

White Paper

Synthetic Data: Changing the Game for Market Research with AI-Powered Insights

Published : January 2025

Copyright © 2025 Knometrix



Table of Contents

Introduction

Applications in Market Research

Understanding Historical Data's Role in Synthetic data

Synthetic Data Generation Methods

Quality and Ethics Considerations

Case Studies

Conclusion

Use of the White Paper This publication is provided by Knometrix for general guidance only and does not constitute the provision of business, technology, investment or product advice. The information is provided “as is” with no assurance or guarantee of completeness, accuracy or timeliness of the information, and, to the extent permitted by law, without warranty of any kind, express or implied.

No part of this publication may be quoted, cited, excerpted, reproduced, stored in a retrieval system or database, distributed or transmitted in any form or by any means without the prior written permission of the Knometrix.

Requests should be submitted in writing to us at info@knometrix.com outlining which excerpts you wish to use and the context in which you wish to use it.
Disclaimer

This white paper does not necessarily reflect the views and recommendations of individual members of the White Paper working group nor the company (“Knometrix”)

Executive summary

In a time of stringent privacy regulations and challenges in accessing real-world data, synthetic data has become a transformative tool for market research. This white paper explores how synthetic data is reshaping the way businesses understand consumer behavior, forecast demand, and refine marketing strategies, all while maintaining privacy compliance and reducing operational costs.

Synthetic data offers a practical alternative to traditional data collection methods by generating high-quality, privacy-safe datasets that retain statistical accuracy without including personal information. Advanced techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) enable organizations to create datasets that reflect complex consumer behavior and market trends, facilitating extensive market research and testing without the costs and risks associated with collecting real-world data.

Real-world examples demonstrate the significant value of synthetic data. Nationwide Building Society used synthetic data to streamline their partner evaluation process, cutting timelines and improving decision-making. Similarly, Amazon utilized synthetic data to expand Alexa's language capabilities into new markets, achieving high comprehension while safeguarding user privacy.

The impact of synthetic data spans multiple areas of market research. Businesses can accelerate research through rapid A/B testing and scenario simulations, optimize marketing strategies, and improve demand forecasting without relying solely on real-world datasets. This approach minimizes risks, reduces costs, and provides a secure environment for testing strategies.

However, successful implementation requires careful consideration. Validation of synthetic datasets is critical to ensure accuracy and prevent biases from affecting results. A balanced approach that combines synthetic and real-world data offers the most effective outcomes. This white paper provides a comprehensive guide to implementing synthetic data solutions, offering actionable insights and best practices for market research professionals looking to harness this innovative approach.



Introduction

The landscape of market research is changing rapidly, with synthetic data leading the way in this evolution. In a time of increasing privacy concerns and tighter regulations on data usage, synthetic data provides a practical and innovative solution. By creating artificial datasets that replicate the characteristics of real-world data, organizations can navigate privacy constraints while achieving significant operational advantages. Synthetic data not only addresses privacy issues but also helps reduce project costs, shortens execution timelines, and facilitates the rapid launch of new products, even when no prior data is available. This approach enables businesses to extract valuable insights quickly and efficiently, making it a game-changer for modern market research.

Synthetic data, first introduced by Donald B. Rubin in 1993, has become a valuable tool for protecting privacy while still allowing for effective data analysis. This approach creates artificial data that mirrors real-world patterns without using any sensitive information directly. Companies like Tesla and Microsoft have taken this concept further, using synthetic data to solve complex problems in a variety of fields, including market research, where privacy concerns and data access limitations often arise. Today, the growth of synthetic data is driven by its ability to generate diverse and realistic datasets, reducing the need for real data that may be costly or difficult to obtain. It also allows businesses to simulate complex scenarios, test ideas, and predict trends more accurately. As industries continue to recognize the benefits of synthetic data, its application has expanded to areas like autonomous driving, customer behavior analysis, and product testing.

This paper delves into the methods, advantages, and real-world applications of synthetic data in market research, highlighting how it is reshaping the way businesses model consumer behavior, forecast demand, and optimize marketing strategies.

