



Generating synthetic data with SAS Data Maker- solution

Jussi Varjus

Principal Technical Account Manager, SAS

AI/ML Applications and Data Issues

Sufficient amounts of High-Quality data is
the basis of all AI development

The AI era is here, and it will determine winners and losers...

\$16
Trillion

AI is projected to contribute an additional \$15.7 trillion to the global economy by 2030 ([PWC](#))

66%
productivity gains

Studies show AI tools can increase worker productivity by up to 66%, equating to decades of natural productivity growth ([Nielsen Norman Group](#))

72%
adoption rate

As of 2024, 72% of organizations have adopted AI in at least one business function ([McKinsey](#))

\$200B
investment

Global AI investment is expected to approach \$200 billion by 2025. ([Goldman Sachs](#))

...yet serious data challenges impede organizations from adopting and succeeding with AI.

Data Privacy & Regulatory Compliance

Highly regulated industries must adhere to strict data privacy laws, making sharing or using real-world data difficult.

Incomplete & Imbalanced Data

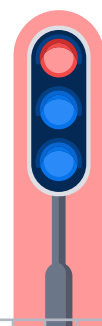
Real-world data often has gaps, missing fields, or imbalanced representations, creating biased or underperforming AI models.

Cost of Data Collection

Collecting, cleaning, and labeling high-quality real-world data is time-consuming and expensive, particularly for large organizations handling massive datasets.

Difficulty in Data Sharing for co-operation

Organizations often face hurdles sharing data across internal teams and partners due to privacy and regulatory constraints.



Synthetic data helps organizations overcome their data challenges quickly and securely.

Simplifying Privacy & Compliance

Anonymized Data for Safe Use: Mimics real-world patterns without exposing sensitive information, ensuring compliance while enabling AI.

Cross-border Data Sharing: Allows easier data sharing across teams, departments, or even countries without breaching data residency laws.

Augmenting & Balancing Data

Fill Data Gaps: Generate realistic samples to fill missing data, enriching datasets.

Balance Data: Balance rare events (like fraud or disease outbreaks), allowing AI models to learn effectively from all scenarios.

Reducing Cost of Data Collection

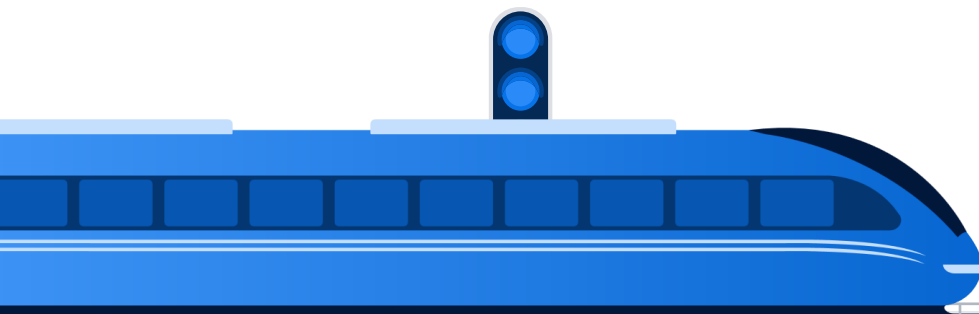
Cost Efficient: Can be generated programmatically, reducing the need for expensive data collection and manual labeling.

Faster AI Deployment: Accelerate data preparation, allowing your organization to move from concept to production more quickly.

Enabling Data Sharing

Facilitate Collaboration: Share synthetic datasets that retain desired statistical properties without compromising privacy.

Accelerate Innovation Ecosystems: Ease of data sharing fosters collaboration across business units or with external partners, driving AI advancements.



