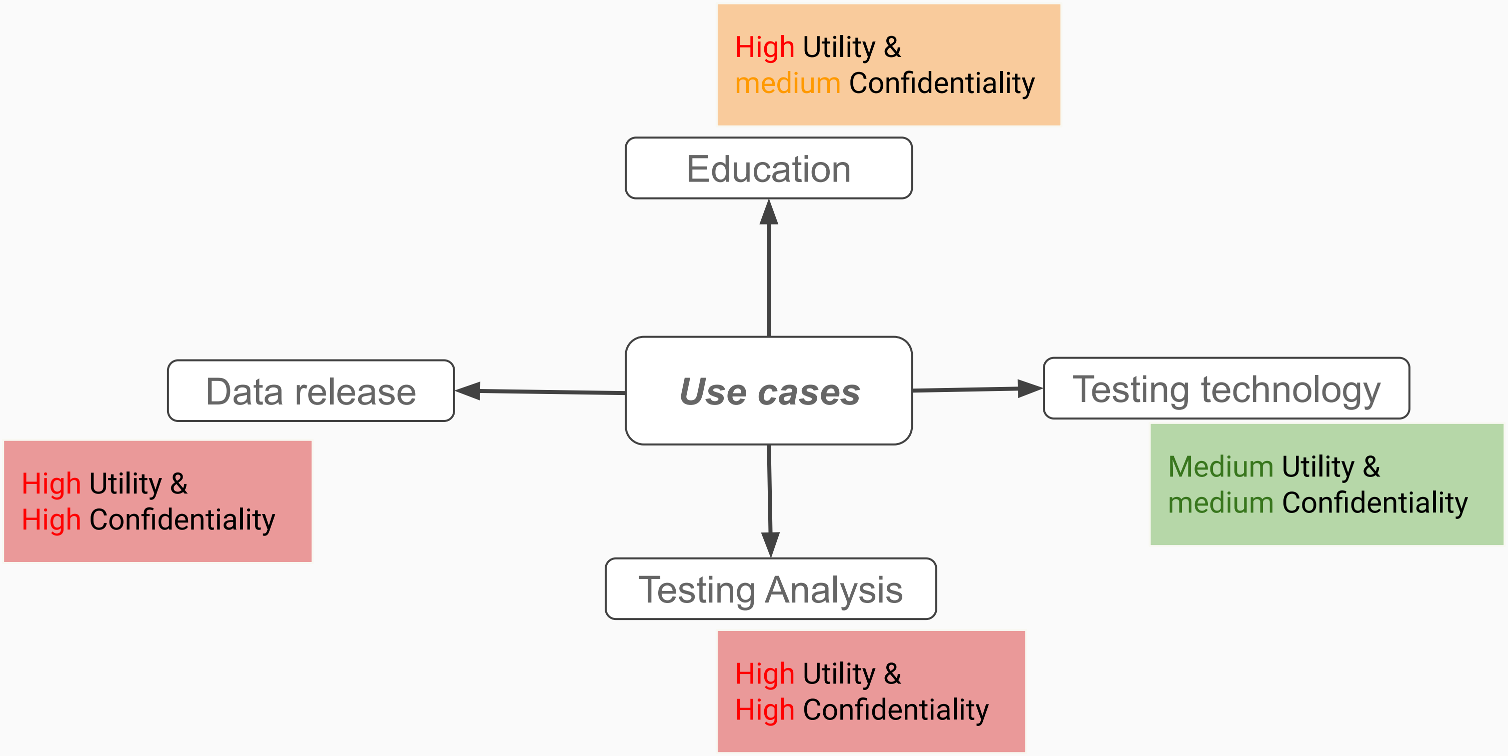
Statistics Netherlands Team

ACS Data set

Outline

- Understanding the data set
- Partially synthetic data generation
  - Analytical Validity
  - Privacy evaluation
  - Utility evaluation
- Fully synthetic data generation
  - Analytical Validity
  - Privacy evaluation
  - Utility evaluation
- Judging the guideline

