

Challenges and Solutions for Generating Synthetic CDISC Data

Cat Briggs, Sundaresh Sankaran; Pritesh Desai

March 23rd, 2026

ET20 PHUSE US Connect 2026





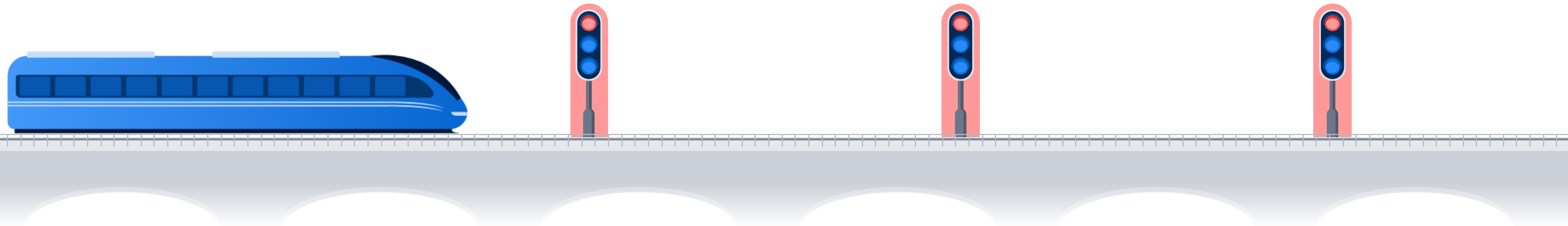
Agenda

- Current State
- Generation Challenges
- Process Paradigm
- Q & A

Serious data challenges impede organizations from adopting and succeeding with AI.

Data Privacy & Regulatory Compliance

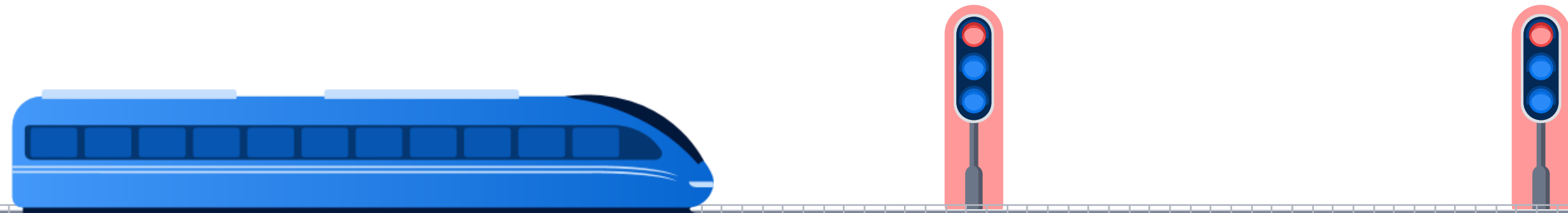
Highly regulated industries must adhere to strict data privacy laws, making sharing or using real-world data difficult.



Serious data challenges impede organizations from adopting and succeeding with AI.

Incomplete & Imbalanced Data

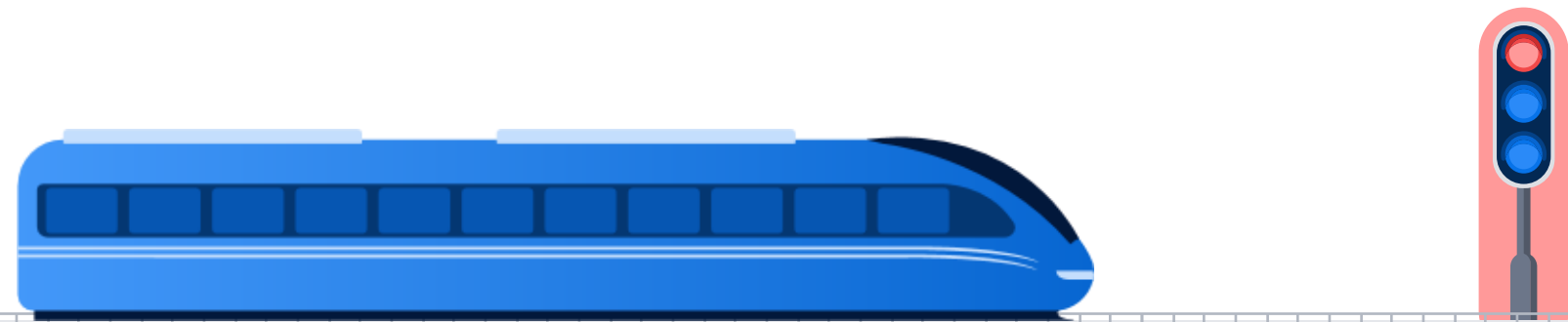
Real-world data often has gaps, missing fields, or imbalanced representations, creating biased or underperforming AI models.



