2024.03.27

Università di Bologna

Generation of Synthetic Mutation Data of Acute Myeloid Leukemia Patients with Large Language Models

Relatore
Prof. Enrico Giampieri
Correlatore
Prof. Gastone Castellani

Candidato
Alessio Giuseppe Ferraioli

SYNTHEMA

www.synthema.eu

Our Mission

Establish a cross-border hub to develop and validate Artificial Intelligence techniques for anonymisation and synthetic data generation in rare hematological diseases.

SYNTHEMA aims to generate reliable, high-quality synthetic data that can shape new **virtual patients** to further enhance dia-gnostic capacity, assess treatment options and predict outcomes in rare hematological diseases.

Overview of the presentation

Introduction on Synthetic Data Generation (SDG) in Medical Research

GReaT: application of language models to SDG

The dataset we modeled: AML patients mutations

Why GReaT does not work for this dataset and our solution to overcome this

Results: synthetic patients visualizations

Techniques for analyzing the results:

- UMAP dimensionality reduction
- Kaplan Meier estimator of Survival Function

Comparisons with other techniques for SDG: C-GAN and VAE

Data-driven Medical Research

- Helping diagnosis e.g.: identifying tumors
- Individual tailoring of treatment pathways
- Recognizing patterns in quantities of data too big for a human to analyze

The Applied Physicist contributes to the whole medical data pipeline

- Development of data acquisition, in particular imaging
- Processing and Analysis of data
- Modelling of biological processes

Availability and quality of data are crucial

- Privacy concerns
- Long process and strict requirements to access data
- Data production is costly and unpractical
- Historical data comes usually with biases

Synthetic data could mitigate these problems

see for example:

Choi et al., *Generating Multi-label Discrete Patient Records using Generative Adversarial Networks*, 2017;
Park et al., *Data synthesis based on generative adversarial networks*, 2018;
Xu et al., *Modeling tabular data using conditional GAN* ,2019;
Borisov et al., *Deep neural networks and tabular data: A survey,* 2021

https://github.com/kathrinse/be_great

GReaT: Generator of Realistic Tabular data

Convert data into **meaningful sentences** in the English language and use **pre-trained language models** to generate new sentences coherent with data.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. *Language models are realistic tabular data generators*, 2022.

Language models read the data as a human would

No loss of contextual knowledge

If “Age” is 7, probably “Marital status” will not be “Married”, “Divorced” nor “Widowed”

This semantic knowledge of the features is lost when encoding them numerically.

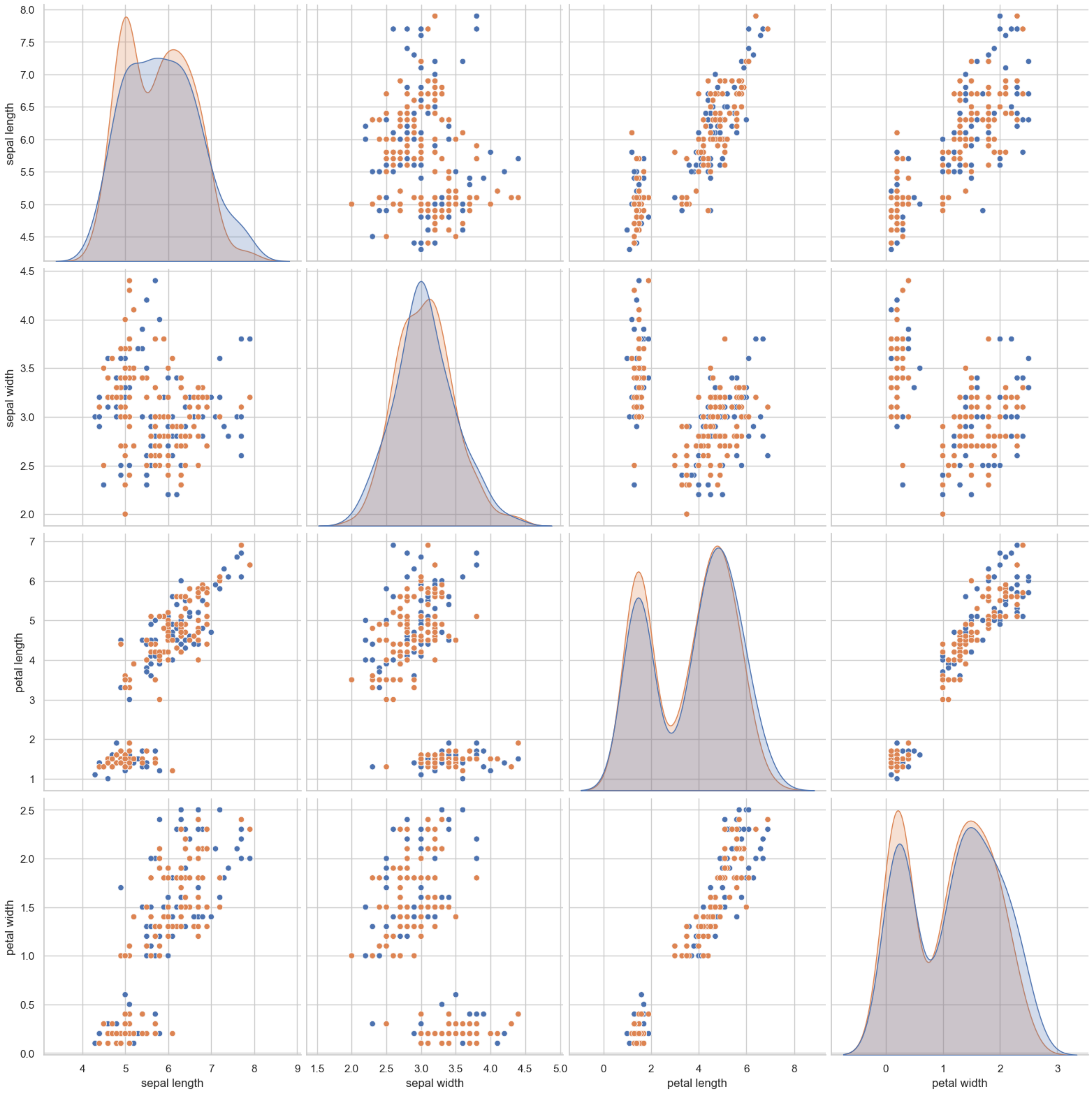
Possibility to encode textual diagnosis data

“The patient was pale, jaundiced and dehydrated. She had weak pulse and fever..”

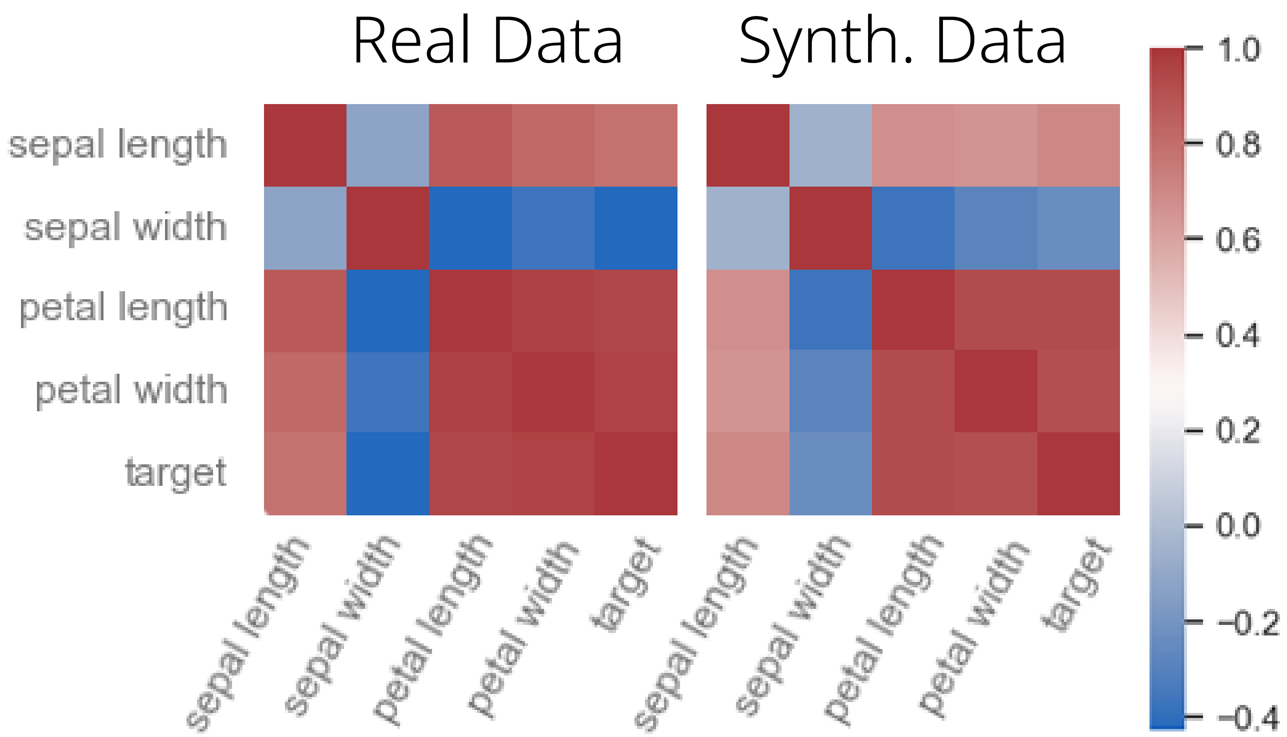
Textual diagnosis are ubiquotous in medical data, but usually ignored in encoding of data

