Synthetic and anonymized data

10th Montréal Industrial Problem Solving Workshop
Desjardins Group

August 27, 2020

Agenda

1. Team's presentation
2. Desjardins context & goal
3. General comments on synthetic datasets
4. Approaches
    a. Fully synthetic approaches (GANs,)
    b. Partially synthetic (De-anonymized)
5. Data-copying as a measure of privacy
6. DP-Auto-GAN

Introduction (Anne-Sophie)

Team

Sébastien Gambs
Professor @ UQAM

Anne-Sophie Charest
Professor @ Université Laval

Arezoo Rajabi
PhD Student @ Oregon State University

Mahdieh Abbasi
PhD Student @ Université Laval

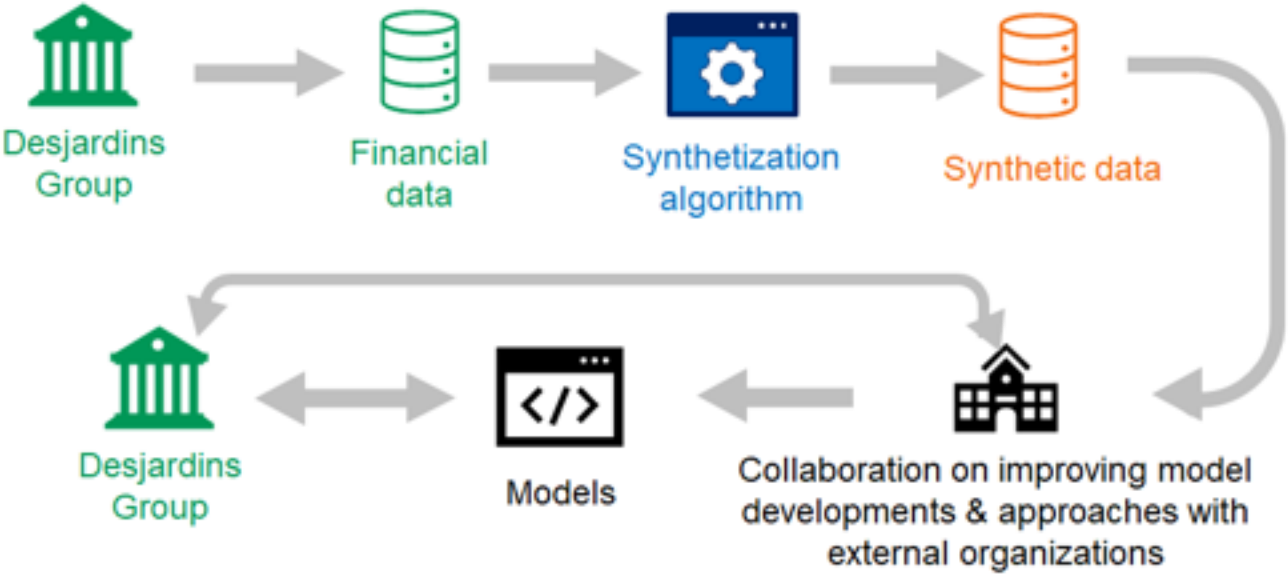
Ehsan Rezaei
PhD Student @ Polytechnique Montreal

Dylan Loader
MSc Student @ University of Calgary

Dena Kazerani
Research scientist @ Isart Digital

Desjardins: Context & Goal

- Explore approaches and develop algorithms to **produce synthetic and anonymized data,** while **retaining a maximum of statistical information** to enable the development of models.



Dataset

- Workshop research conducted on a Kaggle financial dataset

  Home Credit Default Risk : application_train.csv*
  307 511 observations and 122 variables
reduced to 82 variables during data cleaning
(some categories were also modified)

- Specific task in mind : Compute the risk of default on a mortgage loan.

What’s a good synthetic dataset?

Offers privacy

Various ways to measure it!
(either before or after producing data)

- Differential privacy
- Risk of correct prediction of confidential attributes
- Data-copying
- …

Offers utility

Various way to measure it!

- Conservation of summary statistics / statistical estimates
- Conservation of prediction power
- Similarity between the original and synthetic dataset (e.g. KL divergence, log-cluster)
- ...

How to generate a synthetic dataset?

- Classical approaches :
  Learn joint distribution of the variables and generate new data from that model.
  E.g. - R package synthpop (sequential modeling of each variable)
  Could not handle the 80 variables of the dataset on a simple desktop computer
  - Using Bayesian networks (in particular, PrivBayes also provides DP)
  Did not have the time/resources to implement
- Modern deep generative approaches:
  - GAN or VAE

Literature review on GANs approaches (Mahdieh)

Two main tracks

- **Fully synthetic data**: all the features (attributes) are sensitive

    - Estimate the data distribution, then randomly sampling from it
        - MedGAN , (Differential Privacy)-GAN

- **Partially synthetic data**: some features are sensitive, not all
    - Censor or synthesize them