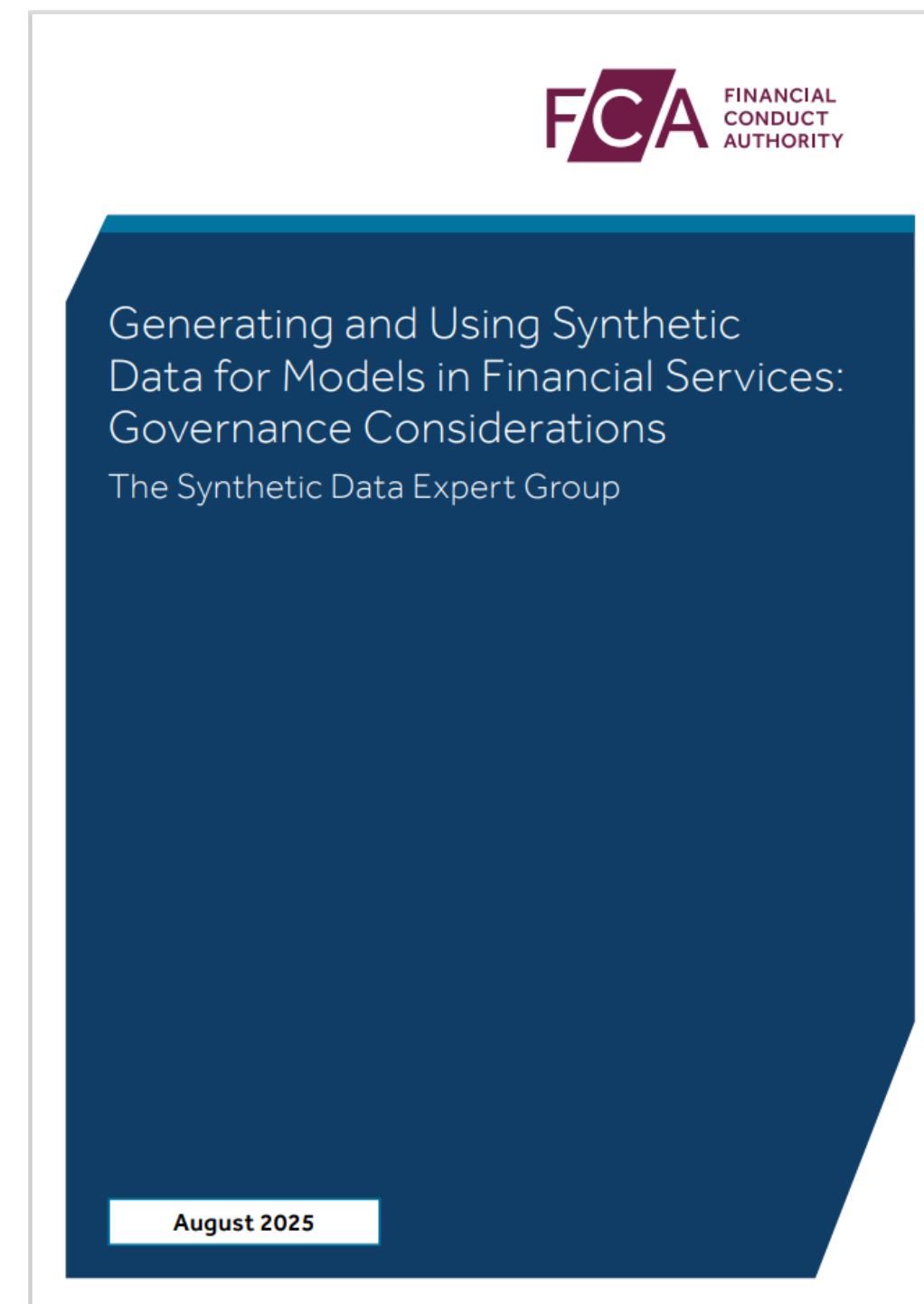
The Synthetic Data era is also here...

Synthetic data should also be generated responsibly

"Synthetic data offers a powerful way to unlock the value of data, enable experimentation, model development, and broader innovation across the financial system – all while maintaining strong privacy protections and public trust.

Recognising the potential of this technology, we convened the FCA's Synthetic Data Expert Group (SDEG) to bring together leaders from across financial services, academia, and the public sector. Our aim was simple: to enable open and practical conversations about how synthetic data is being used, where the challenges lie, and what's needed to move forward responsibly."

