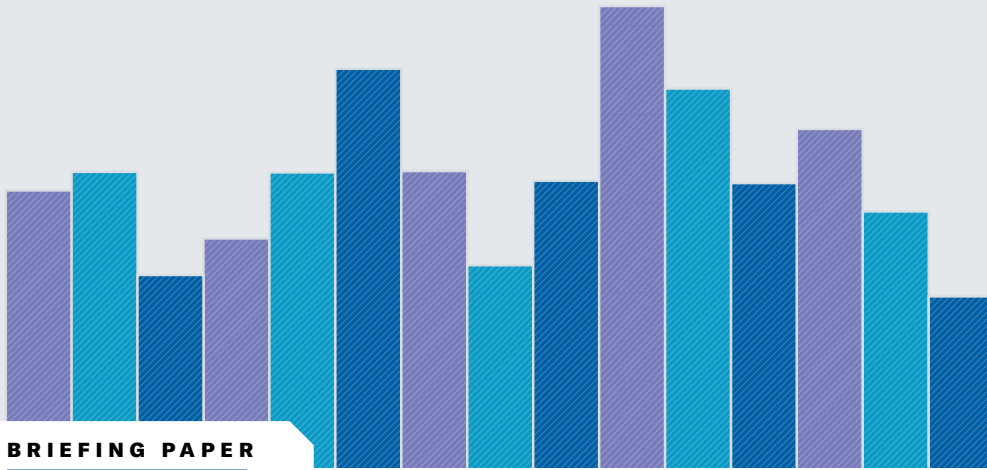




**Harvard
Business
Review**

ANALYTIC SERVICES



BRIEFING PAPER

The Executive's Guide to Accelerating Artificial Intelligence and Data Innovation with Synthetic Data



Sponsored by

MOSTLY·AI

SPONSOR PERSPECTIVE

Organizations today face a challenging environment where it's easy to fall behind. Many enterprises are looking for ways to gain a competitive advantage via technology, synthetic data being one of the hot topics on the table. We sponsored this research by Harvard Business Review Analytic Services to shed light on the emerging synthetic data use cases and the technology's many advantages for businesses worldwide. Erste Group—one of the largest banks in Europe—and others are leveraging synthetic data to increase efficiency and competitiveness, drive digital transformation, offer personalization, accelerate cloud migration, create competitive pricing models, add new revenue streams, and ultimately, to save millions of dollars. The people and the organizations featured in this research are mature synthetic data practitioners, presenting best practices ready to be picked up and success stories ready to be replicated.

Many enterprises choose MOSTLY AI as their trusted partner to make the most of their synthetic data opportunities. These companies, typically coming from the banking, insurance, and telecommunications spaces, are building private-by-design synthetic data pipelines fueling artificial intelligence (AI), advanced analytics, and testing applications across their organizations. Gartner recommends integrating this technology into your stack to accelerate the analytics development cycle, lessen regulatory concerns, and lower the cost of data acquisition.¹ These forward-thinking enterprises are leveraging all of these synthetic data advantages already.

The beauty of synthetic data is that it is highly flexible; you can create, share, and discard this data at will. It's as good as your production data, yet it is exempt from the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act. It's capable of improving data quality for AI and can be used to modify existing datasets, e.g., to correct for present biases. With the upcoming AI regulations, we predict that synthetic data will be a mission-critical part of AI governance, providing fairness to AI systems and explainability to regulators.

MOSTLY AI is one of the early synthetic data pioneers, with years of experience in this emerging technology. We know how to make synthetic data work, especially in the supersensitive industries of finance and insurance. We support our clients in building synthetic data into their existing environment. The MOSTLY AI team plays an important and defining role in the research and development scene globally, participating in the important work of setting up synthetic data standards and ethical AI best practices.

The synthetic data opportunity means something different for every organization, but new revenue streams; faster, easier, GDPR-compliant data access; better pricing models; and scalable, ethical, and explainable AI all are within reach for those business leaders ready to remove their data privacy blind spots.



