MOSTLY·AI

Data Without Barriers

Synthetic Data as a Catalyst for Responsible Innovation

Paul Tiwald
paul.tiwald@mostly.ai

Data Democratization

Data Access

Data Insights

Synthetic Data

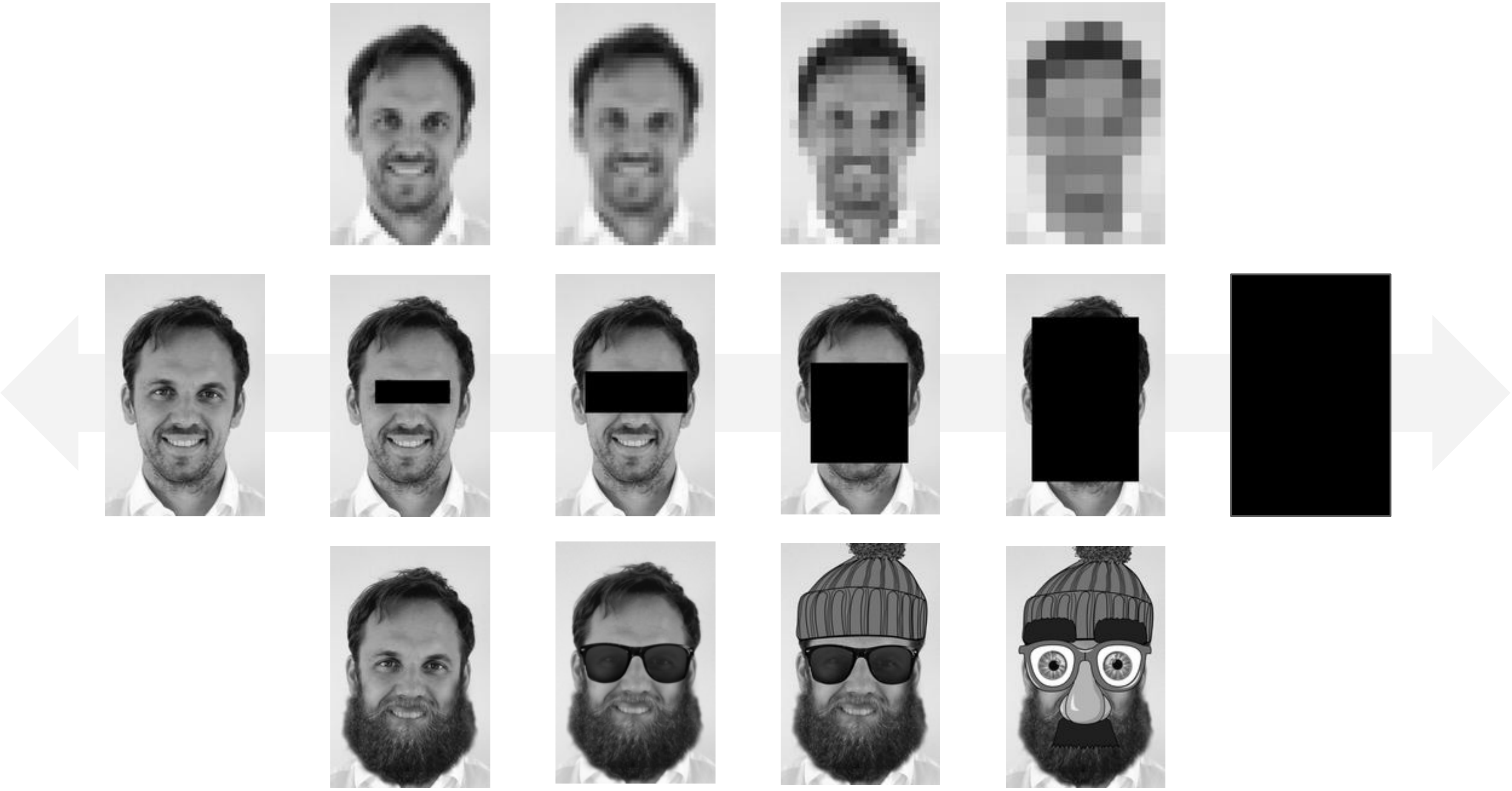
power of LLMs/assistants

… for everyone

MOSTLY·AI

Why Synthetic? … Real Data has its issues

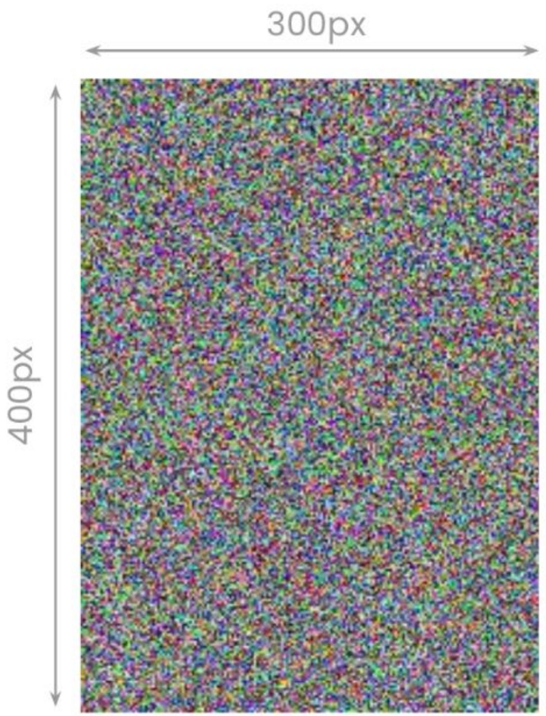
useful, but
re-identifiable



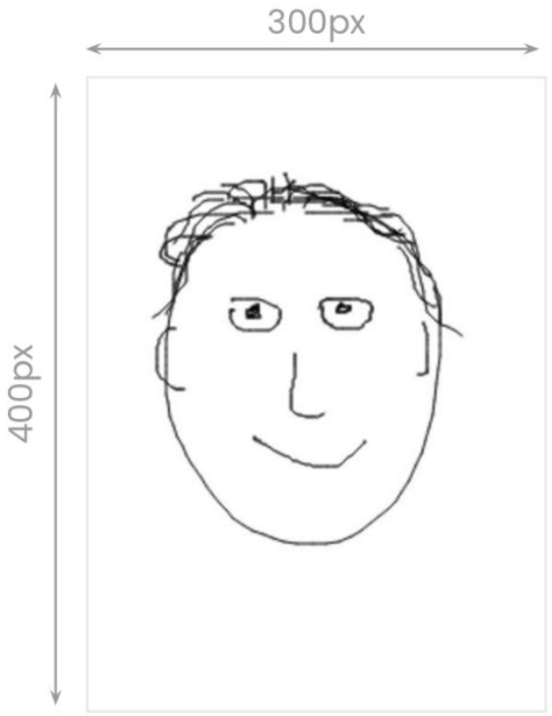
private, but
useless

MOSTLY·AI

What is Synthetic Data



random data



self-generated data

model-generated data
rule-based

AI-generated data
“data-based”

Synthetic Data = Generative AI

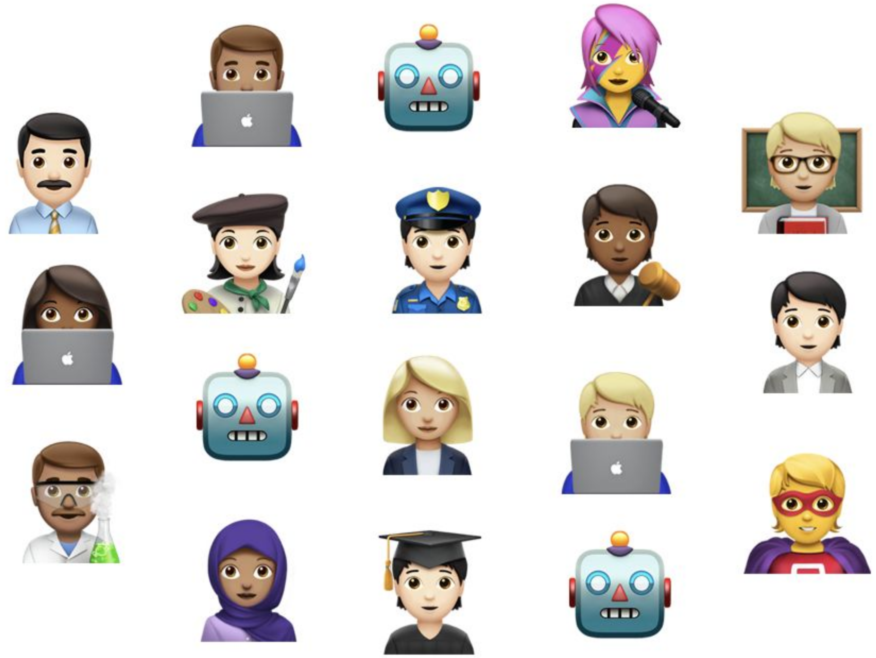
Actual Data

privacy-restricted
biased
incomplete

Generative Model

Synthetic Data

realistic
representative
anonymous
granular level



Data Consumers

people & algorithms

MOSTLY·AI

Tabular ARGN

MOSTLY·AI

Tabular ARGN - Auto-Regressiv Generative Networks

arXiv > cs > arXiv:2501.12012

Search...

Help

Computer Science > Machine Learning

[Submitted on 21 Jan 2025 (v1), last revised 6 Feb 2025 (this version, v2)]

TabularARGN: A Flexible and Efficient Auto-Regressive Framework for Generating High-Fidelity Synthetic Data

Paul Tiwald, Ivona Krchova, Andrey Sidorenko, Mariana Vargas Vieyra, Mario Scriminaci, Michael Platzer

Synthetic data generation for tabular datasets must balance fidelity, efficiency, and versatility to meet the demands of real-world applications. We introduce the Tabular Auto-Regressive Generative Network (TabularARGN), a flexible framework designed to handle mixed-type, multivariate, and sequential datasets. By training on all possible conditional probabilities, TabularARGN supports advanced features such as fairness-aware generation, imputation, and conditional generation on any subset of columns. The framework achieves state-of-the-art synthetic data quality while significantly reducing training and inference times, making it ideal for large-scale datasets with diverse structures. Evaluated across established benchmarks, including realistic datasets with complex relationships, TabularARGN demonstrates its capability to synthesize high-quality data efficiently. By unifying flexibility and performance, this framework paves the way for practical synthetic data generation across industries.

