Easy guide

MOSTLY·AI

Banking on synthetic data

A deep-dive into synthetic data use cases for the banking industry and all the intel on how to get started on going synthetic.

Synthetic financial data is the fuel banks need to become AI-first and to create cutting-edge services. In this ebook, you can read about:

Banking technology trends in 2022 from superapps to personalized digital banking.

Data privacy legislations affecting the banking industry in 2022.

The most valuable data science and synthetic data use cases in banking: customer acquisition and advanced analytics, mortgage analytics, credit decisioning and limit assessment, risk management and pricing, fraud and anomaly detection, cybersecurity, monitoring and collections, churn reduction, servicing and engagement, enterprise data sharing, synthetic test data for digital banking product development.

Synthetic data engineering: how to integrate synthetic data in financial data architectures.

The challenges in AI/ML development, testing and data sharing that synthetic data can solve.

Smarter Synthetic Data

Table of contents

Banks and financial institutions are aware of their data and innovation gaps and AI-generated synthetic data is their best bet. According to Gartner:

By 2030, 80 percent of heritage financial services firms will go out of business, become commoditized or exist only formally but not competing effectively.

A pretty dire prophecy, but nonetheless realistic, with small neobanks and big tech companies eyeing their market. There is nowhere to run but forward.

The future of banking is all about becoming AI-first and creating cutting-edge digital services coupled with tight cybersecurity. In the race to a tech-forward future, most consultants and business prophets forget about step zero: customer data. In this ebook, we will give an overview of the data science use cases in banking and attempt to offer solutions throughout the data lifecycle.

We'll concentrate on the easiest to deploy and highest value synthetic data use cases in banking. We'll cover three clusters of synthetic data use cases: data sharing, AI, advanced analytics, machine learning, and software testing. But before we dive into the details, let's talk about the banking trends of today.

Banking technology trends

The pandemic accelerated digital transformation, and the new normal is here to stay. According to Deloitte, 44% of retail banking customers use their bank's mobile app more often. At Nubank, a Brazilian digital bank, the number of accounts rose by 50%, going up to 30 million. It is no longer the high-street branch that will decide the customer experience. Apps become the new high-touch, flagship branches of banks where the stakes are extremely high. If the app works seamlessly and offers personalized banking, customer lifetime value increases. If the app has bugs, frustration drives customers away. Service design is an excellent framework for creating distinctive personalized digital banking experiences. Designing the data is where it should all start.

A high-quality synthetic data generator is one mission-critical piece of the data design tech stack. Initially a privacy-enhancing technology, synthetic data generators can generate representative copies of datasets. Statistically the same, yet none of the synthetic data points match the original. Beyond privacy, synthetic data generators are also fantastic data augmentation tools. Synthetic data is the modeling clay that makes this data design process possible. Think moldable test data and training data for machine learning models based on real production data.

The rise of superapps is another major trend financial institutions should watch out for. Building or joining such ecosystems makes absolute sense if banks think of them as data sources.
Data ecosystems are also potential spaces for customer acquisition. With tech giants entering the market with payment and retail banking products, data protectionism is rising. However, locking up data assets is counterproductive, limiting collaboration and innovation. Sharing data is the only way to unlock new insights.

Especially for banks, whose presence in their customers' lives is not easy to scale unless via collaborations and new generation digital services. Insurance providers and telecommunications companies are the first obvious candidates.

Other beyond-banking service providers could also be great partners, from car rental companies to real estate services, legal support, and utility providers. Imagine a mortgage product that comes with a full suite of services needed throughout a property purchase. Banks need to create a frictionless, hyper-personalized customer experience to harness all the data that comes with it.

Another vital part of this digital transformation story is AI adoption. In banking, it's already happening. According to McKinsey,

> **"The most commonly used AI technologies (in banking) are:**
> - robotic process automation (36%) for structured operational tasks
> - virtual assistants or conversational interfaces (32%) for customer service divisions
> - machine learning techniques (25%) to detect fraud and support underwriting and risk management."

