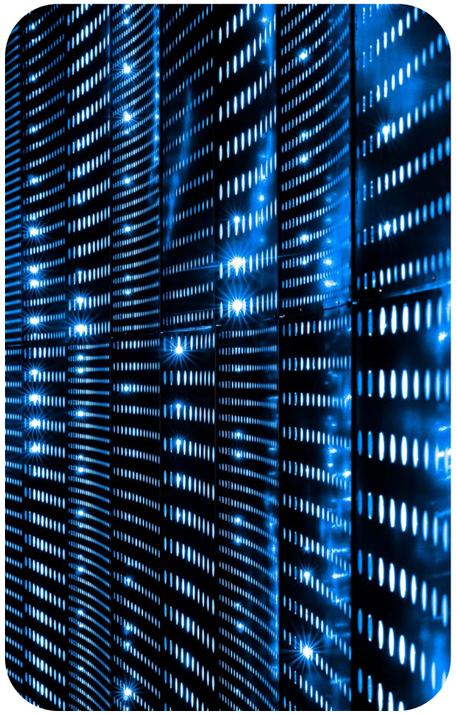


Regulation, Governance and Standards: between barrier and catalyst

Without ethics or strict regulation, the use of synthetic data can violate rights and result in multimillion-dollar fines

The need to balance the protection of rights, algorithmic governance, and the promotion of the digital economy requires robust regulatory frameworks. The GDPR, the AI Act, and other emerging regulations are fundamental to guiding the responsible use of synthetic data. As adoption grows, **clear ethical principles must be established, as well as certification and auditing mechanisms to ensure transparency and explainability in its use.**

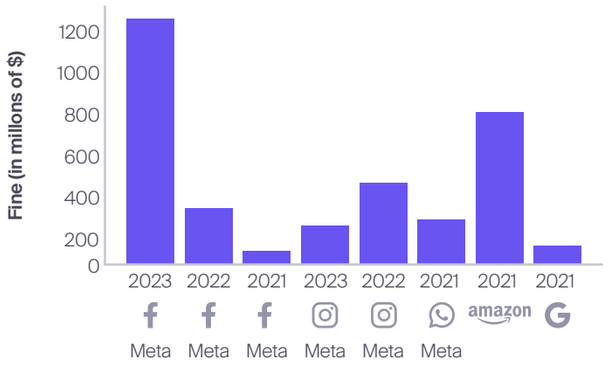
Fines to tech giants underline the importance of complying with these regulations. Meta, for example, was fined 1.2 billion euros for violating the GDPR in the transfer of international data, while Amazon received a fine of 746 million euros for not obtaining proper consent for data tracking. All of this highlights the need for clear regulation to protect privacy and ensure that **technology is used ethically and responsibly.**



Source: Statista

The largest fines for violating one or more articles of the general data protection regulation

(In $ millions)



The role of GDPR, AI act and other emerging regulations

The EU General Data Protection Regulation (GDPR) has probably been the biggest indirect driver of synthetic data technology, severely restricting the processing of personally identifiable information (PII), requiring legal bases, consent, minimization, and imposing severe penalties if someone is reidentified. This regulatory severity has led many organizations to seek solutions such as synthetic data to continue using data without violating the law. The GDPR does not regulate a database that does not include personal data, since, if it is not attributable to an identified or identifiable person, it is outside the scope of the regulation.

Furthermore, GDPR Article 89 **promotes the use of techniques to protect data and enable research**, and explicitly mentions "pseudonymization" and other techniques. While it does not name synthetic data, this privacy by design spirit aligns with the philosophy of generating non-real data to protect privacy.

For its part, the **EU Artificial Intelligence Act (AI Act)**, which came into force on August 1, 2024, is beginning to recognize synthetic data. It mentions that the use of synthetic data can be a measure to ensure the quality of training data and the protection of rights.

Regulations such as GDPR, AI Act, CCPA/CPRA, HIPAA and APACs ensure protection and ethics in the handling of sensitive data

In the U.S., state laws such as CCPA/CPRA **also focus on personal data.** If a company uses synthetic instead of real data, it may reduce its exposure; it is worth noting that there is no explicit mention in these laws, but the principle is similar to GDPR. A special case is the health sector with HIPAA in the U.S.

HIPAA defines de-identification methodologies (list of identifier suppression or expert certification). Synthetic data is not explicitly mentioned but could fit within an expert certification that certifies the data does not identify anyone. In fact, the FDA and health organizations are exploring its use for simulated clinical trials, etc., which suggests future adaptations in health regulations.

At the **international level, organizations such as the OECD and the G7 have highlighted PETs.** The OECD already published information on emerging privacy technologies and included synthetic data, discussing its potential and the need for legal frameworks. Finally, the G7 in its AI forum mentioned the importance of promoting PETs.

In a sense, these regulations acted as an initial barrier, while also serving as a catalyst for innovation in PETs. Moreover, some of them could require certain AI providers to apply PETs in their processes, which would make synthetic data **go from optional to almost mandatory in regulated environments.**

GDPR

Transparency
Informed consent
Right to erasure
Data minimization
Anonymization
Accountability and governance

AI ACT

Risk-based regulation
Ethics and fundamental rights
Transparency
Continuous evaluation

CCPA (U.S.)

Right to know
Right to deletion
Opt-out of data sale
Non-discrimination

HIPAA (U.S.)

Health data protection
Access and correction
Confidentiality
Breach notification

APACs (ASIA & PACIFIC)

Consent
Access and correction
Transparency
Data management
Sensitive data protection

But there is ambiguity and legal gaps, and clear rules are needed to align generators, users, and regulators

Navigating the legal gap for responsible innovation

Despite efforts such as ISO/IEC JTC1 SC42 technical reports, which provide an overview of synthetic data and its applications, **there is currently no universally accepted standard for certification.** This lack of standardization generates uncertainty for both organizations using synthetic data and regulators seeking to protect users' rights.

In this context, some companies, such as Gretel Labs, are implementing tools like "Privacy Reports" and "Utility Reports" **to assess the quality and privacy of generated data**, which is a first step toward normalization.

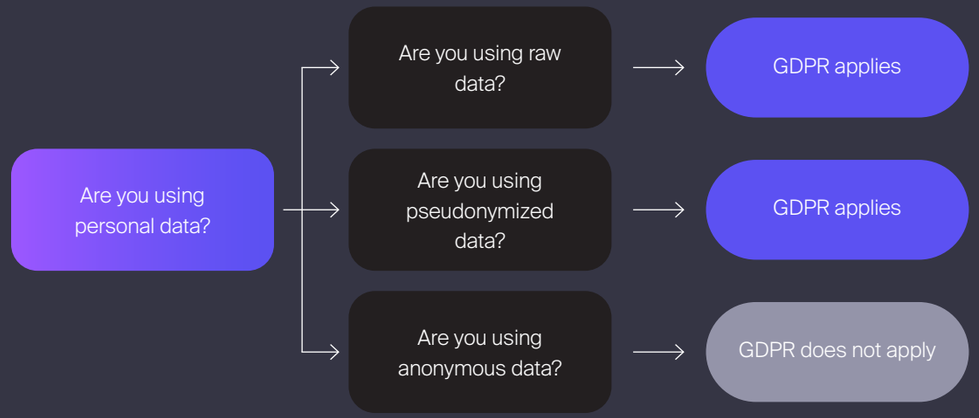
However, there is a need for **clear rules to ensure that generators, users, and regulators are aligned.** Among the most prominent problems, we find:

- **Definition and classification.** The fundamental question is whether synthetic data should always be considered "anonymous." Although many legal frameworks, such as the GDPR and the AI Act, accept that synthetic data should not be subject to privacy regulations if it cannot be linked to individuals, there is still no universally accepted technical or legal test to determine when they meet anonymization requirements.
- **Overlaps and gaps between laws.** Some regulations refer directly to synthetic data, validation conditions, transparency, and risks associated with reverse engineering, but many others remain ambiguous. Data governance laws designed for collected data do not necessarily apply to generated data.
- **Accountability and audit requirements.** The AI Act proposes labeling synthetic content and proving it does not come from real data, but there is still no standardized audit process to verify its privacy or representativeness. Clear standards for the necessary documentation have also not been defined, nor have regulations been established to determine responsibilities in the event of misuse or misrepresentation.
- **Legal risks of re-identification and quality.** Attacks that can de-anonymize synthetic datasets are increasing. High-fidelity synthetic data can inadvertently reproduce personal patterns or information that can be traced back to individuals, generating a legal gray area.
- **Implications for international data exchange and competition.** Synthetic data can reduce access barriers to data in competitive markets, which raises new challenges in terms of antitrust laws. In addition, regulations on cross-border data exchange are not clear on how to handle synthetic datasets generated from individuals' data from different jurisdictions.

