EPI RQ4 Research Update: Privacy Preserving Distributed Machine Learning

Saba Amiri

s.amiri@uva.nl

Supervisor: Adam Belloum

Promoters: Sander Klous, Leon Gommans

Multiscale Networked Systems Group

1 July 2021

Lessons Learned

ML system outputs are vulnerable

Current methods adequate?

Data anonymization is not enough

What is the solution?

Privacy preserving machine learning

Which privacy preserving mechanism?

Differential Privacy

What are the costs?

Added complexity
- Privacy design
- Utility/privacy trade-off

Any independent factors?

Non-i.i.d data
- Class imbalance
- Imbalanced distribution among hospitals

And if all these problems are fixed?

Definitions of privacy are abstract

How to use it in real world?

Deployment and governance mechanism needed

Recap

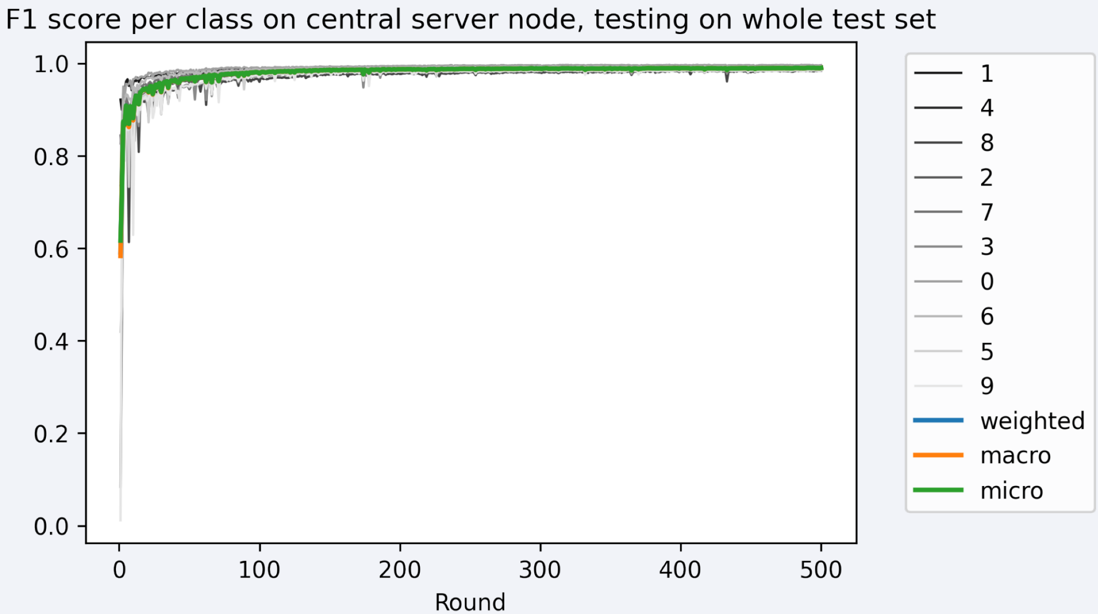
- Differentially private compressive federated learning
  - Simple federated learning setup
  - Add differential privacy through compression mechanism necessary due to constrained communication channel
- Differentially private synthetic data generation
  - Distributed datasets
  - Privacy preserving
  - Non-i.i.d data distribution among nodes
  - Skewed/imbalanced dataset
- Impact of non-i.i.d distribution on the performance of machine learning models
  - Different federated learning fusion schemes
  - Different non-i.i.d data distribution schemes
  - Impact of differential privacy
- Distributed learning pipeline
  - Collaboration with Jamila, Onno on connection of RQ4 with RQ6/BRANE
  - Research on using Vantage6[1] as the distributed machine learning infrastructure (as opposed to more generic solutions, e.g. managing the distributed pipeline through use of Pytorch distributed)

