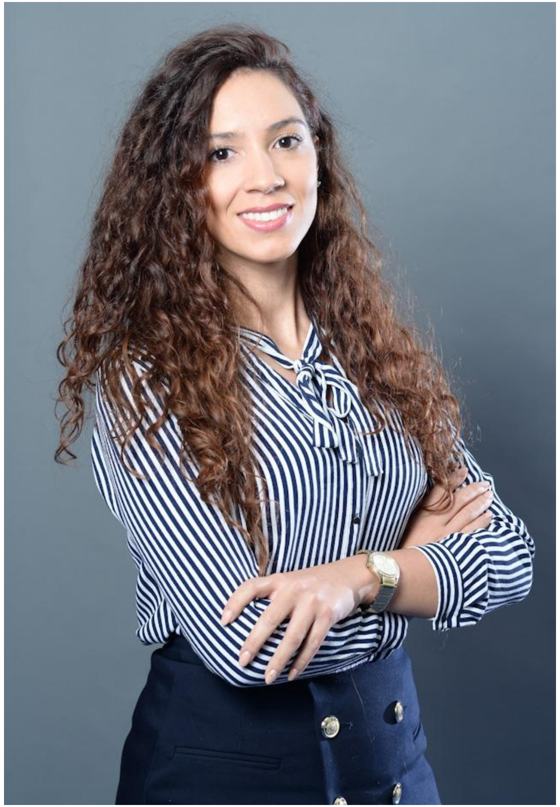
YDATA

Improved data for AI

Deep Learning
Sessions Lisboa

Synthetic tabular data generation

A GAN based approach



Professional experience

Applied Maths & Data Science

From big enterprises to startups

Data Science & Architecture

Co-Founder @YData

Interests

Data Science

Time-Series

Generative Models

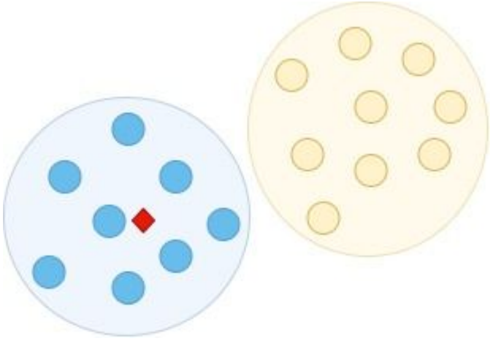
The Definition

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

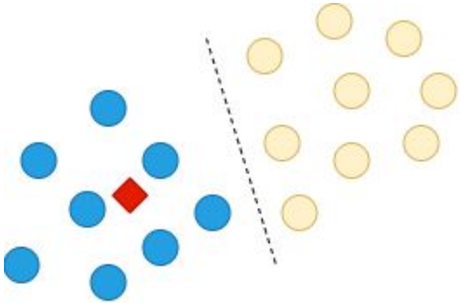
Then, matches the new animal to both cat and dog models.