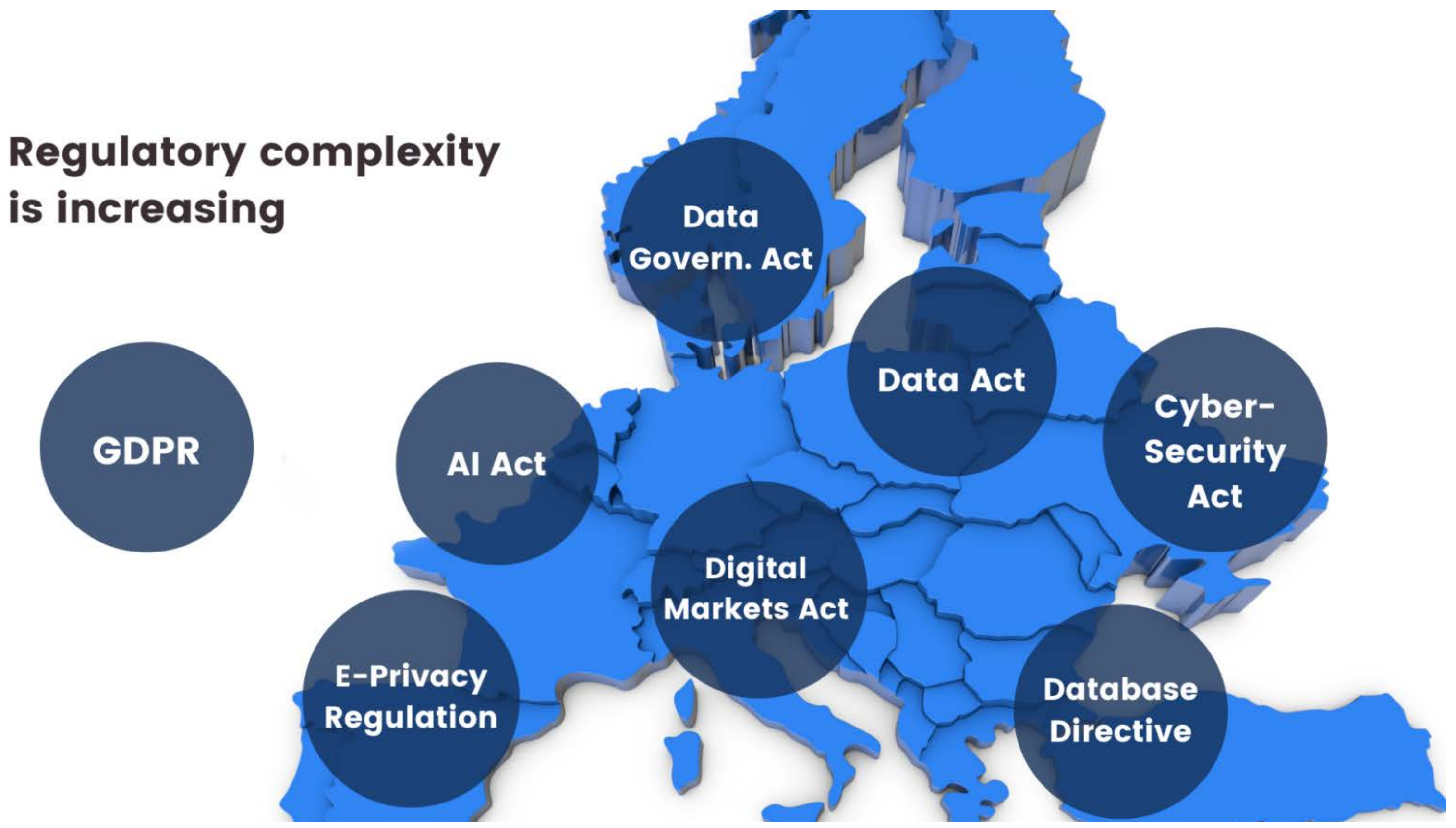
It sounds like banks are running full speed ahead into an AI future, but the reality is more complicated than that. Due to the legacy infrastructures of financial institutions, the challenges are numerous. Usually, there is no clear strategy or fragmented ones with no enterprise-wide scale.

> "McKinsey estimates that AI technologies could potentially deliver up to $1 trillion of additional value in banking each year."

Different business units operate almost completely cut off with limited collaboration and practically no data sharing. These fragmented data assets are the single biggest obstacle to AI adoption. McKinsey estimates that AI technologies could potentially deliver up to $1 trillion of additional value in banking each year. It is well worth the effort to unlock the data AI and machine learning models so desperately need. Let's take a look at the number one reason or rather excuse banks and financial institutions hide behind when it comes to AI/AA/ML innovation: data privacy.

The state of data privacy in banking in 2022

The pandemic accelerated digital transformation, and the new banks have always been the trustees of customer privacy. Keeping data and insights tightly secured has prevented banks from becoming data-centric institutions. What's more, an increasingly complex and restrictive legislative landscape makes it difficult to comply globally.



The European data privacy landscape in 2022

Let's be clear. The ambition to secure customer data is the right one. Banks must take security seriously, especially in an increasingly volatile cybersecurity environment. However, this cannot take place at the expense of innovation. The good news is that there are tools to help. Privacy-enhancing technologies (PETs) are crucial ingredients of a tech-forward banking capability stack.

It's high time for banking executives, CIOs, and CDOs to get rid of their digital banking blindspots. Banks must stop using legacy data anonymization techniques that endanger privacy and hinder innovation. Data anonymization methods, like randomization, permutation, generalization, and pseudonymization, carry a high risk of re-identification or destroy data utility.

Maurizio Poletto, Chief Platform Officer at Erste Group Bank AG, said in The Executive's Guide to Accelerating Artificial Intelligence and Data Innovation with Synthetic Data:

> "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. The problem is that obfuscation is nice, and anonymization is nice, but you can always find a way to get the original data back. We need to be thorough and cautious as a bank 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."

Modern PETs include AI-generated synthetic data, homomorphic encryption, or federated learning. They offer the way out of the data dilemma in banking. Data innovators in banking should choose the appropriate PET for the appropriate use case. Encryption solutions should be looked at when necessary to unencrypt the original data.

Anonymized computation, such as federated learning, is a great choice when models can get trained on users' mobile phones. AI-generated synthetic data is the most versatile privacy-enhancing technology with just one limitation. Synthetic datasets generated by AI models trained on original data cannot be reverted back to the original. Synthetic datasets are statistically identical to the original datasets they were modeled on.

However, there is no 1:1 relationship between the original and the synthetic data points. This is the very definition of privacy. As a result, AI-generated synthetic data is great for specific use cases—advanced analytics, AI and machine learning training, software testing, and sharing realistic but unencryptable datasets.

Synthetic data is not a good choice for use cases where the data needs to be reverted back to the original, such as information sharing for anti-money laundering purposes, where perpetrators need to be re-identified. In such cases, homomorphic encryption would be a safe bet. Let's see a comprehensive overview of the most valuable synthetic data use cases in banking!