Reforms to the laws and clarification of technical standards are widely recognized as necessary **to close these gaps and ensure the safe and ethical use of new technologies.**

The future regulatory framework will require greater responsibility, transparency, and clear controls in the use of synthetic data in enterprises

Application of the General Data Protection Regulation (GDPR)



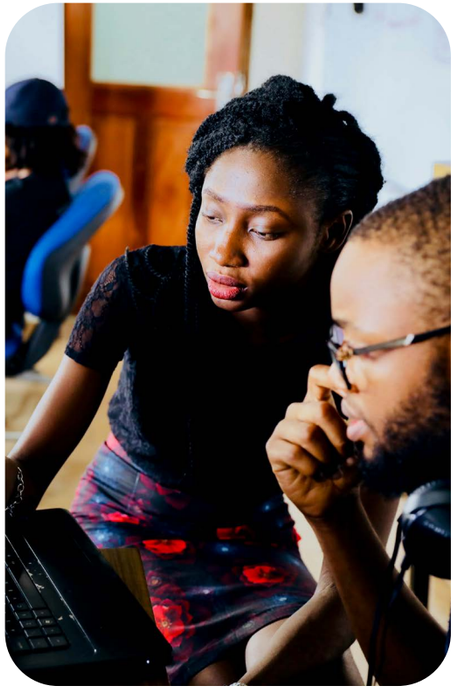
Future trends and recommendations for companies

By 2026, the AI Act will have established a stricter framework, classifying AI uses according to risk level and imposing specific obligations for systems trained with synthetic data. In addition, it **will drive greater transparency requirements regarding the generation and use of such data, demanding that companies demonstrate how privacy and fairness are protected in AI models.**

Moreover, if the use of these synthetic datasets is based on real data, it could **require companies to carry out Data Protection Impact Assessments (DPIAs)**, audit re-identification risks, and document mitigation controls. In other words, traceability and explainability of generation processes will be key to ensuring regulatory compliance.

In addition, the **EU Data Act,** which sets rules for access and use of data, will apply starting at the end of 2025. This signals a more open, transparent, and competitive environment, fostering innovation and protecting users' rights.

To prepare for an increasingly regulated environment, **companies should ensure they comply with restrictions by implementing privacy impact assessments** and establishing a solid governance framework, with risk management controls, clear labeling of datasets, and sound quality validation. Testing for re-identification and quality will remain the best allies for generating organizations.



With this checklist, organizations can prepare their use of synthetic data, avoiding sanctions and showing ethical commitment

Strategic-regulatory checklist to prepare the use of synthetic data

Certain actions should be taken now to minimize risks or sanctions, incidents, or reputational damage in the future, in addition to demonstrating **ethical and technological leadership in the digital economy.**

Risk assessment and analysis

Identify data sources: classify data by origin and justify their use.

Privacy and risk analysis: perform an initial and periodic DPIA (data protection impact assessment).

Consent and legal bases: verify that collection complies with regulations and meets the legal basis.

Model design and configuration

Selection of appropriate techniques: choose generative models based on the use case and data type.

Privacy control configuration: integrate anonymization techniques and reduce reidentification risks.

Define business and compliance objectives: ensure that synthetic data generation aligns with corporate strategic and compliance objectives.

Quality validation and assurance

Fidelity and utility control: perform statistical tests, correlations, and quality and utility controls.

Reidentification risk evaluation: apply benchmarks and simulated attacks.

Equity and bias evaluation: perform analysis to ensure the representativeness and fairness of generated datasets.

Governance, documentation, and auditing

Clear governance: define roles and responsibilities, as well as policies for their life cycle.

Comprehensive documentation: maintain updated, ready-to-use documents.

Recognized standards and certifications: adopt recognized frameworks to demonstrate compliance and strengthen trust with clients, regulators, and partners.

Continuous monitoring and improvement

Regulatory monitoring: track legislative changes (GDPR, AI Act, etc.) and adapt to new requirements.

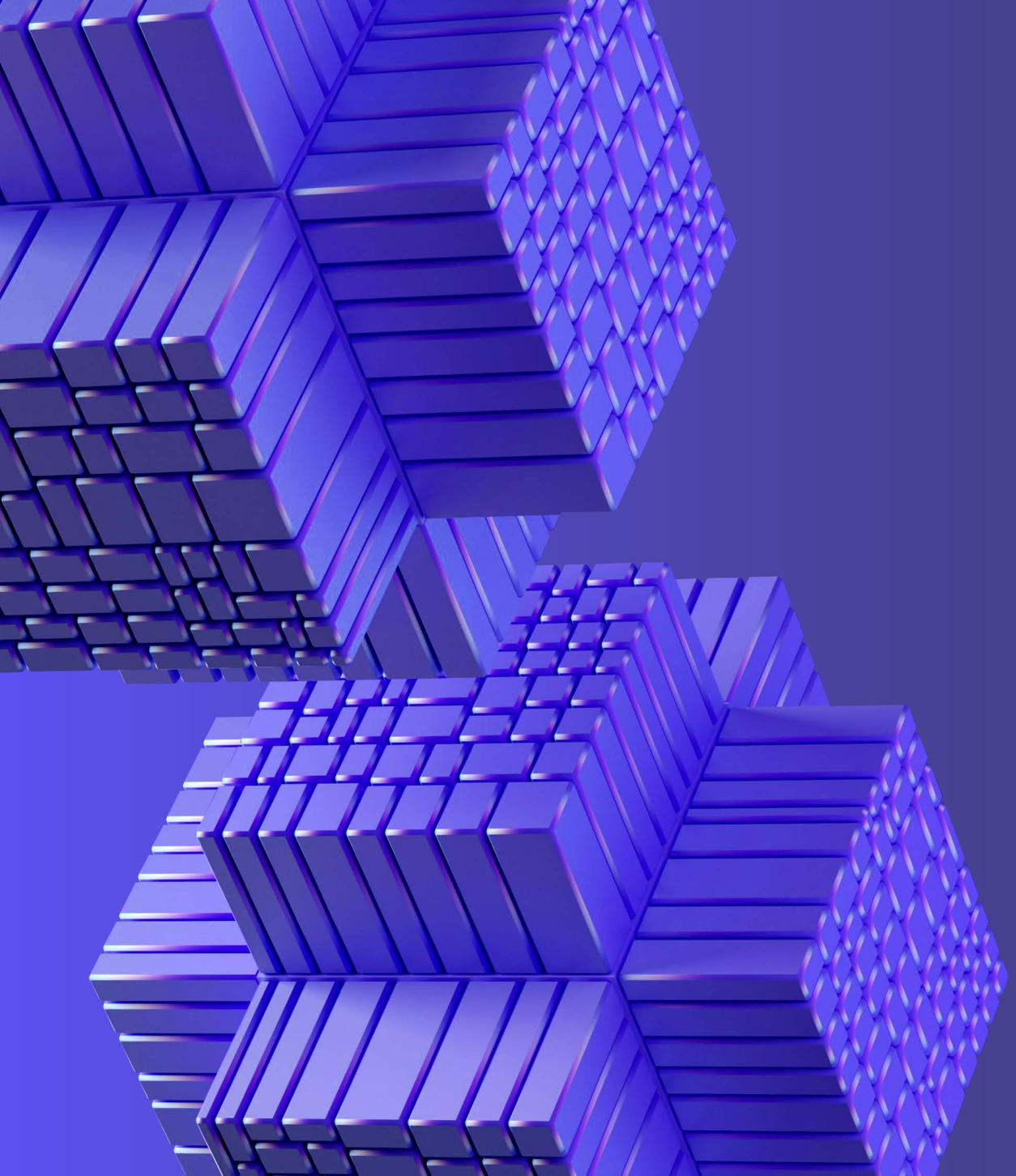
Automated quality control: use AI-driven tools to verify quality, privacy, and utility of data.