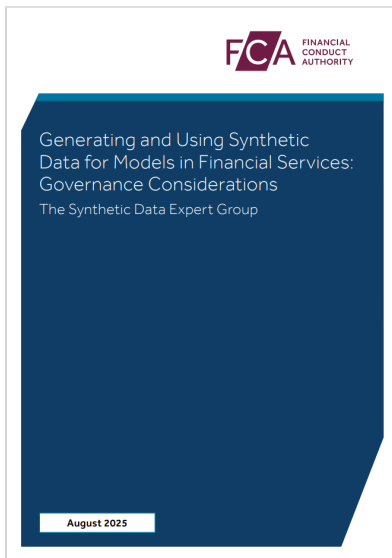
- Jessica Rusu, CDIIO, FCA



9 Key Principles by FCA

Drawing on the MRM and Data & AI Ethics frameworks

1. **Accountability:** Establish clear accountability structures for data, algorithmic and AI systems, defining responsibilities throughout the data and AI lifecycle. Accountability extends to technologies or models from third-party providers and managed service providers, with documented chains of responsibility.
2. **Safety:** Design systems with safety as a priority, encompassing reliability, robustness, and accuracy.
3. **Transparency:** Maximise the information available to a decision-maker validating the system and its outputs.
4. **Explainability and Interpretability:** Ensure system's internal processes are understandable to humans and provide justification for specific outputs.
5. **Security and Privacy:** Design systems to protect both data security and individual privacy rights throughout the data lifecycle.
6. **Fairness:** Systems which process or impact social or demographic data are designed to prevent discriminatory outcomes.
7. **Agency:** Model operators reviewing algorithmic outputs to have meaningful ways to understand, question, and contest these decisions.
8. **Suitability:** Use cases are justified by genuine needs, informed by an understanding of current technological constraints, and considerate of broader socio-technical context
9. **Continuous Monitoring and Improvement:** Regularly assess models and systems to ensure they remain effective, compliant, and fit for purpose



SAS Data Maker

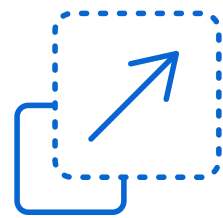
Elevate innovation,
productivity and quality
with trusted
synthetic data

Key Benefits of SAS Data Maker



Innovate Faster

- Ease of implementation
- Improve productivity by democratizing data generation
- Integrate flexibly with existing systems



Scalable Outcomes

- Address data scarcity
- Significantly reduce data acquisition costs
- Unlock the potential of existing data



Data You Can Trust

- Never have to choose between quality & security again
- Enhance privacy and address bias
- Build trusted and reliable AI models

SAS Data Maker Capabilities

Address gaps and
limitations in
real-world data

1

Generate data on demand using algorithms. Enable pre- and post- processing data



2

Replicate data characteristics including statistical properties and distributions



3

Evaluate synthetic data quality with visual evaluation metrics



Customer Experiences

Global Organizations Are Closing Data Gaps
With Synthetic Data



Synthetic data helped improve AI model accuracy, potentially reducing losses.

PROBLEM

- Machine learning credit scoring models must guide personal loan decisions in a **fair, responsible manner**
- Previous models using only real-life data faced **limitations in accuracy, scalability, and data sensitivity**

ACTION

- **Tested synthetic data generated with SAS** to improve machine learning model, comparing model accuracy versus models using only real-life data

RESULT

- **28% improvement** in model accuracy with facilitated machine learning efforts, supporting potential reduction in losses



Synthetic data augmented insufficient real data, helping conserve right whales

PROBLEM

- North Atlantic right whales are endangered, partly due to being struck accidentally by boats
- Need a predictive heat map of whale movement to inform ship captains
- Insufficient whale sighting data to validate predictive model of whale movement

ACTION

- Used SAS to generate synthetic data to achieve sufficient data for validation (expanded set from 40K to 500K data points).
- Fathom was able to [validate their predictive model](#), increasing confidence using machine learning.

