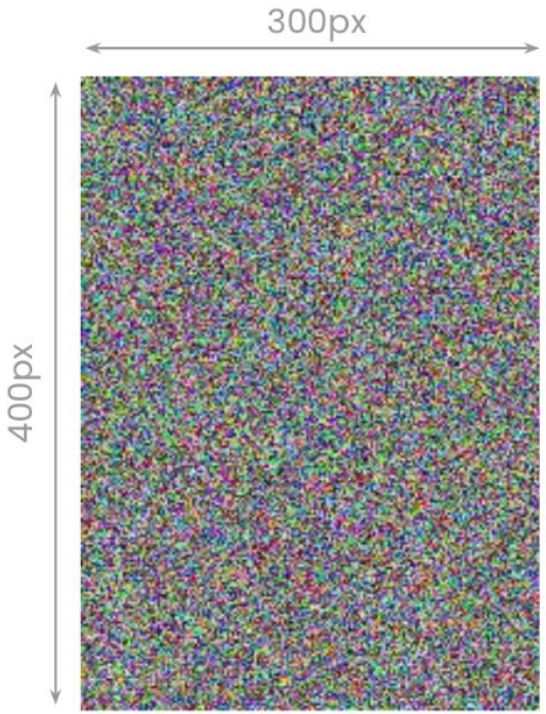
MOSTLY·AI

TabularARGN

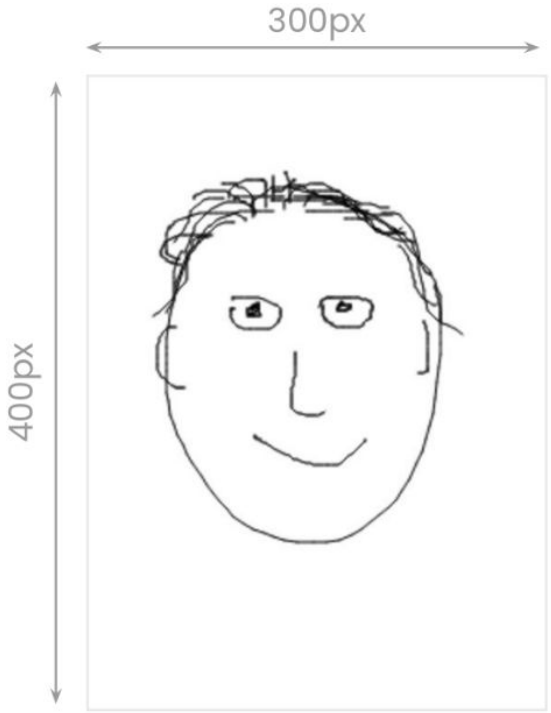
A Framework for High-Quality Synthetic Data Generation

Paul Tiwald
paul.tiwald@mostly.ai

What is Synthetic Data



random data



self-generated data

model-generated data
rule-based

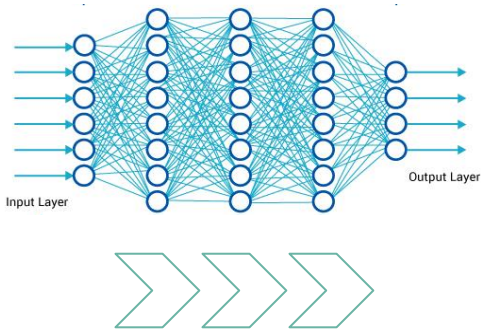
AI-generated data
“data-based”

MOSTLY·AI

What is Tabular Synthetic Data

NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Mary	25y	female	Book	12€	4/2/19	8:12
John	72y	male	Pizza	34€	4/2/19	18:12
...						
Bill	18y	male	Swim	6€	4/4/19	10:02
Bill	18y	male	Shoes	123€	4/4/19	12:32

Real Data



NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Kim	29y	female	Amazon	236€	4/4/19	12:32
Kim	29y	female	Zalando	36€	4/4/19	18:58
...						
Brian	82y	male	Beer	6€	4/2/19	21:32
Sue	24y	female	Sushi	12€	4/2/19	21:32