Application on Iris Dataset:

Generating realistic new observations



Correlations

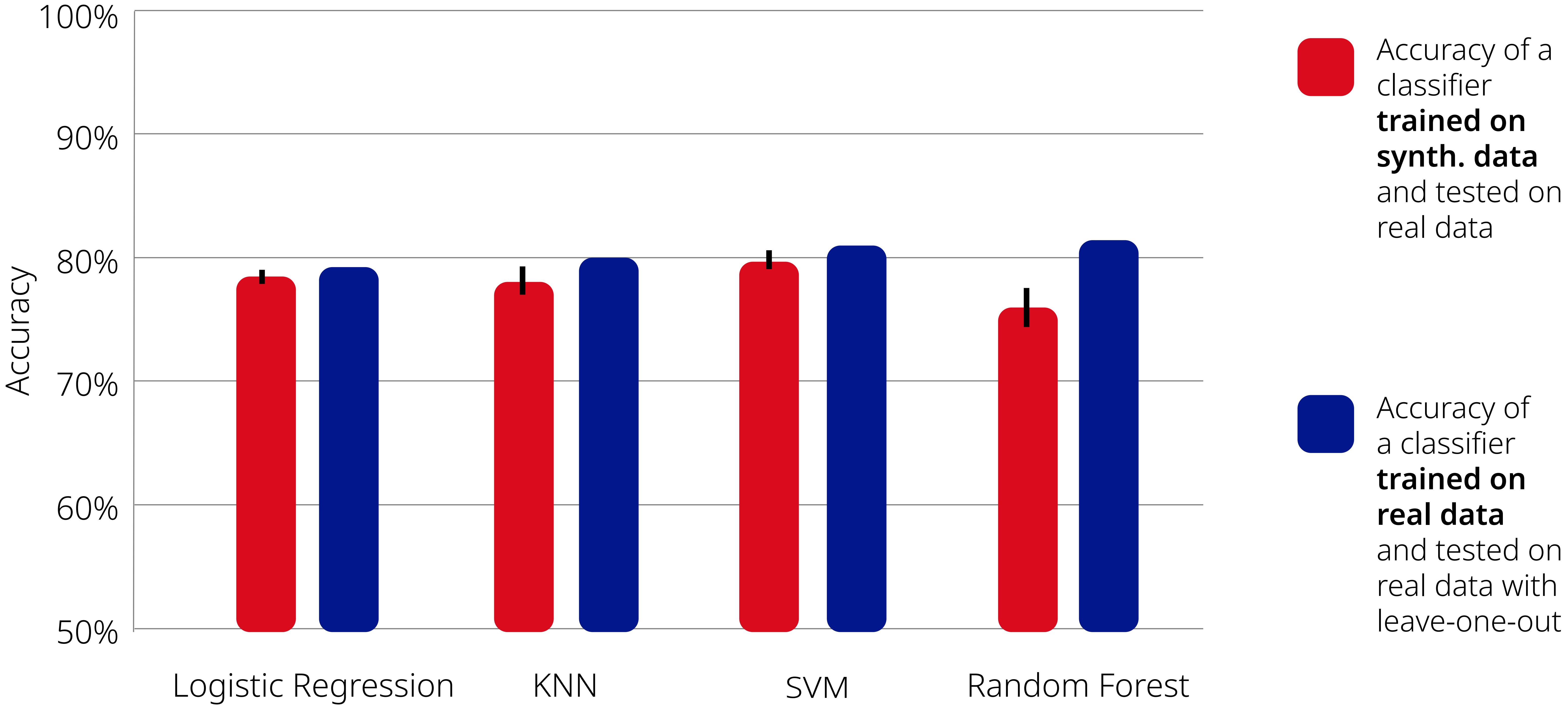


● *Real Data*

● *Synth. Data*

Application on Titanic Dataset:

Training a Classifier without access to real data



🟥 Accuracy of a classifier **trained on synth. data** and tested on real data

🟦 Accuracy of a classifier **trained on real data** and tested on real data with leave-one-out

Our Goal: Modeling a dataset of genetic and cytogenetic mutations of Acute Myeloid Leukemia patients

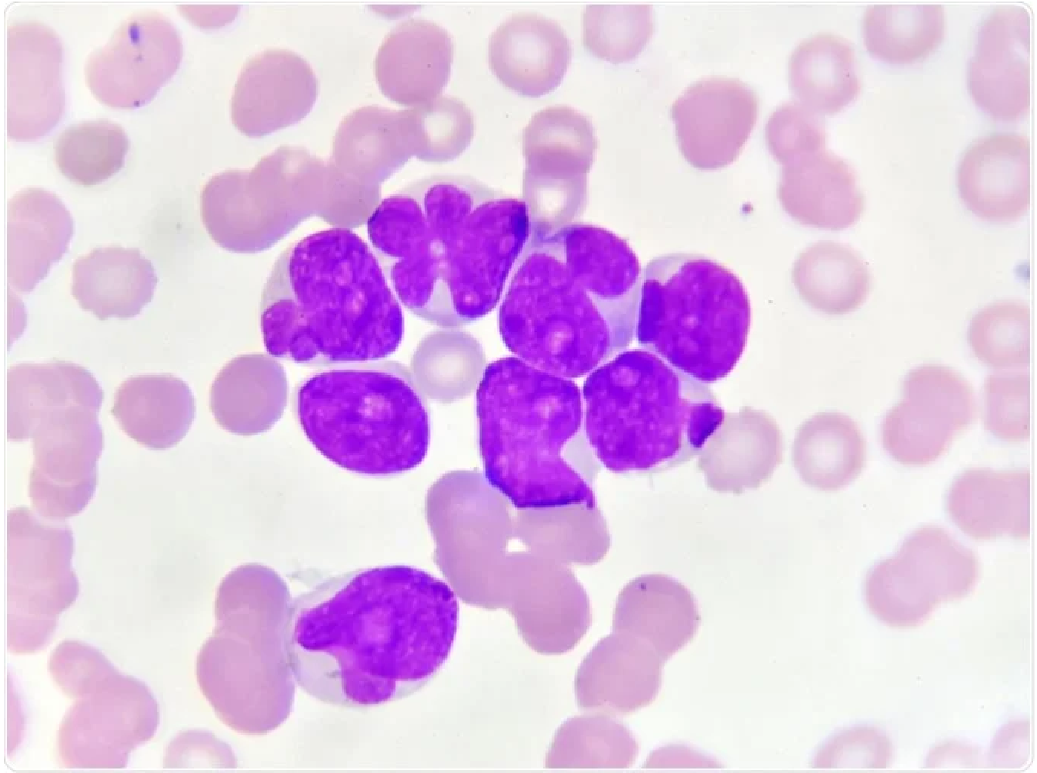


Image by Jarun Ontakrai on Shutterstock.com

Acute Myeloid Leukemia (AML) is an aggressive and often fatal form of leukemia in the bone marrow that obstructs the production of blood cells

Different mutations

↓

Different evolution of disease

↓

Different suggested treatment

The dataset contains 154 mutations as well as the state (dead or alive) and time of the last observation.

We want to model the **time of death of patients depending on their mutations**.

GReaT can not model high-dimensional datasets, such as a **mutation dataset** like this one

The generated sentences are too long to handle and very redundant

senteces like:

“PHF6 is 0, RIT1 is 0, CBL is 0, add_8 is 1, CUL2 is 0, del_18 is 0, t_15_17 is 0, SMG1 is 0, NPM1 is 0, MYC is 0, JAK2 is 0, KANSL1 is 0, KIT is 0, MED12 is 0, MLL is 0, IDH2_.R172 is 0, NOTCH1 is 0. GATA1 is 0, IDH2_p.R140 is 0, FLT3_other is 0, KMT2C is 0, BRAF is 0, CEBPA_mono IS 0, DNMT3A is 0, CUX1 is 0, EED is 0, FBXW7 is 0, t_9_11 is 0, NRAS_p.Q61_62 is 1, ZRSR2 is 0, TP53 is 0, del_16 is 0, WT1 is 0, add_8 is 1, STAG2 is 0, RIT1 is 0, RUNX1 is 0, PTPN11 is 0, RAD21 is 0

To ovecome its limitations, we propose to train on just sentences that signal **present mutations**

