TRENDS REPORT

Synthetic Data Delivers Real Value To Fuel The Future Of AI

Explore The Weird, Wonderful, And As-Yet Unseen Use Case Landscape

September 7, 2022

By Rowan Curran, Jeremy Vale with Srividya Sridharan, Mike Gualtieri, Brandon Purcell, Michele Goetz, Stephanie Liu, Enza Iannopollo, Natalie Schibell, Alex Shlyankevich, Kyle Rybarczyk, Karsten Monteverde

FORRESTER®

Summary

AI applications demand an ever-growing breadth and depth of data to power the models behind intelligence — and enterprises are struggling to get enough of the right data to train them. Technology executives and AI team leaders must understand synthetic data so that they can support their teams with the data they need. Synthetic data for AI duplicates, mimics, or extrapolates real-world data — encompassing everything from transfigured spreadsheets to full-blown simulated worlds. This data allows teams to train models for use cases and circumstances that would otherwise be unattainable due to poor quality data, privacy concerns, security restrictions, or lack of existing data.

There Will Never Be Enough Good, Clean Data

AI applications are critical to becoming an insights-driven business — and datasets that are accurate, complete, and large enough for training models are essential for any AI aspirant. Forrester's 2021 data shows that 21% of data and analytics decision-makers cite a lack of well-curated training data as among the biggest challenges for their AI initiatives. Data quality tools prepare data for use and provide capabilities for labelling and some quality correction, but they cannot fully solve the issue. Getting the needed data for AI use cases is challenging because:

- **Sourcing training data is resource intensive.** ML models can't survive without good data, and for many use cases, such as models that respond to some type of user interaction or experience, generating training data is resource intensive (e.g., humans must manually generate the data by interacting with the system). This is both a monetary drain and a time drain for valuable team members who must plan, organize, and create this data. For ML models that require labelled data, the problem gets worse as humans must manually label vast and complex datasets. This is expensive, slow, and prone to errors, requiring QA controls on the labelling process. Some complex use cases, such as Amazon Go's cashier-less retail experiences or facial expression tracking require precise, consistent labeling beyond what human annotators can easily achieve. And in industries such as healthcare, bad data can change the course of someone's life.
- **Real data introduces personal and organizational risk, and reinforces biases.** As any data professional knows, real-world data is messy, complex, and personal — all issues that can severely hold back ML adoption in the enterprise. Using real data about people for training ML models can introduce risks that are both obvious and obscure: Real datasets require stronger security protocols as well as adherence to regulatory requirements (e.g., the GDPR and the California Consumer Privacy Act). And training on real datasets can accidentally expose user data through membership inference attacks or through inadequate governance. Additionally, modelling on real-world datasets can reinforce existing biases and lead to negative outcomes for customers and enterprises. While synthetic data is not a panacea for data biases, it does enable more biases to be identified and accounted for.
- **Some use cases may be too new or novel.** There are some instances where real data simply does not exist. As the landscape of AI solutions both expands and matures, cutting-edge use cases increasingly demand data to model against

scenarios which do not yet exist. For example, FICO has been using synthetic data in their scenario modelling for years to predict major events that would impact their business and customers. This includes modelling for a global disease outbreak, which made them better prepared than many other enterprises for the COVID-19 pandemic. Synthetaic, in partnership with National Geographic, uses synthetic data to aid in the automatic detection of poachers as well as other seldom photographed threats. Collecting additional real world data in such instances is not only dangerous, but sometimes impossible.