RESULT

- WhaleCast tool can be integrated into existing boat on-board touch screens, helping mariners better understand areas with greater risk of striking whales

Organization: **US Health Care Provider**
Industry: **Health Care**
Region: **United States**



Synthetic data unlocks value from sensitive data without compromising privacy and security

PROBLEM

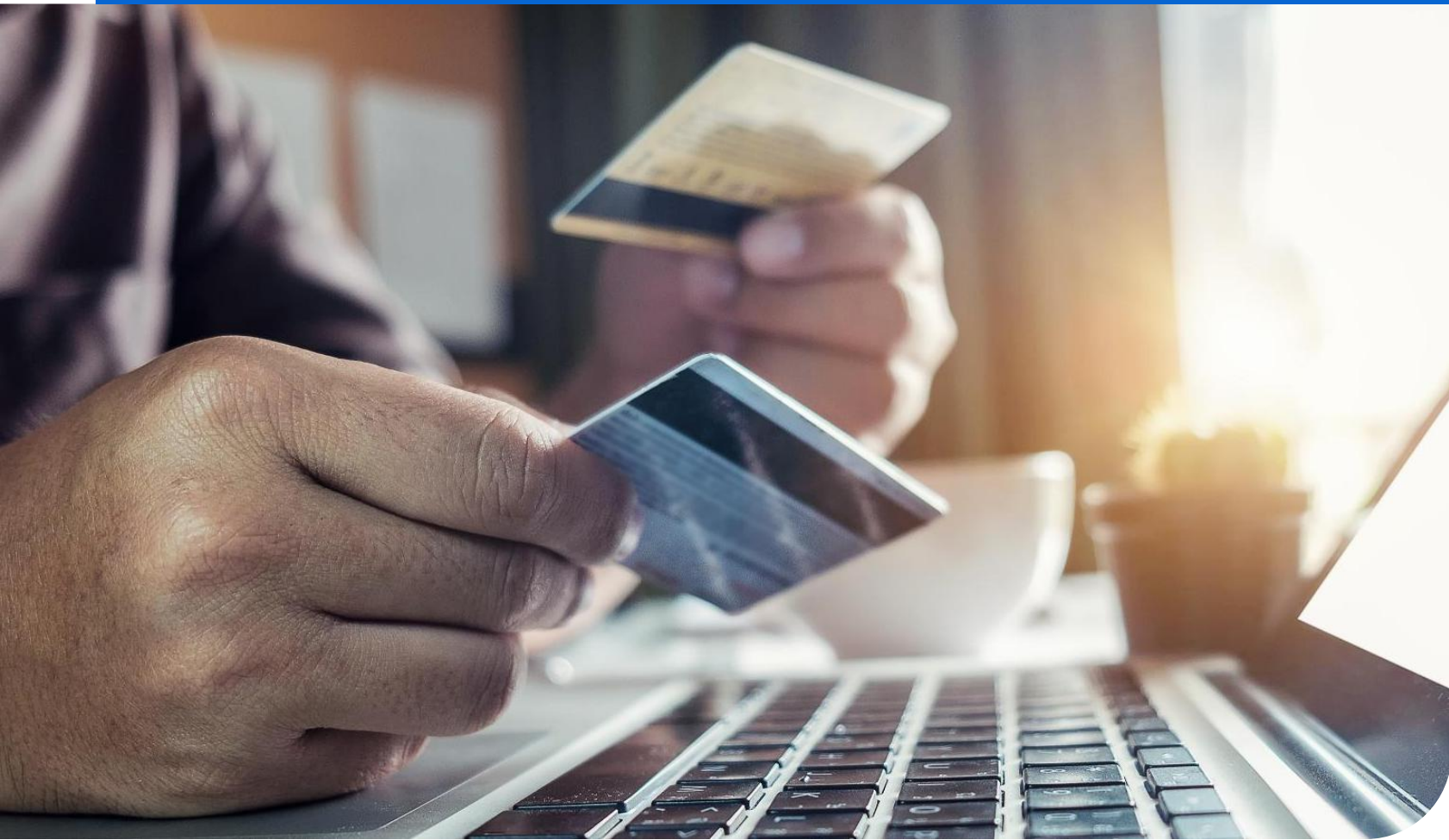
- Patient healthcare data is protected by privacy regulations, restricting access to data and increasing costs.
- These costs made performing analytics with the data cost-prohibitive.
- Provider must support an ecosystem of shared research and innovation involving universities and research groups.