Auditing and incident response: establish alert systems and protocols to address vulnerabilities.

Training and updates

Staff training: invest in training on new technologies and best practices.

Communication and transparency: inform consumers and regulators about the use, benefits, and safeguards of synthetic data.



Stakeholders and perceptions, priorities and resistance

The general perception of industry leaders is positive, but barriers in knowledge, quality, and effective implementation persist

Opinions of industry leaders on the application of synthetic data

Currently, **89% of executives from large corporations consider the use of synthetic data as an essential element** to maintain their market competitiveness.

Industry leaders with knowledge of synthetic data technologies at the forefront **have expressed their confidence in this technology's ability to address critical problems using real-world data.** While recognizing the importance of improving data, more than half (51%) of respondents are not aligned with the explicit technical definition of advanced synthetic data approaches, revealing a critical knowledge gap.

Of those who understood the correct definition, 50% mentioned that one of the main benefits of synthetic data is overcoming the limitation of labels provided through supervised learning and human annotation.

Although many are convinced of the potential benefits of this data, **there are significant barriers to its adoption; moreover, 67% of industry leaders agree** that their organization lacks the necessary knowledge to implement synthetic data effectively. Likewise, 67% acknowledge that users in their industry will not accept synthetic data until they see clear benefits for themselves.

Source: Synthesis

Among the most challenging aspects of its use within organizations:

46%

of respondents **are concerned about the quality of models created with synthetic data**, fearing they may not be as good as those generated with "real" data.

45%

mentioned the **difficulties of creating high-quality synthetic data for complex systems.**

42%

indicated that **integration and implementation costs** represent a significant challenge.

Despite these challenges, **59% of decision-making leaders believe their industry will use synthetic data, either independently or combined with "real" data, in the future.** This suggests that many organizations are beginning to experiment with this technology but still have a long way to go. In short, while there is great interest in synthetic data, greater infrastructure and training are still needed for its widespread implementation.

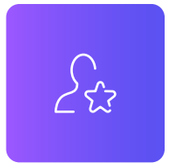
Additionally, the successful adoption of synthetic data requires alignment between CIOs, CDOs, and DPOs to integrate interests and manage change

CIOs, CDOs, DPOs: The business decision-making dashboard

The introduction of synthetic data in an organization involves various stakeholders, each with their own perspectives and concerns. Understanding these perceptions is vital to manage the change toward adopting synthetic data.

The adoption of this technology depends not only on technological innovation but also on the integration of interests. In this context, **the roles of the CIO, CDO, and DPO play a key part in decision-making that directly impacts the adoption and success of synthetic data** within an organization. Each of these actors brings a distinct approach to the decision-making table, and their strategic alignment is crucial to advancing adoption.

Synthetic data can significantly improve model performance and can effectively replace real data in 60% to 80% of cases without losing performance.



CIO (Chief Information Officer): As those responsible for digital strategy and technological infrastructure, CIOs play one of the main roles in adopting synthetic data. For them, this technology represents a strategic opportunity to generate competitive advantages through the creation of personalized AI models trained with exclusive datasets. Their focus is on accelerating innovation and mitigating technological risks. Their decision to adopt this data, however, depends on demonstrating a clear return on investment, which often goes hand-in-hand with improving process efficiency.

DPO (Data Protection Officer): The DPOs, responsible for ensuring compliance with privacy regulations, have a more critical perspective on the use of this data. For them, this technology can be a key tool to comply with the data minimization principle established by the GDPR. Their focus is on ensuring that synthetic data is truly irreversible and unidentifiable, avoiding any possibility of reidentification. They are responsible for implementing internal policies that define when synthetic data meets anonymization requirements, thus avoiding legal risks.

CDO (Chief Data Officer): The CDO is the main advocate for the quality of an organization's data and sees synthetic data as a vital solution to address access and quality challenges. They are interested in how synthetic data can unlock restricted datasets due to privacy regulations, allowing broader and easier data exchange inside and outside the company. In addition, they see in synthetic data monetization an opportunity to create products that do not infringe on privacy.

CIO, CDO and DPO must collaborate by integrating technology, governance and privacy to successfully adopt synthetic data

	CIO: driver of innovation and competitiveness	**CDO:** guardian of data value	**DPO:** ensuring privacy and regulatory compliance
Concerns and needs	• **Modernize the technological infrastructure** to adopt emerging technologies agilely. • **Accelerate the integration** of cutting-edge solutions guaranteeing differentiation in the market. • **Protect the ecosystem** by integrating AI and ML for threat prevention and data integrity. • **Address skills gaps** in teams.	• **Ensure auditability and compliance** with internal and regulatory standards. • **Ensure the maintenance of their analytical value** without compromising quality or representativeness. • **Offer a solution for the secure** exchange of data. • **Manage rapid adaptation** to new regulations and ensure data validation.	• **Ensure synthetic data** are irreversible and unidentifiable, complying with the minimization principle. • **Ensure compliance** with data protection laws and avoid any violations. • **Document generation** processes with robust controls and access to audits.
Steps to follow	1. **Align the company's strategy** with the use of synthetic data. 2. **Implement synthetic data** solutions to improve operational agility and cost reduction. 3. **Ensure that the infrastructure** can handle both synthetic and real data efficiently and at scale.	1. **Develop a clear governance strategy** aligned with internal and regulatory standards. 2. **Evaluate quality and representativeness** to ensure their usefulness in AI and advanced analytics models. 3. **Facilitate integration in data collaboration** processes between internal and external actors.	1. **Verify** synthetic data are generated following applicable regulatory principles. 2. **Implement audit** and transparency mechanisms to oversee the data generation process. 3. **Collaborate internally** to prevent reidentification risks or privacy violations of generated data.

For consumers, synthetic data combine security and progress, but require transparency to ensure acceptance and trust

What do consumers think?

As synthetic data technology continues to gain ground, consumer trust remains a key factor for its adoption and implementation. 94% of organizations acknowledge that if data is not properly protected, consumers will stop buying their products. However, **87% of consumers do not trust companies to handle their data responsibly**, highlighting the importance of building trust.

They are a truly promising solution, especially when seen as a measure to protect privacy and reduce exposure risks, as the **vast majority of consumers prioritize companies that safeguard their privacy**, increasing their loyalty. However, a lack of transparency or the unauthorized use of data can lead to distrust.

91% of consumers demand that companies protect personal data, especially in the use of technologies such as AI and synthetic data.

