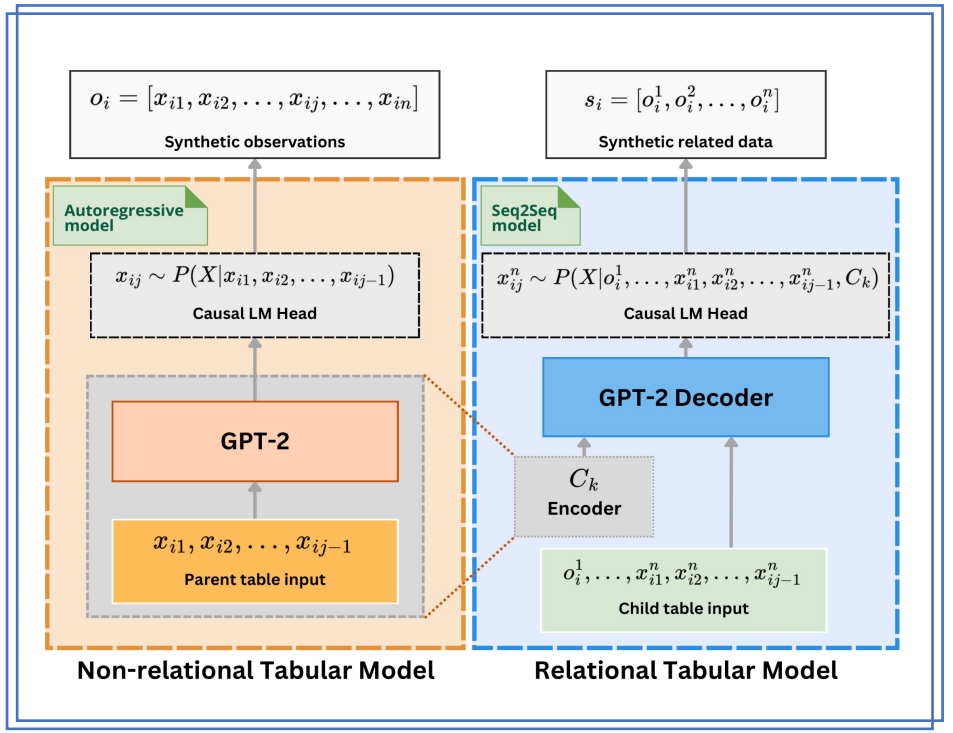
Generating synthetic data using REaLTabFormer, and assessing the probabilistic measure of statistical disclosure risk

Olivier Dupriez | odupriez@worldbank.org
Deputy Chief Statistician (World Bank)

Aivin V. Solatorio | asolatorio@worldbank.org
Data Scientist (World Bank)
GitHub: @avsolatorio

