Really Useful Synthetic Data

A Framework to Evaluate the Quality of Differentially Private Synthetic Data

Christian Arnold (Cardiff University)
Marcel Neunhoeffer (University of Mannheim)

June 9th, 2020

Vision

So You Want to Share Sensitive Data...

Use cases, e.g.

- Science: Reproduce studies
- Government: Accountability
- Business: Externalise services

How Do We Currently Solve this Dilemma?

Don’t publish data.

Publish results without the data themselves.

Sign contracts to work with such data.

Remove ID variables and share anonymized data.

Privacy preserving synthetic data as a promising idea.

Synthetic data for sharing statistical information

Synthetic data for sharing statistical information
Principled data protection with differential privacy

Synthetic data for sharing statistical information

Principled data protection with differential privacy

$\Rightarrow$ *Differentially private synthetic data* to protect privacy with principled guarantees while maximising data utility

Core Challenges

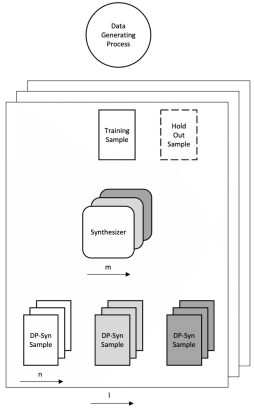
DP Data Synthesis is a Multi Phase Inference Problem

Inference usually

DP Data Synthesis is a Multi Phase Inference Problem

Inference usually

Inference with DP synthetic data



Typical Data Challenges Applied Data Analysts Are Facing

Continuous data
Real values.

Discrete data
Integers.

Structural zeros
Non-logical combinations of data (e.g. pregnant men).

Missing data
Missing data entries. Leads to biased inference if not handled appropriately.

Nested data
Observations are nested in others. E.g. individuals in groups or observations over time. In effect, observations are not iid.

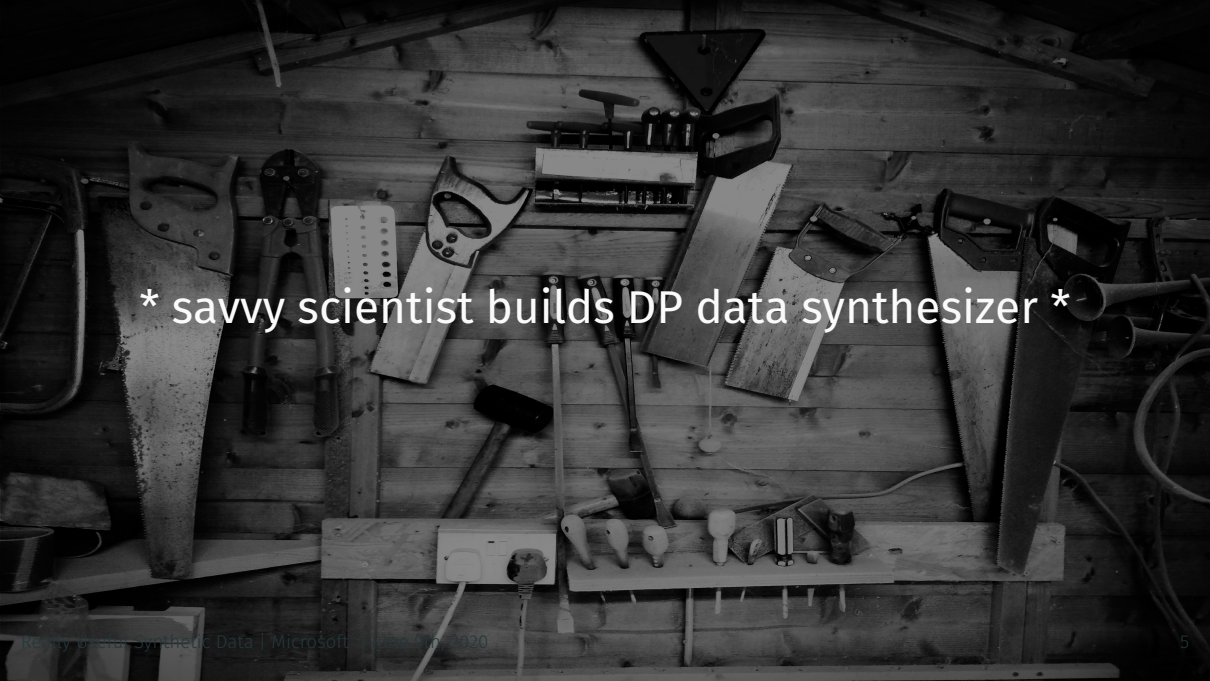
DP Data Synthesizers Ideally Perform Well in Different Settings