Tobias Hann
CEO
MOSTLY AI

¹ Saul Judah, Andrew White, Svetlana Sicular et al., "Predicts 2021: Data and Analytics Strategies to Govern, Scale and Transform Digital Business," Gartner, December 2, 2020. <https://www.gartner.com/en/documents/3993855/predicts-2021-data-and-analytics-strategies-to-govern-sc>.

The Executive's Guide to Accelerating Artificial Intelligence and Data Innovation with Synthetic Data

Senior leadership finds itself grappling with the competing demands of operating in an increasingly digital world. This digital environment requires use of sophisticated technologies like artificial intelligence (AI), machine learning (ML), and data analytics, as well as using data for product development and testing, and sharing it with external partners. Companies need to better utilize data to innovate and deliver for more customer-centric products and services in order to stay competitive. But they also must comply with a growing number of regulations restricting data use in order to protect privacy, such as the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA). Enterprises need an effective solution to bridge this divide.

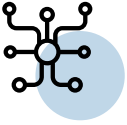
Synthetic data is a tool that addresses many data challenges, particularly AI and analytics issues such as privacy protection, regulatory compliance, accessibility, data scarcity, and bias as well as data sharing and time to data (and therefore time to market). It can speed up the analytics development cycle, lower the cost of data acquisition, bridge information silos, and support data monetization.¹ Generated by AI to simulate the characteristics and behavior of real data, synthetic data serves as a proxy for it in order to not reveal sensitive information.²

HIGHLIGHTS

Synthetic data is a tool that addresses many data challenges, particularly artificial intelligence and analytics issues such as privacy protection, regulatory compliance, accessibility, data scarcity, and bias, as well as data sharing and time to data (and therefore time to market).

With recognition of synthetic data's ability **to support digital transformation, improve efficiency, speed innovation**, and even help talent recruitment, more industries will see the business benefit of utilizing it.

As this is a relatively new field, lack of awareness, education, and accessible, nontechnical research highlighting best practices are major challenges to greater adoption of synthetic data, as are organizational issues.



Companies need to better utilize data to innovate and deliver for more customer-centric products and services in order to stay competitive.

“Getting access to data at scale, especially when data comes from multiple silos, is always a challenge,” says Chalapathy Neti, head of the AI/ML platform at SWIFT, the international payments organization. “One way to address that is to use simulations, to avoid constraints on the amount of historical data that we can access, and at the same time, interpolate missing data.”

Synthetic data is still relatively new—it was first created at the Massachusetts Institute of Technology in 2013.³ Therefore, its use across industries and the types of data it is being used for are uneven. However, by 2024, Gartner predicts that 60% of the data used for the development of AI and analytics solutions will be synthetically generated, and synthetic data generation capability will be expected by then in ML and analytics platforms.⁴

Its use is most advanced in industries such as banking, insurance, health care, and telecommunications as well as parts of the public sector, where customer data is very sensitive and subject to strict regulation. The top use cases are AI training, product development and testing, and third-party data sharing, such as with startups, to validate vendors, or even within an organization when sharing is restricted by data privacy rules.

Assisting AI Training

Synthetic data is vital when the data needed to train AI models must be guaranteed to be anonymous, such as with medical data. Slawek Kierner, senior vice president for enterprise data and analytics at Humana, says, “Humana has roughly 17 million members, and there’s a treasure trove of information about your health care. There is plenty of data to mine in medical records; there are insurance claims and increasingly all kinds of connected medical devices that give us more data even after you leave a doctor’s visit. So, we have a lot of sensitive data on people’s health, including our own employees’—the scope of this data set is huge—and we need to make sure this information is protected.”

Using synthetic data to train AI is transforming the way Humana provides care. “With AI, we can predict what will happen with our members’ health,” Kierner says. “We created accurate models that predict the progression of a disease to the degree that we can predict when you will need an ER, so that we can be two steps ahead and help our members prevent emergencies. It’s totally changing health care—in the way that other industries have changed—from reactive to proactive.”