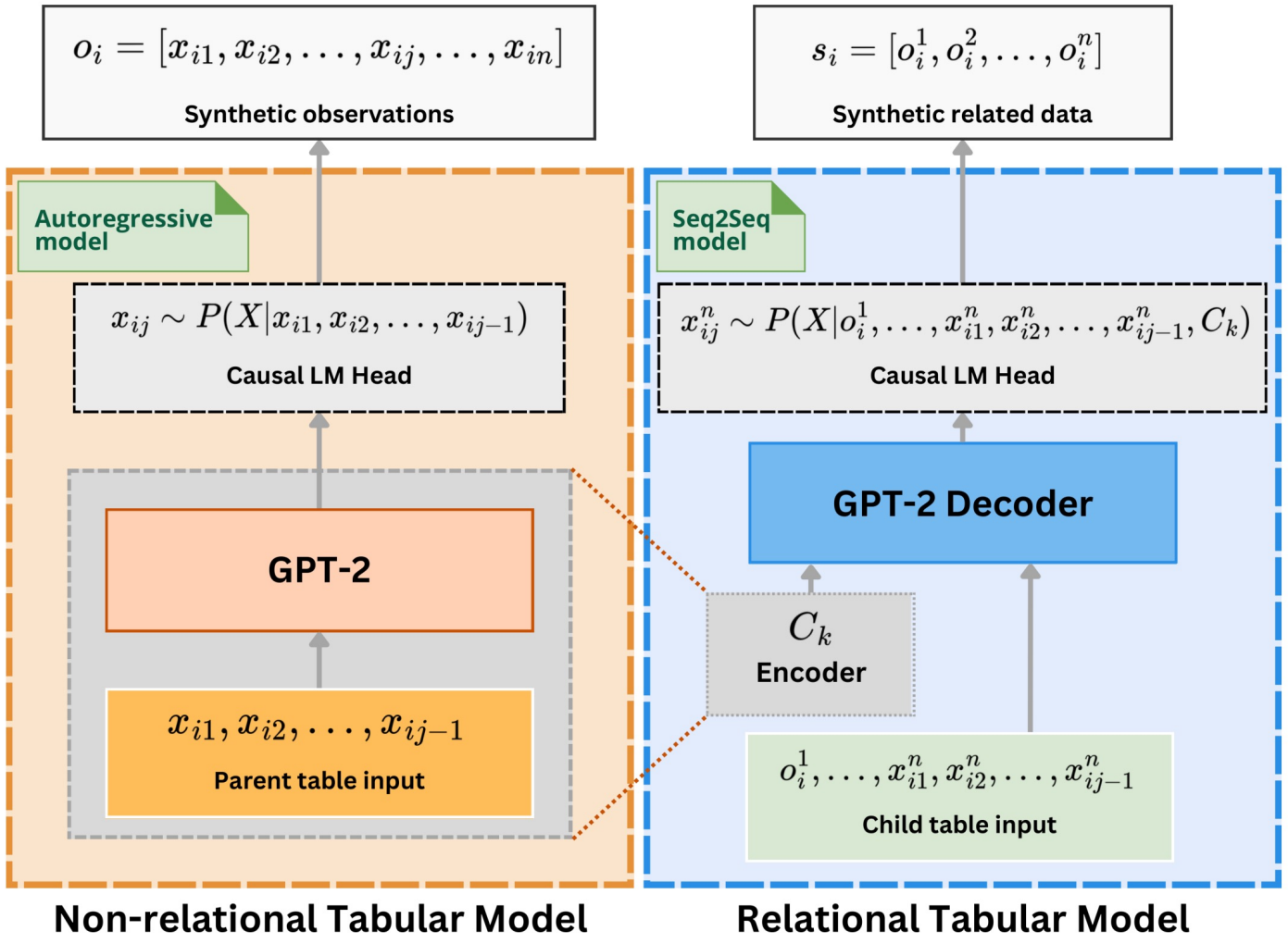


Models for generating synthetic data

- Bayesian Generative Models
- Graphical models
- Variational Autoencoders
- Generative Adversarial Networks (GAN)
- Diffusion models
- **Transformer-based generative models (GPT)**

The REaLTabFormer Model

- A GPT architecture-based synthetic tabular data generator
- Can model both regular tabular data and relational datasets
- Can generate or impute missing values.
- Supports all data types.
- Requires very minimal configuration, and basically works out-of-the-box.
- *No differential privacy guarantees* but implements mechanisms to mitigate data-copying.



Paper: https://arxiv.org/abs/2302.02041
GitHub: https://github.com/worldbank/REaLTabFormer

