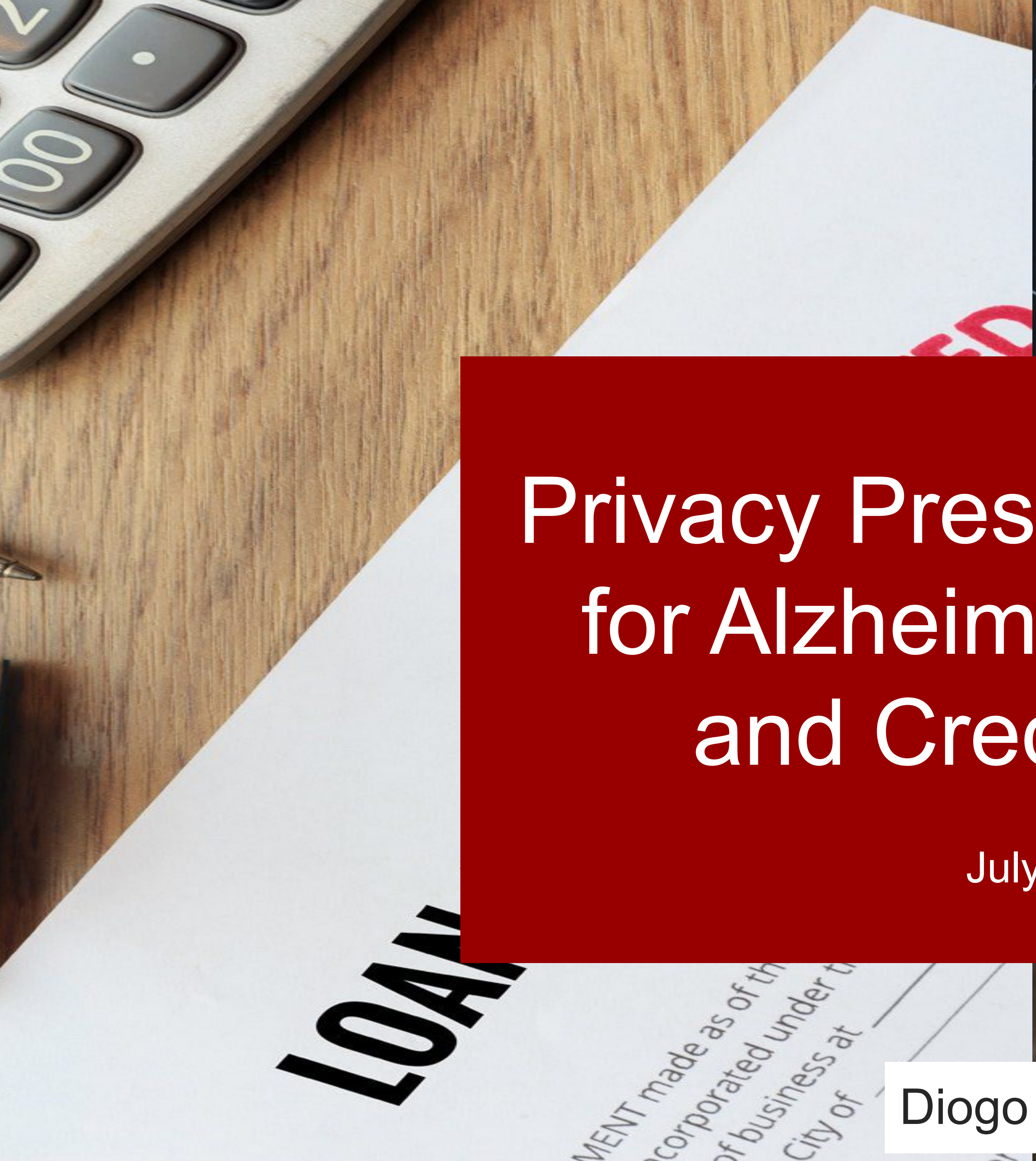
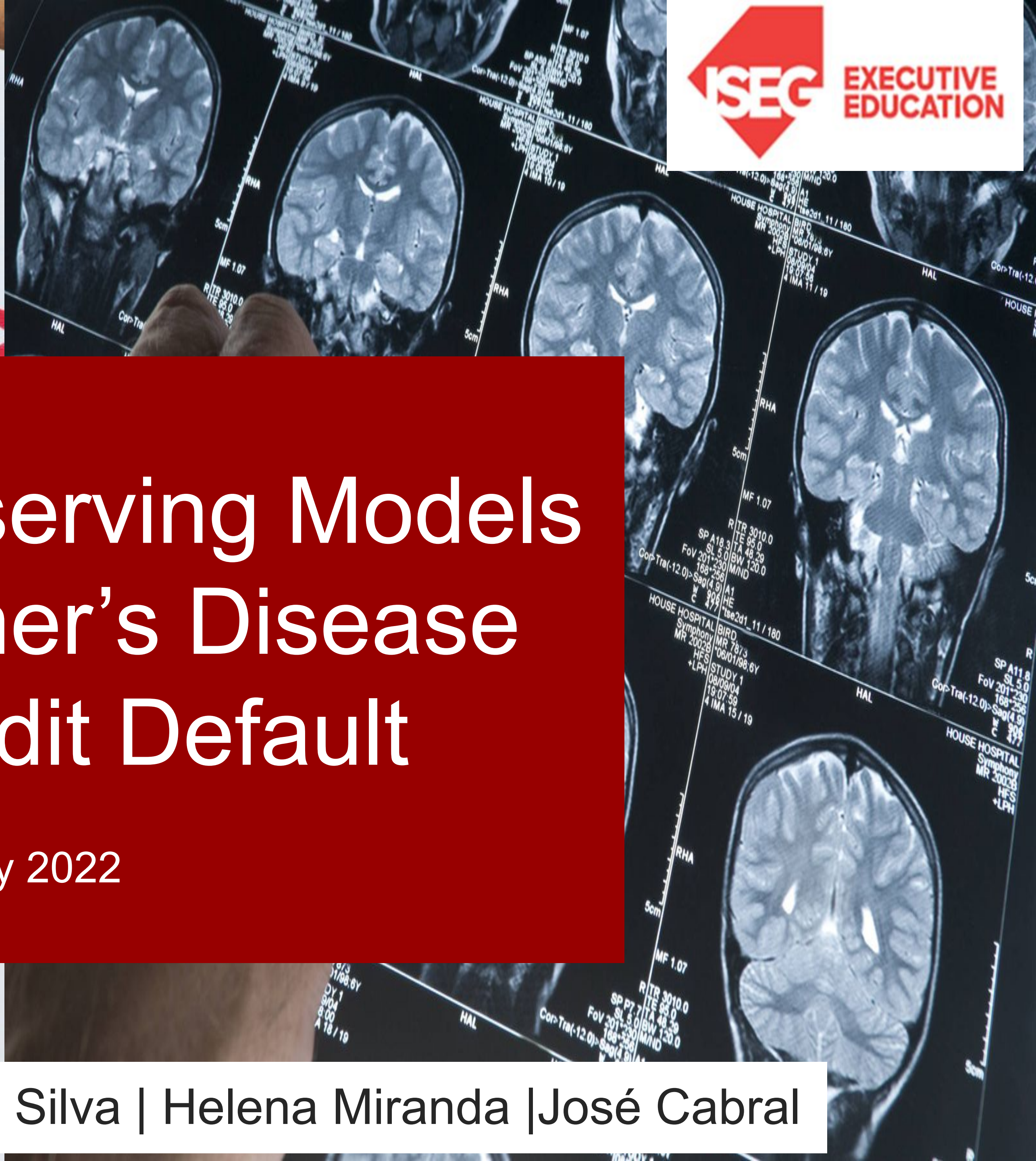