The most valuable synthetic data use cases in banking

Synthetic data generators come in many shapes and forms. In the following, we will be referring to MOSTLY AI's synthetic data generator. It is the market-leading synthetic data solution able to generate synthetic data with high accuracy. MOSTLY AI's synthetic data platform comes with advanced features, such as direct database connection and the ability to synthesize complex data structures with referential integrity. As a result, MOSTLY AI can serve the broadest range of use cases with suitably generated synthetic data. In the following, we will detail the lowest hanging synthetic data fruit in banking. These are the use cases we have seen working well in practice and generating a high ROI.

Challenges

AI/AA/ML

- Data locked away for privacy reasons
- Training data quality is suboptimal due to legacy anonymization
- Training data is erroneous due to embedded historic bias
- Model performance is not good enough to be put into production
- Domain knowledge is missing due to restricted data
- Modern cybersecurity approaches rely heavily on high-performance anomaly detection models
- Mortgage analytics models miss out on next-generation data assets, such as transaction data and location data

How can synthetic data help?

- Synthetic data can be used freely
- Synthetic training data is as good as real with up to 99% accuracy
- Synthetic data generation can fix biases by upsampling minority groups
- Upsampling via synthesization can improve ML performance by 15%
- Synthetic data can be injected into models and linked to other datasets
- Synthetic data replacement improves model performance and limits the impact of intrusions
- Synthetic geolocation data from mobile service providers and accurate synthetic behavioral data of transactions are compliant and accurate

TESTING

- Production data is off-limits for privacy reasons
- Manual data generation misses business rules
- Data can't be shared with third-party test teams or other lines of business
- Fragmented data for testing omnichannel journeys
- Test environments are slow to build (40+ days)
- Complex database structures are impossible to recreate with referential integrity

- Synthetic data can simulate production data accurately
- Synthetic data implicitly picks up on all business rules
- Synthetic data is free to share even across borders and cross-teams tests
- Synthetic data can be shared to create omnichannel testing stories
- Synthetic test data generation is fast and on-demand, shortening sprints
- Advanced synthetic data generators synthesize entire database structure with referential integrity

Challenges

DATA SHARING

- Regulations prohibit cross-border data sharing
- Vendor selection is suboptimal due to lack of bank-specific test data
- Distinct business lines operate siloed data reserves

How can synthetic data help?

- Synthetic data is not personal data and is free to share across borders
- Synthetic data is free to share with third parties and provides realism
- A synthetic data sandbox provides access to data across the enterprise for a 360-degree customer view

The 16 highest value synthetic data use cases in banking

> "MOSTLY AI's synthetic data platform comes with advanced features, such as direct database connection and the ability to synthesize complex data structures with referential integrity."

Synthetic data for AI, advanced analytics, and machine learning

Synthetic data for AI/AA/ML is one of the richest use case categories with many high-value applications. According to Gartner, by 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated. Machine learning and AI unlocks a range of business benefits for retail banks.

- **Advanced analytics** improves customer acquisition by optimizing the marketing engine with hyper-personalized messages and precise next best actions.
- **Intelligence from the very first point** of contact increases customer lifetime value. Since synthetic data is not personal data and contains all the intelligence of the original, no customer consent is needed to harness its power.
- **Operating costs will be lower** if decision-making in acquisition and servicing is supported with well-trained machine learning algorithms. Synthetic training data is better than real data, not only due to privacy compliance. The synthesization process allows room for data augmentation, such as upsampling, for better AI performance. Lower credit risk is also a benefit that comes from early detection of behaviors that signal a higher risk of default.
- **Underserved customer segments** can get the credit they need by fixing embedded biases via data synthesization.
- **Mass-market AI explainability,** increasingly demanded by tech-savvy customers, will be impossible to provide without synthetic data.

Automated, personalized decisions across the entire enterprise can increase competitiveness. The data backbone, the appropriate tools, and talent need to be in place to make this happen. Synthetic data generation is one of those capabilities essential for an AI-first bank to develop. The reliability and trustworthiness of AI is a neglected issue. According to Gartner:

> **“65% of companies can't explain how specific AI model decisions or predictions are made. This blindness is costly. AI TRiSM tools, such as MOSTLY AI's synthetic data platform, provide the Trust, Risk and Security Management needed for effective explainability, ModelOps, anomaly detection, adversarial attack resistance and data protection. Companies need to develop these new capabilities to serve new needs arising from AI adoption.”**

From explainability to performance improvement, synthetic data generators are one of the most valuable building tools. Data science teams need synthetic data to succeed with AI and machine learning use cases. Here is how to use synthetic data in the most common AI banking applications.

CUSTOMER ACQUISITION AND ADVANCED ANALYTICS

CRM data is the single most valuable data asset for customer acquisition and retention. A wonderful, rich asset that holds personal data and behavioral data of the bank's future prospects. However, due to privacy legislation, up to 80% of CRM data tends to be locked away. Compliant CRM data for advanced analytics and machine learning applications is hard to come by. Banks either comply with regulations and refrain from developing a modern martech platform altogether or break the rules and hope to get away with it.

There is a third option. Synthetic customer data is as good as real when it comes to training machine learning models. The model learns the patterns in the original data and extracts granular level insights for advanced analytics. Once the model is trained, the power of personalized messages and landing pages can be unleashed. The insights help identify new prospects and improve sign-up rates significantly.