Comparison of differentially-private model and REaLTabFormer

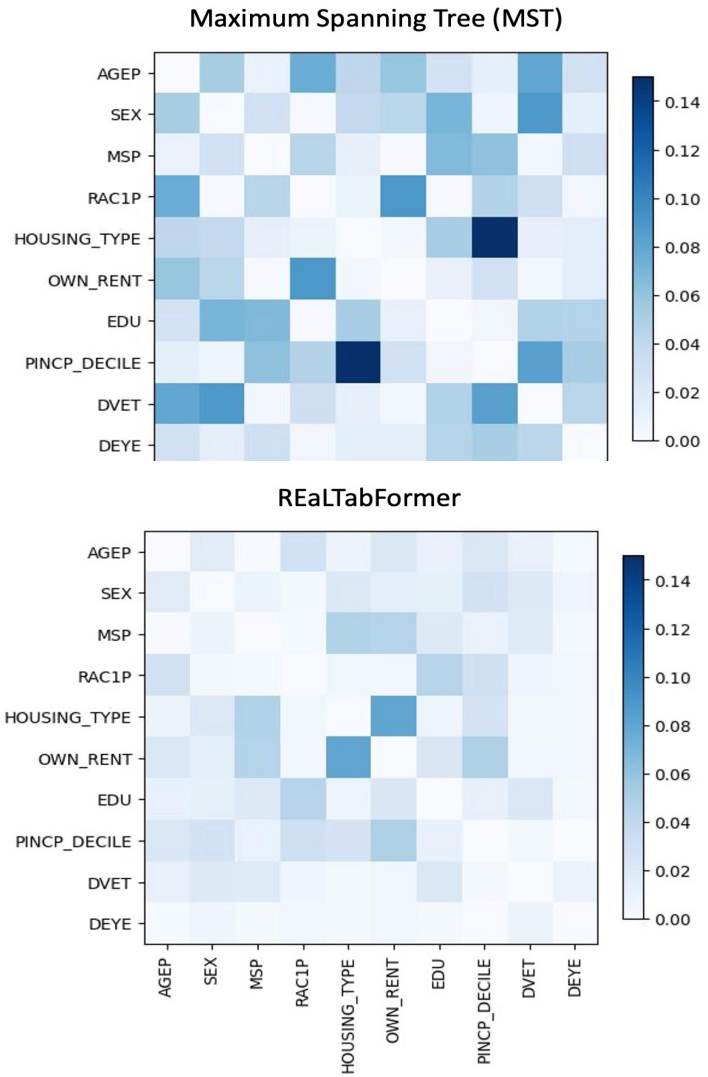
Synthetic data generation using the NIST Diverse Community Data Excerpts

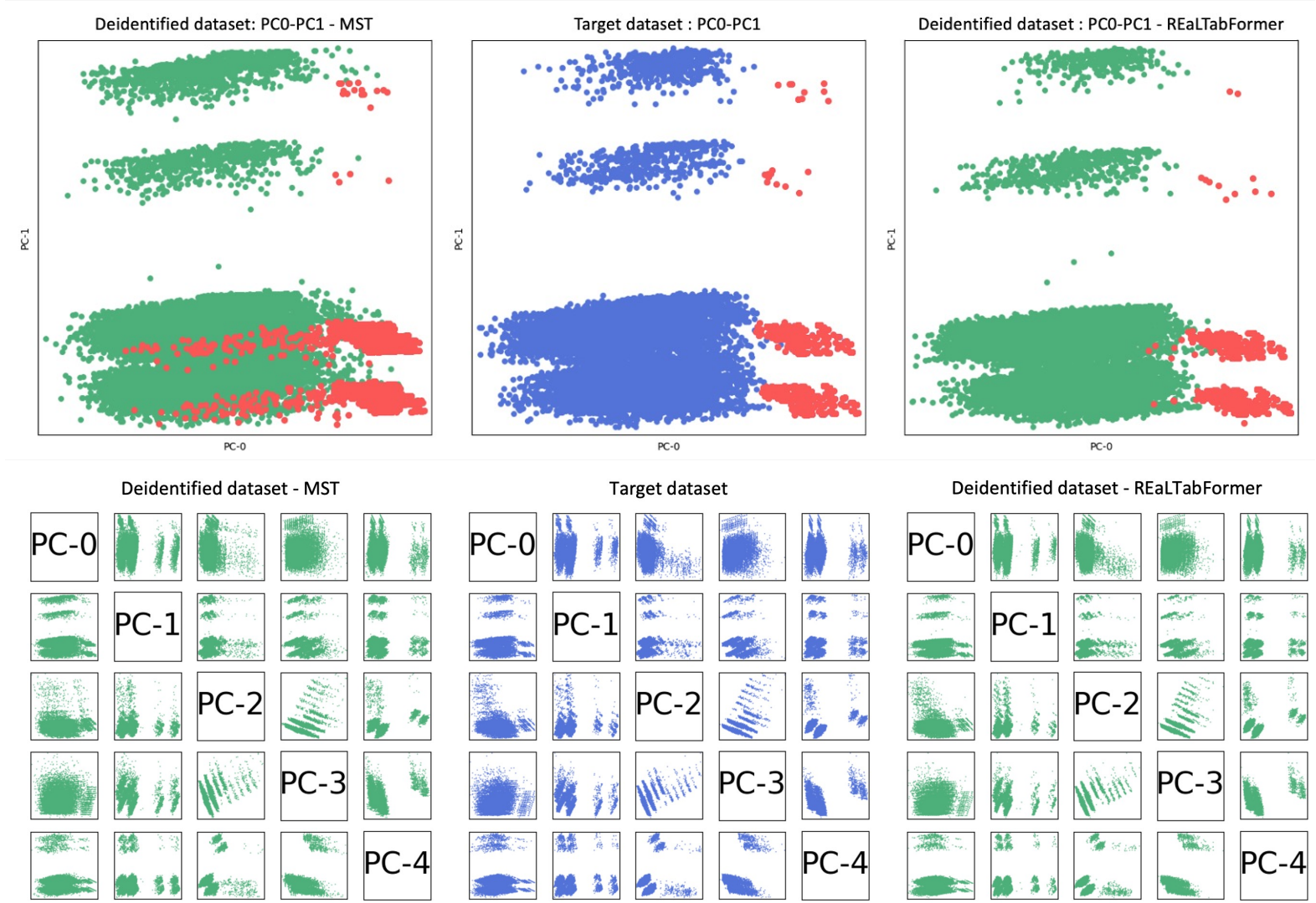
Application to NIST data and baseline model

