

Synthetic Data Will Save the Day: Fact or Fiction

Sherrine Eid, Global Head, Epidemiology, RWE and Observational Research
Sundaresh Sankaran, Global Technology Practice



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

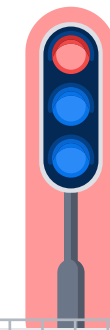
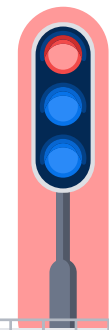
Real-world data often has gaps, missing fields, or imbalanced representations, creating biased or underperforming AI models.

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

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

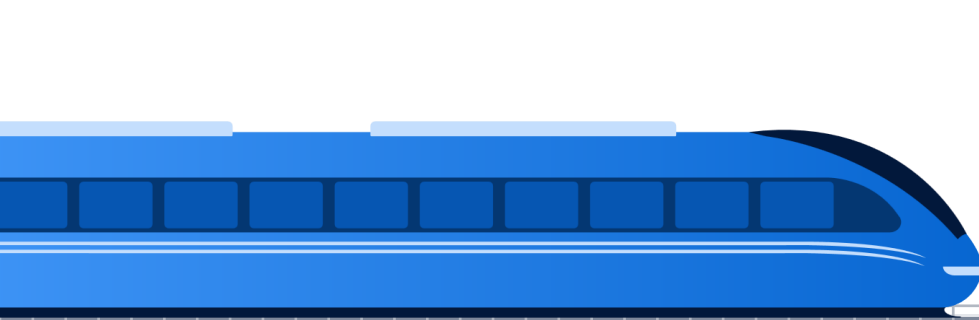
Cost Efficient:
Can be generated programmatically, reducing the need for expensive data collection and manual labeling.

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

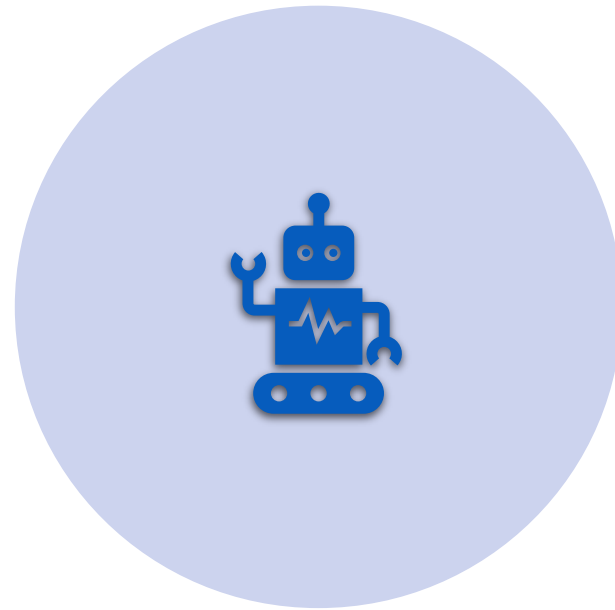
Enabling Data Sharing

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.



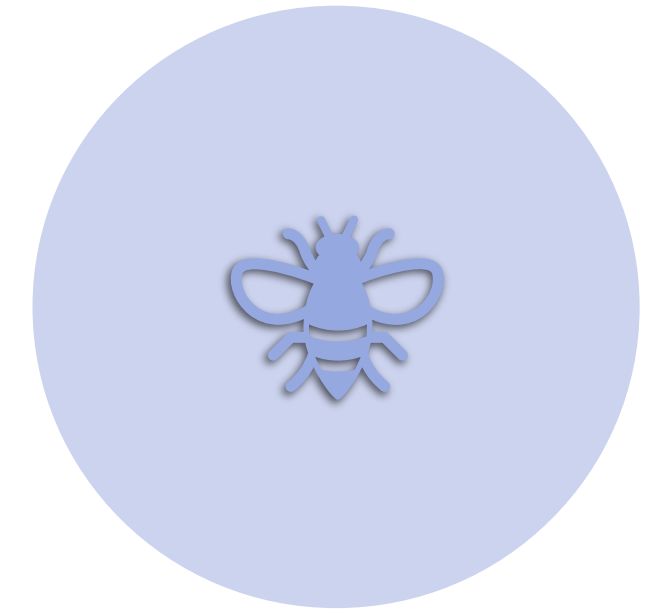
History



EMERGENCY



REGULATORY ACCEPTANCE



APPLICATIONS

Synthetic Data Market Drivers

Increased Data Demand

Growth in data demands associated with machine learning and model training

Data Monetization

Growing awareness of opportunities to monetize and exchange data with third parties

Limitations in traditional data anonymization techniques

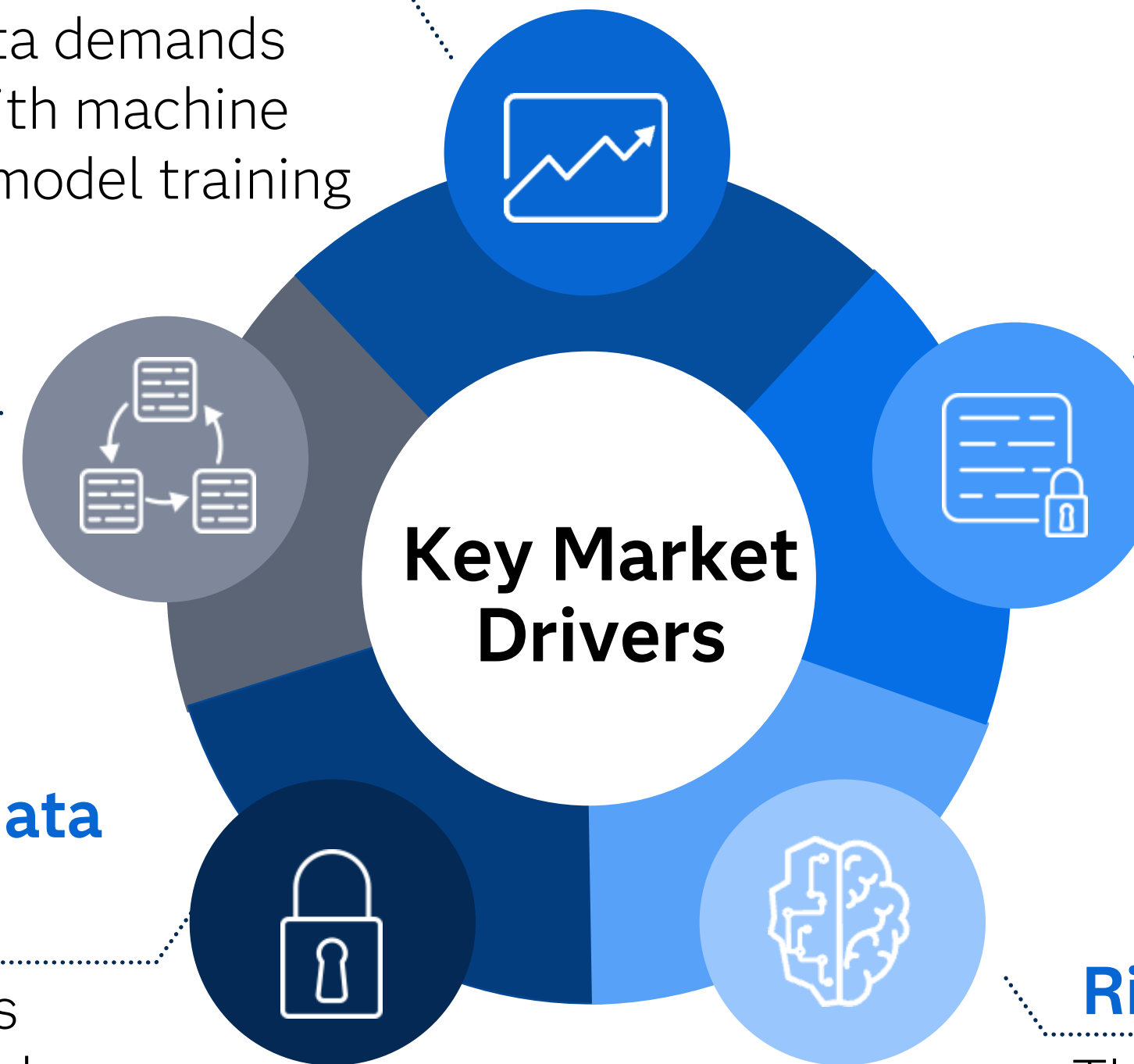
Data anonymization techniques often heavily limit the analytical utility of datasets

Elevated Privacy Risks

Regulatory demands on data privacy and protection are increasing

Rise of Generative AI

The explosion of innovation in AI foundation models is boosting synthetic data creation



Characteristics of Synthetic Data

Privacy Preserving



Privacy-preserving synthetic data preserves to a high degree the properties and statistical information of the original data.

Cost Effective



Cost-effective alternative to manual processes for collecting real-world data and third-party data sets to reduce data acquisition and processing costs

Balanced



Synthetic data can address data bias challenges by utilizing diverse data augmentation and generation methods

Scalable

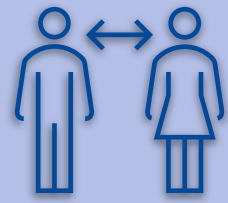


Compared to rules-based generated data, synthetic data is easier to create and offers flexibility, scalability and addresses data scarcity.

Five best practices for generating synthetic data

- 1 Select the right data generation technique for your business problem
- 2 Define and address privacy requirements upfront
- 3 Assess quality of synthetic data
- 4 Consider combining data augmentation and data generation techniques
- 5 Leverage methods and approaches that can scale for your enterprise (such as low code no code, able to run on cloud and on premise)

Examples of Success



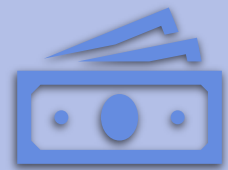
Virtual Control Groups:

Researchers have successfully used synthetic data to create virtual control groups, reducing the need for large-scale placebo trials and accelerating the testing process.



Modeling Drug Responses:

Synthetic data helps simulate how drugs might interact with different demographics and conditions, aiding in personalized medicine research.



Cost Savings:

Companies have reported significant cost savings by using synthetic data to streamline trial designs and reduce logistical challenges.

Examples of Failure



Accuracy Concerns:

The accuracy and validity of synthetic data can sometimes be questioned, especially when the underlying algorithms or models do not sufficiently represent real-world diversity and variability.