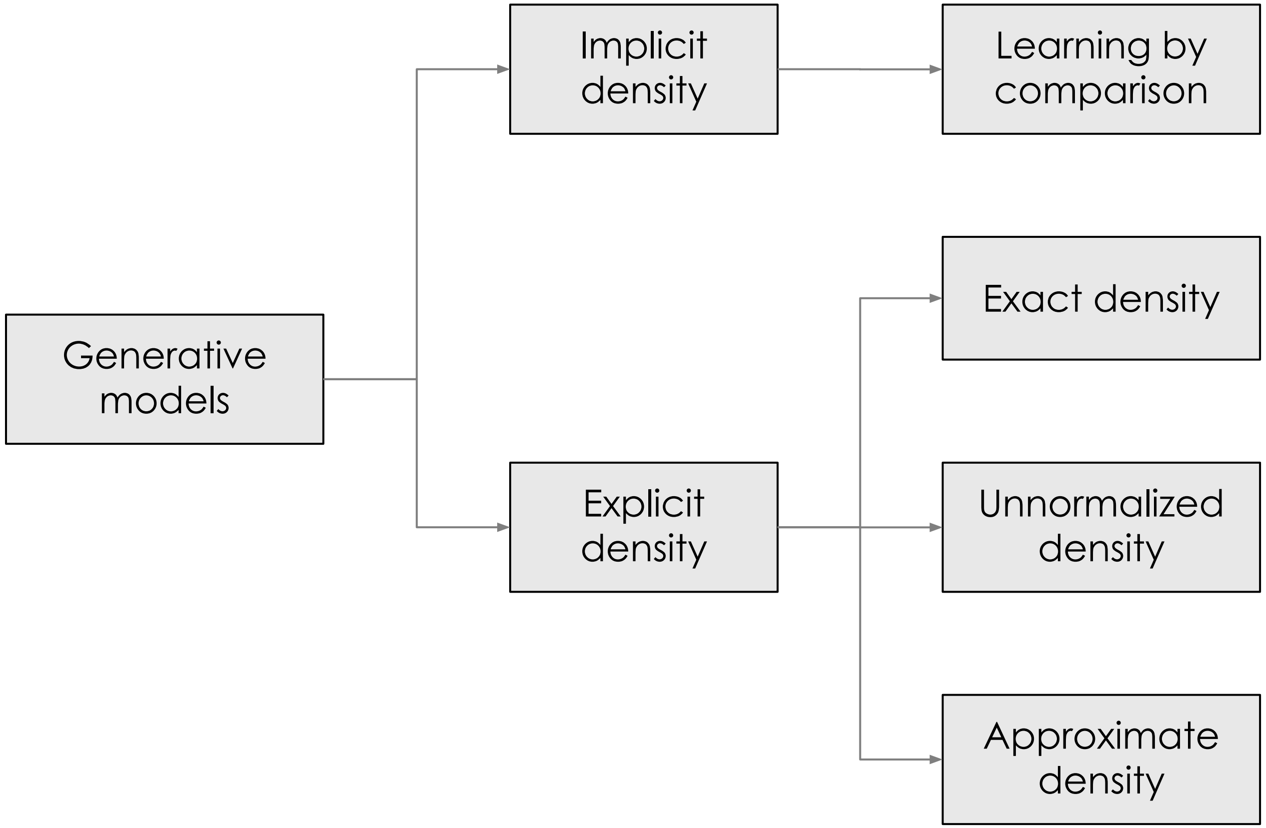
Tabular ARGN is implemented in the Synthetic Data SDK:

https://github.com/mostly-ai/mostlyai

```
pip install -U mostlyai[local]
```

MOSTLY·AI

Taxonomy of deep generative models



Learning by comparison
- only care about generation, not **p(x)**
- instead of maximizing the density, compare real vs generated sample (classification problem)
- examples: GAN, GMMN

Exact density
- directly learn density **p(x)**
- examples: autoregressive models (Transformer, RNNs), flow-based models

Unnormalized density
- learn unnormalized density **E(x)∝ p(x)**
- examples: EBM

Approximate density
- learn approximation (e.g. lower bound) of density **L(x) ≤ p(x)**
- examples: VAE, diffusion models

MOSTLY·AI

Flat Model

Fixed (column) Order Training Phase

patients data set:

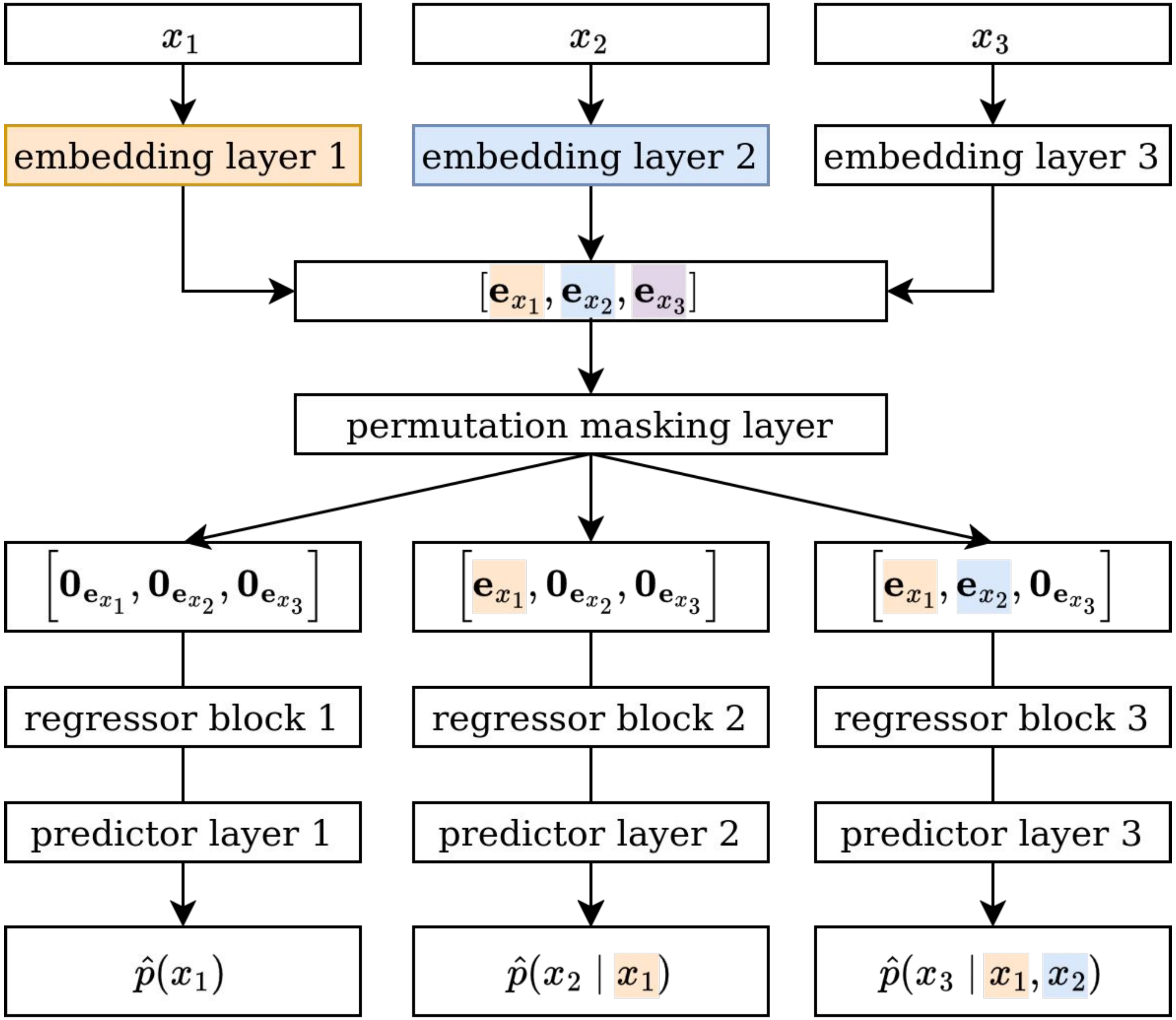
x1 - age

x2 - gender

x3 - blood type

loss function:

$$\max_{\theta} \sum_{i=1}^{D} \log p_\theta(x_i \mid x_{<i})$$



Flat Model

Any (column) Order Training Phase

patients data set:

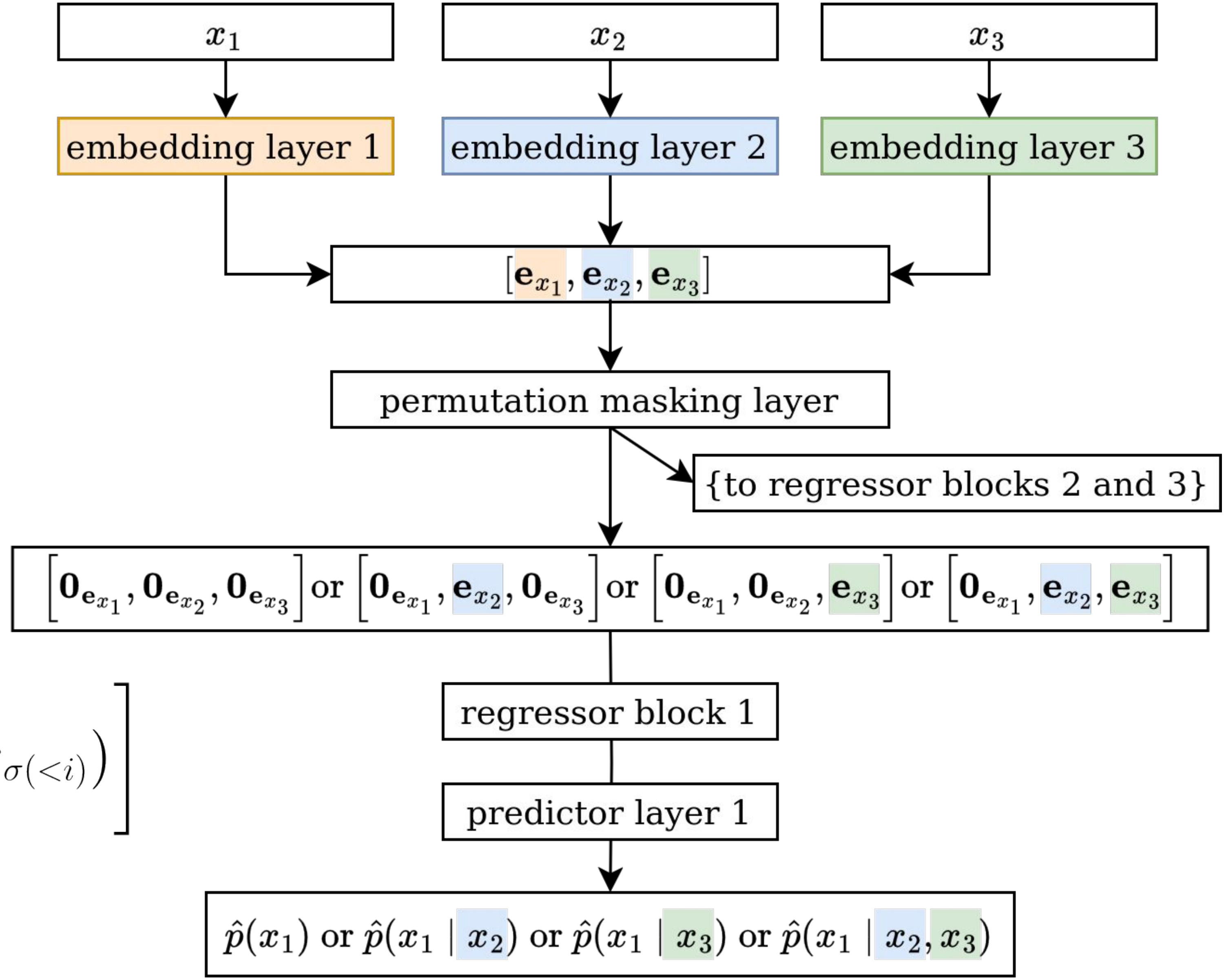
x1 - age

x2 - gender

x3 - blood type

loss function:

$$\max_{\theta} \ \mathbb{E}_{\sigma \in \mathrm{Uniform}(S_D)} \left[\sum_{i=1}^{D} \log p_\theta \left(x_{\sigma(i)} \mid x_{\sigma(<i)} \right) \right]$$



MOSTLY·AI

Flat Model

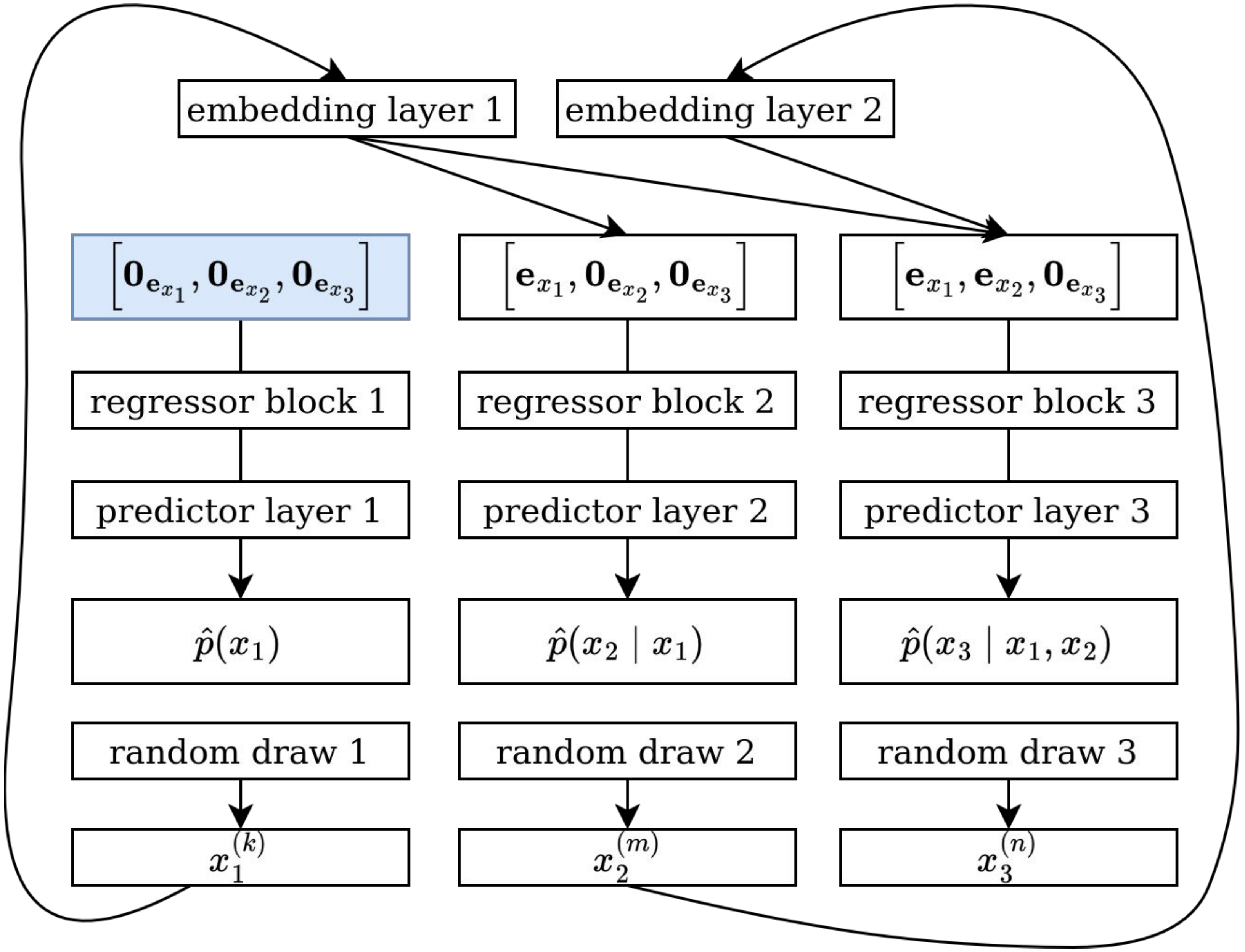
Generation Phase

patients data set:

x1 - age

x2 - gender

x3 - blood type



Sequential Model - doctor visits

auto-regressive along the column and the time dimensions

↓ time

date	systolic blood pressure	body weight
15.1.2022	140	85
19.1.2022	138	85
3.4.2022	135	84
16.7.2022	132	?

history encoder

$$\hat{p}(\text{body weight} \mid \text{date, blood pressure}, \text{encoded history})$$

MOSTLY·AI

Sequential Model with context

auto-regressive along the column, time, and table dimensions

← time

date	systolic blood pressure	body weight
15.1.2022	140	85
19.1.2022	138	85
3.4.2022	135	84
16.7.2022	132	?

age	gender	blood type
36	m	A+

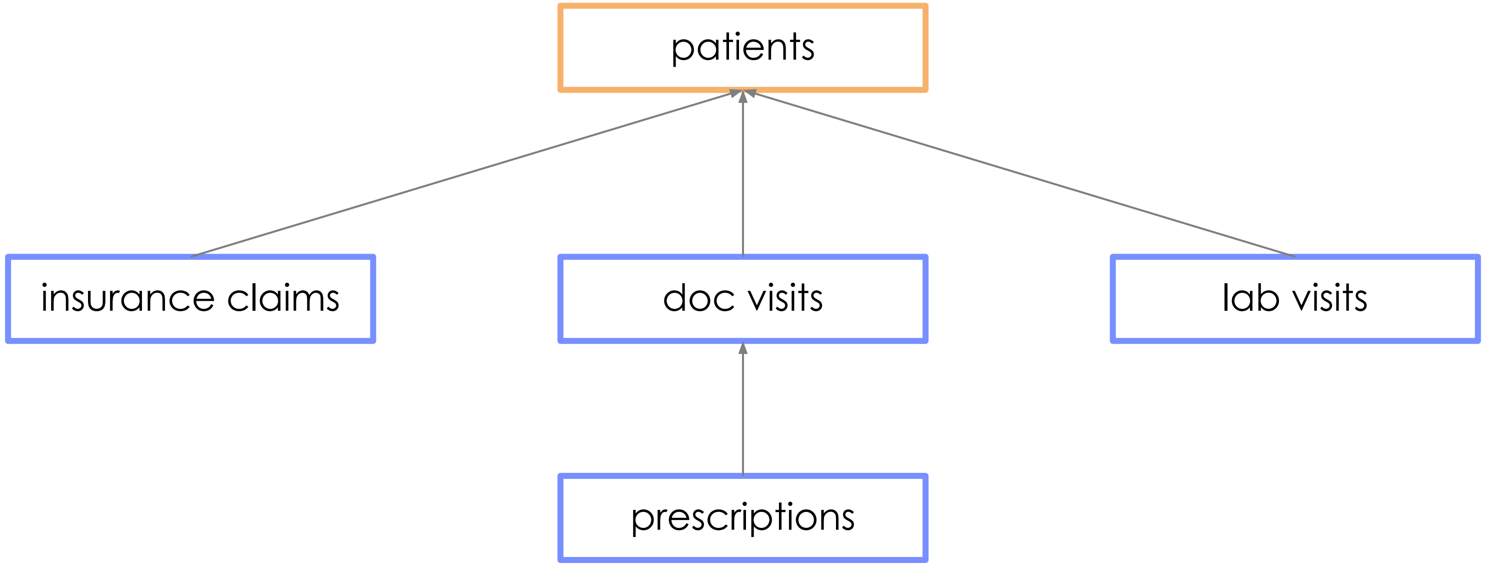
history encoder

context proc.

$$\hat{p}(\text{body weight} \mid \text{date, blood pressure}, \text{encoded history}, \text{context})$$

