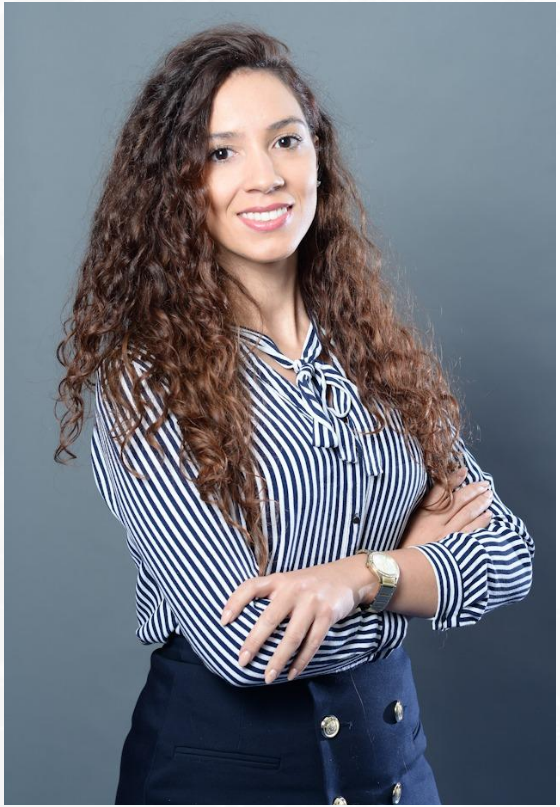
SYNTHETIC TABULAR DATA GENERATION

A GAN based approach

MAKING DATA AVAILABLE WITH PRIVACY BY DESIGN



Professional experience

Applied Maths & Data Science

From big enterprises to startups

Data Science & Architecture

Co-Founder @YData

Interests

Data Science

Time-Series

Generative Models

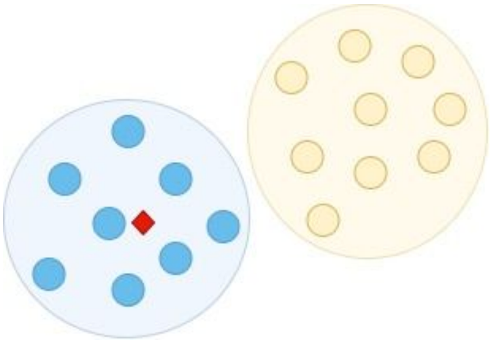
The Definition

Classify whether an animal is a cat or a dog

Generative Models

Build the model for those who look like dogs and then builds the model for those who look like cats

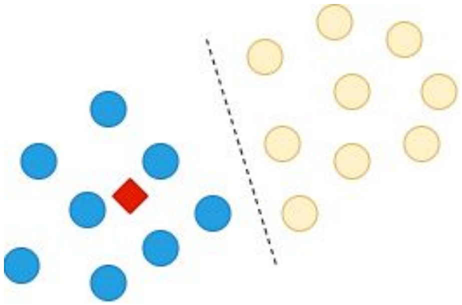
Then, matches the new animal to both cat and dog models.



Discriminative Models

Finds a decision boundary that separates cats and dogs.

Check on which side of the decision will fall the new animal.