MORTGAGE ANALYTICS, CREDIT DECISIONING AND LIMIT ASSESSMENT

AI in lending is a hot topic in finance. Banks want to reach out to the right people with the right mortgage and credit products. In order to increase precision in targeting, a lot of personal data is needed. The more complete the customer data profile, the more intelligent mortgage analytics becomes. Better models bring lower risk both for the bank and for the customer.

Rule-based or logistic regression models rely on a narrow set of criteria for credit decision-making. Banks without advanced behavioral analytics and models underserve a large segment of customers. People lacking formal credit histories or deviating from typical earning patterns are excluded. AI-first banks utilize huge troves of alternative data sources. Modern data sources include social media, browsing history, telecommunications usage data, and more. However, using these highly personal data sources in their original for training AI models is strictly forbidden.

Legacy data anonymization techniques destroy the very insights the model needs. Synthetic data versions retain all of these insights. Thanks to the granular, feature-rich nature of synthetic data, lending solutions can use all the intelligence. Credit and mortgage decision-making, including limit assessment, needs synthetic training data to maximize the value of the underlying machine learning model.

RISK MANAGEMENT AND PRICING

Pricing and risk prediction models are one of the most important to get right. Even a small improvement in their performance can lead to significant savings and competitive pricing. Injecting additional domain knowledge into these models, such as synthetic geolocation data or synthetic text from customer conversations, significantly improves the model's ability to quantify a customer's propensity to default.

MOSTLY AI's ability to provide the accuracy needed to generate synthetic geolocation data has been proven already. A large North American insurance provider synthesized home addresses and linked those synthetic coordinates with weather data to improve the performance of their pricing models. The same idea works with mortgage products. Synthetic text data can be used for training machine learning models in a compliant way on transcripts of customer service interactions. Virtual loan officers can automate the approval of low-risk loans reliably.

It is also mission-critical to be able to provide insight into the behavior of these models. Local interpretability is the best approach for explainable AI today, and synthetic data is a crucial ingredient of this transparency.

FRAUD AND ANOMALY DETECTION

Fraud is one of the most interesting AI/ML use cases. Fraud and money laundering operations are incredibly versatile, getting more and more sophisticated every day. Adversaries are using a lot of automation too to find weaknesses in financial systems. It's impossible to keep up with rule-based systems and manual follow-ups. False positives cost a lot of money to investigate, so it's imperative to continuously improve precision aided with machine learning models.

To make matters even more challenging, fraud profiles vary widely between banks. The same recipe for catching fraudulent transactions might not work for every financial institution. Using machine learning to detect fraud and anomaly patterns for cybersecurity is one of the first synthetic data use cases banks usually explore. The fraud detection use case goes way beyond privacy and takes advantage of the data augmentation possibility during synthesization. Maurizio Poletto, CPO at Erste Group Bank, recommends synthetic data upsampling to improve model performance.

Training and retraining models with synthetic data can improve fraud detection model performance by as much as 10%, leading to millions of dollars in savings on investigating false positives alone.

> “Synthetic data can be used to train AI models for scenarios for which limited data is available—such as fraud cases. 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. The more cases you have, the more detailed the model can be.”

CYBERSECURITY: AI AND THE ZERO TRUST DATA MODEL

Modern approaches limiting the impact of adversarial intrusions or reducing the blast radius heavily rely on high-performing AI systems. Anomaly detection and network intrusion detection can be improved with new, upsampled synthetic data. What's more, the single biggest cybersecurity risk today comes not from the outside but from within. According to Gartner:

> "59% of privacy incidents originate with an organization's own employees. Worse still — 45% of employee-driven privacy failures come from intentional behavior (though it may not be malicious)."

In the age of the zero trust cybersecurity approach, organizations need synthetic data alternatives more than ever to protect the privacy of their customers even within the bank's walls.