Discriminative Models

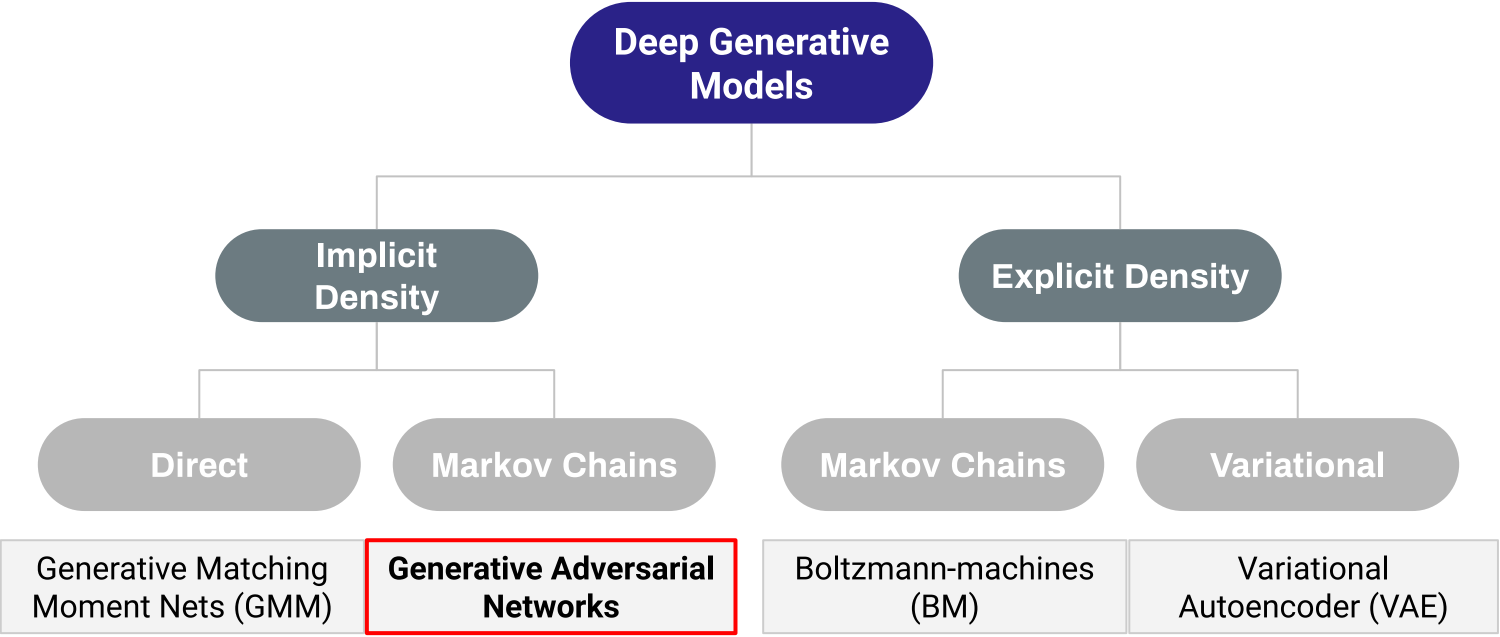
Finds a decision boundary that separates cats and dogs.