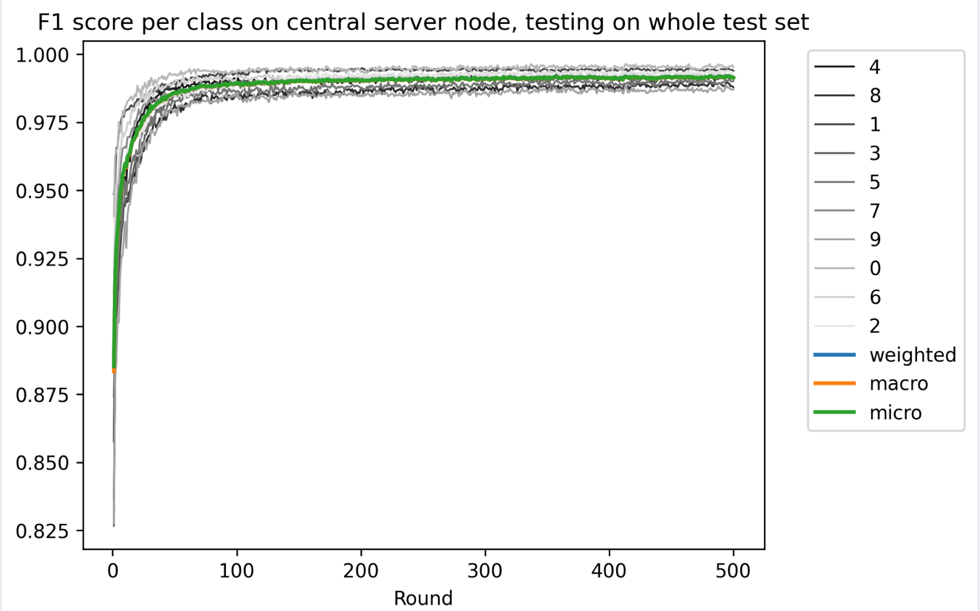
[1] Moncada-Torres, Arturo, et al. "VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange." AMIA Annual Symposium Proceedings. Vol. 2020. American Medical Informatics Association, 2020.

Results – Data Distribution

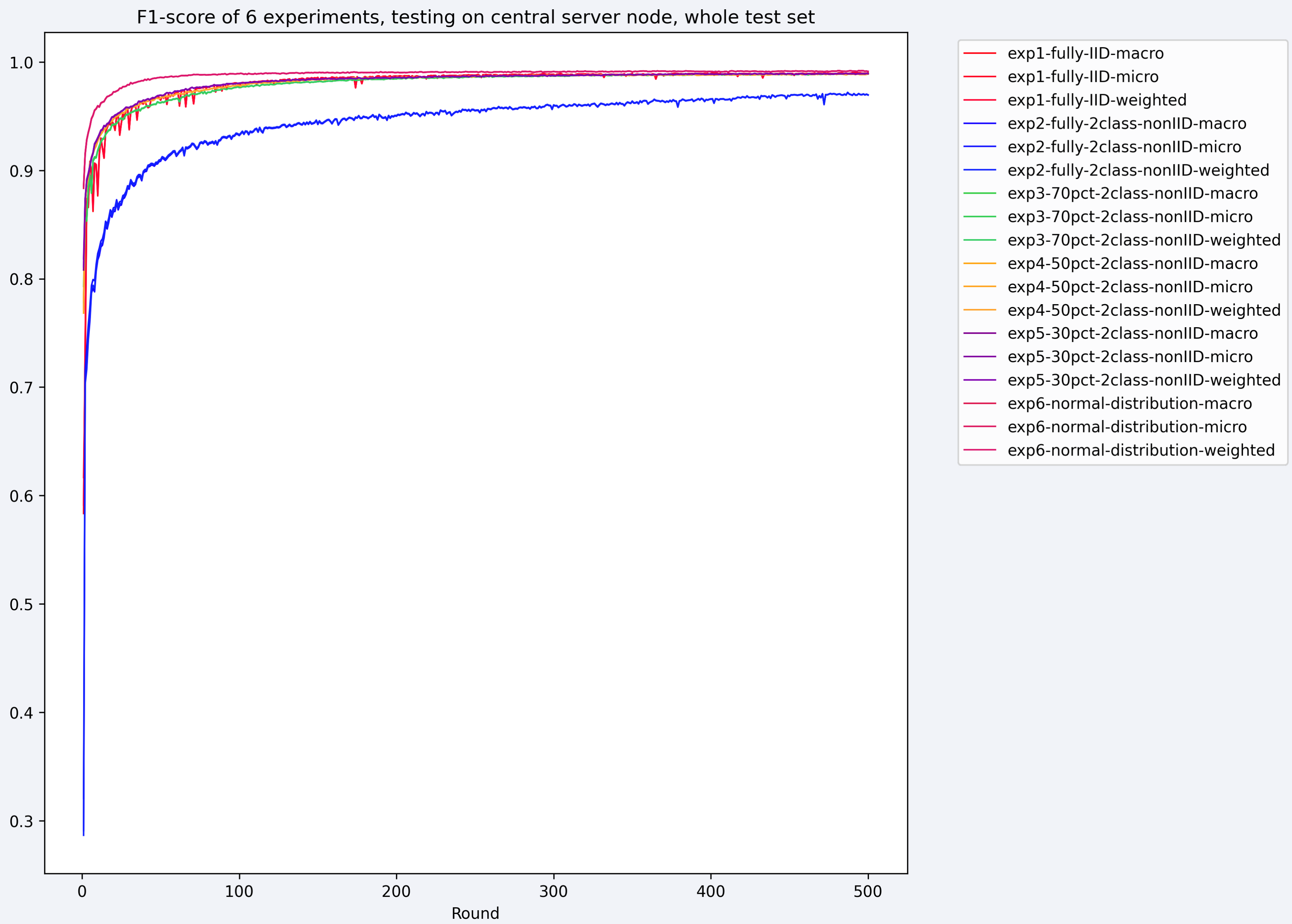
- Toy example
  - Dataset with ten classes
  - Data distributed among 5 different hospitals
- Different distribution schemes being researched
  - Fully i.i.d (each of the 5 hospitals have the same number of each of the 10 classes)
  - Fully non-i.i.d (All the samples of each class reside on **only one** node)
  - Partial non-i.i.d (samples from 5 of the classes are distributed identically among 5 hospitals, the next 5 class each reside only on one hospitals)
  - Statistical distribution (all hospitals have some samples of all classes, the distribution of samples among nodes follows a statistical distribution, e.g. Gaussian)
- Metrics
  - Machine learning utility
  - Class-conditional utility
  - Fairness (utility and/or imbalance in under-represented classes)