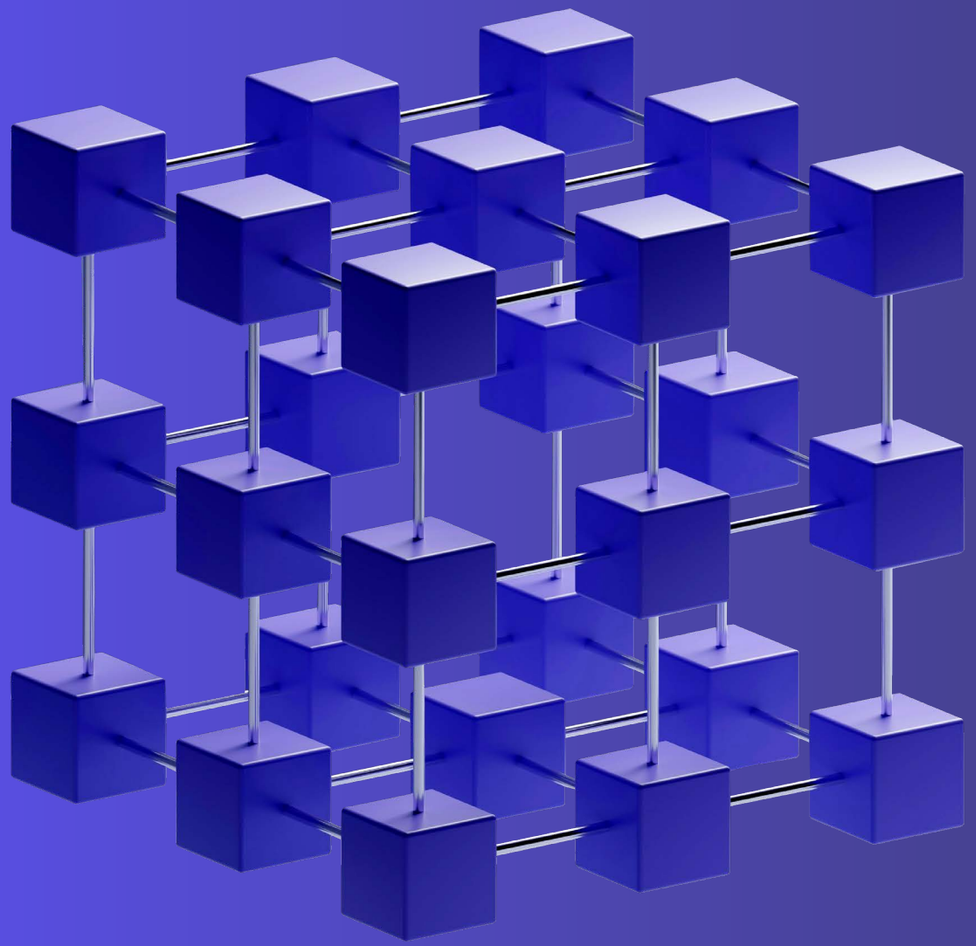
In this regard, **80% believe that privacy policies are positive**, but this depends on ethical management, transparency, and going beyond regulations to ensure the effective protection of data.

Impact of synthetic data on consumer privacy and trust

In environments where companies and end consumers interact, it is essential to protect privacy, and synthetic data achieve this by **eliminating the need to use real personal information.** This allows companies to innovate in products and services and a djust variables such as prices without compromising data security.

The direct benefits for consumers are greater privacy protection, risk reduction from security breaches, and increased transparency in privacy practices. However, the lack of public understanding about what synthetic data are **can lead to distrust if not properly communicated**, especially in algorithmic decision-making processes for data generation.

Therefore, it is essential that companies **clearly explain how synthetic data is generated and used** to ensure consumer trust and guarantee that decisions based on this data are transparent and responsible.



Industry impact: synthetic data as a driver of transformation

Early adoption is concentrated in industries where the pressure to innovate, scale, and protect data is critical for daily operations

The adoption of synthetic data follows a typical technology diffusion curve, where some industries and companies are already leading its implementation, while others remain in a much more delayed point. The following profiles appear in the adoption curve.

Innovators

As with other technological innovations, the first to adopt these solutions — the so-called innovators — are made up of industries with high R&D investment and a strong culture of experimentation, such as AI startups, Big Tech, Insurtechs, the automotive sector, and emerging Fintechs.

Early adopters: leaders in synthetic data adoption

Several industries are leading the way in **capitalizing on the use of synthetic data.** These understand the strategic advantages synthetic data offers, especially in terms of agility, regulatory compliance, and privacy.

- **Financial industry (banking, insurance).** Banks and insurers have been among the first to adopt it, driven by strict privacy regulations and a strong need to efficiently handle large volumes of data.
- **Healthcare and life sciences.** Pharmaceutical companies, as well as research centers, have begun implementing synthetic data in clinical trials and medical studies. In this sector, privacy and precision are extremely important, which has driven early adoption. Medical research is benefiting from synthetic data to conduct studies without exposing sensitive information.
- **Technology.** Large technology companies have pioneered the use of synthetic data, mainly to train AI models in simulation environments. In the automotive sector, synthetic data is being used to train autonomous driving systems.
- **Innovative public sector.** Some governments, such as those of the UK, South Korea, and Singapore, have begun experimenting with synthetic data to improve transparency and data sharing without compromising citizens' privacy.

Early majority: adopting technology after clear results

These industries are beginning to see **the clear benefits of synthetic data**, such as improved operational efficiency and resource optimization.

- **Industry 4.0.** Manufacturing companies are using synthetic data to simulate machine failures and train predictive monitoring systems.
- **Telecommunications.** Telecom companies are using it to test networks and simulate traffic without compromising user privacy.
- **Energy and utilities.** Energy companies are starting to generate synthetic data to simulate power grids and predict future demand.

Traditional industries advance cautiously, limited by structural and technological barriers and lower digital urgency

Late majority

The traditional public sector and education are examples of sectors that will adopt this technology later, **due to their more conservative nature and preference to wait** until solutions are more mature and standardized. These sectors focus on regulation and security, so the adoption of synthetic data will be slower, but it is expected to happen as the technology matures and becomes a more common practice.

Laggards

Finally, **traditional industries are the most lagging** in adopting synthetic data. Many small companies do not see the immediate need to adopt this data, as they do not have large volumes of data to analyze nor teams specialized in AI. However, as the technology becomes a more accessible solution and synthetic data tools integrate into traditional platforms, these sectors will adopt the technology more broadly.

Barriers for lagging industries

The lack of standards for creating synthetic data, regulatory uncertainty, and cultural resistance to technological change are significant obstacles. **Organizations that have not had direct experience with the advantages of synthetic data tend to be more cautious,** waiting for the technology to consolidate further and for their traditional providers to integrate it more easily into their solutions.

In the technology industry, synthetic data accelerate innovation and testing, eliminating legal barriers and real privacy risks

Synthetic data are transforming the tech industries, offering innovative solutions in areas such as advanced analytics, AI/ML, software development, cybersecurity, and simulation. As companies face increasing privacy and data security demands, **synthetic data make it possible to maintain the utility of information without compromising the protection of individuals.**

Tech companies operate with fast release cycles (CI/CD). For this, they need robust testing environments. Synthetic data have revolutionized **testing and QA practices by offering massive, realistic, and varied datasets.** A team could instantly generate hundreds of thousands of diverse cases, covering unusual or borderline conditions. By improving the quality of products, the resulting bugs often appear only in rare use cases. At an advanced level, synthetic data allow companies to provide valuable information without exposing real data, helping organizations gain insights without violating regulations. In AI and ML, they enable the creation of large volumes of high-quality data to train more precise models and improve performance, reducing costs and time. In software development, they make it easier to perform more realistic and effective tests, thus improving usability and reliability.

In cybersecurity, **synthetic data are used to train threat detection systems on sensitive data**, enabling simulations of realistic cyberattacks. Additionally, in simulation and modeling, synthetic data create realistic virtual environments, enhancing user experience and tech testing (especially for AI developers) by training them with a valuable ally to generate innovative products. Examples include autonomous driving software or AI-assisted software development companies.

Interestingly, within the **tech sector, new companies have emerged whose core product is synthetic data** (**Mostly AI, Gretel, Hazy, or Synthetatic**). These offer platforms and APIs for third parties to generate data.

In the healthcare industry, they enable training medical AI and sharing clinical data without compromising patient privacy

The health industry is adopting synthetic data as a key tool to transform areas such as **clinical trial simulation, drug development, and AI-assisted diagnosis**. These data make it possible to work with complete datasets without compromising patient privacy.

In **clinical trials, synthetic data simulate patient populations, optimizing trial designs, reducing waiting times, and lowering the financial pressure** of these services. For the first time, it is possible to share patient data to train AI without revealing any real patient information. For example, a wellness company used synthetic control cohorts and, instead of recruiting a control group, generated synthetic patients with characteristics similar to those in the real control group to compare treatments. They are also used in health prediction models, simulating disease progression and improving health monitoring systems through AI.