MOSTLY·AI

Flexible context allows for synthesis of multi-table setups



Synthetic Data = Generative AI

Actual Data

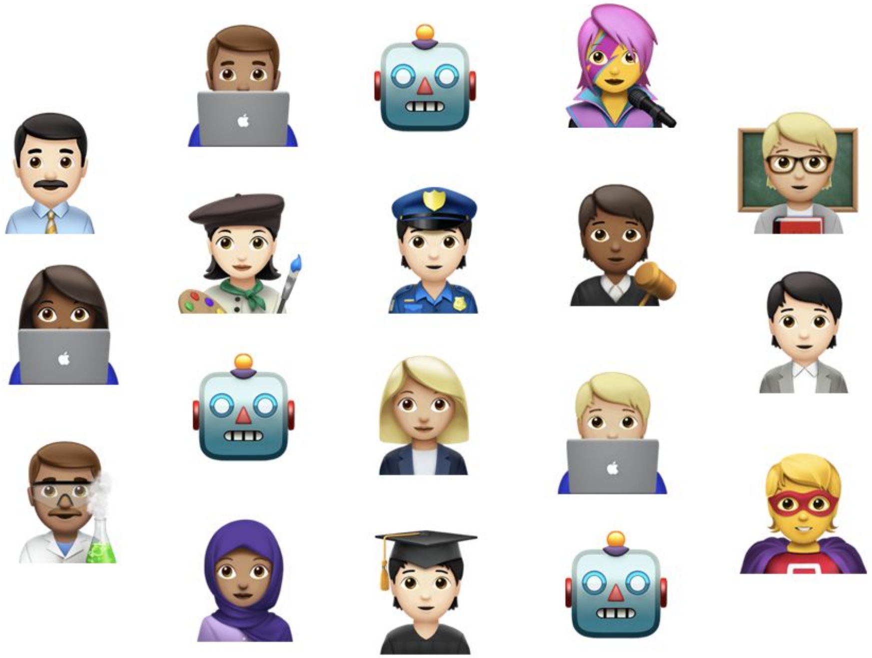
privacy-restricted
biased
incomplete

AI

Generative Model

Synthetic Data

realistic
representative
anonymous
granular level



Data Consumers

people & algorithms

MOSTLY·AI

The main use case: Privacy and Data Access

Privacy/Data Access - the main use case

- reducing the "time-to-data"

Privacy/Data Access - the main use case

- reducing the “time-to-data”
- breaking down data silos within organizations (e.g. synthetic-data products)

MOSTLY·AI

Know your customer

MOSTLY•AI

Privacy/Data Access - the main use case

- reducing the “time-to-data”

- breaking down data silos within organizations (e.g. synthetic-data products)
- share data between subsidiaries in different countries
- share data between organizations (e.g. *clean rooms*)

MOSTLY·AI

Privacy/Data Access - the main use case

- reducing the “time-to-data”

- breaking down data silos within organizations (e.g. synthetic-data products)
- share data between subsidiaries in different countries
- share data between organizations (e.g. clean rooms)

- open-data initiatives by public entities

make the information in EPCs accessible

EPC - Analysis | Synthetic EPC

Energy Performance Certificate - Geoclustering

Powered by MOSTLY·AI

Analysis

Analysis of EPCs

DPR412_classification	construction_year	degree_days	altitude	floors	net_area	heat_loss_surface
13	2017	2981	537	0	622.69	1664
8	1970	3396	840	0	164.15	579.5
8	2005	2426	197	1	129.18	487.23
4	1900	2765	354	1	74.37	228.65
13	1975	2591	265	0	2532.5	5743.9
5	1940	2778	283	0	6678.71	11930.02
6	1930	2741	345	0	299.21	1340.18
4	1900	3224	664	0	145.38	623.28
4	1992	2528	172	2	197.347	413.73
12	2022	2961	405	0	81.68	801.35

Rows per page 10 1-10 of 2519

MODERATE

Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for.

About MODERATE

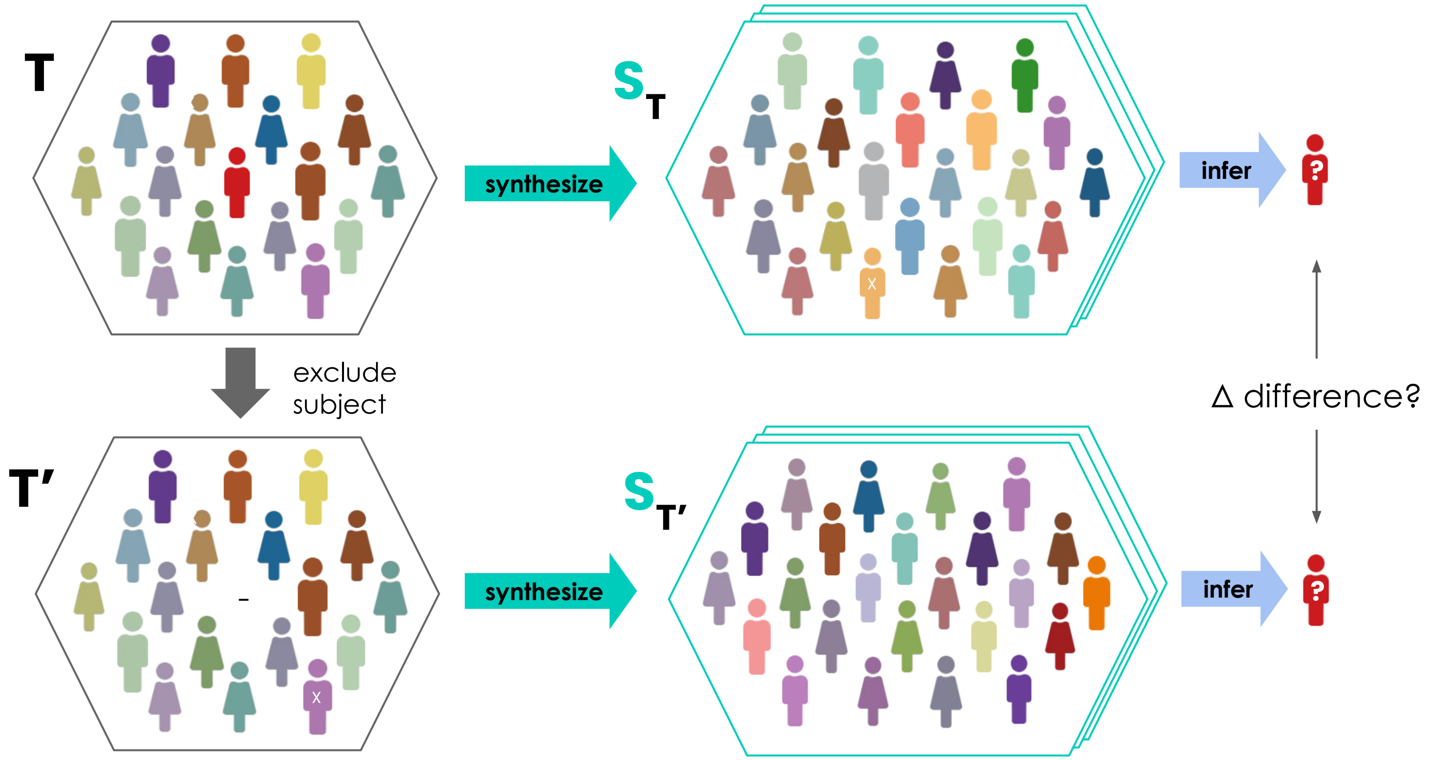
In a nutshell

News

Contact

https://tools.eeb.eurac.edu/epc_clustering/piemonte/

How to test Synthetic-Data Privacy



We must not be able to infer more about an individual, when that person is included in the database used for synthesis.

How to test Synthetic-Data Privacy - Empirically

Accuracy scores for 50 randomly chosen subjects, that were part of training

	NB	SVM	KNN	RF	LR	FRNN	ENS	DUM	**RMEAN**
Target **T**	42.8±5	43±6.6	41.6±9.3	49.8±9.4	38.3±3.7	49.1±9.8	45±6.8	32	**44.2±3.8**
Synthetic $\mathbf{S_T}$	42.1±4.2	39.5±7.3	36.2±6.8	37.9±5.9	36.5±6	37±6.3	39.7±6.1	32	**38.4±2**

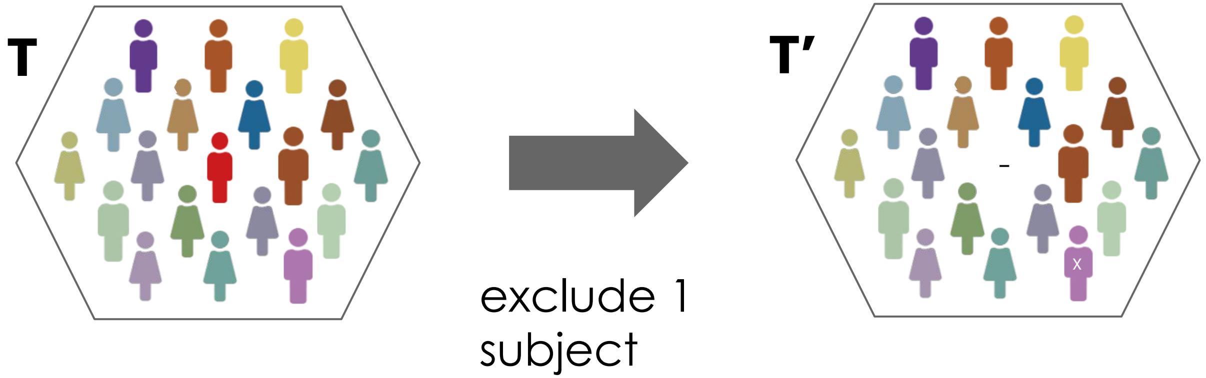
Accuracy scores for 50 randomly chosen subjects, that were NOT part of training

	NB	SVM	KNN	RF	LR	FRNN	ENS	DUM	**RMEAN**
Target **T'**	42.1±5	39.9±7.1	36.9±5.7	39.7±5.1	37.6±4.1	39.1±5	39.7±5.9	32	**39.3±1.6**
Synthetic $\mathbf{S_{T'}}$	43.7±4.2	40.4±6.4	35.9±6.4	39.1±6.2	36.9±4.6	38±5.8	40.5±7	32	**39.2±2.4**