Results – Data Distribution, Fully i.i.d (Ideal) vs. Normal Dist.





Results – Data Distribution, Overall



Results – Privacy Preserving Synthetic Data Generation (PPSDG)

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

Results – PPSDG, Machine Learning Efficacy

- Dataset: Adult income dataset
  - Class label (Income): ”>50k”, “<50k”
- We trained 3 baseline generative models on the dataset
- We generated 3 synthetic datasets using the 3 generative models
- We designed a simple classification model to predict income
- We trained the classification model 4 times using original dataset and 3 synthetic datasets

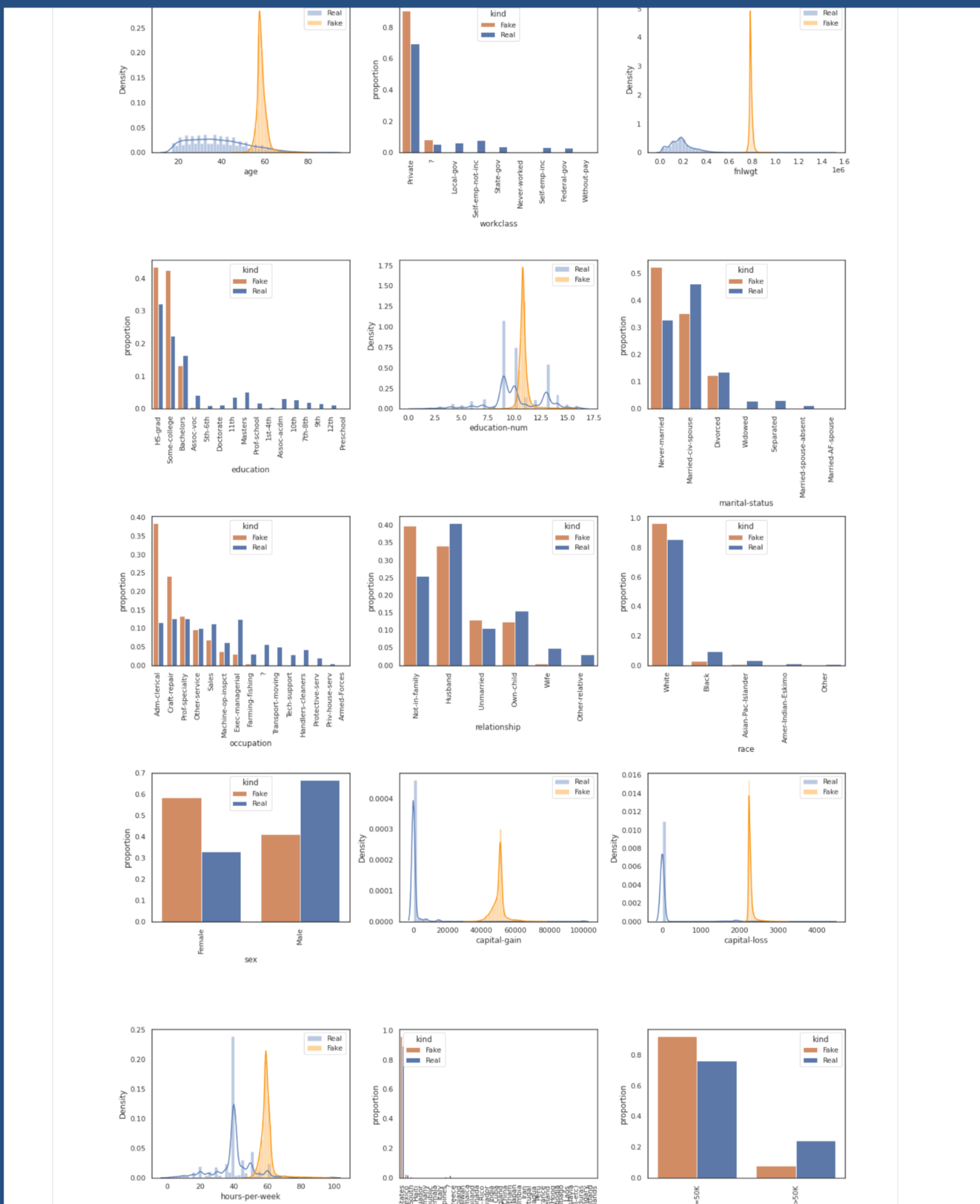
Dataset	SDG Model	Accuracy %
Original	-	83.6
Synthetic 1	GAN	82.1
Synthetic 2	GAN	82.8
Synthetic 3	**GAN**	**98**