Ethical Issues:

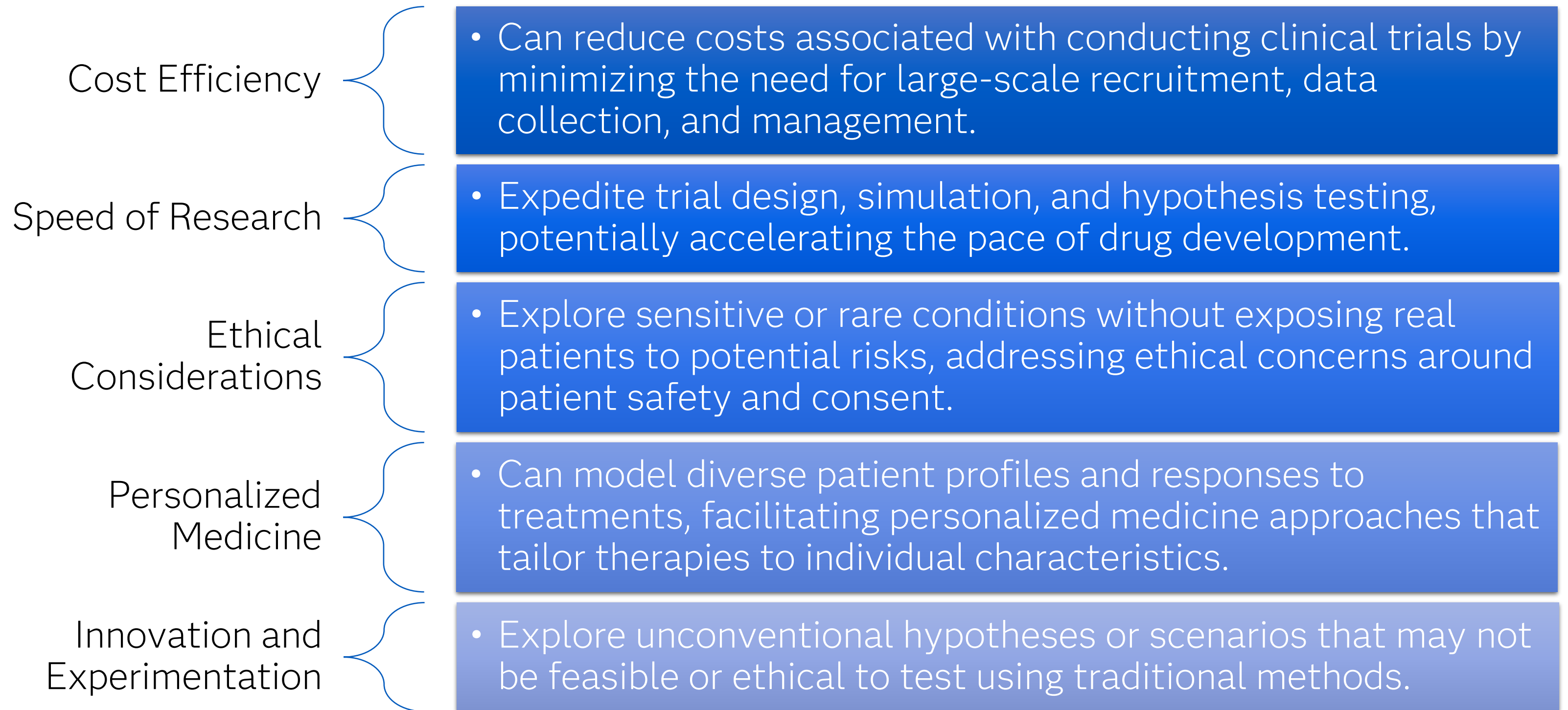
There are ethical concerns regarding the use of synthetic data, particularly in terms of privacy and consent, which need to be carefully addressed to gain public and regulatory trust.



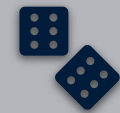
Regulatory Hurdles:

Regulatory bodies are still navigating the acceptance of synthetic data, with challenges in establishing standards and guidelines for its use in clinical trials.

The Promise



The Concern



Data Quality and Representativeness: The accuracy and representativeness of synthetic data are critical. Biases or inaccuracies in modeling real-world populations could lead to flawed conclusions or ineffective treatments.



Regulatory Acceptance: Regulatory bodies may be cautious in accepting synthetic data due to concerns about reliability, validation methods, and the robustness of simulations compared to real-world outcomes.



Privacy and Security: Generating synthetic data requires access to real patient data, raising concerns about privacy breaches or the unintentional re-identification of individuals.

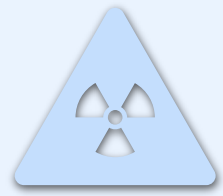


Lack of Transparency: The algorithms and methodologies used to generate synthetic data may not always be transparent or understandable, making it challenging to validate results or identify potential biases.



Integration with Real Data: Integrating synthetic data with real-world evidence or clinical trial data poses challenges in terms of compatibility, validation, and ensuring seamless transition between simulated and actual results.

Not a Silver Bullet



Synthetic data is not a replacement for real data.



It can mimic statistical properties of real data to some extent



It is inherently a distorted version and should be used as a tool in the research pipeline rather than the final substitute.

Opportunities for Synthetic Data Generation

Precision/Personalized
Medicine

Privacy Regulations

Constraints on linkage with
(patient) outcomes

Rare Diseases

Disease Prediction

Risk Assessment

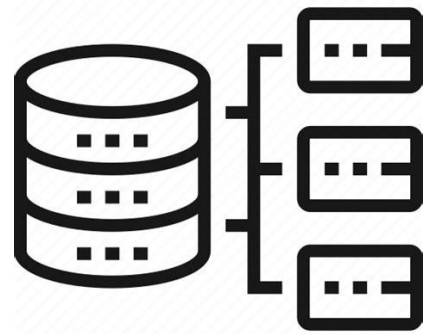
Data Sharing

Collaboration

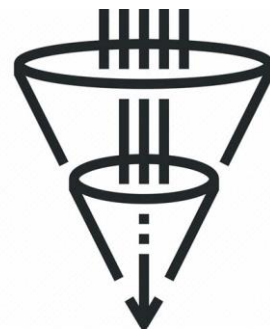
Regulatory Compliance

Train and validate machine
learning models

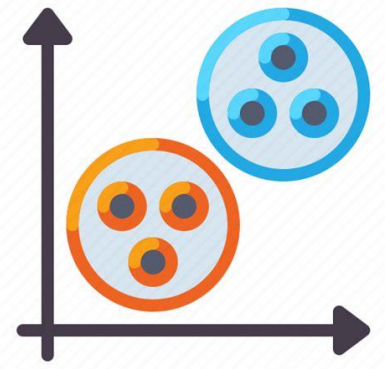
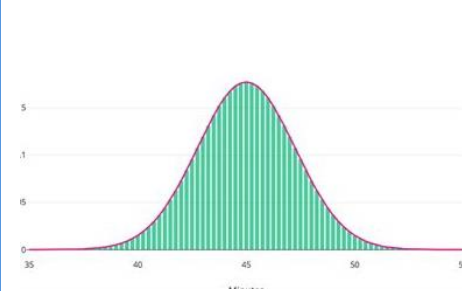
- drug discovery
- disease prediction
- population health



Data Management



Variable Selection



Synthetic Data Strategy



SDG - Train a Synthetic Dat...



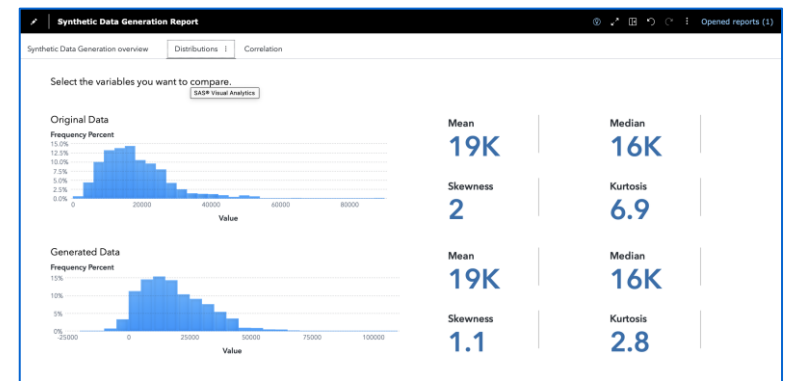
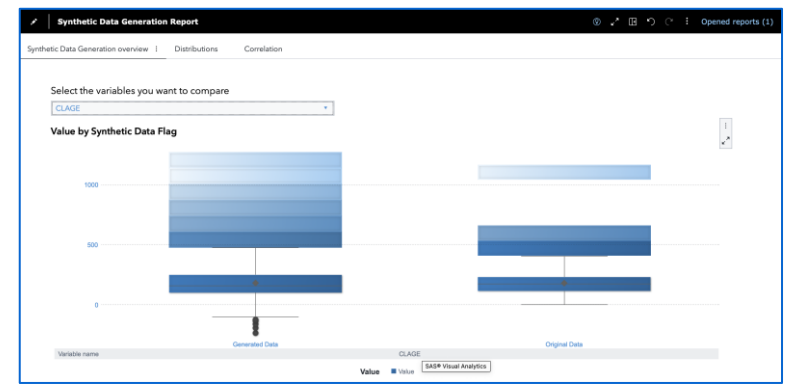
Generate Data



Generate Synthetic Data



Assess, Curate and Control



		y =	BAD	CLAGE	CLNO	DEBTINC	DELINQ	DEROG	LOAN	MORTDUE	NINQ	VALUE	YOJ
Generated Data	Synthetic_data_flag		r	r	r	r	r	r	r	r	r	r	r
	BAD		1	-0.17	0	0.2	-0.33	0.28	-0.08	-0.03	0.17	-0.03	-0.18
	CLAGE			1	0.24	-0.05	0.02	-0.04	-	-	0.13	-	0.2
	CLNO				1	0.18	-	-	-	-	-	-	-
	DEBTINC					1	-	-	-	-	-	-	-
Original Data	DELINQ						1	-	-	-	0.07	-	-
	DEROG							1	-	-	0.17	-	-
	LOAN								1	-	0.23	0.04	0.11
	MORTDUE									1	0.03	-0.04	-0.19
	NINQ										1	-	-
Generated Data	VALUE											1	0.01
	YOJ												1
	BAD												
Original Data	CLAGE												
	CLNO												