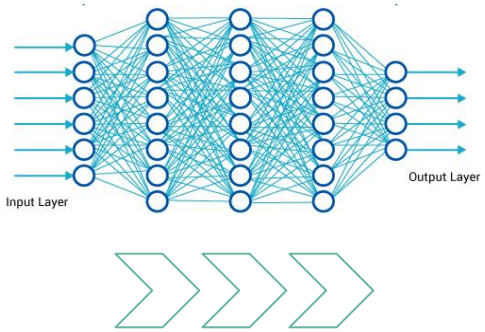
Synthetic Data

MOSTLY·AI

Privacy-preserving Tabular Synthetic Data

NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Mary	25y	female	Book	12€	4/2/19	8:12
John	72y	male	Pizza	34€	4/2/19	18:12
...						
Bill	18y	male	Swim	6€	4/4/19	10:02
Bill	18y	male	Shoes	123€	4/4/19	12:32

Real Data



Private

Open

NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Kim	29y	female	Amazon	236€	4/4/19	12:32
Kim	29y	female	Zalando	36€	4/4/19	18:58
...						
Brian	82y	male	Beer	6€	4/2/19	21:32
Sue	24y	female	Sushi	12€	4/2/19	21:32

Synthetic Data

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)]

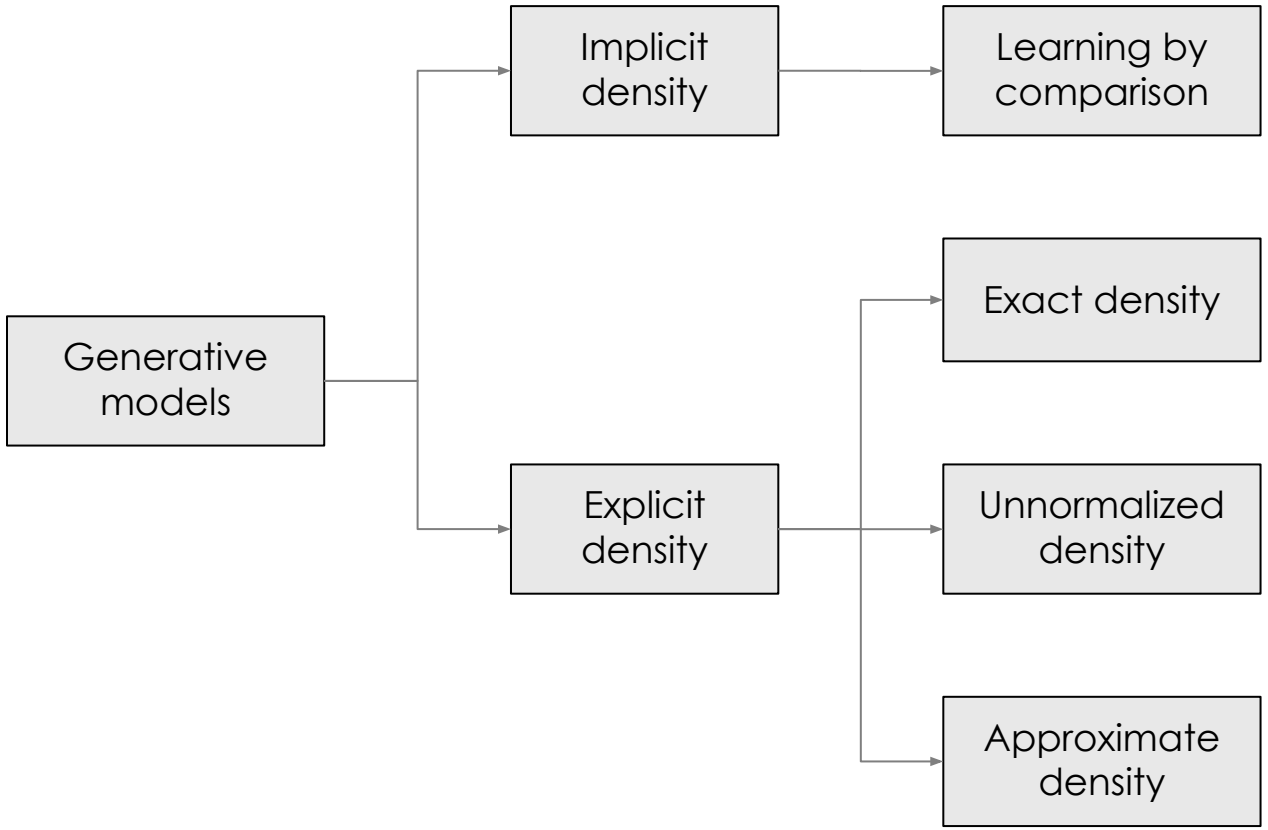
TabularARGN: A Flexible and Efficient Auto-Regressive Framework for Generating High-Fidelity Synthetic Data

