TABULARIS.AI

Advancing Data Privacy with Synthetic Data:
A Novel Approach to Secure Analytics

Vadim Borisov, PhD
tabularis.ai

19.11.2024

About *Tabularis.AI*

Tabularis.AI, a spin-off from the University of Tübingen, specializes in generating synthetic datasets that mirror real data characteristics, ensuring high data security.

Tabularis.AI aims to provide synthetic datasets that replicate real data characteristics, enabling secure data access, usage, and sharing while maintaining the highest standards of data security.

Outline

- The Great Privacy Wall
- Approached for Data Privacy
- Annonymization
- Differential Privacy
- Synthetic Data
- The Tabularis Approach
- Case Study
- Summary

The Great Data Privacy Wall

Raw Data

- Personal Information
- Sensitive Data
- User Behavior

Privacy Wall

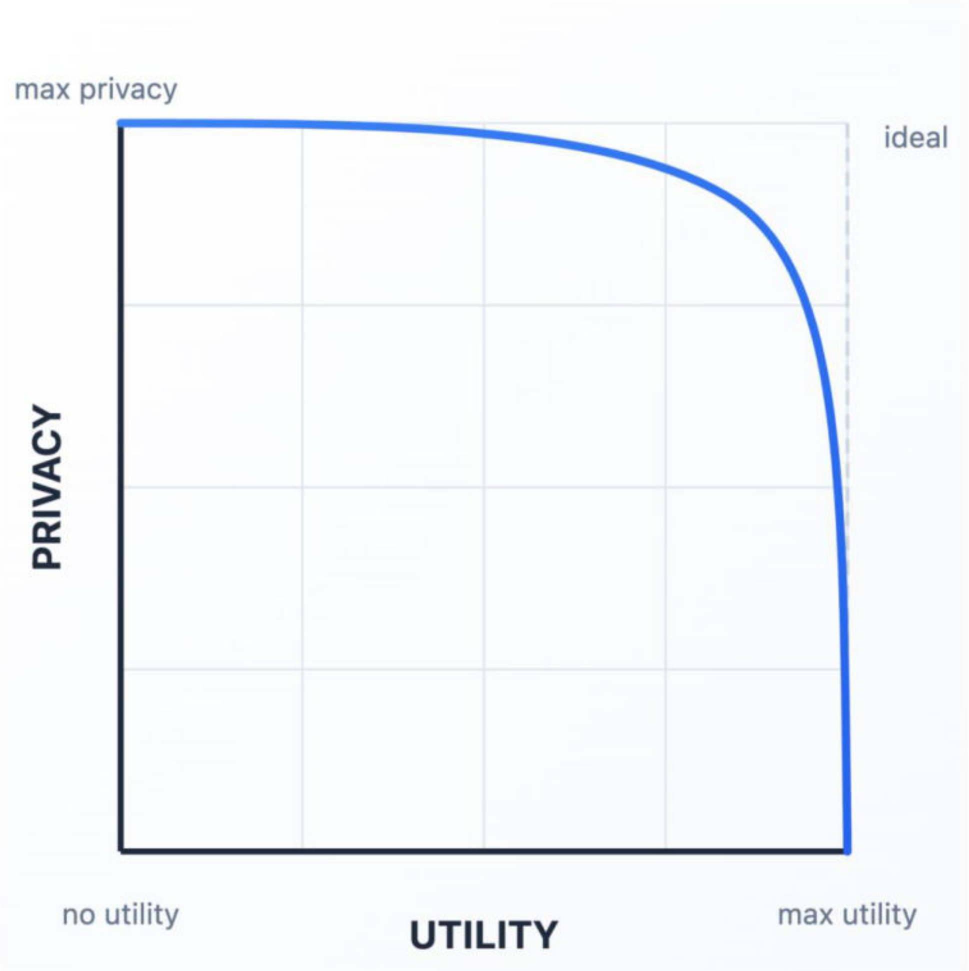
- Privacy Laws
- Data Protection
- GDPR

Algorithms & Insights

- Decision Making
- Improvements
- Innovation

How can we safely bridge this gap?

Approaches for Data Privacy



Nelson, Gregory S. "Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification." In *SAS global forum proceedings*, pp. 1-23. 2015.

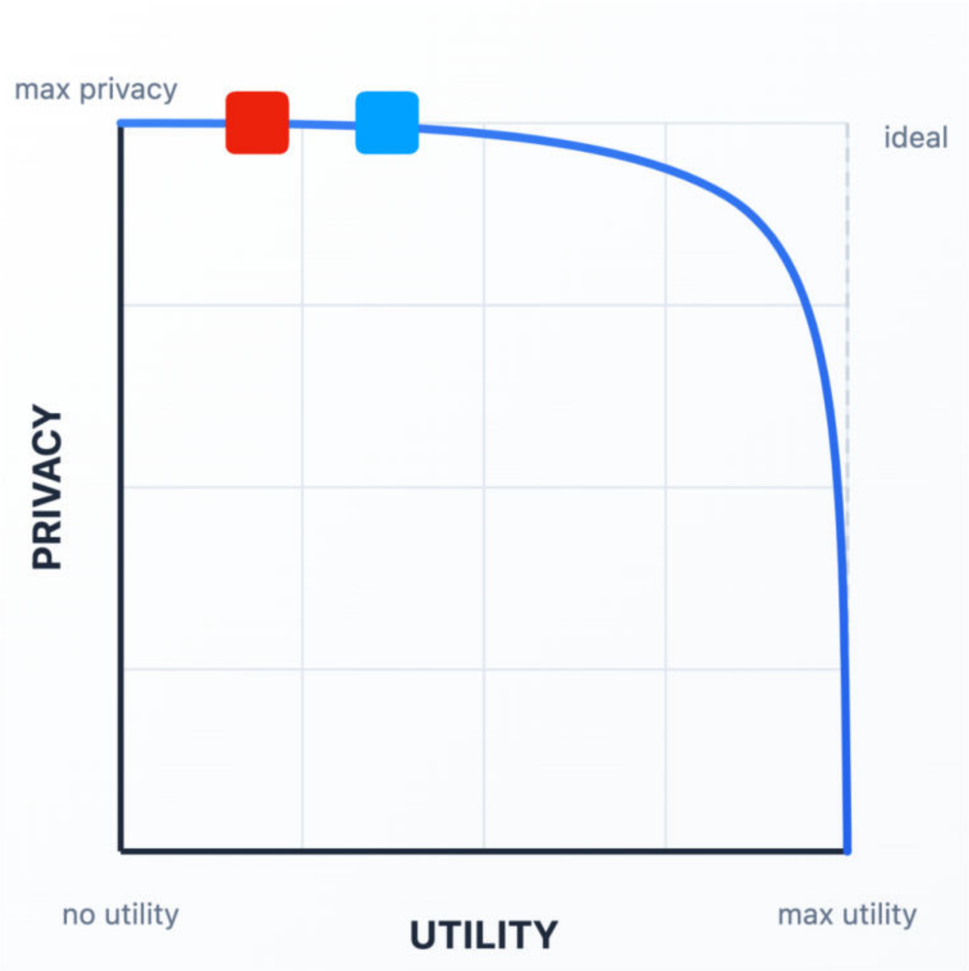
Approaches for Data Privacy

- **Anonymization**
  - Removal of personally identifiable information,
  - Adding "confusing" information to mislead an "attacker"

Nelson, Gregory S. "Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification." In *SAS global forum proceedings*, pp. 1-23. 2015.