Secure, Govern & Administer

sas viya

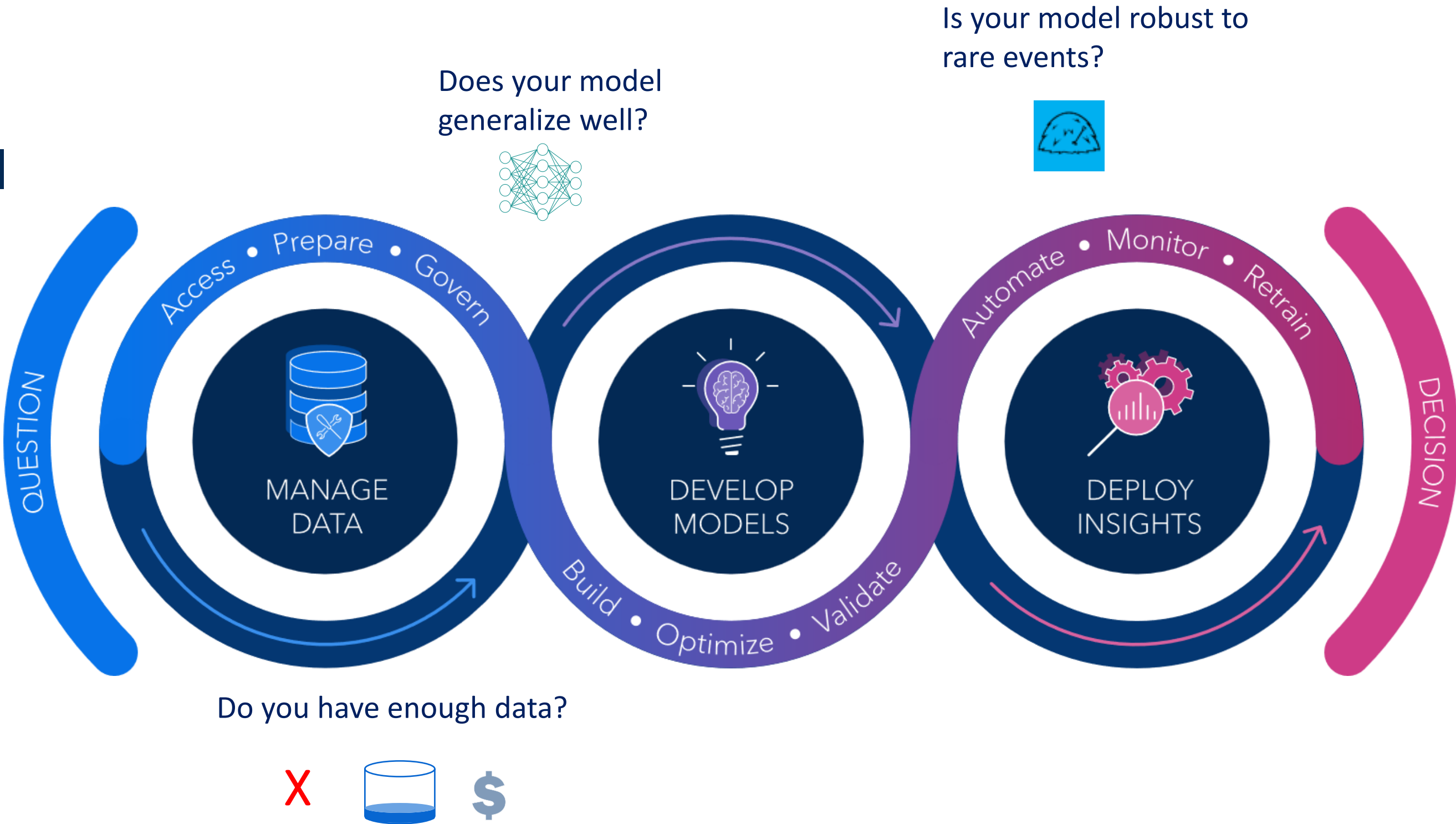
Quick Demo

Synthetic Data Needs Across AI Life Cycle

Is there sufficient representation among sensitive groups?



Do privacy issues prohibit access to or usage of real data?



Pre-generation tasks

Right Data begets Right Data

Data Assessment and Quality Checks

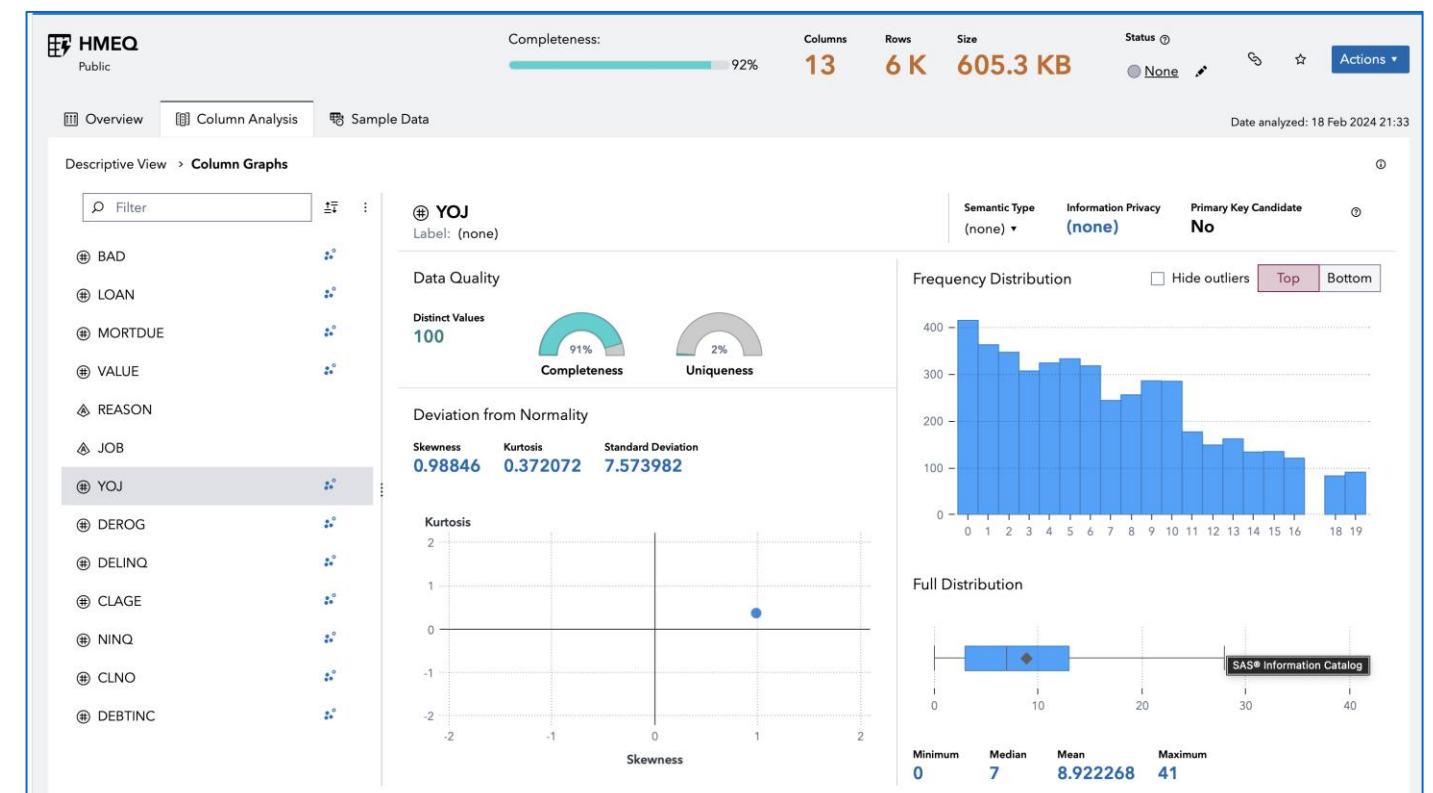
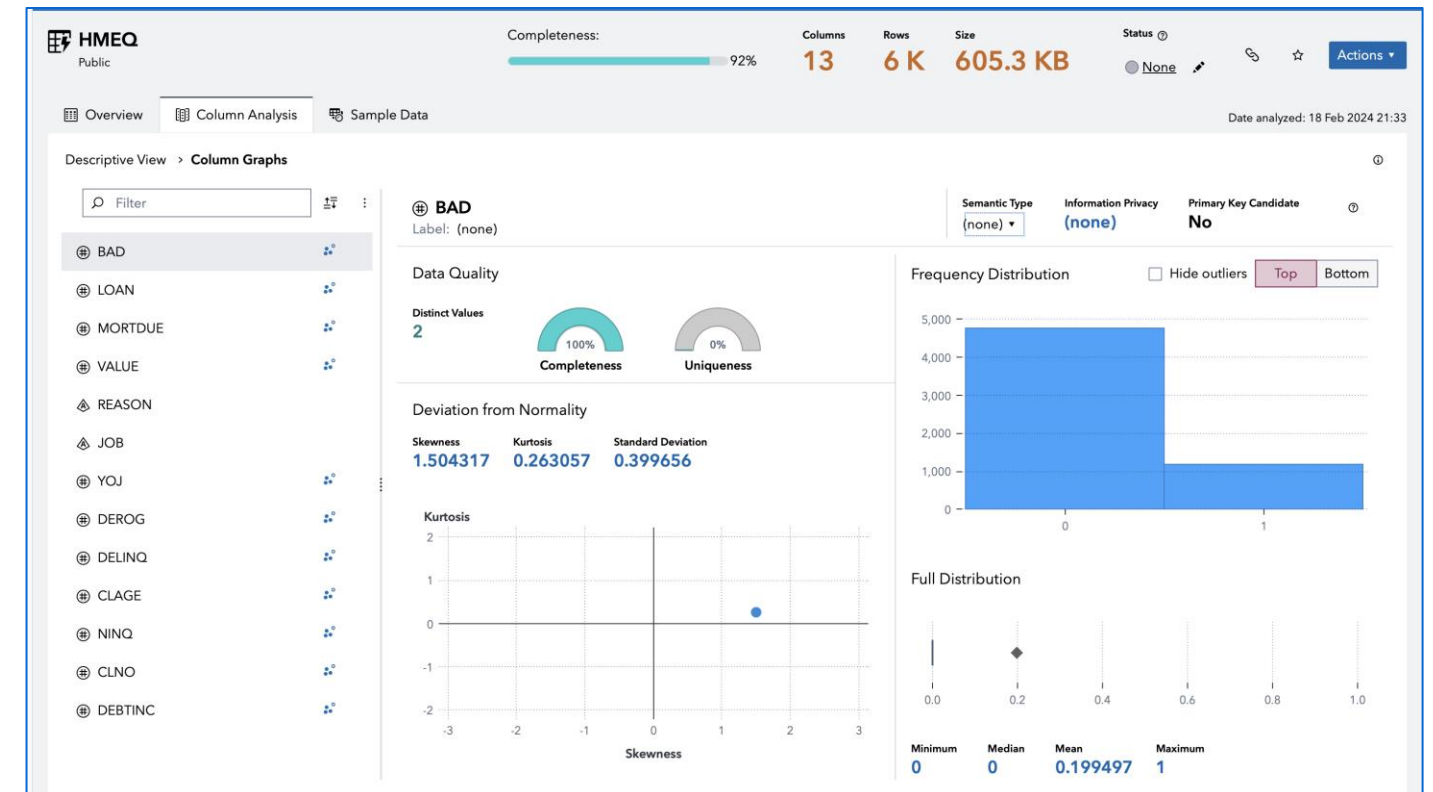
- De-identification
- Imputing missing values
- Identifying anomalies

