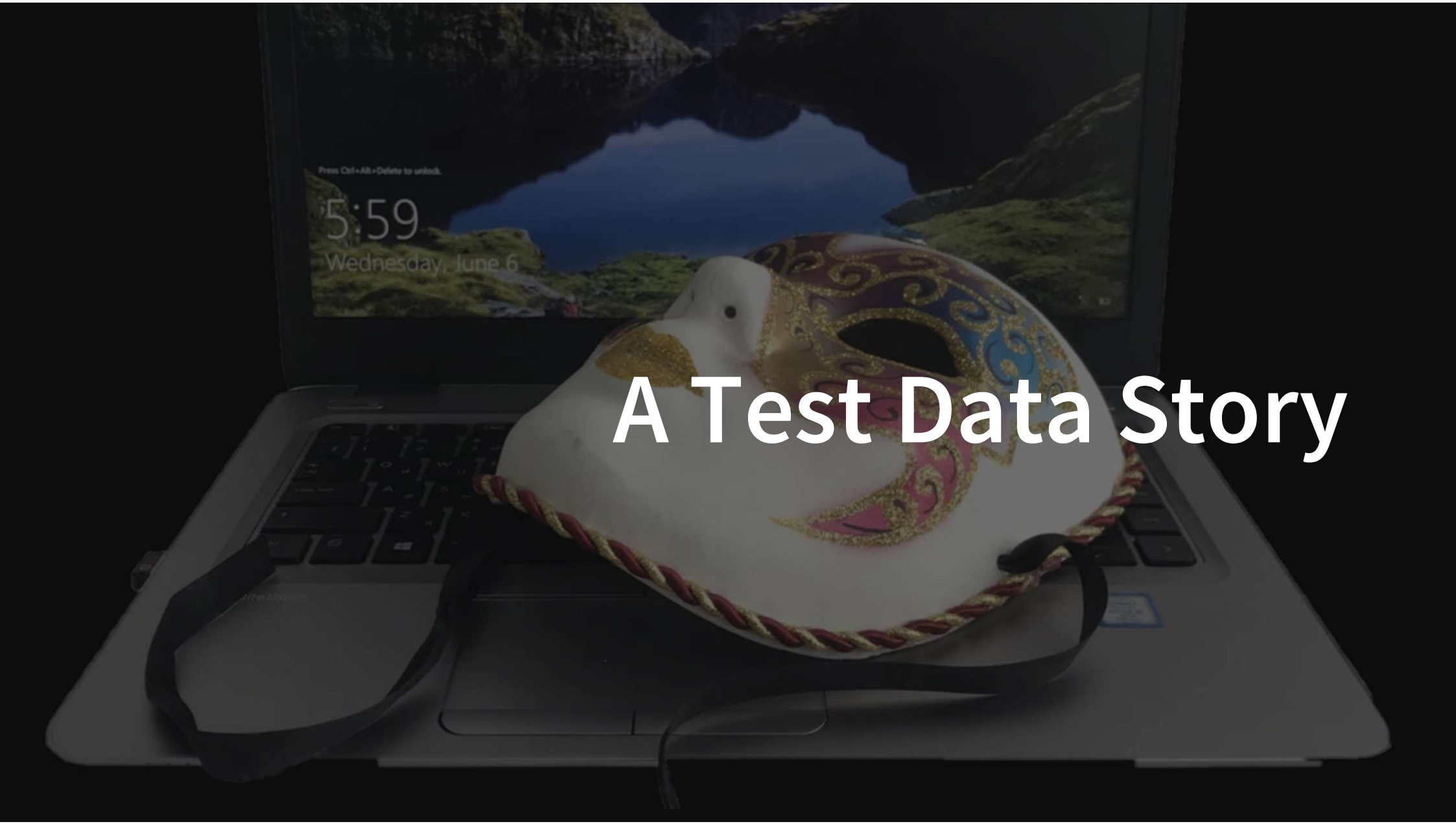
Practical Test Data Generation Adopting Deep Neural Networks

September 2021

Agenda

1. Intro
2. A test data story
3. From masking to machine learning
4. Beyond machine learning
5. Q&A



A Test Data Story

A TEST DATA STORY

Houston, we've had a problem!

You are going to lose access to production data in **3 months**!

1000+	10K+	300+
Related Objects	Fields	Personal and Sensitive Columns

Time to complete – **Yesterday**

Naïve data masking

EI*			Quasi Identifiers					SD*
SSN	First Name	Last Name	Race	Date of Birth	Admission Date	Gender	ZIP	Diagnosis
123-45-6789	John	Galt	African American	1/13/1934	4/10/2018	M	80011	Testicular Cancer
772-12-4636	Horacio	Oliveira	White American	11/1/1957	4/12/2018	M	80019	Broken Arm
NULL	Natalya	Rostova	White American	10/24/1974	5/1/2018	F	80011	Pneumonia
078-05-1120	Mary	Poppins	White American	5/5/1983	5/13/2018	F	80022	Pneumonia
288-99-1020	Veena	Apsara	Asian American	6/7/1951	5/22/2018	F	80019	Broken Arm
430-09-9291	Oskar	Matzerath	White American	7/4/2018	6/3/2018	M	80011	Brain Concussion

EI – Explicit Identifier **SD** – Sensitive Data

Naïve data masking

EI*			Quasi Identifiers				SD*
SSN	First Name	Last Name	Race	Age, Years	Admission Date	Gender	ZIP	Diagnosis
xxx-xx-xxxx	Alfred	Uncle	African American	84	Apr, 2016	M	80000	Testicular Cancer
xxx-xx-xxxx	Benito	Gonzales	White American	61	Apr, 2016	M	80000	Broken Arm
xxx-xx-xxxx	Nadezhda	Zhdanova	White American	43	May, 2016	F	80000	Pneumonia
xxx-xx-xxxx	Deborah	Jonze	White American	35	May, 2016	F	80000	Pneumonia
xxx-xx-xxxx	Mary	Pereira	Asian American	67	May, 2016	F	80000	Broken Arm
xxx-xx-xxxx	Alex	Heilmann	White American	0	Jun, 2016	M	80000	Brain Concussion

Suppression | **Random Substitution** | **Generalization**

EI – Explicit Identifier **SD** – Sensitive Data

Naïve data masking

EI* | Quasi Identifiers | SD*

SSN	First Name	Last Name	Race	Age Group	Admission Date	Gender	ZIP	Diagnosis
xxx-xx-xxxx			Any	Senior	2016	M	80000	Testicular Cancer
xxx-xx-xxxx			Any	Adult	2016	M	80000	Broken Arm
xxx-xx-xxxx			Any	Adult	2016	F	80000	Pneumonia
xxx-xx-xxxx			Any	Adult	2016	F	80000	Pneumonia
xxx-xx-xxxx			Any	Adult	2016	F	80000	Broken Arm
xxx-xx-xxxx			Any	Infant	2016	M	80000	Brain Concussion

Outliers!

More Suppression

More Generalization

EI – Explicit Identifier **SD** – Sensitive Data

Rules-based generation

Let get the data generated!