Serious data challenges impede organizations from adopting and succeeding with AI.

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.



Serious data challenges impede organizations from adopting and succeeding with AI.

Difficulty in Data Sharing

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



Synthetic Data Generation

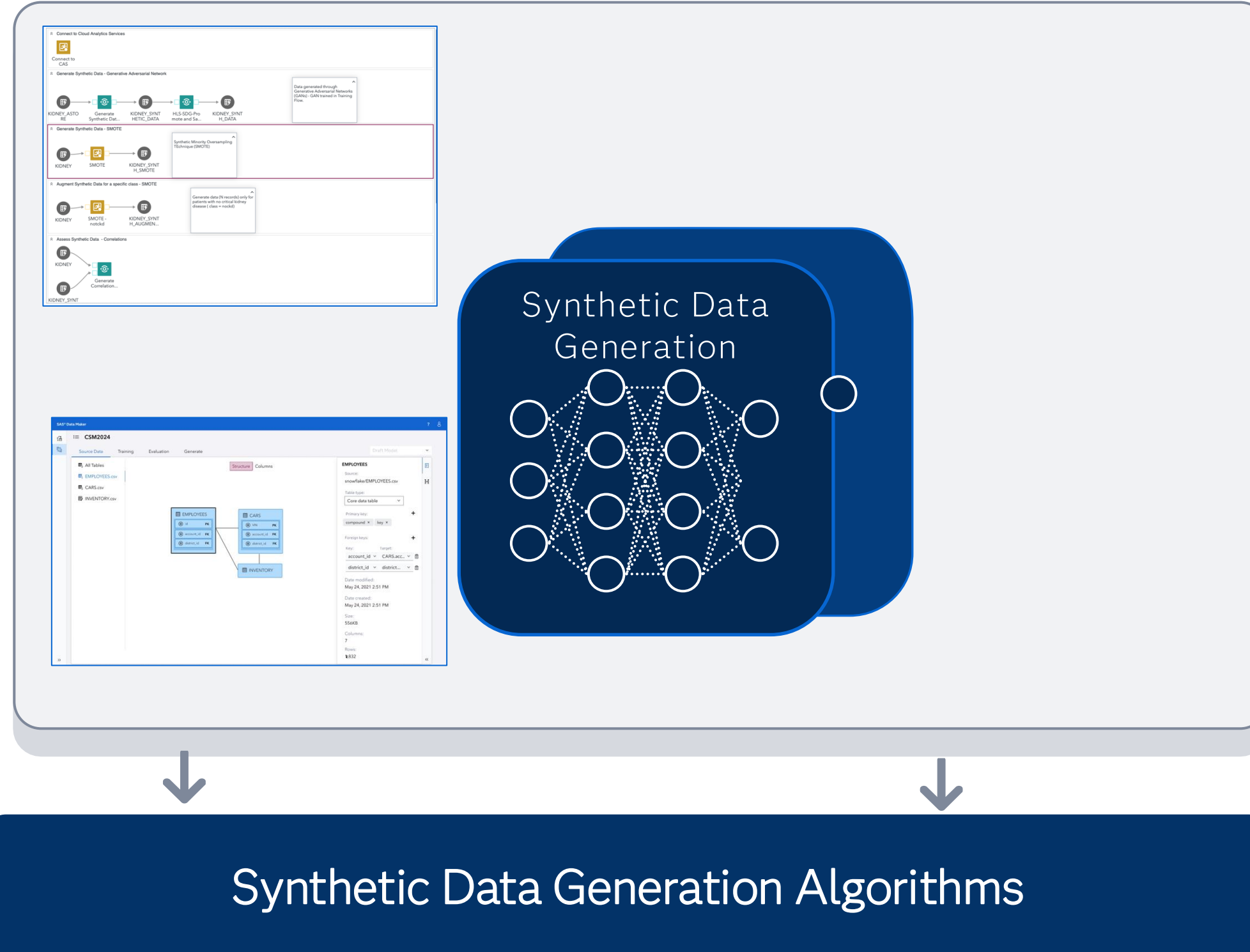
Enables Analytics Innovation

Restricted Data

Imbalanced Data

Rare Events

Insufficient Data



Synthetic Data

ID1

ID2

ID3

ID4

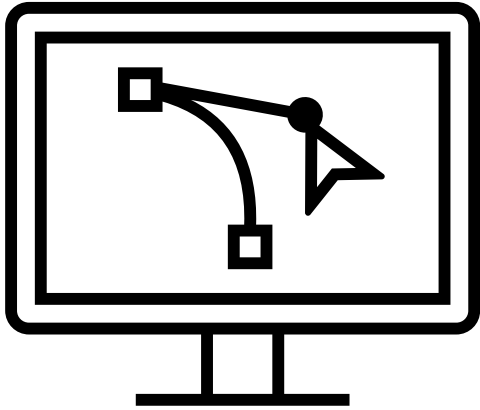
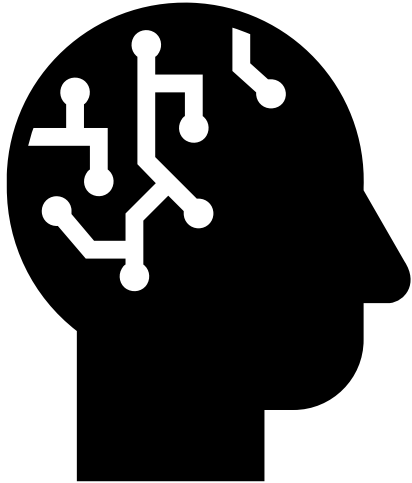
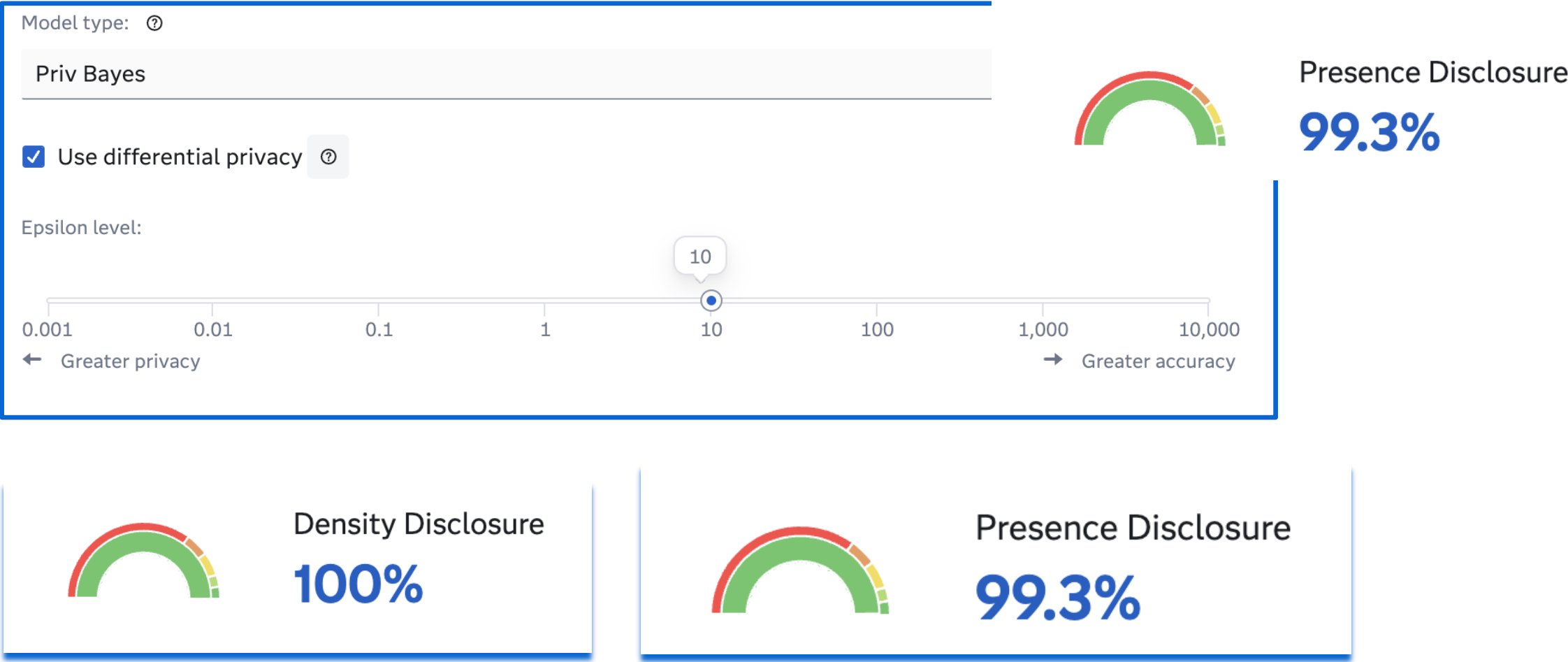
IDN