Quick comparison of Data Types: Real Life, Mock, and AI-Generated Synthetic Data

	Real life data	Mock data	AI-generated Synthetic data
Definition	Data collected from actual events, transactions, or user interactions.	Simplified, manually created data used for testing, training, or demonstration purposes.	Artificially created data that mimics the statistical patterns and properties of real-life data using advanced algorithms, such as machine learning models.
Realism	Fully realistic with natural variations, outliers, and noise found in real-world scenarios. Captures complex relationships and interdependencies between variables that occur organically.	Often lacks complexity and sophistication. Typically contains simplified patterns and relationships, missing the nuanced variations and edge cases found in real data. May follow overly uniform or predictable patterns.	Highly realistic when well-trained, maintaining statistical properties and correlations of the original data. Can generate diverse scenarios and edge cases, though quality depends on training data and model sophistication. May occasionally produce subtle artifacts or unrealistic combinations.
Inputs	Actual data collection methods, such as primary and secondary research methodology.	Does not require data samples, the user needs to define the rules (or randomness) how the data is created with limited reliance on real-world context or patterns.	Real-life data as a reference to train machine learning models such as historical datasets, transaction patterns, or user behaviours.
Output	Insights directly representative of actual customer actions, market trends, or operational outcomes with precision and reliability when processed and cleaned.	Outputs are functional for validating systems (e.g., testing database connections or layouts) but not useful for analysis or decision-making.	Artificially created datasets that replicate real-world patterns and diversity without containing actual records.
Drawbacks	Potential risks of bias, missing data, or privacy violations if not managed properly.	Basic fields and values to simulate structure but lack complexity or authenticity.	Poor inputs or models may result in biased or unrealistic datasets.
Usefulness	Derives actual behaviours, trends, and outcomes.	Validates system behaviour, such as testing how an app handles user inputs and Demonstrates product functionality to stakeholders.	Ensures models learn patterns without using sensitive data. Also, simulates various conditions, such as market shifts or consumer behaviour changes.

Applications in Market Research

01 Consumer Behavior Analysis

Synthetic data is increasingly preferred for understanding customer behavior due to its ability to address key challenges in market research. Various challenges such as limited data or unavailable data and allows businesses to test "what-if" scenarios, such as pricing strategies or marketing campaigns, without real-world risks. Traditionally, companies relied on small sample sizes to derive insights into buying patterns, often constrained by data availability and privacy regulations. Synthetic data allows for the creation of large, representative datasets that simulate real-world behavior, enabling businesses to uncover hidden trends and better predict future actions. Additionally, it is a cost efficient, time saving, customisable method that enables businesses to tailor datasets for specific market conditions or customer profiles.

For example, synthetic datasets enable researchers to model complex decision-making processes, simulate product interactions, and test marketing strategies across diverse customer segments. By mimicking the statistical patterns of real data, synthetic data allows for accurate predictions without the constraints imposed by real-world data limitations.

02 Enhancing Demand Forecasting




Demand forecasting is a vital part of business strategy, helping companies prepare for future market conditions. Traditionally, businesses rely on historical data to make these predictions, but real-world data often falls short, especially when dealing with unusual or unexpected events. Synthetic data helps bridge these gaps by allowing businesses to simulate a variety of scenarios, including market fluctuations or rare occurrences. This approach enables companies to create more accurate and flexible forecasts, improving their ability to adapt to changing dynamics. For instance, retail brands can use synthetic data to model consumer purchasing patterns during peak seasons, or to simulate the impact of a new competitor entering the market. This increased flexibility leads to more accurate and actionable forecasts, helping businesses optimize inventory, pricing, and promotional strategies.

03 Accelerating A/B Testing and Scenario Simulations

A/B testing and scenario simulations are essential for businesses to understand consumer preferences and make informed decisions. However, these processes can be time-consuming and costly, particularly when relying on limited real-world data. Synthetic data accelerates A/B testing by providing an environment where businesses can quickly test different scenarios, such as product variations, pricing models, or marketing tactics. By generating consumer responses to these changes, businesses can evaluate multiple strategies in parallel without waiting for months of real-world data to accumulate. This not only speeds up decision-making but also reduces the costs and risks associated with traditional A/B testing.

Understanding Historical Data's Role in Synthetic data

Synthetic data generation is deeply rooted in the analysis of historical records. These records capture the patterns and behaviors that characterize consumer decision-making, such as purchase timing, payment methods, and seasonal preferences. By analyzing large datasets of past behaviour patterns and preferences, companies can identify recurring patterns and trends that form the basis for generating synthetic data. For instance, historical data can reveal insights like

 <p>Footfall Timing: Identifying the specific times when consumers are most likely to make purchases, allowing businesses to optimize staffing, promotions, and inventory for peak demand periods.</p>	 <p>Payment Preferences: Understanding shifts from traditional payment methods to digital ones.</p>	 <p>Seasonal Trends: Observing how consumer behavior changes during different times of the year.</p>
--	---	--

This historical foundation ensures that synthetic data retains a high degree of realism and relevance, accurately reflecting the patterns of actual consumer behavior.