1000+	10K+	=	10K+
Related Objects	Fields		Columns to generate

Time to complete – **Never**

Age of Machine Learning and Artificial Intelligence

More variability is better

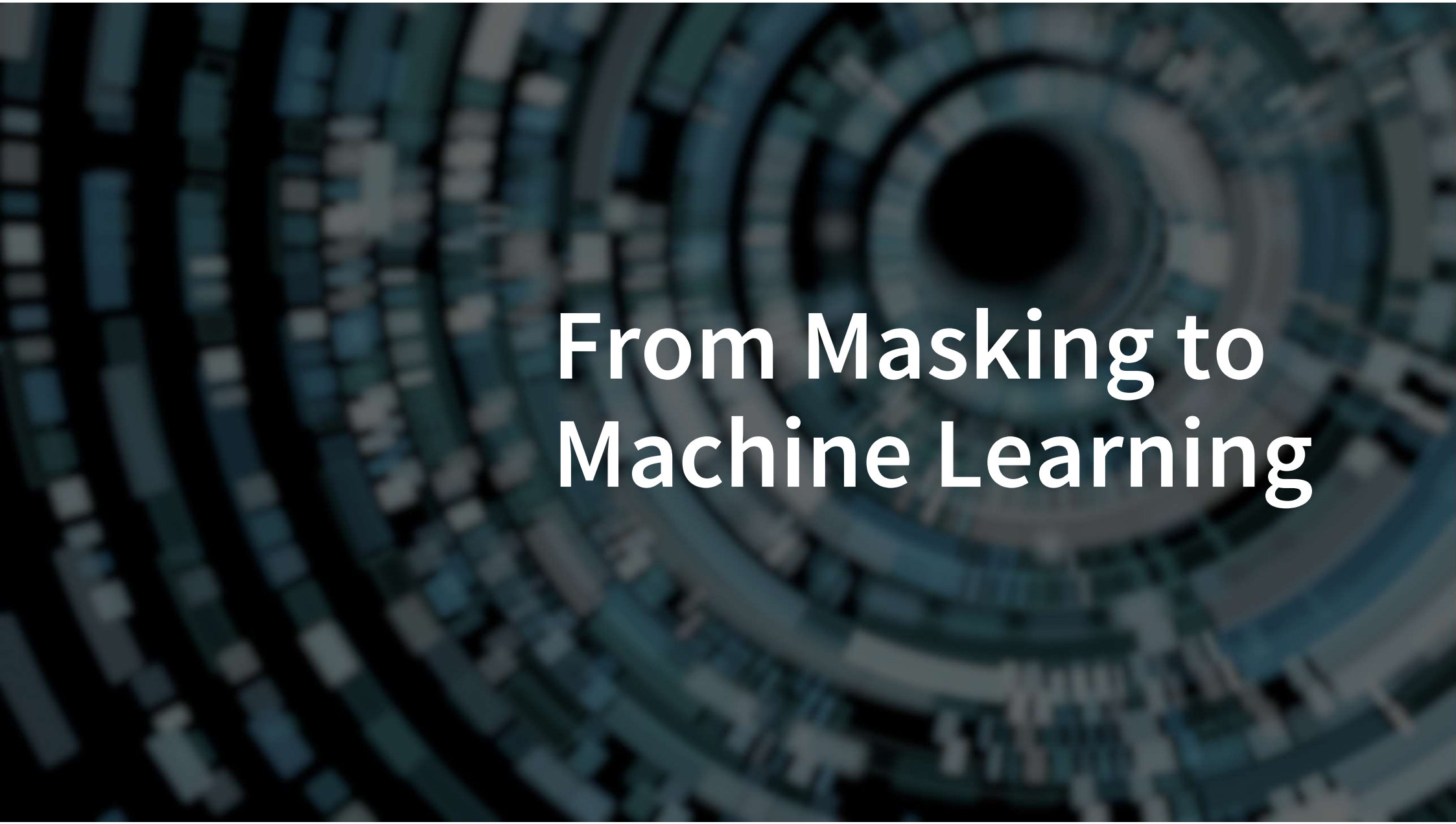
Not-OK data for better test coverage

Get more data than I already have

Data for testing ML/AI projects

Any sufficiently advanced technology is indistinguishable from magic

Arthur C. Clarke

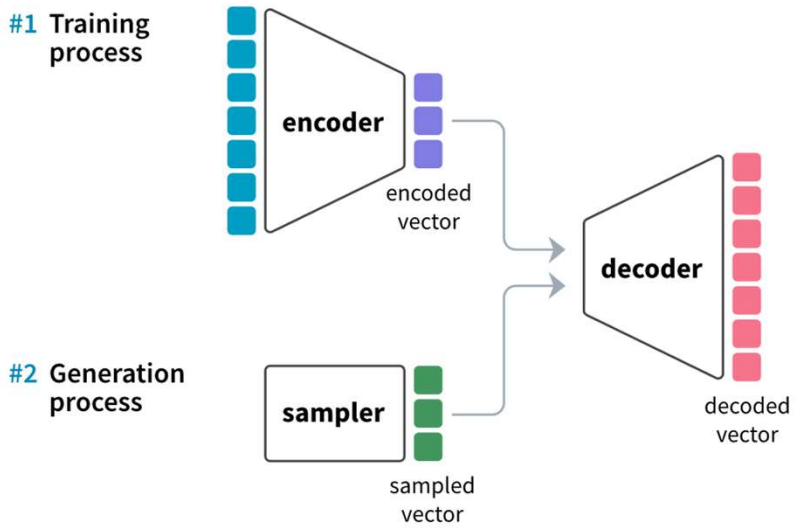

From Masking to
Machine Learning

NEURAL NETWORK MODEL

ML-based synthetic data generation

1. Neural networks: Variational Autoencoder (VAE), Generative Adversarial Network (GAN) and others
2. Automated process: computational resources, not people.
3. Different data types supported (categories, numeric values, free text).
4. Utility or privacy?
5. Customization is possible.
6. Various use cases:
   - generation of test data
   - data for analysis
   - data for ML engineers
   - …

Variational autoencoder scheme



SYNTHETIC DATA EXAMPLES

Original vs. synthetic

	ORIGINAL DATA	DATA GENERATED BY NEURAL NETWORK
Attribute	**Value**	**Value**
Age	39	58
Workclass	State-gov	Self-emp-not-inc
Final weight	77516	134436
Education	Bachelors	Bachelors
Marital status	Never-married	Married-civ-spouse
Occupation	Adm-clerical	Craft-repair
Relationship	Not-in-family	Husband
Race	White	White
Sex	Male	Male
Hours per week	40	34
Native country	United-States	United-States
Capital	2174	3277
Income	<=50K	>50K

Data patterns preservation

UUID

Original	Synthetic
fe52c68a-48ec-4837-a24c-4c2bd191431f	ae92a2d1f-d233-45bb-8415-dc7b84b72477
d19953f6-1f53-4b54-8fa4-d7374ca0607b	ea4c30c0d-dc4d-45a2-ac29-3cd8877516dd
3aea0026-faae-4845-8708-9bf323ff8a77	0-7114383-fe4c-47e5-9079-33463a4b9c16

Address

Original	Synthetic
1108 ROSS CLARK CIRCLE	471O M EOASHA BS E ORT
145 NEWCOMB AVENUE	33 OXNITTHOH2 X1VT
2540 EAST ST	150 1BETM EY STA H TNNEBE

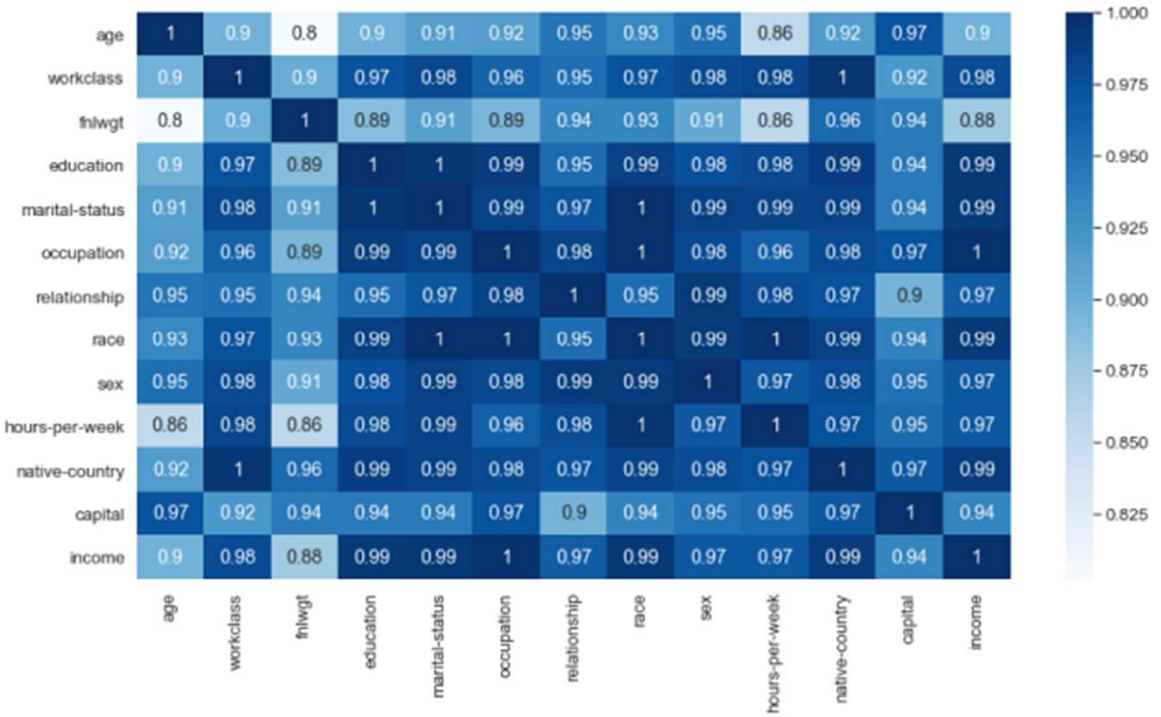
Serial number

Original	Synthetic
S000028330	S000023777
S000026892	S000000850
S000005529	S000043975

Is new synthetic data useful?