ACTION

- Provide secure and unified platform which enables synthetic data generation to simulate patient behavior and outcomes, test treatment plans and choose optimal care paths.
- Platform also offers critical supporting capabilities such as data transformation, generation and post-processing.

RESULT

- Enables patients to benefit from data-driven research on synthetic clinical datasets
- Reduced risk to customer privacy



Synthetic data enhanced customer privacy and collaboration with external vendors

PROBLEM

- Lack of safe and realistic test data. Privacy leakage concerns.
- Needed a way to share data with external vendors without risks to customer privacy and compliance.

ACTION

- Generated synthetic data repository for on-demand test data.
- New products were generated in a sandbox with synthetic data.

RESULT

- Sandbox was safely accessed by over 700 external developers.
- Enhanced privacy, reduced risk, improved collaboration with external vendors.

Reducing data access time for a large telecom provider

PROBLEM

- Customer churn predictive model was encumbered by data access delays
- Long delays to access real data and train model led to out-of-date results

ACTION

- Tested synthetic data into their training and testing process
- Results with synthetic data were identical results to real data

RESULT

- Reduced data access time **from weeks to minutes** while maintaining model accuracy
- Enabled up-to-date predictive churn model to improve customer retention

Test case	ROC AUC (test set)	ROC AUC (train set)	Accuracy (test set)	Accuracy (train set)
Real dataset (benchmark)	64.2%	71.3%	93.9%	93.9%
Synthetic dataset (for both training and testing)	61.9%	70.4%	94.1%	94.1%
Combined set (real and synthetic data together)	61.4%	67.7%	94.0%	93.9%
Training with synthetic dataset, testing with real dataset	64.2%	71.3%	93.9%	93.9%

Feedback on SAS Data Maker has been very positive

“I like the way it was so easy to use and fast”

“An effective solution for creating microdata that can be used for analysis without compromising sensitive information.”

“Gave us quality indicators of microdata generated which [increased] our confidence that the inferred results would reflect results of real-world data”

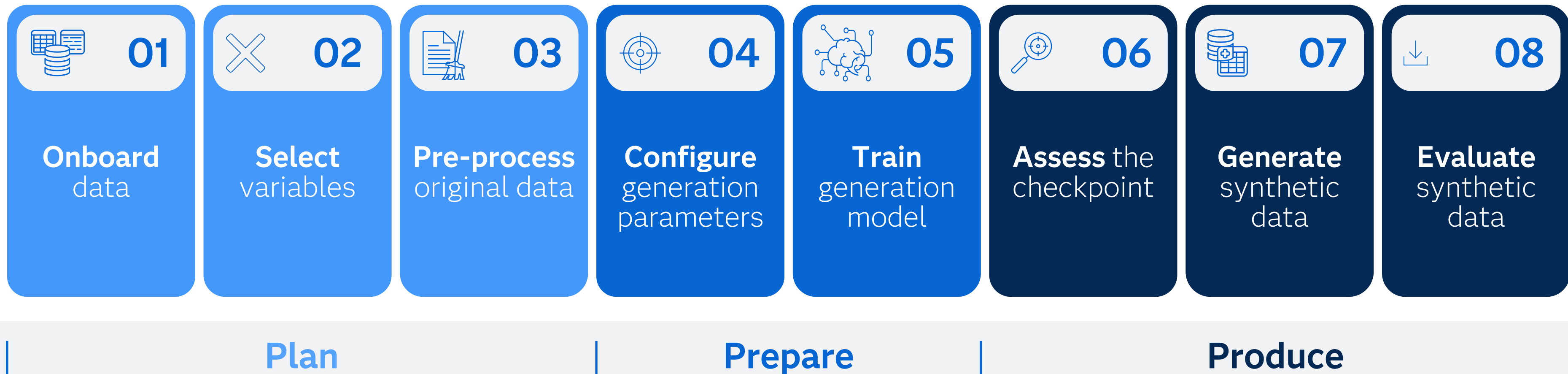
“My overall experience with it has been generally positive, especially when it comes to generating structured, high-quality test data for different use cases.”