ISEG EXECUTIVE EDUCATION

Privacy Preserving Models for Alzheimer’s Disease and Credit Default

July 2022

Diogo Silva | Helena Miranda |José Cabral

Index

- Use Case 1: Predicting Alzheimer’s Disease
    - Synthetic data generation
    - Classification
- Use Case 2: Predicting Mortage Default
    - Architecture: Data Generation
    - Anonymizing Data with ARX
    - Generating Synthetic Data
    - Modelling with Original vs Anonymized vs Synthetic Dataset
- Conclusion
- Further Improvements

Use Case 1:

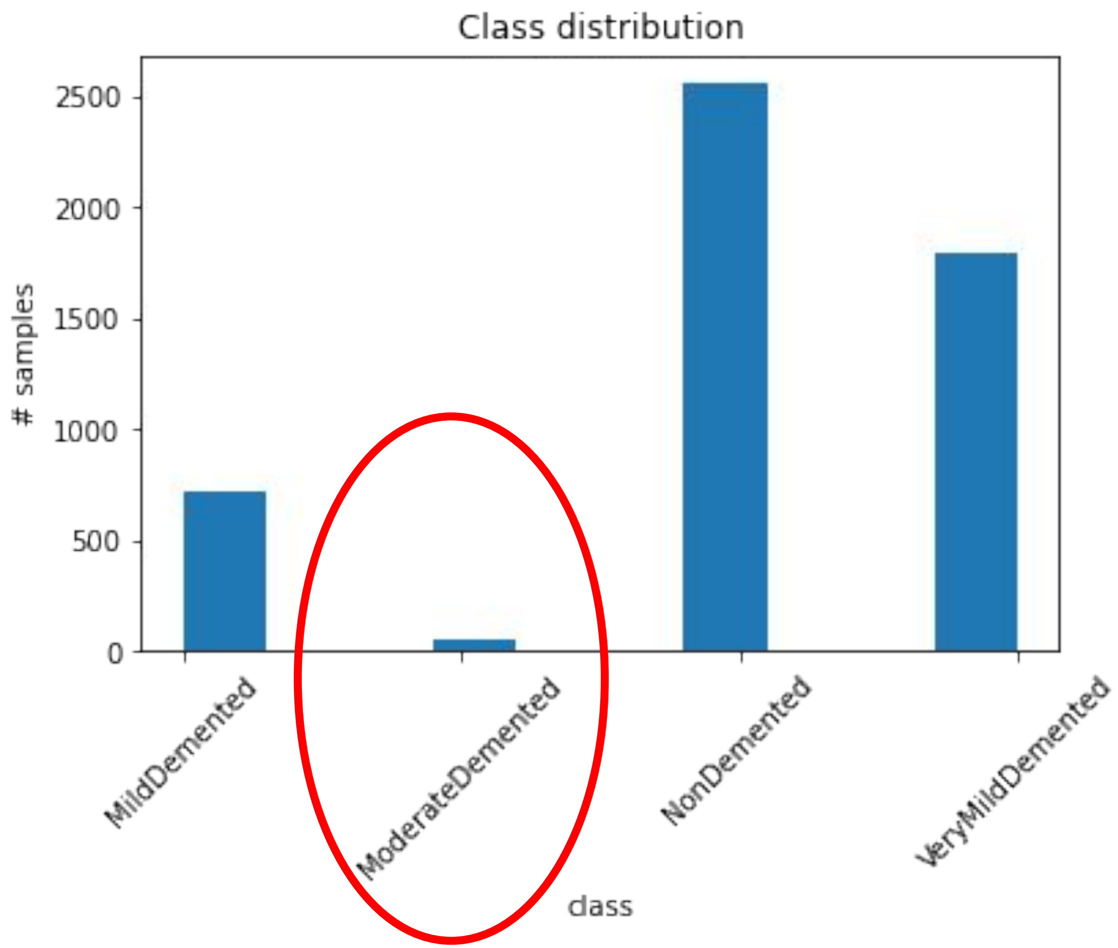
Predicting Alzheimer’s Disease

Synthetic data

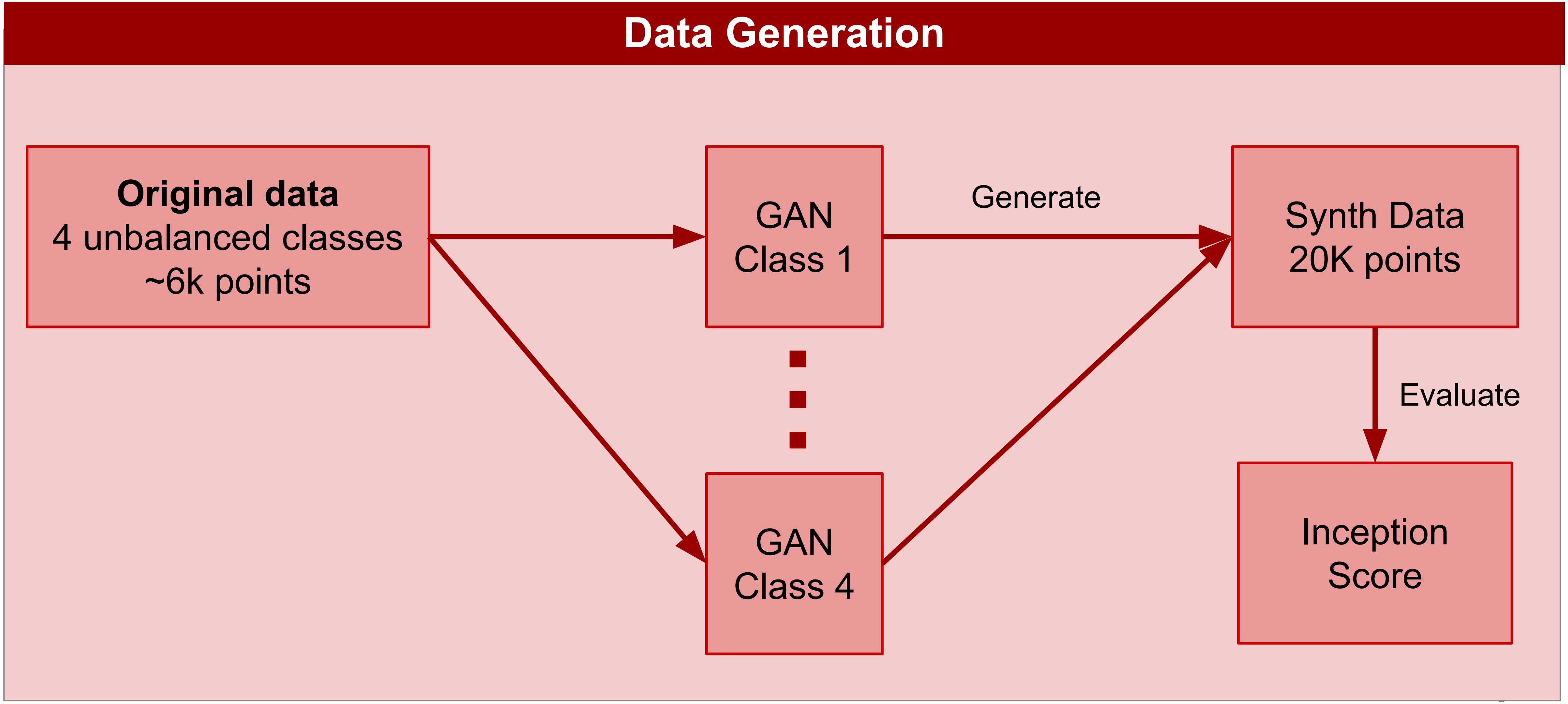
- Why use synthetic data?
    - Develop models without compromising privacy.
    - Data augmentation.

Synthetic data

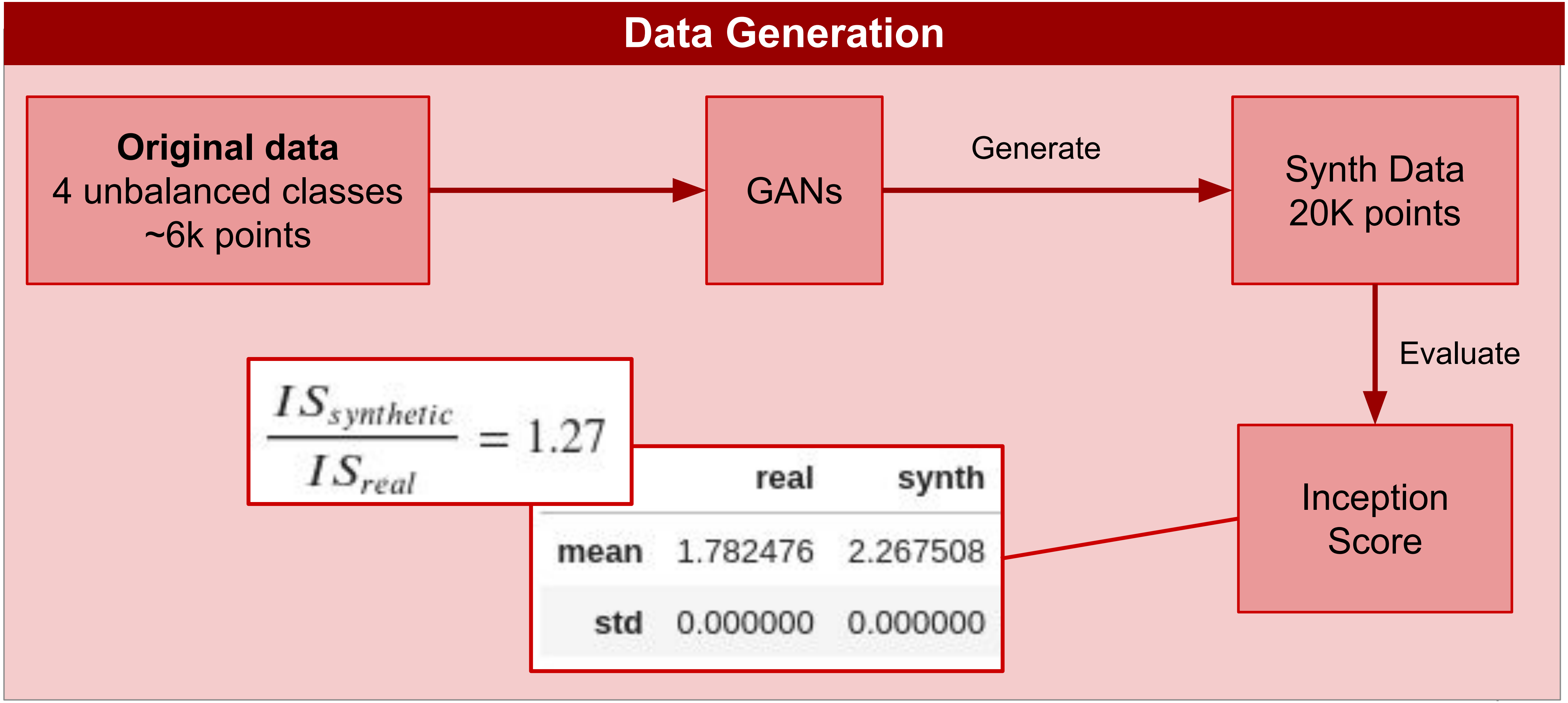
- Why use synthetic data?
  - Develop models without compromising privacy.
  - Data augmentation.