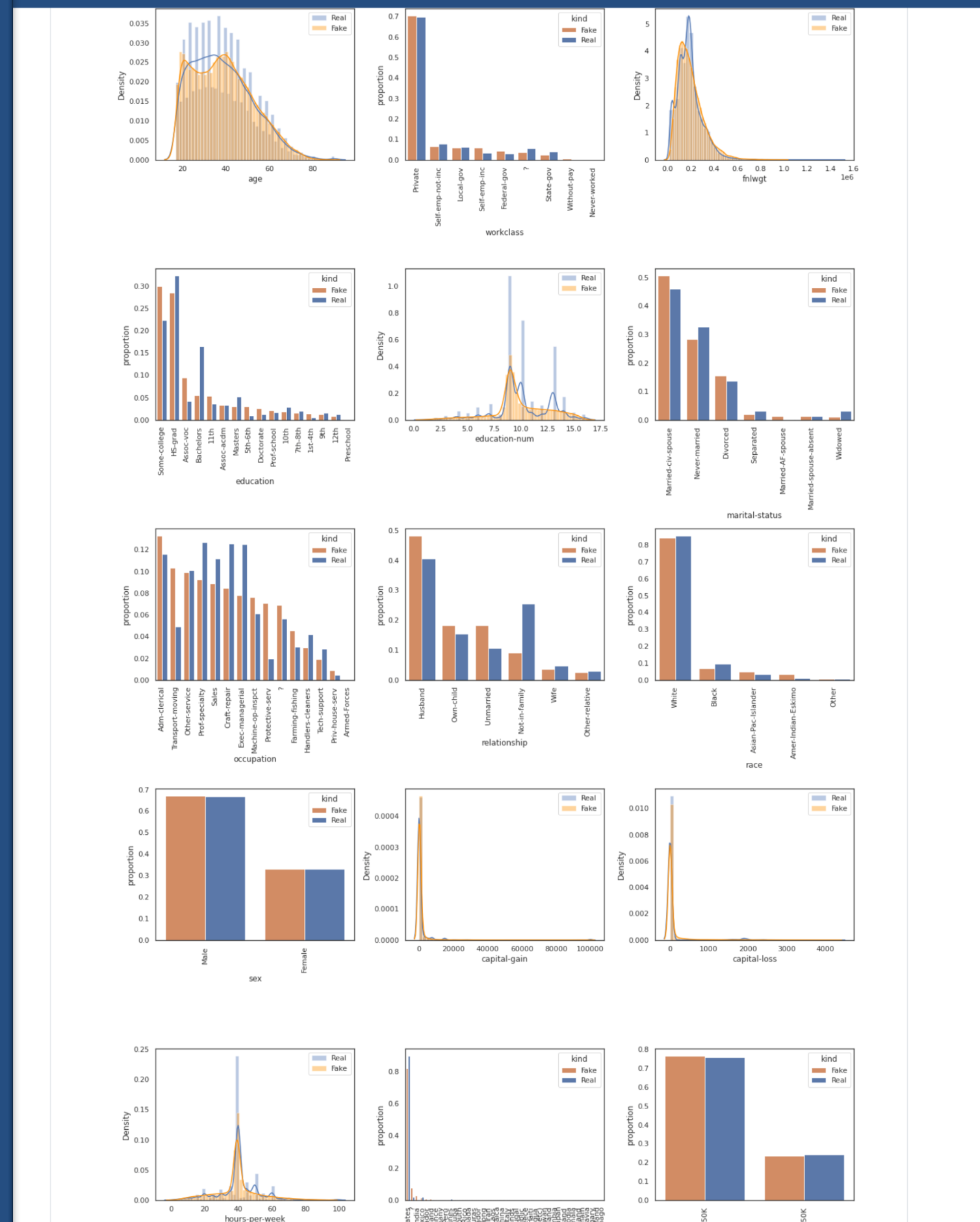
Results – PPSDG, Machine Learning Efficacy

- Dataset: Adult income dataset
  - Class label (Income): ”>50k”, “<50k”
- We trained 3 baseline generative models on the dataset
- We generated 3 synthetic datasets using the 3 generative models
- We designed a simple classification model to predict income
- We trained the classification model 4 times using original dataset and 3 synthetic datasets