MONITORING AND COLLECTIONS

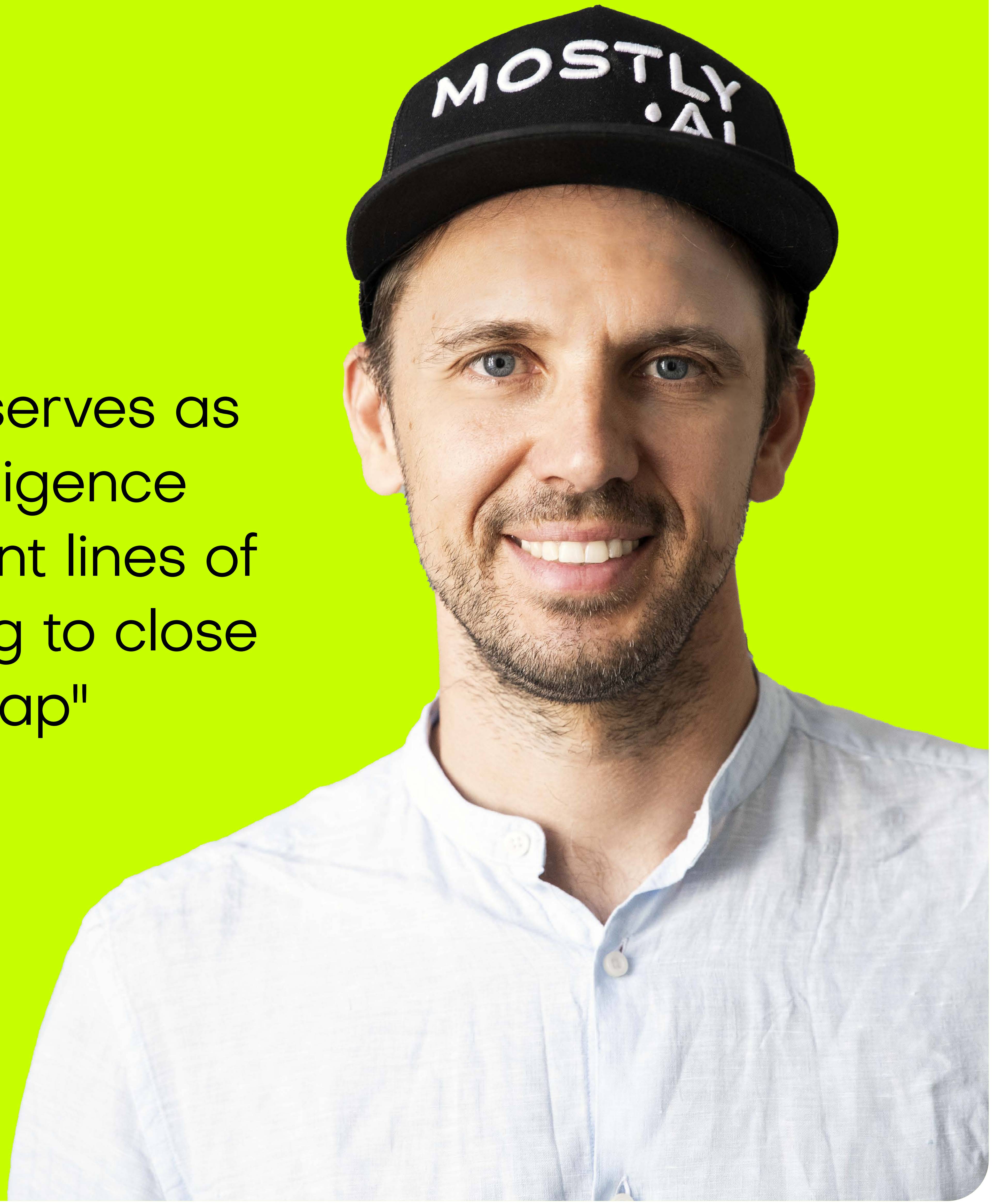
Transaction analysis for risk monitoring is one of the most privacy-sensitive AI use cases banks need to be able to handle. Apart from traditional monitoring data, like repayment history and credit bureau reports, banks should be looking to utilize new data sources, such as time-series bank data, complete transaction history, and location data. Machine learning models trained with these extremely sensitive datasets can reliably microsegment customers according to value at risk and introduce targeted interventions to prevent defaults. These highly sensitive and valuable datasets cannot be used for AI/ML training without effective anonymization.

MOSTLY AI's synthetic data generator is one of the best on the market when it comes to synthesizing complex time-series, behavioral data, like transactions with high accuracy. Behavioral synthetic data is one of the most difficult synthetic data categories to get right, and without a sophisticated AI engine, like MOSTLY AI's, results won't be accurate enough for such use cases.

CHURN REDUCTION, SERVICING, AND ENGAGEMENT

Another high-value use case for synthetic behavioral data is customer retention. A wide range of tools can be put to good use throughout a customer's lifetime, from identifying less engaged customers to crafting personalized messages and product offerings. The success of those tools hinges on the level of personalization and accuracy the initial training data allows. Machine learning models are the most powerful at pattern recognition. ML's ability to identify microsegments no analyst would ever recognize is astonishing, especially when fed with synthetic transaction data. Synthetic data can also serve as a bridge of intelligence between different lines of business: private banking and business banking data can be a powerful combination to provide further intelligence, but strictly in synthetic form. The same applies to national or legislative borders: analytics projects with global scope can be a reality when the foundation is 100% GDPR compliant synthetic data.

> “Synthetic data serves as a bridge of intelligence between different lines of business, helping to close the innovation gap"



Synthetic data for enterprise data sharing

Open financial data is the ultimate form of data sharing. According to McKinsey, economies embracing financial data sharing could see GDP gains of between 1 and 5 percent by 2030, with benefits flowing to consumers and financial institutions. More data means better operational performance, better AI models, more powerful analytics, and customer-centric digital banking products facilitating omnichannel experiences. The idea of open data cannot become a reality without a robust, accurate, and safe data privacy standard shared by all industry players in finance and beyond. This is a vision shared by Erste Bank's Chief Platform Officer:

> 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 this data to someone to analyze the correlation between them.
>
> 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."
>
> *Erste Bank's CPO, Maurizio Poletto*

It's an ambitious plan, and like all grand designs, it's best to start building the elements early. At this point, most banks are still struggling with internal data sharing with distinct business lines acting as separate entities and being data protectionist when open data is the way forward. Banks and financial institutions share little intelligence, citing data privacy and legislation as their main concern.

However, data sharing might just become an obligation very soon with the EU putting data altruism on the map in the upcoming Data Governance Act. While sharing personal data will remain strictly forbidden and increasingly so, anonymized data sharing will be expected of companies in the near future.

In the U.S., healthcare insurance companies and service providers are already legally bound to share their data with other healthcare providers. The same requirement makes a lot of sense in banking too where so much depends on credit history and risk prediction. While some data is shared, intelligence is still withheld. Cross-border data sharing is also a major challenge in banking. Subsidiaries either operate in a completely siloed way or share data illegally. According to Axel von dem Bussche, Taylor Wessing's partner and IT lawyer, as much as 95% of international data sharing is illegal due to strict internal policies and the destruction of the EU-US Privacy Shield by the Schrems II decision. Some organizations fly analysts and data scientists to the off-shore data to avoid risky and forbidden cross-border data sharing. It doesn't have to be this complicated. Synthetic data sharing is compliant with all privacy laws across the globe. Setting up synthetic data sandboxes and repositories can solve enterprise-wide data sharing across borders since synthetic data does not qualify as personal data.