SAS Data Maker -process

Careful preparation and testing produce the most reliable results

SAS Data Maker Process

Synthetic data generation follows these common steps



SAS Data Maker -demo

#SASInnovate

Copyright © SAS Institute Inc. All rights reserved.

sas innovate
on tour 2025



SAS® Data Maker

Your Gateway to Seamless Synthetic Data Generation

Create New Project

Projects



How to

Movies ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 3

✔ Complete

FraudGen ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 1

🔄 Training

EcomMock ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 0

📄 Draft

Cars ⋮

sample data

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 5

✖ Error

RetailSales2025 ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 3

✔ Complete

StreetView x3 ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 7

✔ Complete

CryptoGen ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 3

🔄 Training

Bio Sim Data ⋮

sample data

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 12

🔄 Generating

FauxSet ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 1

✔ Complete

MockWorld ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 14

✔ Complete

PharmaSim Data ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 2

📄 Draft

EHR Data ⋮

Trained dataset with sampled data.

Source data: AZURE/datamakerdev:datamaker-input-data
Date Modified: 27 Feb, 2025 12:55:02
Models: 11

✔ Complete



Overview

See an overview of SAS Data Maker

[Watch Video](#)



Get Started

Learn about Synthetic Data Generator capabilities and how to get started with SAS.

[View documentation](#)