A short intro. on generative models

Instead of using real data records, generate synthetic records

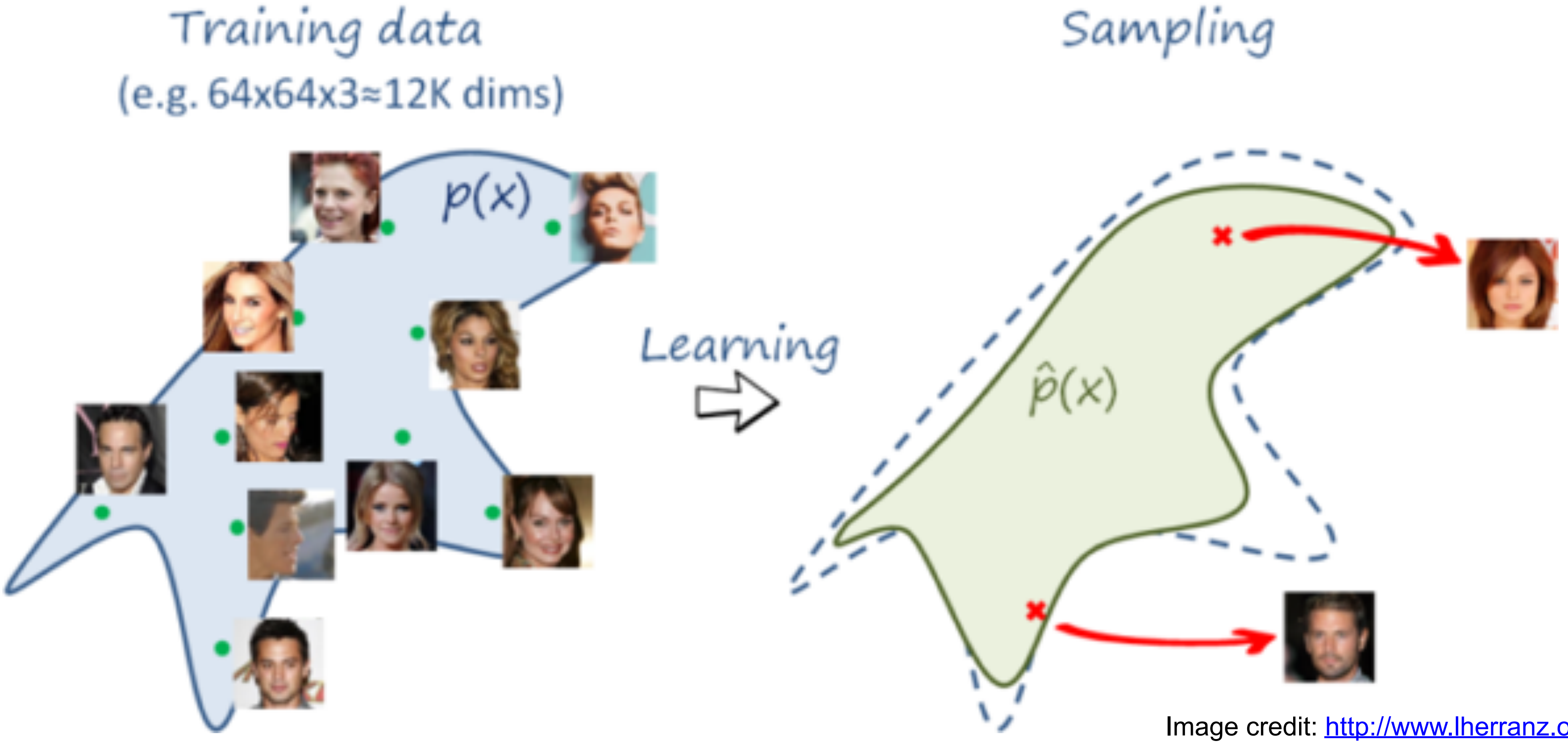


Image credit: http://www.lherranz.org/

1st track: Generative models - MedGAN

MedGAN: handle tabular features by incorporating an auto-encoder into GAN

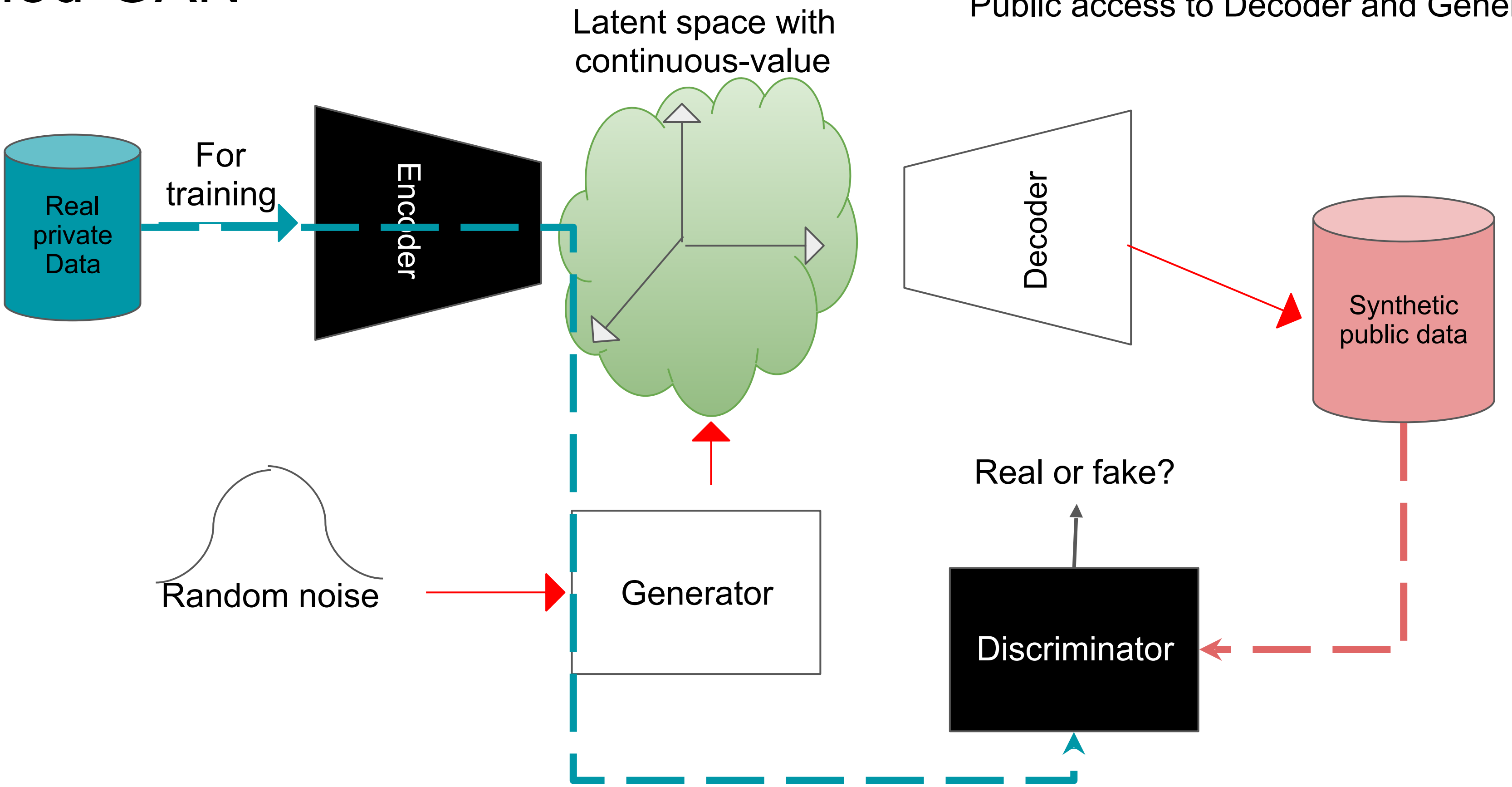
- **Pro:** handle discrete, binary, categorical features in tabular datasets

- **Cons:** No privacy guarantees (except some empirical evidences)
- No explicit privacy objective used for training MedGAN

Med-GAN

No public access to encoder and discriminator