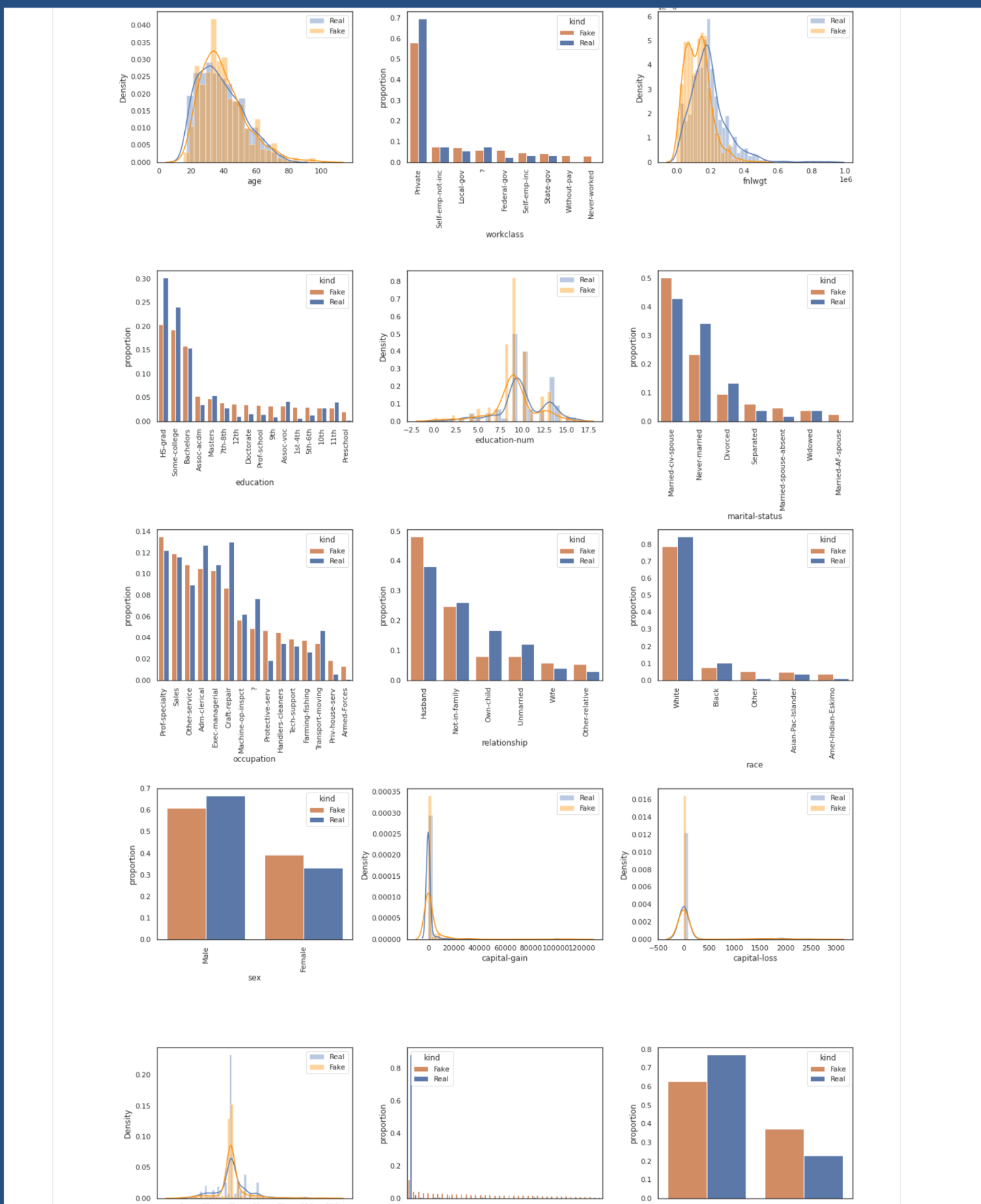
	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	label
0	17	Private	124130	Some-college	9	Separated	Protective-serv	Not-in-family	White	Male	30	0	40	Haiti	<=50K
1	26	Private	168914	HS-grad	10	Married-civ-spouse	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

