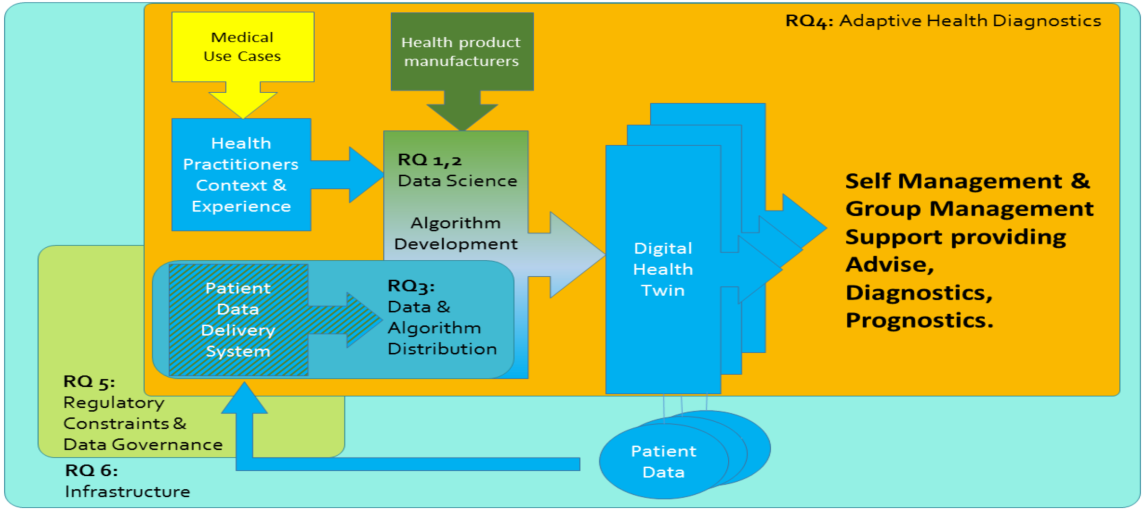
Private Federated Machine Learning
The EPI Project

Saba Amiri
Adam Belloum, Sander Klous, Leon Gommans, Eric Nalisnick

Enabling Personalized Interventions

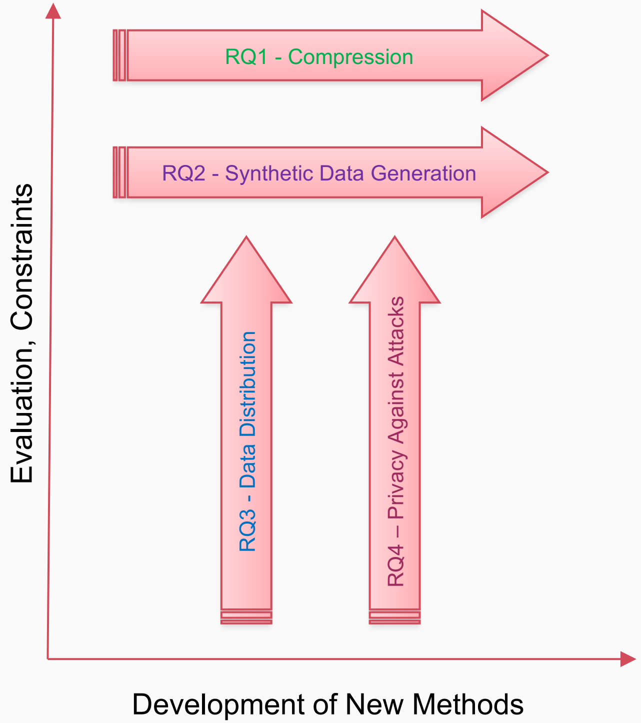
- EPI[*] project broadly aims to create a Digital Health Twin
  - The digital reflection of a person in terms of health related data and allows algorithms
  - Enables distributed processing of disparate relevant data, e.g. perform monitoring or predict outcomes of treatments



https://enablingpersonalizedinterventions.nl/

Basic Setting

- RQ4-1 - How To Achieve Differential Privacy Through Compression?
- RQ4-2 - How to generate differentially-private synthetic tabular data in a distributed setting?
- RQ4-3 - What is the effect of non-i.i.d data distribution on the performance of differentially private machine learning models?
- RQ4-4 - How can we measure the privacy level of DP machine learning methods from the perspective of privacy attacks?