SBA Research, 2020

How to test Synthetic-Data Privacy - Mathematically

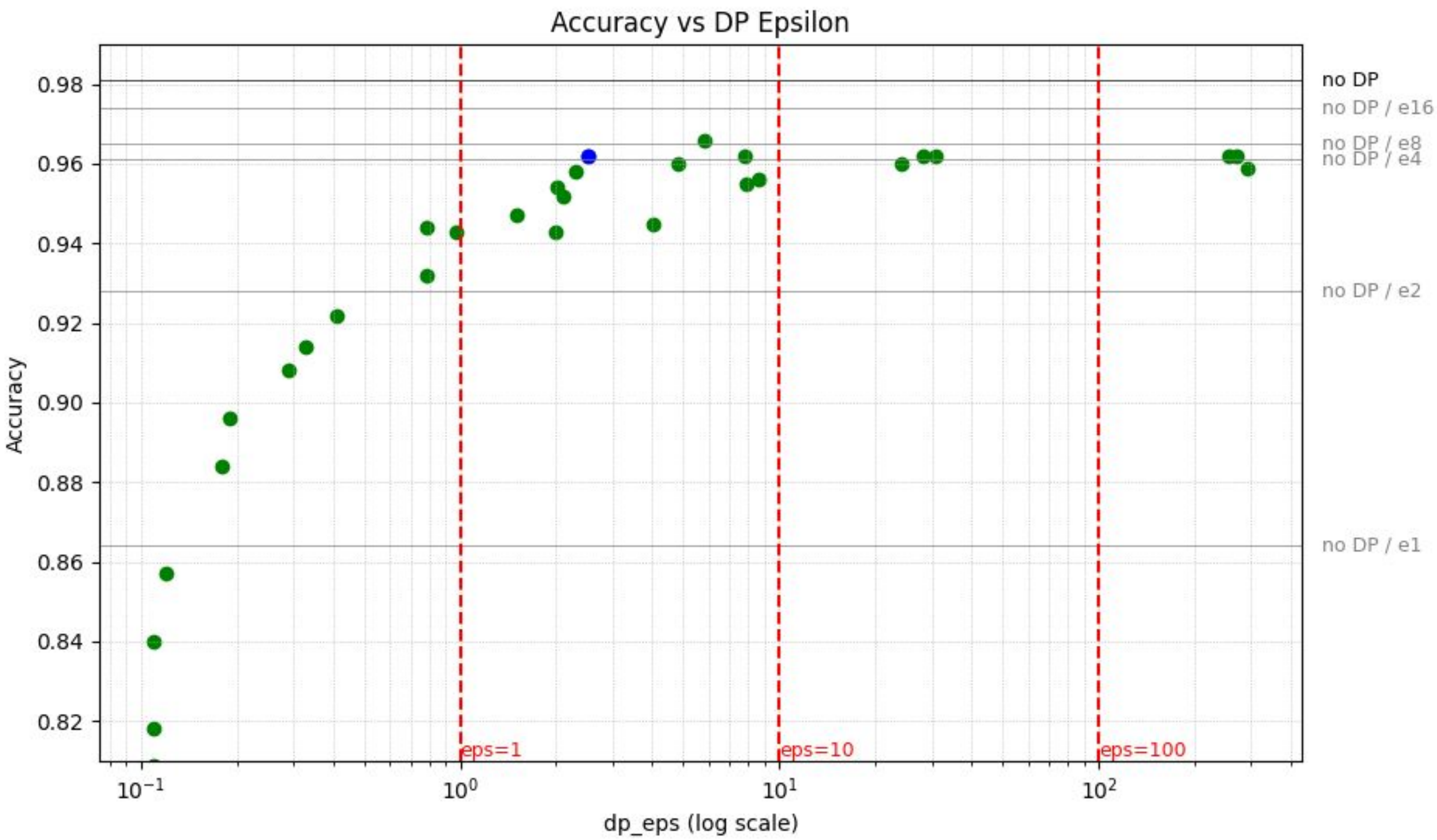
$$\Pr[\mathcal{A}(T)] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(T')]$$



- gold standard definition of privacy
- idea: limit influence of single individuals
- provides a mathematical guaranteed upper bound (**ε**) for the difference in outcomes of an algorithm A applied to the adjacent data sets T and T'

MOSTLY·AI

Differential Privacy



MOSTLY·AI

ML on synthetic data

MOSTLY·AI

Synthetic Switzerland

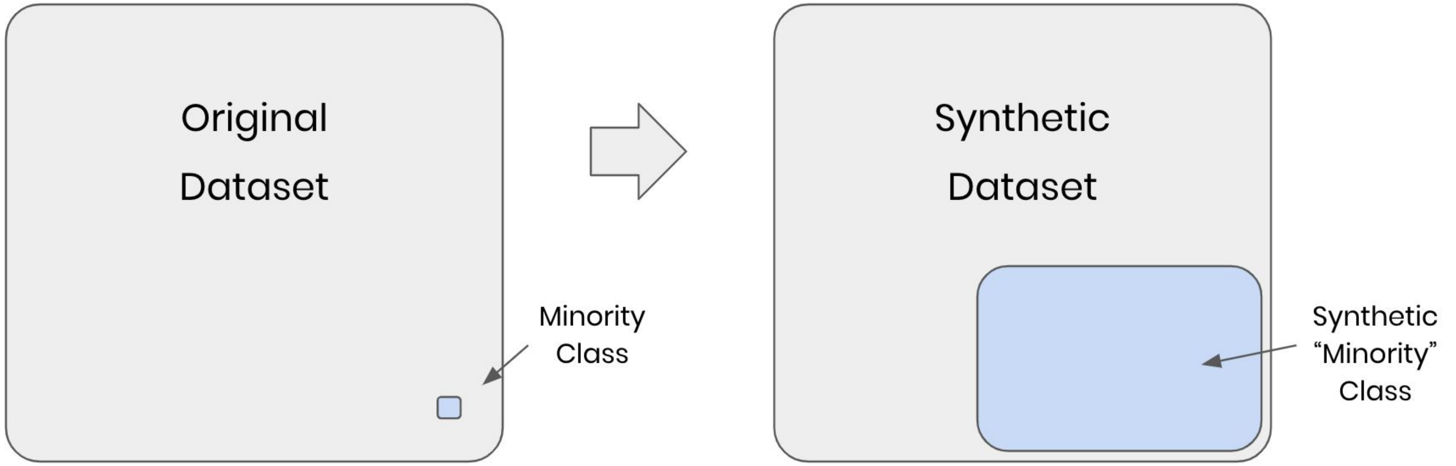
https://www.youtube.com/watch?v=Rt5gXcIc0jY

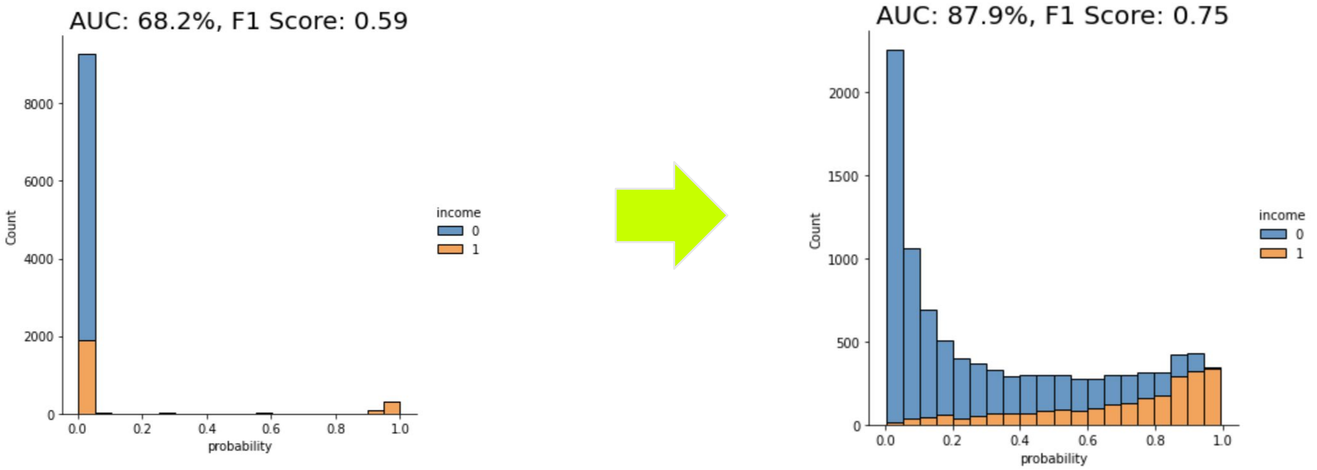
- Switzerland has around 9 Mio. Inhabitants
- Only 15% (1/9) provide explicit consent for marketing use
- 8/9 remain locked behind privacy
- Synthetic Switzerland mirrors population patterns without real identities
- We can now analyse and train models with 9/9 citizens instead of 1/9
- Zero personal data is used