- To demonstrate the synthetic tabular generation capability of the REaLTabFormer model, we trained it on the NIST Diverse Community Data Excerpts (national).
- We used the differentially-private model, based on the maximum spanning tree formulation, which won the 2018 NIST DP synthetic data generation competition as a baseline.
- The open-sourced SDNist package was used to evaluate the quality of both models in generating synthetic version of the original dataset.

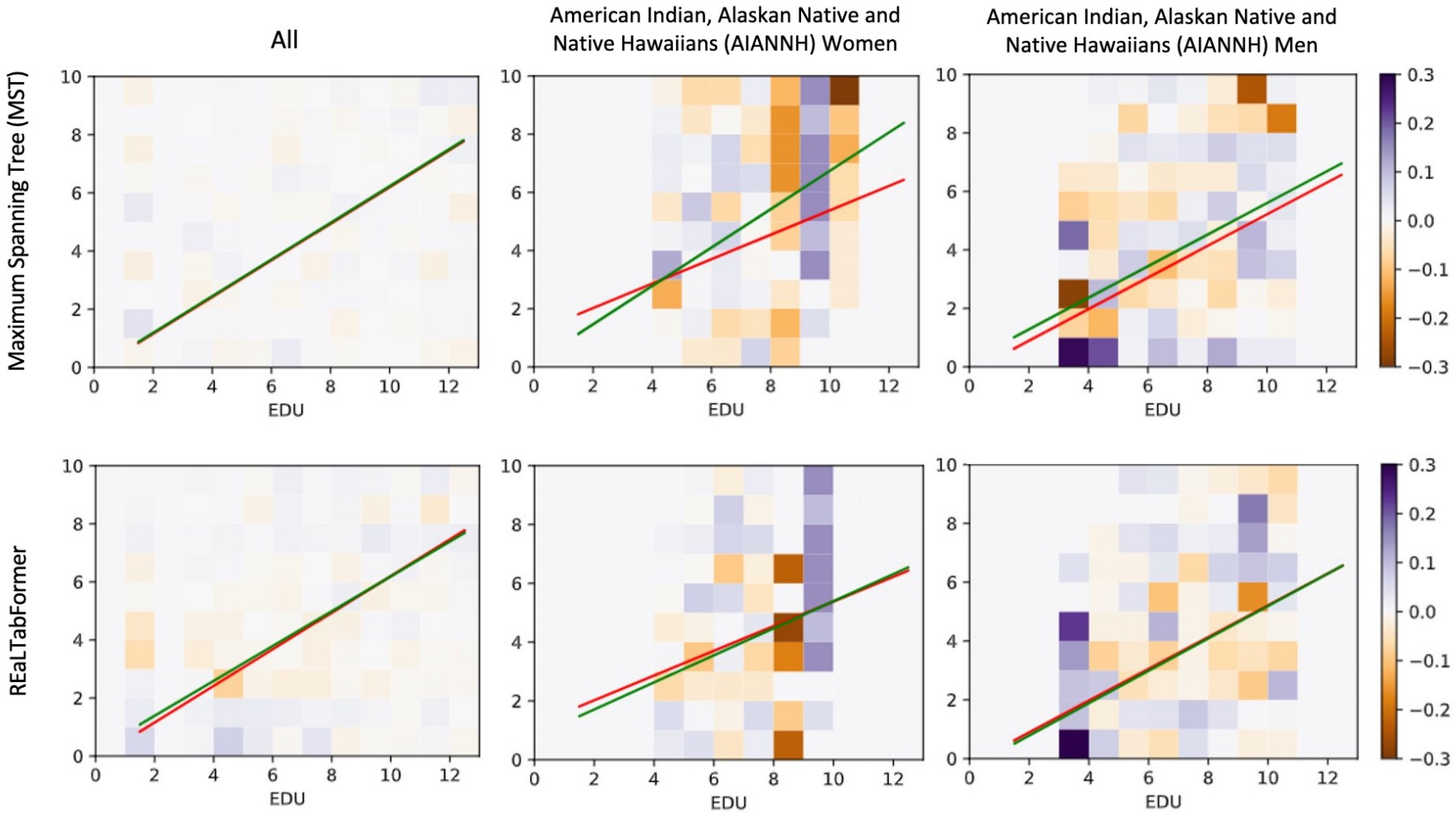
SDNist GitHub: https://github.com/usnistgov/SDNist
MST Docs: https://docs.smartnoise.org/synth/synthesizers/mst.html