Data Transformation

- Variable transformation and encoding
- Normalisation

'Learning' the Distribution

- Joint Probability Distributions
- Clustering & 'right' number of clusters





Recent

 Recycle Bin[illegible]

Select an item to see its information.

The graph displays the percentage of U.S. adults who are obese from 1980 to 2008. The x-axis represents the year, and the y-axis represents the percentage. The data shows a consistent upward trend, starting at about 15% in 1980 and reaching over 30% by 2008. A shaded area around the line indicates the confidence interval.



New Options View Open Save All

GPCI Demos



Steps



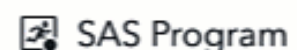
Type to filter list

SAS Steps : Shared

Data (Input and Output)



Develop



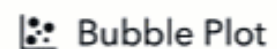
Transform Data

Integrate

Enrichment



Visualize Data



Start Page

Synthetic Data Generation - HMEQ.flw x +

Cancel



Add ▾

View ▾



Flow

Generated Code

Submitted Code and Results

Connect to CAS (if required)

Create CAS
Session and...

Baseline



Input data

Distribution of
Target Variable

Generate Synthetic Data using Generative Adversarial Networks (GAN)

Generate Synthetic Data using Synthetic Minority Oversampling Technique (SMOTE)

Example SMOTE application to boost a class

Assess Generated Data (GAN)

Assess Generated Data (SMOTE)

Persist table

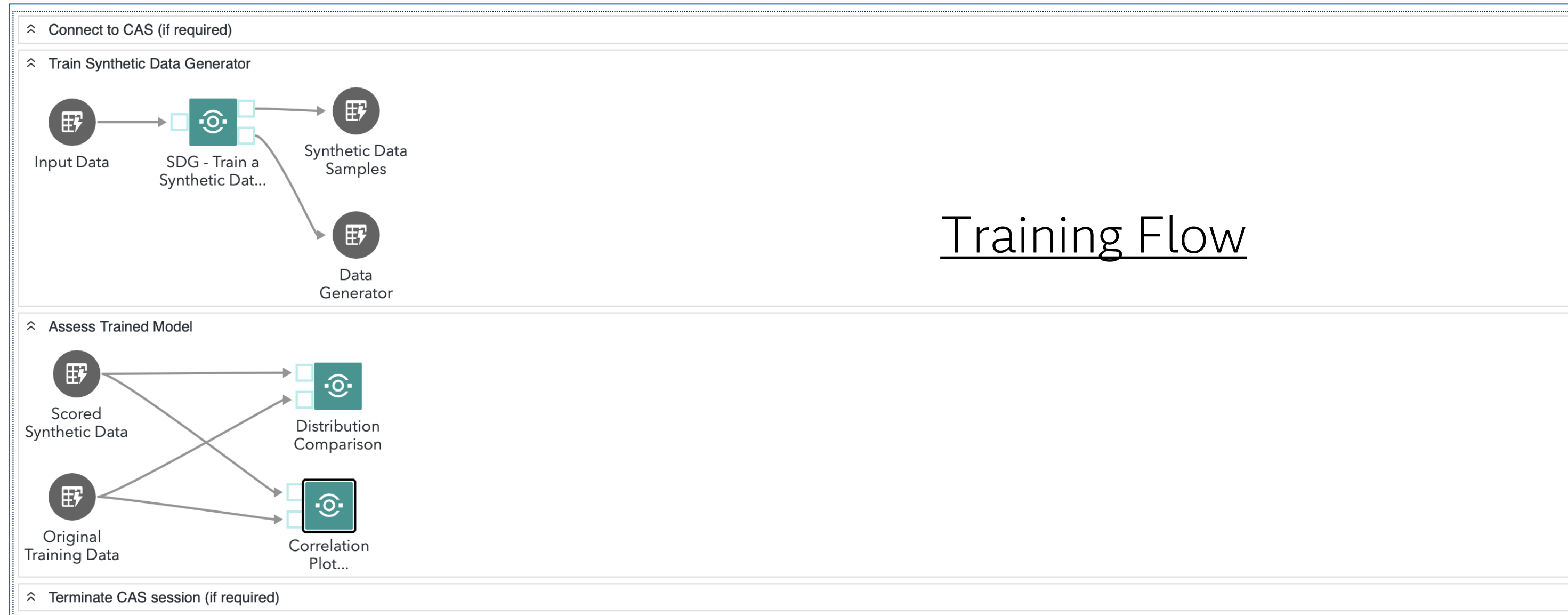
Create Analytical Base Table (ABT)

Create Analytical Base Table (ABT) based on SMOTE

Wind Down (if required)

Easy access = better user experience

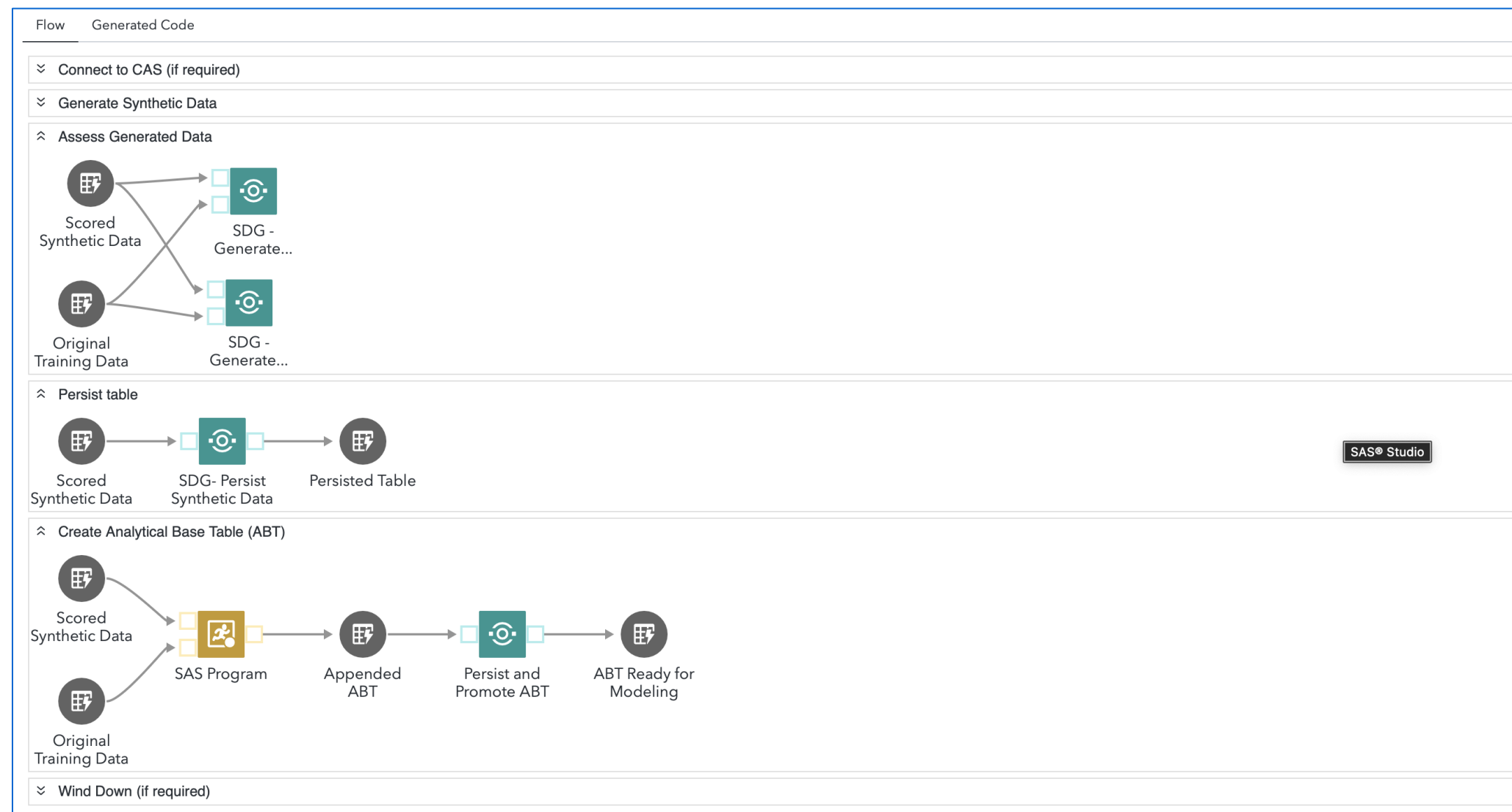
SAS Studio Custom Steps to make functionality available to a larger audience.



<https://github.com/sassoftware/sas-studio-custom-steps/tree/main/Synthetic%20Data%20Generation>

Easy access = better user experience

SAS Studio Custom Steps to make functionality available to a larger audience.