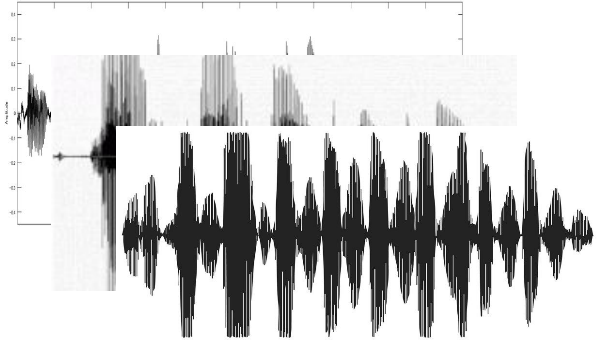


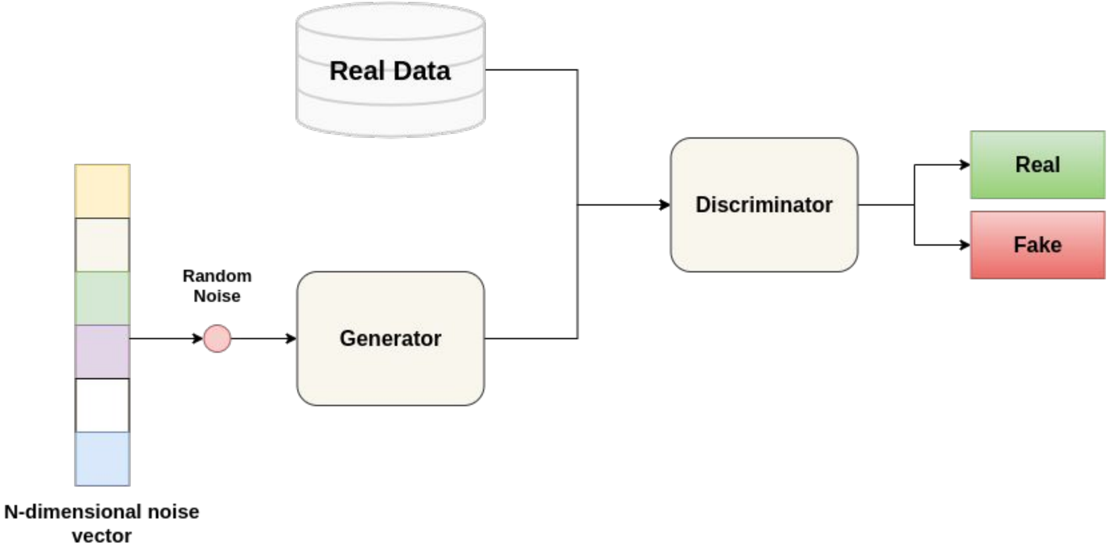
Generative Models

Deep Generative Models

- **Deep Generative Models**
  - **Implicit Density**
    - **Direct**
      - Generative Matching Moment Nets (GMM)
    - **Markov Chains**
      - **Generative Adversarial Networks**
  - **Explicit Density**
    - **Markov Chains**
      - Boltzmann-machines (BM)
    - **Variational**
      - Variational Autoencoder (VAE)

Generative Adversarial Networks (GANs)





Generative Adversarial Networks (GANs)

Human Faces Generation

[This person doesn't exist]

From Human to Anime

[Selfie to Anime]

[Github - taki0112/UGATIT]