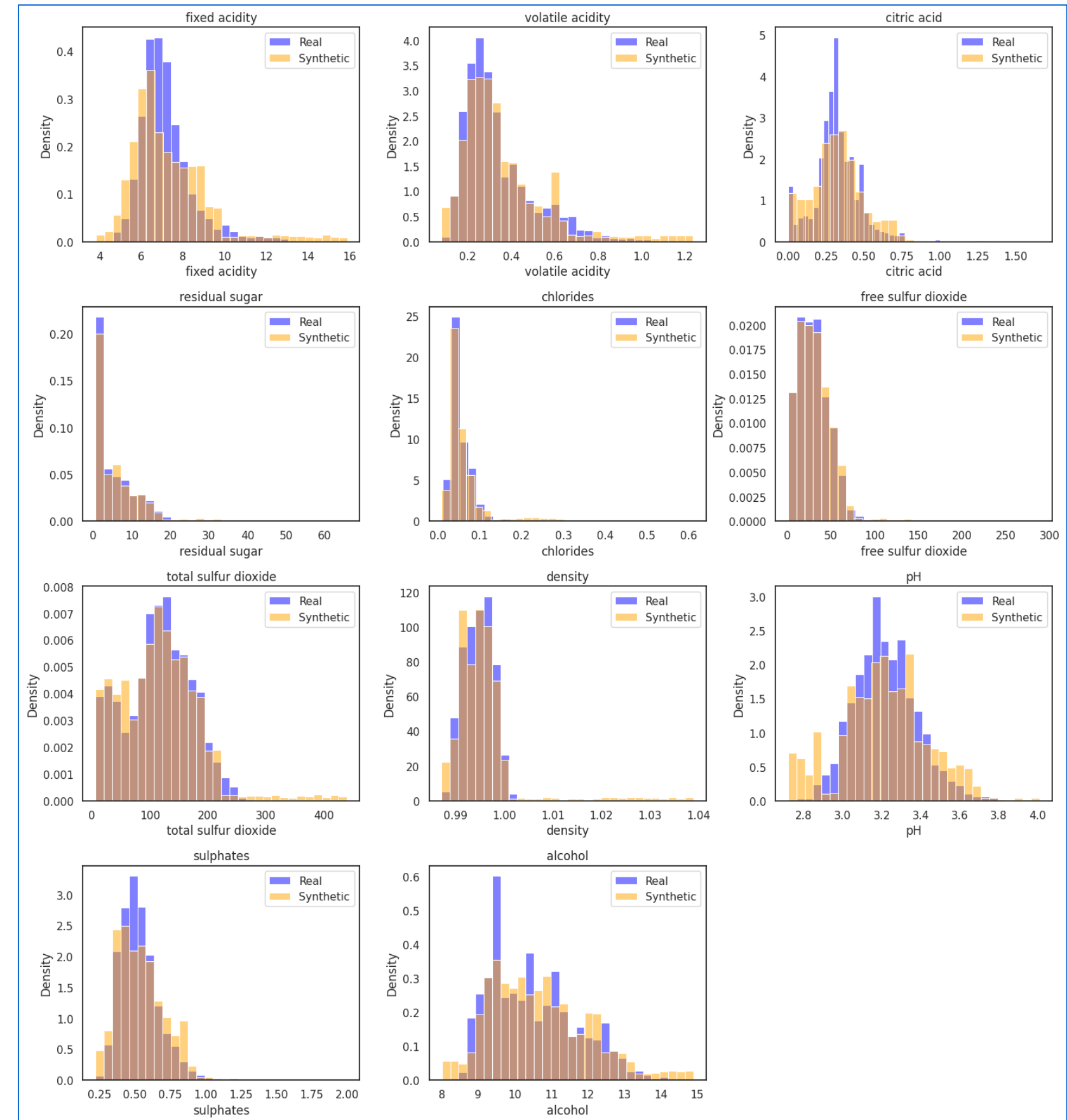
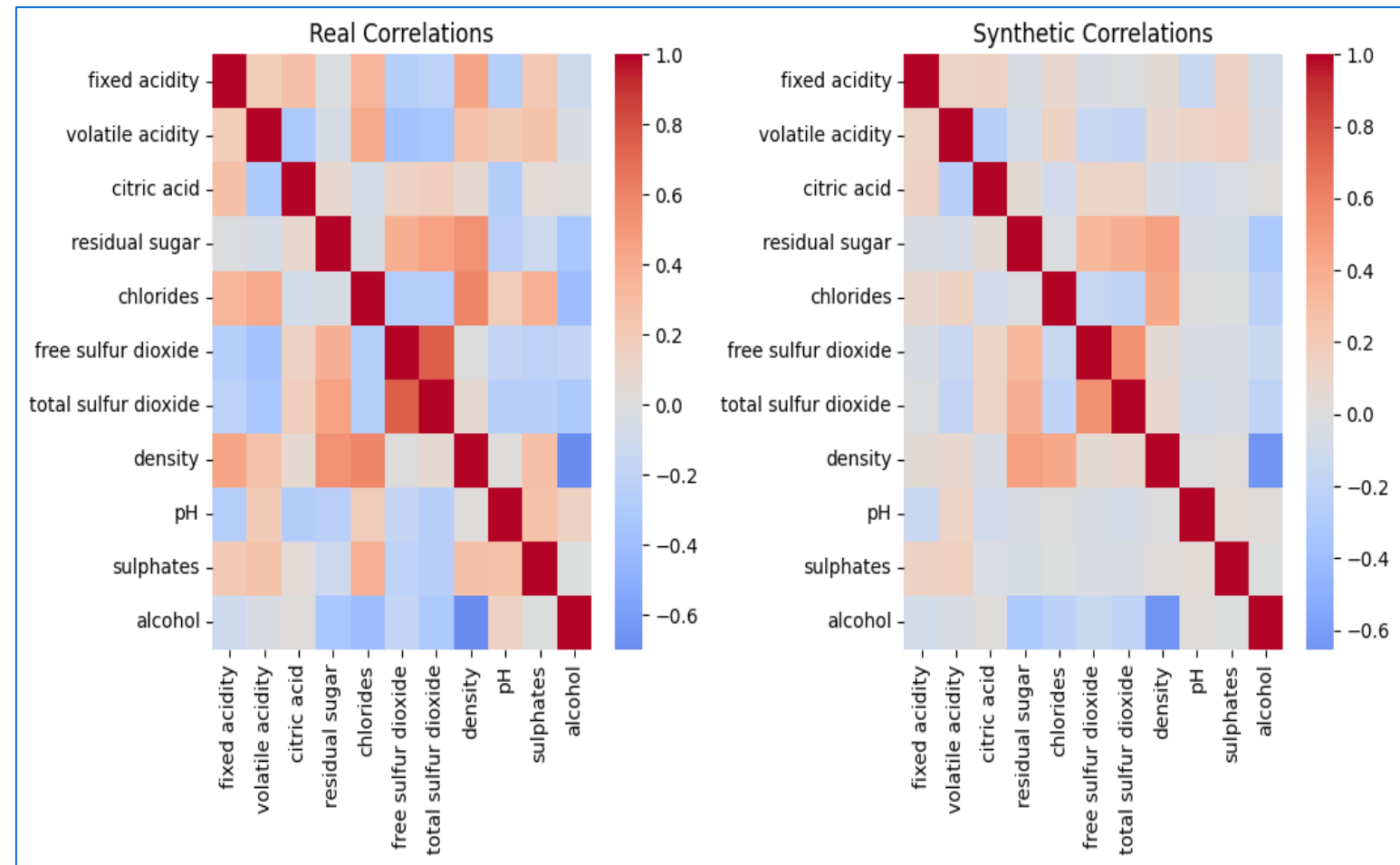
Provide Feedback

