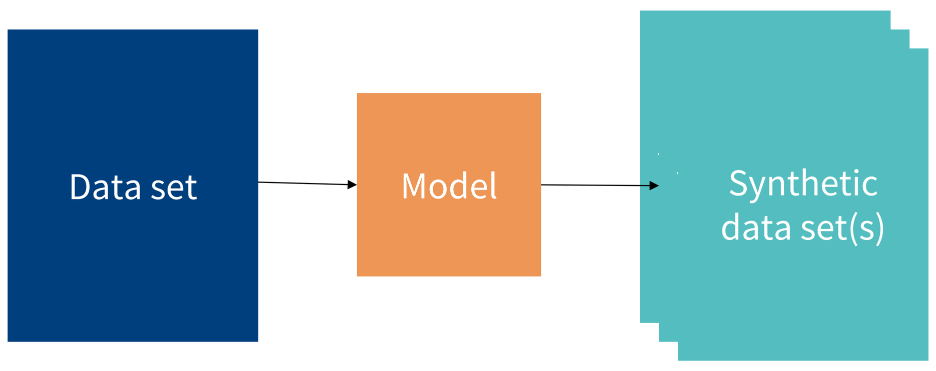
It‘s complicated — Insights into models and privacy guarantees of synthetic data.

Prof. Dr. Jörg Drechsler, Dr. Jonathan Latner & Dr. Marcel Neunhoeffer, E-Mail: marcel.neunhoeffer@iab.de

What are synthetic data?



Two perspectives on synthetic data:

1. Generating synthetic data is easy.

More and more commercial vendors and software packages to generate synthetic data on the “market“. For example:

- Gretel.ai: *“The synthetic data platform for developers. Generate artificial datasets with the same characteristics as real data, so you can develop and test AI models without compromising privacy.”*
- Mostly.ai: *“Synthetic Data. Better than real. Still struggling with real data? Use existing data for synthetic data generation. Synthetic data is more accessible, more flexible, and simply...smarter.”*

2. Generating synthetic data is hard.

- See, e.g., the report of the Alan Turing Institute (Jordan et al. 2022): *“Significant care is required to produce synthetic data that is useful and comes with privacy guarantees.”*
- There is a trade-off between utility/quality of synthetic data and privacy guarantees.
  - From a statistical perspective, synthetic data cannot be used to protect privacy and as an equivalent substitute for original data.

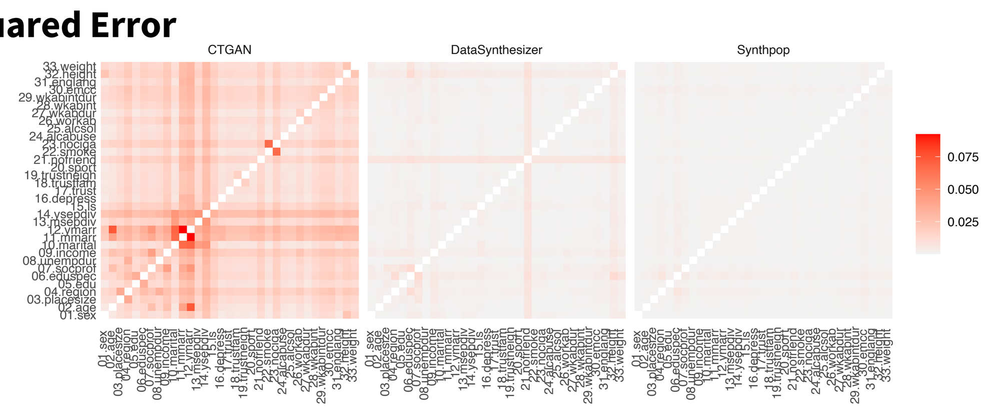
It’s complicated...

- **One** real data set
  - We use real data with all of its quirks for our evaluation.
  - Here the Social Diagnosis 2011 data set.