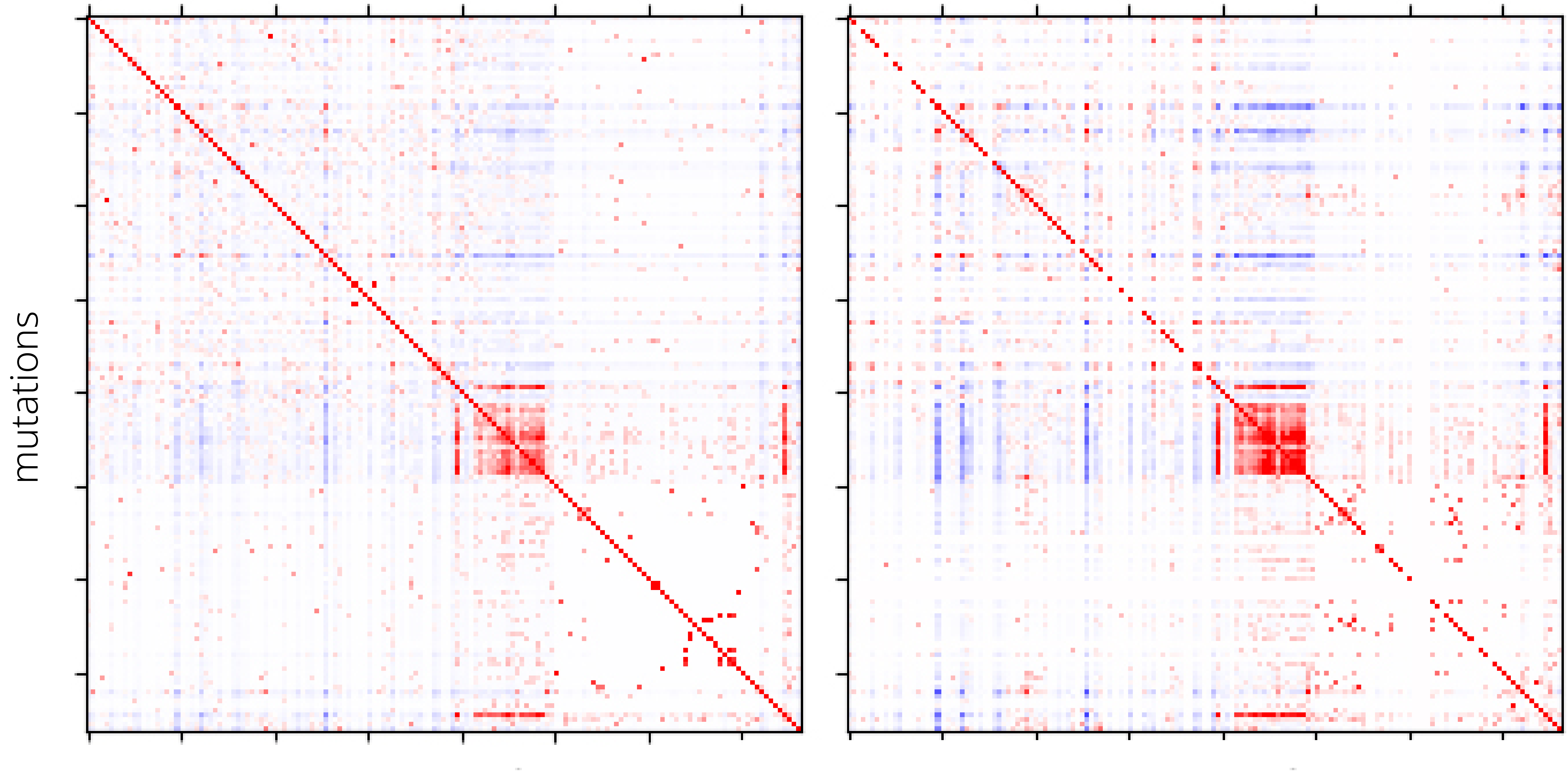
"PHF6 is 0, RIT1 is 0, CBL is 0, add_8 is 1, CUL2 is 0, del_18 is 0, t_15_17 is 0, SMG1 is 0, NPM1 is 0, MYC is 0, JAK2 is 0, KANSL1 is 0, KIT is 0, MED12 is 0, MLL is 0, IDH2_.R172 is 0, NOTCH1 is 0. GATA1 is 0, IDH2_p.R140 is 0, FLT3_other is 0, KMT2C is 0, BRAF is 0, CEBPA_mono IS 0, DNMT3A is 0, CUX1 is 0, EED is 0, FBXW7 is 0, t_9_11 is 0, NRAS_p.Q61_62, ZRSR2 is 0, TP53 is 0, del_16 is 0, WT1 is 0, STAG2 is 0, RIT1 is 0, RUNX1 is 0, PTPN11 is 0, RAD21 is 0, ...

"add_8 is 1, NRAS_p.Q61_62 is 1"

All the others are implictly zero

The modified GReaT model generates realistic patients

Correlations



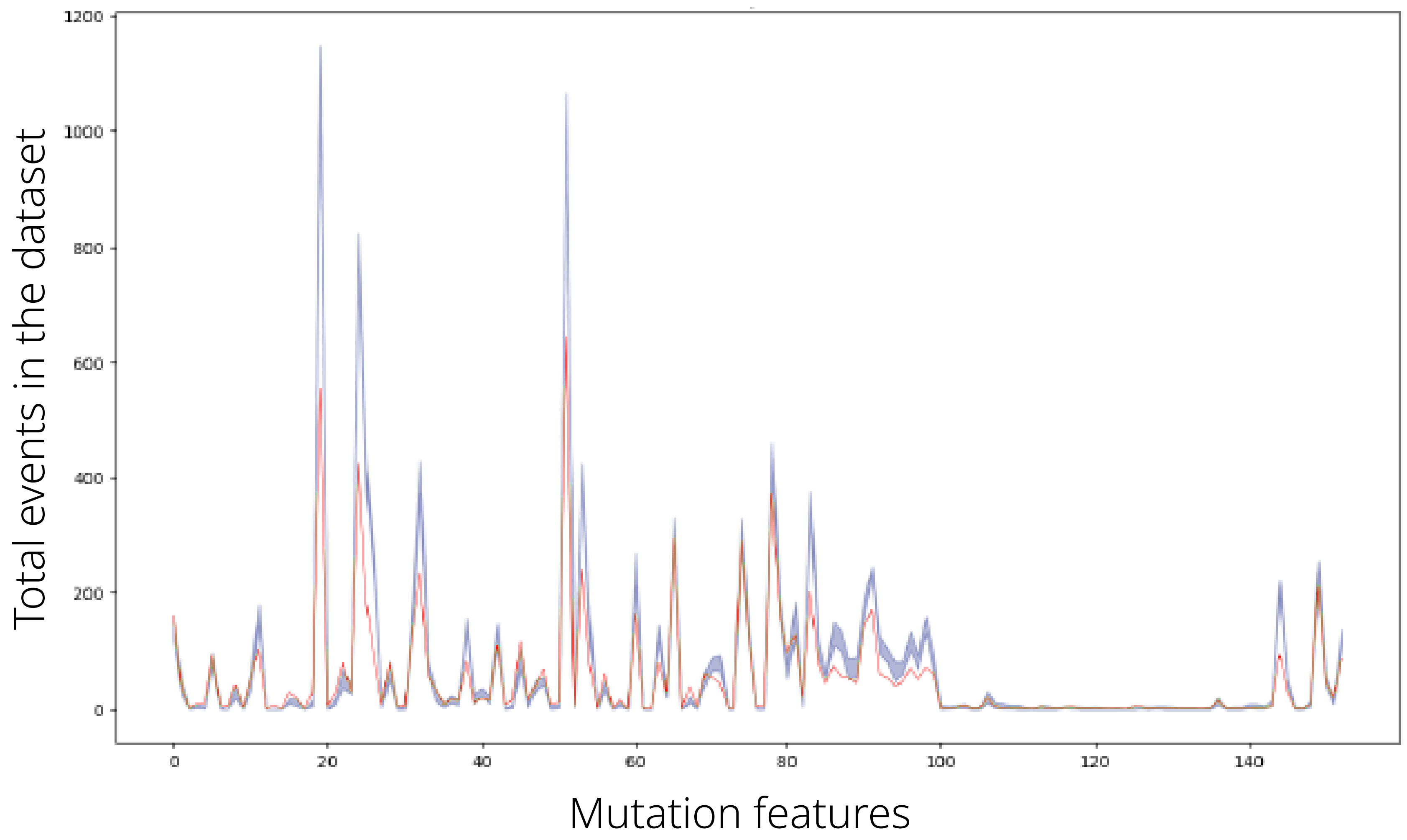
Real Data

Synth. Data

The modified GReaT model generates realistic patients

It slightly over-represents most common values.