Basic Setting

- Medical use-cases (EPI[*])
- Private, distributed, large datasets
- Common goal: train a machine learning model on these datasets while preserving privacy of the individuals in the datasets

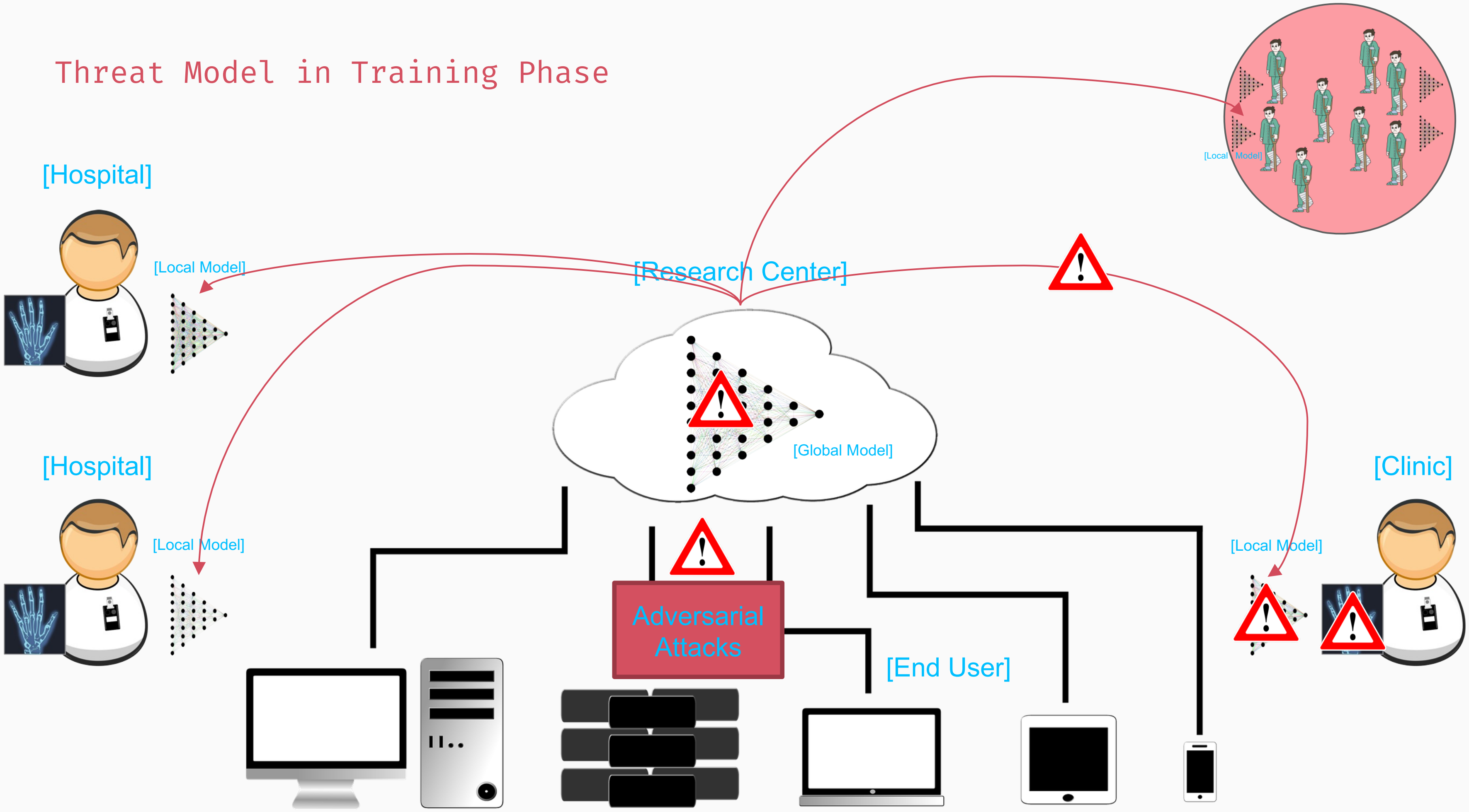
https://enablingpersonalizedinterventions.nl/

Basic Setting

- Medical use-cases (EPI[*])
- Private, distributed, large datasets
- Common goal: train a machine learning model on these datasets while preserving privacy of the individuals in the datasets
- Initial solution: accumulate data, train a centralized model
- Poses challenges, e.g. privacy, communication, etc.

https://enablingpersonalizedinterventions.nl/

Threat Model in Training Phase



Impact of non-i.i.d Distribution on Federated Learning

The Problem w/ Federated Learning

- **Privacy**
  - FL solves the problem of data sharing
  - The training process is vulnerable
  - The model could leak information after being trained
- **Data distribution**
  - i.i.d assumption about data
  - 4 main types of imbalance in the data
    - Feature
    - Label
    - Temporal
    - Node
  - Has disparate impact on performance, fairness