Median accuracy: 0.969

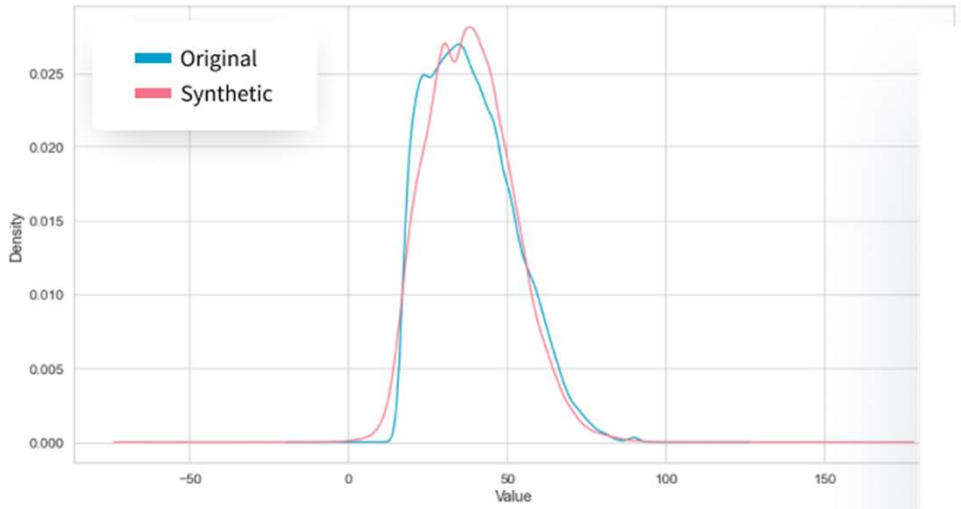
Accuracy: original ~ synthetic



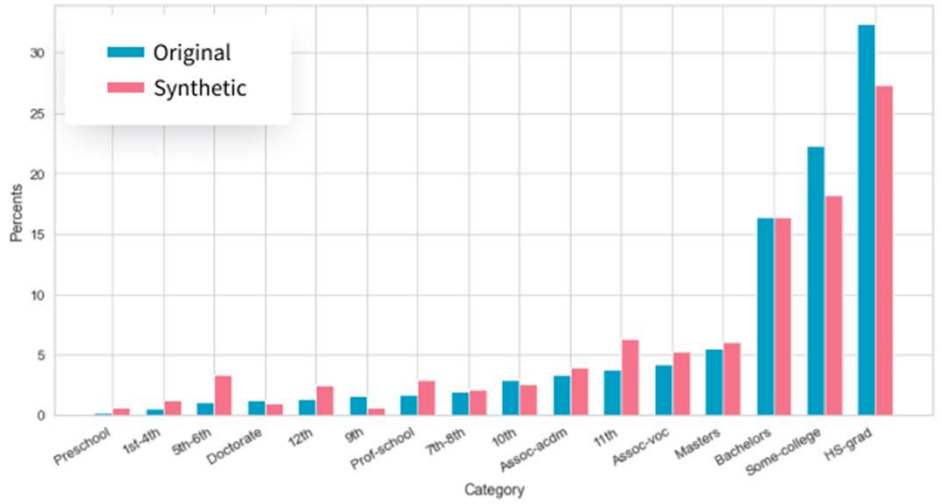
ACCURACY METRICS

Univariate distributions

Age

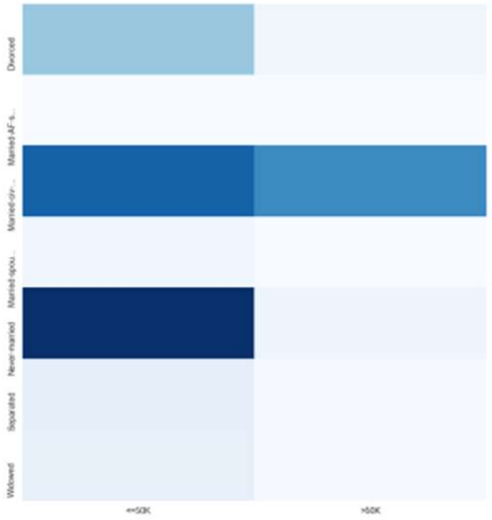


Education

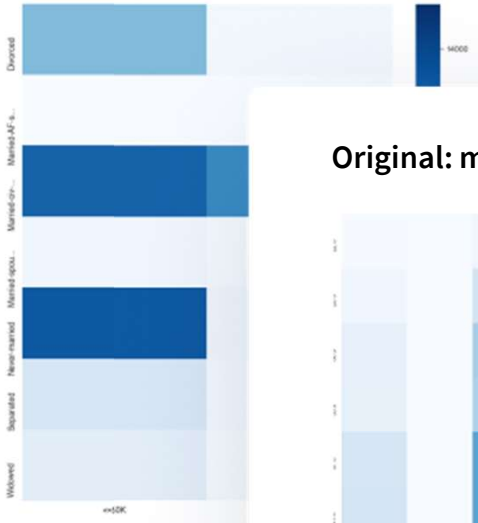


Bivariate distributions

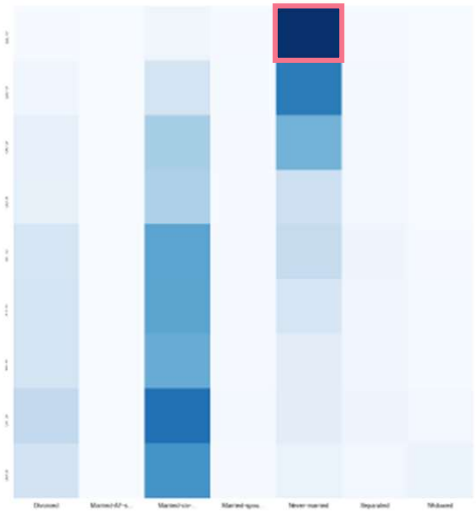
Original: age vs marital-status



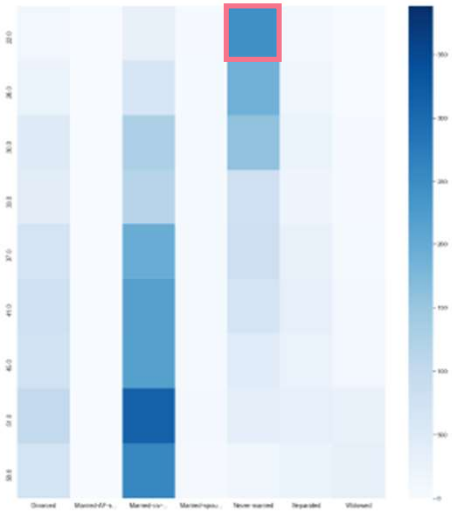
Synthetic: age vs marital-status



Original: marital-status vs income



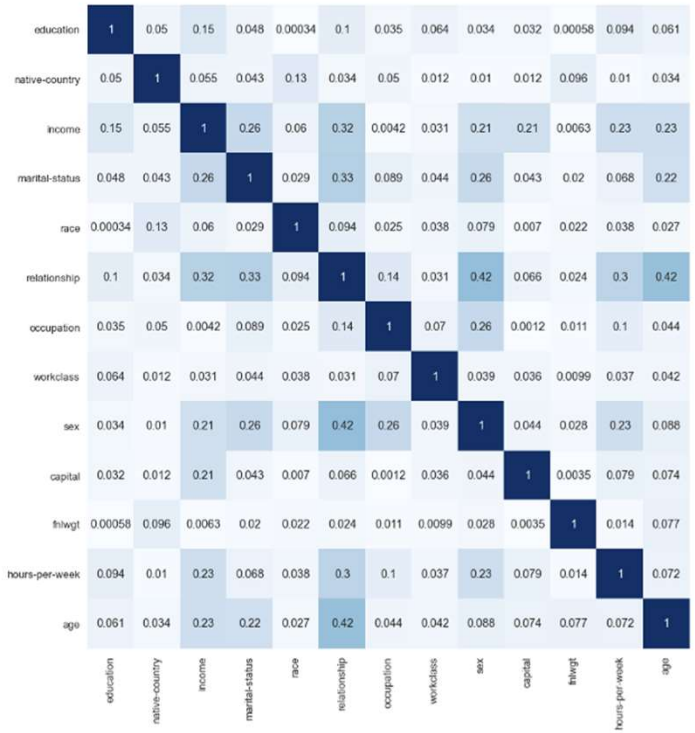
Synthetic: marital-status vs income



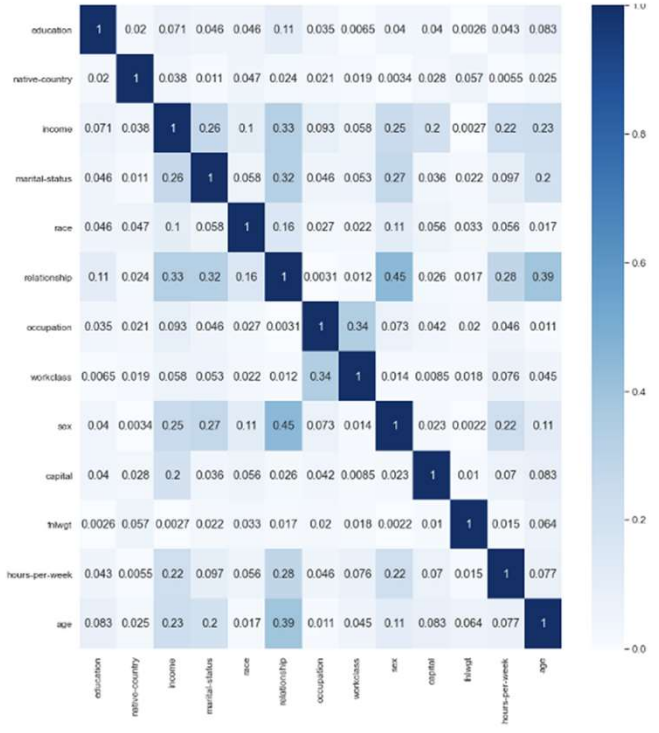
ACCURACY METRICS

Correlations between columns

Original



Synthetic

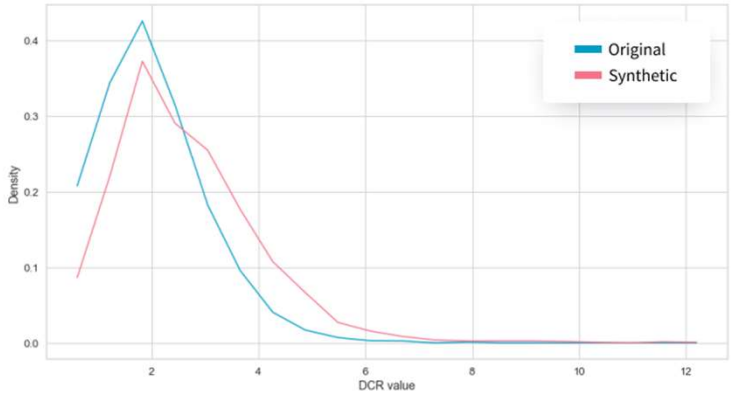


PRIVACY METRICS

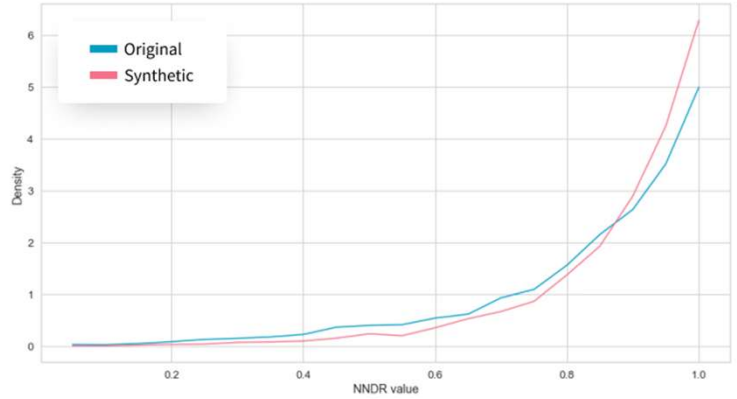
Is new synthetic data private?