- Generate high quality synthetic data
- Validate the quality of the generated data
- 'Market' data downstream

Generation Challenges

Challenge 1: Patient Privacy

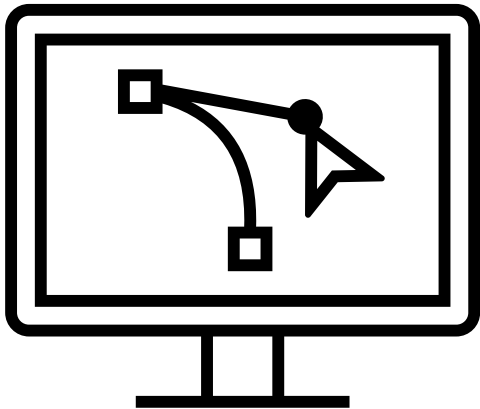
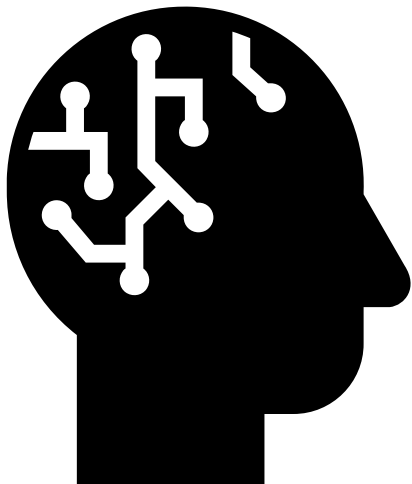
Problem: Need to remove any sensitive and patient information but still retain the structure and patterns of the real data