How does it help?

Builds Consumer Portrait with Demographics

One of the key strengths of synthetic data is its ability to create detailed consumer profiles that reflect real-world diversity. By integrating demographic factors such as age, income, location, and household composition, businesses can simulate how different consumer segments behave under various conditions.

For example, an e-commerce platform could use synthetic data to model purchasing habits by demographic group, such as:

- How younger, tech-savvy consumers prefer shopping via mobile apps.
- How families with children tend to purchase in bulk or prioritize value deals.
- How high-income consumers exhibit brand loyalty in luxury categories.

These consumer portraits enable businesses to better understand their target markets and tailor their marketing efforts more precisely.

Provides Decision-Making Insights

Synthetic data aids in A/B testing and scenario simulations by replicating historical data patterns related to consumer decision-making. It reveals key factors influencing purchasing behavior, such as price sensitivity, brand loyalty, and the impact of external influences like promotions or economic shifts. For instance, historical data can be used to:

For example, consider how historical purchase data can inform:

- Price Sensitivity: How consumers in different regions respond to price changes, helping businesses adjust their pricing strategies.
- Brand Loyalty: How repeat customers behave over time, assisting in customer retention strategies.
- Cross-Selling Opportunities: How purchasing patterns for one category (e.g., food) correlate with the likelihood of purchasing related items (e.g., beverages).

By synthesizing these insights, businesses can create realistic datasets that simulate actual consumer decision-making, driving more informed marketing and product development strategies.

Aids in Policy Formulation:

Due to strict privacy concerns in the healthcare and financial sectors, historical synthetic data plays a crucial role in formulating effective policies. By replicating real-world patterns and outcomes, synthetic data allows policymakers to analyze past trends and predict future scenarios without relying on sensitive or incomplete data. This provides valuable insights into key factors influencing policy decisions, such as treatment effectiveness and financial stability.

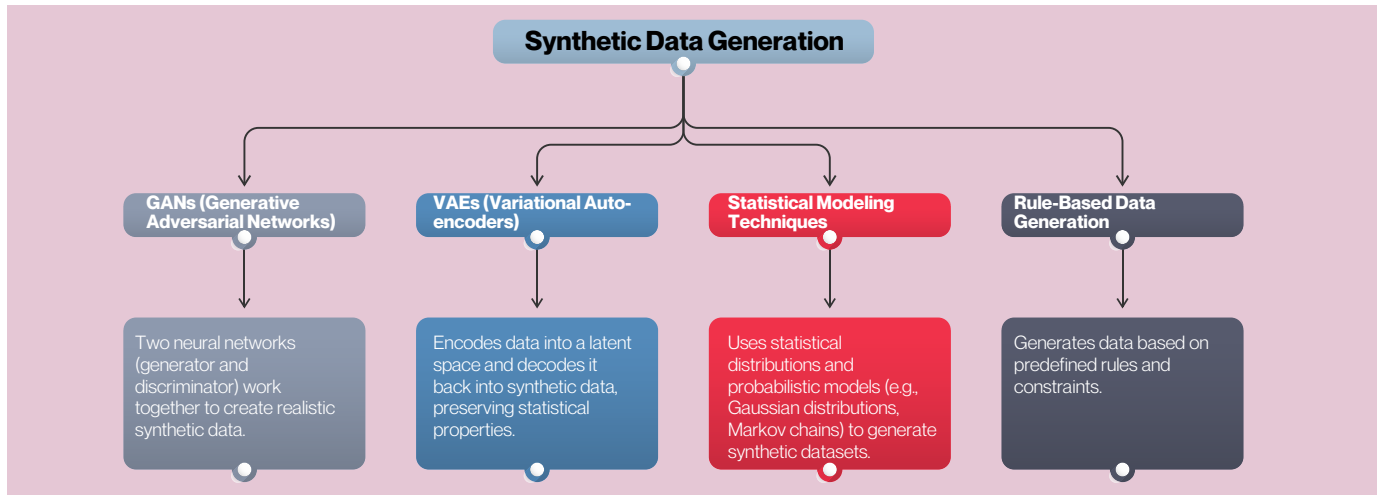
For instance:

- Resource Allocation: By replicating historical data on hospital admissions and insurance claims, synthetic data helps efficiently allocate resources and supports informed policy decisions in healthcare.
- Credit Risk Modeling: Assessing credit risk involves multiple factors, such as economic conditions, market trends, and individual behavior. Using synthetic data can simulate various economic scenarios and borrower profiles, and assist financial institutions to test their credit risk models.