DATA (SD2011)

Number	Variable	Description	Type	Observations	Unique.Values	Missings	Negative.values	Generated	Quirks
1	sex	Sex	factor	5000	2	0	0		
2	age	Age of person, 2011	numeric	5000	79	0	0		
3	agegr	Age group, 2011	factor	5000	7	4	0	Yes	Yes
				...					
7	eduspec	Discipline of completed qualification	factor	5000	28	20	0		Yes
				...					
10	income	Personal monthly net income	numeric	5000	407	683	603		
11	marital	Marital status	factor	5000	7	9	0		
12	mmarr	Month of marriage	numeric	5000	13	1350	0		
13	ymarr	Year of marriage	numeric	5000	75	1320	0		
14	msepdiv	Month of separation/divorce	numeric	5000	13	4300	0		
15	ysepdiv	Year of separation/divorce	numeric	5000	51	4275	0		
				...					
22	nofriend	Number of friends	numeric	5000	44	0	41		
23	smoke	Smoking cigarettes	factor	5000	3	10	0		
24	nociga	Number of cigarettes smoked per day	numeric	5000	30	0	3737		Yes
				...					
27	workab	Working abroad in 2007-2011	factor	5000	3	438	0		
28	wkabdur	Total time spent on working abroad	numeric	5000	33	0	4875		Yes
				...					
33	height	Height of person	numeric	5000	65	35	0		
34	weight	Weight of person	numeric	5000	91	53	0		
35	bmi	Body mass index (weight - kg/(height - cm^2)*10000)	numeric	5000	1396	59	0	Yes	Yes

- **Three** different synthetic data generators
  - DataSynthesizer (Ping, Stoyanovich & Howe 2017)
  - CTGAN (Xu, Skoularidou, Cuesta-Infante & Veeramachaneni 2019)
  - Synthpop (Nowok, Raab & Dibben 2016)
- **Four** quality measures
  - **Propensity Score Mean Squared Error (pMSE)**
  - Ratio of counts (ROC)
  - Confidence Interval Overlap
  - Computational Efficiency



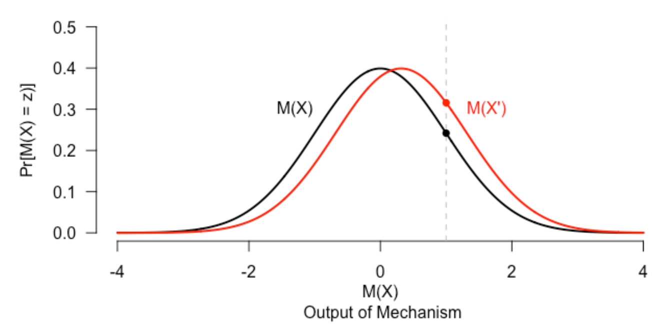
Our conclusion: Generating synthetic data is easier than ever. **But**, to generate good synthetic data, scientists need to have a detailed understanding how and for what purpose the synthetic data are generated.

Can *traditional* synthetic data generators satisfy formal privacy guarantees?

- Traditionally, synthetic data are generated to protect privacy.
- **However**, models to generate synthetic data **do not** automatically satisfy formal privacy guarantees.
- There is relatively little research into whether and how traditional statistical disclosure control (SDL) methods satisfy formal privacy guarantees.
  - First results on SDL methods like swapping show that traditional models can satisfy differential privacy (see Bailie, Gong & Meng 2023).
  - We want to analyze if and under what conditions parametric synthetic data generators can satisfy differential privacy guarantees.

What is Differential Privacy?

- Differential Privacy (Dwork, McSherry, Nissim & Smith 2006) is a formal privacy guarantee.
- “A randomized algorithm M is differentially private if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$, the random variables $M(X)$ and $M(X')$ are similarly distributed.”



First Results

We start with a very simple case: Generating synthetic data from a sample from a univariate normal distribution with known variance using a normal synthetic data generator.

In this case, we can show that releasing synthetic data is equivalent to noise addition with a ρ-zCDP (Bun & Steinke 2016) guarantee (i.e., noise addition with a Gaussian mechanism).

Noise Addition	**Synthetic Data**
$\tilde{x} \sim \bar{x} + \mathcal{N}(0, \frac{\Delta^2}{2\rho})$	$\tilde{x} \sim \bar{x} + \mathcal{N}(0, \frac{\sigma^2}{n_{synth}})$

This means, that the number of synthetic samples n_{synth} can be set such that the synthetic data satisfies a ρ-zCDP guarantee by:

$$n_{synth} = \lfloor \frac{2\rho\sigma^2}{\Delta^2} \rfloor$$

Two observations:

1. To satisfy formal privacy guarantees, the sensitivity (Δ^2) of the sufficient statistics must be bounded.
2. The number of synthetic samples is the privacy parameter. Releasing multiple/more synthetic data samples weakens the privacy guarantee.

Next steps: Moving to the case with unknown variance makes things more complicated (but can be done).

Our conclusion: Under certain conditions, traditional synthetic data generators (here: parametric models) can satisfy differential privacy guarantees.

AnigeD

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Finanziert von der Europäischen Union
NextGenerationEU