Synthetic Data Use Cases



High Utility &
medium Confidentiality

Education

Use cases

Data release

High Utility &
High Confidentiality

Testing technology

Medium Utility &
medium Confidentiality

Testing Analysis

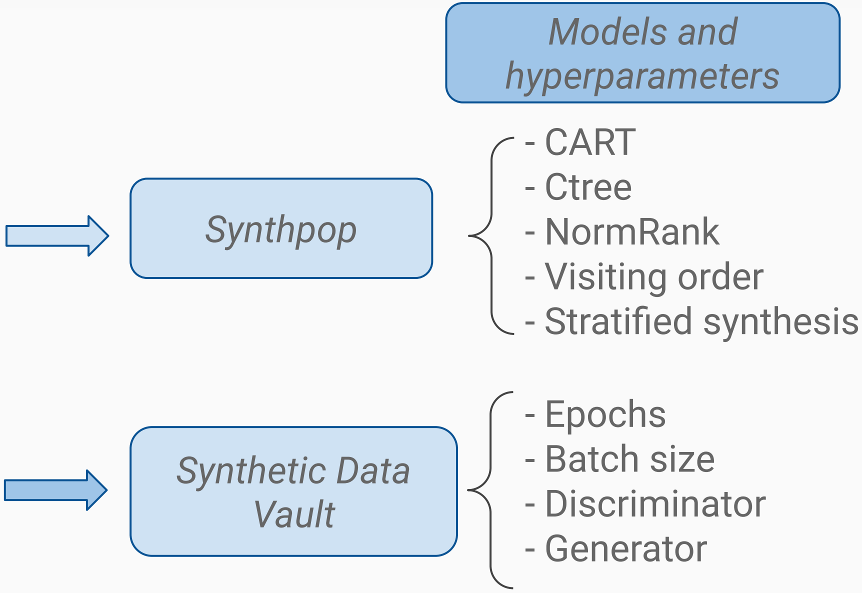
High Utility &
High Confidentiality

Data Synthesis Techniques

- Partially vs Fully synthetic data
- Sequential modeling
  - Fully conditional specification method (FCS)

- Deep Learning
  - Generative adversarial networks
  - Variational AutoEncoder
  - Gaussian Copulas

- Random data



Models and hyperparameters

Synthpop

- CART
- Ctree
- NormRank
- Visiting order
- Stratified synthesis

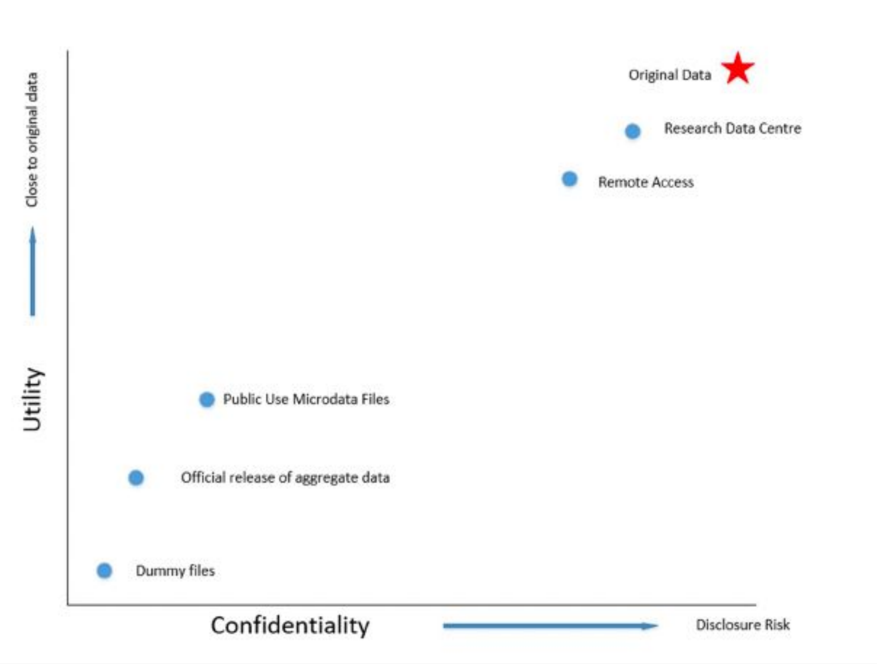
Synthetic Data Vault

- Epochs
- Batch size
- Discriminator
- Generator

Evaluation of synthetic data

How good is the generated data?

How much good is good?



***Credits**: Figure from chapter Introduction of the starter guide.*

EDA~ ACS data

- Large data takes time for synthesize
  - Maybe divide and conquer!
  - Clustering
- Attributes: DEPARTS and ARRIVES
  - ARRIVES > DEPARTS
  - Possibly we need to convert it to minutes!
- Impose constraints
- Quasi identifiable attributes
  - Maybe: sex, age, marst, race, hispan, educ, citizen
- Sensitive attributes:
  - GQ, HCOVANY, HCOVPRIV, HINSEMP, HINSCAID, HINSCARE, EMPSTAT, EMPSTATD, LABFORCE, WRKLSTWK, ABSENT, LOOKING, AVAILBLE, WRKRECAL, WORKEDYR