Tell us about your experience.

[Open feedback form](#)

Some tests with Synthetic Data Generation

OBS! PrivBayes- algorithm



Model	Accuracy (train)	Accuracy (test: real)	Accuracy (test: bag)	Trained and tested on real data
Random Forest	67%	68%	67%	74%
Logistic Regression	61%	65%	64%	63%
Decision Tree	67%	62%	59%	75%

#SASInnovate

Copyright © SAS Institute Inc. All rights reserved.

Delivery & Deployment Targets



Deployment

Azure (Customer's Azure tenant)

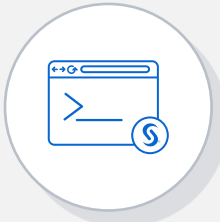


Architecture

Build the service running on the Workbench Split-Plane Architecture

Now

Next →



SAS Data Maker in Microsoft Marketplace

\$1/1k



SAS Data Maker available on AWS Marketplace



SAS Data Maker in SAS Viya Integration:

- Model Studio
- Data Management
- Model Manager



SAS Data Maker available in Snowflake



SAS Data Maker in SAS Viya Workbench Integration



SAS Data Maker SaaS

Customer Benefits with SAS Data Maker

Customers will experience enhanced innovation and research capabilities, a competitive edge through simulation of multiple future scenarios, and development of trustworthy AI systems.



Enhanced innovation and research by providing access to rich synthetic datasets, fostering new opportunities and breakthroughs.



Faster time-to-market via rapid generation of high-quality synthetic data, accelerating the development cycle for AI projects.



Trustworthy AI systems with robust synthetic data processes and diverse synthetic datasets, enabling organizations to develop reliable AI systems that adhere to ethical standards.



Increased data privacy and security by generating synthetic data that doesn't expose real, identifiable information, allowing organizations to operate more confidently.



Cost savings by reducing reliance on costly data collection methods, making data for analytics more accessible.



Kiitoksia!

- More info on SAS Data Maker!

sas innovate
on tour 2025

#SASInnovate

Copyright © SAS Institute Inc. All rights reserved.