Approaches for Data Privacy

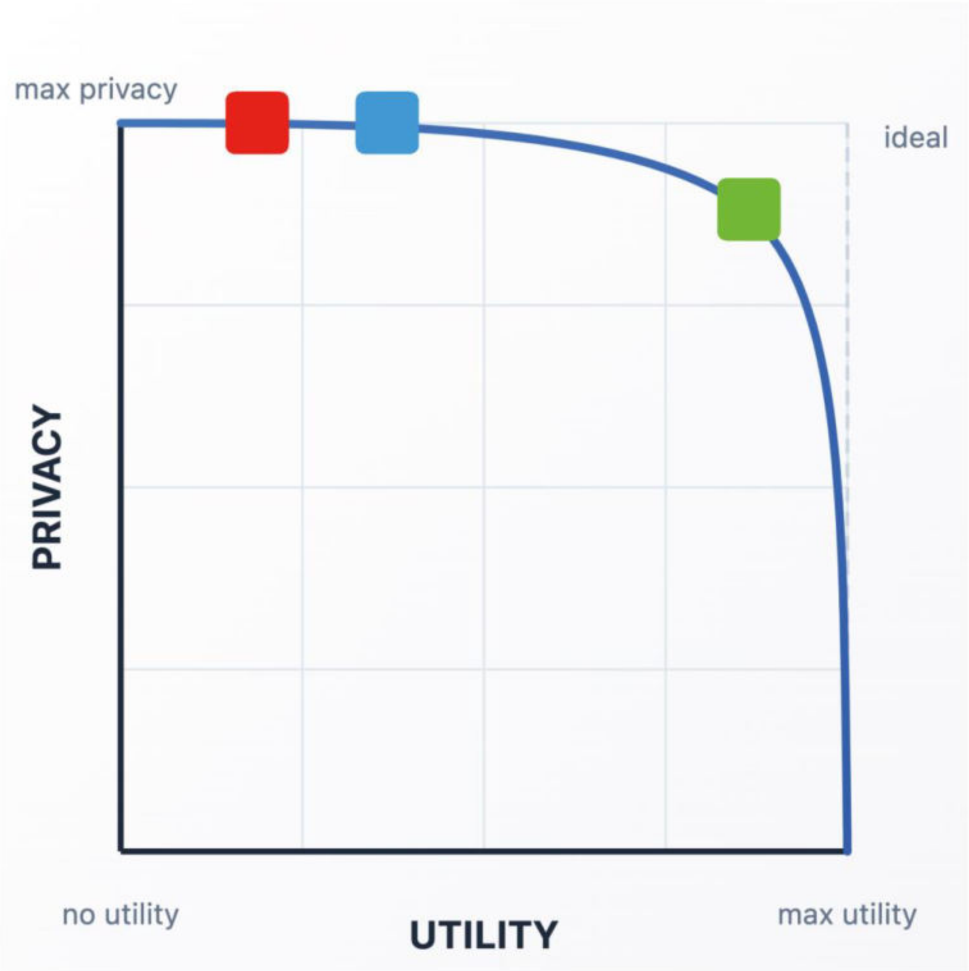
- **Anonymization**
  - Removal of personally identifiable information,
  - Adding "confusing" information to mislead an "attacker"
- **Differential Privacy**
  - Adding controlled noise to data or queries
  - Provides mathematical privacy guarantees



Nelson, Gregory S. "Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification." In *SAS global forum proceedings*, pp. 1-23. 2015.

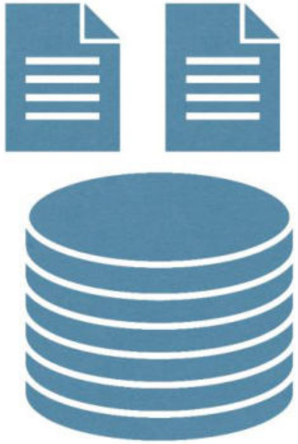
Approaches for Data Privacy

- **Anonymization**
  - Removal of personally identifiable information,
  - Adding "confusing" information to mislead an "attacker"
- **Differential Privacy**
  - Adding controlled noise to data or queries
  - Provides mathematical privacy guarantees
- **Synthetic Data**
  - Artificially generated data mimicking real data
  - Preserves statistical properties
  - Can be combined with anonymization and differential privacy



Nelson, Gregory S. "Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification." In *SAS global forum proceedings*, pp. 1-23. 2015.

Anonymization



Original Data

Data Anonymisation

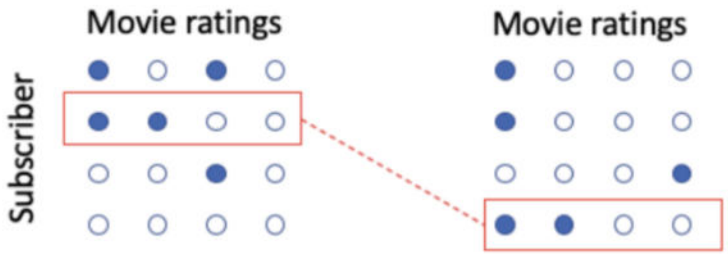
Current anonymization approaches include techniques like k-anonymity, l-diversity, and t-closeness, which remove or obscure identifiable information to protect individual privacy.

Name	Age	Gender	Height	Weight	State of domicile	Religion	Disea
*	20 < Age ≤ 30	Female	165cm	72kg	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	162cm	70kg	Kerala	*	Viral infe
*	20 < Age ≤ 30	Female	170cm	68kg	Tamil Nadu	*	Tubercu
*	20 < Age ≤ 30	Male	170cm	75kg	Karnataka	*	No illnes
*	20 < Age ≤ 30	Female	165cm	71kg	Kerala	*	Heart-re
*	20 < Age ≤ 30	Male	160cm	69kg	Karnataka	*	Tubercu
*	Age ≤ 20	Male	167cm	85kg	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	180cm	81kg	Karnataka	*	Heart-re
*	Age ≤ 20	Male	175cm	79kg	Kerala	*	Heart-re
*	Age ≤ 20	Male	169cm	82kg	Kerala	*	Viral infe

https://blog.pangeanic.com/discover-k-anonymity

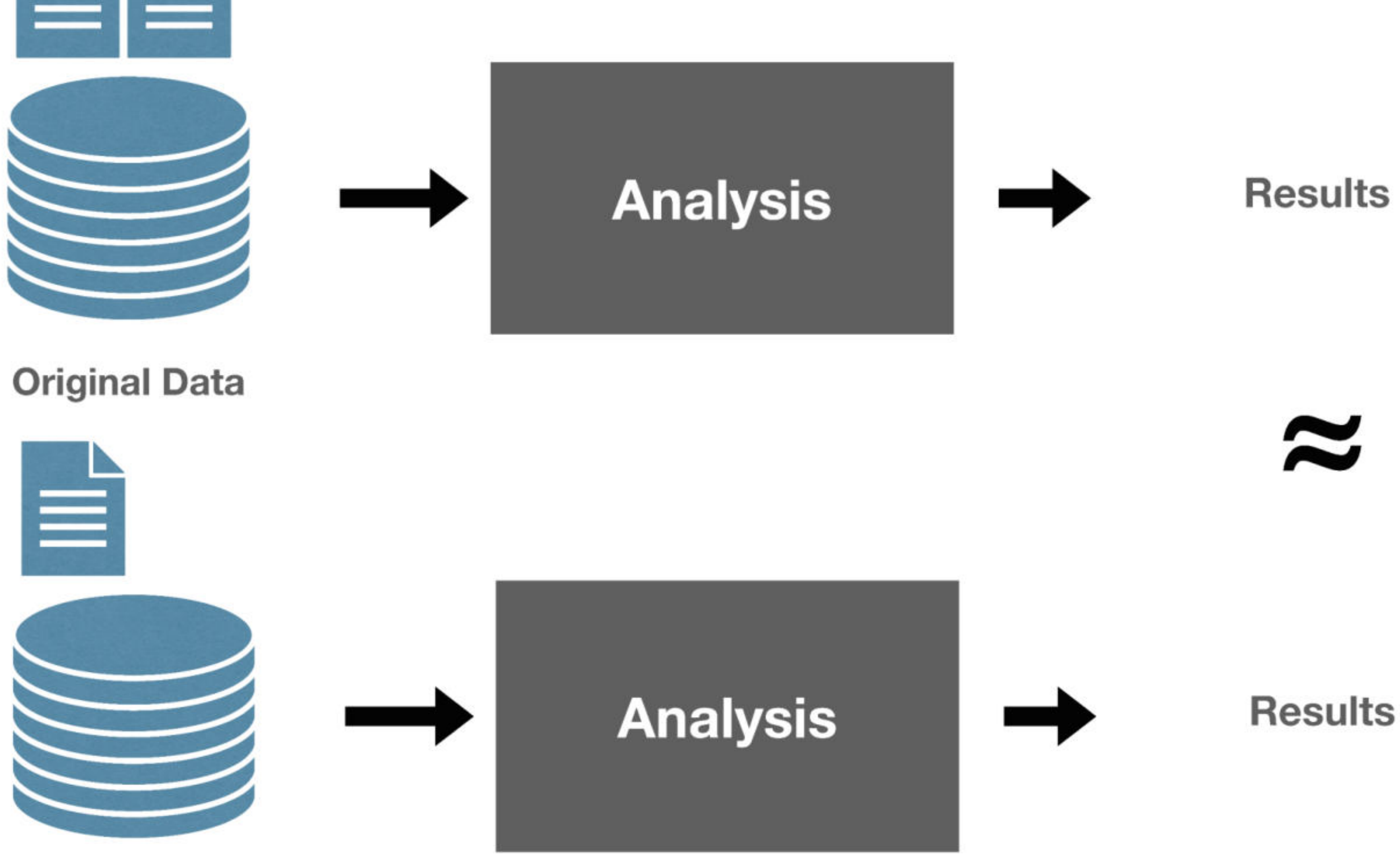
Example: The Netflix Prize

- Netflix released 100M ***anonymized*** movie ratings from 480k users.
- Goal was to improve Netflix's movie recommendation algorithm by 10%, winner gets 1M $
- Researchers cross-referenced with public IMDb ratings
    - 99% of users identifiable with 8 ratings + dates
    - 68% identifiable with just 2 ratings + dates
- Simple anonymization is insufficient; even "safe" data can compromise privacy when combined with external information.