Challenge 2: Ensuring Internal Logic

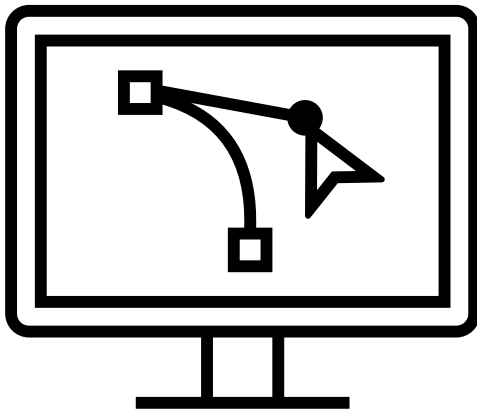
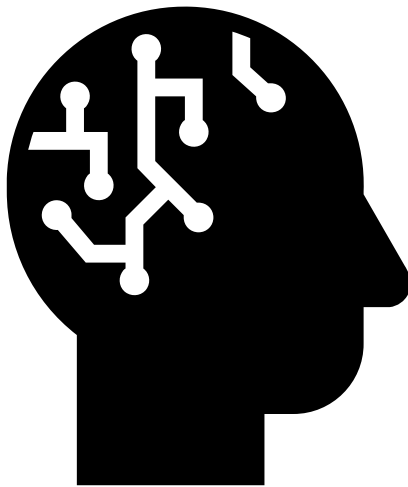
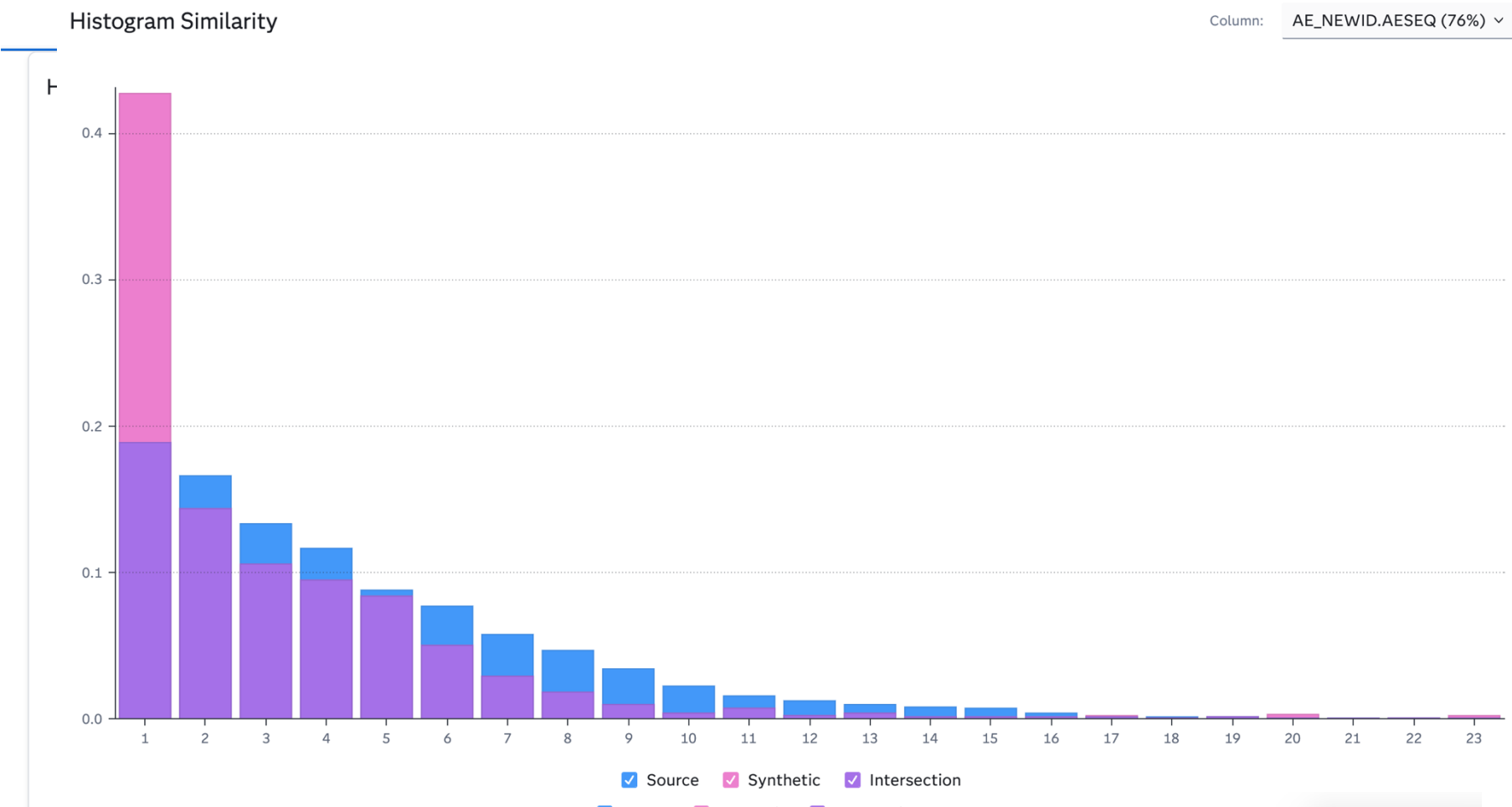
Problem: Without proper inputs into the algorithm , synthetic data may not match required internal logic

	Table of ARM by ARMCD					
	ARM	ARMCD				
		Pbo	Scrnfail	Xan_Hi	Xan_Lo	Total
Placek	Placebo	89	0	0	0	89
Scree	Screen Failure	0	72	0	0	72
Xanon	Xanomeline High Dose	0	0	66	0	66
Xanon	Xanomeline Low Dose	0	0	0	76	76
Total	Total	89	72	66	76	303



Challenge 3: Replicating Variable Distributions

Problem: Without proper control, synthetic data may not match required distributions



Implementation Techniques

Tabular data
simulating using
learned
distributions and
correlations

Integration of newer
generative AI models to
construct complex,
privacy-preserving
synthetic datasets for
domain-specific needs.

Differential privacy
algorithms to
ensure no sensitive
information is
leaked

Support for
synthetic
oversampling
algorithms, such as
SMOTE for
imbalanced data.