Synthetic data generation



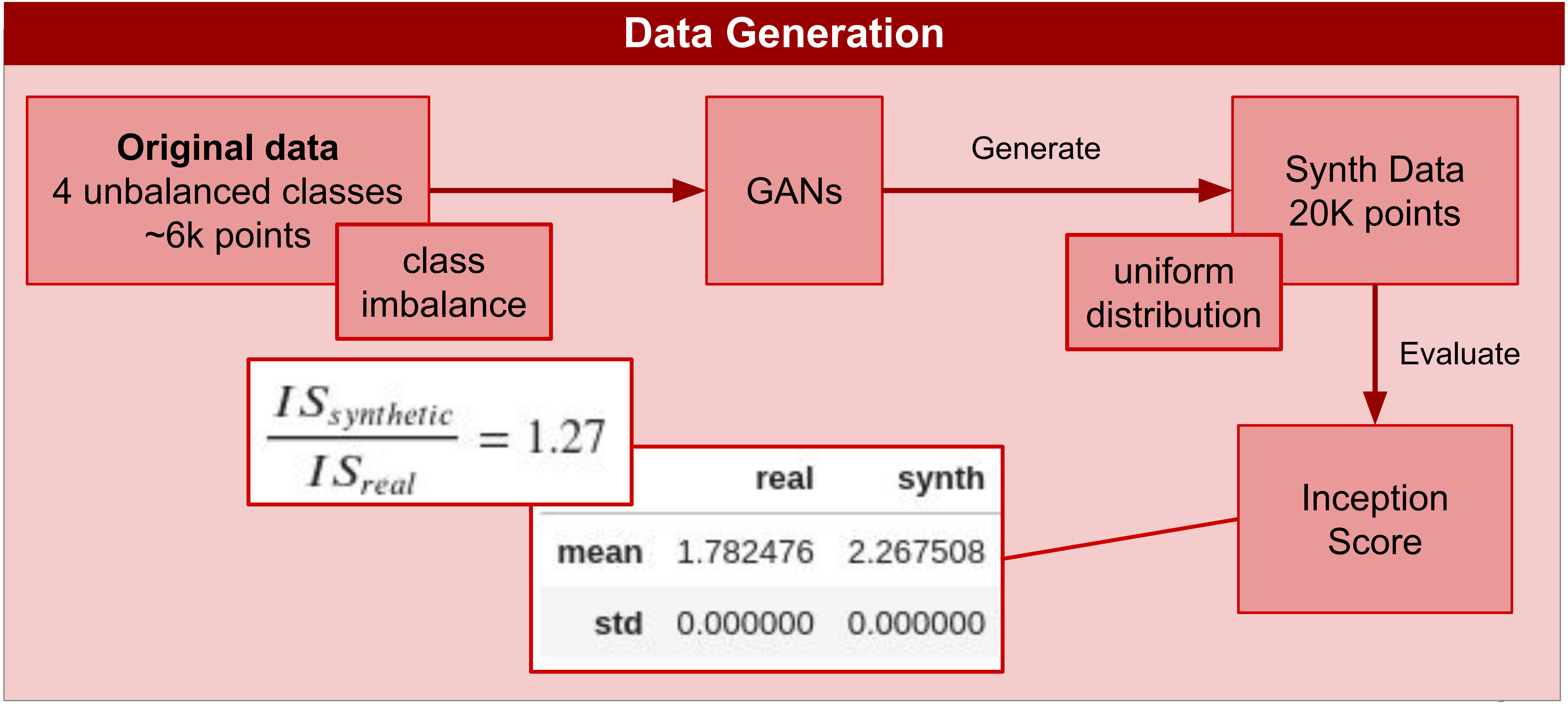
Synthetic data generation



$$\frac{IS_{synthetic}}{IS_{real}} = 1.27$$

	real	synth
mean	1.782476	2.267508
std	0.000000	0.000000

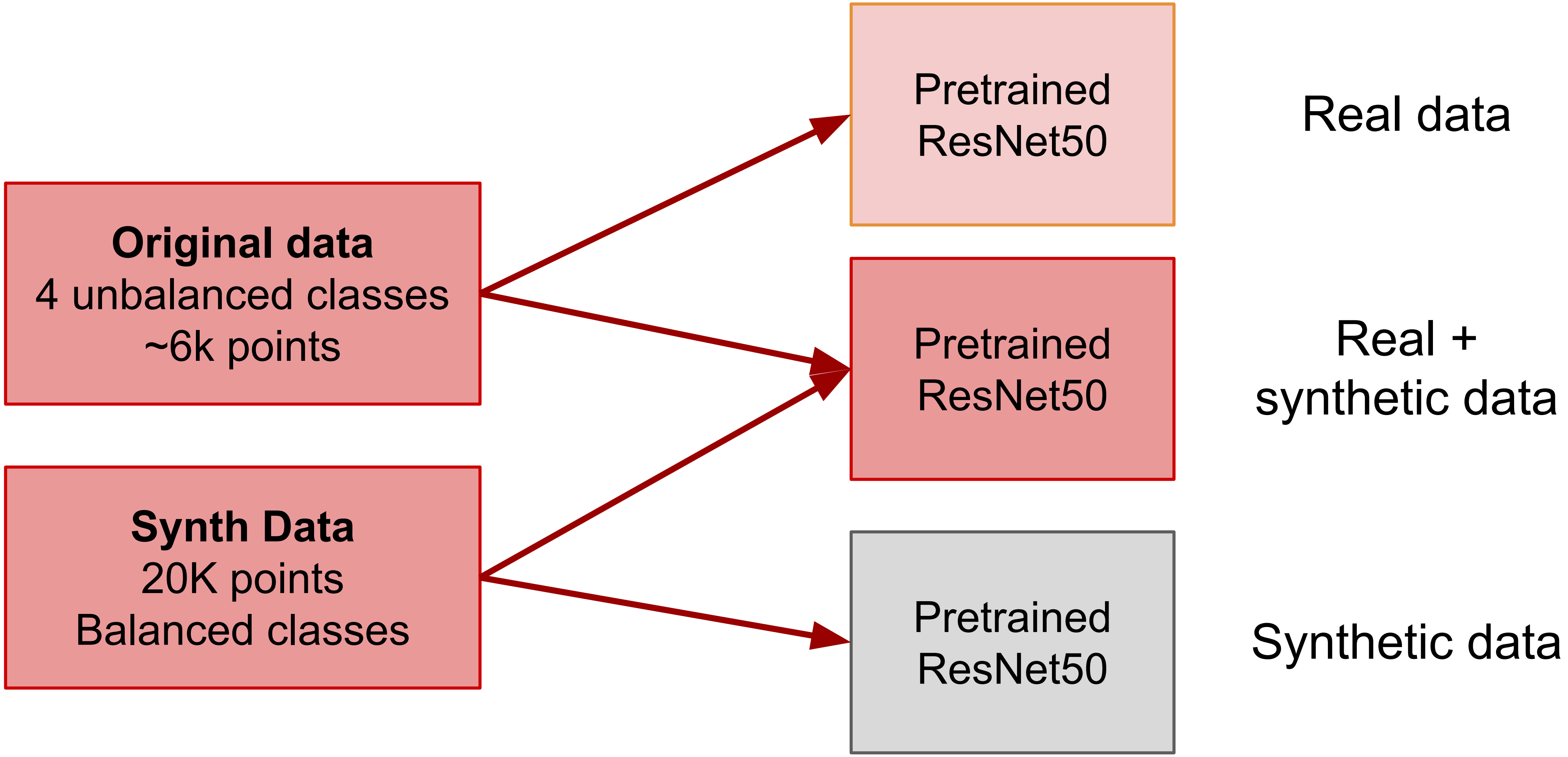
Synthetic data generation



Classification | Architecture

- SGD
- Cross Entropy
- 100 epochs

- 0.001 LR
- LR decay scheduler



Classification | Results

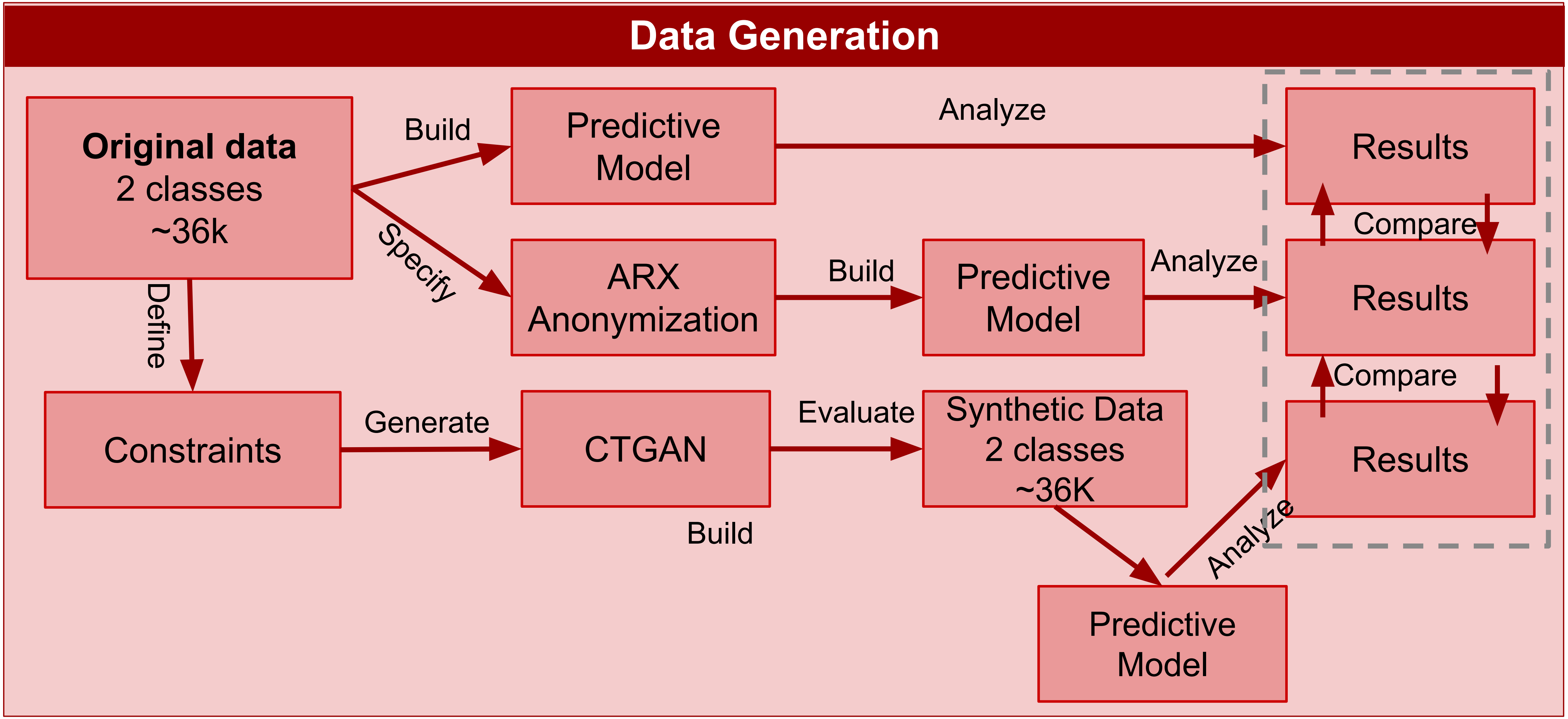
- Overall poor model performance on real data test set.
- Better performance with hybrid model.
- Good performance on synthetic data test set.
- Possible existence of identifying synthetic artifacts

Model	Eval. dataset	Accuracy	Loss
real data, 0.001LR, SGD, 100e	0 real train	56,6%	0,882
real data, 0.001LR, SGD, 100e	**1 real test**	**53,8%**	**0,915**
real data, 0.001LR, SGD, 100e	2 synth train	25,2%	1,722
real data, 0.001LR, SGD, 100e	**3 synth test**	**26,0%**	**1,695**
real data, 0.001LR, SGD, 100e	4 real synth train	31,8%	1,551
real+synth data, 0.001LR, SGD, 100e	0 real train	60,0%	0,817
real+synth data, 0.001LR, SGD, 100e	**1 real test**	**58,2%**	**0,875**
real+synth data, 0.001LR, SGD, 100e	2 synth train	99,0%	0,029