Differential Privacy

Differential Privacy ensures that removing or adding one data point doesn't significantly change the analysis results, preserving privacy.



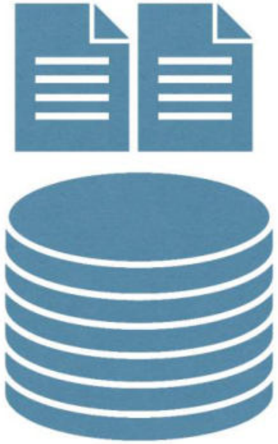
Results are approximate the same.

The differential privacy provides mathematical privacy guarantees.

Original Data without a record x_i

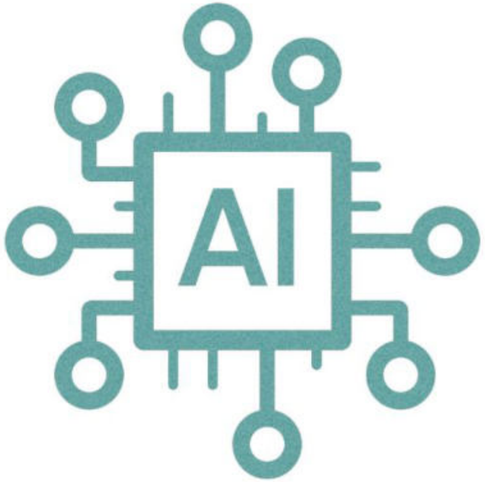
https://blog.pangeanic.com/discover-k-anonymity

Synthetic Data



Original Data

Generative Model



Current approaches are mostly based on *Bayesian networks* (Chow & Liu, 1968), *GANs* (Choi et al., 2017), VAEs (Xu et al., 2019), *diffusion models* (Lee et al., 2023), and *LLMs* (Borisov et al., 2023).

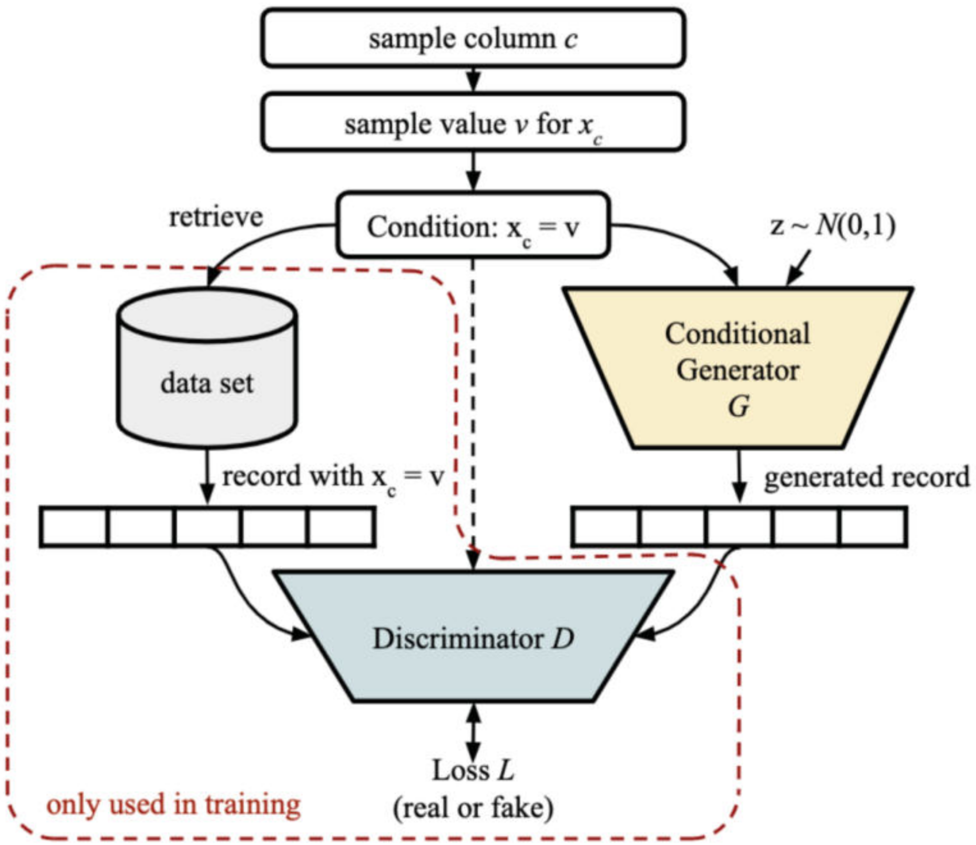
Synthetic Data

- Ensures Privacy Protection
- Facilitates Safe Data Sharing
- Reduces Bias in Data
- Scalable for Large Datasets

https://www.fleksy.com/solutions/generative-ai/

Synthetic Data Generation Methods

Conditional Tabular GAN (CTGAN)

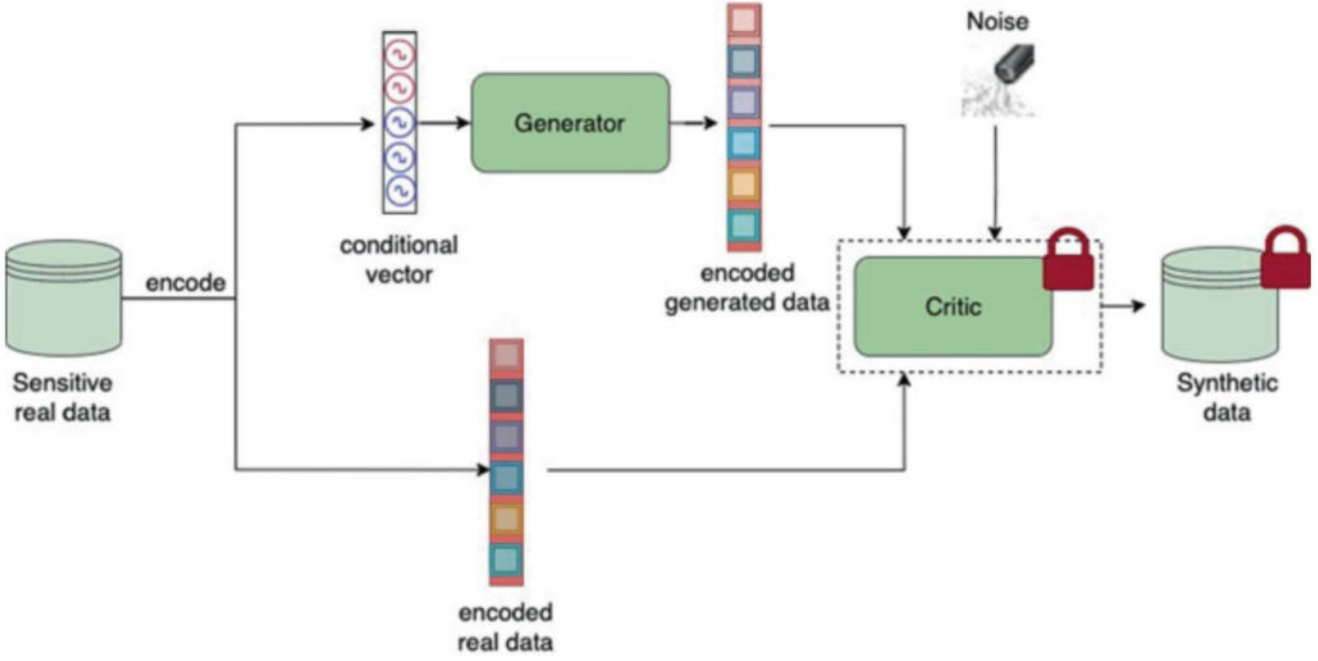


A condition is sampled first and passed to the conditional generator ***G*** along with a random input ***z***. The generated sample is opposed to a randomly picked example from the data set that also fulfills the condition and assessed by the conditional discriminator ***D***. This approach allows to preserve dependency relations.