(Synthetic
Minority
Oversampling
Technique)

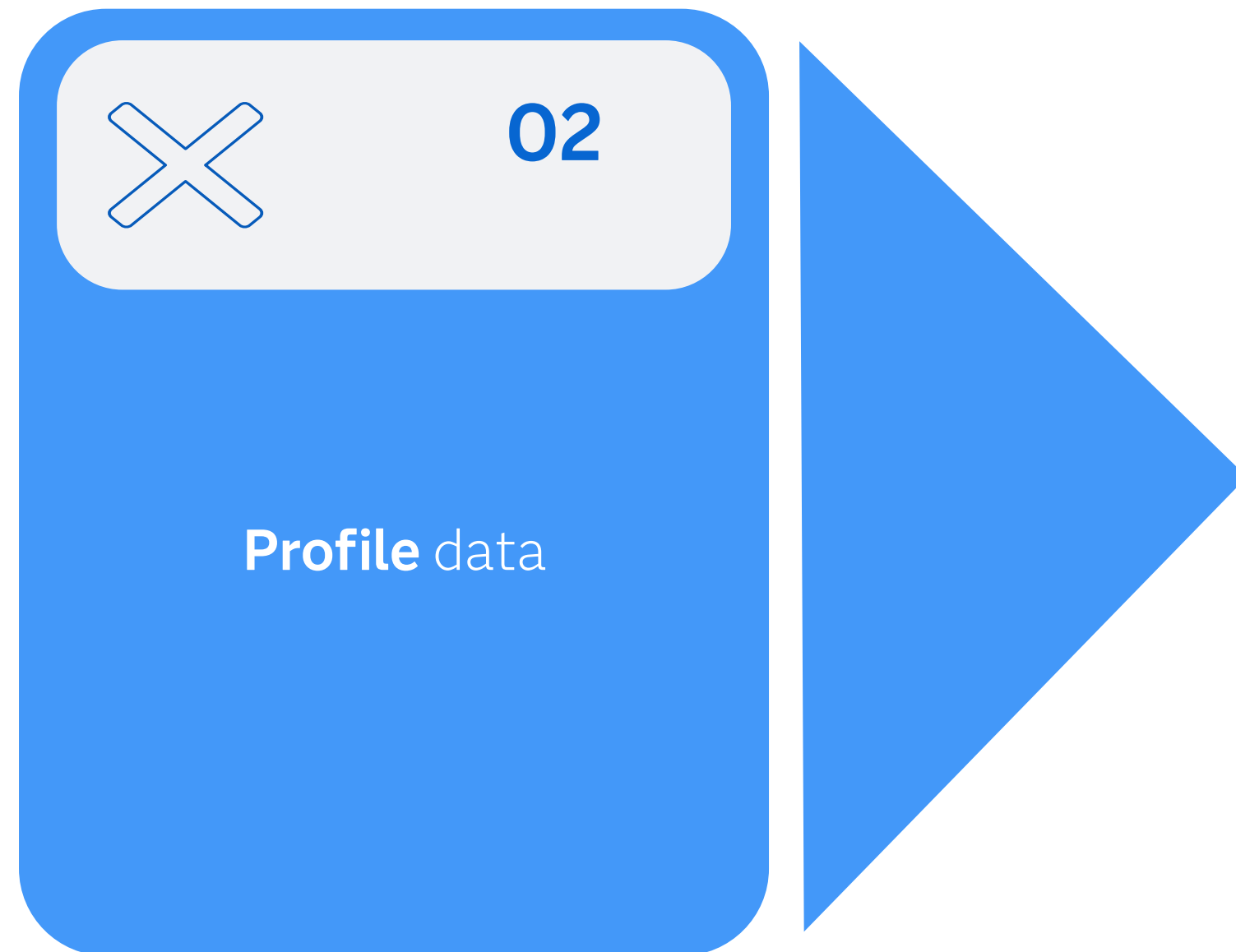
Process Paradigm

A Process Paradigm for Synthetic Data

Some common steps recommended for synthetic data generation



Profile Data



- Work with Data Expert
- Determine which variables need to be mimicked exactly
- Determine which variables have internal logic
- Determine data table type
- Map data and reference tables

Pre-process original data



- Work with Data Expert and data scientist
- Map data and reference tables
- Label variables types
- Indicate repeated information and linked variables

Configure Generation Parameters



04

Configure generation
parameters

- Work with Data Scientist
- Use models with differential privacy, if necessary
- Use proper bin sizes for sequential data modeling
- Set maximum number of categories to cover all variables
- Set seed for robust and reliable results