1. Distance-based metrics (comparing original hold-out and generated data):
   - Full matches
   - Distance to closest record (DCR)
   - Nearest neighbor distance ratio (NNDR)
2. ML-based (training an attacker model)

Passed DCR test



Passed NNDR test

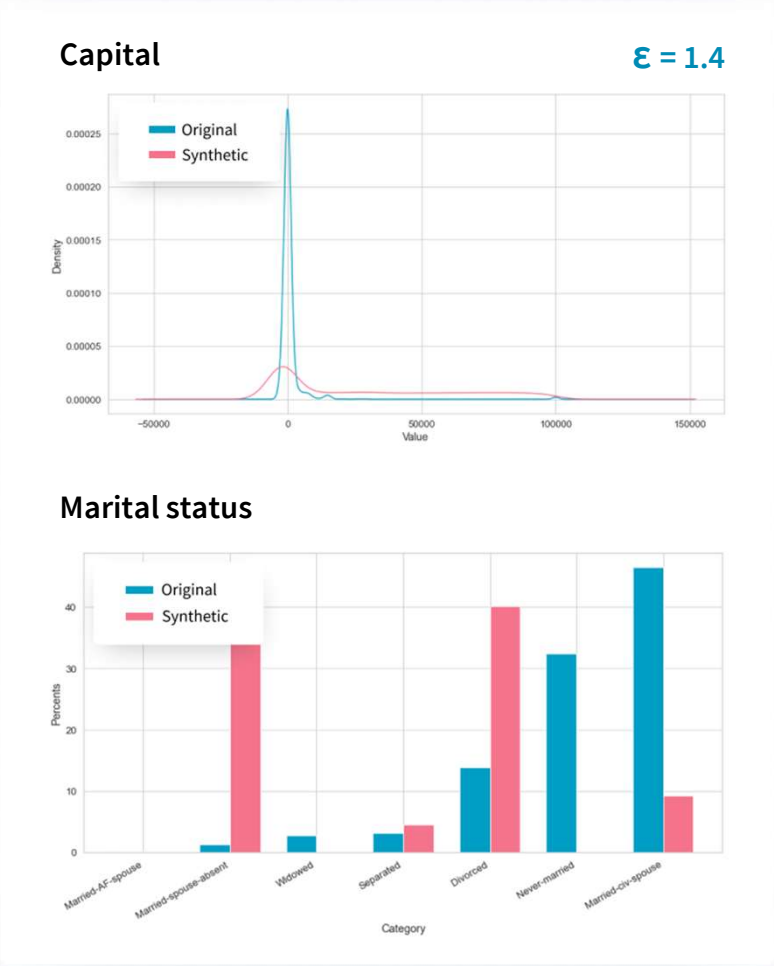


Need more privacy?

Differential privacy – a methodology of making computations private with quantifiable and provable worst case scenario leakage of the original data.

Variational Autoencoder with differentially private training process: **DP-SGD**

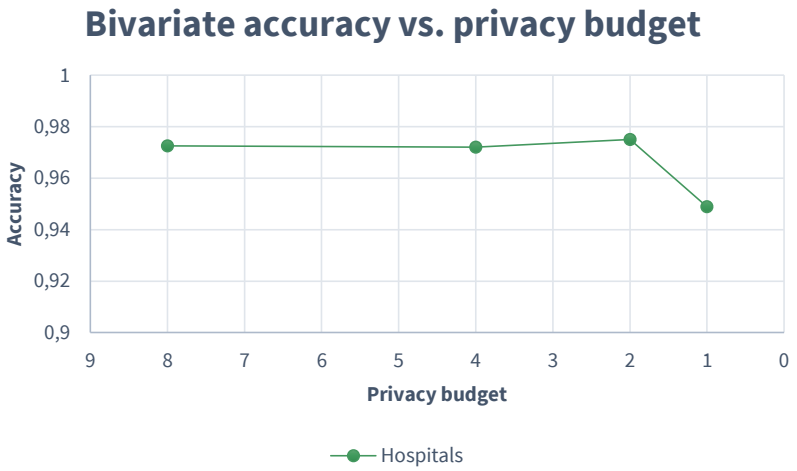
Privacy level can be controlled with parameter **ε** (privacy budget).

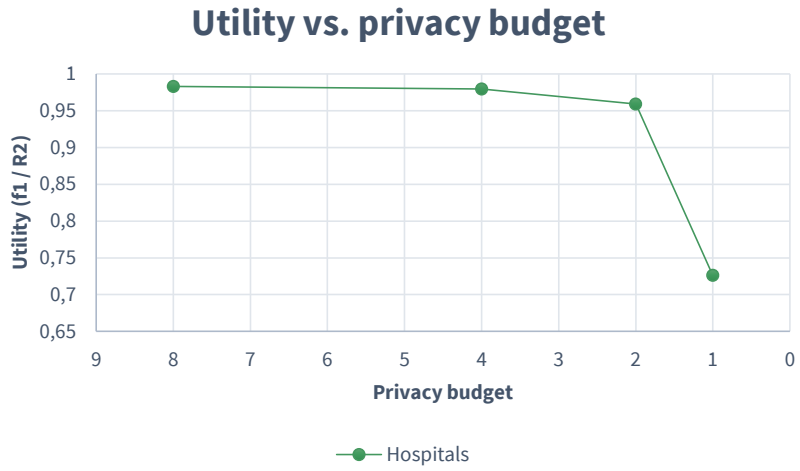


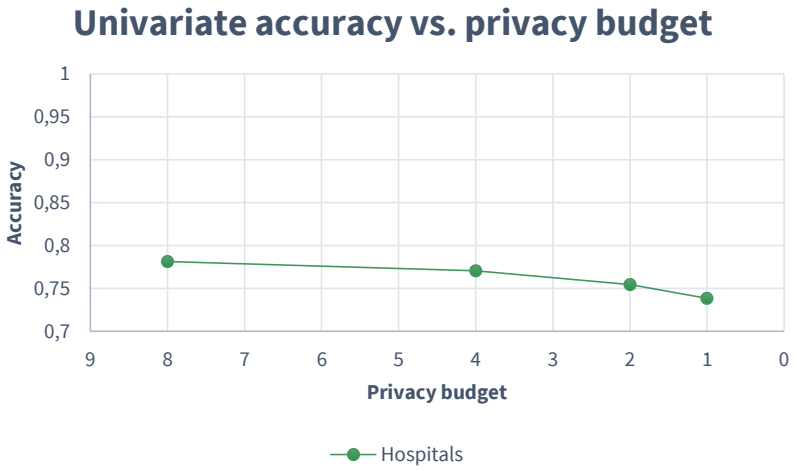
What is the utility/privacy exchange rate?

The experiments show that

- Considerable data utility drop happens for privacy budgets less than 2
- In some cases, the utility of the generated data might be higher than the original







REQUIRED RESOURCES

What do I need to get the synthetic data?

Dataset with hospital data:

- ~50000 rows
- 5 free text columns
- 7 numeric columns
- 7 categorical columns

Dataset with us census income data:

- ~50000 rows
- 4 numeric columns
- 9 categorical columns

Google Cloud VM shape: n1-standard-16 (16 vCPU, 60 GB RAM)

Accelerator: 1x NVIDIA TESLA P100

Hourly rate: $2.27

Dataset	Epochs	Rows	Time	Minute/Epoch	Cost
Hospital: complications and deaths	200	50000	22h 1m	~6m 25s	$50
US census income	200	50000	7h 53m	~2m 20s	$18

What do I need to get the synthetic data?