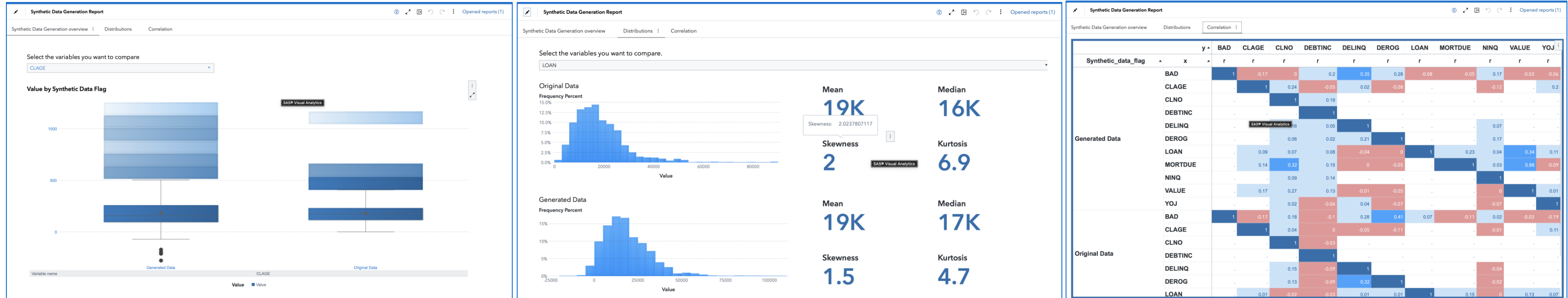
Scoring Flow



<https://github.com/sassoftware/sas-studio-custom-steps/tree/main/Synthetic%20Data%20Generation>

Post-generation tasks

I'll still poke it to see if it's the right data



Assessment

- Is my generated data the same as original data?

Leakage (and Leaching)

- Is it still possible to identify a random record?

Human-in-the-loop

- Impose limits and business rules



Algorithms for Generation

Synthetic Minority Oversampling Technique (SMOTE)

- New examples from the minority class
- Relates (draws a line) examples close together in feature space
- Identifies new examples on the relationship function ('line')
- Can select a class to augment

Generative Adversarial Networks (GANs)

- Correlation-preserving conditional tabular GAN
- Consists of a “generator” and a “discriminator” component
- Uses cluster centroids from encoded input data to learn the distribution
- Generates new data which is then assessed by the discriminator for the “probability of being fake”

Let's play a game!

Cops or Robbers? Match the following

Generator

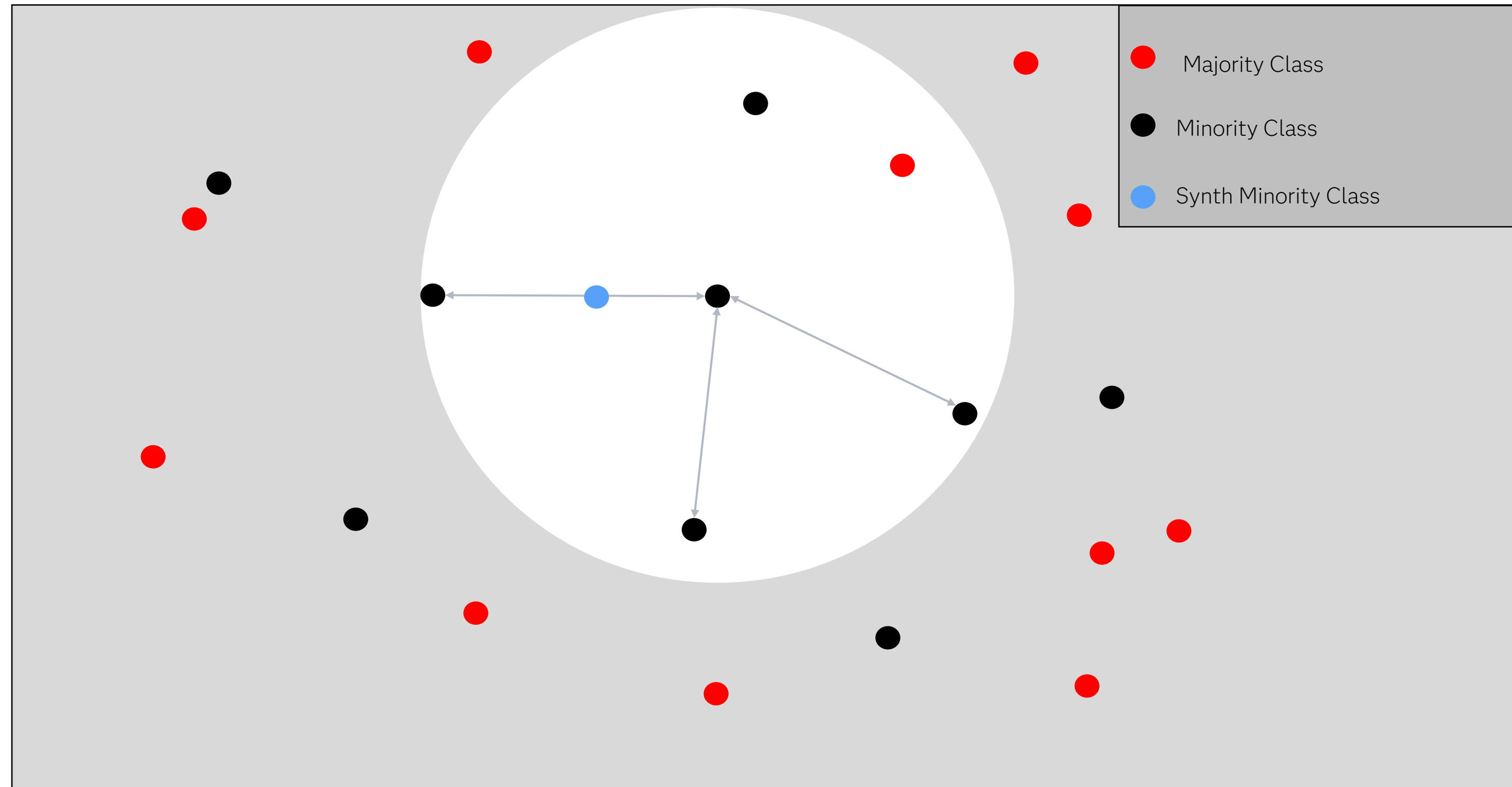
Discriminator



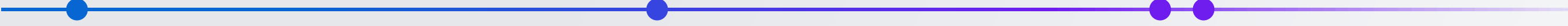
Cop? Robber? RobberCop? Copbler ?