real+synth data, 0.001LR, SGD, 100e	**3 synth test**	**100,0%**	**0,002**
real+synth data, 0.001LR, SGD, 100e	4 real synth train	91,3%	0,189
synth data, 0.001LR, SGD, 100e	0 real train	13,8%	7,253
synth data, 0.001LR, SGD, 100e	**1 real test**	**14,0%**	**7,768**
synth data, 0.001LR, SGD, 100e	2 synth train	97,5%	0,079
synth data, 0.001LR, SGD, 100e	**3 synth test**	**99,7%**	**0,023**
synth data, 0.001LR, SGD, 100e	4 real synth train	80,5%	1,540

Use Case 2:

Predicting Mortgage Defaults

Architecture



Anonymizing Data with ARX: Data Types Definition

Identifying

- ID

Quasi-Identifiers

- CODE_GENDER
- CNT_CHILDREN
- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- NAME_FAMILY_STATUS
- OCCUPATION_TYPE
- CNT_FAM_MEMBERS

Sensitive

- AMT_INCOME_TOTAL
- DAYS_BIRTH

Privacy models | Population | Costs and benefits

Type	Model	Attribute
k	10-Anonymity	
ℓ	Distinct-4-diversity	AMT_INCOME_TOTAL
ℓ	Distinct-4-diversity	DAYS_BIRTH

We have applied **k-anonymity of 10** to all our data and a **L-diversity of 4** the variables identified as sensible (Income and Days of Birth). Those parameters were chosen based on the **trade off analysis of our anonymized model quality and risk vs information loss**.