By using synthetic data, policymakers can make more informed decisions that effectively address current issues while anticipating future challenges in both healthcare and finance.

Key Methods for Generating Synthetic Data

Several advanced techniques are used to generate synthetic data, each contributing unique advantages depending on the specific application. Below are some of the most prominent methods:



Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) consist of two neural networks: the generator and the discriminator that work in opposition to produce realistic synthetic data. The generator creates new instances, while the discriminator evaluates their authenticity against real datasets. This adversarial training process enables GANs to capture complex patterns and distributions, resulting in high-quality synthetic datasets.

Applications in Market Research:

- **Consumer Behavior Simulation:** GANs can generate realistic consumer profiles and behavior patterns, aiding in segmentation and targeting strategies.
- **Market Trend Analysis:** By creating synthetic datasets that reflect potential market scenarios, researchers can better understand consumer responses to new products or services.

Statistical Modeling Techniques

Statistical modeling techniques involve using established statistical distributions or probabilistic models to generate synthetic datasets. These methods are particularly useful when researchers have a clear understanding of the underlying statistical properties of the original dataset.

Applications in Market Research:

- **Scenario Simulation:** Statistical models can simulate various market scenarios based on historical data, providing insights into potential future trends.
- **Demographic Data Generation:** By applying statistical techniques, researchers can create synthetic demographic datasets that reflect specific population characteristics relevant to their studies.

Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) are generative models that encode input data into a latent space and then decode it back into synthetic outputs. VAEs are particularly effective at preserving the statistical properties of the original dataset while generating new samples, making them suitable for various applications.

Applications in Market Research:

- **Survey Response Generation:** VAEs can be employed to create synthetic survey responses that reflect real-world variability, enhancing the robustness of survey-based studies.
- **Customer Profile Creation:** Researchers can use VAEs to generate diverse customer profiles that help in understanding different market segments.

Rule-Based Data Generation (Non-ML Method)

Rule-based data generation involves creating synthetic data based on predefined rules and constraints set by users. For instance, if a researcher wants to generate fake customer records, they might specify rules like "age must be between 18 and 65" or "income must be a positive number." The system then generates data that adheres to these rules.

Applications in Market Research:

- **Data Enrichment:** Rule-based generation can create additional rows or columns in existing datasets, helping researchers expand their datasets efficiently.
- **Data Cleansing:** This method can also assist in cleaning existing datasets by generating consistent values that correct inconsistencies or fill missing values.

Maintaining Realism and Representativeness in Synthetic Data

Ensuring that synthetic data accurately mirrors real-world patterns is crucial for making reliable and actionable market research decisions. For instance, when a company generates synthetic consumer behavior data to test reactions to a new product, the quality of this data impacts key strategic decisions such as pricing, marketing, and product features. If the synthetic data fails to reflect the target audience's demographics, preferences, and purchasing behaviors, the insights derived may lead to flawed strategies and failed market performance. Therefore, the accuracy of synthetic datasets is vital for aligning business decisions with market realities.

The effectiveness of synthetic data, particularly in market analysis, also depends on the quality of synthetic data. For example, a financial institution developing a credit scoring model risks inaccurate risk assessments or discriminatory practices if the synthetic data lacks realistic income distributions or demographic diversity. This can result in financial losses or reputational harm. By ensuring that synthetic data maintains realistic patterns, organizations can improve model performance and make more accurate predictions in real-world scenarios.

Ethical concerns also arise with synthetic data, especially in sectors like healthcare and finance. If synthetic data fails to represent diverse demographic groups, it can perpetuate inequalities. In healthcare, for instance, using synthetic data that doesn't reflect varied patient demographics can lead to biased treatment protocols. Prioritizing realism and representativeness in synthetic data not only enhances analytical outcomes but also ensures fairness and equity in decision-making, fostering trust among stakeholders and upholding ethical standards.

Case Studies & Real world implementations

Amazon Alexa: Breaking Language Barriers with Synthetic Data

When Amazon tackled the challenge of expanding Alexa into new languages, they encountered a significant obstacle: the lack of real-world training data for languages yet to be launched. Their innovative solution came in the form of "golden utterances" a synthetic data generation system that created countless realistic voice commands from basic templates. By leveraging their extensive music catalog and developing smart variation algorithms, Amazon generated thousands of training examples that maintained natural language patterns. The approach proved to be successful, enabling Alexa to launch in Hindi, U.S. Spanish, and Brazilian Portuguese with comprehension levels that exceeded expectations, while maintaining user privacy throughout the development process.