Quality assessment using correlation difference and PCA





Synthetic data quality using regression model on income rank against education level



Summary of quantitative comparison from the SDNist Deidentified Data Report Generator between MST and the REaLTabFormer model

Table 1. Summary measures from the SDNist Deidentified Data Report Generator, MST and REaLTabFormer

	k-marginal score	Propensity mean square error	Number of inconsistencies	Unique target data records exactly matched in deidentified data*	Number of target data records exactly matched in deidentified data on quasi-identifiers**
MST	**966**	0.010	1017	**7.41%**	101 (0.37%)
ReaLTabFormer	957	**0.003**	**18**	9.48%	**90 (0.33%)**

* Refers to sample uniques (considering all variables) that are present in the synthetic data.

** RAC1P, OWN_RENT, MSP, SEX, EDU

Generating synthetic microdata for an imaginary country

Application of REaLTabFormer
on a collection of datasets

Large-scale synthetic data generation for synthesizing an imaginary country

- Our goal is to generate a realistic population representing an imaginary country. This dataset can then be used for various use cases, e.g., training for various statistical methods in census data, simulation, etc.
- We collected a diverse collection of IPUMS census data, the Demographic and Health Surveys (DHS), and national household expenditure surveys, from various countries so that the population cannot be mapped precisely to any one country.
- We harmonized the data as needed so the values are consistent across the collection.
- We used the dataset to train a customized/earlier version of the REaLTabFormer model to learn the statistics of the data, and realistically synthesize observations. The synthetic population consists of ~10 million individuals.
- The dataset and more information about it are publicly available from the World Bank Microdata Library.

Synthetic Dataset (English Documentation): https://doi.org/10.48529/78M1-AE09
Synthetic Dataset (French Documentation): https://doi.org/10.48529/X5BG-SD13