Machine specifications:

- **CPU:** Intel(R) Core(TM) i5-10310U CPU @ 1.70GHz, 2208 Mhz, 4 Core(s), 8 Logical Processor(s)
- **RAM:** 32 GB

Hospital: complications and deaths:

- 5 free text columns
- 7 numeric columns
- 7 categorical columns

Rows	Size (KB)	RAM	Time	Rows/s
1,000	320	400	1m 2s	31
10,000	3200	630	5m 12s	32
100,000	32000	2200	1h 20m	21

US census income:

- 4 numeric columns
- 9 categorical columns

Rows	Size (KB)	RAM	Time	Rows/s
1,000	112	400	1s	1000
10,000	1097	400	1s	10000
100,000	11220	400	4s	25000

Training of synthetic data generation models

Resources required for training **scale liner** and depend on:

- Size of the dataset – number of rows and columns
- Types of the columns: texts, categorical, numeric
- Number of epochs
- Differential privacy settings

Training specs	
Platform	AWS SageMaker
Shape	ml.g4dn.xlarge
Cost, USD/hour	0.74

Number of rows	10000	5000	10000	5000	500
Differential privacy	Y	Y	N	N	N
Time, hours	3	1.5	0.2	0.08	0.06
Cost for 20 epochs, USD	**2.22**	**1.11**	**0.12**	**0.03**	**0.02**
Cost per epoch	0.11	0.06	0.006	0.0015	0.001
Estimated cost for 50 epochs	5. 5	3	0.3	0.075	0.05

Data set	
Churn_modelling.csv	
Columns	14
Text	1
Numeric	6
Categorical	7

Example:

The cost of a model training on the dataset with **100K** rows and **15** columns is **2.22 * 10 = 22 USD** in AWS SageMaker

Note: Prices for the regular GPU-enabled VMs are lower but require careful planning of the workload. We recommend using on-spot instances for non-time-critical calculations.

Synthetic data generation estimations

Resources required for generation **scale liner** and depend on:

- Size of the dataset – number of rows and columns
- Types of the columns: texts, categorical, numeric

Generation specs	
Platform	AWS SageMaker
Shape	ml.m5.large
Cost, USD/hour	0.12

Number of Rows	5000	2000	500
Time, hours	0.1	0.08	0.08
Cost, USD	**0.012**	**0.0096**	**0.0096**

Data set	
Churn_modelling.csv	
Columns	14
Text	1
Numeric	6
Categorical	7

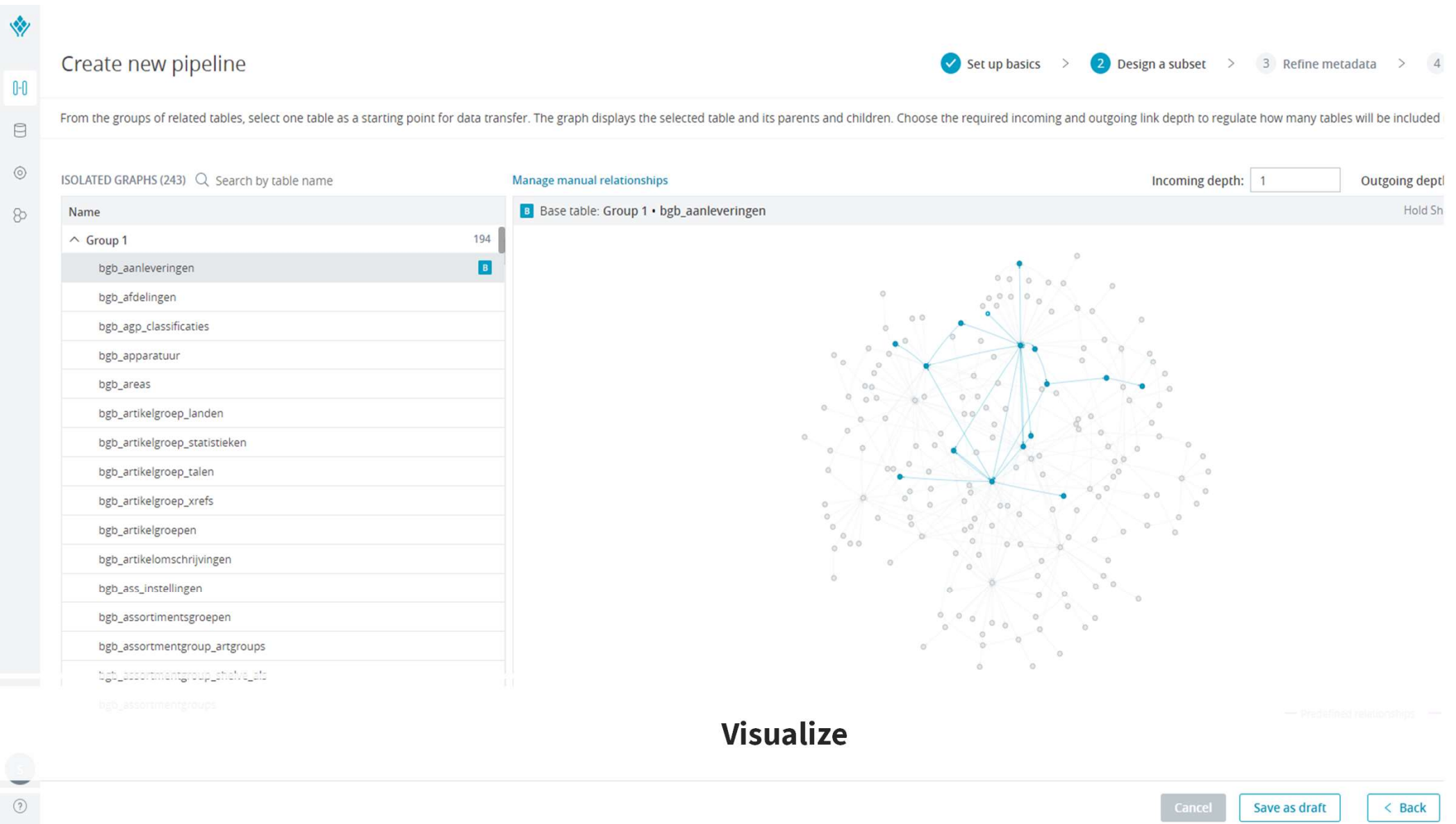
Example:

The cost of a dataset generation with **100K** rows and **15** columns is **0.012 * 20 = 0.24 USD** in AWS SageMaker

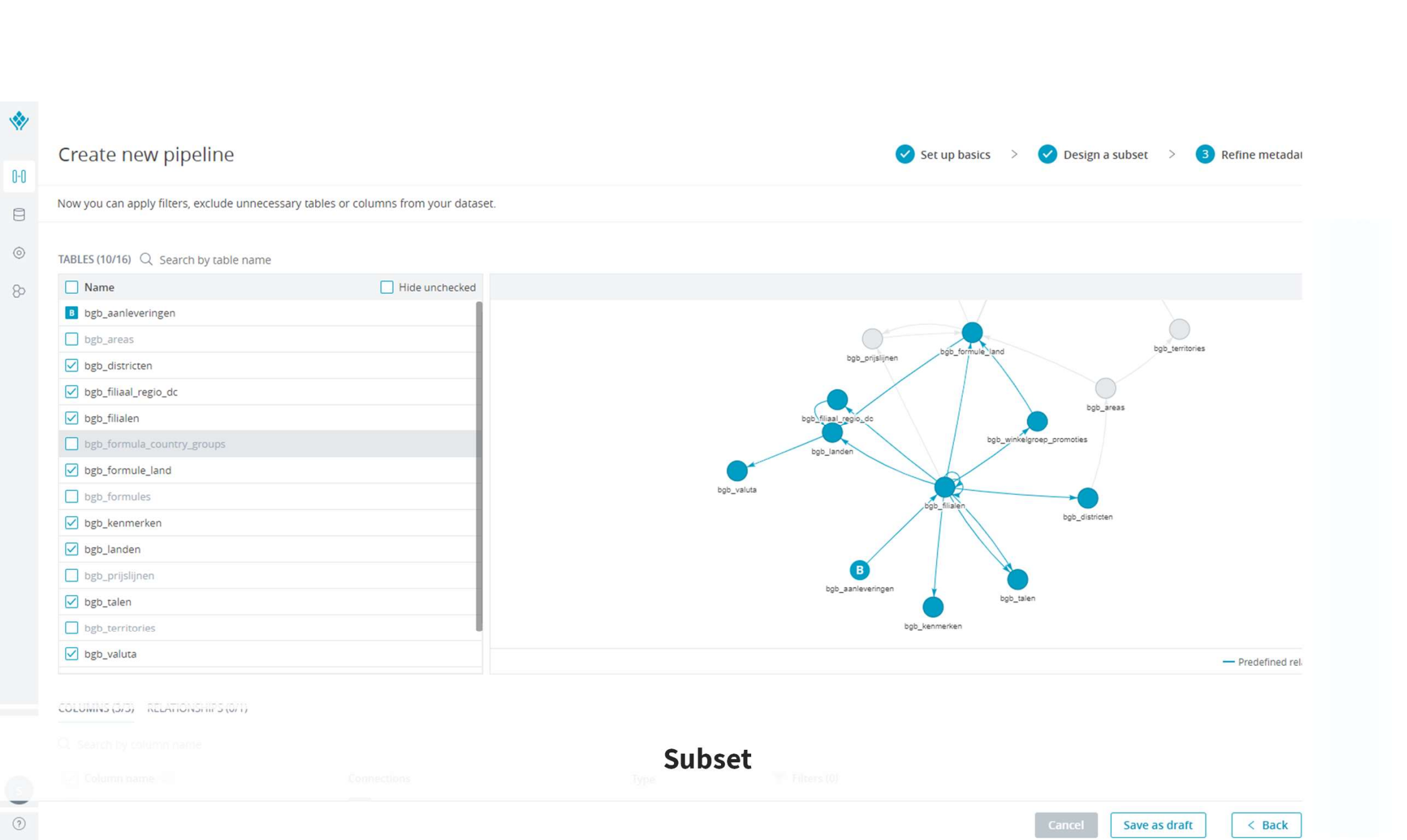
Note: Data generation does not require GPU-enabled VM and can be done on most types of machines, including workstations.