Quasi-Identifiers Transformations

Masking Approach

NAME_FAMILY_STATUS *

Level-0	Level-1	Level-2	Level-3	Level-4
Civil marriage	Civil marriage *	Civil marriage **	Civil marriage ***	Civil marriage ****
Married	Married *	Married **	Married ***	Married ****
Separated	Separated *	Separated **	Separated ***	Separated ****
Single / not marr...	Single / not marr...	Single / not marr...	Single / not marr...	Single / not mar*...
Widow	Widow *	Widow **	Widow ***	Widow ****

NAME_INCOME_TYPE *

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
Commercial asso...	Commercial asso...	Commercial asso...	Commercial asso...	Commercial asso...	Commercial asso...
Pensioner	Pensioner *	Pensioner **	Pensioner ***	Pensioner ****	Pensioner *****
State servant	State servant *	State servant **	State servant ***	State servant ****	State servant *****
Student	Student *	Student **	Student ***	Student ****	Student *****
Working	Working *	Working **	Working ***	Working ****	Working *****

NAME_OCUPATION_TYPE *

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
	*	**	***	****	*****
Accountants	Accountants *	Accountants ...	Accountants ...	Accountants *...	Accountants **...
Cleaning staff	Cleaning staff *	Cleaning staff **	Cleaning staff *...	Cleaning staff *...	Cleaning staff **...
Cooking staff	Cooking staff *	Cooking staff **	Cooking staff *...	Cooking staff *...	Cooking staff **...
Core staff	Core staff *	Core staff **	Core staff ***	Core staff ****	Core staff *****
Drivers	Drivers *	Drivers **	Drivers ***	Drivers ****	Drivers *****
HR staff	HR staff *	HR staff **	HR staff ***	HR staff ****	HR staff *****
High skill tech st...	High skill tech st...	High skill tech st...	High skill tech st*...	High skill tech s*...	High skill tech **...
IT staff	IT staff *	IT staff **	IT staff ***	IT staff ****	IT staff *****
Laborers	Laborers *	Laborers **	Laborers ***	Laborers ****	Laborers *****

NAME_EDUCATION_TYPE *

Level-0	Level-1	Level-2	Level-3
Lower secondary	{Lower secondar...	{Lower secondar...	*
Secondary / seco...	{Lower secondar...	{Lower secondar...	*
Incomplete higher	{Incomplete high...	{Lower secondar...	*
Higher education	{Incomplete high...	{Lower secondar...	*
Academic degree	{Academic degre...	{Academic degre...	*

The variables were anonymized using a masked approach which implies suppressing letters from in a word.

*Sample example of the interval levels created in ARX

Quasi-Identifiers Transformations

Intervals Approach

CNT_FAMILY_MEMBERS *

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
1.0	[1, 2[	[1, 4[	[1, 8[	[1, 16[	*
2.0	[2, 4[	[1, 4[	[1, 8[	[1, 16[	*
3.0	[2, 4[	[1, 4[	[1, 8[	[1, 16[	*
4.0	[4, 6[	[4, 8[	[1, 8[	[1, 16[	*
5.0	[4, 6[	[4, 8[	[1, 8[	[1, 16[	*
6.0	[6, 8[	[4, 8[	[1, 8[	[1, 16[	*
7.0	[6, 8[	[4, 8[	[1, 8[	[1, 16[	*
9.0	[8, 10[	[8, 12[	[8, 16[	[1, 16[	*
15.0	[14, 16[	[12, 16[	[8, 16[	[1, 16[	*
20.0	[20, 21[	[20, 21[	[16, 21[	[16, 21[	*

CNT_CHILDREN

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
0	[0, 1[	[0, 2[	[0, 4[	[0, 8[	*
1	[1, 2[	[0, 2[	[0, 4[	[0, 8[	*
2	[2, 3[	[2, 4[	[0, 4[	[0, 8[	*
3	[3, 4[	[2, 4[	[0, 4[	[0, 8[	*
4	[4, 5[	[4, 6[	[4, 8[	[0, 8[	*
5	[5, 6[	[4, 6[	[4, 8[	[0, 8[	*
7	[7, 8[	[6, 8[	[4, 8[	[0, 8[	*
14	[14, 15[	[14, 16[	[12, 16[	[8, 16[	*
19	[19, 20[	[18, 20[	[16, 20[	[16, 20[	*

The variables anonymization technique groups data into intervals.

*Sample example of the interval levels created in ARX

Quasi-Identifiers Transformations

Ordering Approach

NAME_EDUCATION_TYPE *

Level-0	Level-1	Level-2	Level-3
Lower secondary	{Lower secondar...	{Lower secondar...	*
Secondary / seco...	{Lower secondar...	{Lower secondar...	*
Incomplete higher	{Incomplete high...	{Lower secondar...	*
Higher education	{Incomplete high...	{Lower secondar...	*
Academic degree	{Academic degre...	{Academic degre...	*