Synthetic data can be used to train AI models for scenarios for which limited data is available—such as fraud cases. This is a use case Maurizio Poletto, chief platform officer at Erste Group Bank AG, is considering. “We could take a fraud case using synthetic data to exaggerate the cluster, exaggerate the amount of people, and so on, so the model can be trained with much more accuracy,” he says. “The more cases you have, the more detailed the model can be.”

It can also help when past data is no longer a reliable guide to the future, resulting in a significant decay in AI model accuracy. “Until recently, AI and model training were based on past data,” says Svetlana Sicular, a research vice president at Gartner. “But as we saw during Covid-19, a lot of models stopped working. Past experiences and past examples became invalid under the current circumstances. And that can happen anytime. Synthetic data, on a philosophical level, relieves AI from the limitations of looking only at the past and learning from the past data. With synthetic data, you can dream up the future, create the data that you think might come in the future, and create the models to deal with the future.”

SWIFT is in the early days of using synthetic data, starting with “transaction intelligence, so gaining insights that are valuable to our entire customer base” has been the organization’s first priority, says Neti, head of AI/ML. He sees other areas—enterprise and customer intelligence—where he plans to apply synthetic data beyond transaction intelligence.

When it comes to operational intelligence, he explains, we’re “looking at the infrastructure that underpins this financial network and being able to predict demand so we can optimize services efficiency for our customers.”

Because of the wide spectrum of use cases across the organization where Neti believes synthetic data will be beneficial, there may be more opportunity to utilize it for machine learning purposes. In fact, Gartner predicts that by 2024, the use of synthetic data will halve the volume of real data needed for ML—accelerating data-driven innovation.⁵

Accelerating Product Development and Testing

The main use case for Erste Group is creating synthetic segments and communities. The banking group is trying to use synthetic data to build new features and test how certain

“Leveraging simulated data is one of my biggest drivers of using synthetic data because it enables engagement with external technology partners and leading labs around the world to create state-of-the-art solutions for our community.”
—Chalapathy Neti, head of the AI/ML platform, SWIFT

types of customers would react to these features. “Normally, the data we use is static. We see everything from the past,” says Erste Group’s Poletto. “But for features like notifications and triggers—like receiving a notification when your salary comes in—they can only be tested with dynamic test users. With synthetic data, you push a button to generate that user with an unlimited number of transactions in the past and a limited number of transactions in the future, and then you can put into your system a user which is alive.”

He has big ambitions for how synthetic data will support product development and testing. “I hope we’re going one day to be able to create synthetic environments,” Poletto asserts. “Imagine a synthetic replica of your bank, same location, same type of customer, and then you can put in a new product and see how the environment reacts, which customers, with which characteristics, will be attracted to that product.”

For Humana, synthetic data has accelerated the process of testing. “Before we had synthetic data, we had to wait for months to validate security and compliance of new technologies to allow us to use real data, which slowed down innovation,” says Humana’s Kierner. “Or we manually created test data sets, but that is less useful, less representative of real data sets, and rarely allows us to validate the complete experience of new technologies at scale. Synthetic data has also been helpful to double-check what has been done with our internal system testing.”

The value of using synthetic data for product testing was highlighted by the fine given to the Norwegian Confederation

of Sport by the Norwegian Data Protection Authority (NDPA) for a GDPR violation. When testing solutions for moving a database from a physical server environment to the cloud, the sports organization shared the personal data of about 3.2 million Norwegians online in error. The NDPA stated that the situation could have been avoided by using synthetic data.⁶

Enabling Faster, Privacy-Compliant Data Sharing

Synthetic data facilitates the sharing of data within enterprises—helping support innovation, as seen at Humana—but also helps them share data outside the organization. With the GDPR in the European Union, the CCPA, and other emerging privacy regulations around the world making the sharing of personal information so complicated, if not impossible, synthetic data is vital to support collaboration. As it is fully anonymous, it is exempt from these rules. It is particularly necessary after the recent European Court of Justice decision, known as Schrems II, invalidated the EU-US Privacy Shield Framework—making the majority of EU-US data transfers illegal.