Trained models as a data source

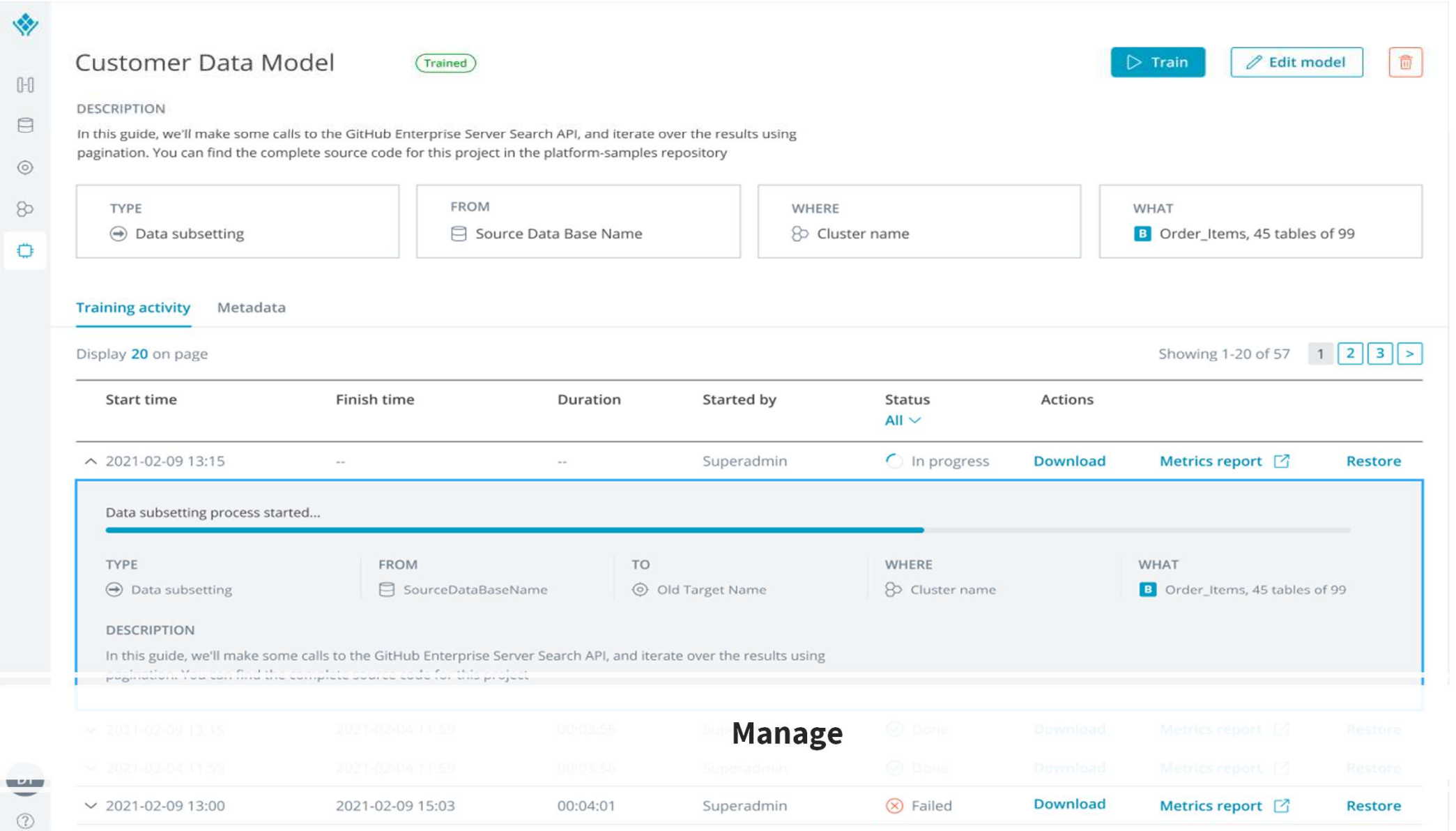
Train	Train one model per table
Relate	Pass information on relationships between tables to the models
Preserve	Preserve data types of the attributes
Manage	Manage set of trained models as a single data source
Evaluate	Evaluate models performance after training
Control	Control the training cost
Generate	Generate data and load it into various data stores



Visualize


Create new pipeline
Set up basics
Design a subset
3 Refine metadat
Now you can apply filters, exclude unnecessary tables or columns from your dataset.
TABLES (10/16)
Search by table name
Name
Hide unchecked
bgb_aanleveringen
bgb_areas
bgb_districten
bgb_filiaal_regio_dc
bgb_filialen
bgb_formula_country_groups
bgb_formule_land
bgb_formules
bgb_kenmerken
bgb_landen
bgb_prijslijnen
bgb_talen
bgb_territories
bgb_valuta
bgb_prijslijnen
bgb_formule_land
bgb_territories
bgb_areas
bgb_filiaal_regio_dc
bgb_winkelgroep_promoties
bgb_landen
bgb_valuta
bgb_filialen
bgb_districten
B
bgb_aanleveringen
bgb_talen
bgb_kenmerken
Predefined rel
Search by column name
Column name
Connections
Type
Filters (0)
Cancel
Save as draft
Back

Subset



Manage

Likely limitations

Dependencies between relational entities might be broken

Too much training data increases requirements to the hardware resources

Too little training data affects accuracy of the generated data

The model maintains correlations within datasets but not with the outside world objects

Machine readable formats such as JSON and XML

Thank you!

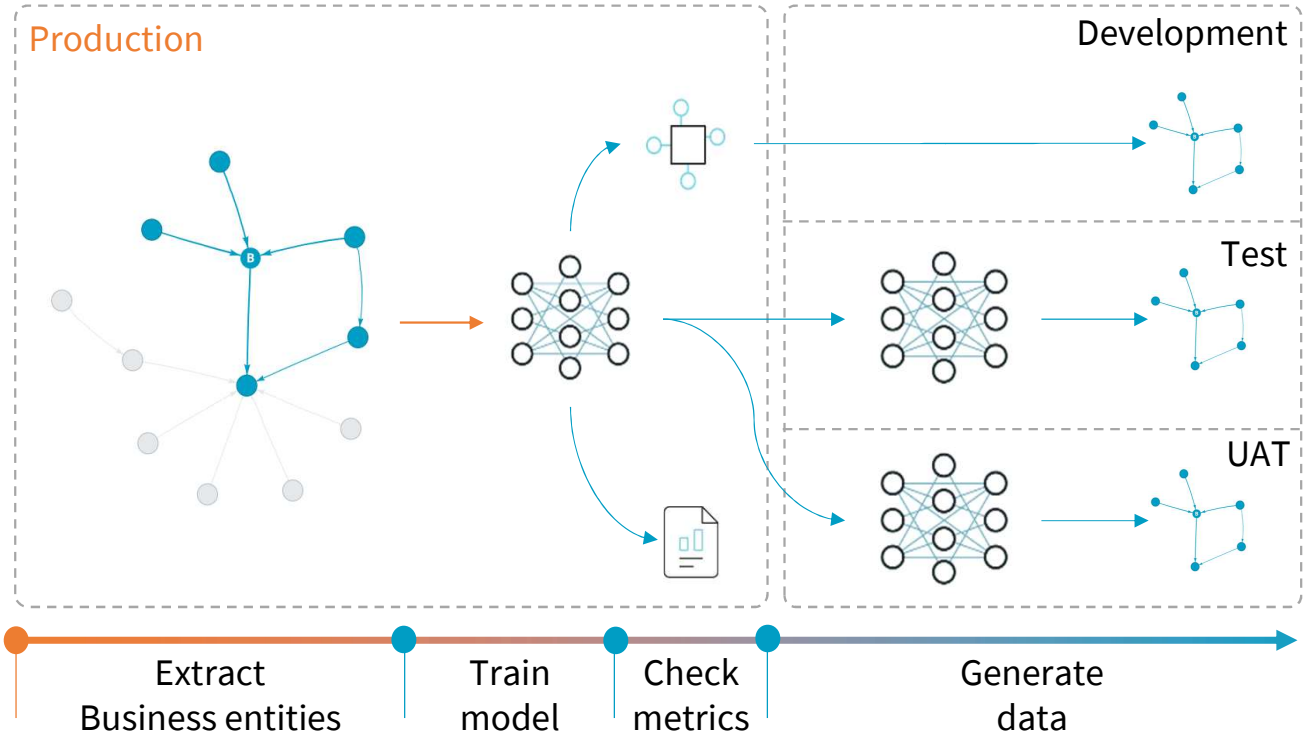
For more information, please contact:
supportepmc-tdm@epam.com