Public access to Decoder and Generator

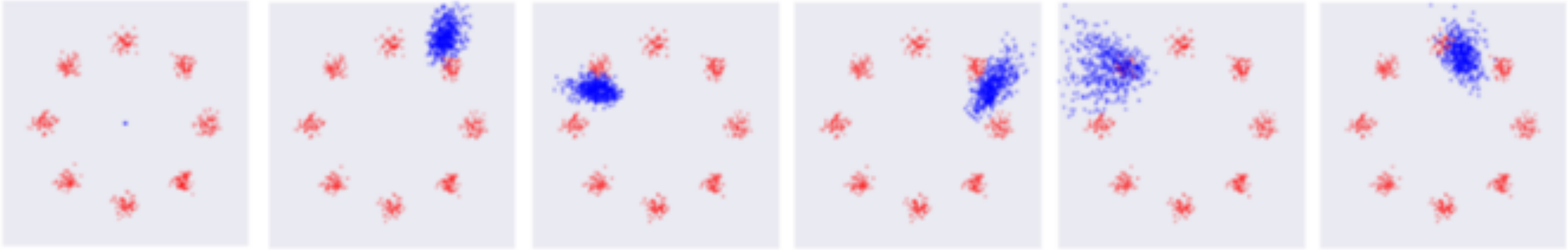


(Differential Privacy) GAN

- DP-SGD (stochastic gradient descent)
  - Clipping the gradient
  - Adding noise to the gradient

Challenges associated with generative models

Mode-collapse



1st training step Step 5k Step 10k Step 15k Step 20k Step 25k

Overfitting & memorization

Image credit: https://arxiv.org/pdf/1709.03831.pdf

Adversarial Noise: An approach for de-anonymizing datasets / partially synthetic method

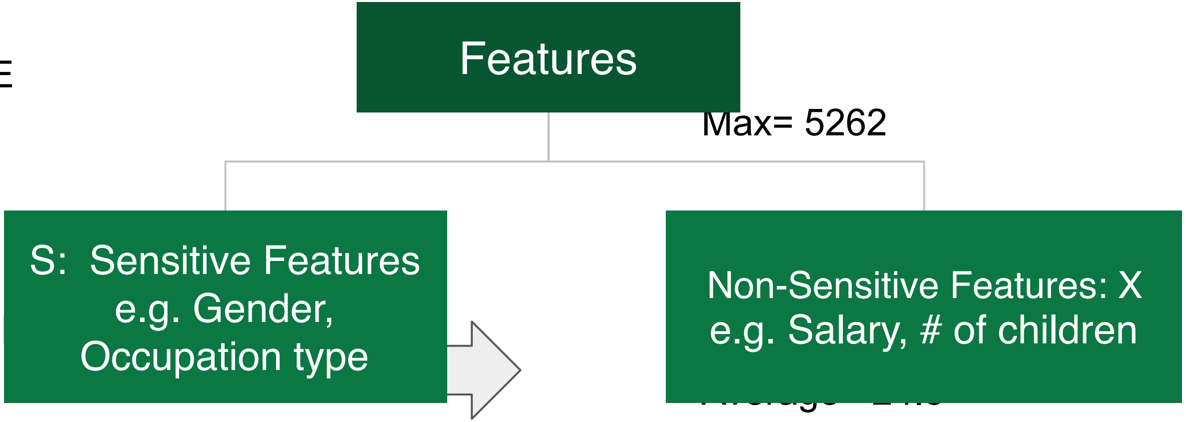
(Arezoo)