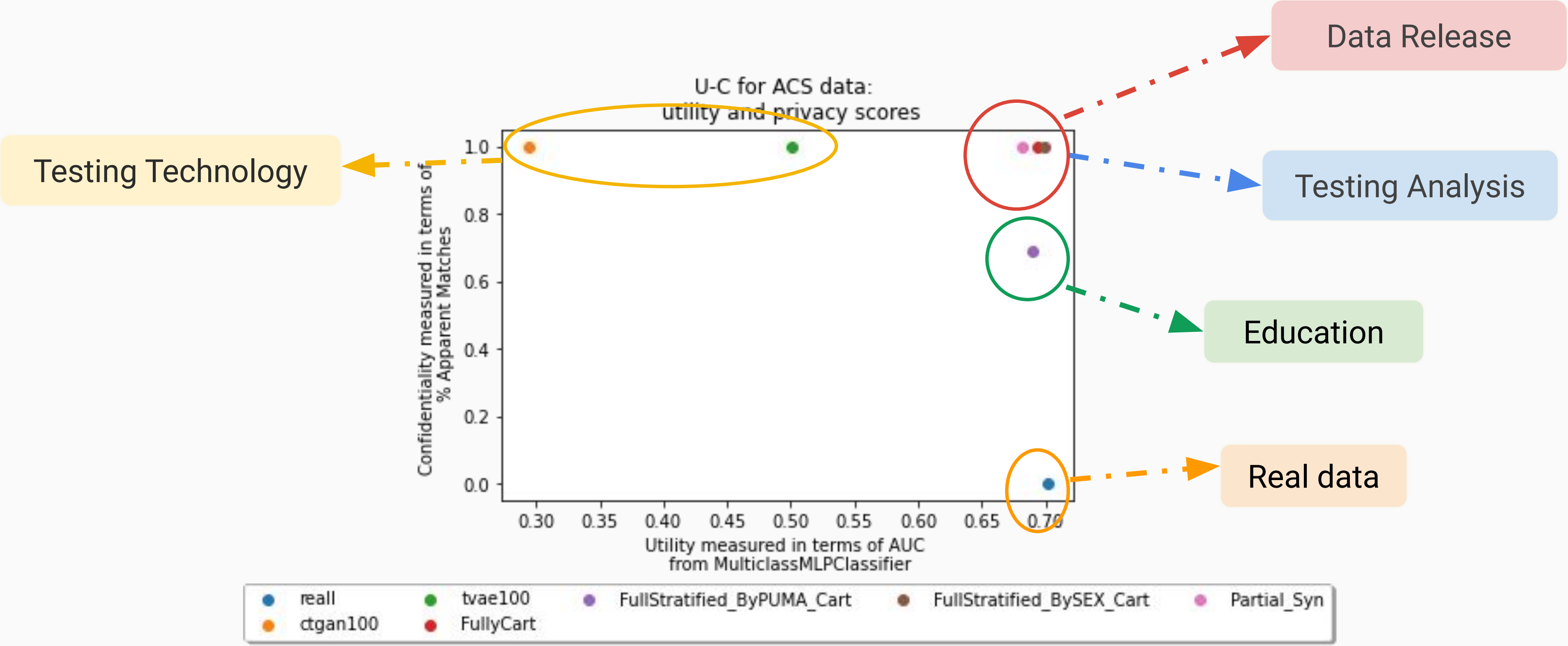
EDA ~ ACS data

Note: Red== sensitive, green= non sensitive, blue== quasi-identifiable

The columns are as follows:

- PUMA (str)
- YEAR (uint32)
- HHWT (float)
- GQ (uint8)
- PERWT (float)
- SEX (uint8) AGE (uint8) MARST (uint8) RACE (uint8) HISPAN (uint8) CITIZEN (uint8) EDUC (uint8)
- SPEAKENG (uint8)
- HCOVANY, HCOVPRIV, HINSEMP, HINSCAID, HINSCARE (uint8)
- EMPSTAT, EMPSTATD, LABFORCE, WRKLSTWK, ABSENT, LOOKING, AVAILBLE, WRKRECAL, WORKEDYR (uint8)
- INCTOT, INCWAGE, INCWELFR, INCINVST, INCEARN (int32)
- POVERTY (uint32)
- DEPARTS, ARRIVES (uint32)

Utility vs. Confidentiality on ACS real vs synthetic data sets



Testing Technology

Data Release

Testing Analysis

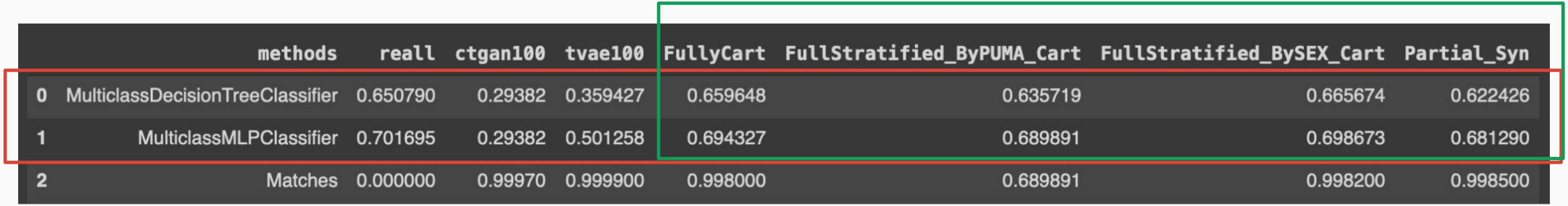
Education

Real data

The target class for the classifier is “SEX” !