Pix2Pix

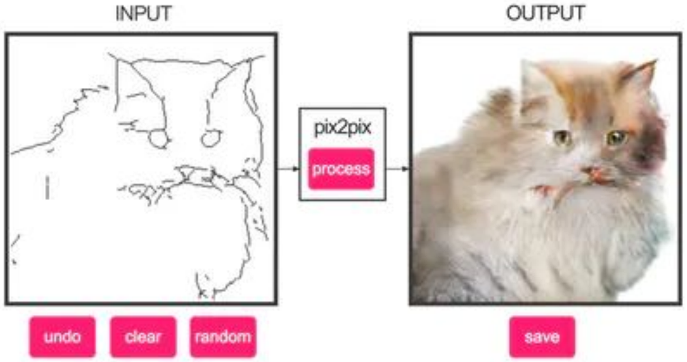
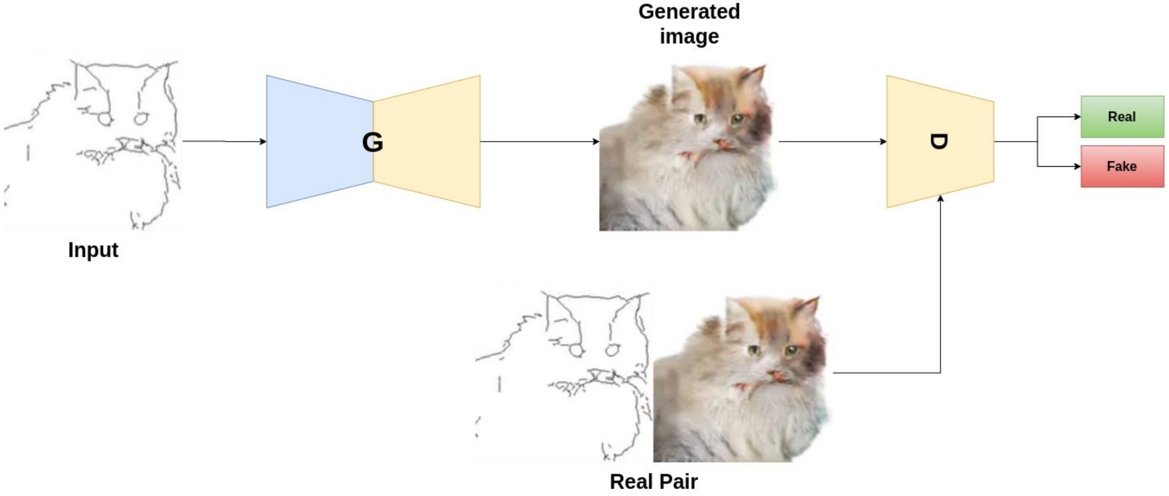
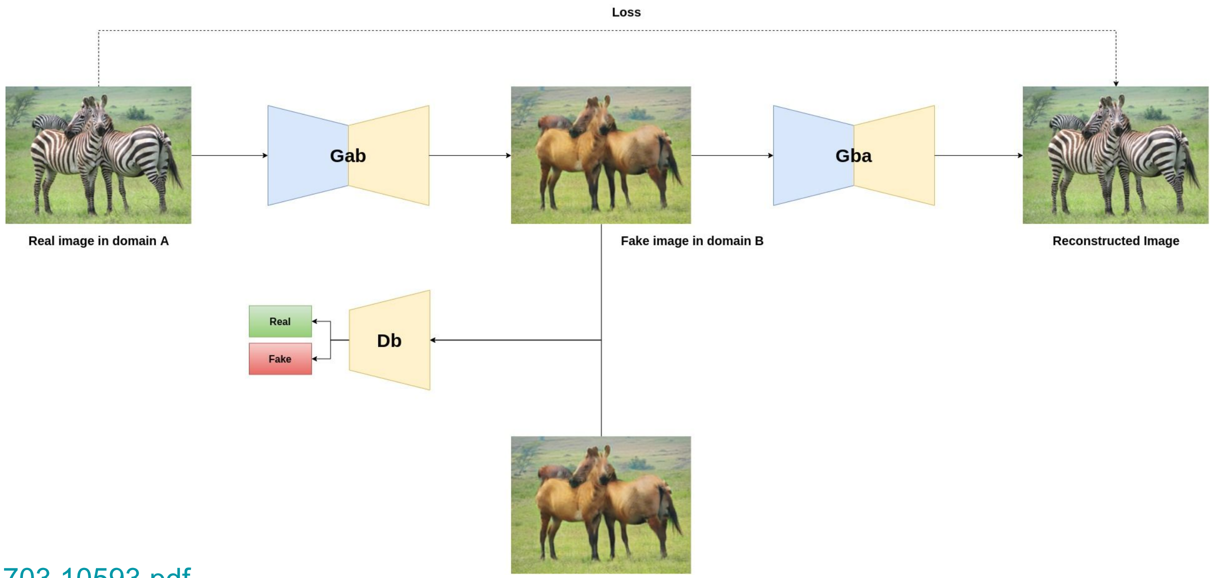


Image-to-image translation



https://arxiv.org/abs/1611.07004

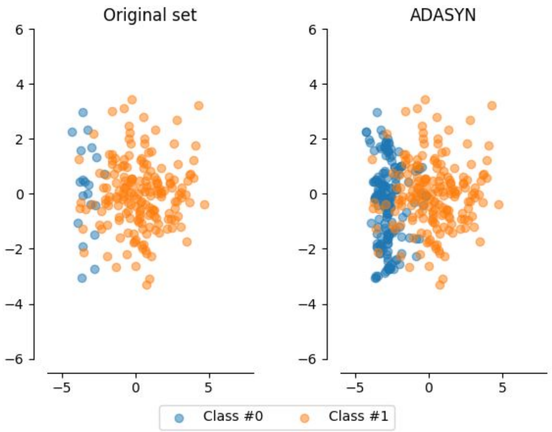
CycleGAN



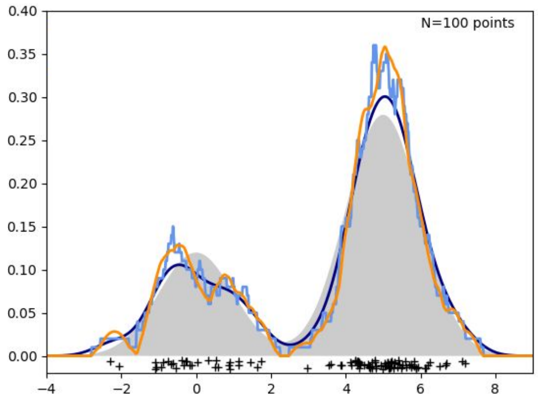
https://arxiv.org/pdf/1703.10593.pdf

But what about Tabular data?