Determining Sensitive Features

How to find features that can pose a privacy risk

Categorial Features

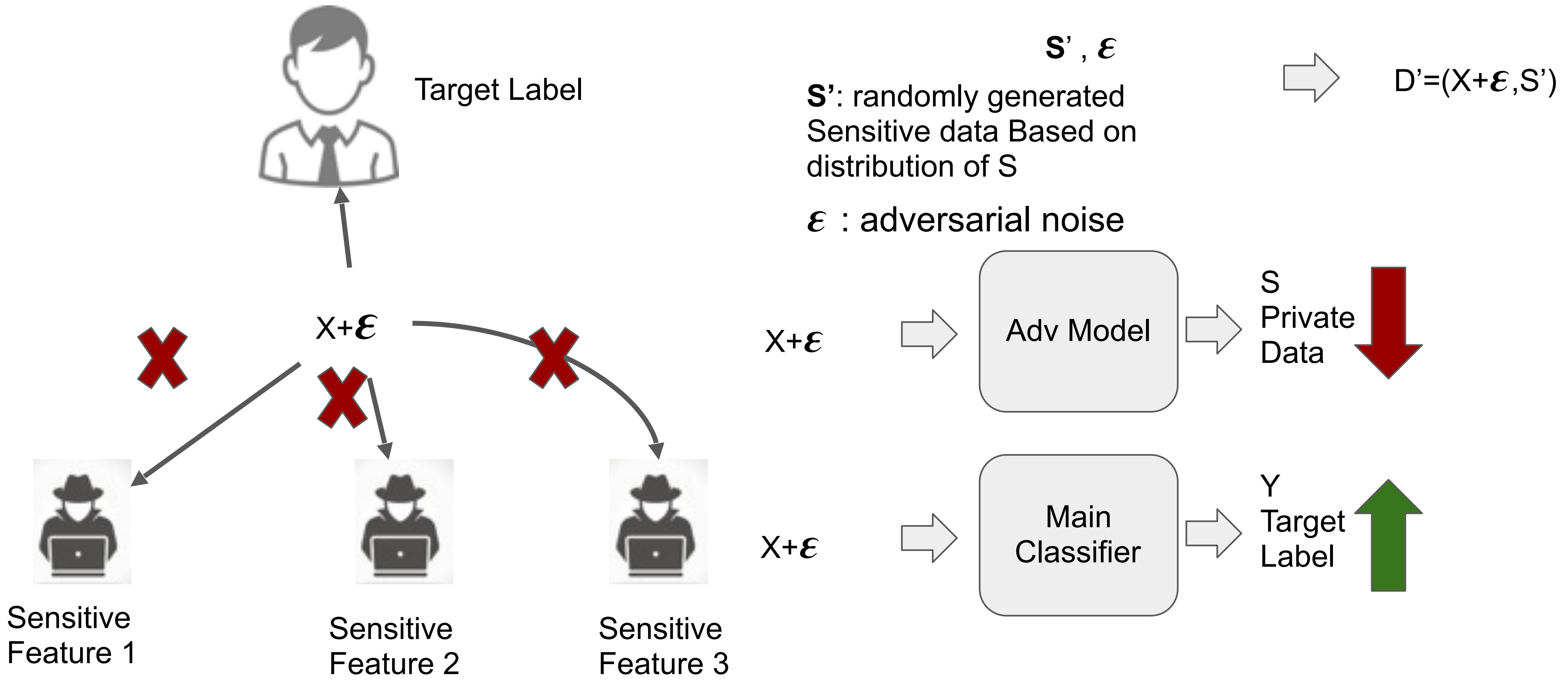
NAME-CONTRACT-TYPE
CODE-GENDER
FLAG-OWN-CAR
FLAG-OWN-REALTY
NAME-TYPE-SUITE
NAME-INCOME-TYPE
NAME-EDUCATION-TYP
NAME-FAMILY-STATUS
NAME-HOUSING-TYPE
OCCUPATION-TYPE
WEEKDAY-APPR-PROCESS-START
ORGANIZATION-TYPE
FONDKAPREMONT-MODE
HOUSETYPE-MODE
WALLSMATERIAL-MODE
EMERGENCYSTATE-MODE



Max= 5262

std= 136.47

Adversarial Noise for Deanonymization



Why and Why not Adversarial Noise

Pros:

- The final dataset looks like to the main dataset
- The relation between non-sensitive data mostly would be preserved

Cons:

- Adding many constraints to keep the relations could be computationally expensive
  - For example if one is 16 years old or less can not have a several children
- Accuracy is the main metric to measure when to stop

Classification

	precision	recall	f1-score	support
0	0.97	0.28	0.43	93362
1	0.10	0.90	0.18	8117

ROC AUC score is: 0.5887238259950801

GaussianNB

	precision	recall	f1-score	support
0	0.92	1.00	0.96	93362
1	0.54	0.01	0.02	8117

ROC AUC score is: 0.5043280697209536

LogisticRegression

Classification using Only Non-Sensitive Features

The classification report is as follows:

	precision	recall	f1-score	support
0	0.92	0.99	0.95	282686
1	0.07	0.01	0.01	24825
accuracy			0.91	307511