Synthetic Data Bridges The Gap Between Ideas And Insights

Enter synthetic data for AI. This type of synthetic data goes far beyond the data generated for load and performance testing in software development. Synthetically generated data can be anything from images of industrial components for modelling predictive maintenance to patients for modelling in healthcare use cases to humans for facial recognition (see Figure 1). Even enterprises that consider themselves data-rich should examine if and how leveraging synthetic data might benefit their AI projects. In the fall of 2021, Meta acquired synthetic data platform AI.Reverie, signaling that even with their vast stores of user data, synthetic data might have a role to play in their strategy. While the word “synthetic” might conjure up images of cheap acrylics and stuffy polyester sweaters, today’s synthetic data is creating high-fidelity artificial representations of the world. From the seemingly banal to the groundbreaking, the widening landscape of use cases is an increasingly essential tool for data science projects and AI teams. Generated using a variety of different techniques, synthetic data is used in AI to:

Generate training data of any type (structured, transactional, image, audio, etc.) that duplicates, mimics, or extrapolates from the real world but maintains no direct link to it, particularly for scenarios where real-world data is unavailable, unusable, or strictly regulated.

Synthetic Data Can Benefit Data-Hungry AI In Multiple Ways

Data science teams and the AI applications they build rely on a foundation of connected intelligence that provides clean, discoverable datasets for model training, and teams should consider synthetic data an integral component of this data. Synthetic data enables enterprise AI teams to circumvent some of their most fundamental data challenges — and in a way that no other technique can. The use of synthetic data sets

can lead directly to business outcomes that would otherwise be overwhelmingly expensive or simply unachievable. AI teams can use synthetic data to:

- **Amplify (or even replace) existing sources of data.** Synthetic data can support the training of more robust models by amplifying the number of rare events and datapoints. For example, an insurer may have an abundance of legitimate claims to draw from in building a fraud detection model but not have many examples of fraudulent claims and could use a synthetic dataset with additional synthetically generated examples of fraud to boost the signal. Mostly.AI has published work demonstrating that in some circumstances, even where real data is available, an entirely synthetic data set may outperform the real data in terms of accuracy. In another example, Qure.ai uses style transfer techniques to take X-ray scans of healthy individuals and add in lung nodules — an early indicator of lung cancer that is difficult to detect and underrepresented in existing data sets. Their models have a 17% improved sensitivity compared to human radiologists, saving thousands of dollars and leading to quicker diagnosis.
- **Mitigate data privacy concerns.** Both customers and regulators are pressuring enterprises to be more responsible, ethical, and intentional with their use of personal data. This is driving increasing demand for technologies that preserve customer privacy, while also enabling enterprises to still build the intelligent and engaging experiences those same customers demand. Synthetic data is being increasingly employed as a privacy preserving technology to generate datasets that eliminate downstream privacy concerns.
- **Lower data governance challenges.** Data governance processes are critical, but they are not infallible: They can be too rigid and subject to issues of implementation or upkeep. Synthetic data-driven startups like Gretel.ai and Statice aim to capitalize on this by creating statistically equivalent synthetic data sets, which mitigate concerns around sharing and use. Synthetic data can't solve the tug-of-war between regulatory compliance and data accessibility, but it can help. MDClone, a digital health company, is approaching this issue head-on by helping clients generate synthetic datasets based on data from their real patients. The clients are then able to share this data with partners for analysis without externally exposing patient data.

Figure 1

Synthetically Generated Humans From Beyond The Uncanny Valley

Source: This Person Does Not Exist

Source: This Person Does Not Exist

Recognize The Risks And Limitations Of Synthetic Data

While the use of synthetic data for AI presents tremendous opportunities, there are circumstances where data scientists and AI developers will be stymied when using it. Some of these challenges will be overcome by more mature tools and technology, but there are some AI applications where synthetic data falls short or can create negative value, namely:

- **Where data generation/labelling requires subjectivity.** Some ML models that require labelled data need some human judgment to meaningfully apply those labels. For example, a fast-fashion company might want to build an AI application that helps customers identify the most fashionable outfits that they can build with a particular item they chose. For such an application, the input data would need to be labelled with the subjective characteristic "fashionable" or "not fashionable," something that at present can only be effectively done by a human.
- **Where real data is necessary to promote equitable treatment.** In many use cases, synthetic data can be a powerful tool even when real data is extremely limited. For AI projects with a human focus, however, using limited real data can be dangerous and inadvertently lead to discrimination. For example, a small data set consisting of images of a rare medical condition might make accurate predictions for people with similar physical attributes (e.g., race), but the same condition may manifest itself in materially different ways across other sub-populations. With that in mind, it's essential that sufficient real data is first collected across demographic groups before using synthetic data for amplification.