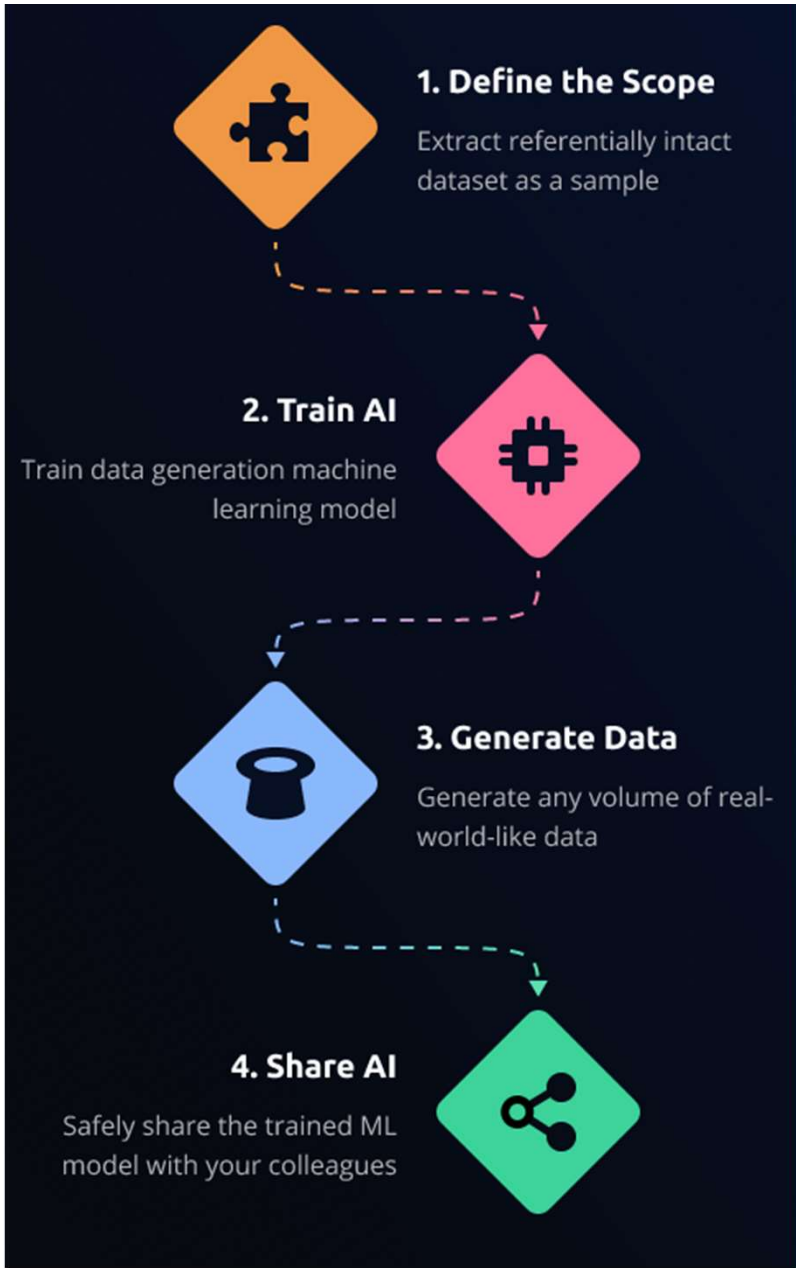
Appendix

TDSPORA: OVERVIEW

High level flow

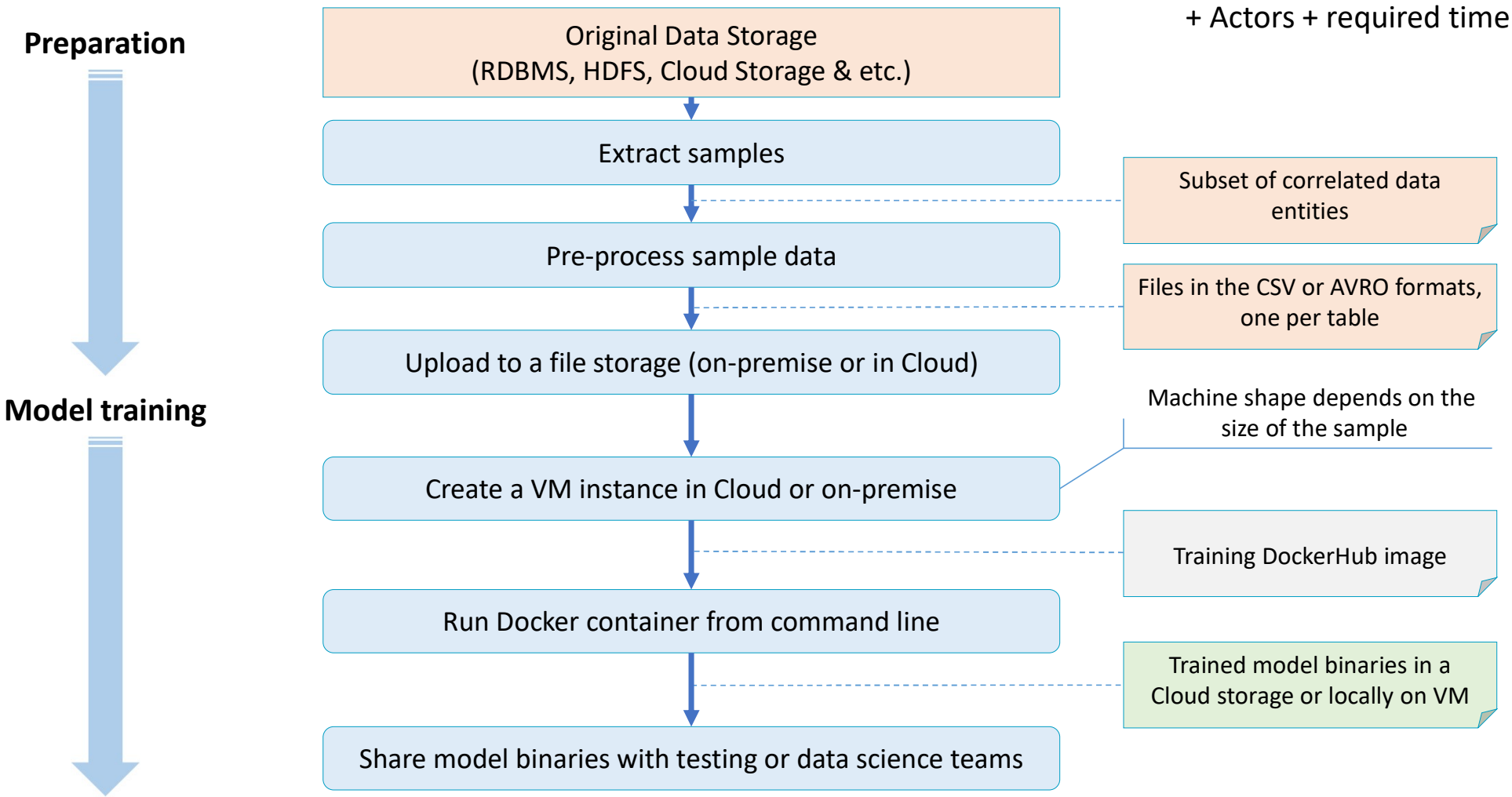
PRODUCTION DATA STAYS WITHIN SECURED PERIMETER



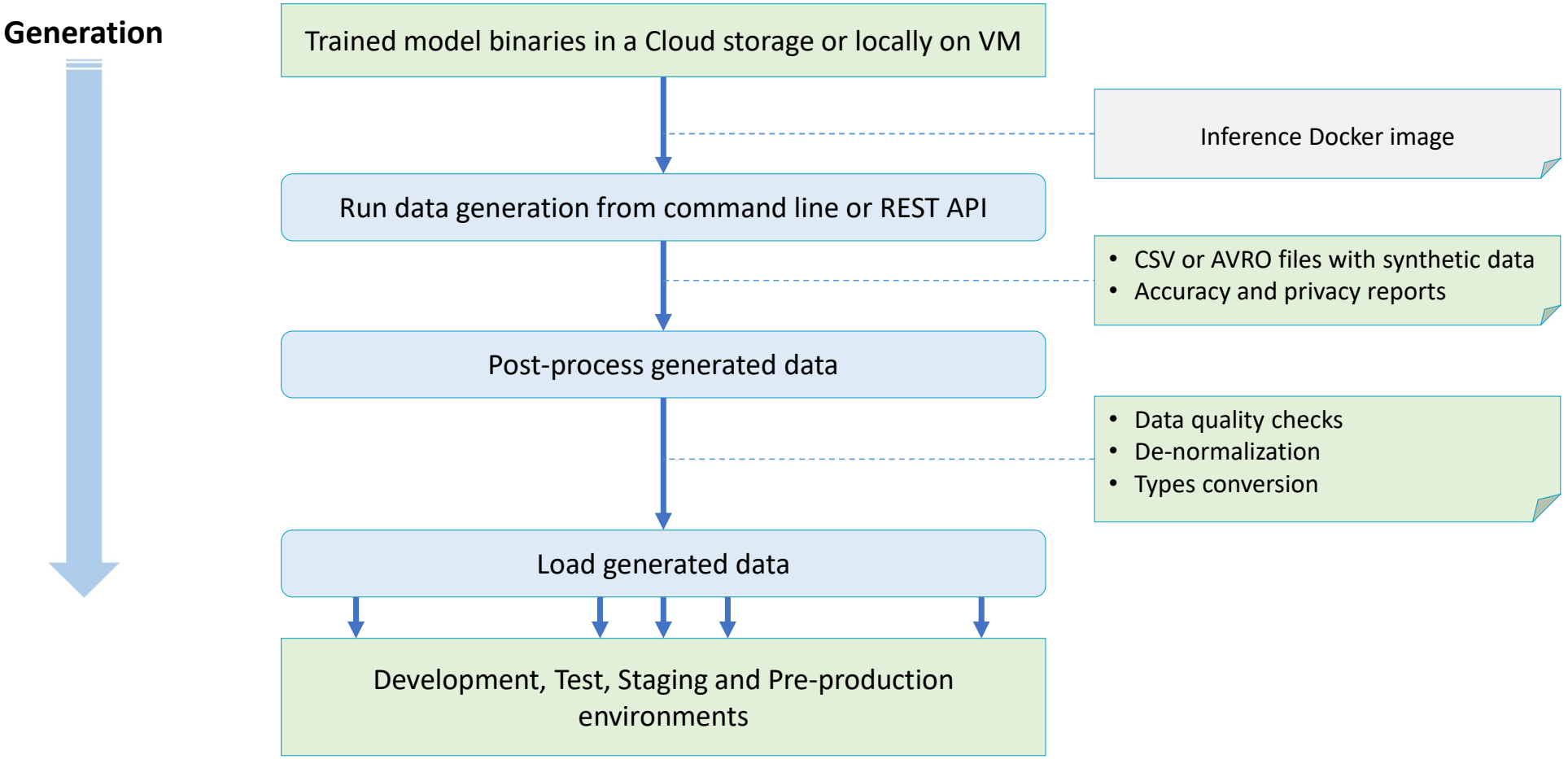


<epam> | TDspora

Reference implementation: Training under your control



Reference implementation: Synthetic data generation



Compare with the “Classic” flow

CLASSIC

Pros

- Precision and control over the data classification and mutation
- High data utility preservation (masking)
- Flexible rules and techniques
- High performance and low resource consumption
- Consistency and repeatability