MOSTLY·AI

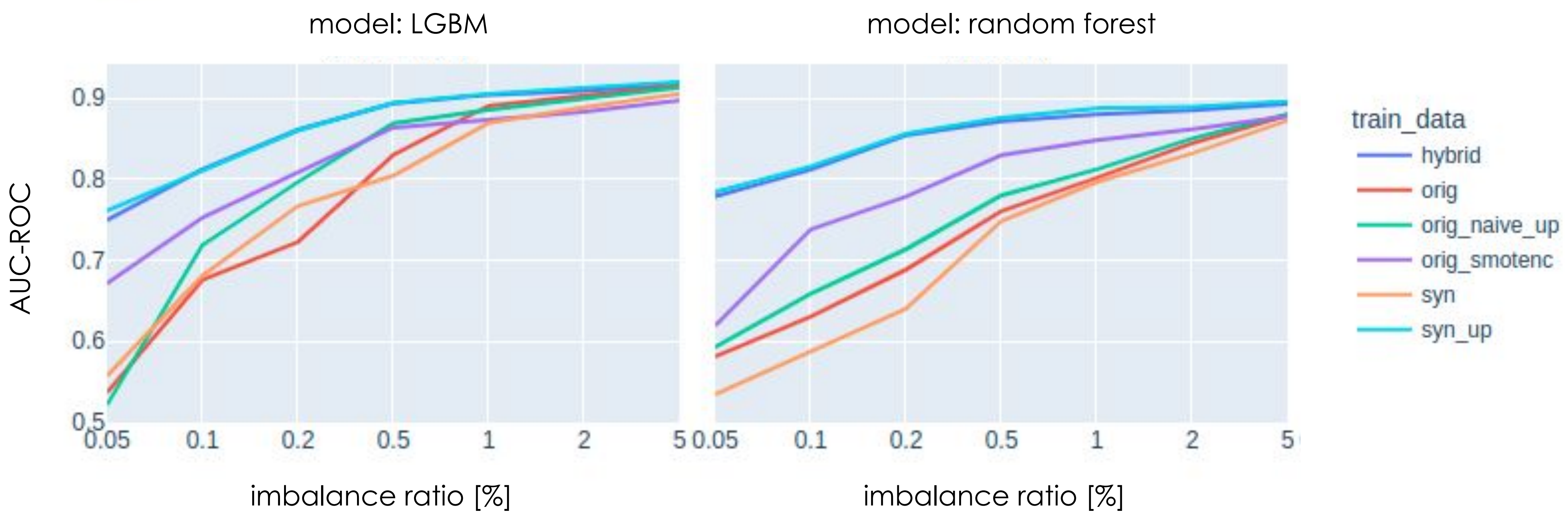
Rebalancing of underrepresented Classes

Open in Colab



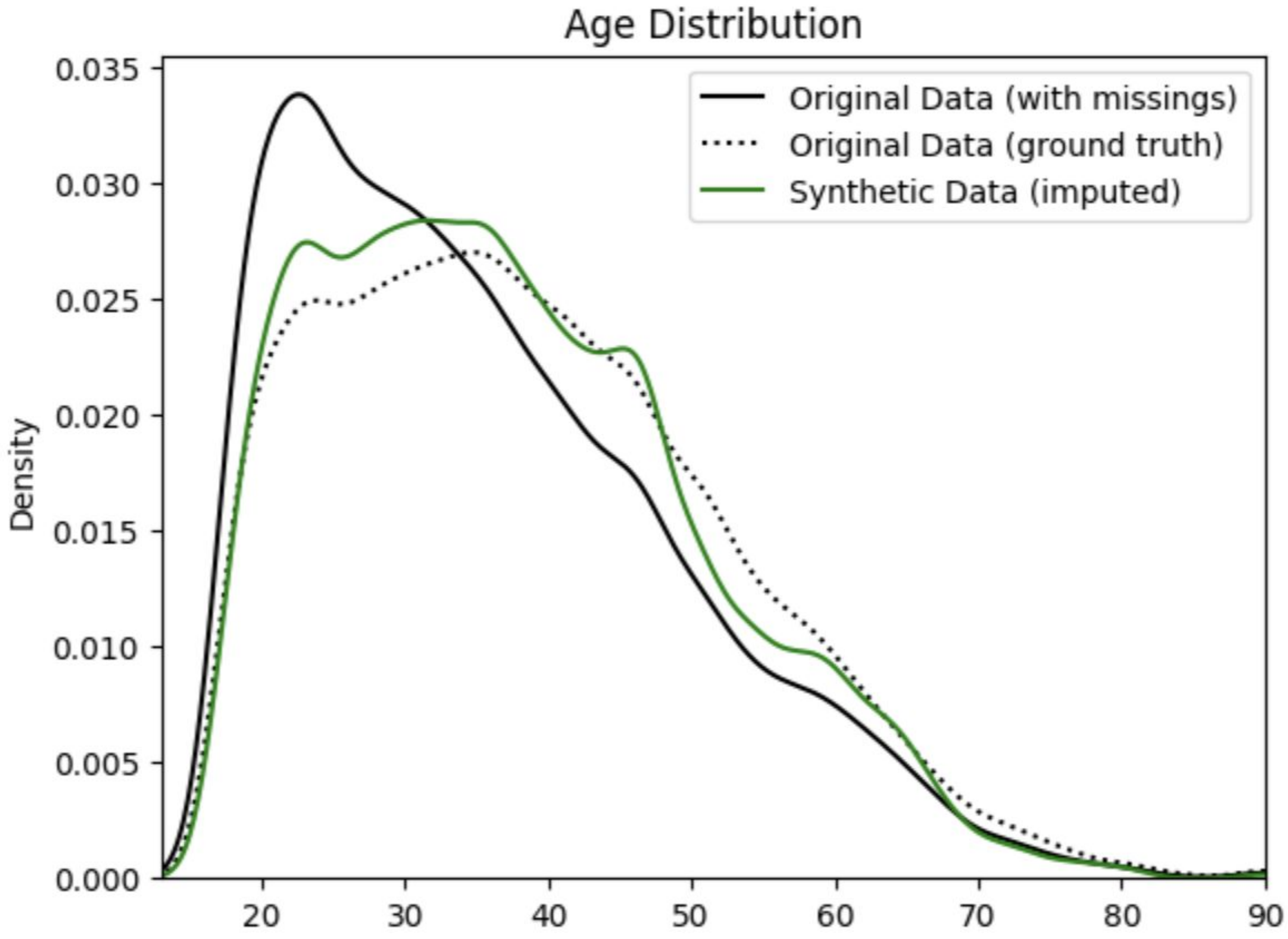


Rebalancing of underrepresented Classes



Imputation

Smart Imputation of Missing Data



Open in Colab

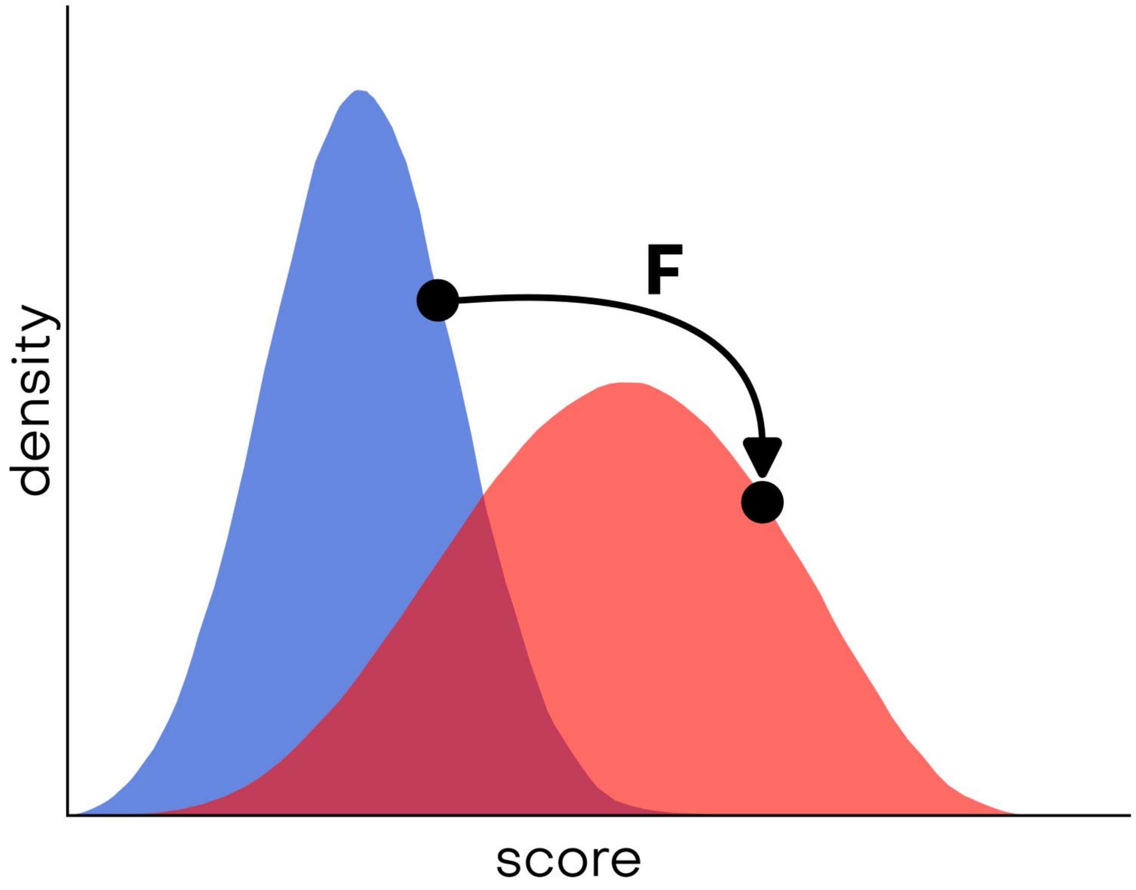