The ability to share data has huge benefits. According to Gartner, organizations that share data externally with their partners generate three times more measurable economic benefits than those that do not share.⁷

“Leveraging simulated data is one of my biggest drivers of using synthetic data because it enables engagement with

How to Start Utilizing Synthetic Data

For those who are new to synthetic data but keen to understand how it can benefit their organizations, here are some guidelines for how to get started:

- Identify use cases for synthetic data where there is a lack of data, data cannot be used because of privacy issues, or where real-world data would benefit from being supplemented.
- Look at situations where data sharing, either between different departments within the organization or with a third party, was not possible due to internal governance or privacy regulations.
- Work with a trusted synthetic data expert to identify, examine, and quantify the advantages and limitations of synthetic data.

Start with tabular data, as it is the best understood, contains sensitive information that is of great value, and is generally the most regulated data:

- Leverage the expertise of data scientists to ensure the statistical validity of the sample and distribution of the synthetic data.
- Work with data security and legal teams to understand the potential for synthetic data usage and to create guidelines for its use.
- Monitor successes and failures at generating and using synthetic data; then adjust your data strategy and guidance as your experience evolves.

Svetlana Sicular, research vice president at Gartner, emphasizes the importance of establishing governance early on. “Start with figuring out overall how the process looks, with establishing your governance,” she says. “Governance doesn’t mean command and control governance; it means practical decision making and standards about what tools to use and how to collaborate over this data. If you don’t have those standards early on, everybody will invent their own, and when it’s time to scale, you’ll find that there are 10 different approaches across the company.”

external technology partners and leading labs around the world to create state-of-the-art solutions for our community,” according to Neti. “My aspiration is to use SWIFT as a membership-driven entity for banks to create a common dataset to unleash innovation.”

Humana also sees the advantage in spurring innovation through sharing. “We want to be one of the friendliest companies in terms of innovation, and that means working with startups that want access to our data,” says Kierner. “But the legal and compliance process to allow access to our longitudinal health records or the feature store of our machine learning platform was legally complex, lengthy, and costly. So, we set up Humana Data Exchange, where we preloaded synthetic data, and then access became very simple. That has accelerated innovation with our partners.”

The Business Impact of Synthetic Data

For those who are using synthetic data, the business benefits are clear, and they have built up trust through the transition. But as it is a relatively new field, lack of awareness, education, and accessible, nontechnical research highlighting best practices are major challenges to greater adoption of synthetic data, as are organizational issues.

“Trust is an internal challenge,” says Poletto. “The question is, do we trust the synthetic data as an accurate data source? I was very skeptical at the beginning because it told me something counter to what I believed about our customer base. But comparing it with real customers’ data, I learned its accuracy; it was right, and so I learned to trust it. But it’s a process because it’s a rather new thing.”

For him, trust goes hand in hand with organizational issues. “To make it really work, you need to position this new technology within the existing workflow, you need to make it available to all the engineers, you need to make it part of the way you build stuff,” he says. “Building pipelines is really complicated. It is done over years, and there are a lot of stakeholders involved, a lot of dependencies. You need to get people using it so they begin to trust it, and then further integrate it in the workflow.”

Some executives, of course, still question whether AI-generated synthetic data can be trusted to be completely anonymous. Just generating synthetic data does not guarantee total anonymity—it is the combination of synthetization with various privacy mechanisms that leads to perfectly anonymous data. “It’s always a conversation, and it requires governance, deliberation, and questioning,” according to Gartner’s Sicular. “It requires a conscious approach to how you use synthetic data. For instance, what do you do about outliers? If you are creating data for a rural area and it’s one person per 100 miles, even though I can create a synthetic

Just generating synthetic data does not guarantee total anonymity—it is the combination of synthetization with various privacy mechanisms that leads to perfectly anonymous data.

person, it doesn't hide anything. Synthetic data on its own is not a data strategy, but it is an excellent new tool that is extremely promising."