The Problem w/ Federated Learning

- **Adult dataset**

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	label
0	17	Private	124130	Some-college	9	Separated	Protective-serv	Not-in-family	White	Male	30	0	40	Haiti	<=50K
1	26	Private	168914	HS-grad	10	Married-civ-spouse	Handlers-cleaners	Husband	Asian-Pac-Islander	Female	21	1	39	Yugoslavia	<=50K
2	33	Self-emp-not-inc	218757	HS-grad	11	Married-civ-spouse	Machine-op-inspct	Not-in-family	White	Male	29	0	24	United-States	>50K
3	62	Self-emp-not-inc	558635	Bachelors	9	Never-married	Prof-specialty	Wife	White	Male	51	1	40	United-States	<=50K
4	27	?	143612	Masters	13	Separated	Priv-house-serv	Unmarried	White	Male	89	-2	40	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	44	Private	179779	HS-grad	9	Never-married	Adm-clerical	Husband	White	Male	2	-3	40	United-States	<=50K
996	28	Self-emp-not-inc	180882	Bachelors	11	Married-civ-spouse	Adm-clerical	Other-relative	Black	Female	43	5	40	United-States	<=50K
997	15	Private	166548	Bachelors	6	Married-civ-spouse	Protective-serv	Other-relative	White	Female	23	7	38	United-States	<=50K
998	19	Private	158057	Doctorate	8	Never-married	Other-service	Not-in-family	White	Male	9	-1	40	United-States	>50K
999	19	Private	119228	Bachelors	13	Divorced	Other-service	Unmarried	White	Male	69	5	40	United-States	<=50K

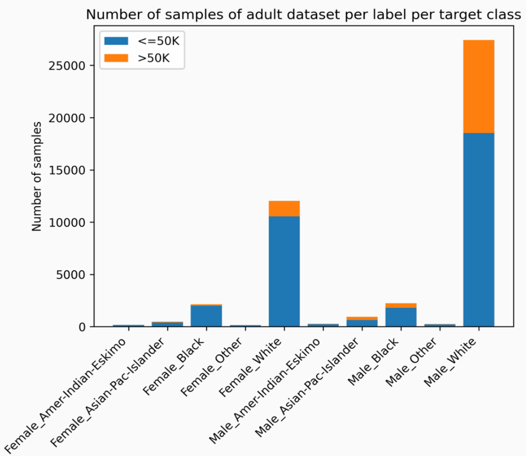
1 age Min 17. 1st Qu 28. Median 37. Mean 38.5816 3rd Qu 48. Max 90.	2 workclass Private 22 696 Self-emp-not-inc 2541 Local-gov 2093 ? 1836 State-gov 1298 Self-emp-inc 1116 (Other) 981	3 education HS-grad 10 501 Some-college 7291 Bachelors 5355 Masters 1723 Assoc-voc 1382 11th 1175 (Other) 5134	4 education-num Min 1. 1st Qu 9. Median 10. Mean 10.0807 3rd Qu 12. Max 16.	5 marital-status Married-civ-spouse 14 976 Never-married 10 683 Divorced 4443 Separated 1025 Widowed 993 Married-spouse-absent 418 Married-AF-spouse 23
6 occupation Prof-specialty 4140 Craft-repair 4099 Exec-managerial 4066 Adm-clerical 3770 Sales 3650 Other-service 3295 (Other) 9541	7 relationship Husband 13 193 Not-in-family 8305 Own-child 5068 Unmarried 3446 Wife 1568 Other-relative 981	8 race White 27 816 Black 3124 Asian-Pac-Islander 1039 Amer-Indian-Eskimo 311 Other 271	9 sex Male 21 790 Female 10 771	10 capital-gain 1st Qu 0. 3rd Qu 0. Median 0. Min 0. Mean 1077.65 Max 99 999.
11 capital-loss 1st Qu 0. 3rd Qu 0. Median 0. Min 0. Mean 87.3038 Max 4356.	12 hours-per-week Min 1. 1st Qu 40. Median 40. Mean 40.4375 3rd Qu 45. Max 99.	13 native-country United-States 29 170 Mexico 643 ? 583 Philippines 198 Germany 137 Canada 121 (Other) 1709	14 income <=50K 24 720 >50K 7841	