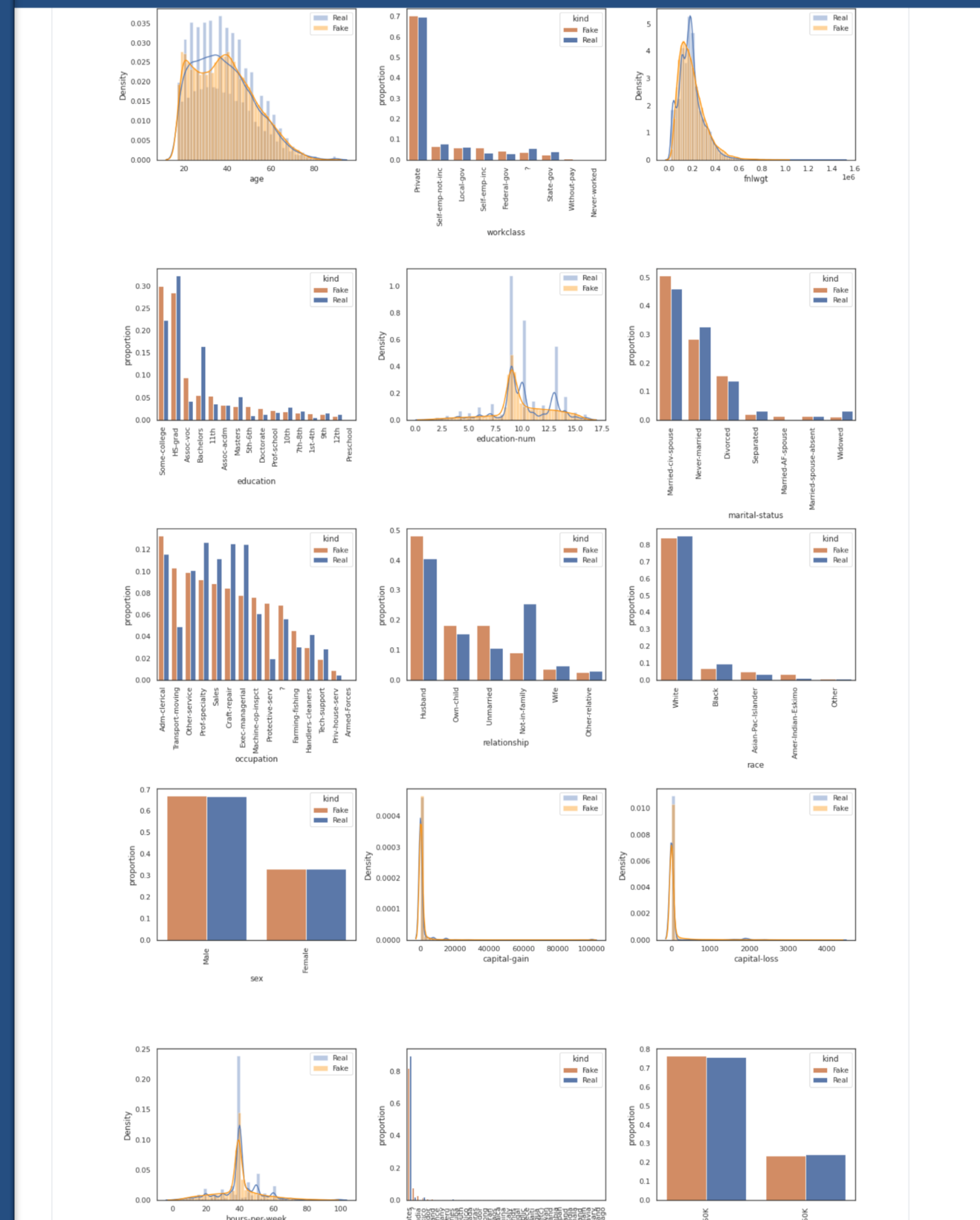
Results – PPSDG, Shortcomings in Baseline, 2 models





Results – PPSDG, Shortcomings in Baseline, Same Model, Skewed Data





Future Works

- [1] August/September 2021
  - Finish phase 1 of research on effect of non-i.i.d distribution in federated learning, submit paper
  - Finish phase 1 of research on PPSDG, submit paper
  - Make the codebase public
- [2] December 2021/January/2022
  - Finish phase 2 of research on PPSDG, submit paper
  - Apply results from phase 1 of ‘non-i.i.d’ research in PPSDG (either the same research or separate one)
- TBD
  - Link with infrastructure/BRANE
  - Extend collaboration with Vantage6 if feasible

Employing Results in Practice

- What we can offer right now
  - Measure the privacy risks, vulnerabilities of current machine learning systems
  - Pipeline to perform federated learning on distributed private datasets
  - Train a machine learning model in a privacy preserving manner (differentially private)
  - Generate privacy preserving synthetic data (conditioned on being analyzed)
- What we will be able to offer in the future
  - Generate privacy preserving synthetic data <u>with theoretical guarantees</u>
  - Measurement and analysis of fairness and robustness of machine learning models against different data distribution scenarios
- How do we test our methods?
  - Datasets
    - Image datasets
      - MNIST, CIFAR-10
    - Tabular datasets (non-medical)
      - adult, census, covertype, intrusion and news
    - Tabular datasets (medical)
      - MIMIC-III
  - Interpretation, domain expertise
    - Following standard in ML research on ML-related aspects of the work
    - Following already existing research for domain-specific interpretation

Tailoring Results for EPI Use-Cases

- Access to the data
- Domain expertise
- Practical use-case
- Resources
- Evaluation framework
- Plan to incorporate results in practice
- Update standards on privacy in machine learning
- Extend differential privacy to any data analysis method (going beyond anonymization)

Thank you!

My direct collaborators in chronological order

- Serge van Haag (AI)
- Boris Egelie (AI)
- Tidi Stamatiou (AI)
- Carlijn Nijhuis (Computer Science)
- Mike Schouw (Computer Science)
- Jetske Beks (Computer Science)
- Willemijn Beks (Computer Science)
- Yu Wang (Computer Science)
- Simon Tokloth (Data Science)