Utility Evaluation

Synthpop based synthesize techniques has a high utility compared to **SDV** based techniques.



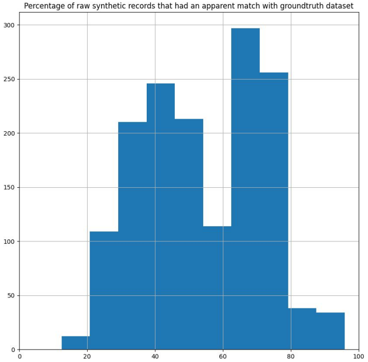
	methods	reall	ctgan100	tvae100	FullyCart	FullStratified_ByPUMA_Cart	FullStratified_BySEX_Cart	Partial_Syn
0	MulticlassDecisionTreeClassifier	0.650790	0.29382	0.359427	0.659648	0.635719	0.665674	0.622426
1	MulticlassMLPClassifier	0.701695	0.29382	0.501258	0.694327	0.689891	0.698673	0.681290
2	Matches	0.000000	0.99970	0.999900	0.998000	0.689891	0.998200	0.998500

The target class for the classifier is “SEX” !

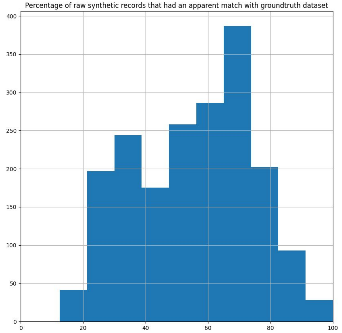
Privacy Evaluation

- Apparent Matches
- Replications

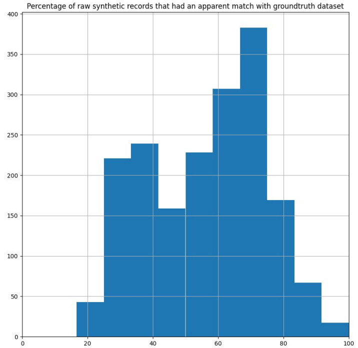
Partially_Stratified_**ByPUMA** **_Ctree**_QuasySensitive



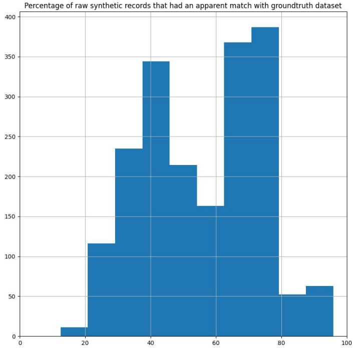
FullStratified_BySEX_**Cart**



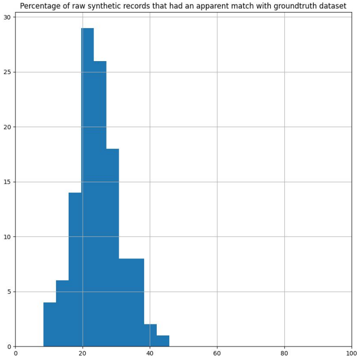
Fully_Stratified_ByPUMA_ **CART**



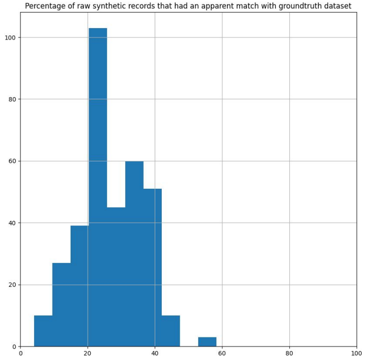
Fully**Cart**



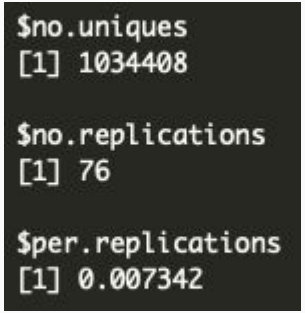
tvae_epoch100



ctgan_epoch100



Partially_Stratified_**ByPUMA**
_**Ctree**_QuasySensitive

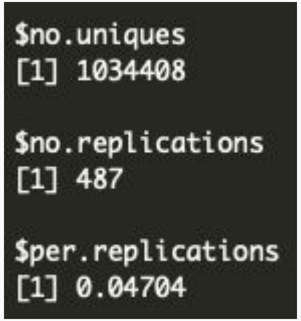


```
$no.uniques
[1] 1034408

$no.replications
[1] 76

$per.replications
[1] 0.007342
```