MOSTLY·AI

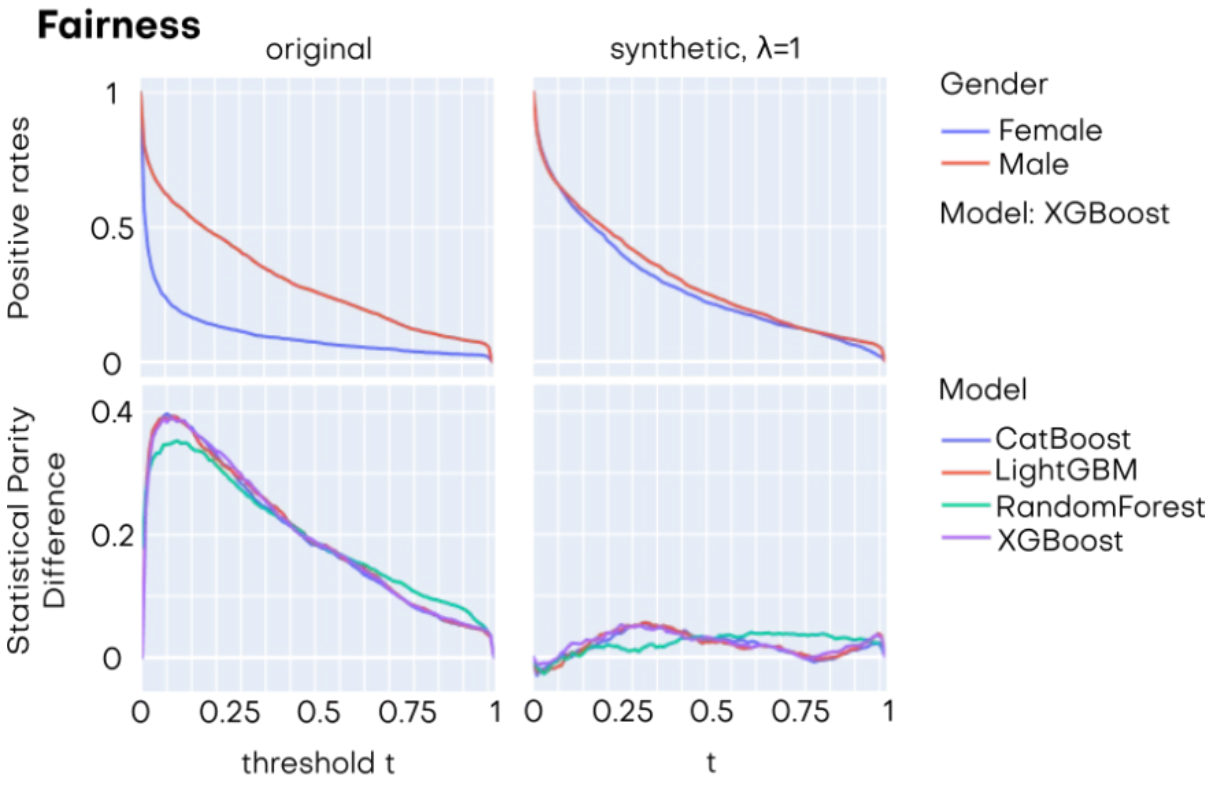
Fair synthetic data

Fair Synthetic Data

… subject to Strong Statistical Parity

https://openreview.net/pdf?id=HbU5QuPZj6



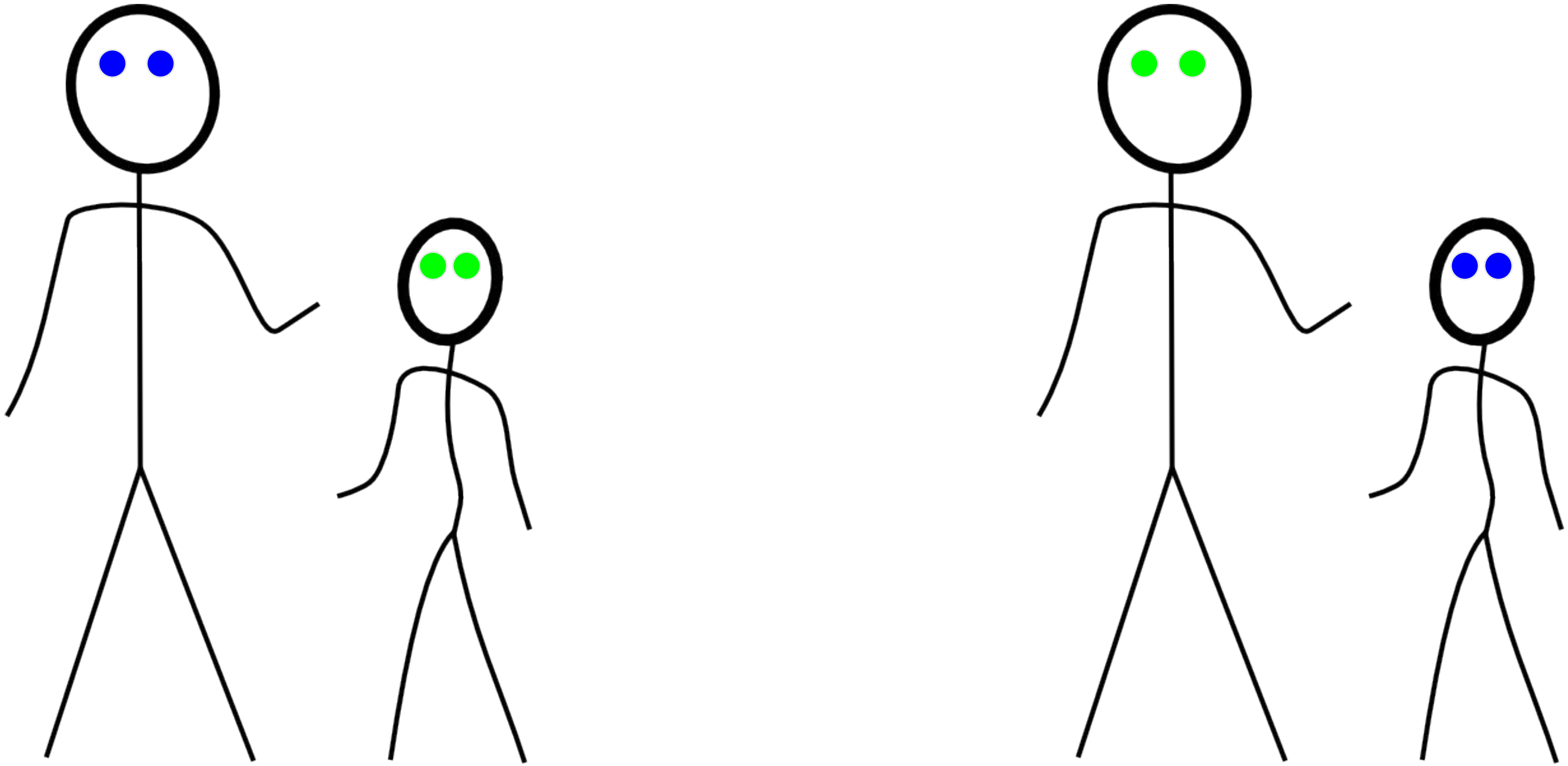


MOSTLY·AI

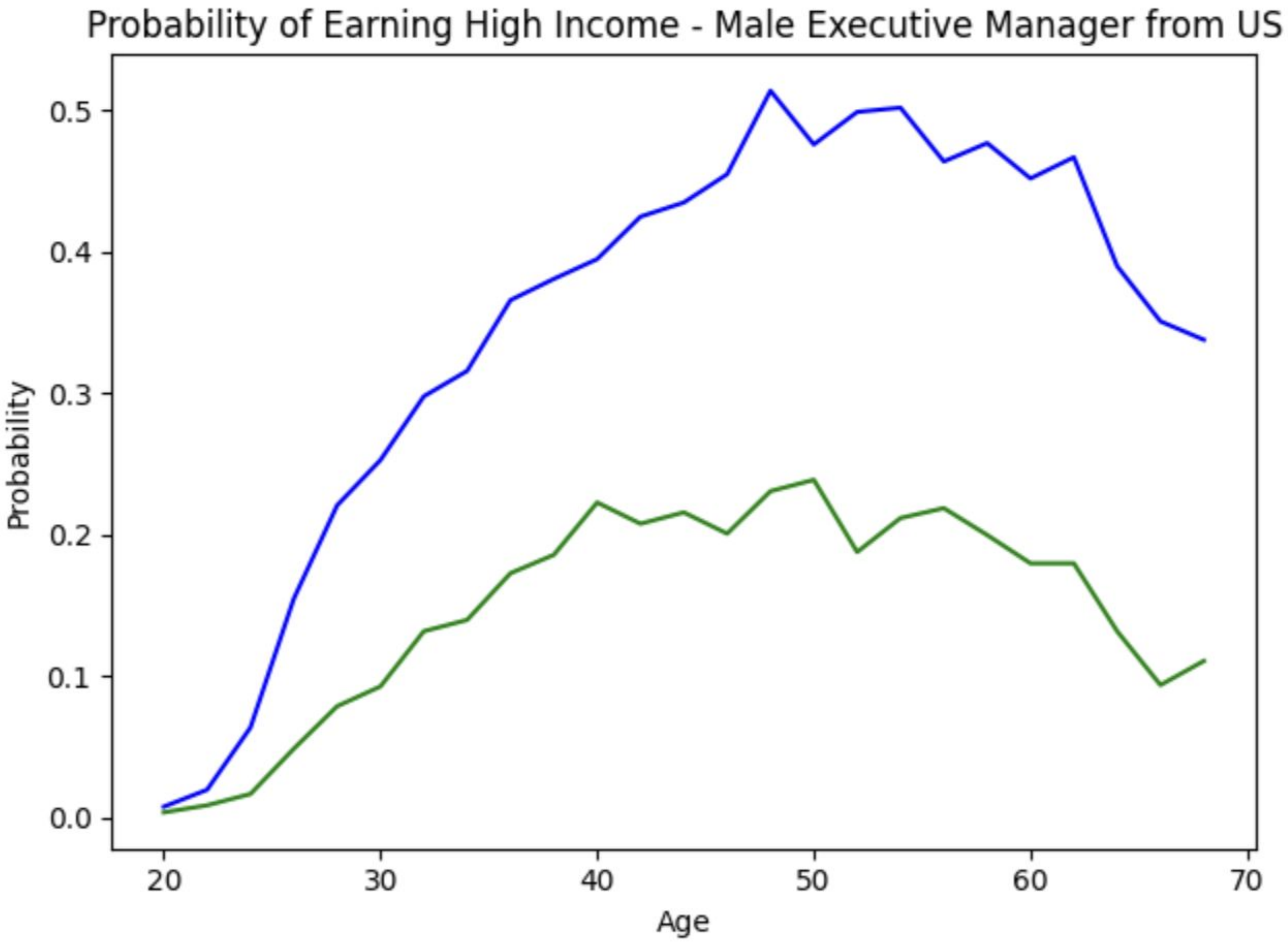
“Simulation”