The Problem w/ Federated Learning

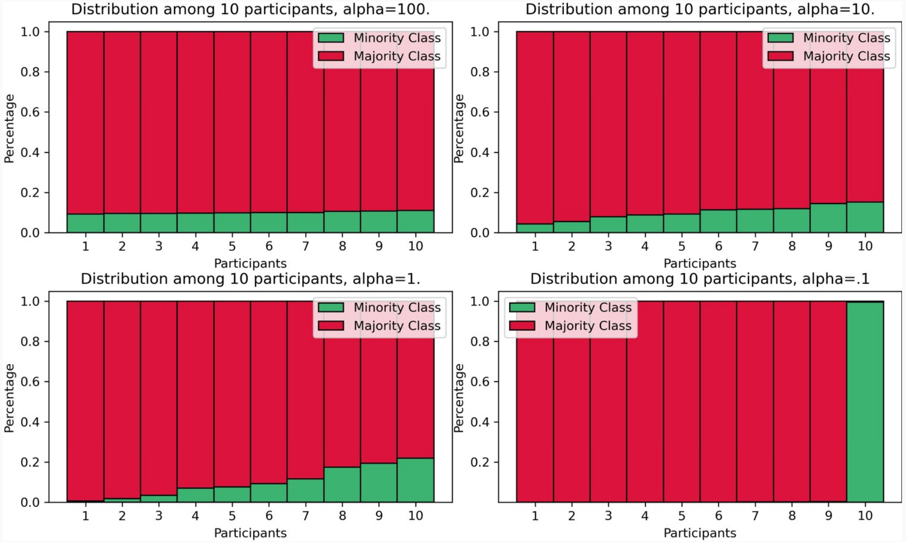
- **Adult dataset**

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	label
0	17	Private	124130	Some-college	9	Separated	Protective-serv	Not-in-family	White	Male	30	0	40	Haiti	<=50K
1	26	Private	168914	HS-grad	10	Married-civ-spouse	Handlers-cleaners	Husband	Asian-Pac-Islander	Female	21	1	39	Yugoslavia	<=50K
2	33	Self-emp-not-inc	218757	HS-grad	11	Married-civ-spouse	Machine-op-inspct	Not-in-family	White	Male	29	0	24	United-States	>50K
3	62	Self-emp-not-inc	558635	Bachelors	9	Never-married	Prof-specialty	Wife	White	Male	51	1	40	United-States	<=50K
4	27	?	143612	Masters	13	Separated	Priv-house-serv	Unmarried	White	Male	89	-2	40	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	44	Private	179779	HS-grad	9	Never-married	Adm-clerical	Husband	White	Male	2	-3	40	United-States	<=50K
996	28	Self-emp-not-inc	180882	Bachelors	11	Married-civ-spouse	Adm-clerical	Other-relative	Black	Female	43	5	40	United-States	<=50K
997	15	Private	166548	Bachelors	6	Married-civ-spouse	Protective-serv	Other-relative	White	Female	23	7	38	United-States	<=50K
998	19	Private	158057	Doctorate	8	Never-married	Other-service	Not-in-family	White	Male	9	-1	40	United-States	>50K
999	19	Private	119228	Bachelors	13	Divorced	Other-service	Unmarried	White	Male	69	5	40	United-States	<=50K

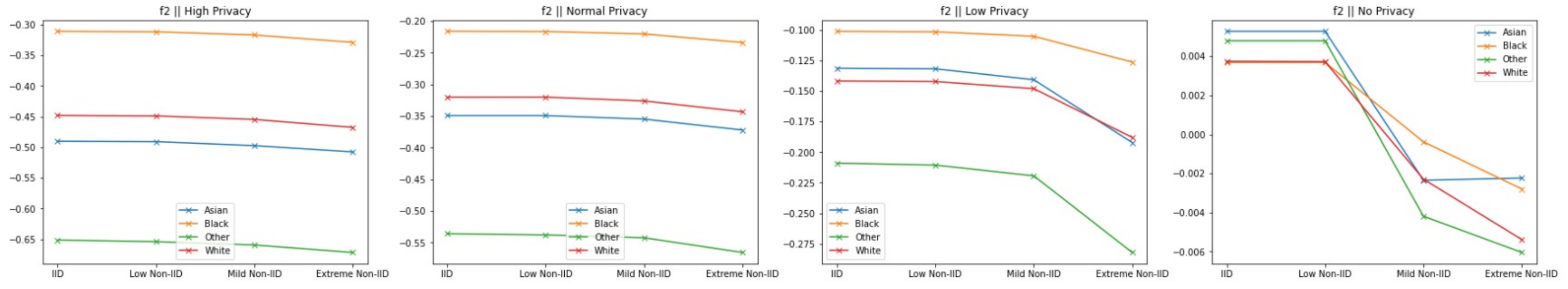
1 age Min 17. 1st Qu 28. Median 37. Mean 38.5816 3rd Qu 48. Max 90.	2 workclass Private 22 696 Self-emp-not-inc 2541 Local-gov 2093 ? 1836 State-gov 1298 Self-emp-inc 1116 (Other) 981	3 education HS-grad 10 501 Some-college 7291 Bachelors 5355 Masters 1723 Assoc-voc 1382 11th 1175 (Other) 5134	4 education–num Min 1. 1st Qu 9. Median 10. Mean 10.0807 3rd Qu 12. Max 16.	5 marital–status Married-civ-spouse 14 976 Never-married 10 683 Divorced 4443 Separated 1025 Widowed 993 Married-spouse-absent 418 Married-AF-spouse 23
6 occupation Prof-specialty 4140 Craft-repair 4099 Exec-managerial 4066 Adm-clerical 3770 Sales 3650 Other-service 3295 (Other) 9541	7 relationship Husband 13 193 Not-in-family 8305 Own-child 5068 Unmarried 3446 Wife 1568 Other-relative 981	8 race White 27 816 Black 3124 Asian-Pac-Islander 1039 Amer-Indian-Eskimo 311 Other 271	9 sex Male 21 790 Female 10 771	10 capital–gain 1st Qu 0. 3rd Qu 0. Median 0. Min 0. Mean 1077.65 Max 99 999.
11 capital–loss 1st Qu 0. 3rd Qu 0. Median 0. Min 0. Mean 87.3038 Max 4356.	12 hours–per–week Min 1. 1st Qu 40. Median 40. Mean 40.4375 3rd Qu 45. Max 99.	13 native–country United-States 29 170 Mexico 643 ? 583 Philippines 198 Germany 137 Canada 121 (Other) 1709	14 income <=50K 24 720 >50K 7841	