ROC AUC score is: 0.4993022716147093

GaussianNB

Overfitting as a measure of privacy

(Ehsan)

Overfitting as a measure of privacy

Overfitting and data memorization in generative models is a serious threat for data privacy:

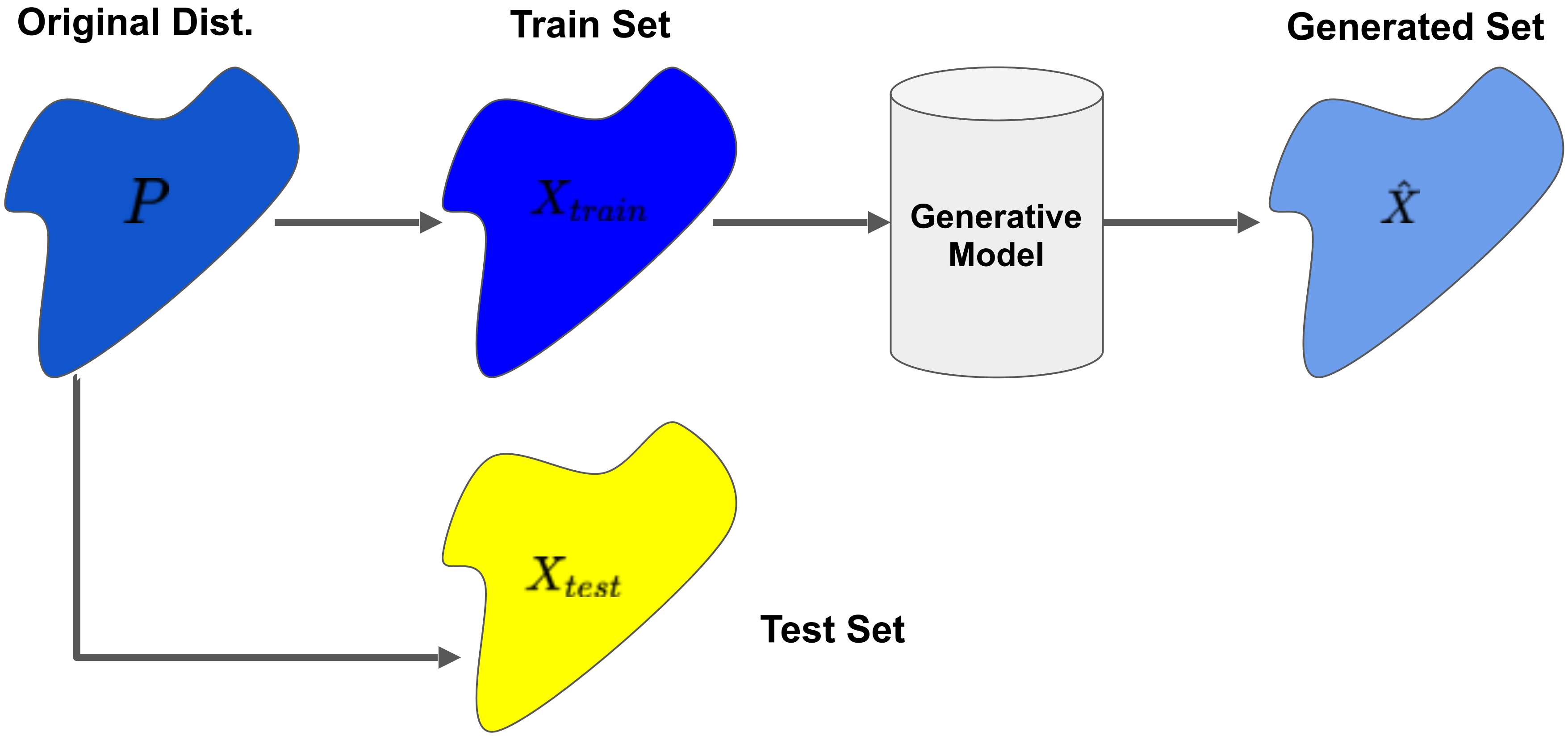
- Increasing the identity risk
- Increasing attribute disclosure risk
- ....

Solution:

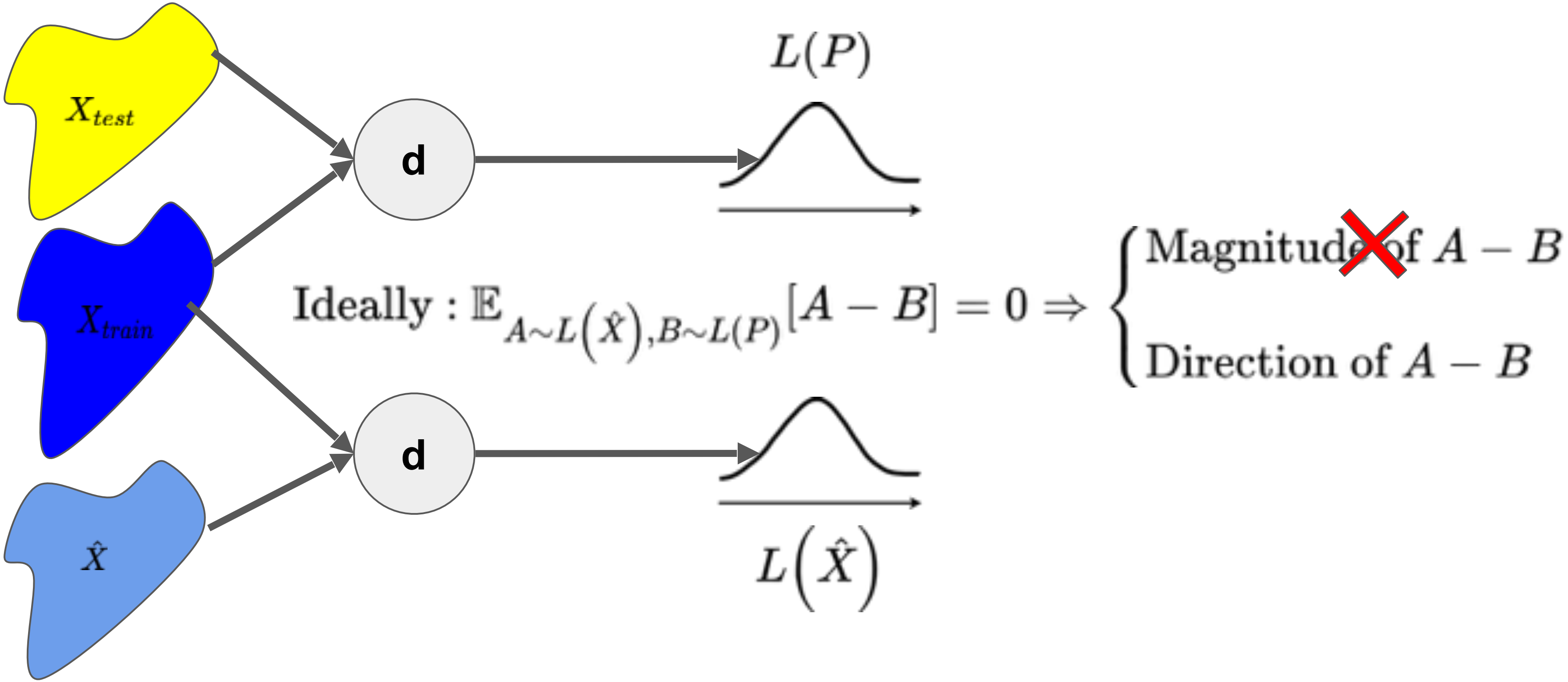
- Data-copying
- Over-representing
- ...

★ "A Non-Parametric Test to Detect Data-Copying in Generative Models", C. Meehan, K. Chaudhuri, S. Dasgupta

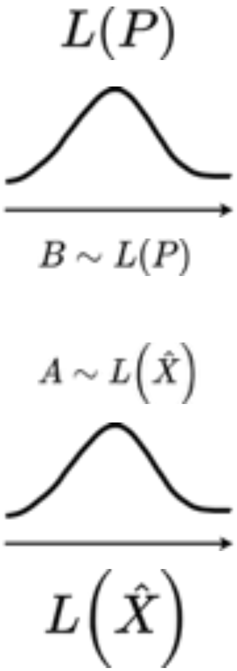
Data-copying



Data-copying

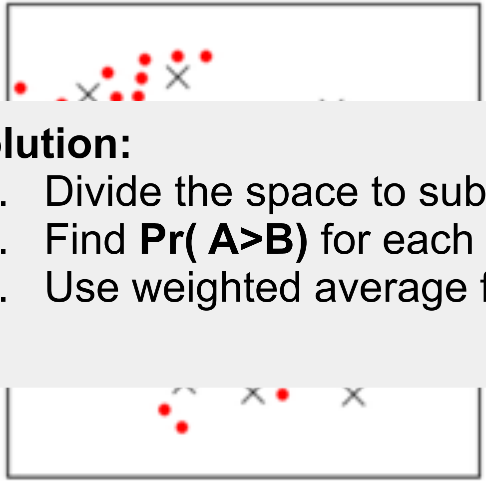


Data-copyin



Solution:

1. Divide the space to subspaces.
2. Find **Pr(A>B)** for each subspace.
3. Use weighted average for data-copying value.



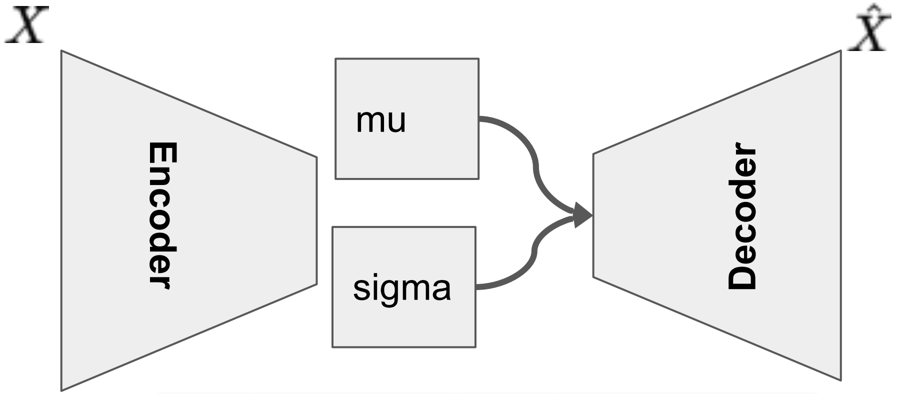
(a) Illustration of over-/under-representation Training sample: ×, Generated sample: •