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

Taxonomy of deep generative models



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

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

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

- 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:

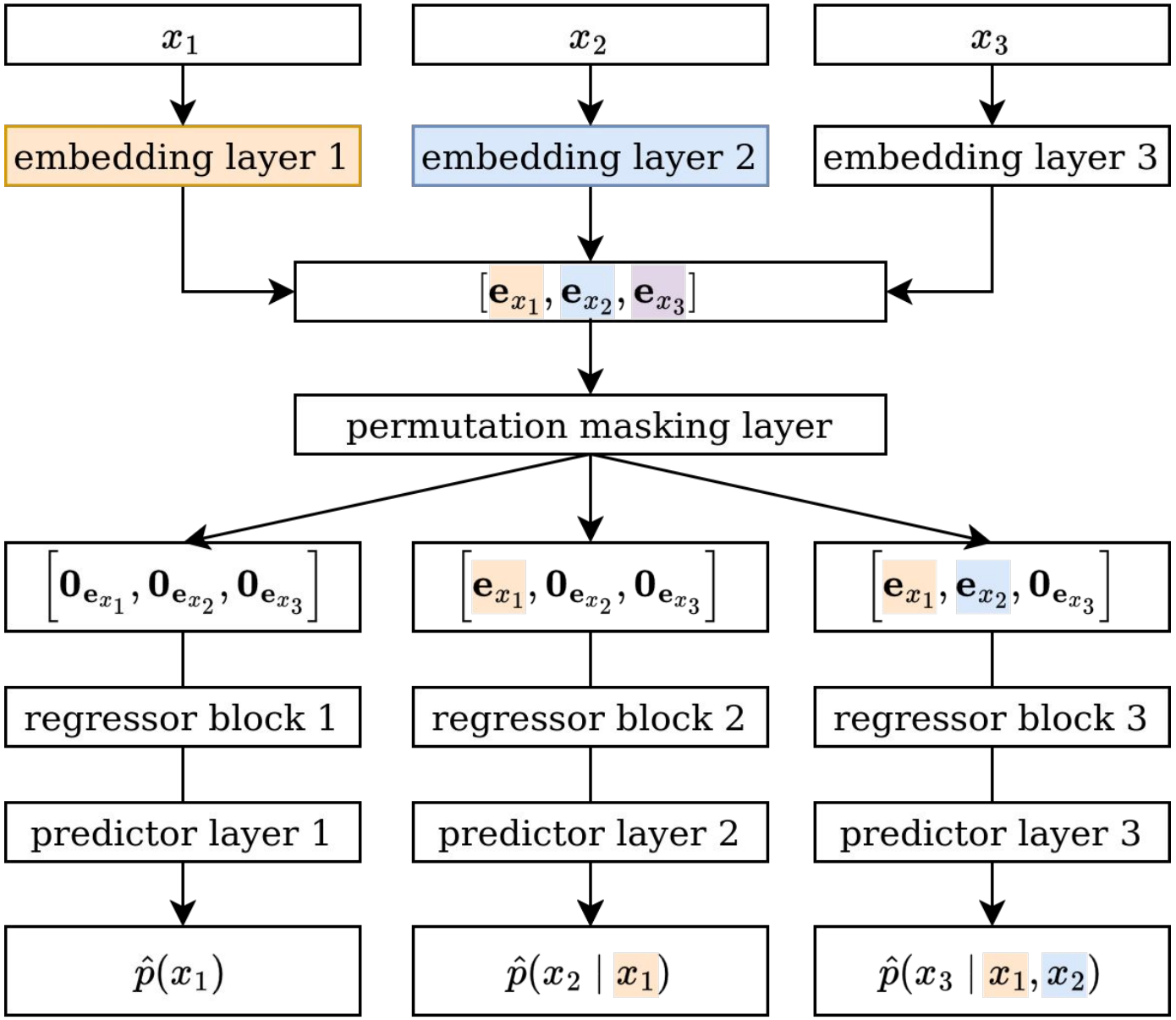
x1 - age

x2 - gender

x3 - blood type

loss function:

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



MOSTLY·AI

Flat Model

Any (column) Order Training Phase

patients data set:

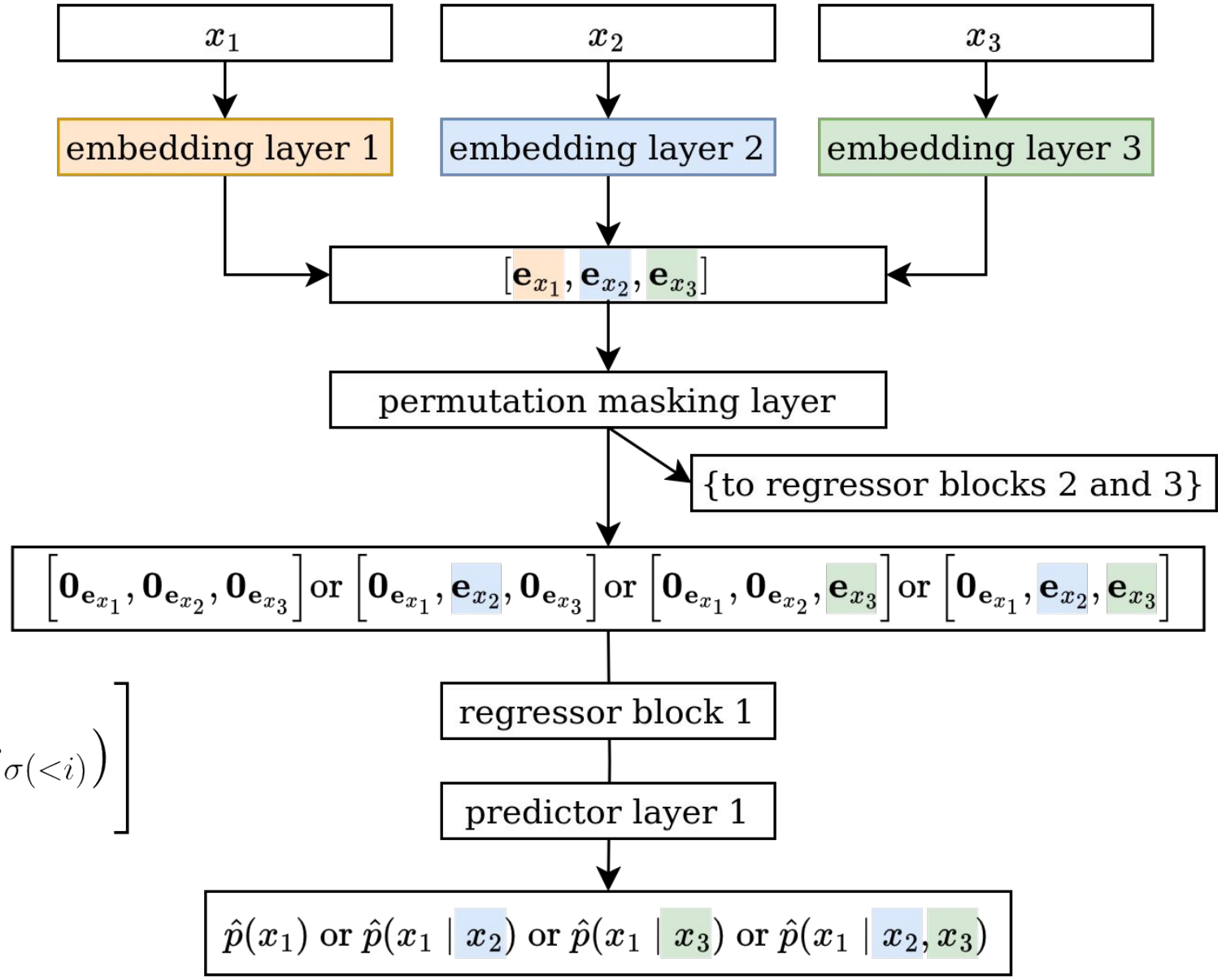
x1 - age

x2 - gender

x3 - blood type

loss function:

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



MOSTLY·AI

Flat Model

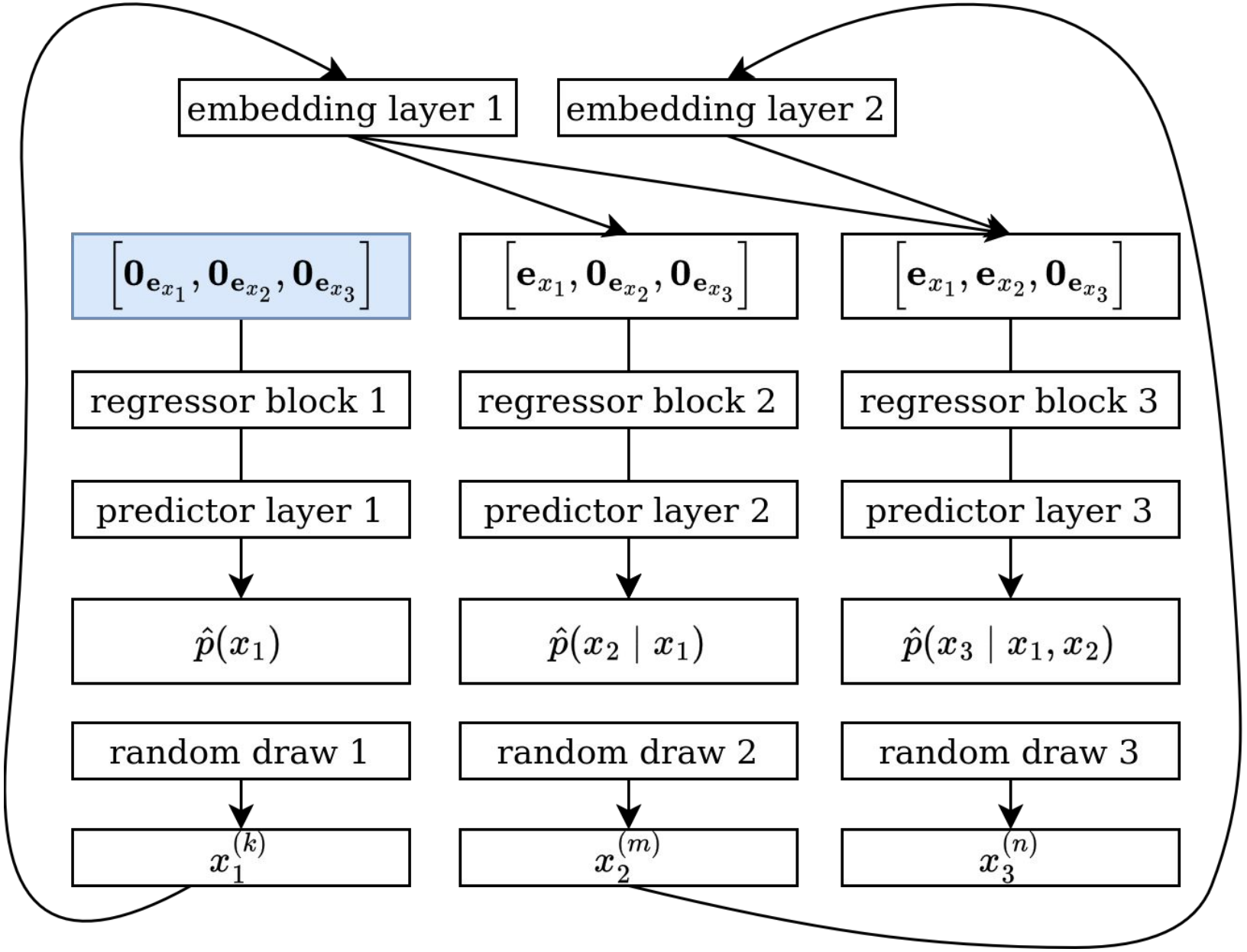
Generation Phase

patients data set:

x1 - age

x2 - gender

x3 - blood type



MOSTLY·AI

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

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

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+

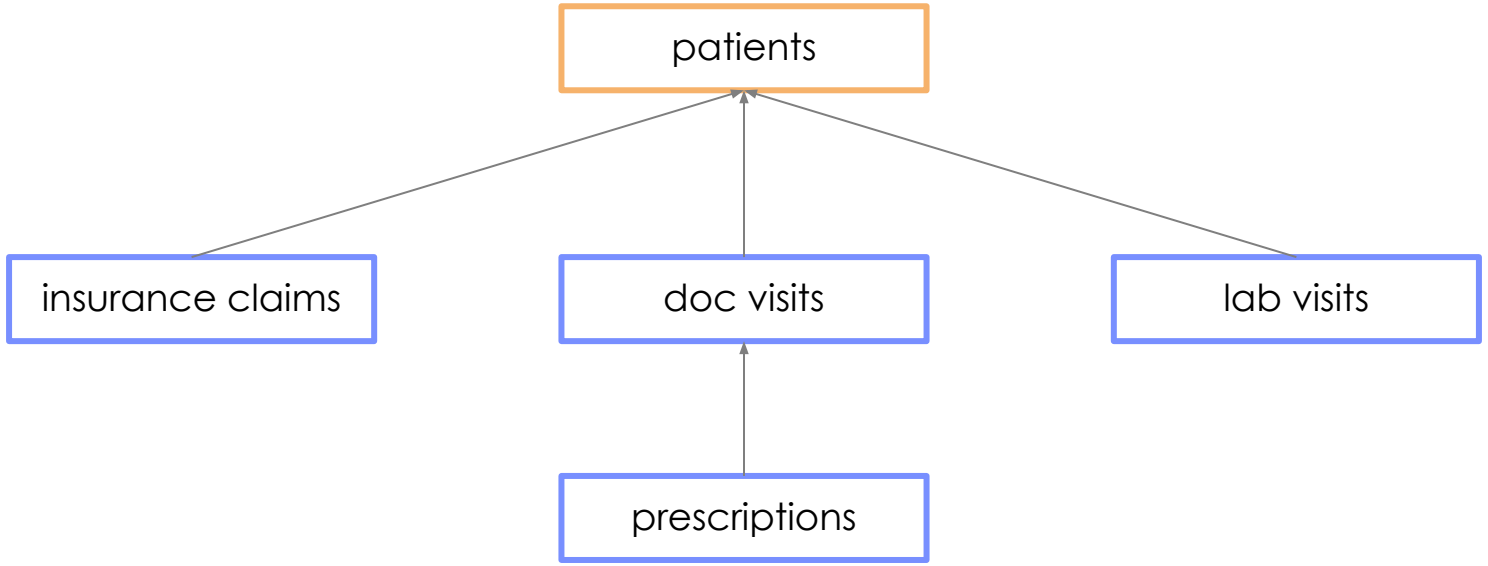
history encoder

context proc.