One of the first method dedicated to tabular data

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling tabular data using conditional gan." Advances in neural information processing systems 32 (2019).

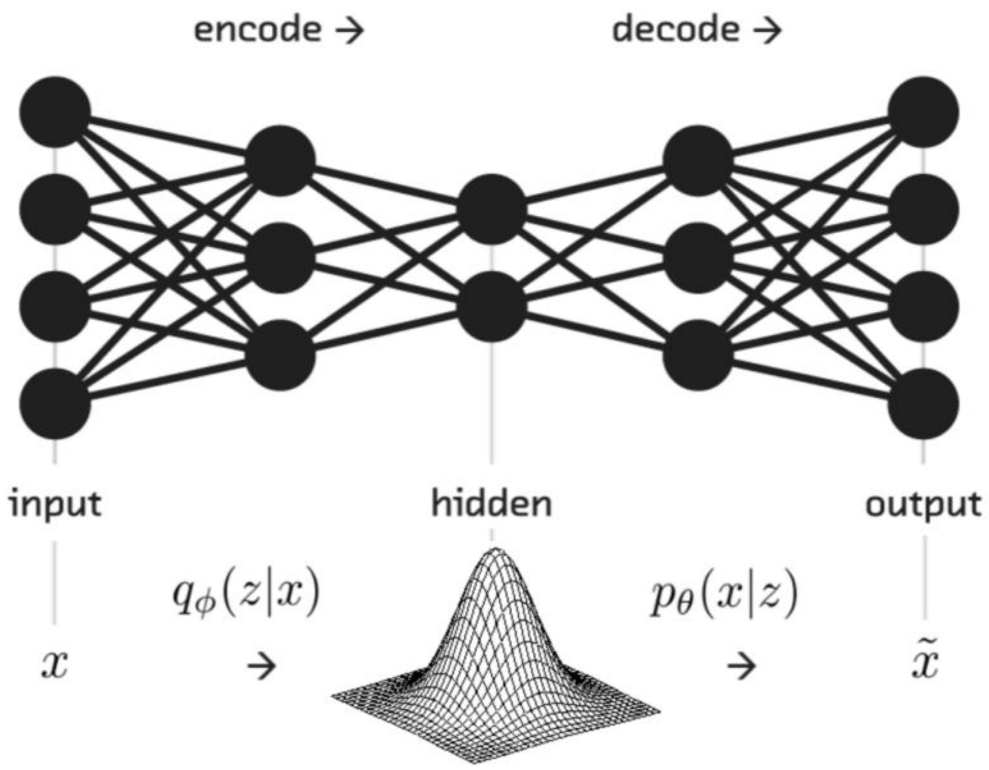
DP-CTGAN: Private CTGAN



Building on the advancements of CTGAN, DP-CTGAN incorporates differential privacy into the conditional tabular generative model to enhance data privacy. This model surpasses current state-of-the-art models under the same privacy budget and leverages federated learning for secure synthetic data generation without centralizing data.

Fang, Mei Ling, Devendra Singh Dhami, and Kristian Kersting. "Dp-ctgan: Differentially private medical data generation using ctgans." In *International Conference on Artificial Intelligence in Medicine*, pp. 178-188. Cham: Springer International Publishing, 2022.

Tabular Variational Autoencoder (TVAE)



Tabular Variational Autoencoder (TVAE) generates synthetic tabular data by learning the underlying data distribution through a variational autoencoder framework.

https://medium.com/@muhammedbuyukkinaci/variational-autoencoder-nedir-vae-t%C3%BCrk%C3%A7e-c1d92bdd3468

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling tabular data using conditional gan." Advances in neural information processing systems 32 (2019).

Re*imagin*ing Data Generation Techniques

- **Computer Vision Techniques:**
  - GANs and VAEs are popular for generating synthetic data in **homogeneous** data domains like images.
- **Tabular Data Complexity:**
  - Tabular data is highly **heterogeneous**, including numerical, categorical, and missing values.
- **The key issue:**
  - Existing methods struggle with this diversity and complexity of tabular data

Picture: *These people are not real – they were produced by our generator that allows control over different aspects of the image.*

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and improving the image quality of stylegan." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8110-8119. 2020.

The Tabularis Approach

Tabular Data Preprocessing

Original Tabular Dataset

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	90000
35	Nurse	Associate	65000
40	Teacher	Master	55000
NaN	Engineer	Bachelor	80000
22	Student	Some College	15000

Tabular Data Preprocessing

Original Tabular Dataset

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	90000
35	Nurse	Associate	65000
40	Teacher	Master	55000
NaN	Engineer	Bachelor	80000
22	Student	Some College	15000

Preprocessed Tabular Dataset

F1	F2	F3	F4
28	0	0	90000
35	1	1	65000
40	2	2	55000
31.25	3	0	80000
22	4	3	15000

Can we do it differently?

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	80,000
35	Nurse	Associate	65,000
40	Teacher	Master	55,000
NaN	Engineer	Bachelor	80,000
22	Student	Some College	15,000

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	80,000
35	Nurse	Associate	65,000
40	Teacher	Master	55,000
NaN	Engineer	Bachelor	80,000
22	Student	Some College	15,000

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	80,000
35	Nurse	Associate	65,000
40	Teacher	Master	55,000
NaN	Engineer	Bachelor	80,000
22	Student	Some College	15,000

Age is 28,

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	80,000
35	Nurse	Associate	65,000
40	Teacher	Master	55,000
NaN	Engineer	Bachelor	80,000
22	Student	Some College	15,000

Age is 28, Occupation is Data Scientist,

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	80,000
35	Nurse	Associate	65,000
40	Teacher	Master	55,000
NaN	Engineer	Bachelor	80,000
22	Student	Some College	15,000

Age is 28, Occupation is Data Scientist, Education is Bachelor,

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	80,000
35	Nurse	Associate	65,000
40	Teacher	Master	55,000
NaN	Engineer	Bachelor	80,000
22	Student	Some College	15,000

Age is 28, Occupation is Data Scientist, Education is Bachelor, Salary is 80,000

Textual Encoding

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	80,000
35	Nurse	Associate	65,000
40	Teacher	Master	55,000
NaN	Engineer	Bachelor	80,000
22	Student	Some College	15,000

Age is 28, Occupation is Data Scientist, Education is Bachelor, Salary is 80,000

Age is 35, Occupation is Nurse, Education is Associate's, and Salary is 65,000.

Age is 40, Occupation is Teacher, Education is Master's, and Salary is 55,000.
Age is missing, Occupation is Engineer, Education is Bachelor's, and Salary is 80,000.

Age is 22, Occupation is Student, Education is Some College, and Salary is 15,000.

Textual Encoding

Age	Occupation	Education	Salary
28	Data Scientist	Bachelor	80,000
35	Nurse	Associate	65,000
40	Teacher	Master	55,000
NaN	Engineer	Bachelor	80,000
22	Student	Some College	15,000

Age is 28, Occupation is Data Scientist, Education is Bachelor, Salary is 80,000

Age is 35, Occupation is Nurse, Education is Associate's, and Salary is 65,000.

Age is 40, Occupation is Teacher, Education is Master's, and Salary is 55,000.

Age is missing, Occupation is Engineer, Education is Bachelor's, and Salary is 80,000.

Age is 22, Occupation is Student, Education is Some College, and Salary is 15,000.

Pre-trained Large Language Models For Synthetic Tabular Data Generation

- **Utilize** advanced pretrained autoregression Large Language Models (LLMs), (GPT2, GPT3, etc) for tabular data generation task
- **Benefit** from the LLM's extensive training on diverse data sources for better data representation
- Commonly, the probability of natural-language sequences is factorized in an auto-regressive manner in LLMs (Jelinek, 1980; Bengio et al., 2000). It is represented as a product of output probabilities conditioned on previously observed tokens:

$$p(\mathbf{t}) = p(w_1, \ldots, w_j) = \prod_{k=1}^{j} p(w_k \mid w_1, \ldots, w_{k-1}) .$$

Textual Representation of Tabular Data

Age is 28, Occupation is Data Scientist, Education is Bachelor, Salary is 80,000

Age is 35, Occupation is Nurse, Education is Associate's, and Salary is 65,000.

Age is 40, Occupation is Teacher, Education is Master's, and Salary is 55,000.

Age is missing, Occupation is Engineer, Education is Bachelor's, and Salary is 80,000.

Age is 22, Occupation is Student, Education is Some College, and Salary is 15,000.