As a result, it is out of scope for GDPR and the infamous Schrems II. ruling, which effectively prohibited all sharing of personal data outside the EU.

Third-party data sharing within the same legislative domain is also problematic. Banks buy many third-party AI solutions from vendors without adequately testing the solutions on their own data.

The data used in procurement processes is hard to get, causing costly delays and heavily masked to prevent sensitive data leaks through third parties. The result is often bad business decisions and out-of-the-box AI solutions that fail to deliver the expected performance. Synthetic data sandboxes are great tools for speeding up and optimizing POC processes, saving 80% of the cost.

Synthetic test data for digital banking products

One of the most common data sharing use cases is connected to developing and testing digital banking apps and products. Banks accumulate tons of apps, continuously developing them, onboarding new systems, and adding new components. Manually generated test data for such complex systems is a hopeless task, and many revert to the old dangerous habit of using production data for testing systems. Banks and financial institutions tend to be more privacy-conscious, but their solutions to this conundrum are still suboptimal. Time and time again, we see reputable banks and financial institutions roll out apps and digital banking services after only testing them with heavily masked or manually generated data. One-cent transactions and mock data generators won't get you far when customer expectations for seamless digital experiences are sky-high.

To complicate things further, complex application development is rarely done in-house. Data owners and data consumers are not the same people, nor do they have the full picture of test scenarios and business rules. Labs and third-party dev teams rely on the bank to share meaningful test data with them, which simply does not happen. Even if testing is kept in-house, data access is still problematic. While in other, less privacy-conscious industries, developers and test engineers use radioactive test data in non-production environments, banks leave testing teams to their own devices. Manual test data generation with tools like Mockaroo and the now infamous Faker library miss most of the business rules and edge cases so vital for robust testing practices. Dynamic test users for notification and trigger testing are also hard to come by. To put it simply, it's impossible to develop intelligent banking products without intelligent test data. The same goes for the testing of AI and machine learning models.

Testing those models with synthetically simulated edge cases is extremely important to do when developing from scratch and when recalibrating models to avoid drifting. Models are as good as the training data, and testing is as good as test data. Payment applications with or without personalized money management solutions need the synthetic approach: realistic synthetic test data and edge case simulations with dynamic synthetic test users. Synthetic test data is fast to generate and can create smaller or larger versions of the same dataset as needed throughout the testing pyramid from unit testing, through integration testing, UI testing to end-to-end testing.

Erste Bank's main synthetic data use case is test data management. The bank is creating synthetic segments and communities, building new features, and testing how certain types of customers would react to these features.

> "Normally, the data we use is static. We see everything from the past. But features like notifications and triggers—like receiving a notification when your salary comes in—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."
>
> *Erste Bank's CPO, Maurizio Poletto*

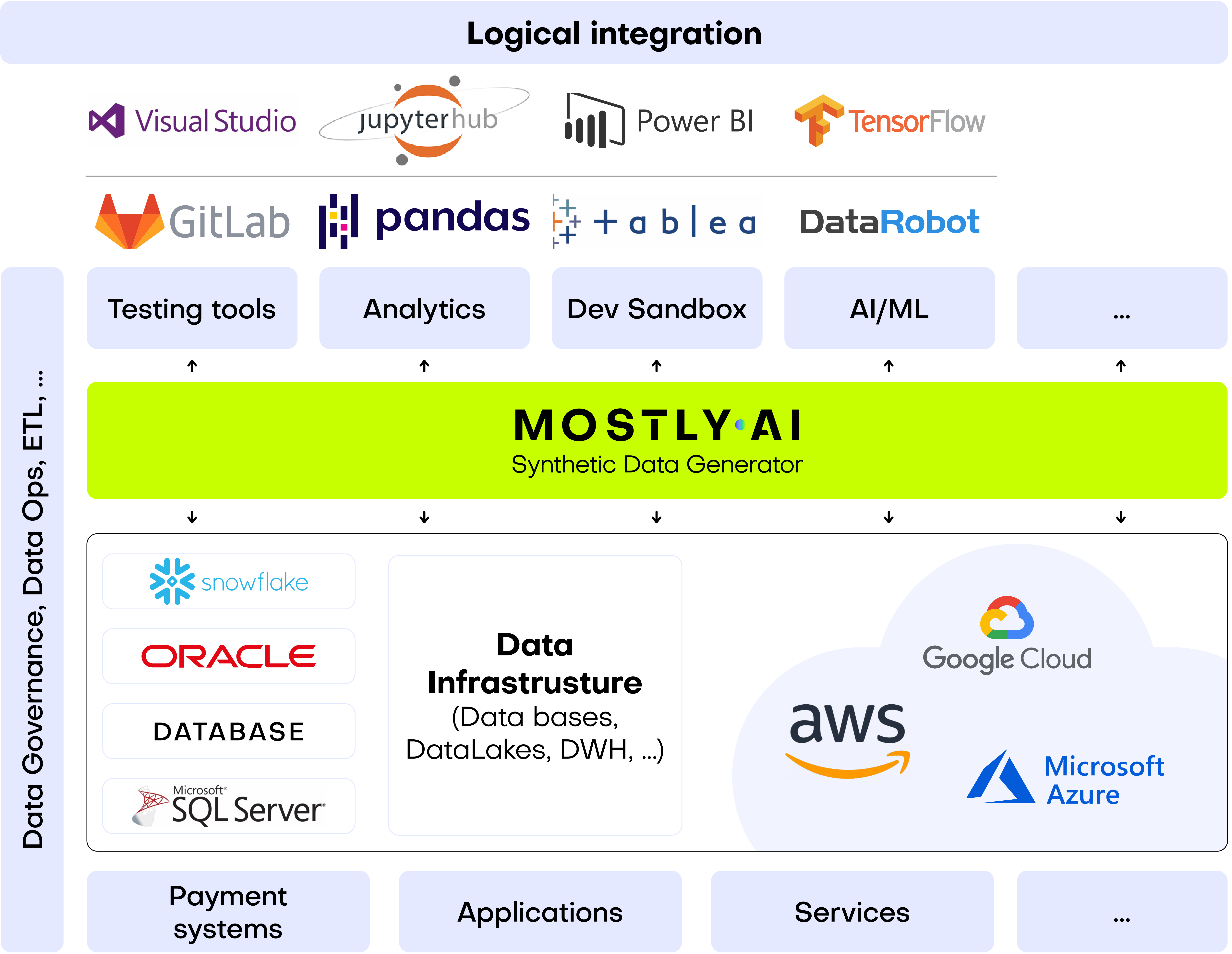
These live, synthetic users can stand in for production data and provide a level of realism unheard of before while protecting customers' privacy. The Norwegian Data Protection Authority issued a fine for using production data in testing, adding that using synthetic data instead would have been the right course to take.