SMOTE

Synthetic Minority Oversampling TEchnique

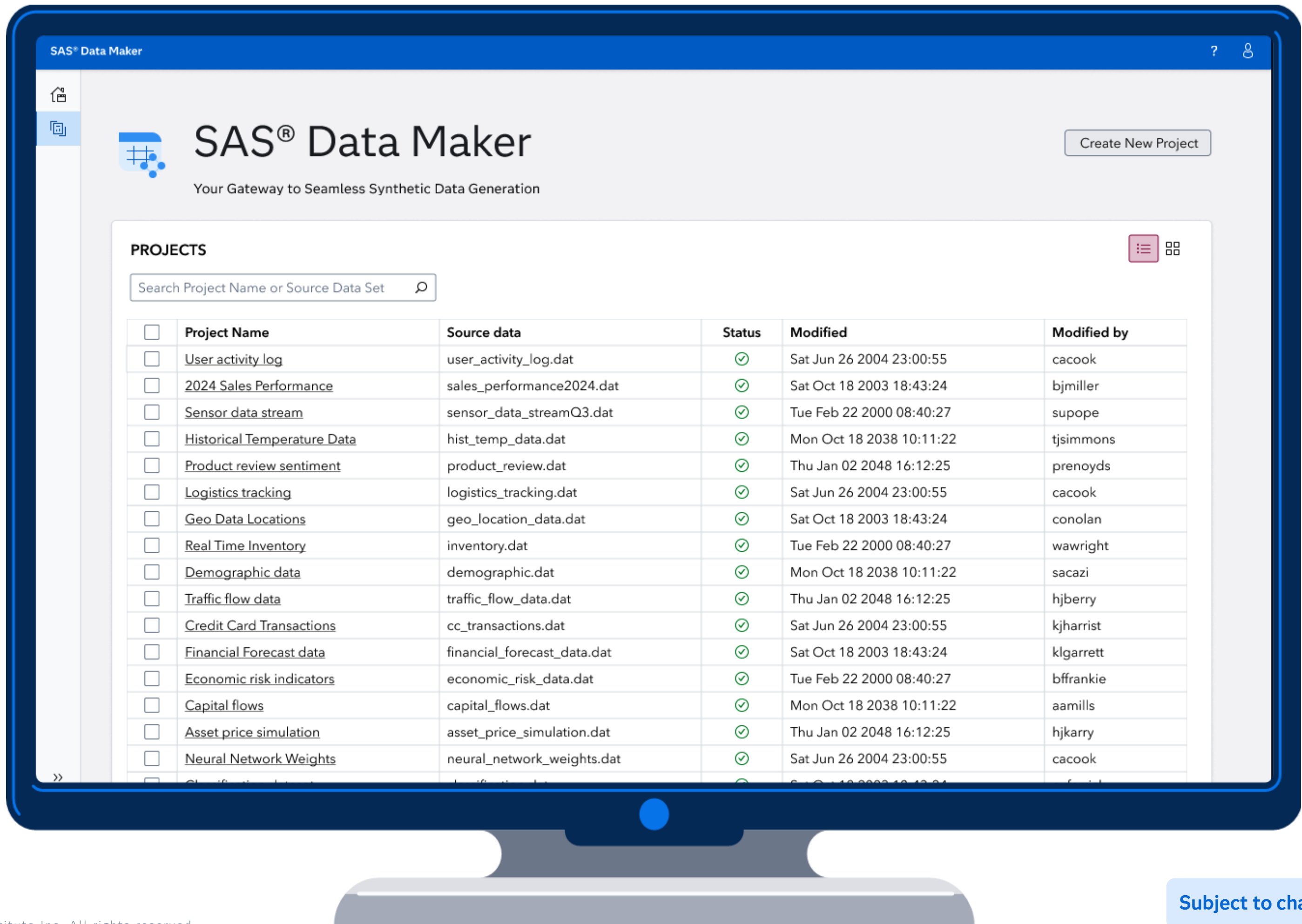


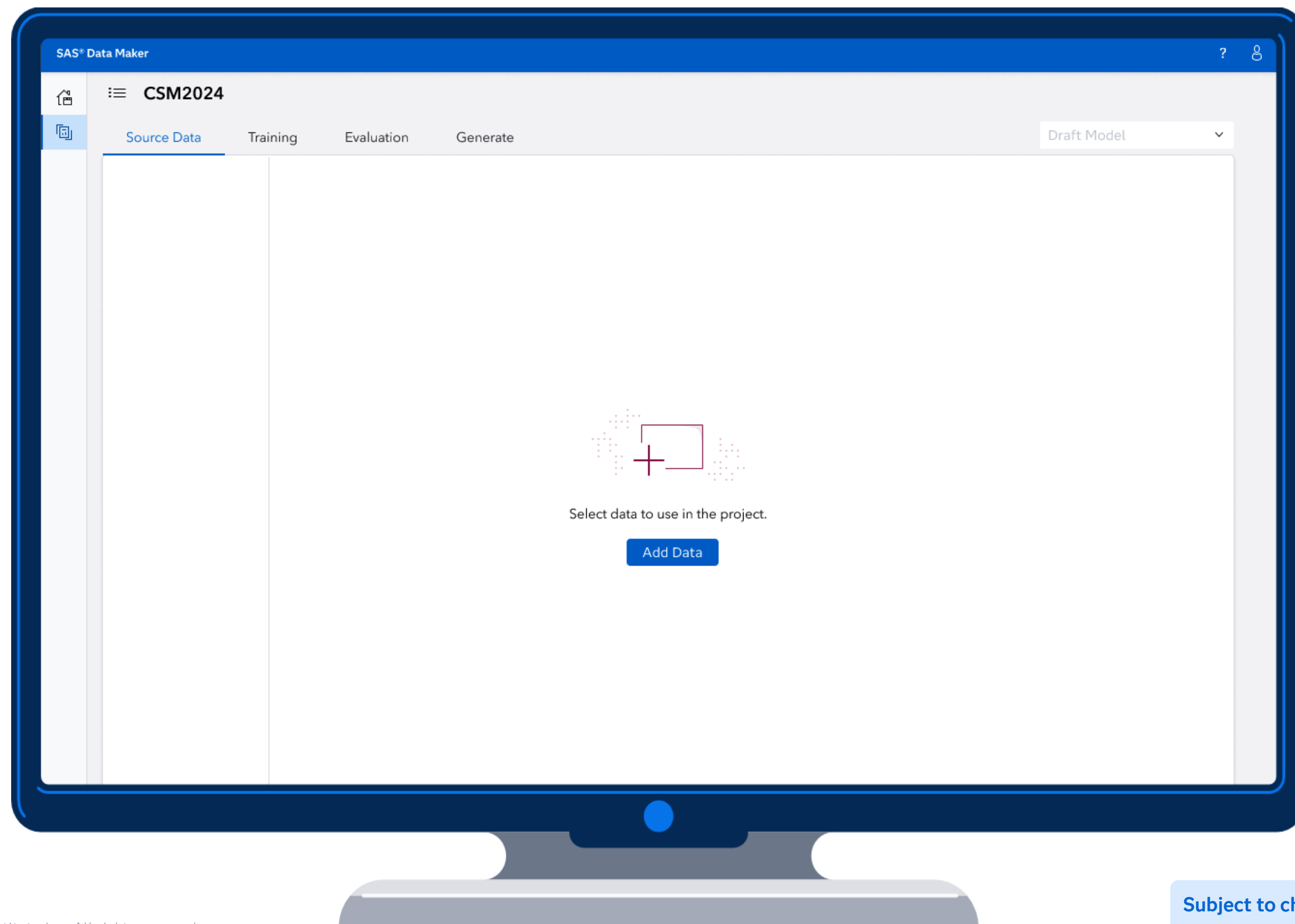
SAS Data Maker Demo

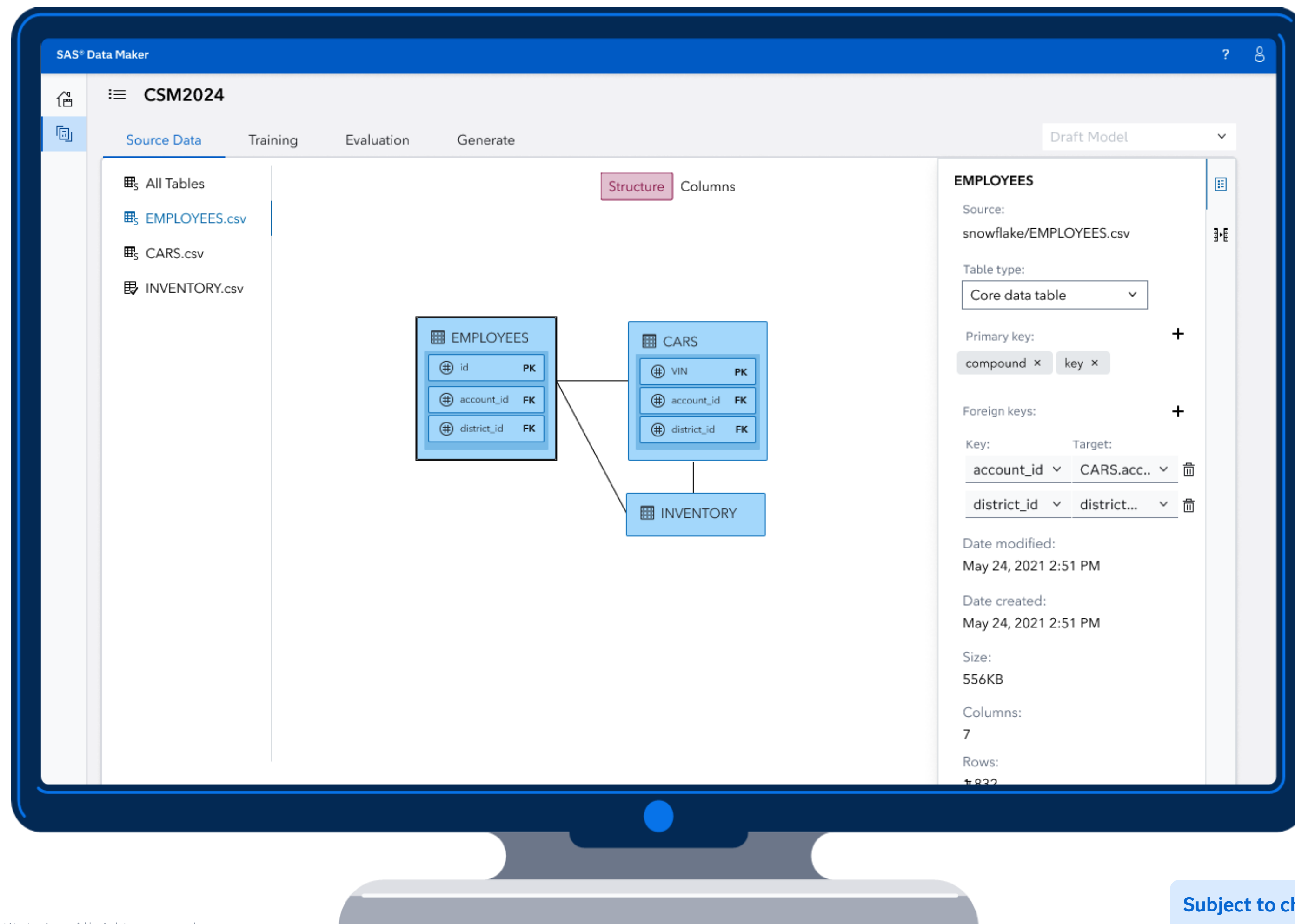


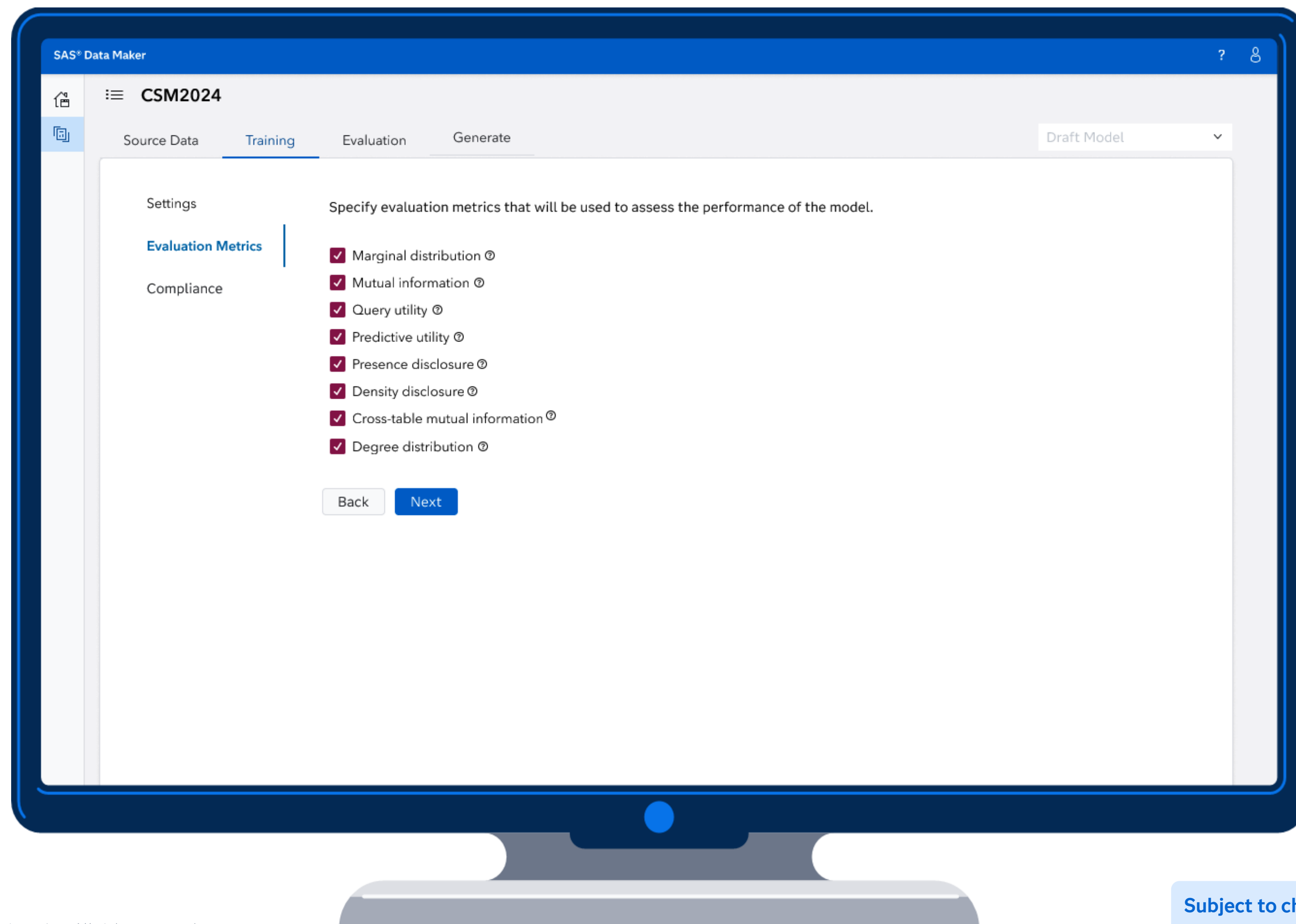
Subject to change

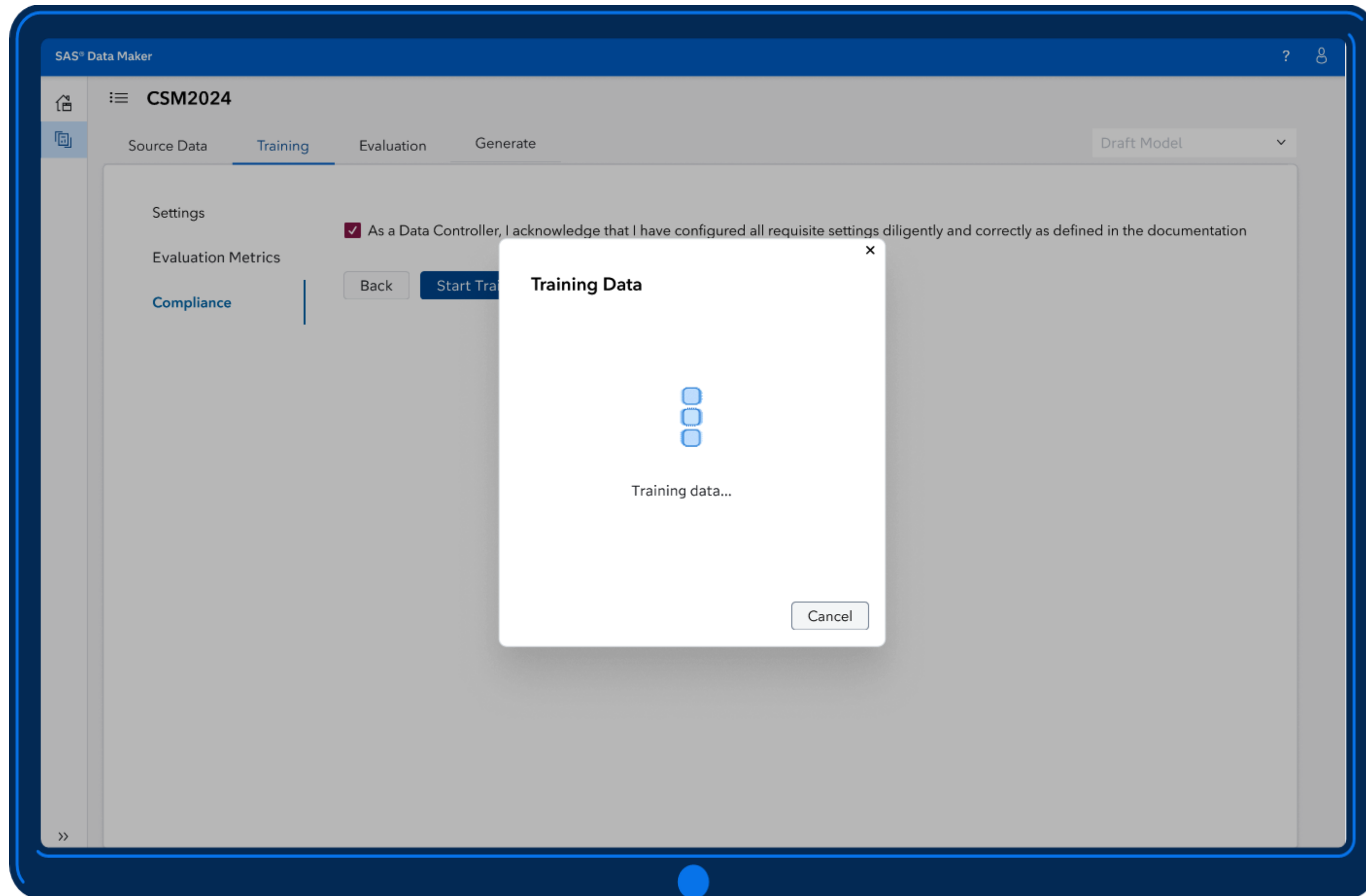
Video recording

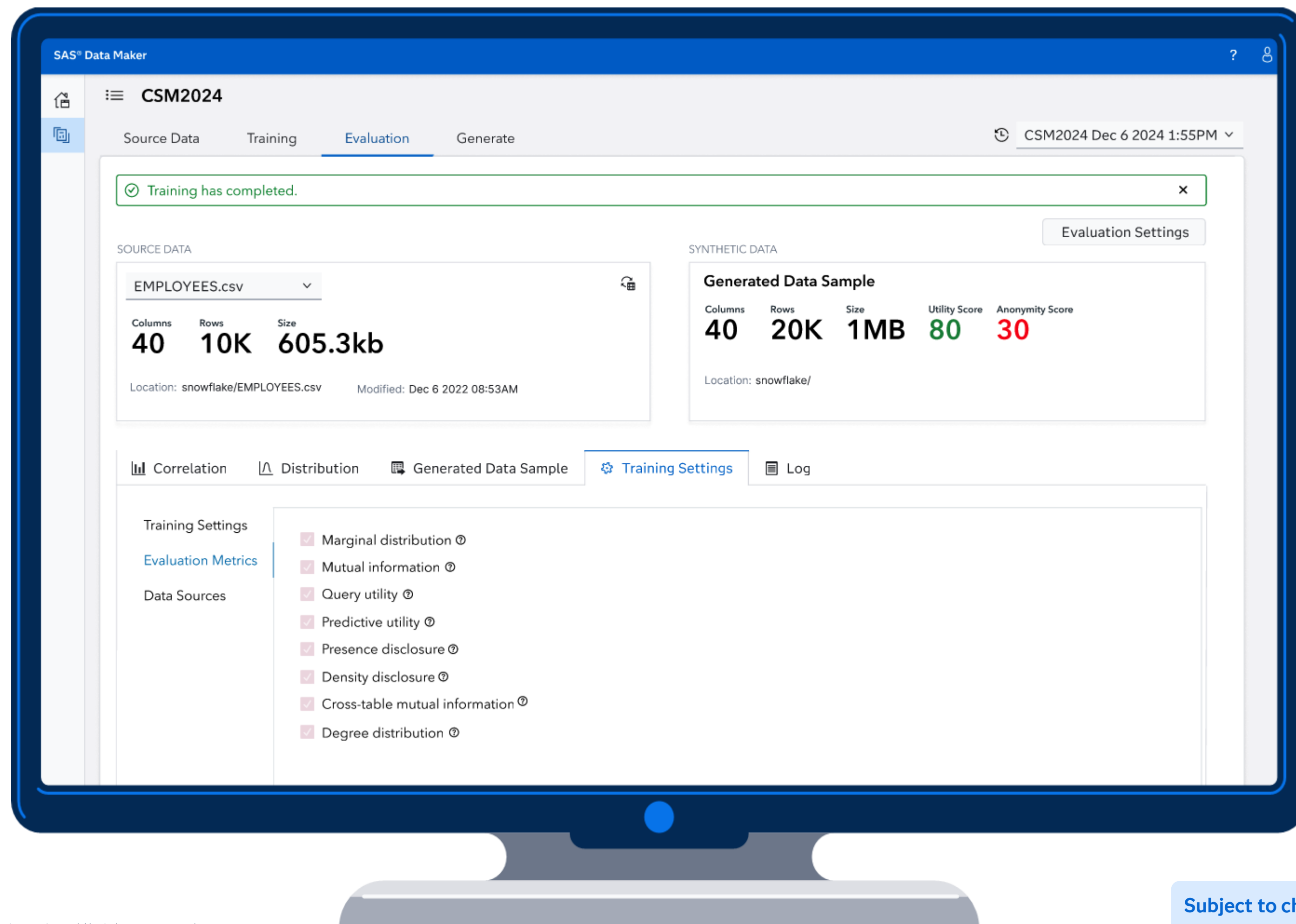


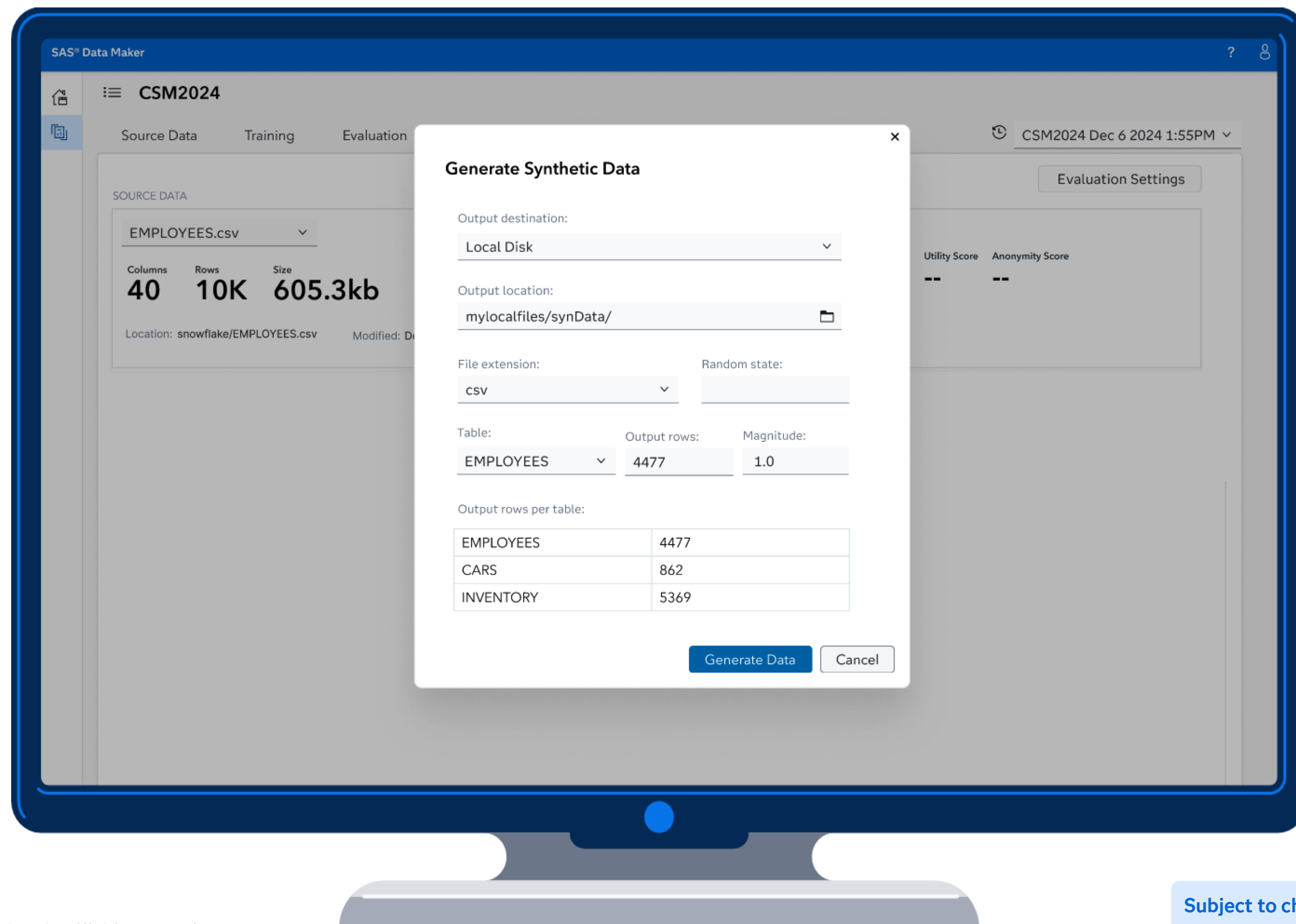