Synthetic superpopulation for estimating disclosure risk

Using REaLTabFormer to extrapolate a population from a sample microdata

Using synthetic superpopulation for risk estimation

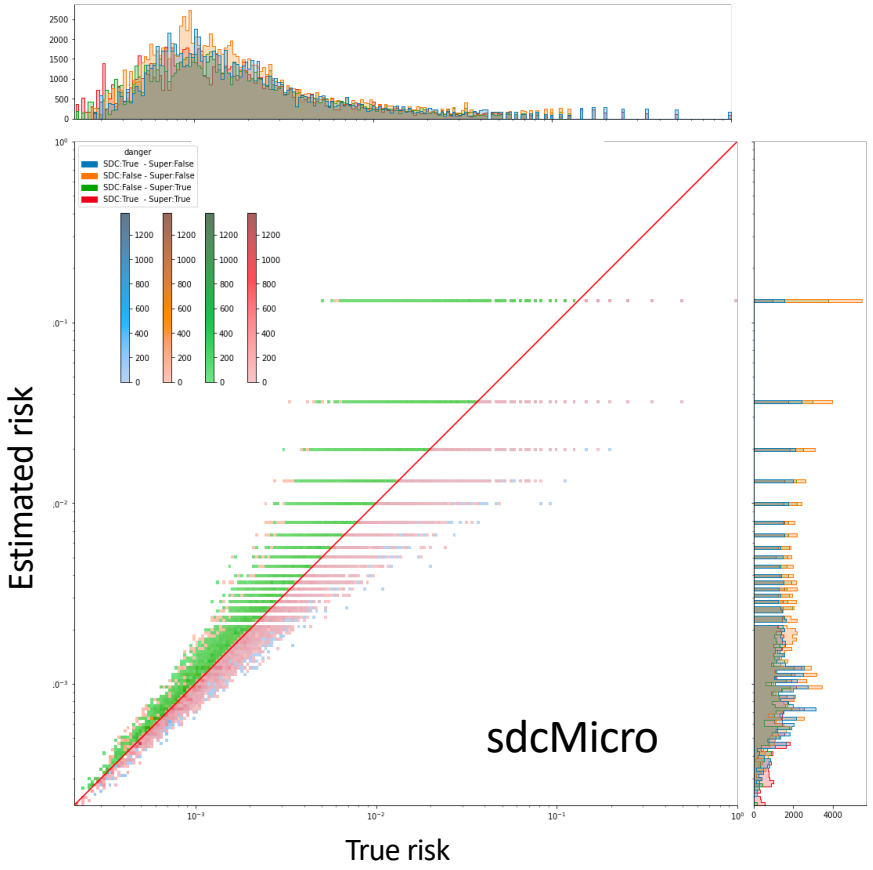
- Traditional methods for assessing disclosure risk leverage analytical Bayesian models. The models typically aim to infer the total population, commonly referred to as the superpopulation (F_k) as a posterior distribution estimated from the sample keys distribution (f_k).

- Here, we propose to generate a synthetic superpopulation, equivalent to the size of the true population, by training the REaLTabFormer on the sample data.

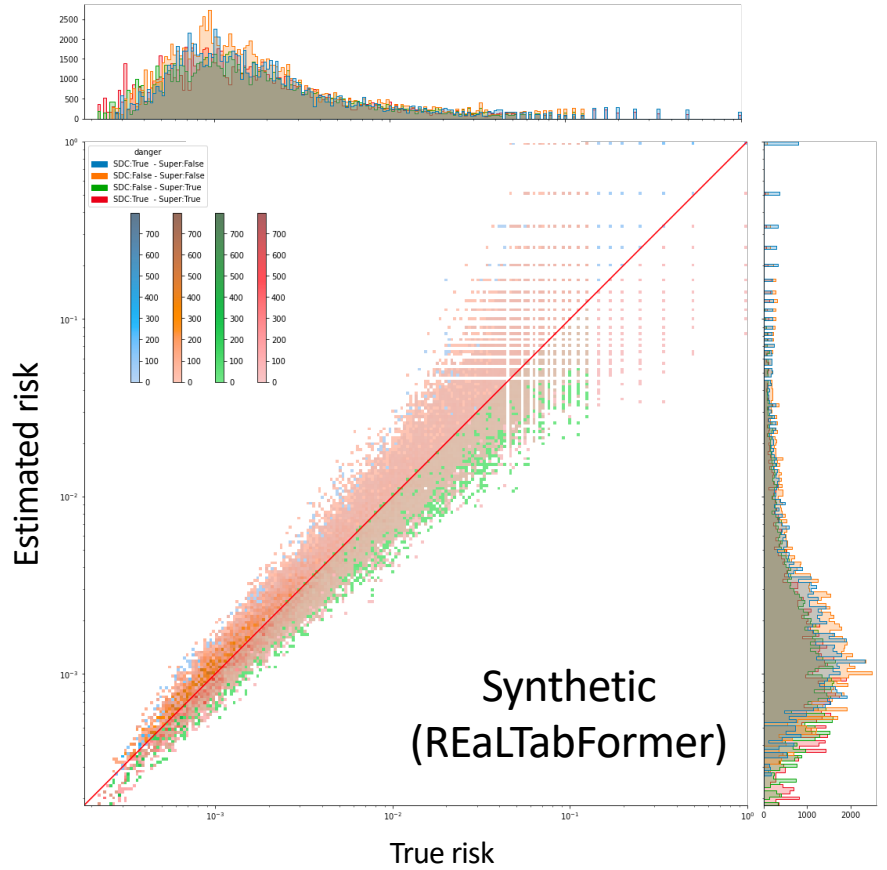
- We then use the resulting superpopulation to compute the “population” risk estimates given the relevant keys.