The Data Pipeline For The Fine-Tuning Step

Original tabular data set

Age	Education	Occupation	Gender	Income
39	Bachelors	Adm-clerical	Male	$\leq$ 50K
50	HS-grad	Exec-managerial	Female	$\geq$ 50K
53	Bachelors	Prof-specialty	Female	$\geq$ 50K

The Data Pipeline For The Fine-Tuning Step

Textual Encoding

Original tabular data set

Age	Education	Occupation	Gender	Income
39	Bachelors	Adm-clerical	Male	$\leq$ 50K
50	HS-grad	Exec-managerial	Female	$\geq$ 50K
53	Bachelors	Prof-specialty	Female	$\geq$ 50K

→

"Age is 39, Education is Bachelors, Occupation is Adm-clerical, Gender is Male, Income is $\leq$ 50K.",
"Age is 50, Education is HS-grad, Occupation is Exec-managerial, Gender is Female, Income is $\geq$ 50K.",
"Age is 53, Education is 11th, Occupation is Handler-cleaners, Gender is Female, Income is $\geq$ 50K."

The Data Pipeline For The Fine-Tuning Step

Original tabular data set

Age	Education	Occupation	Gender	Income
39	Bachelors	Adm-clerical	Male	$\leq$ 50K
50	HS-grad	Exec-managerial	Female	$\geq$ 50K
53	Bachelors	Prof-specialty	Female	$\geq$ 50K

"Age is 39, Education is Bachelors, Occupation is Adm-clerical, Gender is Male, Income is $\leq$ 50K.",
"Age is 50, Education is HS-grad, Occupation is Exec-managerial, Gender is Female, Income is $\geq$ 50K.",
"Age is 53, Education is 11th, Occupation is Handler-cleaners, Gender is Female, Income is $\geq$ 50K."

Random Feature Order Permutation

"Education is Bachelors, Income is $\leq$ 50K, Age is 39, Occupation is Adm-clerical, Gender is Male.",
"Income is $\geq$ 50K, Occupation is Exec-managerial, Age is 50, Education is HS-grad, Gender is Female.",
"Occupation is Handler-cleaners, Education is 11th, Age is 53, Income is $\geq$ 50K, Gender is Female."

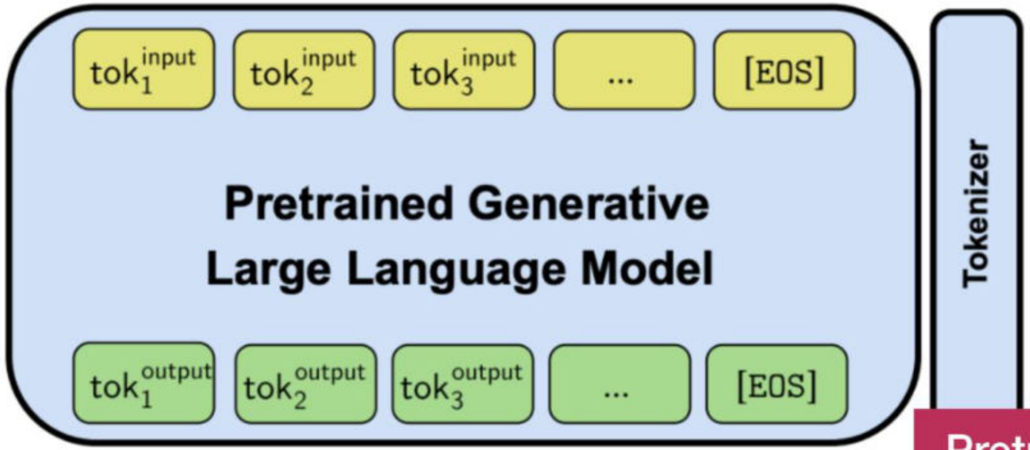
The Data Pipeline For The Fine-Tuning Step

Original tabular data set

Age	Education	Occupation	Gender	Income
39	Bachelors	Adm-clerical	Male	$\leq$ 50K
50	HS-grad	Exec-managerial	Female	$\geq$ 50K
53	Bachelors	Prof-specialty	Female	$\geq$ 50K

"Age is 39, Education is Bachelors, Occupation is Adm-clerical, Gender is Male, Income is $\leq$ 50K.",
"Age is 50, Education is HS-grad, Occupation is Exec-managerial, Gender is Female, Income is $\geq$ 50K.",
"Age is 53, Education is 11th, Occupation is Handler-cleaners, Gender is Female, Income is $\geq$ 50K."

"Education is Bachelors, Income is $\leq$ 50K, Age is 39, Occupation is Adm-clerical, Gender is Male.",
"Income is $\geq$ 50K, Occupation is Exec-managerial, Age is 50, Education is HS-grad, Gender is Female.",
"Occupation is Handler-cleaners, Education is 11th, Age is 53, Income is $\geq$ 50K, Gender is Female."



Pretrained LLMs Fine-tuning

The Sampling Procedure With Arbitrary Conditioning

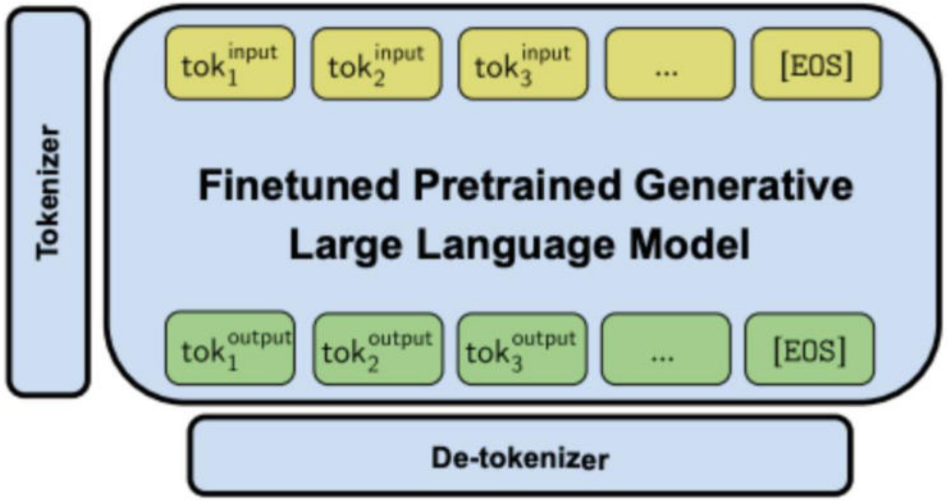
Input text sequences (Arbitrary conditioning)

[" ***Age*** "]

[" ***Age is 26,*** "]

[" ***Education is Masters, Age is 59,*** "]

(a) →



(b) ↓

"***Age*** *is 23, Occupation is Adm-clerical, Income is $\leq$ 50K, Gender is Missing, Education is 11th,* "

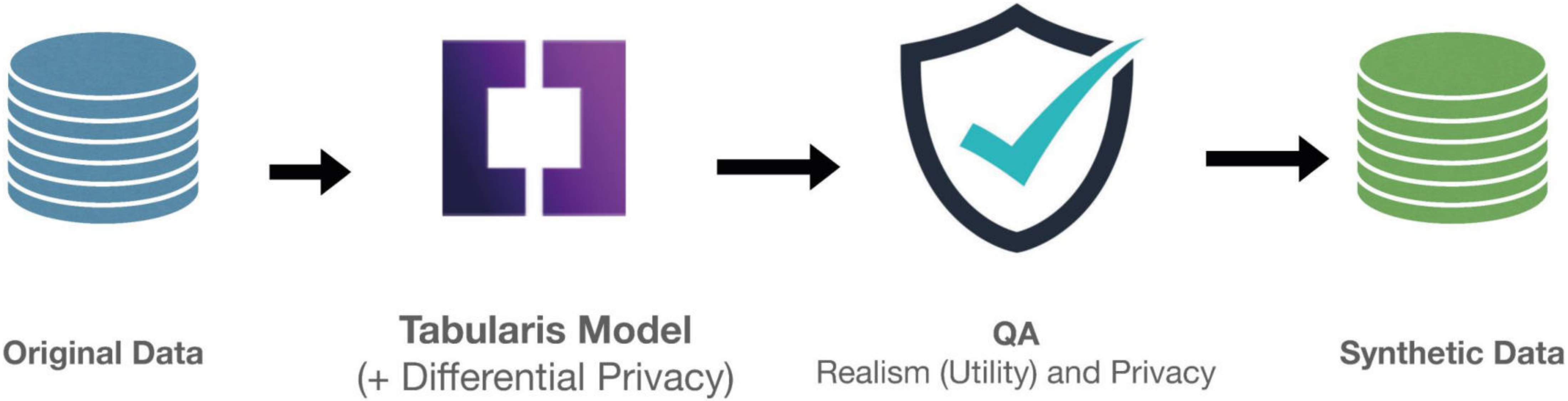
"***Age is 26****, Income is $\geq$ 50K, Occupation is Sales, Education is HS-grad, Gender is Female*"