CNT_GENDER

Level-0	Level-1
F	{F, M}
M	{F, M}

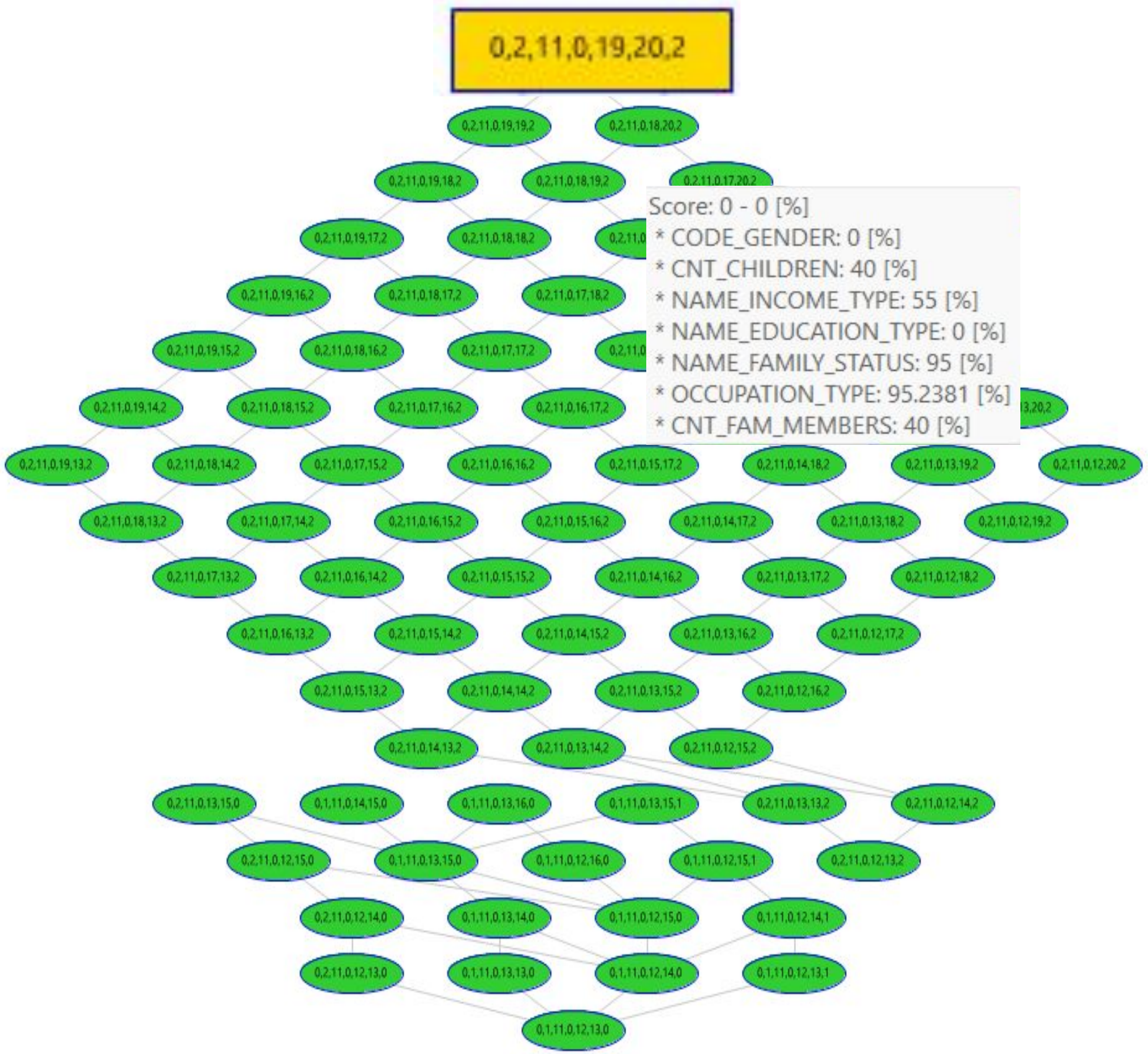
The variables were transform using by grouping levels of generalization (e.g. in the gender variable we can group Man and Women in a level, a level that only considers person).

*Sample example of the interval levels created in ARX

Anonymizing Data with ARX

Our optimal anonymization solution satisfies the conditions required to guarantee a successful anonymization of our data.

	Level	Score
Gender:	0	0%
Children:	2	40%
Income Type:	11	55%
Education Type:	0	0%
Family Status:	19	95%
Occupation Type:	20	95%
Family Members:	2	40%



Anonymization Models

Optimal Solution Evaluation

- Information Loss: 7%
- Intensity: 50%
- Granularity: 86%

Output data | Classification performance | Quality models

Attribute-level quality

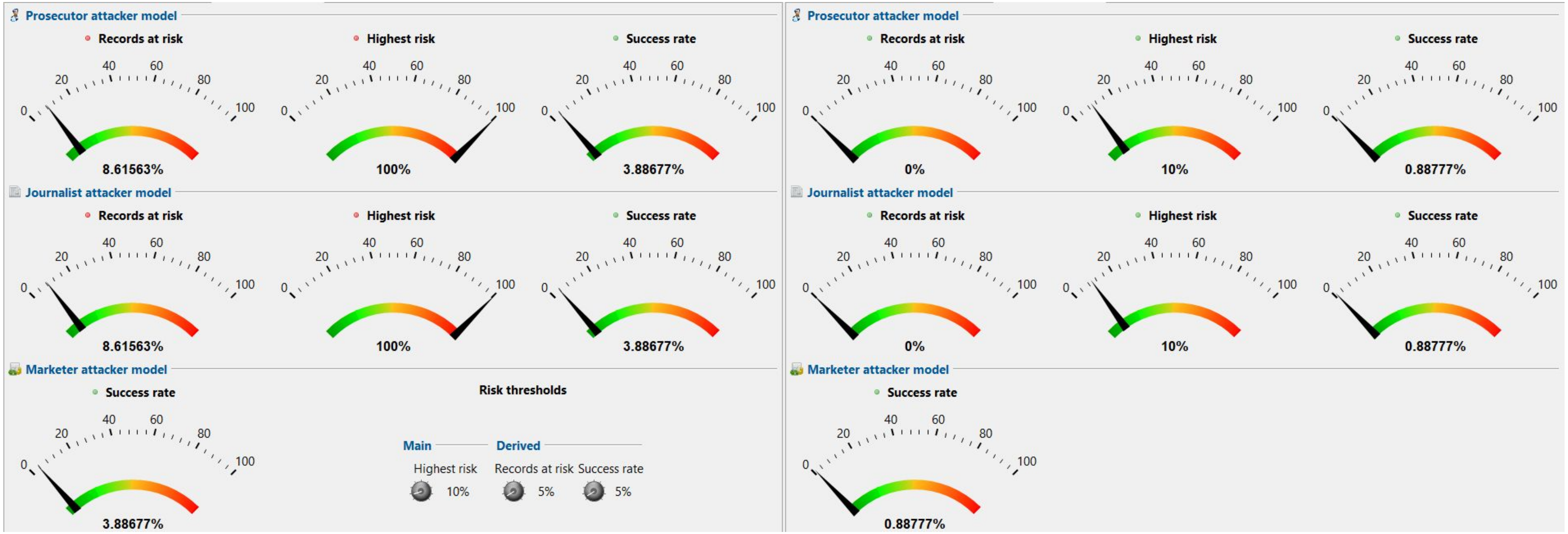
Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
CODE_GENDER	String	7.00003%	92.99997%	92.99997%	92.59166%	92.99997%
CNT_CHILDREN	String	7.00003%	55.79998%	81.37498%	33.73913%	93.07429%
NAME_INCOME_...	String	7.00003%	41.84999%	92.99997%	92.64762%	93.56698%
NAME_EDUCATI...	String	7.00003%	92.99997%	92.99997%	88.39652%	93.30174%
NAME_FAMILY_S...	String	7.00003%	4.65%	88.64073%	79.71472%	89.90591%
OCCUPATION_TY...	String	7.00003%	4.42857%	88.71776%	80.8133%	92.99362%
CNT_FAM_MEMB...	String	7.00003%	55.79998%	71.45526%	22.75955%	92.64913%

Dataset-level quality

Model	Quality
Gen. intensity	49.78978%
Granularity	86.38782%
N.-U. entropy	70.00943%
Discernibility	92.52251%
Average class size	99.76143%
Record-level squared error	87.13097%
Attribute-level squared error	92.88309%
Aggregation-specific squared error	70.86463%

Risk Evaluation

With this model our success rate in the <u>**prosecutor risk**</u> (which is the biggest danger) after anonymization is almost 0% while in the original dataset is of **3.8%.**



Synthetic Data Generation

Constraints Definition

1. IDs are unique
2. Days_Birth is lower than Days employed and individuals must have at least 16 years old to work.
3. Work:Phone implies that individuals are not unemployed.
4. Unemployed (positive work days) do not have information regarding their occupation.
5. Pensioners have positive work days.