FullStratified_BySEX_**Cart**

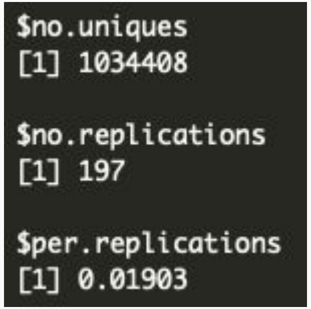


```
$no.uniques
[1] 1034408

$no.replications
[1] 487

$per.replications
[1] 0.04704
```

Fully_Stratified_ByPUMA_
CART

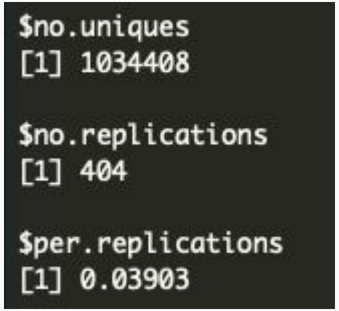


```
$no.uniques
[1] 1034408

$no.replications
[1] 197

$per.replications
[1] 0.01903
```

Fully**Cart**

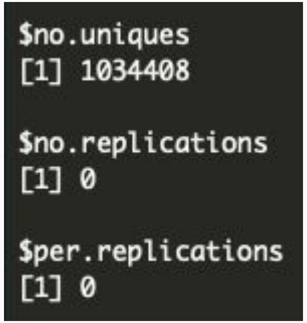


```
$no.uniques
[1] 1034408

$no.replications
[1] 404

$per.replications
[1] 0.03903
```

tvae_epoch100

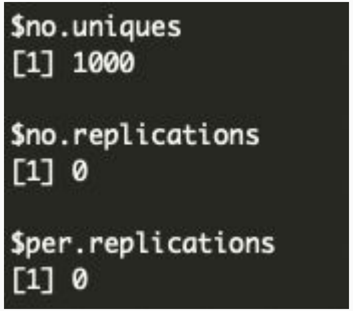


```
$no.uniques
[1] 1034408

$no.replications
[1] 0

$per.replications
[1] 0
```

ctgan_epoch100

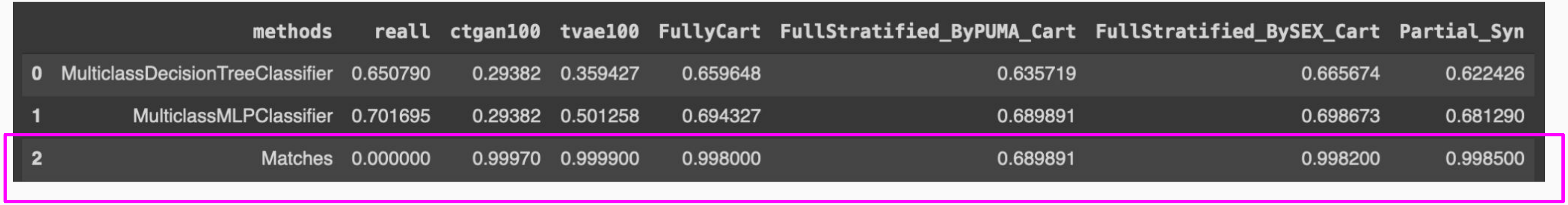


```
$no.uniques
[1] 1000

$no.replications
[1] 0

$per.replications
[1] 0
```

Privacy Evaluation



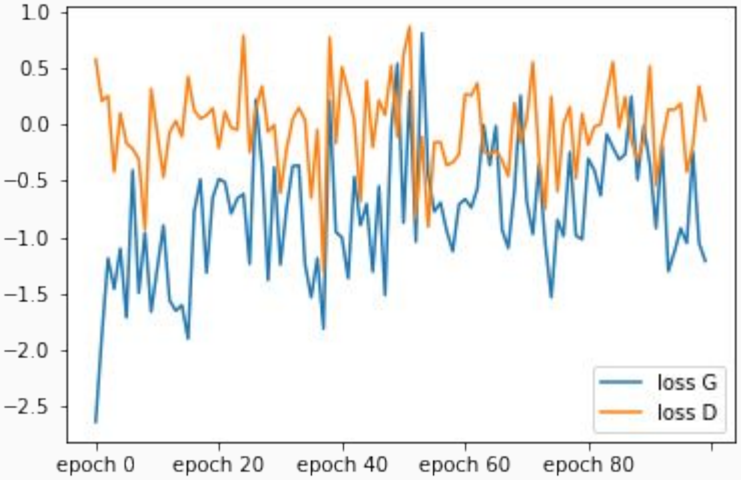
	methods	reall	ctgan100	tvae100	FullyCart	FullStratified_ByPUMA_Cart	FullStratified_BySEX_Cart	Partial_Syn
0	MulticlassDecisionTreeClassifier	0.650790	0.29382	0.359427	0.659648	0.635719	0.665674	0.622426
1	MulticlassMLPClassifier	0.701695	0.29382	0.501258	0.694327	0.689891	0.698673	0.681290
2	Matches	0.000000	0.99970	0.999900	0.998000	0.689891	0.998200	0.998500

Note that “Matches” refer to one minus the apparent matches metric. A high score means a better privacy protection.