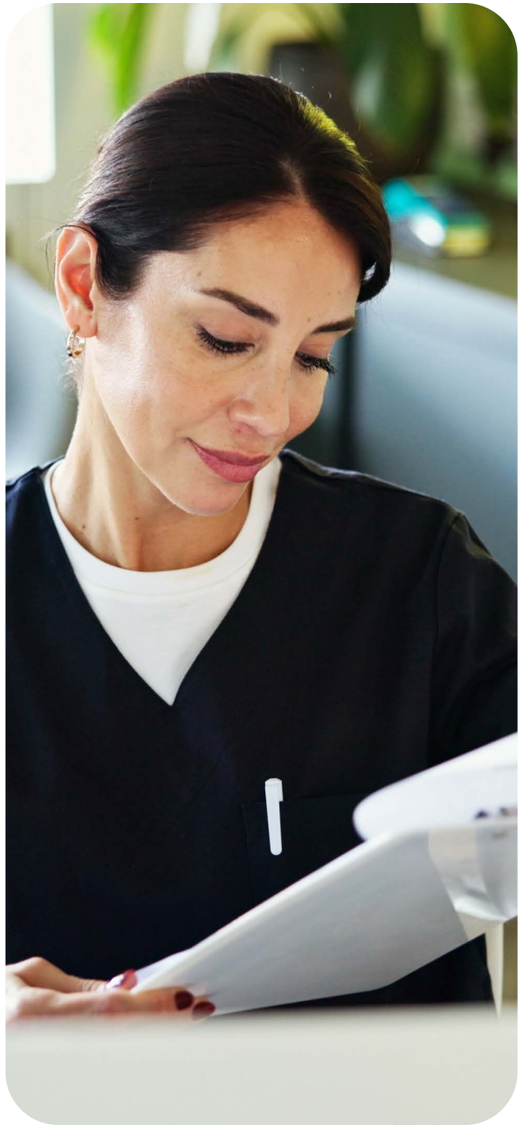
The medical academic community often relies on valuable datasets (hospital records, epidemiological registries) that cannot be freely shared for multicenter analyses. Synthetic data make it possible to create public repositories of synthetic health datasets, such as the UK Biobank, which contains genetic and clinical data from 500,000 people and is exploring the creation of a synthetic data bank showing the same genotype–phenotype correlations without exposing any real individual.

In the medical imaging field, they provide **realistic images and multimedia content to train diagnostic and early detection models without using personal data,** ensuring privacy and improving diagnostic accuracy. They also enable the creation of disease risk models, including rare populations and conditions, which allows identifying at-risk patients earlier.

In **drug development**, they **accelerate the treatment testing phase** and patient tracking for new therapies, reducing research-related costs. In public health, synthetic data are used to simulate disease outbreaks and support health policy planning.

A key factor is ensuring that synthetic data **maintain clinical precision.** In other words, any model or conclusion derived from synthetic data should reflect medical reality. So far, evidence suggests this is achievable.

30% of all data worldwide are healthcare data, and this figure is increasing.



Sources: Leeway Hertz I Reply I Syntheticus I Syntho I V7

In the financial industry, synthetic data enables fraud simulation, risk assessment, and model testing without using real customer confidential information

The financial industry (banks, insurers, capital markets) is **one of the sectors most rapidly transforming with the use of synthetic data**, applying it to fraud detection, financial stress testing, and risk modeling.

In fraud detection, synthetic data simulate fraudulent transactions, which improves the accuracy of AI models without compromising customer privacy. For example, Visa conducted a study in 2024 where it generated synthetic data of anomalous behaviors (such as fictitious merchants attempting to scam) and trained a detection model. The result was a **15% improvement in the fraud detection rate without using any real cardholder transactions.** Similarly, in credit scoring, synthetic data is used to train credit scoring models without accessing real financial data, which allows for greater equity and accuracy.

Another area is **AML (Anti-Money Laundering). Banks have started using synthetic data to simulate money laundering transaction networks** and t hus test their monitoring systems.

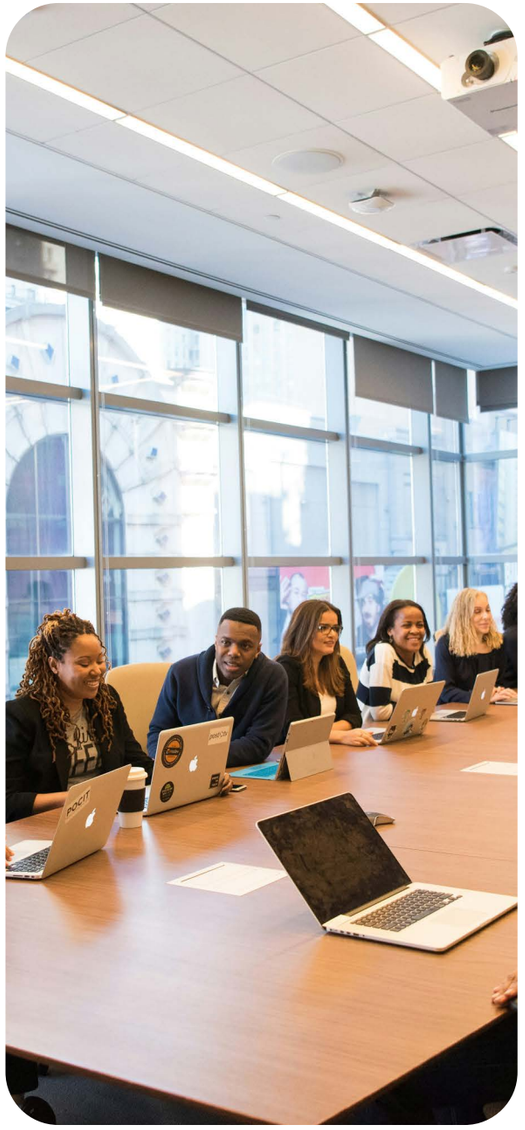
Global banks have mentioned that they have been able to improve the accuracy of their AML alerts by adjusting parameters on these synthetic datasets, which replicate money laundering structures (such as "**smurfing**" or fractional transfers), producing no risk and simulating a real case.

Stress models and scenario analysis are other areas where synthetic data prove extremely useful, enabling financial institutions to generate datasets that simulate market fluctuations and adverse scenarios, helping organizations test market volatility resilience and regulatory compliance. For example, generating one million synthetic client datasets with their jobs and assuming the unemployment rate rises to 15%, how many would default? This estimate can only be made by having aggregated projections, but synthetic d ata allow this heterogeneity to be modeled, producing more realistic models.

When creating scoring or provisioning models, sometimes there is insufficient crisis behavior data. Synthetic data fill that gap and **improve model robustness to unseen events.**

In addition, synthetic data **facilitate test management in development or acceptance environments,** significantly reducing test times and operational load. They also improve process innovation by enabling faster testing of new workflows without compromising sensitive data.

50% reduction in QA team effort thanks to AI-driven test data generation, enabling faster test cycles and fewer defects.