"***Education is Masters, Age is 59,*** *Occupation is Other-service, Gender is Male, Income is $\geq$ 50K*"

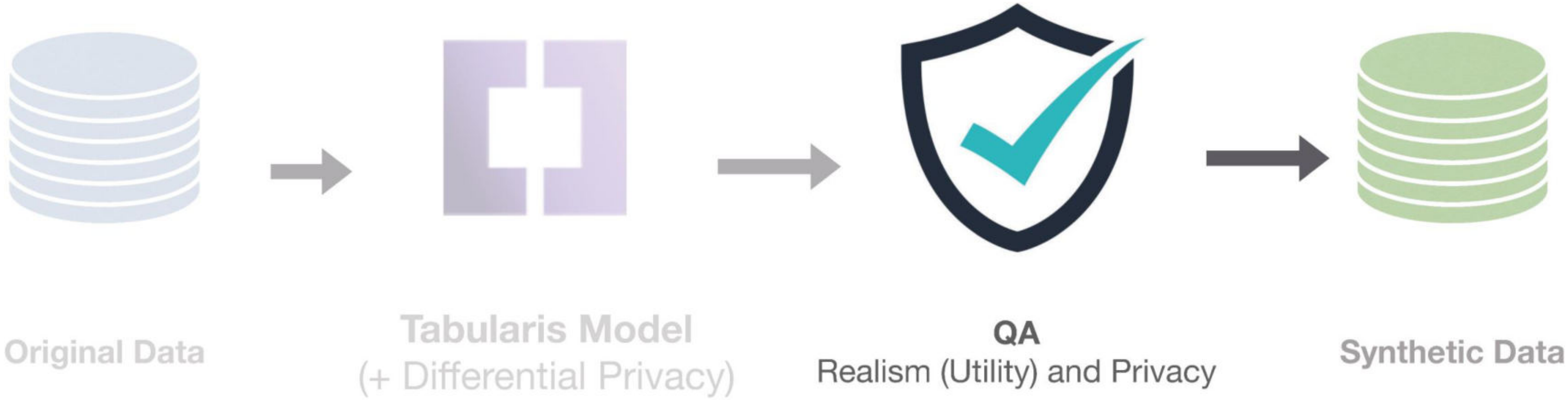
(c) ←

Synthetic tabular data set

Age	Education	Occupation	Gender	Income
23	11th	Adm-clerical	Missing	$\leq$ 50K
26	HS-grad	Sales	Female	$\geq$ 50K
59	Masters	Other-service	Male	$\geq$ 50K



https://www.istockphoto.com/de/grafiken/quality-assurance-logos



https://www.istockphoto.com/de/grafiken/quality-assurance-logos

Quality Assurance (QA) of Synthetic Data

📈 **Statistical Properties**:
Ensure synthetic data mirrors the mean, variance, covariance, and correlations of the original.

📊

📅

🔗

💯

🔐

Quality Assurance (QA) of Synthetic Data

📈 **Statistical Properties**:

Ensure synthetic data mirrors the mean, variance, covariance, and correlations of the original.

📊 **Data Diversity**:

Cover real-world scenarios to avoid overfitting models.

📅

🔗

💯

🔐

Quality Assurance (QA) of Synthetic Data

📈 **Statistical Properties**:
Ensure synthetic data mirrors the mean, variance, covariance, and correlations of the original.

📊 **Data Diversity**:
Cover real-world scenarios to avoid overfitting models.

📅 **Rare Events**:
Represent infrequent occurrences like rare diseases.

🔗

💯

🔐

Quality Assurance (QA) of Synthetic Data

📈 **Statistical Properties**:
Ensure synthetic data mirrors the mean, variance, covariance, and correlations of the original.

📊 **Data Diversity**:
Cover real-world scenarios to avoid overfitting models.

📅 **Rare Events**:
Represent infrequent occurrences like rare diseases.

🔗 **Feature Relationships**:
Maintain original relationships to preserve complexity.

💯

🔐

Quality Assurance (QA) of Synthetic Data

📈 **Statistical Properties**:
Ensure synthetic data mirrors the mean, variance, covariance, and correlations of the original.

📊 **Data Diversity**:
Cover real-world scenarios to avoid overfitting models.

📅 **Rare Events**:
Represent infrequent occurrences like rare diseases.

🔗 **Feature Relationships**:
Maintain original relationships to preserve complexity.

💯 **Realism**:
Ensure high quality, error-free, and realistic data for unbiased ML algorithms.

🔐

Quality Assurance (QA) of Synthetic Data

📈 **Statistical Properties**:
Ensure synthetic data mirrors the mean, variance, covariance, and correlations of the original.

📊 **Data Diversity**:
Cover real-world scenarios to avoid overfitting models.

📅 **Rare Events**:
Represent infrequent occurrences like rare diseases.

🔗 **Feature Relationships**:
Maintain original relationships to preserve complexity.

💯 **Realism**:
Ensure high quality, error-free, and realistic data for unbiased ML algorithms.

🔐 **Privacy**:
Remove sensitive details; apply k-anonymity and l-diversity

Case Study: French Motor Third-Party Liability Claims

Background: In the dataset freMTPL2freq risk features and claim numbers were collected for 677,991 motor third-part liability policies (observed on a year).

Utilise the ***tabularis.ai*** approach, we have generated a synthetic version of this dataset.

Original Data

	ClaimNb	Exposure	Area	VehPower	VehAge	DrivAge	BonusMalus	VehBrand	VehGas	Density	Region
0	1.0	0.10	D	5.0	0.0	55.0	50.0	B12	Regular	1217.0	R82
1	1.0	0.77	D	5.0	0.0	55.0	50.0	B12	Regular	1217.0	R82
2	1.0	0.75	B	6.0	2.0	52.0	50.0	B12	Diesel	54.0	R22
3	1.0	0.09	B	7.0	0.0	46.0	50.0	B12	Diesel	76.0	R72
4	1.0	0.84	B	7.0	0.0	46.0	50.0	B12	Diesel	76.0	R72

Artificial Data

	ClaimNb	Exposure	Area	VehPower	VehAge	DrivAge	BonusMalus	VehBrand	VehGas	Density	Region
0	0.0	0.20	D	4.0	9.0	52.0	50.0	B3	Regular	586.0	R41
1	0.0	0.12	B	5.0	6.0	20.0	95.0	B3	Diesel	98.0	R53
2	0.0	0.08	C	4.0	1.0	66.0	50.0	B12	Regular	301.0	R82
3	0.0	0.08	E	11.0	0.0	66.0	50.0	B12	Regular	3023.0	R11
4	0.0	1.00	C	6.0	12.0	39.0	50.0	B1	Regular	214.0	R26