Convergence of CTGAN

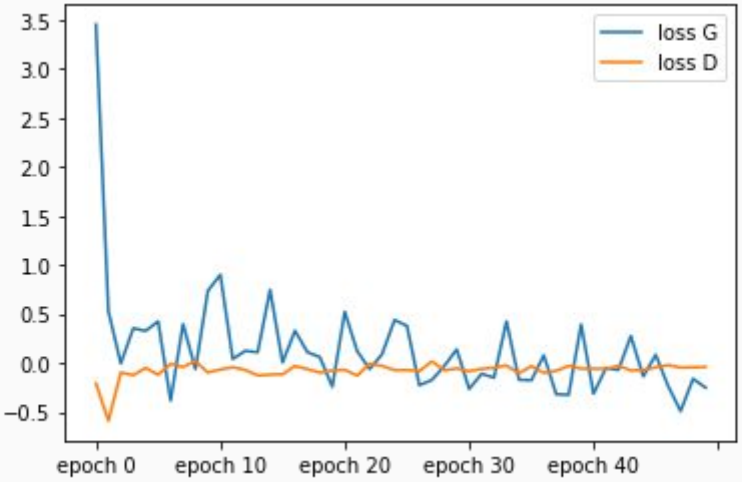
Tuning parameters of CTGAN

Training CTGAN on ACS, not even close to convergence after **100 epochs**

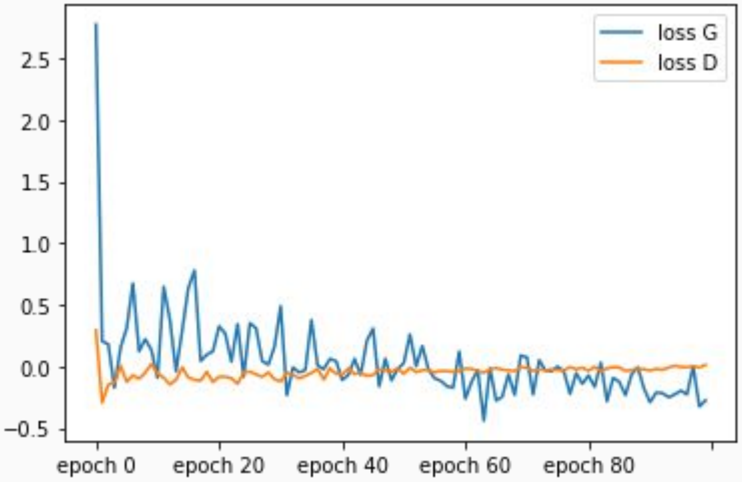


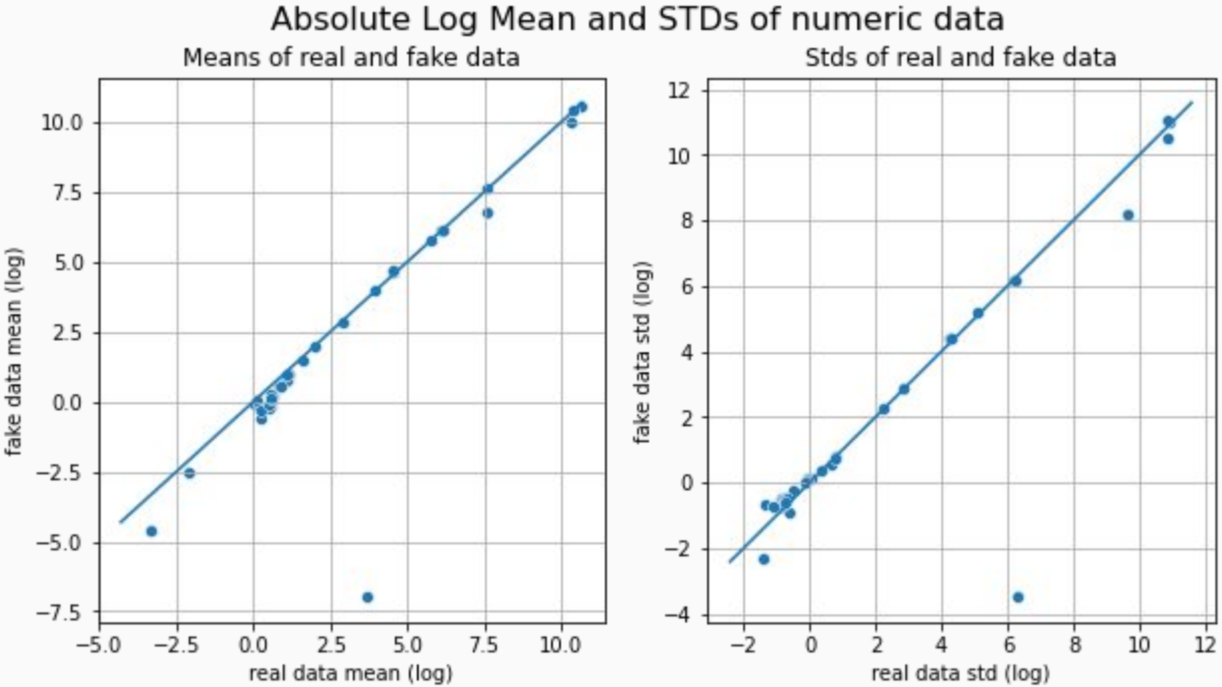
Advanced Tuning parameters of CTGAN

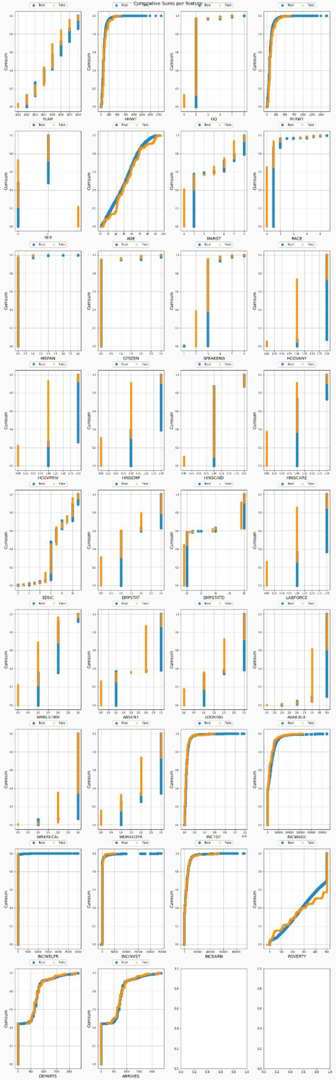
Training CTGAN on ACS,
after 100 epochs



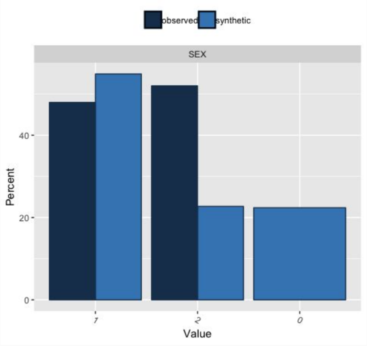