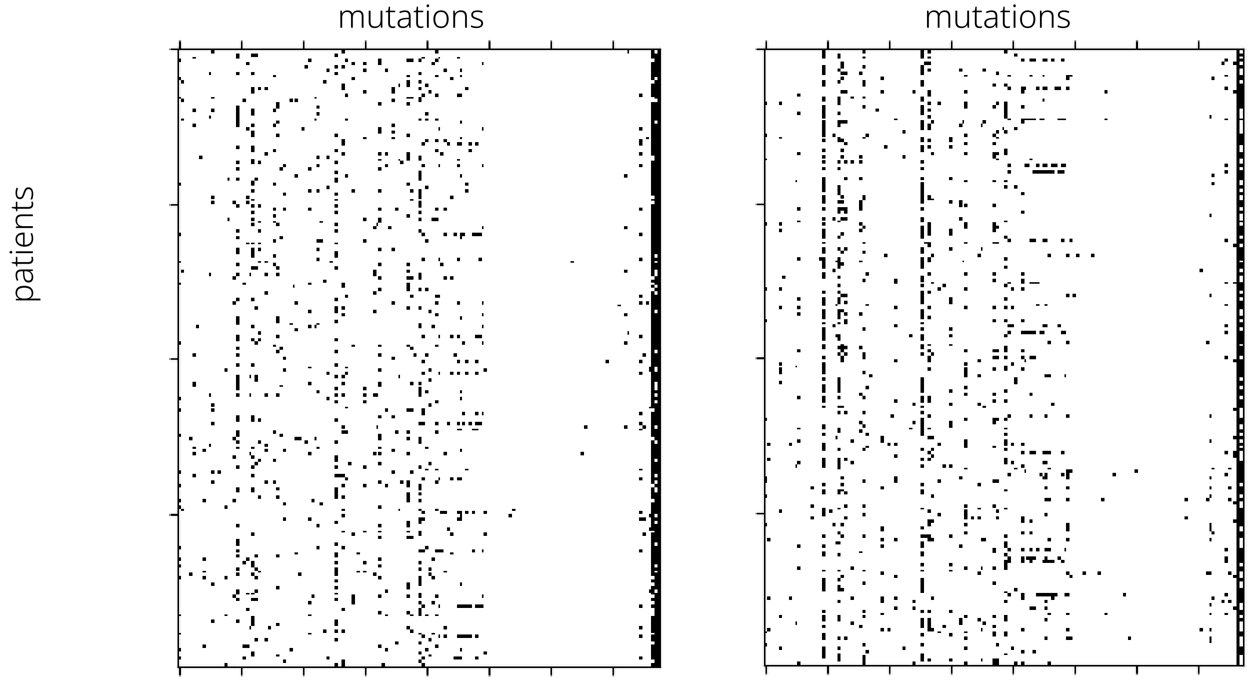
N. of events per mutation feature



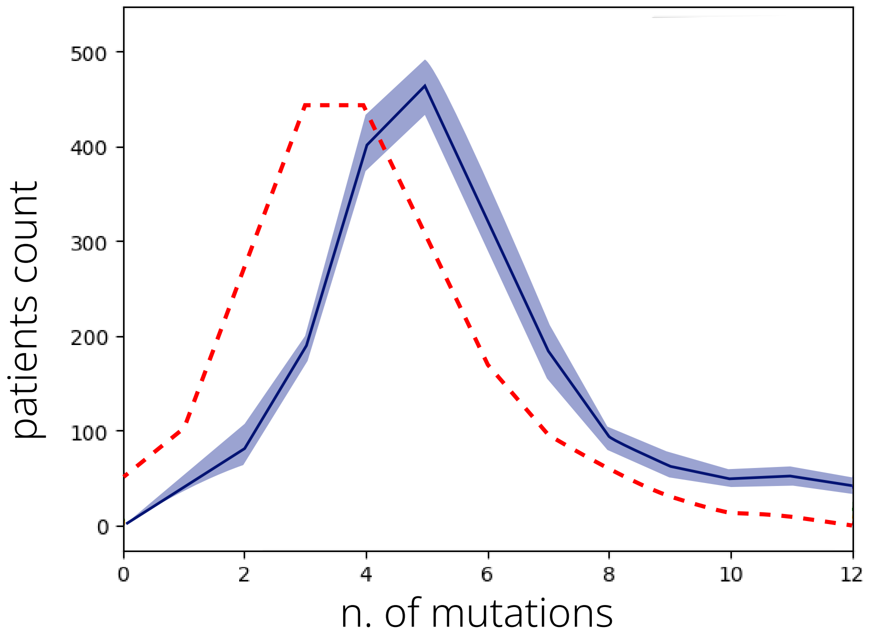
Sparse matrix diagrams

Real Data

Synth. Data



N. of mutations per patient



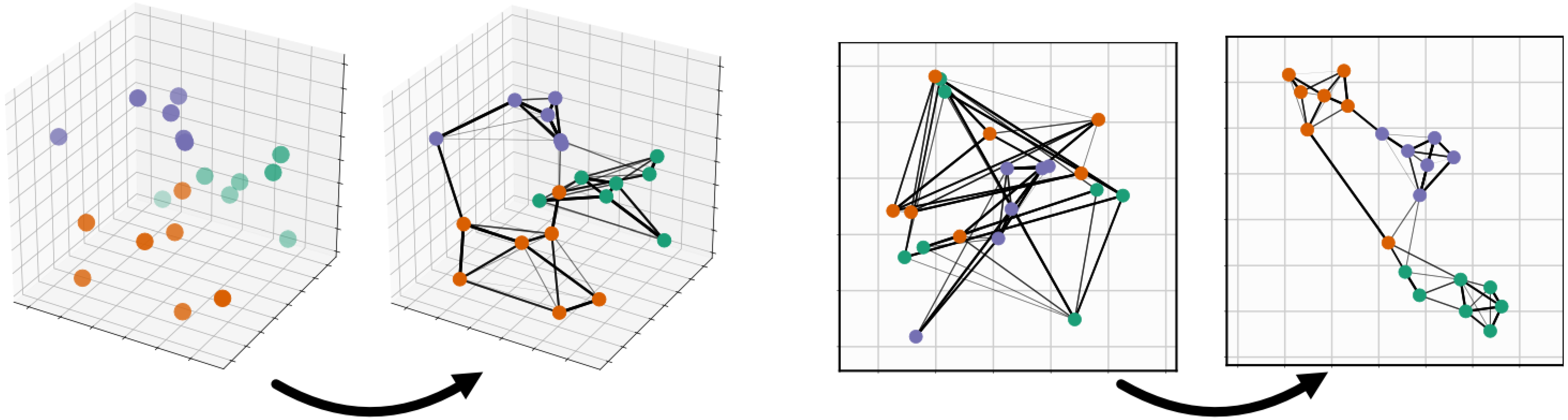
🟥 **Real**

🟦 **Synthetic**

Dimensionality Reduction with UMAP and Kaplan-Meier Survival Curves estimates

UMAP

Dimensionality reduction technique that represents **data as vertices of a weighted graph** that is projected onto a 2D space



Compute a graphical representation of the dataset

Learn an embedding that preserves the structure of the graph

Dimensionality Reduction with UMAP and Kaplan-Meier Survival Curves estimates

Kaplan Meier Estimate

Estimator for the survival curve *S(t)*, the fraction of **patients alive at time *t***, based on maximization of likelihood on discrete observations.