Nationwide Building Society: Revolutionizing Partner Innovation

Nationwide Building Society faced a critical challenge in evaluating innovation partners while adhering to GDPR regulations. Unable to share real customer data without compromising privacy, they turned to Accenture and Hazy for a solution. Through advanced synthetic data generation, they created artificial datasets that perfectly mirrored their customer patterns while eliminating privacy risks. The result showed that the partner evaluation timelines shortened dramatically, and innovation discussions became more dynamic and productive. By removing data privacy barriers, Nationwide could focus entirely on identifying and developing groundbreaking solutions with their partners.

Conclusion

Synthetic data has become an essential tool across various industries, offering key benefits like preserving privacy, identifying patterns of behaviour, launching new products based on already available data of competitor products and enabling the creation of large datasets for training machine learning models. Synthetic data allows organizations to improve model accuracy without exposing personal information and can be customized for specific use cases, leading to cost savings and faster development cycles.

Despite its benefits, synthetic data has its challenges, particularly in ensuring realism and accuracy. Poorly designed methods can produce datasets that fail to reflect real-world patterns, resulting in unreliable outcomes. Validating synthetic data is another concern, as inadequate checks can allow biases or inaccuracies to persist. Over time, over-reliance on synthetic data may cause models to drift away from real-world trends, reducing their effectiveness. Additionally, if biases exist in the original data used to generate synthetic datasets, these biases can carry over, leading to skewed or unfair outcomes. To achieve accurate and reliable insights, synthetic data should complement real data, ensuring a balanced and comprehensive approach to analysis.

In conclusion, synthetic data offers significant advantages in terms of privacy, cost, and flexibility but requires careful handling to address its limitations. By balancing synthetic and real data, organizations can optimize its benefits while mitigating risks, ensuring that data-driven decisions remain reliable and innovative.

Sources

IBM Research. "What Is Synthetic Data?" IBM Research Blog, research.ibm.com/blog/what-is-synthetic-data

Turing. "Synthetic Data Generation Techniques." Turing Knowledge Base, turing.com/kb/synthetic-data-generation-techniques

Hazy. "Accenture Unlocks Customer Project with Synthetic Data." Hazy, hazy.com/resources/accenture-unlocks-customer-project-with-synthetic-data

MOSTLY AI. "What is Synthetic Data?" MOSTLY AI, mostly.ai/what-is-synthetic-data/

Amazon. "Tools for Generating Synthetic Data Helped Bootstrap Alexa's New-Language Releases." Amazon Science, www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexa-s-new-language-releases.

Contact Us

Knometrix provides businesses with global market intelligence & market insights to facilitate intelligent decision making. Our in-depth analysis and unique data-driven insights helps leading organizations in strategic decision-making. We are committed to providing innovative solutions tailored to your business needs, with an emphasis on actionable insights and measurable outcomes.

At Knometrix, we believe in building meaningful partnerships that go beyond traditional client relationships. Whether you are looking for in-depth market analysis, strategic advisory, or custom research solutions, we are here to provide you with the intelligence that powers decision-making and fosters growth. Let's explore how we can collaborate to shape your business's future.

Contact Us Today

We would love to learn more about your challenges and objectives. Our team of expert consultants is ready to offer you personalized advice and develop a tailored research roadmap for your business.

Email: info@knometrix.com

For inquiries or project discussions, drop us a line and we'll connect with you promptly.

Phone: [+91 9742110161](tel:+919742110161)

Prefer to speak with us directly? Call us and we'll arrange a call at your convenience.

Website: www.knometrix.com

Explore our case studies, whitepapers, and client success stories. Learn more about how we help businesses succeed.

Visit us in person or schedule a meeting at one of our global offices.

Global Headquarters

USA
102, 2nd Floor, 447 Broadway,
New York, NY, US, 10013

For all ASEAN Region consulting

Bangalore
91springboard, 13, 80 Feet Rd,
Indiranagar, Bengaluru 560038,
India.

For all MENA Region consulting (Partner office)

Muscat
Office #109, Oman Oil,
South Al Manooma,
Muscat, Oman.

Knometrix refers to the global organization, and may refer to one or more, of the member firms of parent company, each of which is a separate legal entity.

Copyright © 2025 by Knometrix. All rights reserved.