(b) Illustration of data-copying/underfitting Training sample: ×, Generated sample: •

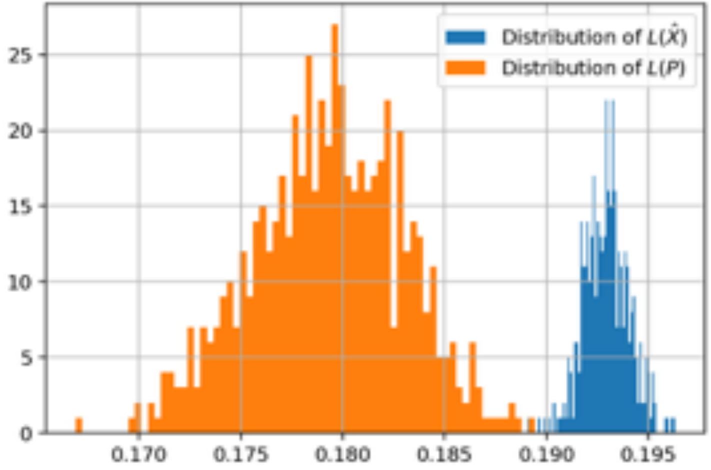
Adopted from the same reference

Simulation Results

VAEs + Data-copying performance:
N_features : 203
X_training : 210k, X_test = 60k, X_Gen = 40k
Scenario 1: Layers = [80,50,30], k-mean cluster = 2



$$Z_U\left(L\left(\hat{X}\right), L(P)\right) = \begin{cases} << 0 & \text{data-copying} \\ >> 0 & \text{underfitting} \\ o.w. & \hat{X} \text{ is an appropriate data set} \end{cases}$$



Z_u = 2.4*1e9

Data-copying

Pros:

- Can be used as an initial test for checking the performance of an existed synthetic data set.
- Can be used to measure the privacy of data

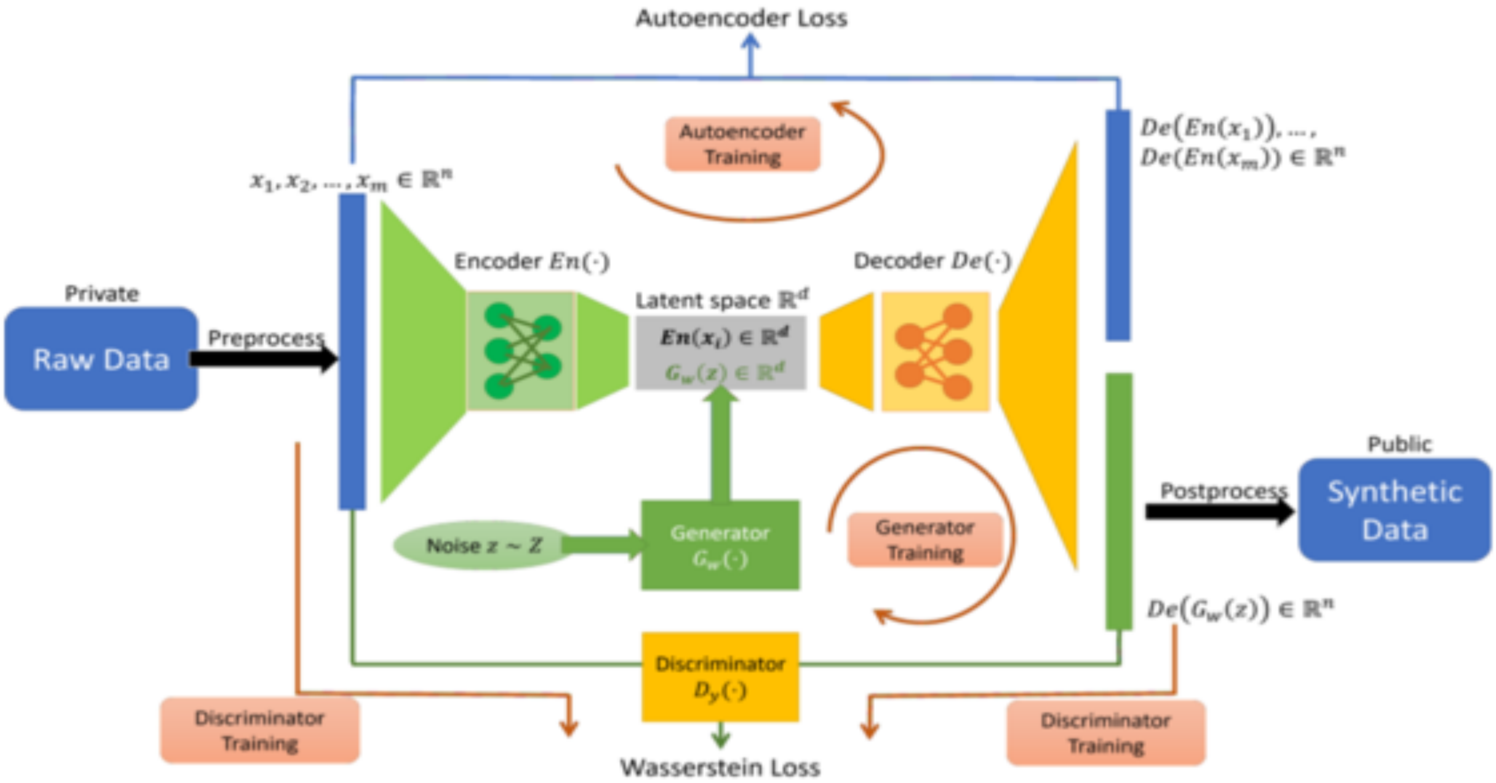
Cons:

- Need to tune hyper-parameters precisely
- It needs to be merged with some other tools for better performance.