Training CTGAN on ACS,
after 100 epochs

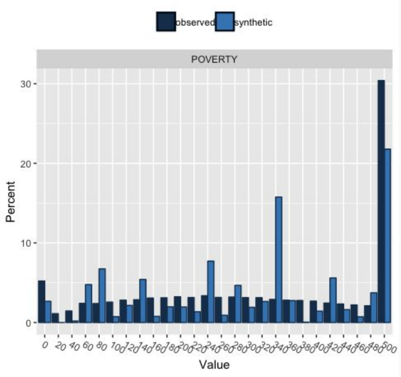


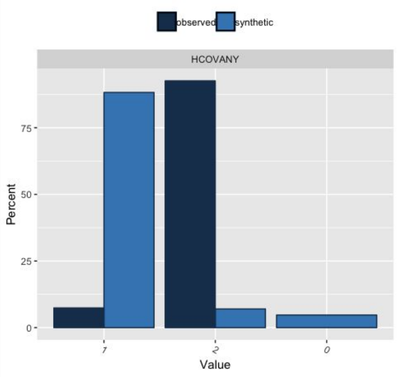

Absolute Log Mean and STDs of numeric data
Means of real and fake data
fake data mean (log)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
real data mean (log)
Stds of real and fake data
fake data std (log)
12
10
8
6
4
2
0
−2
−4
−2
0
2
4
6
8
10
12
real data std (log)