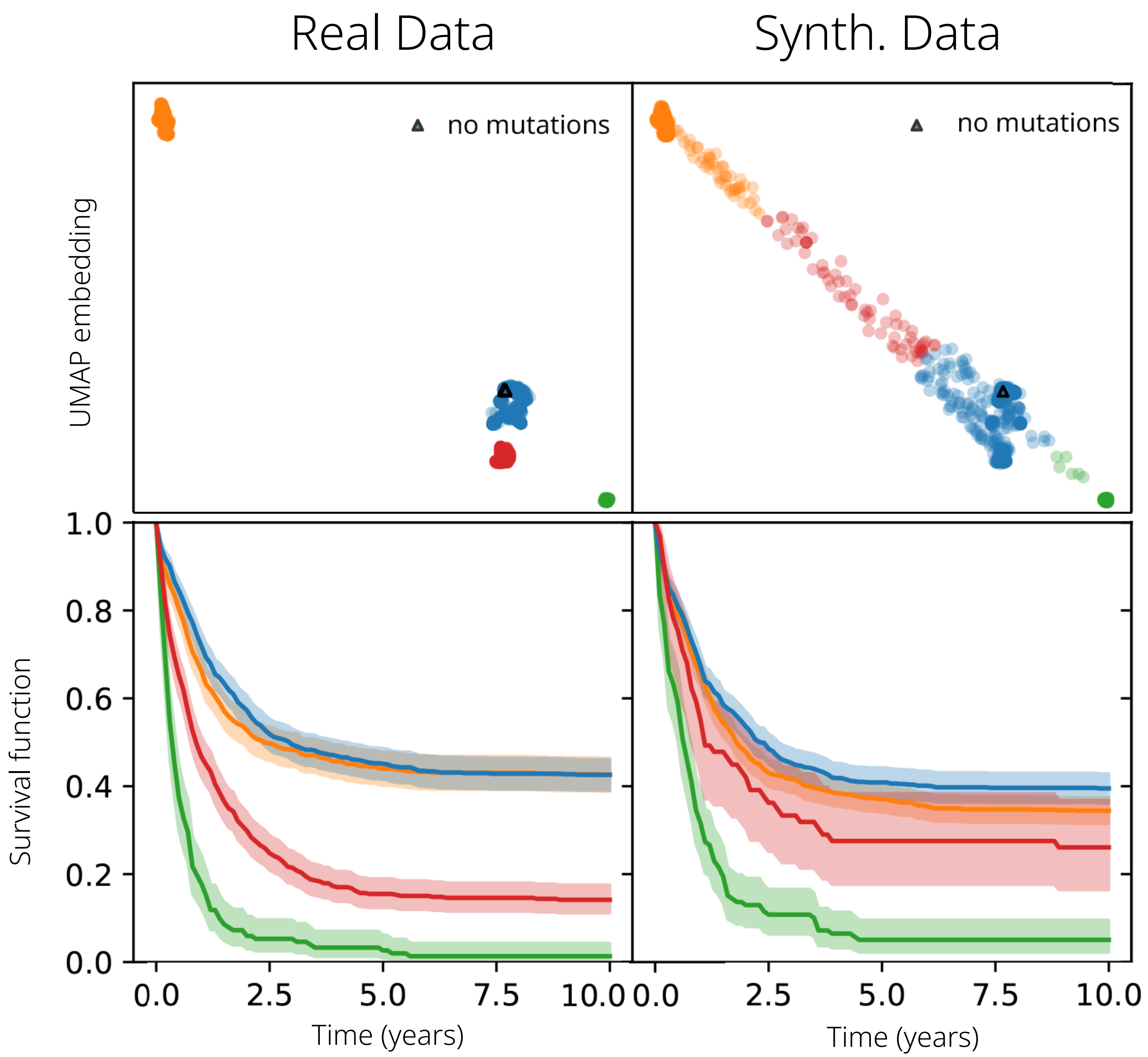
number of deaths happening at time t_j

$$\widehat{S}(t) = \prod_{j:t_j<t} (1 - \frac{d_j}{n_j})$$

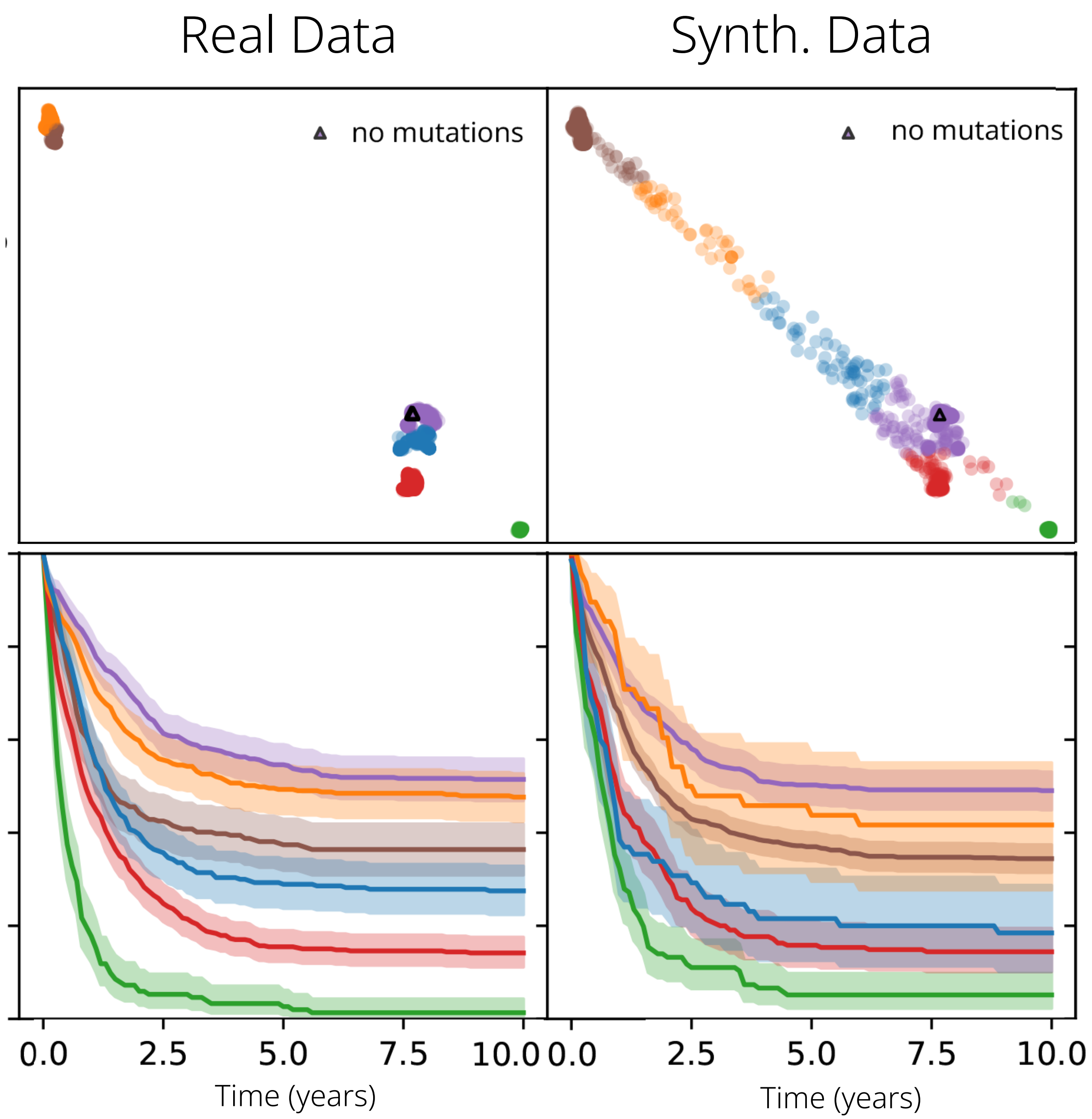
number of patients at risk immediately before t_j

discrete time observations t_j

Dimensionality Reduction with UMAP and Kaplan-Meier Survival Curves estimates

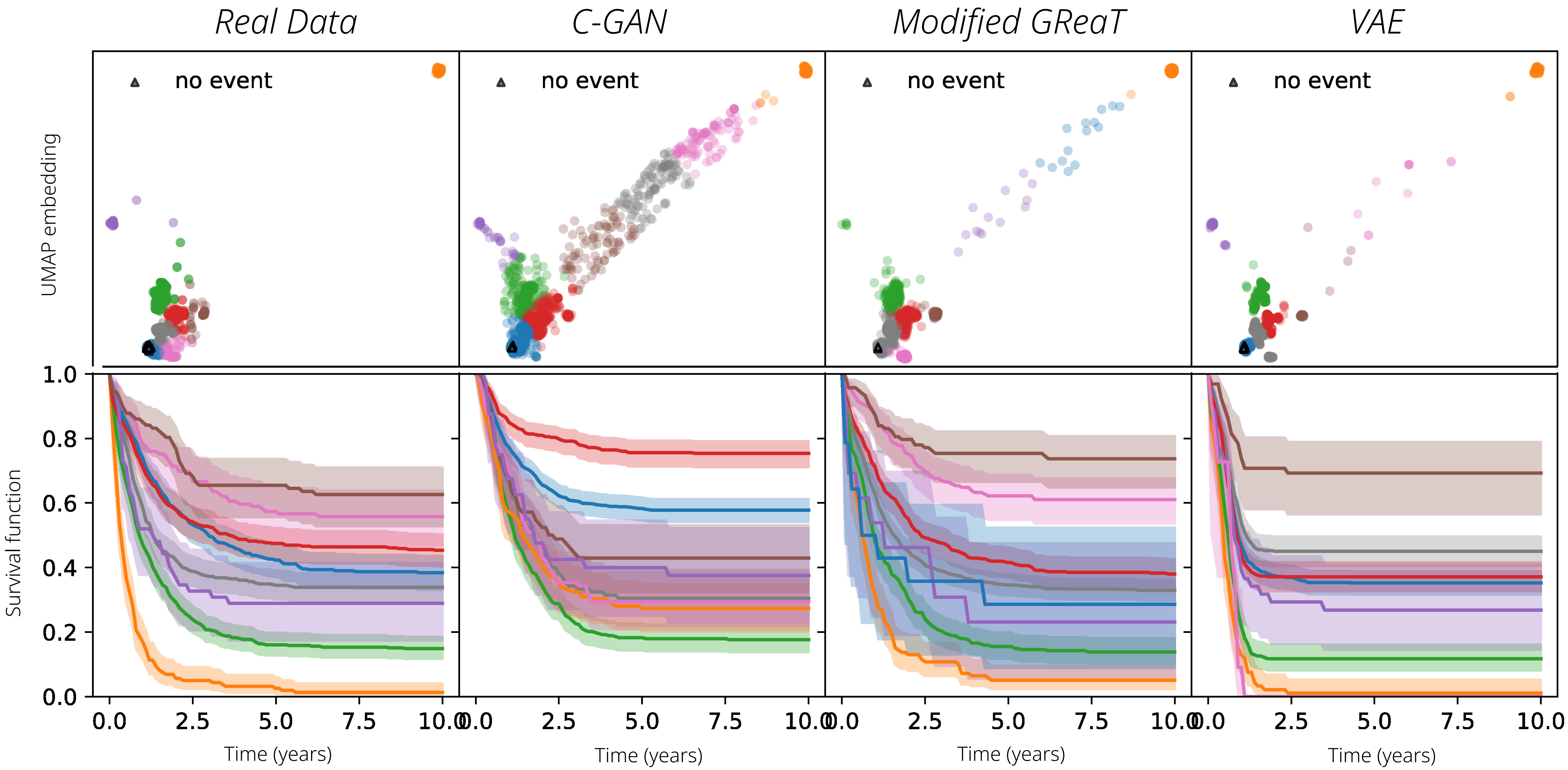


n clusters = 4



n clusters = 6

Comparison with other techniques



Thanks to Adriano Zaghi (adriano.zaghi.2@unibo.it) and Ettore Rocchi (ettore.rocchi3@unibo.it for providing C-GAN and VAE data, and to Mauro Faccin (mauro.faccin@unibo.it) for the contribution in producing this visualization.

In conclusion: the modified GReaT produces promising results comparable to state-of-the-art methods.

Newer and more powerful language models could perform even better.

Thank you for your attention.