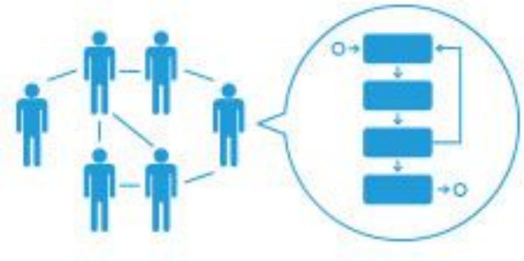
What is Synthetic data?



Oversampling methods



Multivariate statistical methods



Agent-based simulation

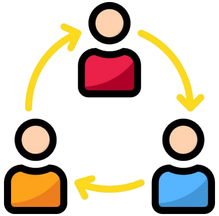
Why Synthetic data?

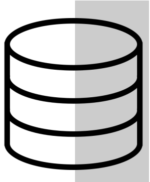
YDATA

GDPR

The General Data Protection

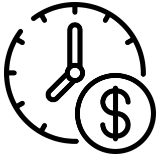
CALIFORNIA CONSUMER PRIVACY ACT



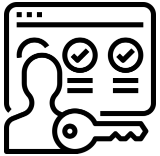


Lack of data

Imbalanced datasets



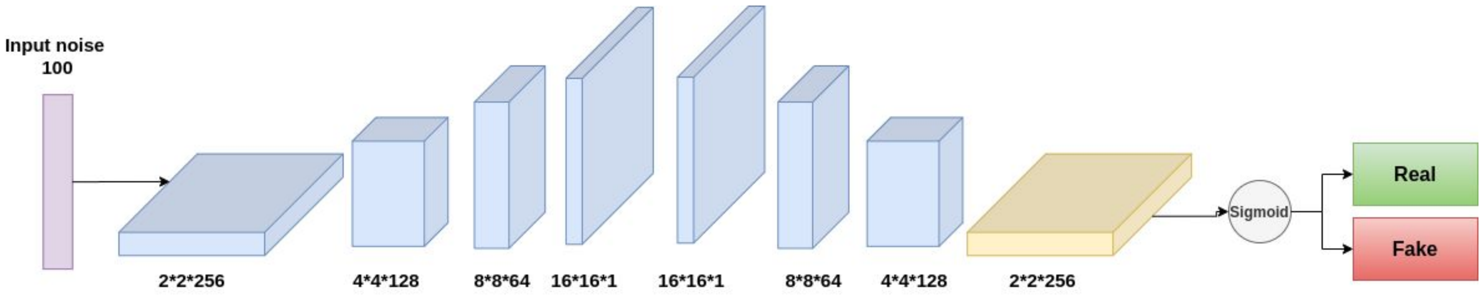
Data **acquisition** and **labelling**



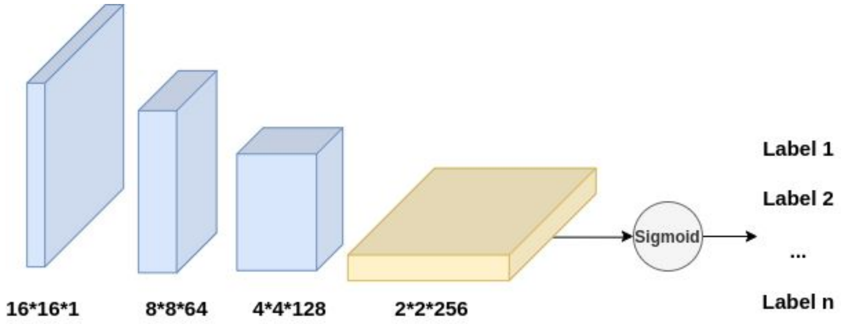
Fast access

DCGAN

Deconvolution and Convolution process



Auxiliary classifier



WGAN - Wasserstein GAN

Wasserstein GAN vs Vanilla GAN differences

- Introduction of a new loss function, based on Wasserstein distance
- Discriminator output is no longer the probability of a record being real or not, but rather a score in the domain
- The optimization problem constrains the discriminator to be a -lipschitz function
- Use of an alternative optimizer, RMSProp.