MOSTLY·AI

Simulation with Synthetic Data



Simulation through Flexible Conditional Generation



Open in Colab

MOSTLY·AI

Simulation with Synthetic Data

ELSEVIER

International Journal of Research in Marketing

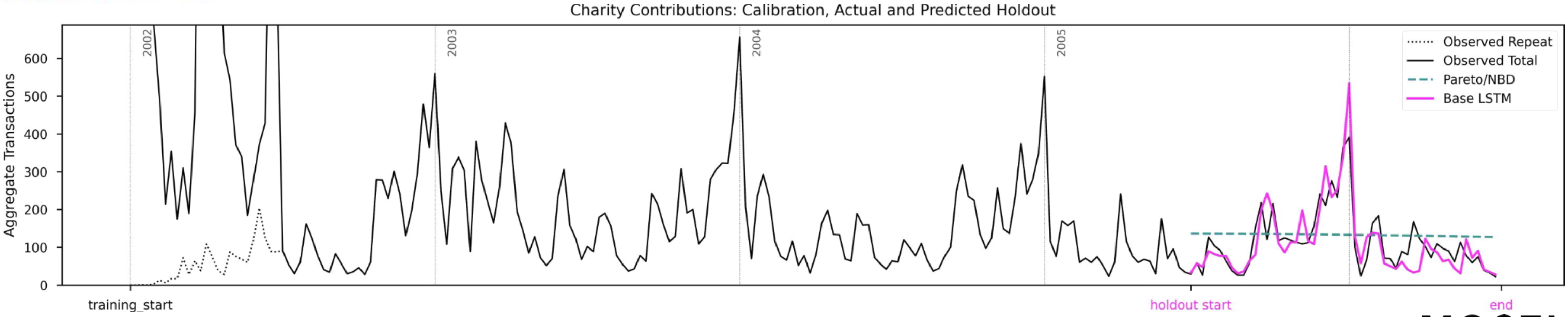
Volume 39, Issue 4, December 2022, Pages 988-1018

Full length article

Customer base analysis with recurrent neural networks

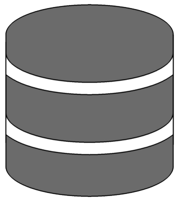
Jan Valendin [a], Thomas Reutterer [a], Michael Platzer [b], Klaudius Kalcher [b]

Customer base analysis with recurrent neural networks



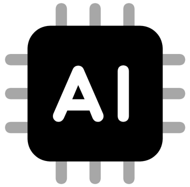
MOSTLY·AI

Synthetic Data = Generative AI

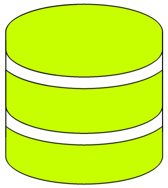


Actual Data

privacy-restricted
biased
incomplete

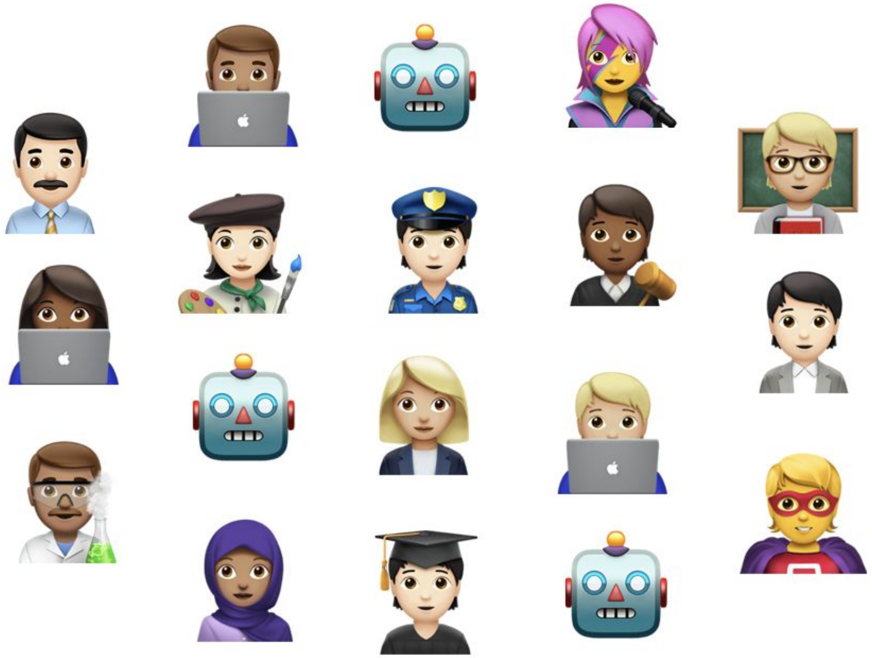


Generative Model



Synthetic Data

realistic
representative
anonymous
granular level

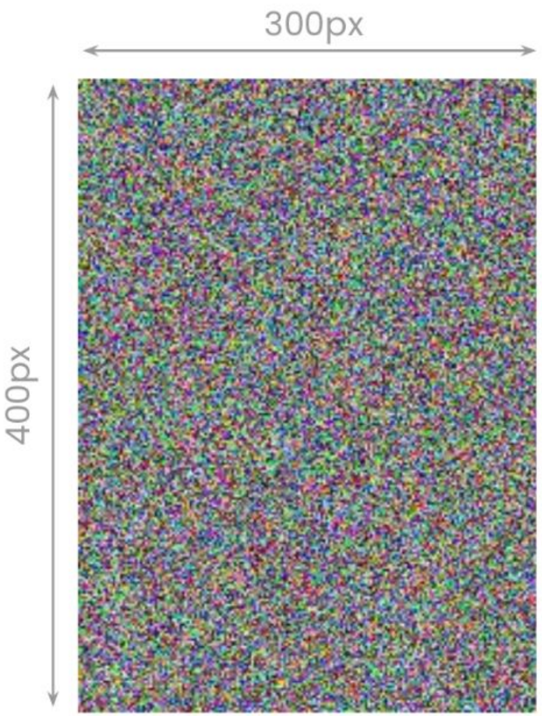


Data Consumers

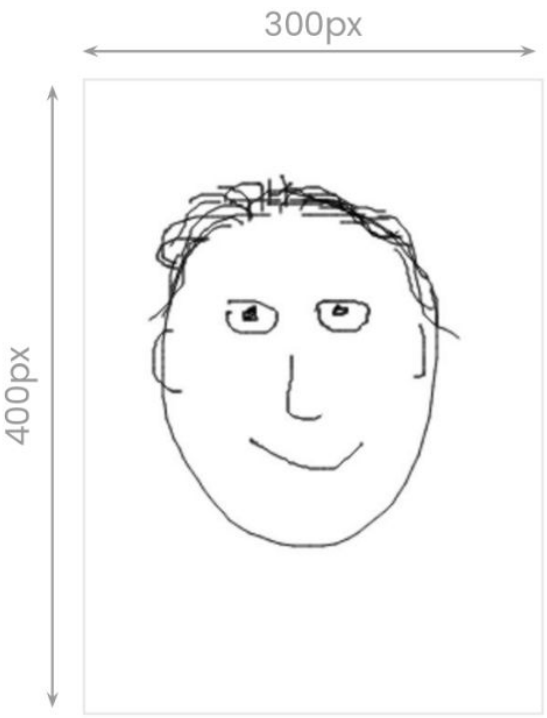
people & algorithms

MOSTLY·AI

What is Synthetic Data



random data

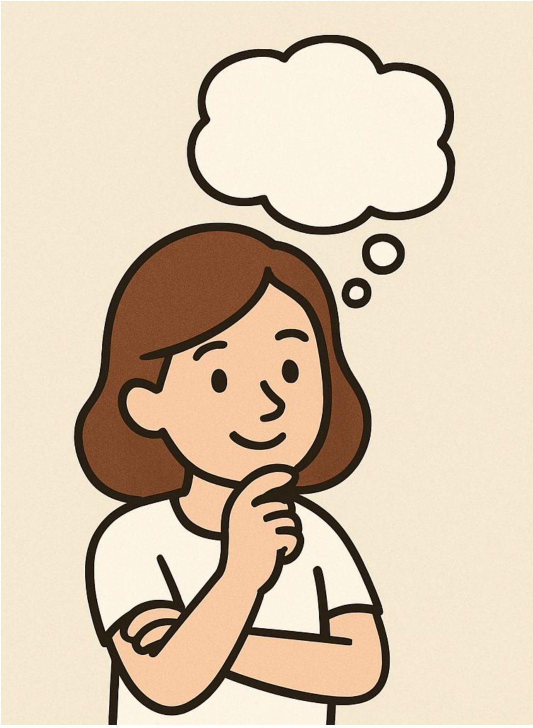


self-generated data

model-generated data
rule-based

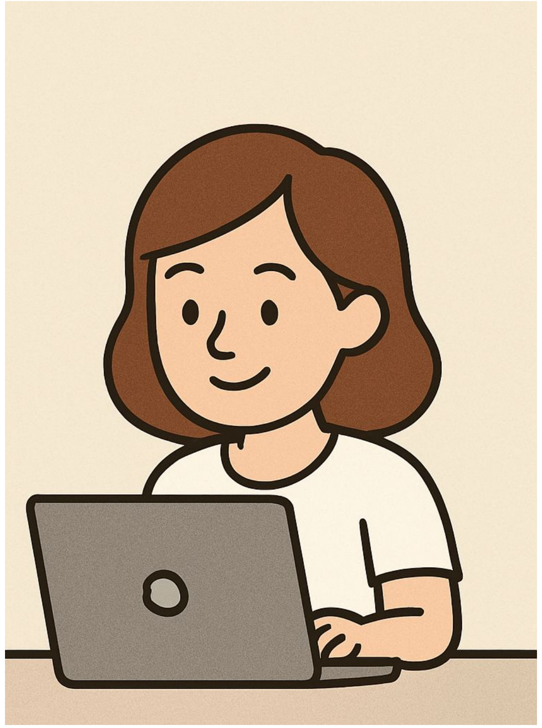
AI-generated data
“data-based”

Get reasonable mock data out of “nothing”



Generative Model
LLM

Mock Data
reasonable
rule-adhering
granular level



https://github.com/mostly-ai/mostlyai-mock

```
pip install -U mostlyai-mock
```

MOSTLY·AI

Data Democratization

Data Access

Data Insights

Synthetic Data

power of LLMs/assistants

… for everyone

Use natural language to get insights

app.mostly.ai

MOSTLY·AI

THE MOSTLY AI PRIZE

CASH PRIZE OF

100k USD

bitly

https://www.mostlyaiprize.com/