DP-Auto-GAN

(Dylan)

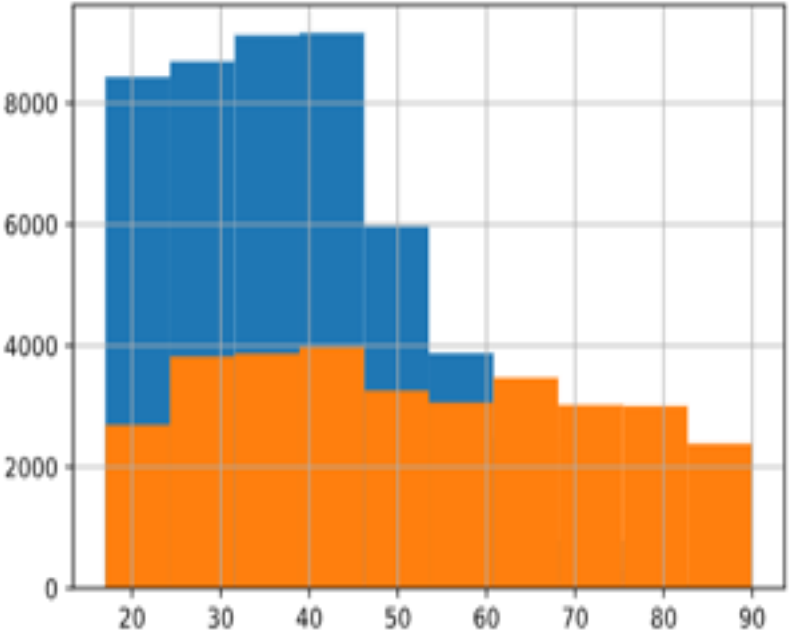
Model Architecture



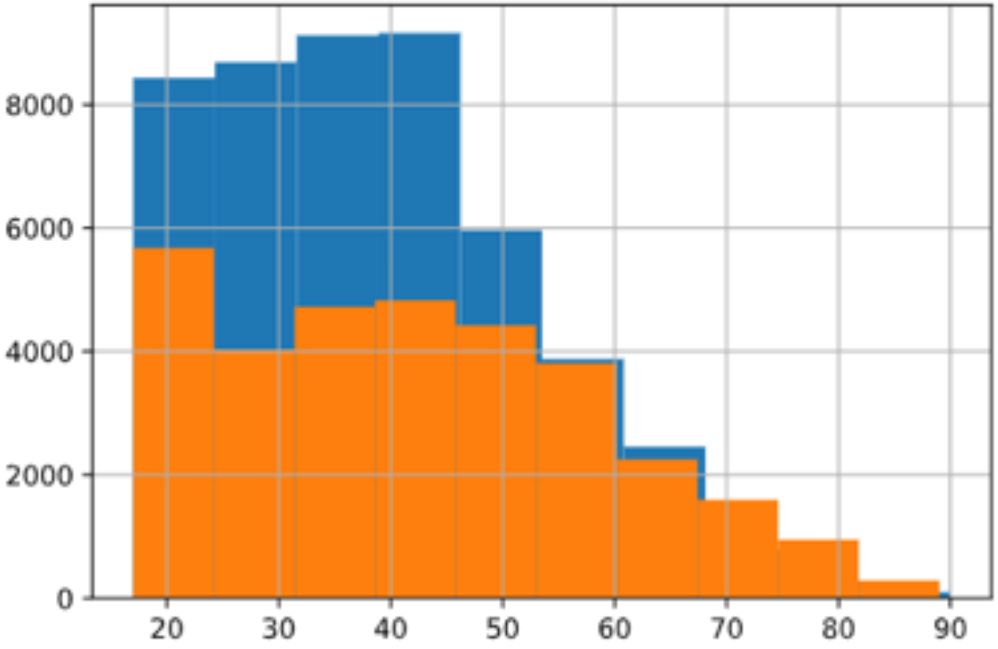
Source: arXiv:1912.03250 [cs.LG]

Graphical Evaluation

Histogram For Age (Initial)



Histogram For Age (15,000 Iterations)



Initial Thoughts on DP-Auto-GAN

Pros:

- Works on binary and mixed data types
- Shows good results for low epsilon values, epsilon ~1.

Cons:

- Training an autoencoder and GAN are computer intensive even for small datasets such as ADULT.
- There is limited information on the amount of data that can be generated
- The proposed methods of evaluation are mostly visual

Conclusion

(Anne-Sophie)

Future work

- Measuring privacy
  - Need to decide exactly what protection is desired
  - Data-copying is a promising criterion; need to study more and compare to DP
- Measuring utility
  - Need to make a complete list of specific utility measures which are desired (if only one classification task is of interest, synthetic data generation is not ideal)
- Generating the synthetic datasets
  - With more time / computing resources it should be possible to get synthpop to work
  - GAN-like methods are promising, but also require a lot of time and knowledge
  - Any method used will have to be personalized to the specific dataset of interest (non-Kaggle)

Thanks for your attention

Q&A