Train Generation Model

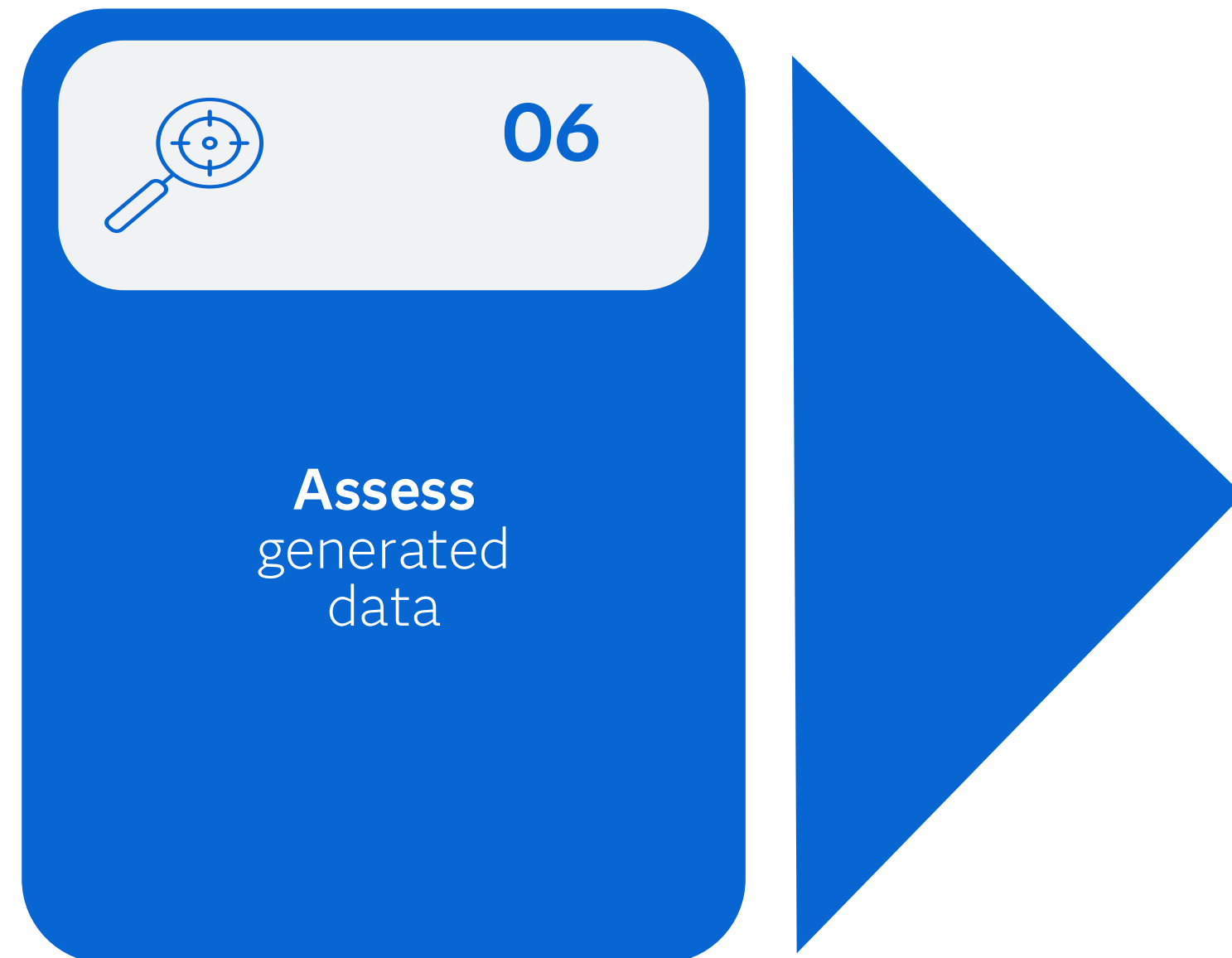


05

Train generation model

- Work with Data Scientist
- Monitor compute usage
- Watch for errors

Assess Generated Data



- Work with Data Expert and Scientist
- Check distributions from original and synthetic data
- Assure internal logic is maintained
- Determine if Density and Presence Disclosure metrics are within acceptable limits

A Process Paradigm for Synthetic Data

Some common steps recommended for synthetic data generation



Thank You

Catherine.Briggs@sas.com

Sundaresk.Sankaran@sas.com

Pritesh.Desai@sas.com