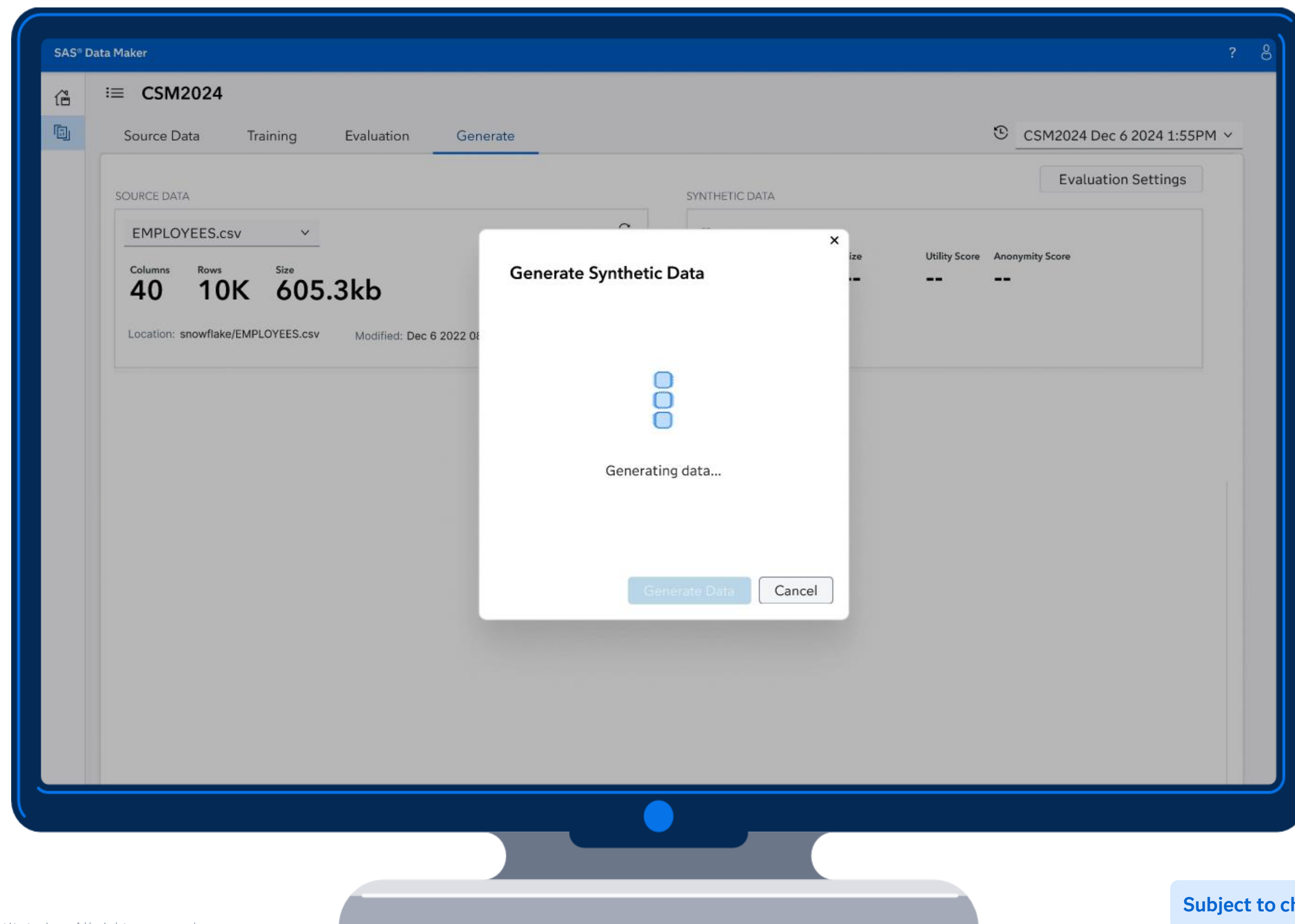


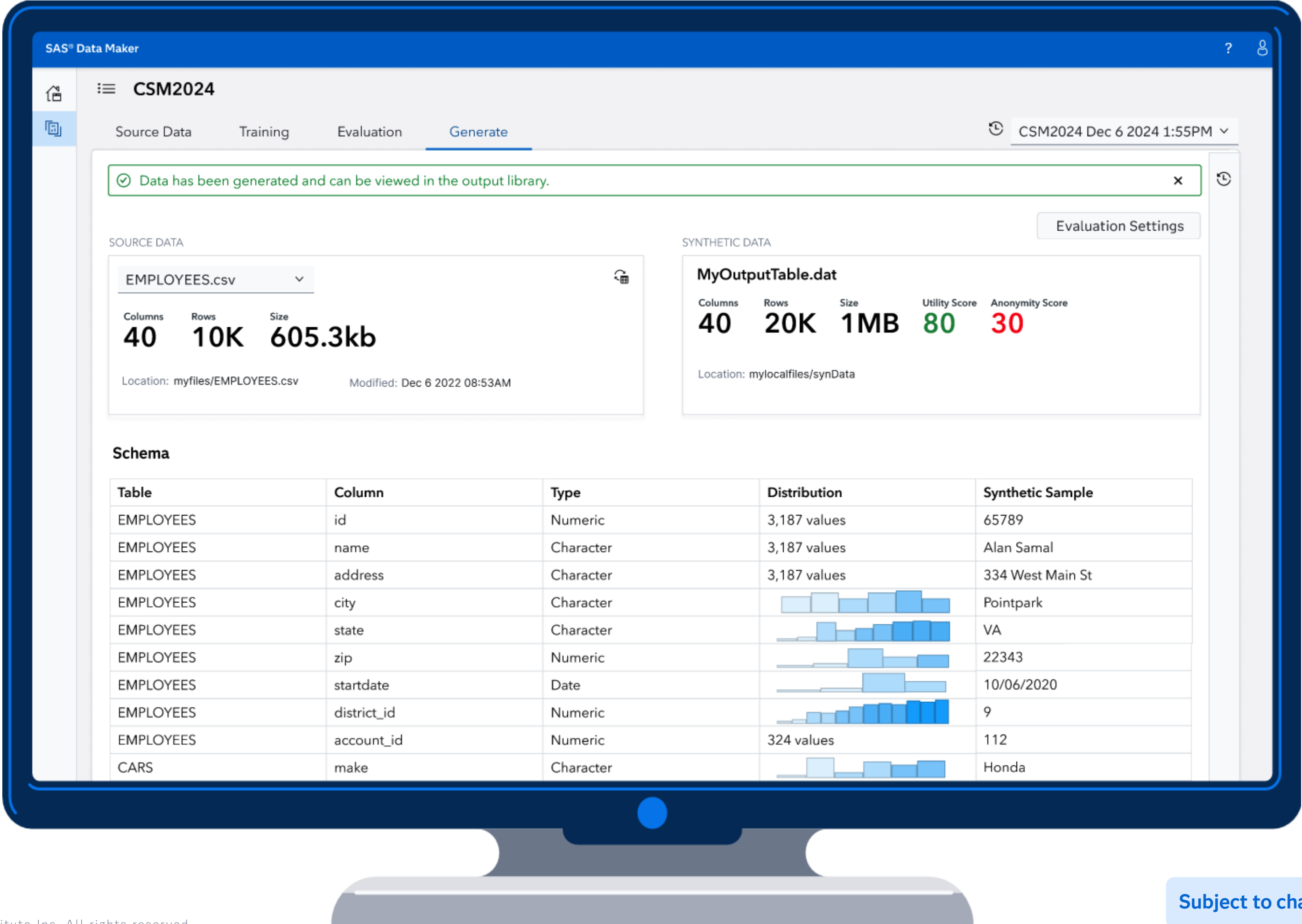












Conclusion



Synthetic data is not a replacement for real data.



Accelerate drug development and clinical trial simulations.



Enables research in areas with limited real-world data.



Enhances privacy protections for sensitive patient information.



Enables the early phases of research and development.



Final models and decisions should be validated and possibly fine-tuned on real data.



Robust validation, strict privacy guarantees, and a careful balance between utility and fidelity are crucial to avoid pitfalls.



Its success depends on responsible deployment, continuous empirical validation, and close integration with traditional data to ensure accuracy and safety in clinical applications.