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

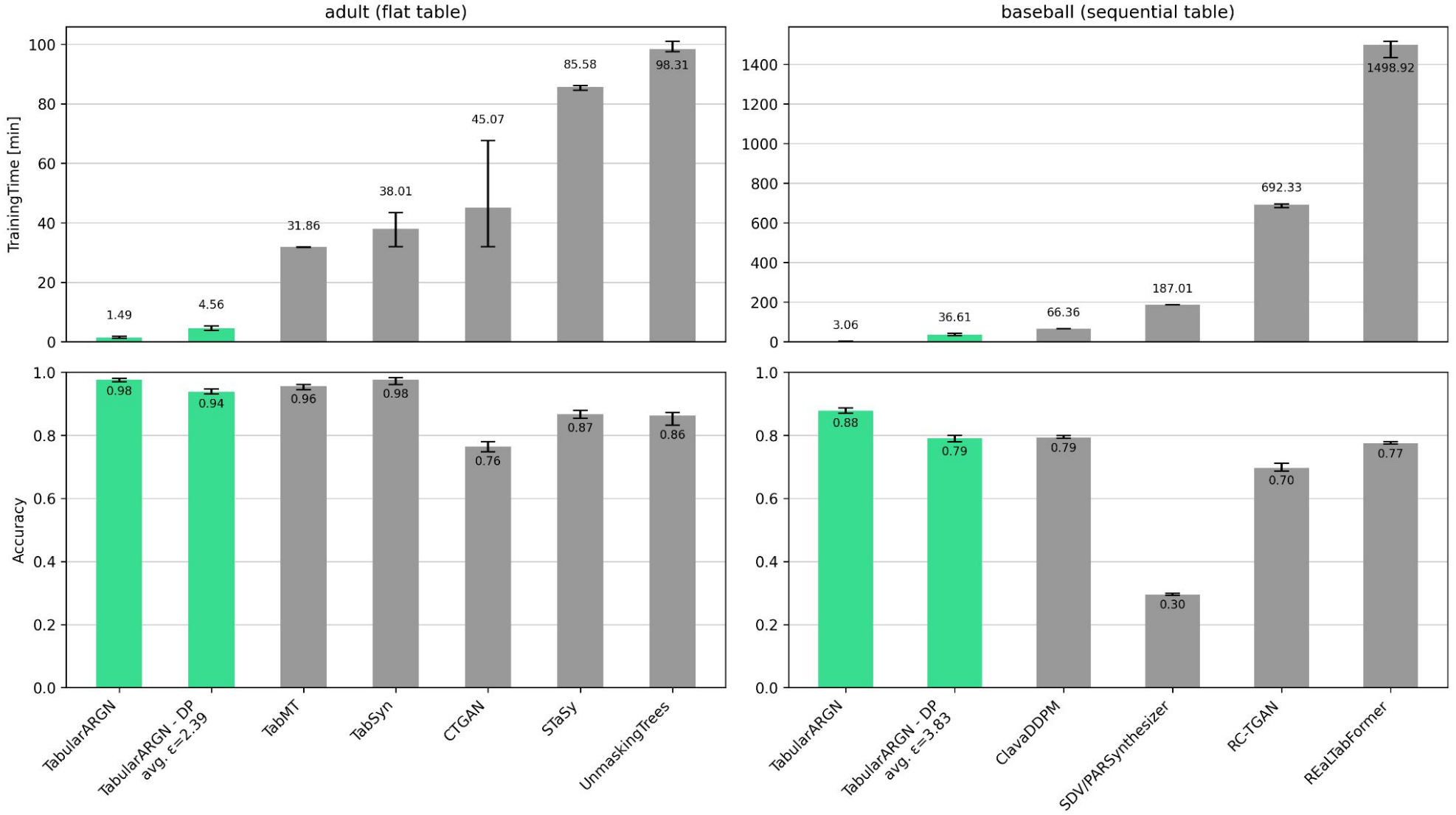
MOSTLY·AI

Flexible context allows for synthesis of multi-table setups



MOSTLY·AI

Benchmarking TabularARGN against SOTA methods



MOSTLY AI

Text-extension allows for leveraging the power of LLMs

combining TabularARGN with text models

date	systolic blood pressure	body weight	diagnosis
15.1.2022	140	85	“The patient presents with …”

TabularARGN

LanguageARGN

training data:

{ "input": "15.1.2022; 140; 85", "output": "The patient presents with …" }

We are hiring

Developer Advocate (Remote)

Permanent employee, Full-time · EMEA Remote

We are looking for a captivating Developer Advocate to help us spread the word, engage with the Python / Data / AI community, and create some fun content. Somewhere at the **intersection of a Data Enthusiast, Open Source Advocate, and a Content Creator**.

Tabular ARGN - Auto-Regressiv Generative Networks

Tabular ARGN is implemented in the Synthetic Data SDK ⭐

for more infos, check out our blog

- Synthetic Behavioral Data
- Synthetic Geo Data
- Differentially Private Synthetic Data
- Fair Synthetic Data
- Synthetic Data Benchmarks
- JRC Report on Synthetic Data
- AI-based Re-Identification Attacks
- Privacy Assessment of Synthetic Data

and many more

MOSTLY·AI