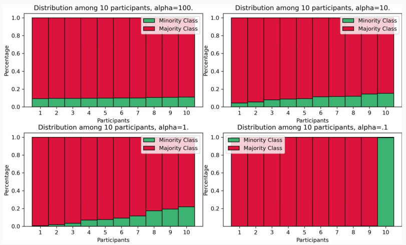


Our Research: Impact of non-i.i.d data on private FL



Our Research: Impact of non-i.i.d data on private FL





Differentially Private Synthetic Data Generation

Our Research: Differentially Private Synthetic Data Generation

- Adding DP to ML models is costly
- Alternatively, we can make the data "privacy preserving"
- How?
  - Use a differentially private generative model to estimate the distribution of the data
  - Train the model on real data
  - Use model to generate a synthetic dataset
  - Due to post-processing theorem, any model trained on our synthetic data is at least differentially private with the same level as our generative model

Our Research: Differentially Private Synthetic Data Generation

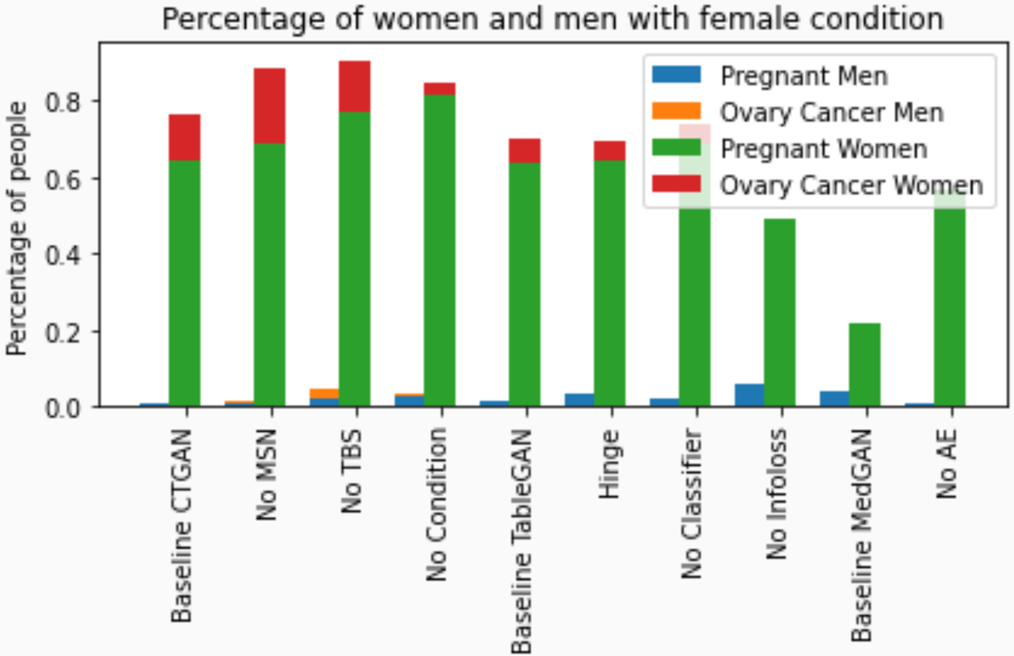
- Generate privacy preserving synthetic data from original data
- Differentially private with an acceptable privacy budget
- On tabular data
- Preserve statistical properties
- Maintain machine learning efficacy
- Distributed environment
- No i.i.d assumptions about data distribution