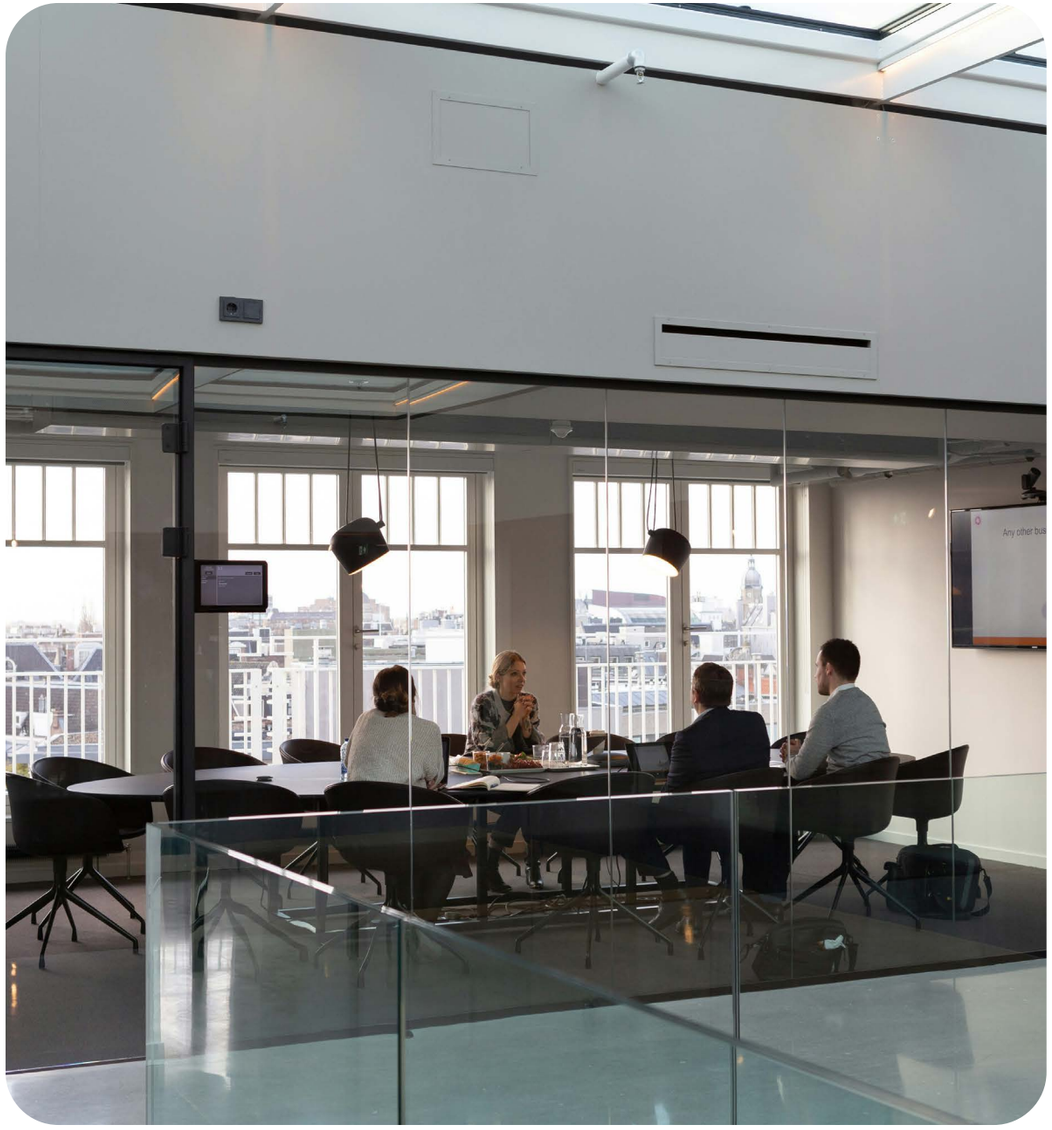
Source: Reply

In the insurance industry, they help assess claims, model extreme scenarios, and personalize offers without exposing sensitive data

The insurance industry is greatly benefiting from the use of synthetic data to optimize its operations, from risk assessment to fraud detection and claims management. **This data allows insurers to create more accurate predictive models, simulate high-risk events, and comply with regulations.**

One of the biggest challenges for insurers is risk assessment and policy rating, areas that benefit enormously from this data. It allows for profiling clients, improving the accuracy of risk models. **When it comes to catastrophic or high-risk events, insurers can better anticipate and plan for unforeseen situations,** reducing their exposure to risk.

Fraud detection is another key area. Synthetic data enables the creation of large volumes of simulated fraudulent transaction data, improving the effectiveness of detection models without exposing clients' sensitive information. And, in claims management, synthetic data allows for scenario testing without the need for real client data. Insurers can also use this data to enhance customer experience personalization, developing products and services tailored to their needs.



In gaming and entertainment, they enable the creation of realistic profiles and scenarios, improving design without compromising user privacy

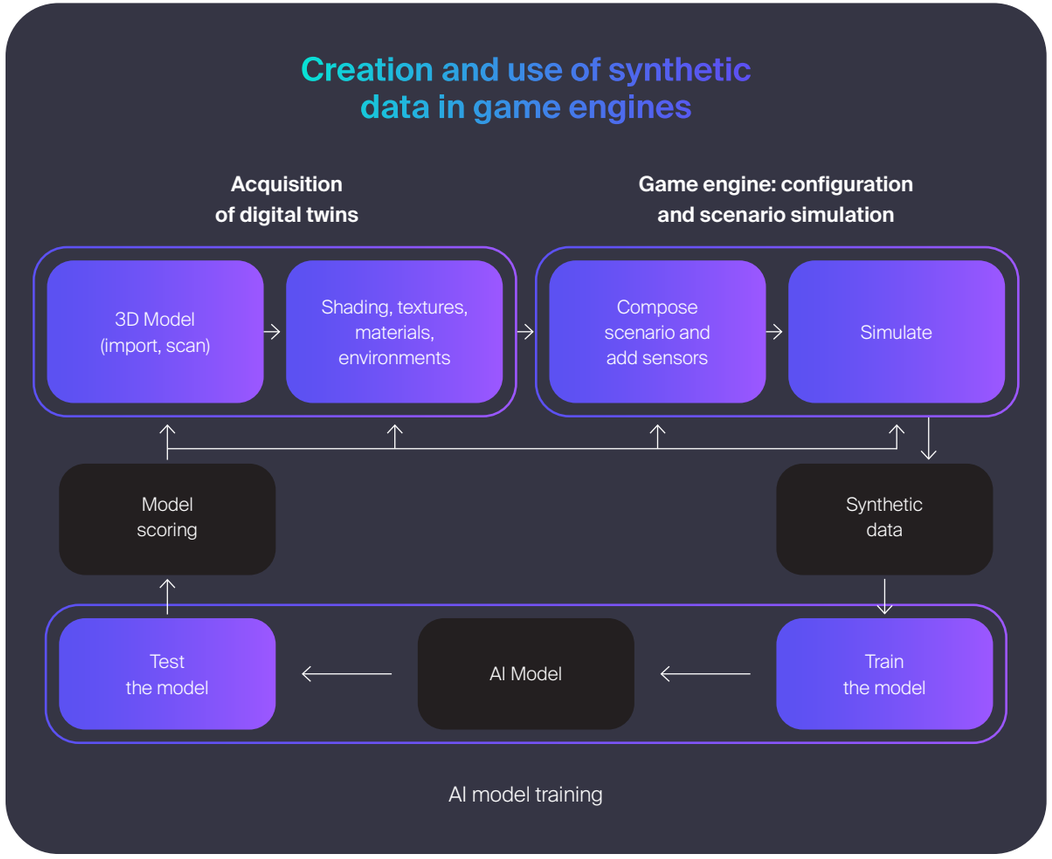
Within the video game and entertainment industry, AI-based technologies have always been strong drivers. In the case of synthetic data, there are several areas where it is useful to implement technology that **not only improves efficiency in creation processes but also enhances the player experience and optimizes content creation.**

One of the biggest challenges in video game development is creating detailed characters and realistic environments that maintain immersion and creativity. Synthetic data allows **developers to generate characters and artificial intelligence scenarios that accurately replicate the desired aspects** of the world without the need to design them manually from scratch, significantly accelerating production timelines.

Another aspect of gameplay experience is interaction with intelligent and challenging opponents. The use of this data to **simulate player and opponent behaviors** allows AI opponents to be trained without needing difficult-to-collect real data. All of this facilitates the creation of more realistic opponents that respond to a more dynamic game flow and adapt to player strategies.

Video game development also requires extensive testing to **detect errors and improve game balance.** This technology streamlines and simplifies the simulation process by generating game scenarios and synthetic player behaviors to detect problems such as server failures or gameplay difficulties. For example, in a multiplayer game, thousands of synthetic players can be simulated to test server load, identify bottlenecks, and adjust gameplay balance.

Finally, **dynamic levels and personalized gameplay experiences** are key in many modern titles, and through procedurally generated content with synthetics, developers can create adaptive game levels that change according to player skill, offering a personalized and evolving experience. Synthetic data enables the creation of a wide variety of environments and situations that maintain player interest without the need for manual design.