DP privacy guarantees

Accepted ϵ values range from $[0.01, 5]$

Sample sizes

Typical sample sizes range from $[500, 100'000]$

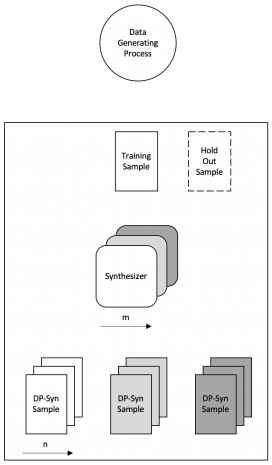

* savvy scientist builds DP data synthesizer *

Now That We Have DP Synthetic Data, How Can We Know It Is Useful?

A General Benchmark for Differentially Private Data Synthesizers

	General Utility	Specific Utility
Training Data Similarity	Marginals: Wasserstein randomisation test	Regression coefficients: Average percent bias
	Joint distribution: *pMSE*	Variances: Average variance and covariance ratio
Generalisation Similarity	Marginals: Wasserstein randomisation test	Regression coefficients: Average percent bias
	Joint distribution: *pMSE*	Variances: Coverage
		Prediction *RMSE* w.r.t. further samples from population

Training Data Similarity



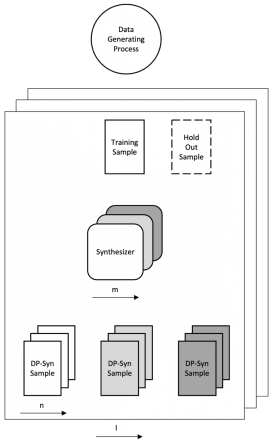
General utility

- pMSE for joint distribution
- Wasserstein randomisation test for marginals

Specific utility

- Average percent bias of regression coefficients
- Average variance and covariance ratio of Variances

Generalisation Similarity



General utility

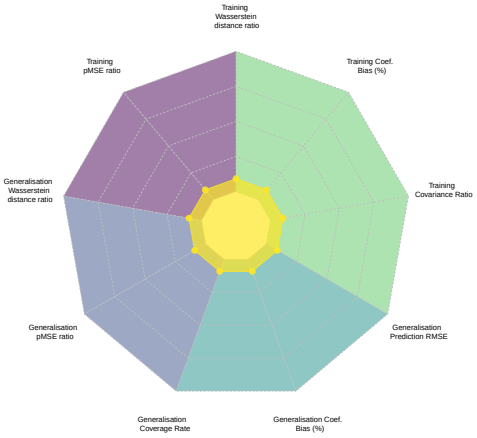
- pMSE for joint distribution
- Wasserstein randomisation test for marginals

Specific utility

- Average percent bias of regression coefficients
- Coverage for Variances
- Prediction RMSE w.r.t. further samples from population

An Application

A Visualization for All Utility Measures



Outlook

What We Would Like to Do Next

Utility measures

- A utility challenge for submitting DP data synthesizers to OpenDP?
- Include further application specific utility measures.

Statistical properties of DP data synthesizers

- What are the statistical properties of DP GANs (or other synthesizers)?
- What do we know about statistical validity of DP data synthesis?

Dr Christian Arnold
Cardiff University
arnoldc6@cardiff.ac.uk
@chrisguarnold

Marcel Neunhoeffer
University of Mannheim
mneunhoe@mail.uni-mannheim.de
@mneunho