Testing is becoming a continuous process. Deploying fast and iterating early is the new mantra of DevOps teams. Setting up CI/CD (continuous integration and delivery) pipelines for continuous testing cannot happen without a stable flow of high-quality test data. Synthetic data generators trained on real data samples can provide just that – up-to-date, realistic, and flexible data generation on-demand.

How to integrate synthetic data generators into financial systems

First and foremost, it's important to understand that not all synthetic data generators are created equal. It's particularly important to select the right synthetic data vendor who can match the financial institution's needs. If a synthetic data generator is inaccurate, the resulting synthetic datasets can lead your data science team astray. If it's too accurate, the generator overfits or learns the training data too well and could accidentally reproduce some of the original information from the training data.

Open-source options are also available. However, the control over quality is fairly low. Until a global standard for synthetic data arrives, it's important to proceed with caution when selecting vendors. Opt for synthetic data companies, which already have extensive experience with sensitive financial data and know-how to integrate synthetic data successfully with existing infrastructures.



Synthetic data engineering in banking

The future of financial data is synthetic

Our team at MOSTLY AI has seen large banks and financial organizations from up close. We know that synthetic data will be the data transformation tool that will change the financial data landscape forever, enabling the flow and agility necessary for creating competitive digital services. While we know that the direction is towards synthetic data across the enterprise, we know full well how difficult it is to introduce new technologies and disrupt the status quo in enterprises, even if everyone can see the benefits. One of the most important tasks of anyone looking to make a difference with synthetic data is to prioritize use cases in accordance with the needs and possibilities of the organization.

Analytics use cases with the biggest impact and generate the biggest value can serve as flagship projects, establishing the foundations of synthetic data adoption. In most organizations, mortgage analytics, pricing, and risk prediction use cases can generate the highest immediate monetary value, while synthetic test data can massively accelerate the improvement of customer experience and reduce compliance and cybersecurity risk. It's good practice to establish semi-independent labs for experimentation and prototyping: Erste Bank's George Lab is a prime example of how successful digital banking products can be born of such ventures. The right talent is also a crucial ingredient of success. According to Erste Bank's CPO, Maurizio Poletto:

> “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. 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.”

Once you have the attraction of a state-of-the-art tech stack enabling agile data practices, you can start building cross-functional teams and capabilities across the organization. The data management status quo needs to be disrupted, and privacy, security, and data agility champions will do the groundwork. Legacy data architectures keeping banks and financial institutions back from innovating and endangering customers' privacy need to be dealt with soon. The future of data-driven banking is bright, and that future is synthetic.

How to start your synthetic data journey

1

START WITH TABULAR DATA

Tabular data has always been the single most important format for financial institutions. Gartner recommends starting your synthetic data exploration by synthesizing tabular data.[1] Identify valuable tabular assets, synthesize, and publish. It's fast, easy, and rewarding.

2

FIND THE RIGHT SYNTHETIC DATA VENDOR

Although open source synthetic data generators are available, they come with serious limitations regarding accuracy and privacy guarantees. Their performance can be volatile and is highly dependent on the community behind it. Closed source offers more sophisticated capabilities and commercial services you can count on.

Choose a vendor with in-depth experience in the financial industry, capable of augmenting as well as synthesizing data. Demand automated privacy and accuracy quality assurance. Use third party research from trusted sources, such as Gartner and Forrester, to identify viable and robust players.

3

SET UP A SYNTHETIC DATA EXCELLENCE CENTRE

Managing data access requests takes up the majority of CDOs' time and resources. By setting up a Synthetic Data Excellence Centre, you can provide a quick, painless, compliant, and fully audited process to request synthetic versions of data.

4

CREATE SYNTHETIC DATA LAKES

Set up synthetic data lakes to mirror your most valuable and insightful data assets. Colleagues across your organization can use it as a self-service data center to access decision-ready data.

By making synthetic data flow freely throughout your organization, true data-centricity is born: data-driven decision making and data literacy increases and works in a self-reinforcing fashion.