Cons

- Labor-intensive steps
- Risks associated with a human error
- Requires training of personnel to use the tool efficiently

ML-DRIVEN Experimental

Pros

- Short test data delivery cycle
- No or minimal human involvement
- Preserves hidden patterns in data and suitable for ML tasks
- Allows moving models between security perimeters rather than data
- Comprehensive metrics
- Increases test data variability and injects boundary cases

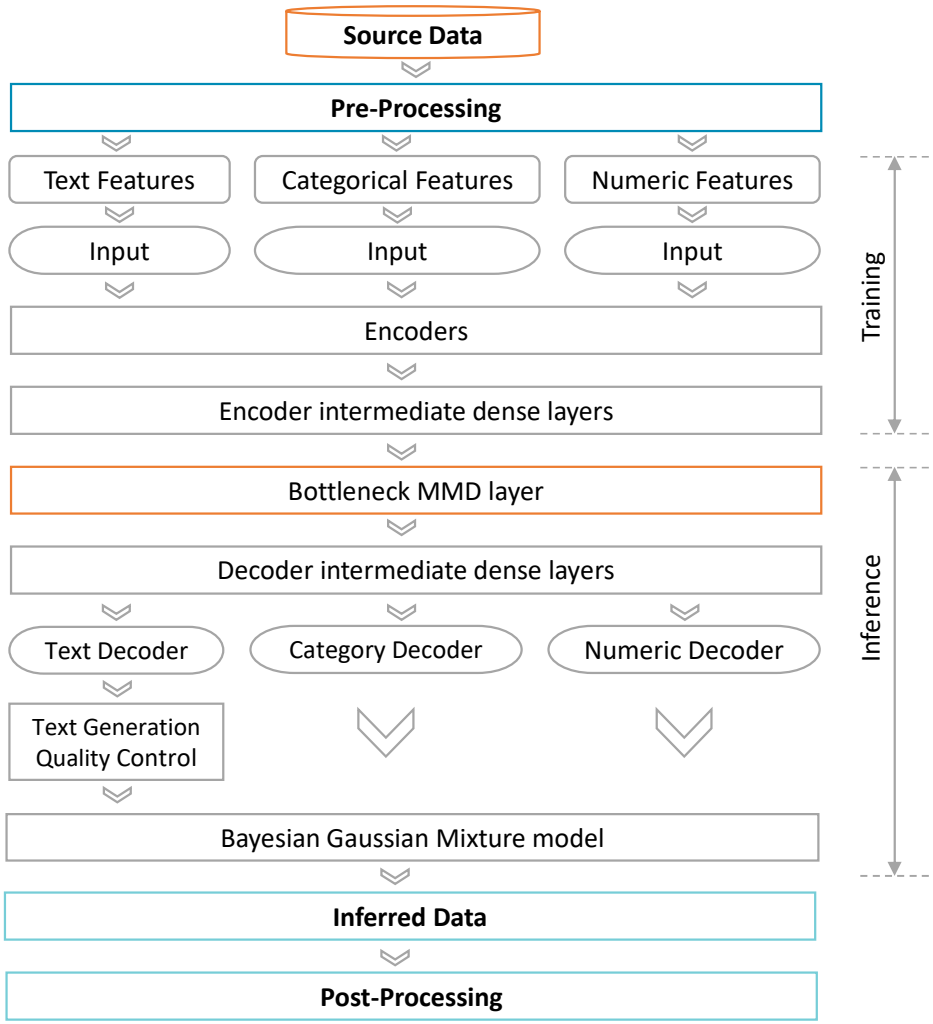
Cons

- High resource consumption for training and inference
- No consistency or repeatability
- Limited support of the relationships between datasets

Internal Model Architecture

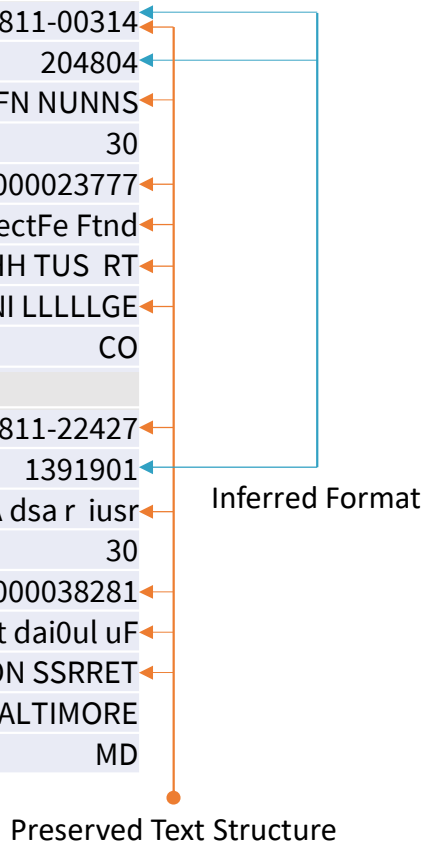
In the core of the data generation module lies a customized deep learning model, based on the Info VAE (Variational Autoencoder) with the following modifications:

- MMD bottleneck layer is added to regularize the latent space
- Bayesian gaussian mixture model (BGMM) on the latent space identifies latent Gaussian regimes
- Unsupervised datatypes detection is employed for preprocessing pipeline
- Model transformers are deployed to generate complex text when needed
- Differential Privacy (DP) enabled as part of the model training process



Samples – Investment Series

	ORIGINAL	GENERATED
Attribute	**Value**	**Value**
Reporting File Number	811-00582	811-00314
CIK	44402	204804
Name of Registrant	NEUBERGER BERMAN EQUITY FUNDS	TRRESERWESREE M F FN NUNNS
Org Type	30	30
Series ID	S000007847	S000023777
Series Name	Neuberger Berman Socially Responsive Fund	GranmaaeEiHT MMh ti roaTemnad ectFe Ftnd
Address_1	605 THIRD AVENUE	1523 SOOHH TUS RT
City	NEW YORK	DRNTENONI LLLLLGE
State	NY	CO
Reporting File Number	811-21714	811-22427
CIK	1317146	1391901
Name of Registrant	MML Series Investment Fund II	ToS e slA dsa r iusr
Org Type	30	30
Series ID	S000028330	S000038281
Series Name	MML Short-Duration Bond Fund	B..Rn nRlRrriSplo 0n2pdeet dai0ul uF
Address_1	1295 STATE STREET	101 EAUH TTON SSRRET
City	SPRINGFIELD	BALTIMORE
State	MA	MD



Pre- and post- processing as the way to improve the accuracy

Before sending data to the model

- Convert dates to timestamps
- Remove dashes from the UUIDs
- De-normalize dimension (small) tables

After data generation

- Filter out unwanted records
- Restore dates from timestamps
- Normalize dimension (small) tables

References on privacy metrics

- «Fidelity and Privacy of Synthetic Medical Data (Review of Methods and Experimental Results)», Ofer Mendelevitch, Michael D. Lesh SM MD FACC, июнь 2021
  https://arxiv.org/ftp/arxiv/papers/2101/2101.08658.pdf
- «Bootstrap confidence intervals», Jeremy Orloff and Jonathan Bloom
  https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading24.pdf
- Short overview of metrics above
  https://mostly.ai/2020/11/04/truly-anonymous-synthetic-data-legal-definitions-part-ii/
- Implementation of quality, accuracy, and privacy metrics for the ML data generation models
  https://sdv.dev/SDV/user_guides/evaluation/index.html
  https://github.com/sdv-dev/SDMetrics

Some examples

Original table:

Age	Workclass	fnlwght	Education	Marital status	Occupation	Relationship	Race	Sex	Hours per week	Native country	Capital	Income
39	State-gov	77516	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	2174	<=50K
50	Self-emp-not-inc	83311	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	13	United-States	0	<=50K
28	Private	338409	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	40	Cuba	0	<=50K

Table generated by neural network:

Age	Workclass	Fnlwght	Education	Marital status	Occupation	Relationship	Race	Sex	Hours per week	Native country	Capital	Income
32	Private	249175	Some-college	Never-married	Other-service	Not-in-family	White	Female	40	United-States	40	<=50K
50	Private	218399	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	48	United-States	-54	>50K
25	Local-gov	114298	Bachelors	Divorced	Prof-specialty	Unmarried	White	Female	34	Canada	40	<=50K