Our Research: Differentially Private Synthetic Data Generation

- Generate privacy preserving synthetic data from original data
- Differentially private with an acceptable privacy budget
- On tabular data
- Preserve statistical properties
- Maintain machine learning efficacy
- Distributed environment
- No i.i.d assumptions about data distribution
- **Data quality: Semantic integrity**

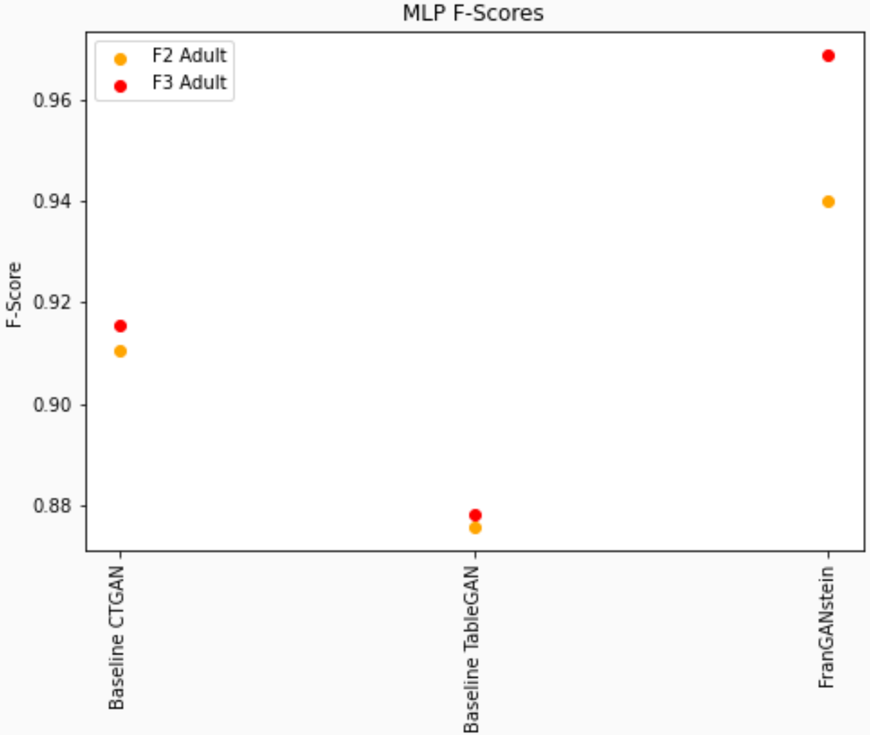
Our Research: Differentially Private Synthetic Data Generation

- Why semantic integrity?



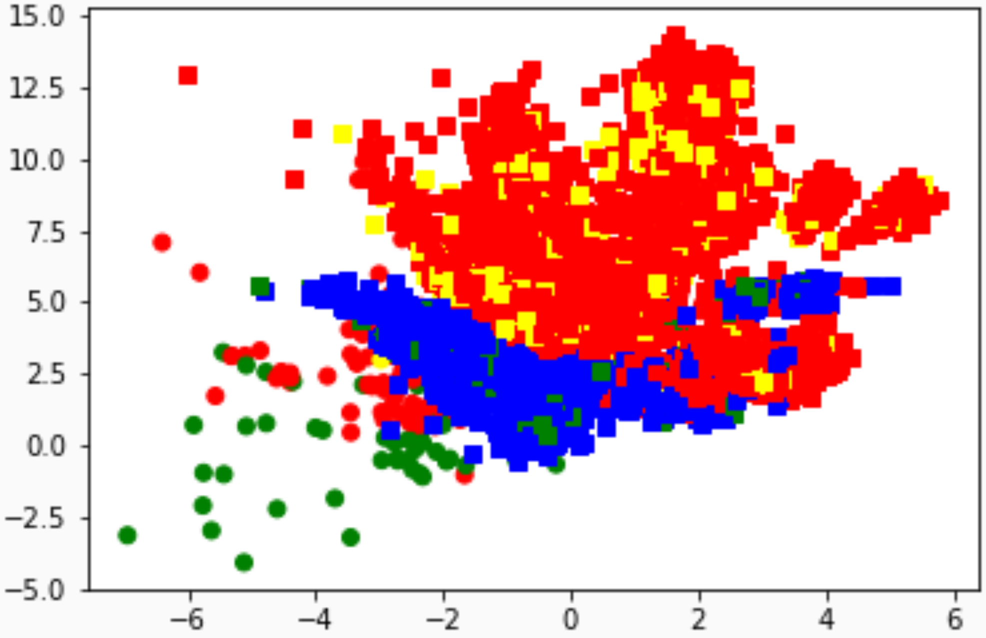
Our Research: Differentially Private Synthetic Data Generation