Case Study: French Motor Third-Party Liability Claims

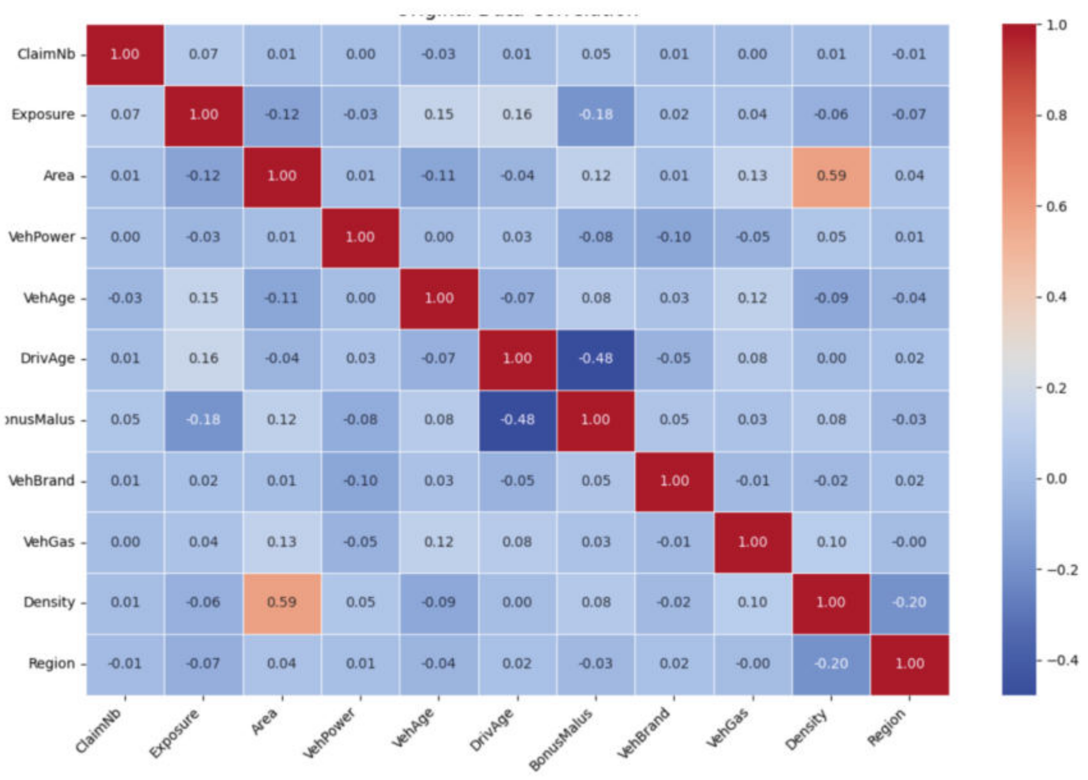
	ClaimNb	VehPower	VehAge	DrivAge	Density
count	678013.000000	678013.000000	678013.000000	678013.000000	678013.000000
mean	0.053247	6.454631	7.044265	45.499122	1792.422405
std	0.240117	2.050906	5.666232	14.137444	3958.646564
min	0.000000	4.000000	0.000000	18.000000	1.000000
25%	0.000000	5.000000	2.000000	34.000000	92.000000
50%	0.000000	6.000000	6.000000	44.000000	393.000000
75%	0.000000	7.000000	11.000000	55.000000	1658.000000
max	16.000000	15.000000	100.000000	100.000000	27000.000000

Original Data

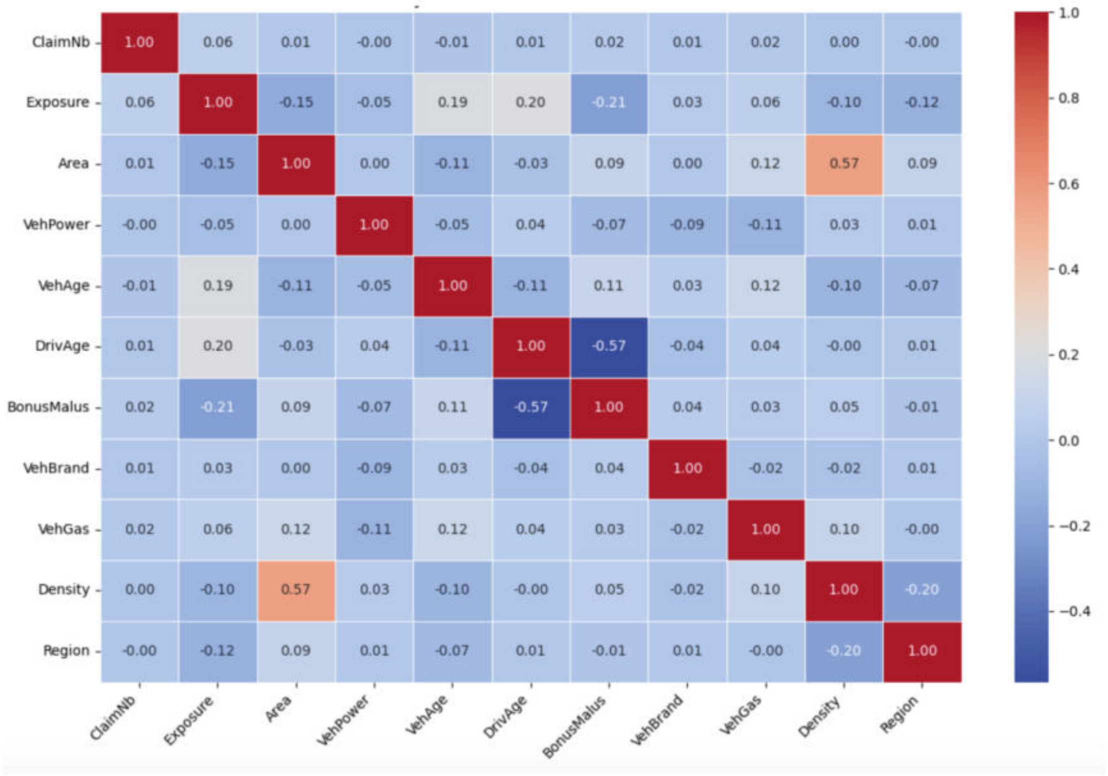
	ClaimNb	VehPower	VehAge	DrivAge	Density
count	45691.000000	45691.000000	45691.000000	45691.000000	45691.000000
mean	0.045458	6.215185	6.330459	44.731129	2201.560833
std	0.237103	1.623775	5.166845	13.360688	4644.534550
min	0.000000	4.000000	0.000000	18.000000	1.000000
25%	0.000000	5.000000	2.000000	34.000000	120.000000
50%	0.000000	6.000000	5.000000	43.000000	583.000000
75%	0.000000	7.000000	10.000000	53.000000	2715.000000
max	16.000000	15.000000	50.000000	99.000000	27000.000000