In the education industry, synthetic data improves learning and, in transportation, plans mobility while respecting citizen privacy

In the educational field, synthetic data can serve various purposes. From educational research, academics studying student performance, school dropout, or other factors often have limited data or difficulties accessing real records due to privacy concerns. With synthetic data, they could obtain **artificial datasets that replicate student characteristics, grades, or socio- economic contexts**, allowing them to test hypotheses, for example, how certain interventions would affect different types of students.

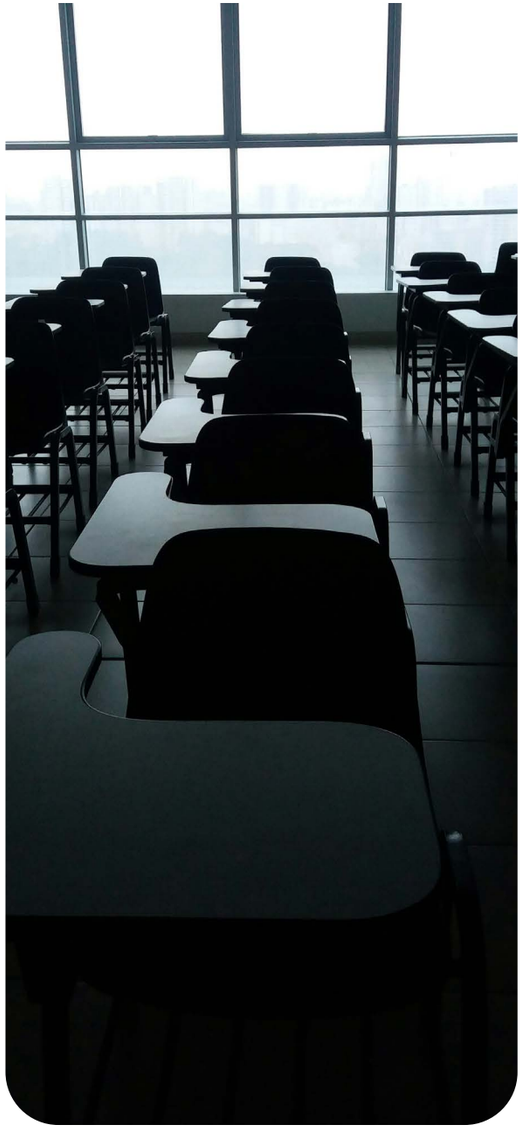
Edtech companies (learning platforms, intelligent tutoring systems) can use **synthetic student interaction data to train their recommendation or difficulty detection algorithms,** without the need to wait to collect thousands of hours of actual usage. For example, generating synthetic sequences of student responses to exercises based on observed patterns to pre-train a system that identifies when a student is stagnating. This improves these systems before launching them, with fewer real data points.

In the urban transport and mobility sector, synthetic data is used to generate synthetic mobility data, calibrated with real counts, to **test infrastructure changes and plan smart cities.** For example, a city can generate 100,000 synthetic citizens with movement patterns based on mobility survey data, then simulate what happens if a main street is closed or a new subway line is introduced and observe how traffic is rerouted.

This same approach helps planners **predict the impact on public transport** without having to experiment in real life first. Smart cities like Singapore already use digital twins and synthetic agents for such purposes.

50% adoption of synthetic data in Finance and Smart Cities projected for 2026

Public transportation agencies can use synthetic data **to evaluate safety policies,** for example, by simulating millions of synthetic vehicle trips in an area with different speed limits to estimate accidents, instead of implementing and waiting for years, or public transportation can simulate synthetic passengers in a bus network to optimize schedules or routes without affecting real passengers during experimentation.



Source: LinkedIn

In public administration, synthetic data improves planning and inter-institutional collaboration without exposing protected personal data

The implementation of synthetic data in the public sector represents an opportunity **to improve efficiency, accelerate decision-making, and ensure the protection of sensitive data.** Governments manage large volumes of confidential data, and synthetic data makes it possible to create datasets that preserve the characteristics of real data without compromising privacy.

They also facilitate **collaboration between public and private organizations**, by removing legal and privacy restrictions on data sharing. This speeds up innovation in public policies, research, and inter-institutional projects. A clear example is the use of synthetic data in fraud detection, where data can be generated without exposing actual information.

Additionally, synthetic data enables **simulation of public policies**, such as budget reassignment, providing a more accurate analysis than traditional assumptions and can be used to create safe training environments where public officials can practice with data without exposure risks.

Moreover, in crisis or disaster situations, governments can **simulate different scenarios to optimize their emergency response**, from epidemics to natural disasters.

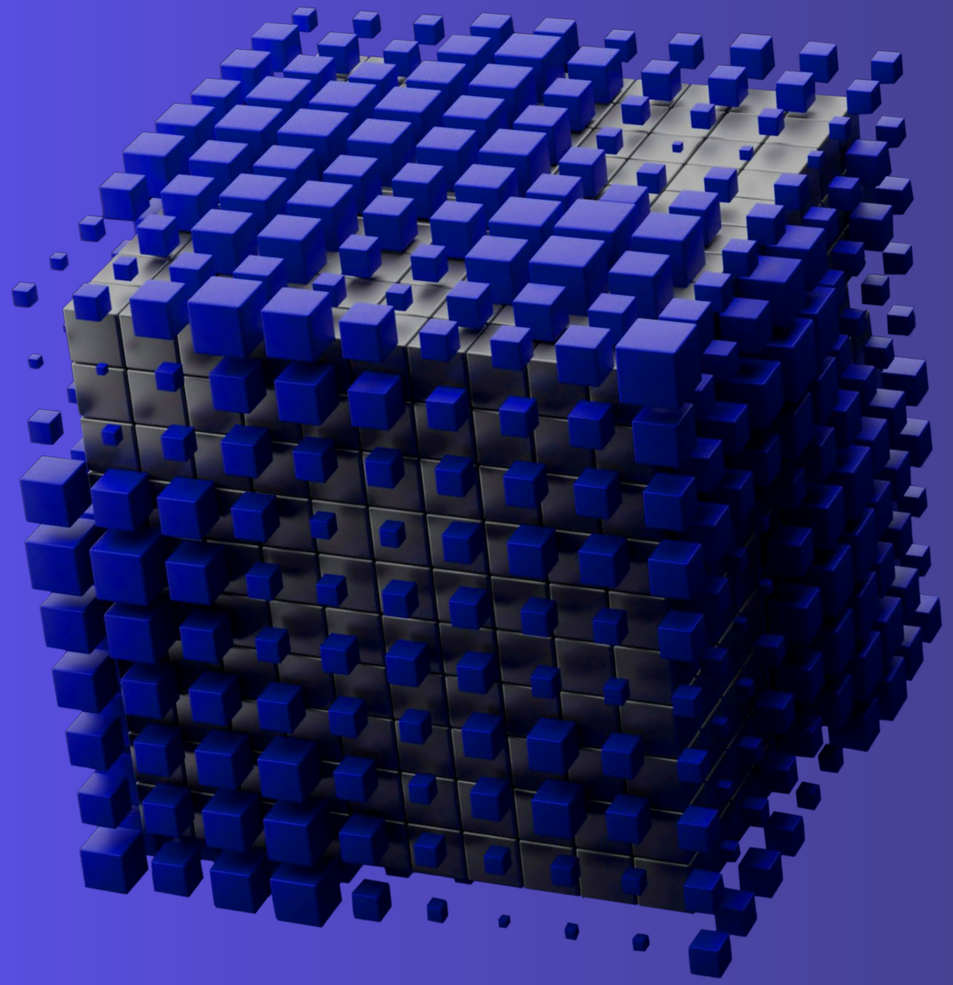
45% of government IT leaders point to data infrastructure as a barrier to digitalization

56% of public organizations mentioned data sharing and privacy as a challenge

Source: Reply

All sectors are driven by privacy, compliance, innovation, cost reduction, scalability, and risk mitigation