- Proposed model's performance



Our Research: Differentially Private Synthetic Data Generation

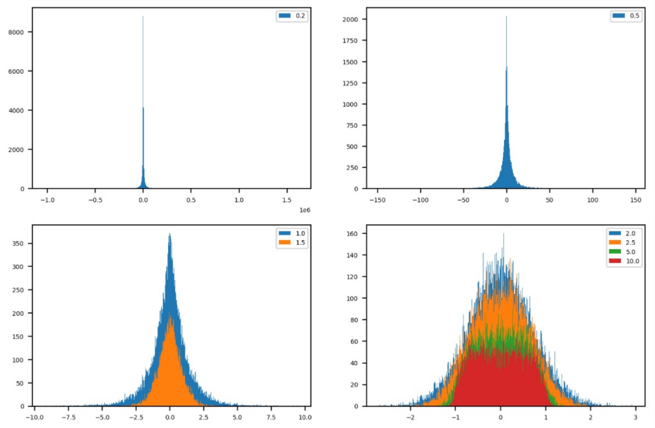
- Proposed model's quality



Synthetic data generation for data with long tailed distributions

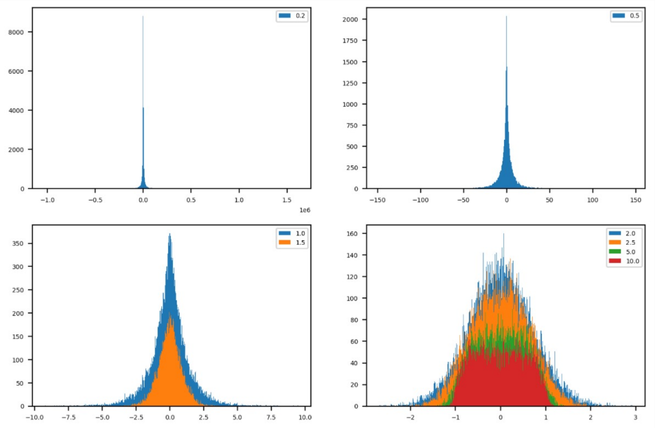
Synthetic data generation for data with long tailed distributions

- Long tailed data
    - Have a generative model that is able to capture the tail behavior of long-tailed distributions



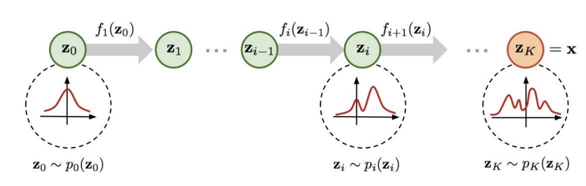
Synthetic data generation for data with long tailed distributions

- Long tailed data
  - Have a generative model that is able to capture the tail behavior of long-tailed distributions
- Initial approach: GANs with a differentiable generalized Gaussian base distribution



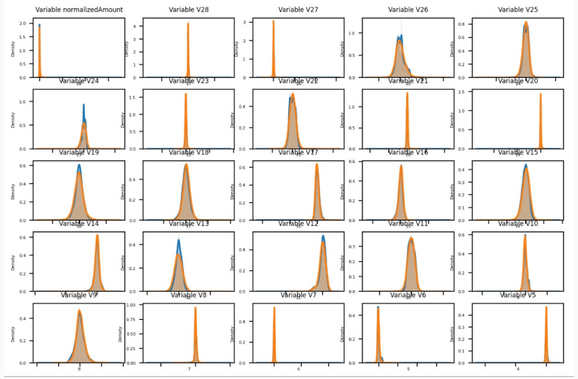
Synthetic data generation for data with long tailed distributions

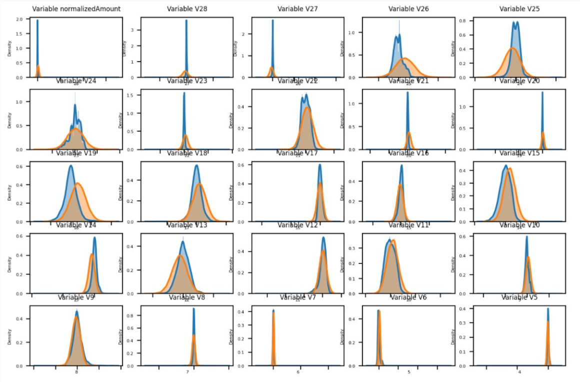
- Second approach: normalizing flows
  - A normalizing flow
    - describes the transformation of a probability density through a sequence of invertible mappings.
    - Transforms a simple distribution into a complex one by applying a sequence of invertible transformation functions.
    - Flowing through a chain of transformations, we repeatedly substitute the variable for the new one according to the change of variables theorem and eventually obtain a probability distribution (i.e. normalized) of the final target variable
    - Normalizing flows can exactly estimate the density function
    - There is theory on capabilities of NFs on capturing the tail behavior of long-tailed distributions