5

REVIEW CURRENT PRACTICES OF DATA ANONYMIZATION

What was sufficient to protect data a few years ago no longer suffice. Classic anonymization techniques, like randomization, pseudonymization, generalization, or permutation do not protect against linkage attacks.[2]

90% Of a banks' data can be utilized in its synthetic form and the remaining 10% sensitive pii data can remain heavily protected, minimizing risk in the event of a data leak.[3] Educate citizen data scientists about data anonymization risks and synthesize wherever possible.

[1] Predicts 2021: Data and Analytics to Govern, Scale and Transform Digital Business, Gartner https://www.gartner.com/en/documents/3993855/predicts-2021-data-and-analytics-strategies-to-govern-sc

[2] Semantic Web Enabled Record Linkage Attacks on Anonymized Data, Jacob Miracle and Michelle Cheatham http://ceur-ws.org/Vol-1750/paper-03.pdf

[3] Cracking FIs' 90/10 Data Monetization Problem, PYMNTS https://www.pymnts.com/news/security-and-risk/2019/synthetic-data-financial-services-arm-insight/

Notes for your data science team

Important questions and answers about AI-generated synthetic data.

What data types can you synthesize?
MOSTLY AI's synthetic data platform can synthesize numerical, categorical, datetime, short text (ex. transaction text), and geographic data. All data must be provided in a tabular format.

Is time series data supported?
Yes. Time series data is modeled in a two-table setup. The first table, called the subject table, contains unique identifiers. The second table, called the linked table, contains events belonging to a unique identifier. For example, user accounts and their transactions.

How is privacy guaranteed?
Privacy is built into the generation process in multiple ways. The model uses a random generative process to avoid direct duplicates in the synthetic data. Outlier handling protects column-wise privacy by ensuring that unique values don't occur in the synthetic data.

How do you guarantee that outliers don't persist in the synthetic data, potentially leaking sensitive information?
Outliers are handled in two different ways, depending on the data type. For numerical data, any values between the 10th and 90th percentile are clipped away. For categorical data, values appearing more than n times are replaced based on a sliding scale.

How are the privacy and accuracy metrics defined?
After each synthetic generation is complete, a custom QA report is generated. We have open-sourced our metrics in a Python library.[4]

Can synthetic data preserve referential integrity?
Yes. Generation is done by first generating identifiers, then their associated attributes. This way, keys referenced in the subsequent tables are guaranteed to exist.

What is the quality of MOSTLY AI's synthetic data for AI/ML model training?
Always highly accurate. Small fluctuations in the accuracy depend on how much data the model was trained with, how long the model is trained for, and how complex the model becomes. Overall, synthetic data can capture 80-99% underlying patterns of the original data. We have done extensive research[5][6] covering the use of synthetic data in ML training. The results are consistently on par or better than training with real data.

Can synthetic data capture patterns of fraudulent transactions?
Fraudulent transactions often suffer from two related problems; lack of fraudulent cases and inability to correctly identify fraud. Synthetic data can capture the complex relationships within the data. It can retain the patterns associated with fraud, even if the number of examples is minimal.

Can synthetic data reproduce business rules?
MOSTLY AI's synthetic data platform can reproduce business rules implicitly. Due to the random nature of the generative process, small violations may occur, but this does not affect the overall quality of the synthetic dataset produced.

Is there a UI?
MOSTLY AI's synthetic data platform has an intuitive interface, with drag and drop functionality, interactive runtime graphs, and the ability to queue multiple runs. Once a model is created, it is possible to generate more sets of synthetic data without having to wait through training time again.

Is there a non-UI feature that allows automated data pipelines in production?
MOSTLY AI's synthetic data platform has an API feature and the ability to read from network drives. These two features allow easy integration for automated data pipelines in production.

- API
- Read data from network drives

Is the application a cloud solution or on-premise installation? Does MOSTLY AI's synthetic data platform interact with the external internet?

- MOSTLY AI's synthetic data platform is available for both as an on-cloud and an on-premise installation.
- No internet connection is required to use MOSTLY AI's synthetic data platform.

[4] Virtual Data Lab, https://github.com/mostly-ai/virtualdatalab
[5] The World's Most Accurate Synthetic Data Platform? Let's check the Numbers! https://mostly.ai/2020/09/25/the-worlds-most-accurate-synthetic-data-platform/
[6] Boost your Machine Learning Accuracy with Synthetic Data https://mostly.ai/2020/08/07/boost-machine-learning-accuracy-with-synthetic-data/

About MOSTLY AI

▷ Talk to one of our experts

MOSTLY AI is the leading synthetic data company globally. Its platform enables enterprises across industries to unlock, share, fix and simulate data.

Thanks to the advances in artificial intelligence MOSTLY AI's synthetic data looks and feels just like real data, is able to retain the valuable, granular-level information, yet guarantee that no individual is ever getting exposed. This enables businesses to drive innovation and digital transformation, overcome data silos, improve machine learning models as well as application testing capabilities. MOSTLY AI was founded in 2017 and is headquartered in Vienna, Austria. Its global operation includes customers in a variety of verticals, including banking, insurance and telecommunications.

Contact: hello@mostly.ai

Vienna office (HQ)
MOSTLY AI Solutions MP GmbH
Landstrasser Hauptstrasse · 71/2 · Wien · Austria · 1030

New York office
MOSTLY AI Inc.
500 7th Ave 8th floor · New York · NY 10018 · United States

MOSTLY·AI