Artificial Data

QA Measure: Correlation

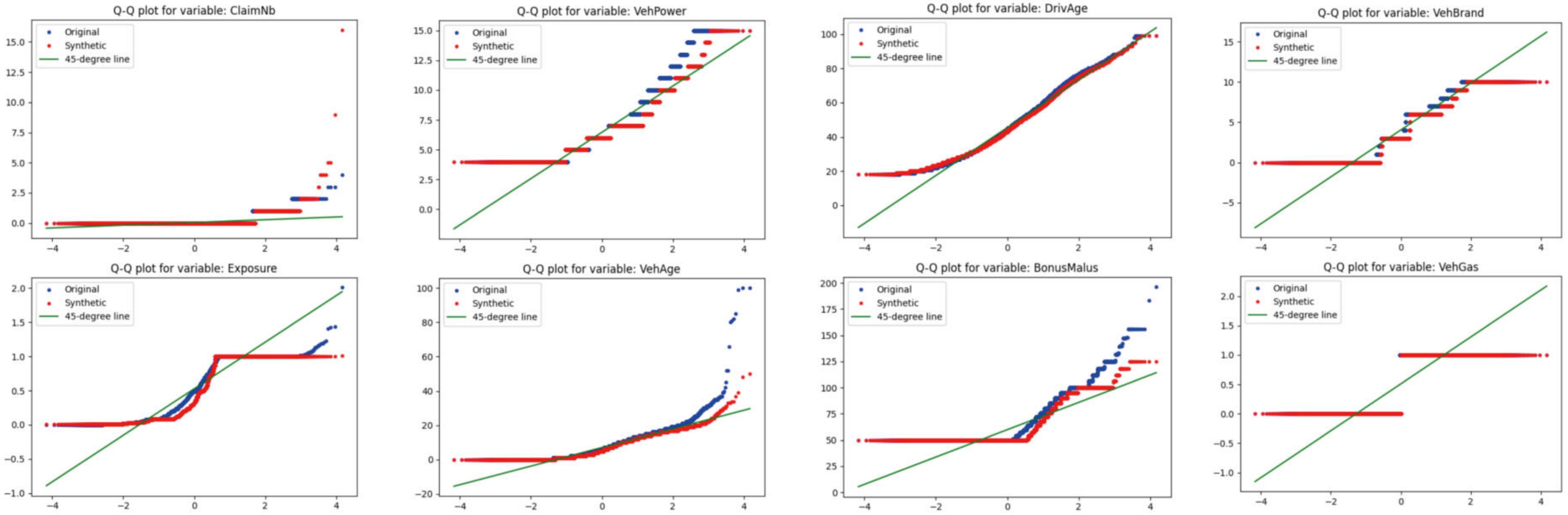


Original Data

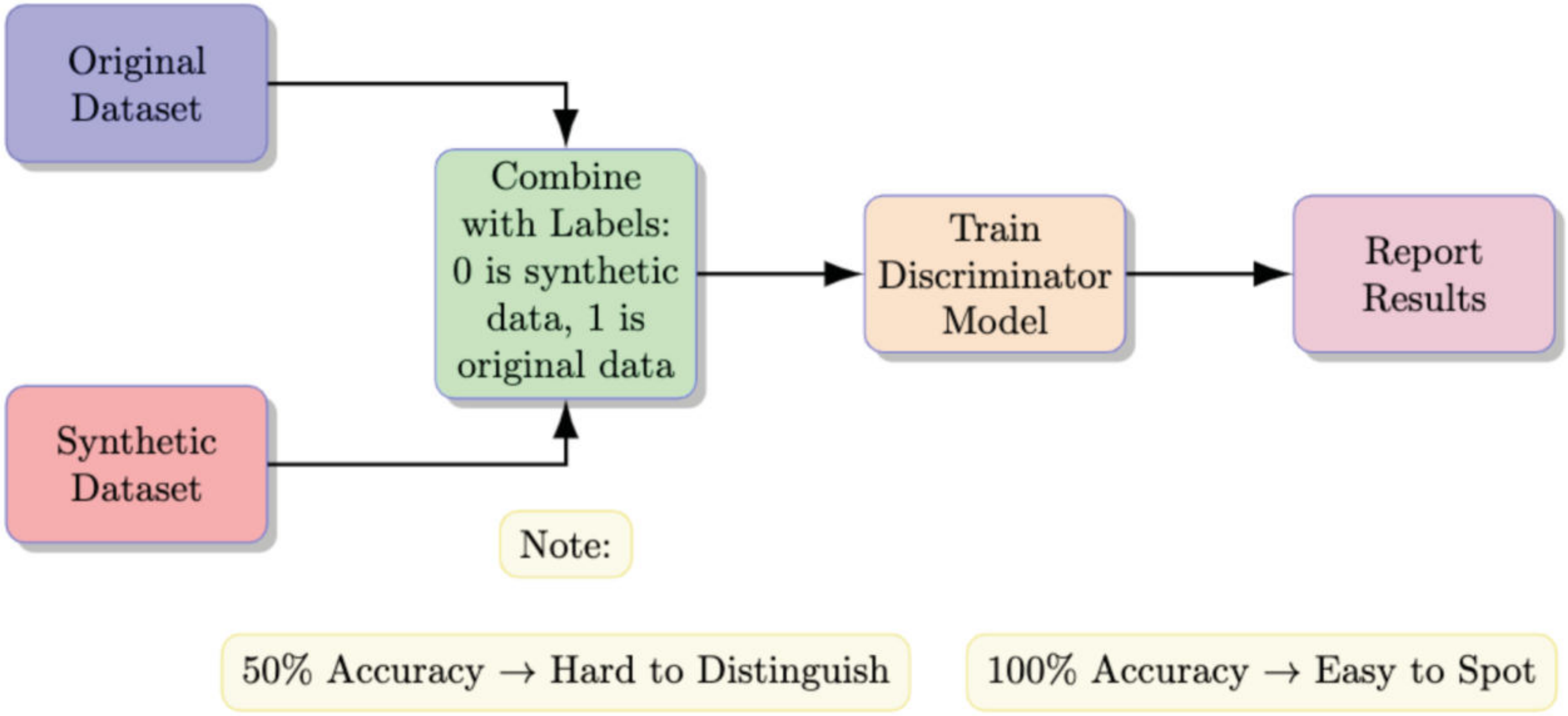


Synthetic Data

QA Measure: Q-Q Plots



QA Measure: Data Realism I



Results

Discriminator Model	Accuracy
Logistic Regression	0.55
Decision Tree	0.60
Random Forest	0.66

QA Measure: Data Realism II

Results

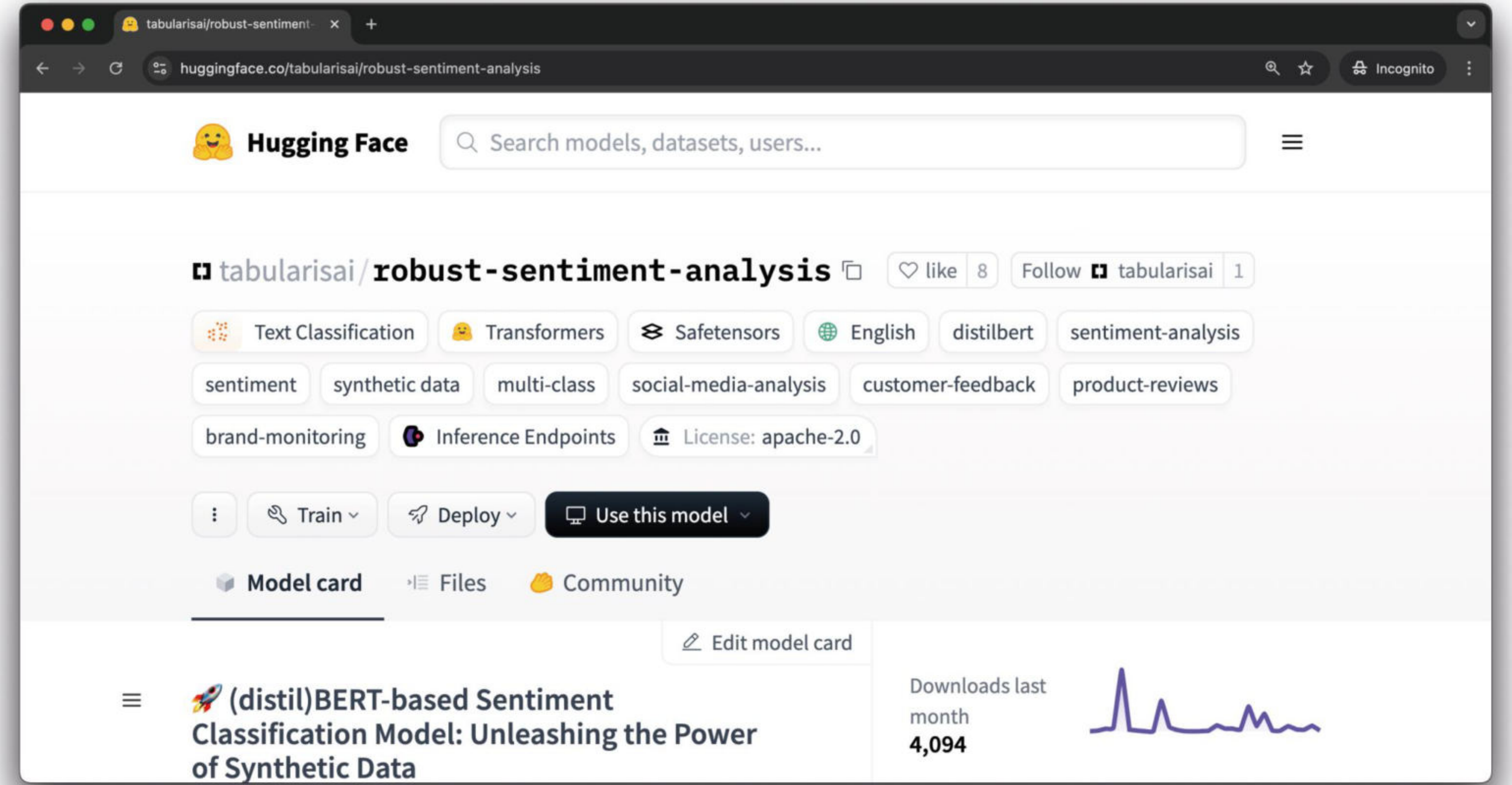
ML Model	Original Train Data	Synthetic Train Data
Logistic Regression	0.27+0.00	0.27+0.01
Random Forest	0.23+0.01	0.24+0.02

- **Goal:** Assess synthetic data's ability to replace real data in training by comparing model performance
- **Idea:** Test discriminative models trained on synthetic train data with real test data, using the mean squared error (regression)

QA Measure: Privacy

- k**-Anonymity Score**: ensures that each record is indistinguishable from at least $k - 1$ other records based on certain identifying attributes, thus preventing re-identification.
    - *Results: Original Data: 45, Synthetic Data: 63*
- l**-Diversity Score:** extends k-anonymity by ensuring that each group of indistinguishable records contains at least $l - 1$ well-represented sensitive values, thereby protecting against attribute disclosure.
    - *Results: Original Data: 32, Synthetic Data: 39*
- **DOMIAS Membership Inference Attack (MIA)**: MIA measures the vulnerability of a dataset to attacks that determine whether a specific record was part of the training data, assessing the robustness of privacy preservation [van Breugel, 2023]
    - *Results: Accuracy: 0.5009, AUC-ROC: 0.5782*

Synthetic Data is Useful Too



Summary

Traditional Anonymization Techniques:

- Often fail to protect against sophisticated attacks

Large Language Models (LLMs):

- Generate hyperrealistic synthetic data
- Preserve statistical properties and enhance privacy

Evaluation of Synthetic Data:

- Essential to ensure data quality and privacy
- Must maintain key statistical properties, cover diverse scenarios, and represent rare events

Thank you for your attention!