- We compare the estimates generated by the synthetic superpopulation method with the commonly used SDC risk computations based on the Bayesian formulation.

Comparison of predicted versus actual risk values

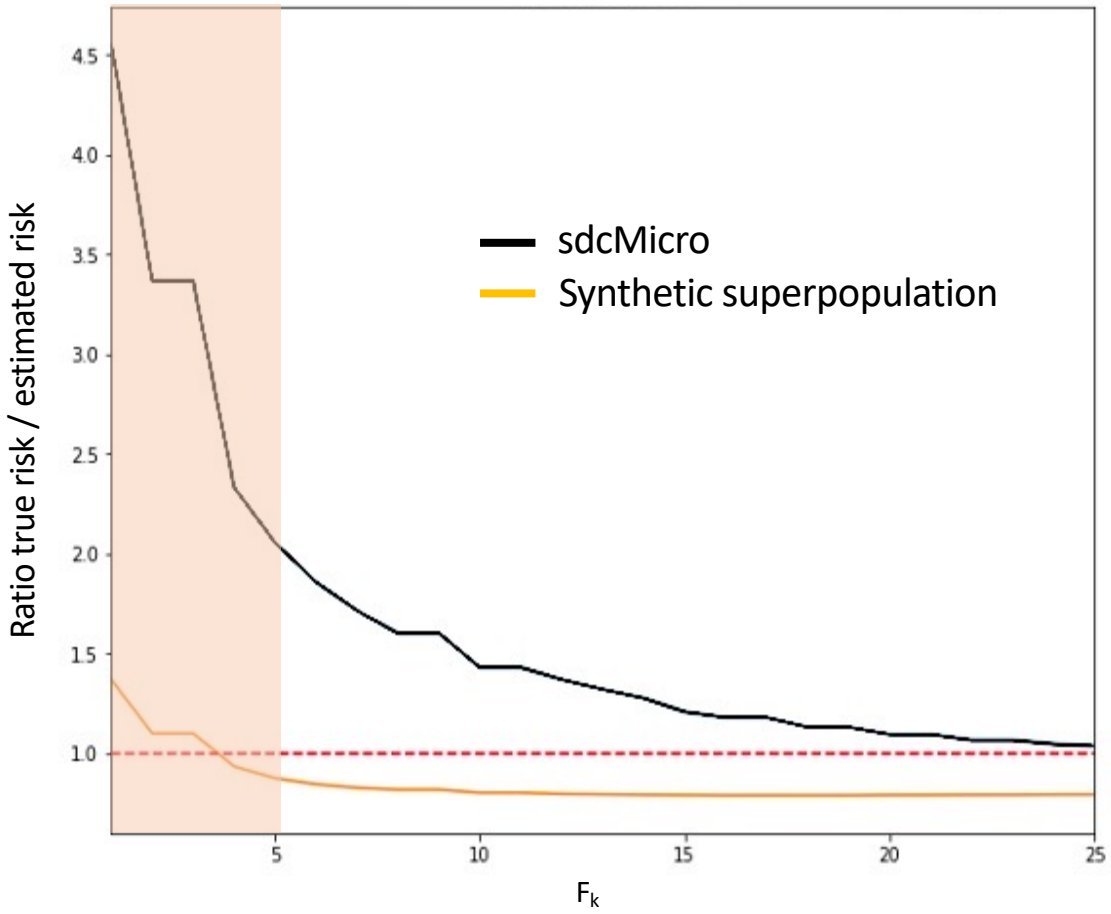


sdcMicro	Metrics	Synthetic
0.877	Accuracy (↑)	**0.965**
0.123	Error rate (↓)	**0.035**
0.119	False positive rate (↓)	**0.108**
0.838	Sensitivity (↑)	**0.973**
0.451	Precision (↑)	**0.987**
0.881	Specificity (↑)	**0.892**



* Risk assessed using six key variables and a sample of 100,000 households

The Bayesian method used in sdcMicro is generally more conservative in estimating risk for population-unique and riskier groups (F_k <= 5) than the estimates from the synthetic superpopulation method.