Generate Synthetic Data Based On Your Use Case

The approach to generating a synthetic data set can vary greatly depending upon the needs of the use case, which dictates the types of synthetic data that you need, such as tabular, geospatial, or 3D scenes (see Figure 2). External factors, such as privacy and ethical concerns, regulatory requirements, and business needs (such as data that must be modeled from an application that is not yet in production), will inform the direction that enterprises take to generate synthetic data. Synthetic data providers can help enterprises overcome these issues by providing both technological and practice guidance for using the data. There are three common scenarios for which we see providers helping companies with synthetic data:

- **Protecting customer data by killing the link to — not the value of — real PII.** Enterprises are often between a rock and a hard place when seeking to use

customer data for data science. On the one hand, there is demand to deliver increasingly intelligent, engaging, and profitable applications and products, while on the other is stronger pressure from regulators and the public to protect personal data and privacy. Synthetic data presents one solution by allowing enterprises to use their knowledge of existing customers to create synthetic data that mirrors the original data but preserves no identifiable connection to the original data or to individual customers.

- **Expanding and enhancing existing datasets.** Even if there is data available for training ML models, it may not be of sufficient volume, or there may be critical cohorts that are underrepresented. Synthetic data can be used to broaden datasets that are too small for proper modelling or to increase smaller populations within a dataset to a size where meaningful analysis can be run. For example, Hyperscience is using samples of anonymized and synthesized driver's licenses and passports to create large training datasets for their intelligent document extraction, which means their customers need less customization for use cases involving those types of documents.
- **Programmatically generating net-new data.** Synthetic data can also be generated from scratch for scenarios that do not yet exist or are hard to model in the real world. While this requires a much better understanding of the type of data needed and how to create it, building synthetic data in this way allows the user to fully define the parameters of the dataset. Such data can be especially valuable in circumstances where a product or service needs to be tested against a known population with novel behavior such as driver monitoring systems (DMS) in cars, which regulations are increasingly requiring for car manufacturers. Vendors like Datagen are working with customers to populate a wide range of synthetic dynamic humans demonstrating various behaviors in vehicles to train the computer vision models powering the DMS.

Three Types Of Synthetic Data Providers Can Help You Get Started

The synthetic data market is still relatively small, but it is seeing growth and investment at a prodigious rate. Providers of synthetic data recognize that enterprises are at different points in their data science and AI journeys, and a variety of offerings have bloomed on the market. Today, you can source synthetic data from three types of providers:

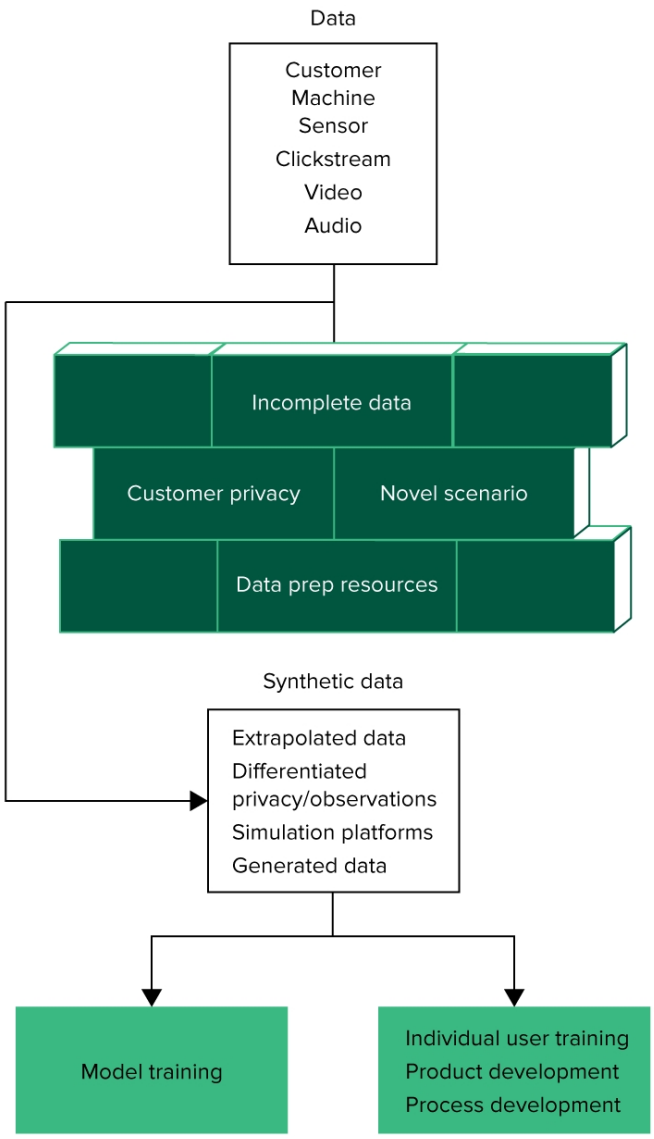
- **Synthetic dataset providers.** Enterprise teams who just want clean, ready-to-go datasets that accurately represent specific customers and scenarios can look to synthetic dataset providers. Some vendors, like Anyline, are working directly with customers to understand their existing data and business needs and delivering