The GReaT workflow

1. **Convert features into elemental sentences**

Occupation	Age	Gender
Sales	59	Male
Developer	33	Female

↓

“Occupation is Sales”
“Age is 59”
“Gender is Male”

“Occupation is Developer”
“Age is 33”
“Gender is Female”

The GReaT workflow

1. **Convert features into elemental sentences**

2. **String them together in random order**

Occupation	Age	Gender
Sales	59	Male
Developer	33	Female

↓

"Occupation is Sales"
"Age is 59"
"Gender is Male"

"Occupation is Developer"
"Age is 33"
"Gender is Female"

↓

"Occupation is Sales, Gender is Male, Age is 59"

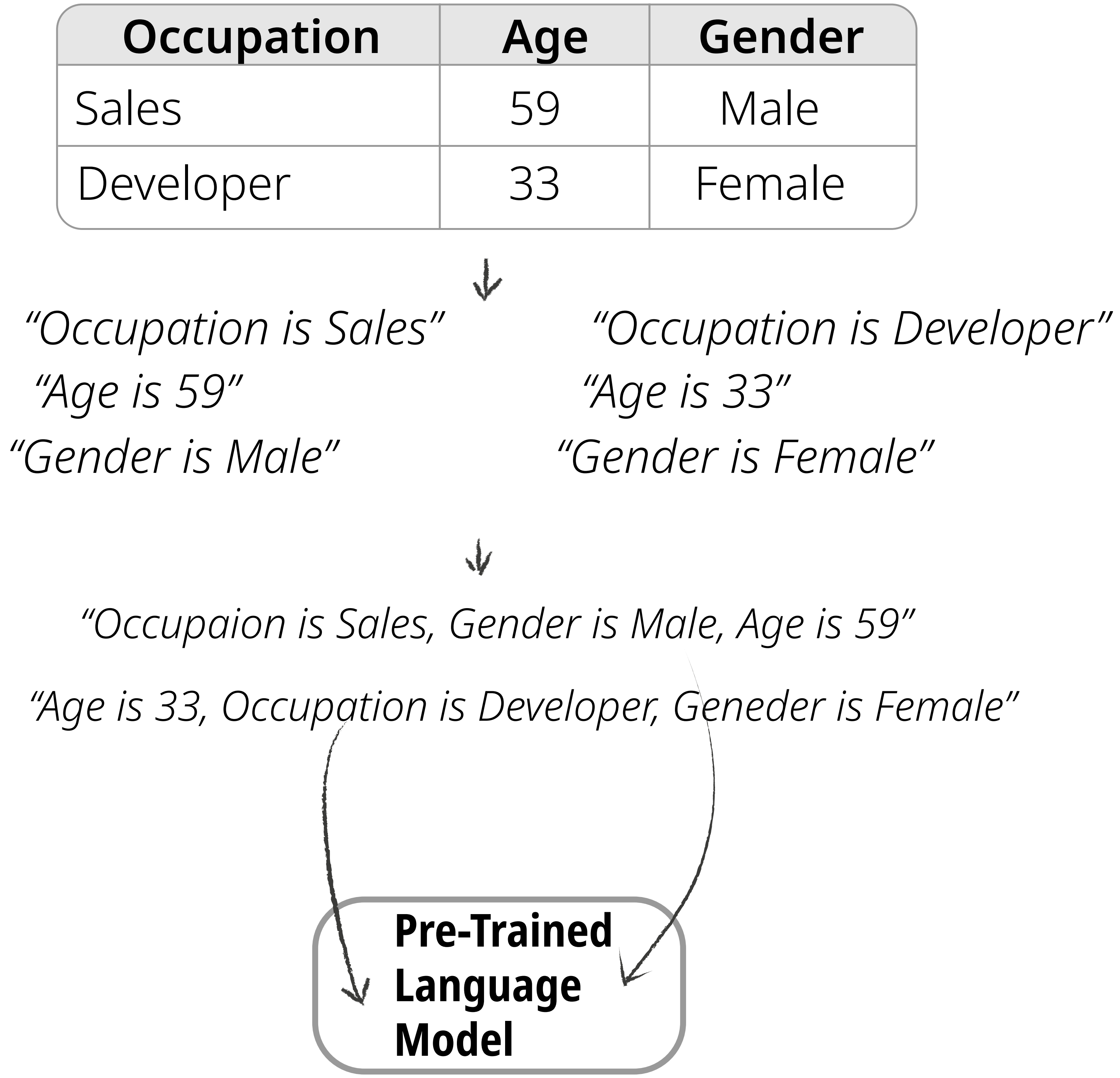
"Age is 33, Occupation is Developer, Geneder is Female"

The GReaT workflow

1. **Convert features into elemental sentences**

2. **String them together in random order**

3. **Fine-tune a pre-trained language model on the sentences**



Occupation	Age	Gender
Sales	59	Male
Developer	33	Female

"Occupation is Sales"
"Age is 59"
"Gender is Male"

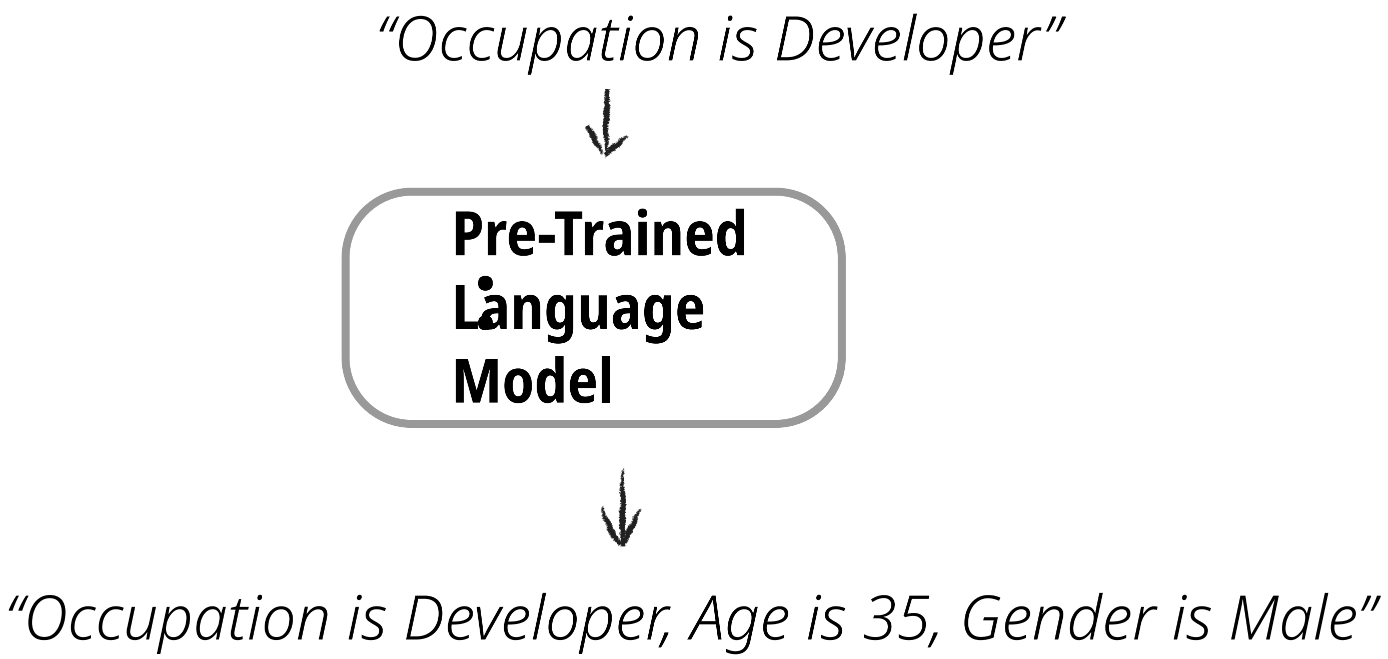
"Occupation is Developer"
"Age is 33"
"Gender is Female"

"Occupaion is Sales, Gender is Male, Age is 59"

"Age is 33, Occupation is Developer, Geneder is Female"

The GReaT workflow

4. **Generate new sentences by drawing a random value for a feature and letting the model complete it**