* Risk assessed using six key variables and a sample of 100,000 households

Ongoing and Future Work

Ongoing developments

- Application of REaLTabFormer for modeling and synthesizing microdata representing a real country.
- Development of a framework to improve the calibration of synthetic microdata for statistical use.
- Improvement of the REaLTabFormer model for efficiency and accuracy.

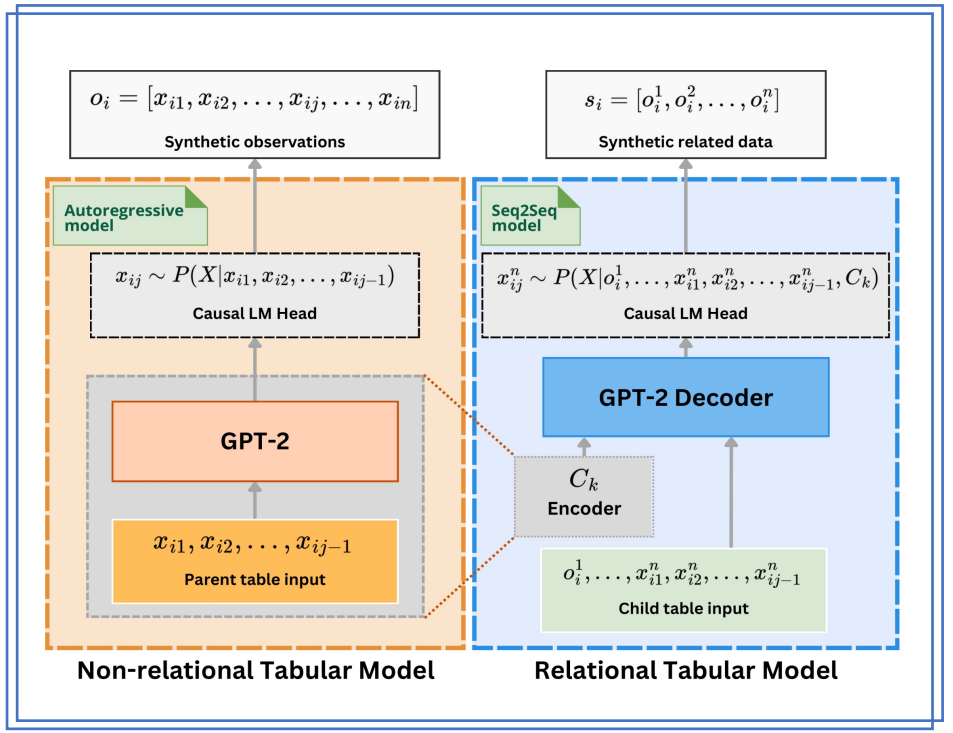
Future research and recommendations

- Look at methods to integrate differential privacy into the REaLTabFormer effectively.
- Investigate existing and or develop novel frameworks to provide robust guaranties for privacy in synthetic data.
- Find ways to make the superpopulation method more efficient so it can easily be integrated in existing risk estimation/assessment workflows.

Generating synthetic data using REaLTabFormer, and assessing the probabilistic measure of statistical disclosure risk

Olivier Dupriez | odupriez@worldbank.org
Deputy Chief Statistician (DECDG)

Aivin V. Solatorio | asolatorio@worldbank.org
Data Scientist (DECDG)
GitHub: @avsolatorio



What is Synthetic Data?

- Synthetic data is a set of data generated by some generative model.
- Ideally, the synthetic data is modeled based on empirical data.
- Generative models are trained (or fitted) to capture the statistical properties of the empirical data.
- Synthetic data are useful as a proxy for using the original data when certain constraints are imposed on the original dataset.
- Generative models can generate an arbitrary number of synthetic data, which can be useful for data augmentation and simulations.