Synthetic data is considered better than traditional anonymization methods, which destroy the utility of the data while still allowing it to be easily reidentified. "In theory, in banking, you could take real account data, scramble it, and then put it into your system with real numbers so it's not traceable," says Poletto. "The problem is that obfuscation is nice, and anonymization is nice, but you can always find a way to get the original data back. As a bank, we need to be thorough and cautious because it is sensitive data. Synthetic data is a good way to continue to create value and experiment without having to worry about privacy, particularly because society is moving toward better privacy. This is just the beginning, but the direction is clear."

But privacy is just one issue when it comes to data and its role in the digital transformation of organizations. The ongoing process of digital transformation has been on top of CEOs' agendas for years, and proper use of data for analytics and AI is key to truly innovating the organization; however, many initiatives are hindered by siloed data, governance that makes sharing difficult, and data scarcity, all of which slow innovation.

"Two years ago, when I joined Humana, most of our infrastructure was fragmented," says Kierner. "Data was sitting in many older systems. We didn't have access to a modern cloud infrastructure to unlock the potential of

analytics on our combined data set to be able to mine this data, and there was a lot of concern about changing because of data security and privacy concerns. Synthetic data was the key to accelerating our own adoption of a public cloud."

For Erste Group, synthetic data has enhanced its competitiveness by increasing efficiency. "In a strictly regulated environment like ours, when you want to do a project with data, a large amount of time is dedicated not to work on the data but to get permission to use the data, making sure all safety and privacy measures are properly in place," says Poletto. "I'm not overestimating when I say this is around 50% of the effort. But with synthetic data, that aspect is completely gone, and you can dedicate 100% of your time to the project itself. We're talking about thousands of hours. We don't measure the ROI of synthetic data, but my gut feeling is that, on a project, it can have an impact of 30% to 50% improvement on the process."

That improved efficiency is helping accelerate innovation. "The fact that we could create our machine learning platform three months earlier because we could preload it with synthetic data and start testing straight away—that is worth tens of millions of dollars in terms of savings and hundreds [of millions of dollars], if not more, when you think about the ecosystem that will be built and its impact in the future," Kierner explains.

Increasing the ability to innovate also helps talent recruitment. "Synthetic data is driving innovation of our data science team by helping them move on to new projects,"



“Talented data engineers want to spend 100% of their time in data exploration and value creation from data. They don’t want to spend 50% of their time on bureaucracy. If we can eliminate that, we are better able to attract talent,” says Maurizio Poletto, chief platform officer at Erste Group Bank AG.

he continues. “We now have a number of other deep learning models, because the team members who have been working closely on this synthetic data generation project have learned and mastered new technology. This is how we unlock further innovation, and it helps us to attract top talent to join and stay at Humana. Top talent values vast yet rich data sets, cloud native technology, [and] sophisticated methods as well as our mission to ethically use AI to improve health care and save lives. That people impact is strategically important even if not directly translated into ROI.”

Poletto agrees. “Talented data engineers want to spend 100% of their time in data exploration and value creation from data,” he says. “They don’t want to spend 50% of their time on bureaucracy. If we can eliminate that, we are better able to attract talent. At the moment, we may lose some or they are not even coming to the banking industry because they know it’s a super-regulated industry and they won’t have the same freedom they would have in a different industry.”

Conclusion


With recognition of synthetic data’s ability to support digital transformation, improve efficiency, speed innovation, and even help talent recruitment, more industries will see the business benefit of utilizing it. And as understanding of synthetic data by senior leadership grows, there will be new demand beyond the current main use cases of assisting AI training. This growing understanding will manifest itself through accelerated product development and testing and faster, privacy-compliant data sharing to new use cases such as audited AI models and mitigated bias in AI.