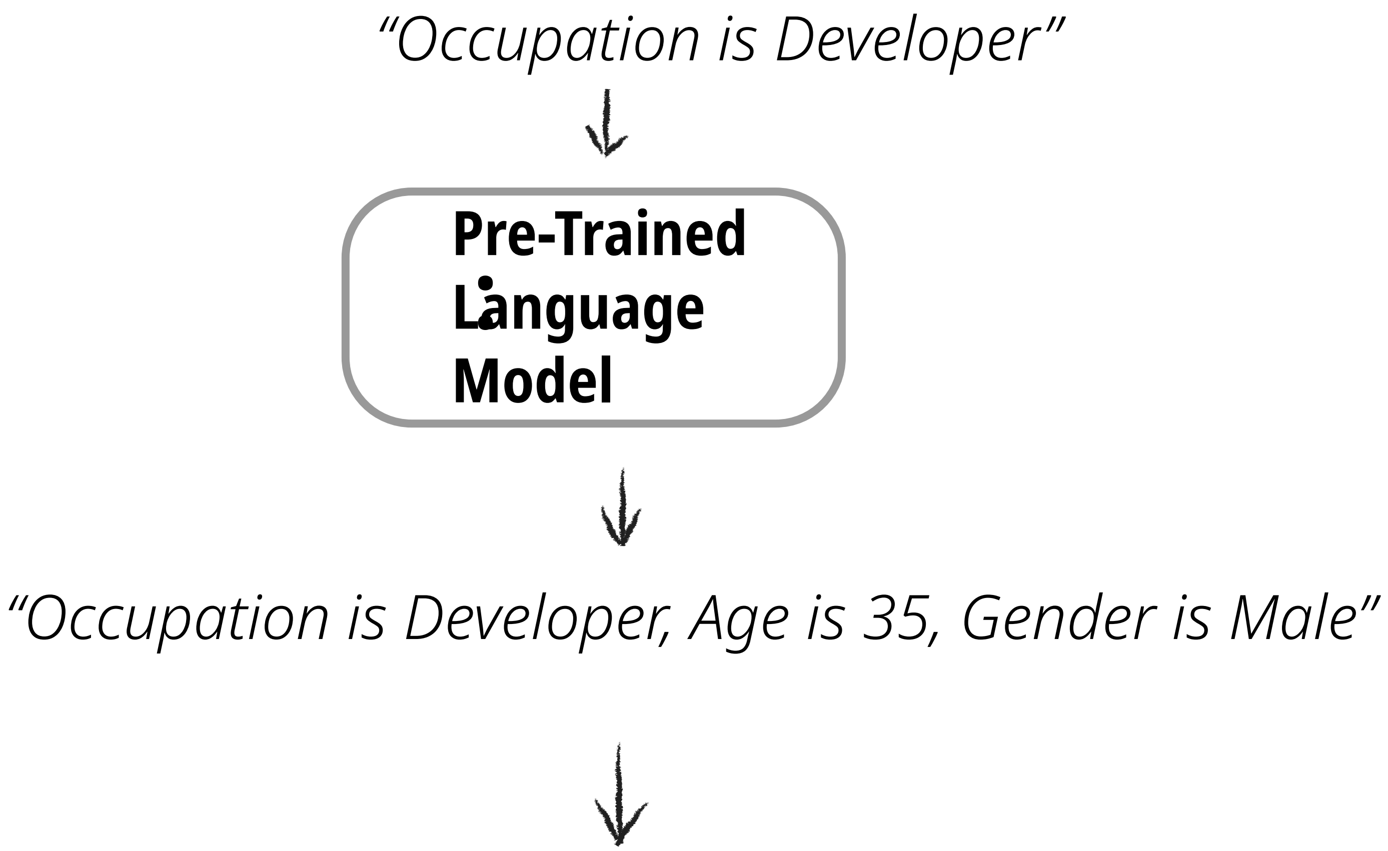
“Occupation is Developer”

Pre-Trained Language Model

“Occupation is Developer, Age is 35, Gender is Male”

The GReaT workflow

4. **Generate new sentences by drawing a random value for a feature and letting the model complete it**



“Occupation is Developer”

Pre-Trained Language Model

“Occupation is Developer, Age is 35, Gender is Male”