Vanilla GAN loss

$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Wasserstein loss

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} \big[\|x - y\| \big]$$

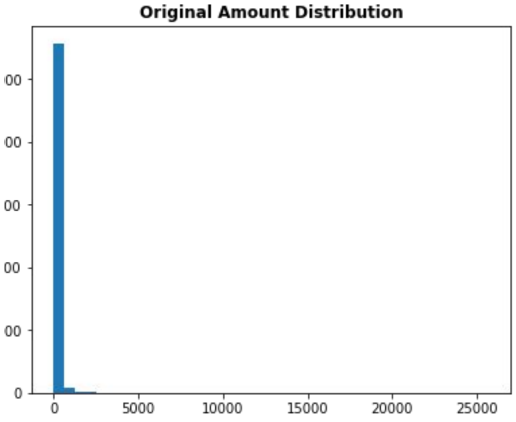
Synthetic Credit Fraud data

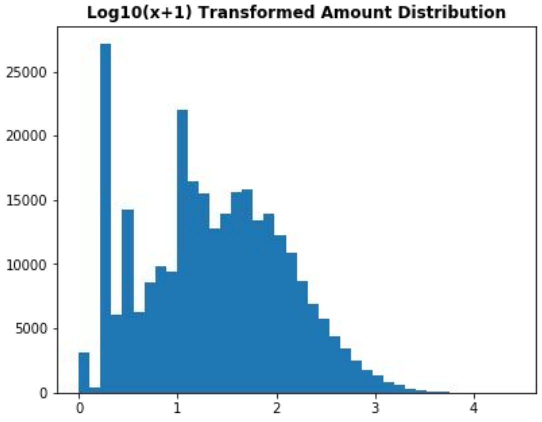
Where can you find the dataset: Kaggle Credit Fraud

Highly imbalanced classes

Non fraudulent event	284315
Fraudulent events	492
Total	284807

Presence of highly skewed variables





Synthetic Credit Fraud data

Vanilla GAN specification

Generator

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 32)]	0
dense (Dense)	(None, 128)	4224
dense_1 (Dense)	(None, 256)	33024
dense_2 (Dense)	(None, 512)	131584
dense_3 (Dense)	(None, 30)	15390

Discriminator

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 30)]	0
dense_4 (Dense)	(None, 512)	15872
dense_5 (Dense)	(None, 256)	131328
dense_6 (Dense)	(None, 128)	32896
dense_7 (Dense)	(None, 1)	129

Training parameters:

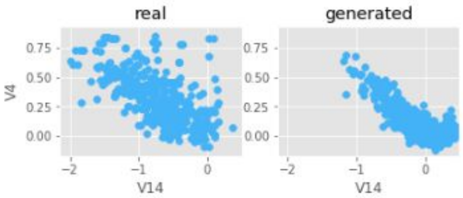
Batch size: 128

Epochs num: 5000

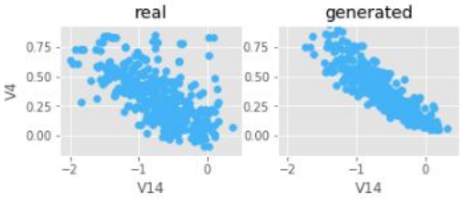
Gen LR: 5e-4

Disc LR: 5e-4

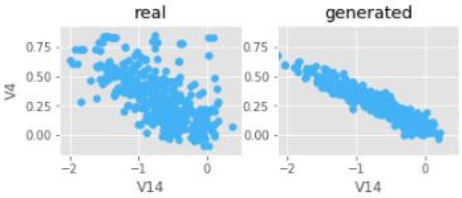
```
Step: 300 of 501.
Losses: G, D Gen, D Real, Xgb: 1.0937, 0.5411, 0.4982, 0.9878
D Real - D Gen: -0.0429
```




```
Step: 400 of 501.
Losses: G, D Gen, D Real, Xgb: 0.9822, 0.6214, 0.7255, 0.9898
D Real - D Gen: 0.1041
```




```
Step: 500 of 501.
Losses: G, D Gen, D Real, Xgb: 0.9689, 0.6660, 0.6171, 0.9776
D Real - D Gen: -0.0488
```