fully modeled datasets directly to them. Anyline worked with a utility company in India to create data sets that combine real-world data with a synthetic dataset to detect fraudulent electric meter readings where customers were evading billing. They needed synthetic data for reliable training examples of both real data that was low quality (e.g., images of legitimate readings taken from an odd angle) and examples of true fraud or potential fraud. This approach removes the need for user expertise in architecting synthetic datasets and can be helpful for teams without their own data engineering capabilities.

- **Synthetic data platform providers.** Synthetic data platforms address the needs of teams who have deeper, more specialized use cases within industries and who want to take a more sophisticated internal approach to synthetic data. Synthetic data platforms currently come in three main flavors: synthetic visual data (e.g., Anyverse, Datagen), synthetic textual data (e.g., Gretel.ai, Dataomize) and synthetic structured data (e.g., MDClone, Mostly.ai). These platforms are used in a range of use cases within their specific domains and may produce some data types, but typically do not address all in a single vendor offering.
- **Open source synthetic data generation options.** If you're not ready to jump in with a commercial vendor, or you think that you've got the chops on your team to create your own data, there are open source tools for generating synthetic data. The open source landscape for generating synthetic data ranges from tools for generating synthetic financial transactions to synthetic images and video. Enterprises are finding success using open source tools like these, provided their data scientists and developers have a good understanding of ML data quality practices (and feedback from subject-matter experts in the business) — no specialized knowledge is required to get started creating synthetic datasets.

Figure 2

Avoid Hitting A Wall In ML With Synthetic Data



Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Supplemental Material

Companies We Interviewed For This Report

We would like to thank the individuals from the following companies who generously gave their time during the research for this report.

Datagen

FICO

Hyperscience

MDClone

Mirage

Skyengine.ai

Tonic.ai

FORRESTER®

We help business and technology leaders use customer obsession to accelerate growth.

FORRESTER.COM

Obsessed With Customer Obsession

At Forrester, customer obsession is at the core of everything we do. We're on your side and by your side to help you become more customer obsessed.

Research

Accelerate your impact on the market with a proven path to growth.

- Customer and market dynamics
- Curated tools and frameworks
- Objective advice
- Hands-on guidance

Learn more.

Consulting

Implement modern strategies that align and empower teams.

- In-depth strategic projects
- Webinars, speeches, and workshops
- Custom content

Learn more.

Events

Develop fresh perspectives, draw inspiration from leaders, and network with peers.

- Thought leadership, frameworks, and models
- One-on-ones with peers and analysts
- In-person and virtual experiences

Learn more.

FOLLOW FORRESTER

Contact Us

Contact Forrester at www.forrester.com/contactus. For information on hard-copy or electronic reprints, please contact your Account Team or reprints@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
Tel: +1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com