5. **Convert the sentence back into tabular data**

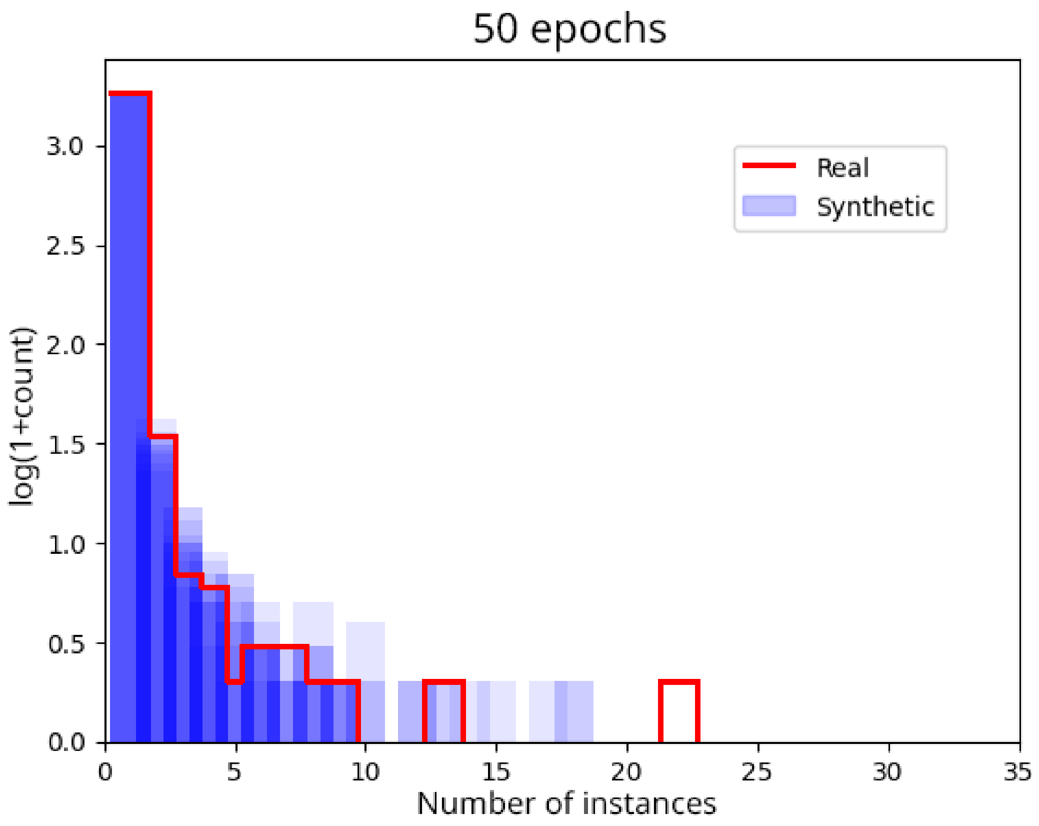
Occupation	Age	Gender
Developer	35	Male

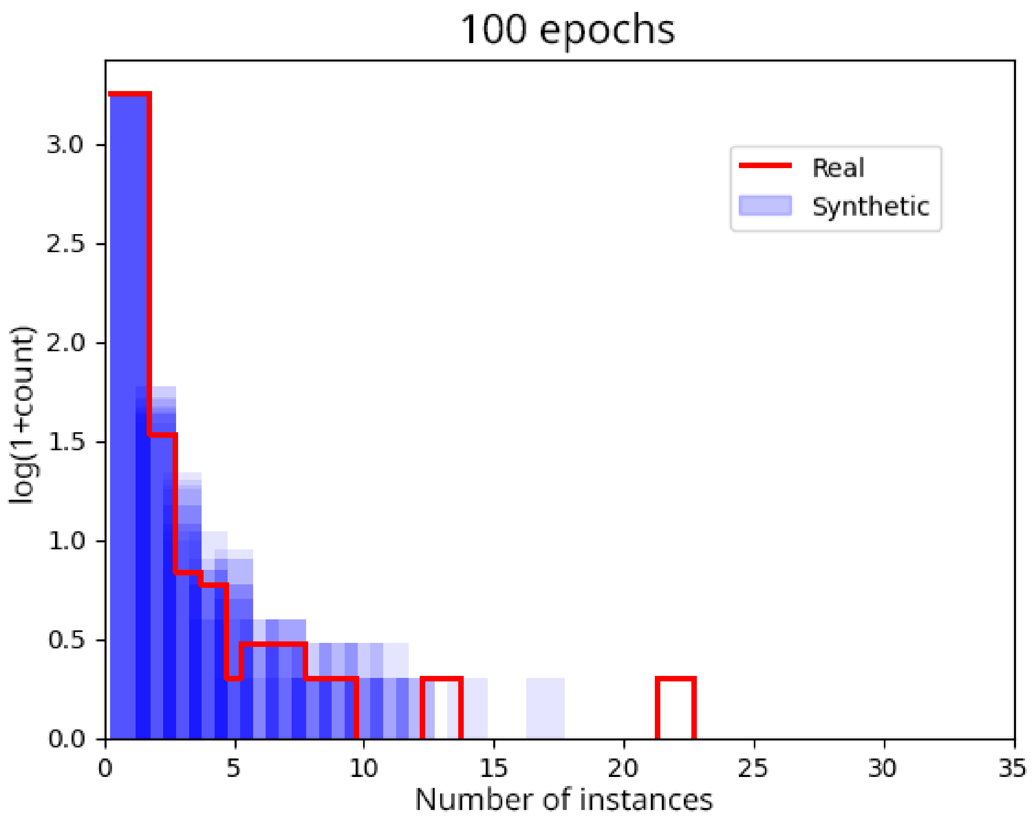
Instances of the same patients

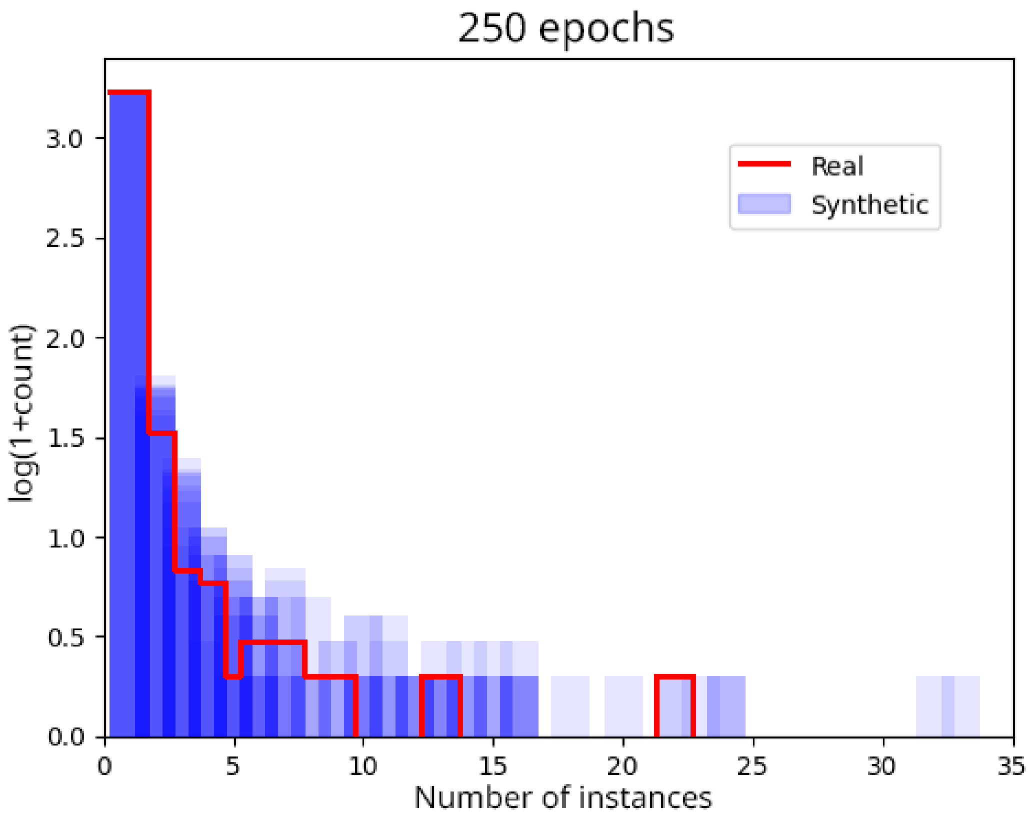
Model	copies of real data	duplicates
50 epochs	23 %	22 %
100 epochs	31 %	27 %
250 epochs	35 %	28 %
Real data	/////////	22 %

Rows with more than 20 instances

Real data		**50 epochs**		**100 epochs**		**250 epochs**	
mutations	count	mutations	count	mutations	count	mutations	count
None	50	['DNMT3A’, ’ITD’, ’IDH2_p.R140’, ’NPM1’]	40	['DNMT3A’, ’ITD’, ’IDH2_p.R140’, ’NPM1’]	31	['DNMT3A’, ’IDH2_p.R140’, ’NPM1’]	22
['DNMT3A’, ’ITD’, ’NPM1’]	39	['DNMT3A’, ’ITD’, ’NPM1’]	34	['DNMT3A’, ’ITD’, ’NPM1’]	39	['DNMT3A’, ’ITD’, ’NPM1’]	40
		['DNMT3A’, ’ITD’, ’NPM1’, ’TET2’]	29	['DNMT3A’, ’ITD’, ’NPM1’, ’TET2’]	36		







Generation of values not seen at training

Examples of anomalous mutation values

minusy	1.99863107460643
ITD	1.1492128678
ITD	1.0075290896
ITD	1.2621492128679
ITD	1.5667351129
NPM1	1.61533196440794
ITD	1.06776180698152
SF3B	1.07323750855
NPM1	1.472963723477
FLT3TKD	1.42915811088
FLT3TKD	1.70020533880904
ITD	1.2368240930869
ITD	1.150581793292
NPM1	1.2676249144422
minusy	1.91170431211499
IDH2_p.R172	1.16632443531828
NPM1	1.2568788501026
NRAS_p.G12_13	1.4517454132786
ITD	1.109514031485284
PTPN11	1.2087611225188
IDH1	1.936344976044
NPM	1.927446954141
ITD	1.30595482546
ITD	1.3901437371663
ITD	1.007529089664
ITD	1.779603011636
minusy	1.782340862423
ITD	1.2785763175906
NPM1	1.060917180013
KIT	1.092402464065
ITD	1.2765229295
ITD	1.2648870636
ITD	1.147843942505
ITD	1.05065023956

N. of anomalous mutation values

50 epochs		100 epochs		250 epochs	
2.3 ± 1.5	0.0006%	34.5 ± 2.8	0.01%	19.2 ± 3.8	0.006%

Very rarely a completely wrong value

It happened just 3 times on 9 millions generated values

11.20876411225188

very similar to the max value found in time, t=11.20876112

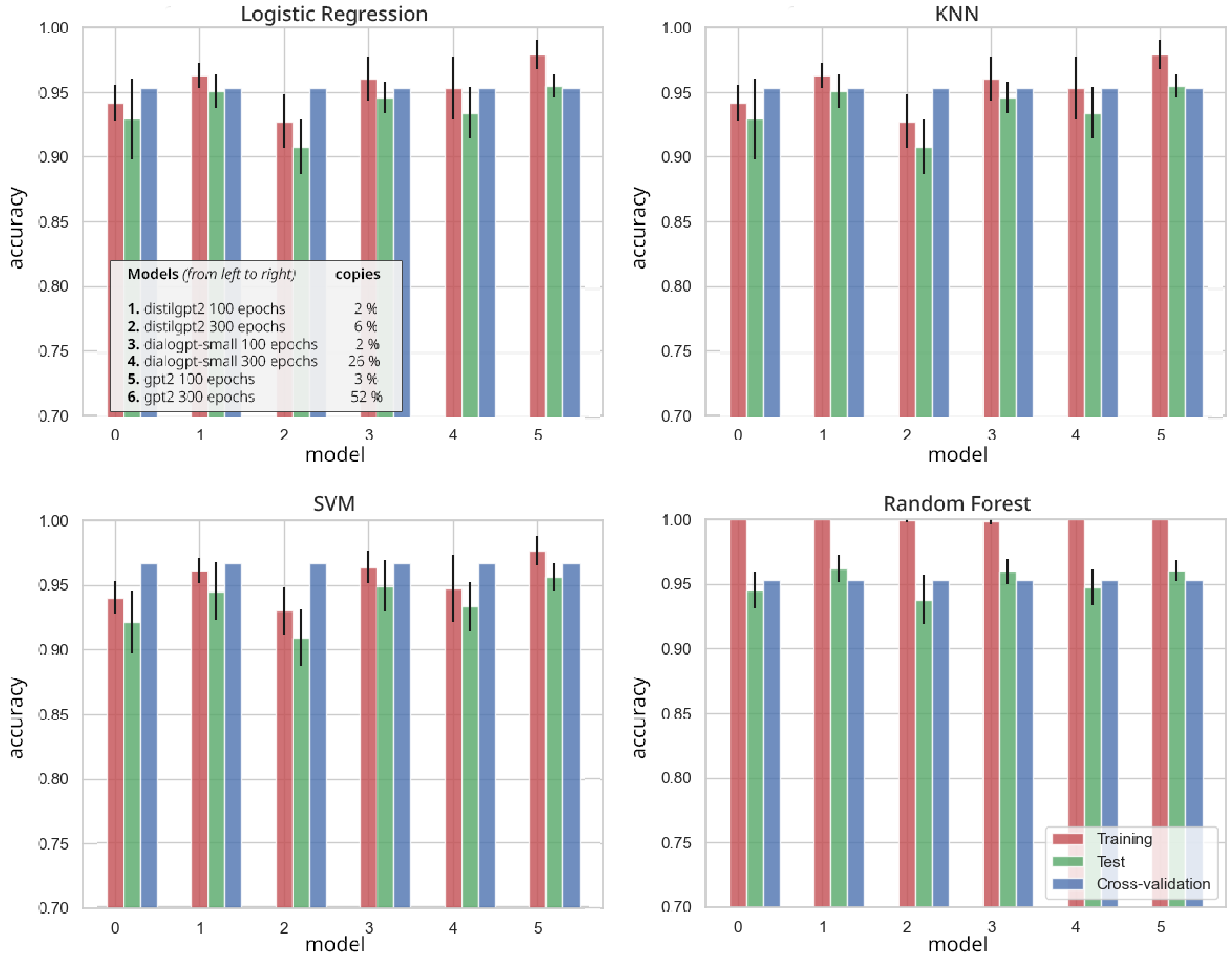
19110198494.0

11111123123

extremely large, but still starting with 1,
so maybe the point was misplaced

Testing Different **Large Language Models (LLMs)** *on the Iris Dataset*

LLM	copies from real data	
distilGPT-2 100 epochs	2.4 ± 1.8	2 %
distilGPT2 300 epochs	9.1 ± 3.3	6 %
GPT-2 100 epochs	3.6 ± 2.4	2 %
GPT-2 300 epochs	38.6 ± 4.3	26 %
DialoGPT 100 epochs	4.7 ± 1.6	3 %
DialoGPT 300 epochs	77.8 ± 6.2	52 %


Logistic Regression
KNN
SVM
Random Forest
accuracy
model
Models (from left to right) copies
1. distilgpt2 100 epochs 2 %
2. distilgpt2 300 epochs 6 %
3. dialogpt-small 100 epochs 2 %
4. dialogpt-small 300 epochs 26 %
5. gpt2 100 epochs 3 %
6. gpt2 300 epochs 52 %
Training
Test
Cross-validation