Synthetic Data Generation

Interpreting our Synthetic Data Quality

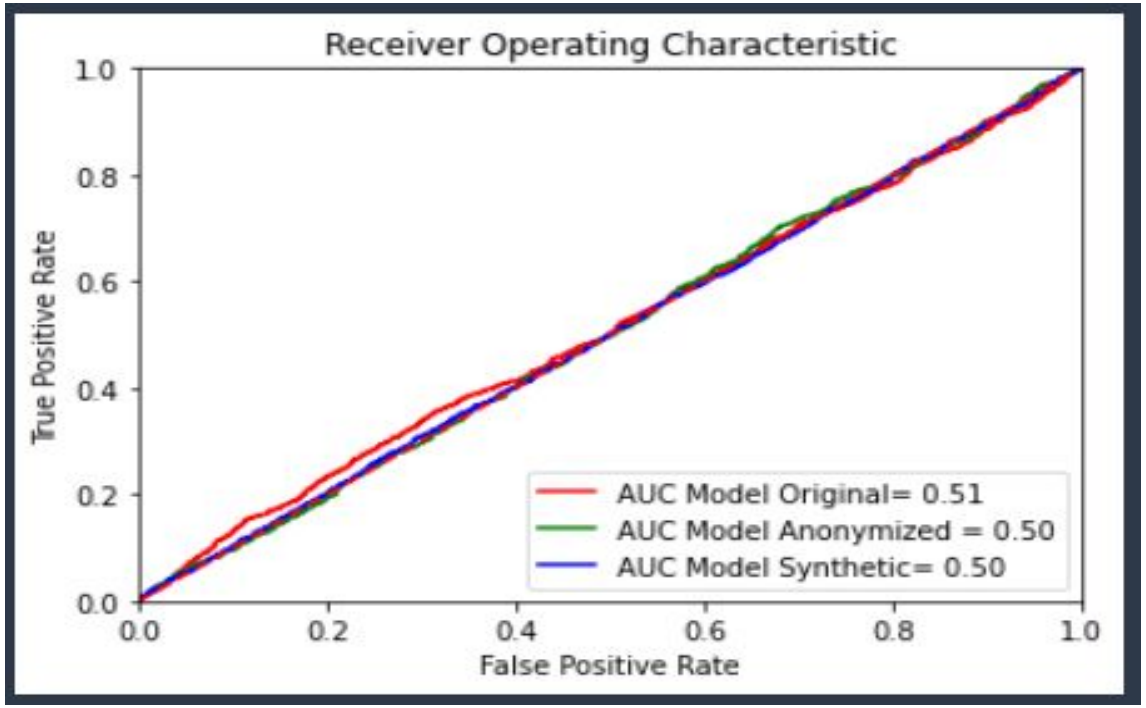
```
evaluate(synthetic_data, data)

0.5597955030608549
```

	metric	name	raw_score	normalized_score	min_value	max_value	goal	error
0	CSTest	Chi-Squared	0.972954	0.972954	0.0	1.0	MAXIMIZE	None
1	KSTest	Inverted Kolmogorov-Smirnov D statistic	0.834501	0.834501	0.0	1.0	MAXIMIZE	None

The results of the evaluate function - which is a combination of several methods - is **satisfactory**. Nevertheless, if we look separately to the our **categorical variables, CSTest reveals a pretty good result** (0.97%) as well our numerical variables according to KSTest (0.83).

Modelling with Original vs Anonymized vs Synthetic Data

Interpreting our Model Results



	Model	Test MAE	Test MSE	Test RMSE	Test ACCURACY
0	Linear Regression Original	0.126174	0.126174	0.355210	0.873826
1	Linear Regression Anonymized	0.120208	0.120208	0.346711	0.879792
2	Linear Regression Synthetic	0.116026	0.116026	0.340625	0.883974

Our models do not discriminate credit default well.

Synthetic data achieved the highest accuracy while our **Original** data allow us to have the best results in terms of AUC. This results are due to the fact that our data is unbalanced and our model predicts much better one class than the other.

Nevertheless, results are quite similar. So, using a synthetic or an anonymized dataset might be an advantage since it allows us to protect our data and achieve very similar results.

Conclusion

In a nutshell, we have conclude that:

1. Synthetic data is a powerful tool for both privacy and data augmentation. Evaluation must be done carefully.
2. Hybrid (real+synth trained) models perform better. Possible existence of identifying synthetic features.
3. ARX and CTGAN allow us to preserve our data patterns and develop models as good as the models obtained using real data. Improving constraints definition in CTGAN would enable our model to better capture our data patterns.
4. Since our data in biased (default weight is 12%), we could have used synthetic data to reduce this bias and improve models capacity to predict both classes.

Further Improvements

Predicting Alzheimer’s Disease

1. Tune data generation GAN.

2. Evaluate diversity and quality separately because of class imbalance.

3. Classification models:
   a. different pretrained models
   b. image transformation
   c. train first on real dataset and then on synthetic dataset
   d. change proportion of synthetic and real data
   e. general parameter tuning (grid search)

Further Improvements

Predicting Mortgage Defaults

1. Generate synthetic data in order to reduce the bias in our target variable and improve our chances to correctly identify credit default.

2. Improve constrains definition in CTGAN in order to be able to better capture data patterns (e.g. Number of Children and Family Members).

3. Test other evaluation methods to assess the quality of our data (e.g. Logistic Detection)

4. Test if with other predictive model the results obtained were the same.

Thank you!

Deleted / changed slides

Further Improvements

We have identified opportunities for improvement, such as:

1	1. xxxxxxxxxxxxxxxxxxxxxxx.
2	Generate synthetic data in order to reduce the bias in our target variable and improve our chances to correctly identify credit default.
3	Improve constrains definition in CTGAN in order to be able to better capture data patterns (e.g. Number of Children and Family Members).
4	Test other evaluation methods to assess the quality of our data (e.g. Logistic Detection)
5	Test if with other predictive model the results obtained were the same.