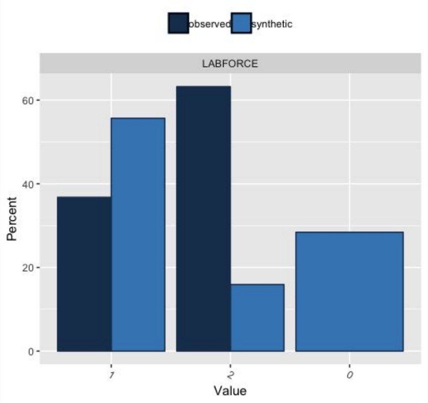

Cumulative Sums per feature
Real
Fake
Cumsum
YEAR
HHWT
GQ
PERWT
SEX
AGE
MARST
RACE
HISPAN
CITIZEN
SPEAKENG
HCOVANY
HCOVPRIV
HINSEMP
HINSCAID
HINSCARE
EDUC
EMPSTAT
EMPSTATD
LABFORCE
WRKLSTWK
ABSENT
LOOKING
AVAILBLE
WRKRECAL
WORKEDYR
INCTOT
INCWAGE
INCWELFR
INCINVST
INCEARN
POVERTY
DEPARTS
ARRIVES

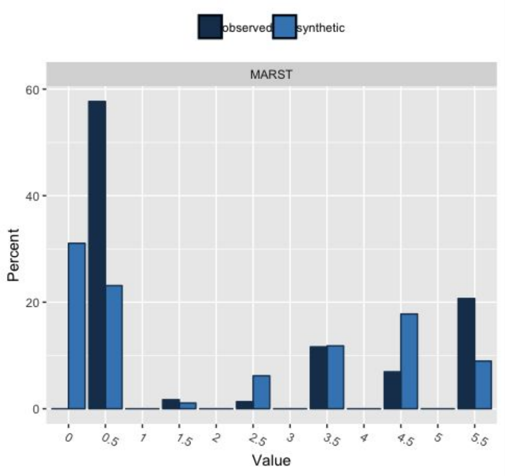
CTGAN with 20 epochs

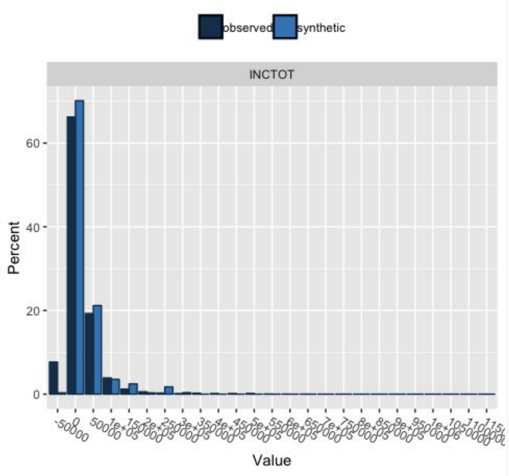


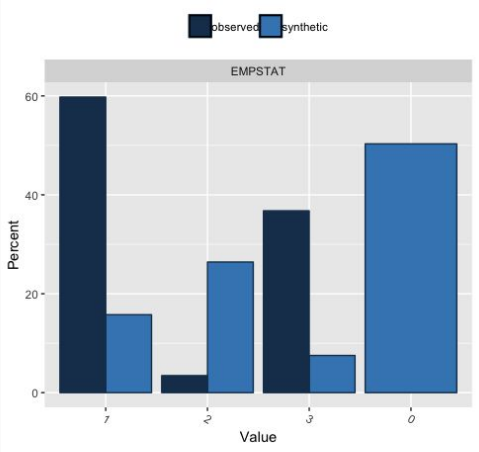












For the presentation:

- Which methods we used and why
- Comparison of results using different methods
  - Utility
  - Privacy
- Interpretation of the results:
  - When is the utility ‘ good enough’ for the different use cases
  - When is the privacy ‘ good enough’ for the different use cases
- What would we recommend to management
- Evaluation of the guidebook
  - What worked well
  - What is missing or needs improving

Observations (1)

- Computing power: despite the fact that we had quite a lot of computing power, it took time to generate the synthetic data for the ACS dataset. We sometimes ran out of memory and had to start again. Especially at the beginning, when we were tweaking parameters and re-running the models this required a lot of time.
- Deciding what is sensitive and non-sensitive: the decision which variables to mark as ‘ sensitive’, non-sensitive and quasi identifiers is subjective. The decision depends on cultural and legal context and prior knowledge of the subject matter. This means two organisations/people who generate a synthetic dataset using the same methods can generate two different results. Combining these results could potentially make it possible to reconstruct the original data. The guide could provide more guidance on this.

Observations (2)

- Utility-privacy for different usecases: (see chapter 2 of guidebook) NSO's can make policy decisions on what types of synthetic data to release (high/low utility). For example for releasing synthetic microdata to the public an NSO could choose to only disseminate lower utility microdata in order to safeguard privacy., especially if the usecase is unknown.
- Thresholds: the guidebook could provide more guidance on thresholds for deciding when a synthetic dataset has ' sufficient' or 'good' utility/privacy per usecase.
- Synthetic data could be 'better' than the original data: P. 58 of the guidebook: there could also be a third reason to evaluate the utility of synthetic data: to evaluate the extent to which the original dataset is an accurate representation of the real world (i.e. bias in the original dataset)