Synthetic data generation for data with long tailed distributions

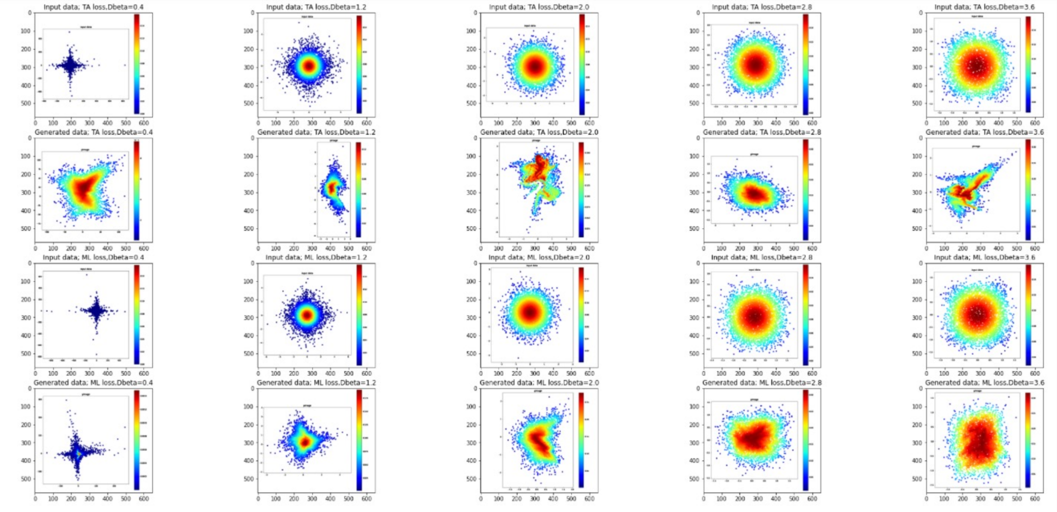
- Second approach: normalizing flows
    - How does a flow-based model compare to a GAN?
- FLOW GAN





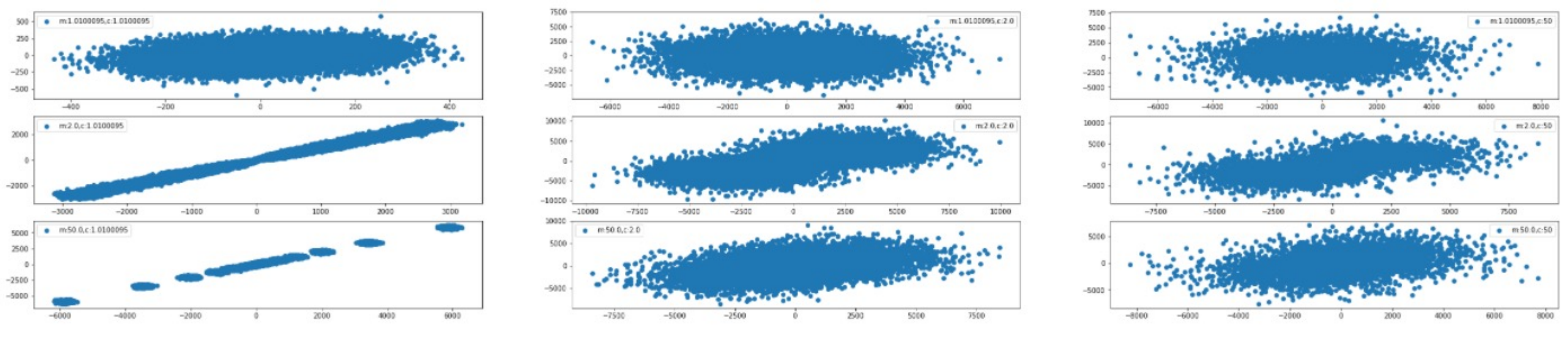
Synthetic data generation for data with long tailed distributions

- Second approach: normalizing flows
  - Next step: dual training with ML-based training for the flow model and a loss function utilizing tail-adaptive alpha divergence for the base parameters



Synthetic data generation for data with long tailed distributions

- Last step: flexible mixture base distribution
  - Smooth contraction/expansion of the base mixture distribution to help the flow-based model capture the tail properties of the target distribution



Thank You!