Synthetic Credit Fraud data

Conditional GAN specification

Generator

Layer (type)	Output Shape	Param #	Connected to
input_6 (InputLayer)	[(None, 32)]	0	
input_7 (InputLayer)	[(None, 1)]	0	
concatenate_2 (Concatenate)	(None, 33)	0	input_6[0][0] input_7[0][0]
dense_16 (Dense)	(None, 128)	4352	concatenate_2[0][0]
dense_17 (Dense)	(None, 256)	33024	dense_16[0][0]
dense_18 (Dense)	(None, 512)	131584	dense_17[0][0]
dense_19 (Dense)	(None, 30)	15390	dense_18[0][0]
concatenate_3 (Concatenate)	(None, 31)	0	dense_19[0][0] input_7[0][0]

Discriminator

Layer (type)	Output Shape	Param #
input_8 (InputLayer)	[(None, 31)]	0
dense_20 (Dense)	(None, 512)	16384
dense_21 (Dense)	(None, 256)	131328
dense_22 (Dense)	(None, 128)	32896
dense_23 (Dense)	(None, 1)	129

Training parameters:

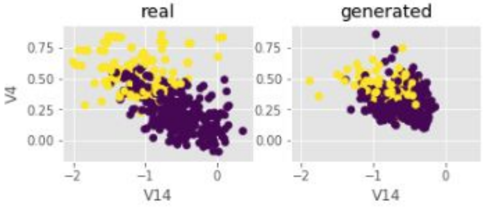
Batch size: 128

Epochs num: 5000

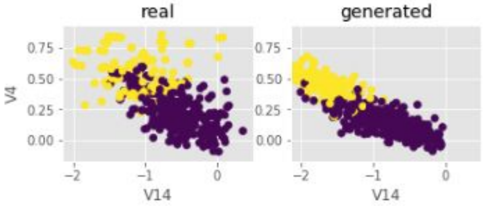
Gen LR: 5e-4

Disc LR: 5e-4

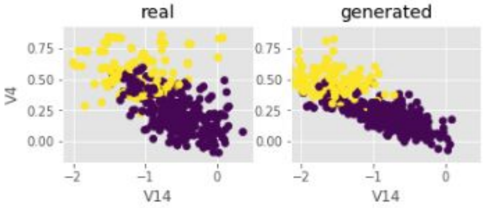
```
Step: 200 of 501.
Losses: G, D Gen, D Real, Xgb: 1.0783, 0.6315, 0.5332, 0.9898
D Real - D Gen: -0.0983
```



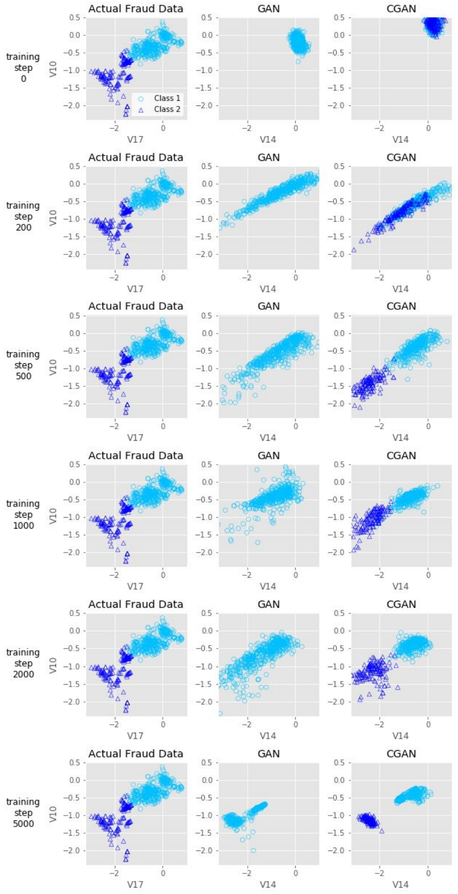

```
Step: 300 of 501.
Losses: G, D Gen, D Real, Xgb: 0.8913, 0.7646, 0.6432, 0.9837
D Real - D Gen: -0.1213
```




```
Step: 400 of 501.
Losses: G, D Gen, D Real, Xgb: 1.0660, 0.5937, 0.6696, 0.9837
D Real - D Gen: 0.0759
```