“You can synthesize less-biased data to train unbiased or minimally biased models, such as if you have an uneven distribution for men and women,” Sicular says. “Or, if you take phone data, it’s mostly collected from the younger generation, so if you apply it to seniors, you put them at a disadvantage, whether it’s with typing speed or even how they speak. It’s not only bias in people—gender, race, and so on—but business biases, such as ‘we will pay more attention to this region versus that region because we have more records from this region.’ You can generate non-biased data sets, and

you can generate more biased data sets to validate the model and make sure that it’s unbiased.”

Poletto sees potential in cross-industry and private- and public-sector collaboration with synthetic data. “Imagine if we in banking use synthetic data to generate realistic and comparable data from our customers, and the same thing is done by the transportation industry, the city, the insurance company, and the pharmaceutical company, and then you give all these data to someone to analyze the correlation between them,” he says. “Because the relationship between well-being, psychological health, and financial health is so strong, I think there is a fantastic opportunity around the combination of mobility, health, and finance data.”

The European Commission’s proposed AI regulation, released in April 2021, will also drive greater demand for synthetic data. The regulation’s definition of AI includes software utilizing machine learning, rules-based AI approaches, and traditional statistical techniques used in creating models. Its remit covers providers of AI systems in the EU regardless of where the provider is located, users of AI systems within the EU, and providers and users based outside the EU “where the output produced by the system is used in the Union”—meaning anyone, anywhere who processes data about EU citizens.⁸ The compliance with rules, documentation, and transparency required by the proposed regulation will add additional complexity to using real data—making synthetic data even more attractive.



With recognition of synthetic data's ability to support digital transformation, improve efficiency, speed innovation, and even help talent recruitment, more industries will see the business benefit of utilizing it.

Endnotes

- 1 Saul Judah, Andrew White, Svetlana Sicular et al., "Predicts 2021: Data and Analytics Strategies to Govern, Scale and Transform Digital Business," Gartner, December 2, 2020. <https://www.gartner.com/en/documents/3993855/predicts-2021-data-and-analytics-strategies-to-govern-sc>.
- 2 Jim Hare, Svetlana Sicular, and Erick Brethenoux, "Tech Providers 2025: Why Small Data Is the Future of AI," Gartner, September 17, 2020. <https://www.gartner.com/en/documents/3990260/tech-providers-2025-why-small-data-is-the-future-of-ai>.
- 3 Laboratory for Information and Decision Systems, "The Real Promise of Synthetic Data," MIT News, October 16, 2020. <https://news.mit.edu/2020/real-promise-synthetic-data-1016>.
- 4 Judah, White, Sicular et al., "Predicts 2021: Data and Analytics Strategies," Gartner.
- 5 Farhan Choudhary, Afraz Jaffri, Svetlana Sicular et al., "Predicts 2021: Artificial Intelligence Core Technologies," Gartner, December 22, 2020. <https://www.gartner.com/en/documents/3994810/predicts-2021-artificial-intelligence-core-technologies>.
- 6 The European Data Protection Board, "Norwegian DPA: Norwegian Confederation of Sport Fined for Inadequate Testing," June 15, 2021. https://edpb.europa.eu/news/national-news/2021/norwegian-dpa-norwegian-confederation-sport-fined-inadequate-testing_en.
- 7 Judah, White, Sicular et al., "Predicts 2021: Data and Analytics Strategies," Gartner.
- 8 Mark MacCarthy and Kenneth Propp, "Machines Learn that Brussels Writes the Rules: The EU's New AI Regulation," Brookings Institute, May 4, 2021. <https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/>.



Harvard Business Review

ANALYTIC SERVICES

ABOUT US

Harvard Business Review Analytic Services is an independent commercial research unit within Harvard Business Review Group, conducting research and comparative analysis on important management challenges and emerging business opportunities. Seeking to provide business intelligence and peer-group insight, each report is published based on the findings of original quantitative and/or qualitative research and analysis. Quantitative surveys are conducted with the HBR Advisory Council, HBR's global research panel, and qualitative research is conducted with senior business executives and subject matter experts from within and beyond the *Harvard Business Review* author community. Email us at hbranalyticsservices@hbr.org.

hbr.org/hbr-analytic-services