Industry	Use cases	Drivers
Technology sector	Network optimization, cybersecurity, software testing, AI training	Implementation of 5G, edge computing, strengthening of security
Healthcare	Clinical trials, diagnostic imaging, patient data protection, pharmaceutical research	Regulatory compliance, data scarcity, acceleration in research
Finance, insurance, and financial regulators	Fraud prevention, risk modeling, algorithmic trading, compliance audits	Regulatory requirements, real-time risk assessment
Gaming and entertainment	Video game development, immersive experiences, AR and VR simulation, graphics optimization, player behavior analysis	Demand for interactive content, adoption of immersive technologies, streaming platforms, monetization of online experiences
Education	Adaptive learning, online education platforms, educational simulations, real-time evaluation, virtual collaboration tools	Personalization of education, expansion of online learning, continuous training, emerging technologies
Transportation and urban mobility	Autonomous vehicle simulation, extreme scenario testing, sensor calibration	Advances in autonomous driving, safety validation, cost optimization
Government and public administration	Simulation training, intelligence analysis, public safety, smart cities	National protection, citizen privacy, operational continuity
Retail and e-commerce	Consumer behavior modeling, experience personalization, inventory management	Customer privacy protection, competitive analysis, market expansion
Manufacturing industry	Digital twins, quality control, predictive maintenance, process optimization	Industry 4.0, Internet of Things (IoT) integration, operational efficiency improvement

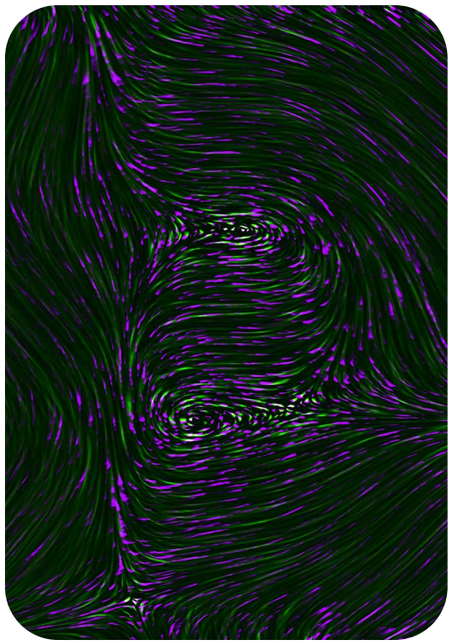


Conclusions and predictions: the future of synthetic data

Synthetic data is an immediate response to the limits of real data and, in the coming years, will dominate corporate data strategy

The emerging promise of synthetic data has already become a **critical tool in business data strategies.** Over the next 5 to 10 years, exponential adoption is expected, driven by regulatory trends that require data protection by design, the growing need for secure and scalable data for AI and advanced analytics, and the practical limitations of using real data and traditional anonymization.

Future trends in the adoption of synthetic data point to a **landscape in which businesses and governments will benefit from the creation of more robust, competitive, and agile artificial intelligence models.**



What to expect in the coming years

Business competitive advantage,
where organizations that master synthetic data generation innovate faster than those relying on real data.

National sovereignty and competitiveness,
with countries like South Korea and Singapore using this data to boost their global competitiveness.

Creation of synthetic data markets,
with companies offering data packs, opening the possibility for global exchanges in situations of global interest.

Resilience and business continuity,
enabling operation against disruptions or attacks and speeding up overall adaptation to change.

Accelerated innovation in sectors,
such as healthcare, technology, and cybersecurity, where testing times can be reduced, speeding up the development of new products.

Potential new gap in access to the best data generators,
which could be dominated by large technology corporations.

To transform risks into opportunities, companies must strengthen controls, quality, and talent in synthetic data

Risks associated with synthetic data and priority actions for companies

Although its adoption offers multiple opportunities, it also **presents risks that companies must recognize and address** to maximize the benefits of this technology without compromising its integrity.

Companies must adopt a **proactive approach to synthetic data adoption**, which includes investing in technologies and generative models, collaborating with specialized partners, and creating ethical and governed frameworks for the responsible use of this data. Furthermore, building controlled and simulated testing environments will enable a smoother transition without compromising privacy and security.

Disinformation and misuse of synthetic data can lead to manipulations that generate misinformation or bias results.	**Establish clear ethical frameworks** for its creation and use, and implement rigorous procedures for monitoring data quality.
Gaps in data quality can lead companies to make wrong decisions, impacting both their operations and customer experience.	**Regularly monitor data quality** to ensure that models trained on it are not defective, and establish an internal regulatory framework to govern how data is generated and evaluated.
Security and privacy risks arise because both the generation and processing of this data can be vulnerable to cyberattacks or allow information inference.	**Invest in cutting-edge technology infrastructure** to ensure synthetic data platforms are protected against threats, and implement privacy-by-design policies.
Concentration of power in large technology corporations could lead them to dominate access to this technology, creating a new gap.	**Foster collaboration** with open ecosystems and promote the creation of accessible platforms where all stakeholders can participate.

By 2030, data protection will require adaptive governance, using synthetic data to ensure compliance and trust

The future of privacy: an imminent change

We are on the verge of a transformational shift in how we manage privacy. As the regulatory landscape for data protection continues to evolve, laws are becoming stricter, more complex, and with greater implications. As we move toward 2030, **the future of privacy is shaping up as a space full of challenges and changes, but also strategic opportunities.** We have already seen early signs of this change: the GDPR has set a strong precedent, imposing multi-million-dollar fines on companies that fail to meet its strict requirements. Yet, this is only the beginning.

This new approach will not stop at protecting personal data; **future regulations will cover much broader areas, including AI ethics and the responsible use of emerging technologies.**

Privacy will no longer be just a matter of personal data, but of how organizations use it to analyze and make decisions. In the future, we will witness a shift toward **“hyper-regulation”**, where organizations will need to be more prepared than ever to meet increasingly strict and complex regulatory requirements.

As regulation becomes **more global and interconnected**, data management will no longer be just about complying with a single national law but navigating a complex maze of multijurisdictional regulations. However, those able to adapt quickly to regulatory changes will benefit from enhanced transparency and compliance as competitive advantages.

Data sovereignty is becoming increasingly important. Information fragmentation can hinder innovation, as companies may be unable to access or share data between regions. But not everything is negative.

Opportunities arise when companies adapt to this change with a localized strategy in each jurisdiction, demonstrating their commitment to national regulations.

This is where synthetic data comes into play as an **enabling tool for the companies of the future**, allowing organizations to comply with regulatory requirements while continuing to innovate. Entities adopting them will be better positioned to navigate the increasingly complex regulatory landscape and capitalize on the opportunities that come with compliance.

Starting well requires 5 critical steps and leading means institutionalizing synthetic data as a driver of responsible innovation

5 Immediate actions to get ready

1. **Audit the current use of sensitive data**
   Identify critical processes that rely on personal data and detect legal risks.
2. **Form a cross-functional synthetic data team**
   Include IT, legal, compliance, and analytics profiles to coordinate responsible and effective adoption.
3. **Select a first synthetic use case pilot**
   Choose a high-impact, low-risk area (testing, R&D, or internal analytics) to generate synthetic data.
4. **Evaluate synthetic data generation providers or technologies**
   Reduce the burden on internal resources and accelerate timelines.
5. **Design an ethical framework and governance for synthetic data**
   Establish core principles around transparency, traceability, validation, and limits of use.

5 Actions to lead the trend

1. **Integrate synthetic data as part of the data and AI strategy**
   Consider them a pillar of innovation, compliance, and scalability.
2. **Develop internal generation and advanced validation capabilities**
   Invest in talent, R&D, and in-house tools to avoid dependence on third parties.
3. **Drive public policies and industry standards**
   Actively participate in industry forums and regulatory sandboxes to help shape sector norms.
4. **Create new synthetic-data-based solutions**
   Develop simulators, train AI, and testing platforms that ensure privacy, innovation, and competitive advantage.
5. **Become a responsible-use success story**
   Publish results, share lessons learned, and position yourself as a benchmark in ethical and innovative use.

Softtek®

softtek.com