Synthetic Credit Fraud data



Generated vs Original dataset statistics

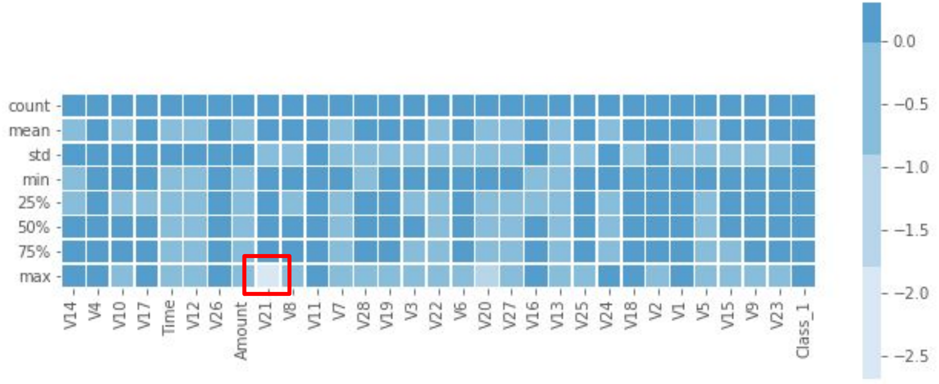
Training parameters:

Batch size: 128

Epochs num: 500

Gen LR: 5e-4

Disc LR: 5e-4

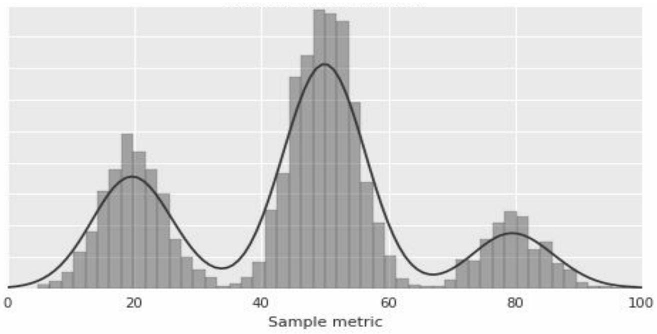


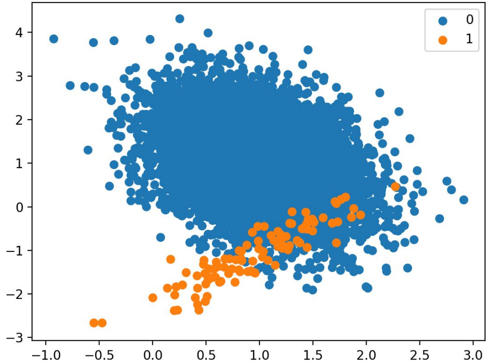
Challenges

Tabular data particular challenges

Order ID	Product	Category	Amount	Date	Country
1	Carrots	Vegetables	$4,270	1/6/2012	United States
2	Broccoli	Vegetables	$8,239	1/7/2012	United Kingdom
3	Banana	Fruit	$617	1/8/2012	United States
4	Banana	Fruit	$8,384	1/10/2012	Canada
5	Beans	Vegetables	$2,626	1/10/2012	Germany
6	Orange	Fruit	$3,610	1/11/2012	United States
7	Broccoli	Vegetables	$9,062	1/11/2012	Australia
8	Banana	Fruit	$6,906	1/16/2012	New Zealand
9	Apple	Fruit	$2,417	1/16/2012	France
10	Apple	Fruit	$7,431	1/16/2012	Canada
11	Banana	Fruit	$8,250	1/16/2012	Germany
12	Broccoli	Vegetables	$7,012	1/18/2012	United States
13	Carrots	Vegetables	$1,903	1/20/2012	Germany

No.	Attribute	Original Type	Range	Type Used
1	age	continuous	17–90	categorical
2	workclassge	categorical	1–8	categorical
3	final weight (fnlwgt)	continuous	12,285–1,484,705	numeric
4	education	categorical	1–16	categorical
5	education-num	continuous	1–16	categorical
6	marital-status	categorical	1–7	categorical
7	occupation	categorical	1–14	categorical
8	relationship	categorical	1–6	categorical
9	race	categorical	1–5	categorical
10	sex	categorical	1–2	categorical
11	capital-gain	continuous	0–99,999	numeric
12	capital-loss	continuous	0–4356	numeric
13	hours-per-week	continuous	1–99	categorical
14	native-country	continuous	1–41	categorical
15	class	categorical	1–2	categorical





Things you can explore

GANs hyperparameters tuning and improved stability

- Hyperparameters tuning - Open-sourced Google's Vizier
- Introducing Gradient Penalty - check this and this article
- Coevolution of Generative Adversarial Network

Avoiding mode collapse

- Packing - PacGAN
- Defining the generator objective with respect to unrolled optimization of the discriminator - Unrolled GAN

GANs for missing data imputation

- Missing data imputation - GAIN

(RE)CREATING ELECTROCARDIOGRAMS

NEED

Data from patients

Develop a model to identify arrhythmias

PROBLEM

Data is sensitive and private

Data is scarce and dirty

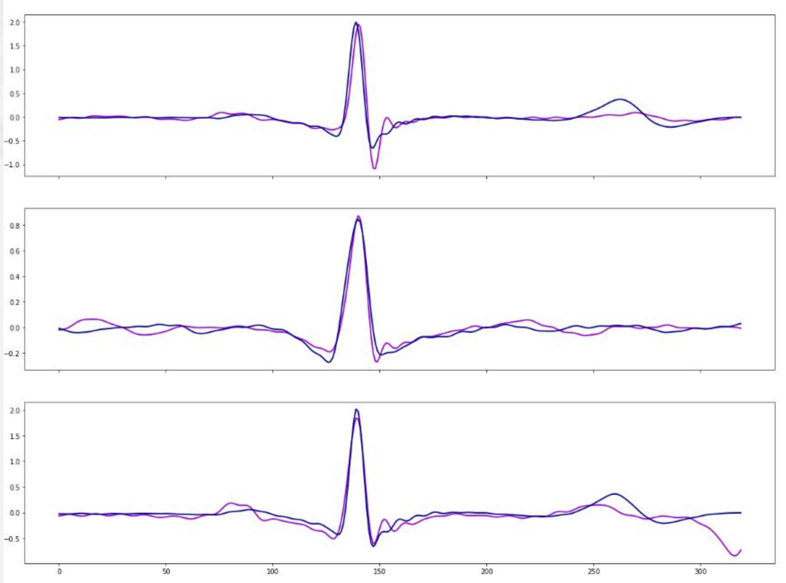
Data is unbalanced and unlabelled

SOLUTION

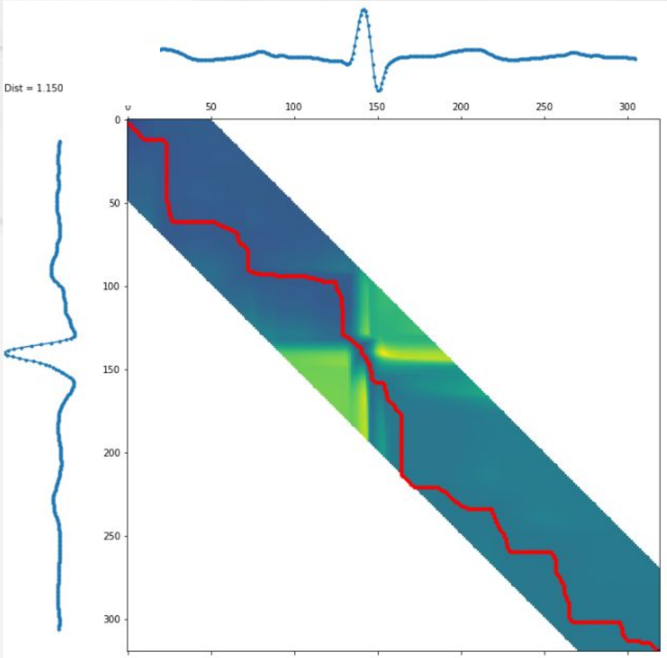
Creation of synthetic ECG from small amounts of data that can be used as the reals ones, without concerns around privacy and security

(RE)CREATING ELECTROCARDIOGRAMS

Real vs. AI generated ECG



Time (s)



Dist: 1.15

Total patients

48

Number of heartbeats

~100,000

Training set:

~65,000 (65%)

Validation set:

~20,000 (20%)

Test set:

~15,000 (15%)

DSPT
DATA SCIENCE PORTUGAL

YDATA

Fabiana Clemente

fabiana.clemente@ydata.ai

@fab_clemente

fabiana.clemente

MAKING DATA AVAILABLE WITH PRIVACY BY DESIGN