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

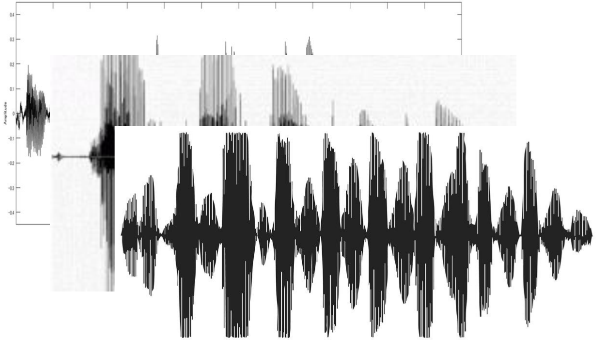


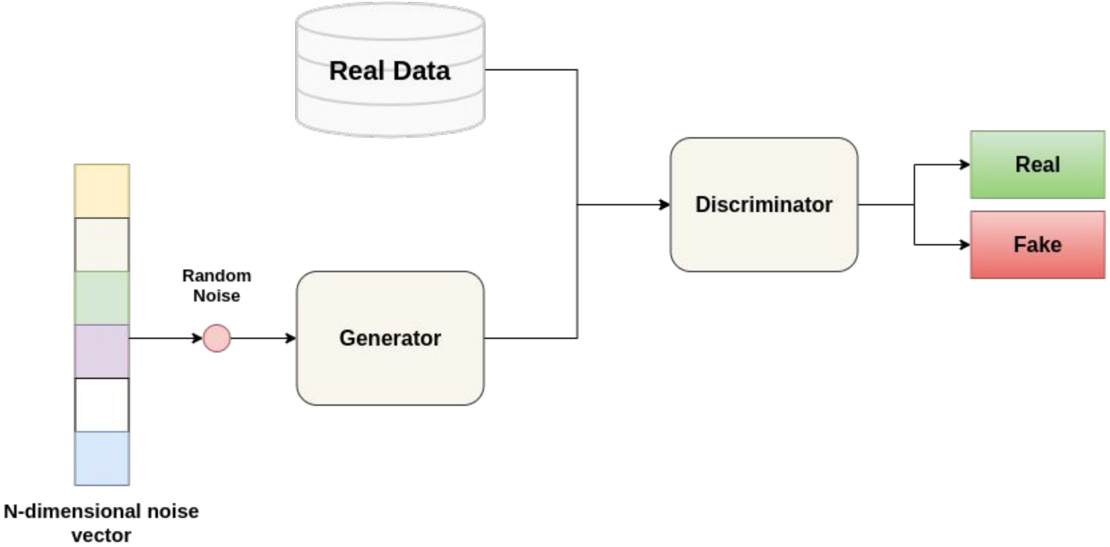
Generative Models

Deep Generative Models



Generative Adversarial Networks (GANs)



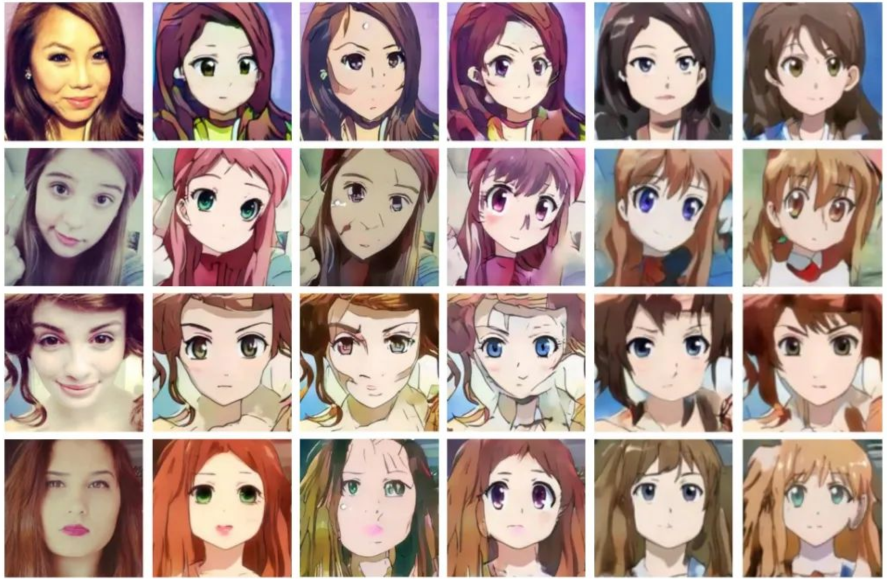


Generative Adversarial Networks (GANs)

Human Faces Generation

This person doesn't exist

From Human to Anime



Selfie to Anime

Github - taki0112/UGATIT

Pix2Pix

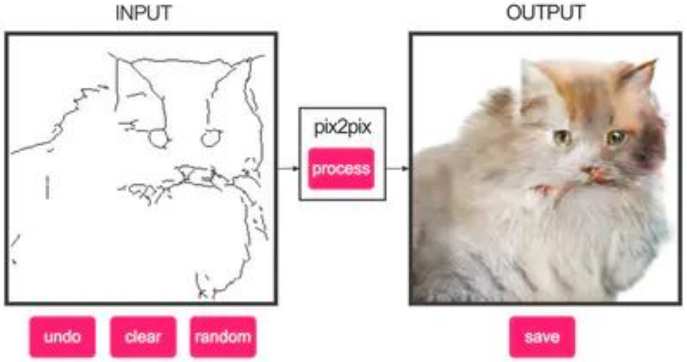
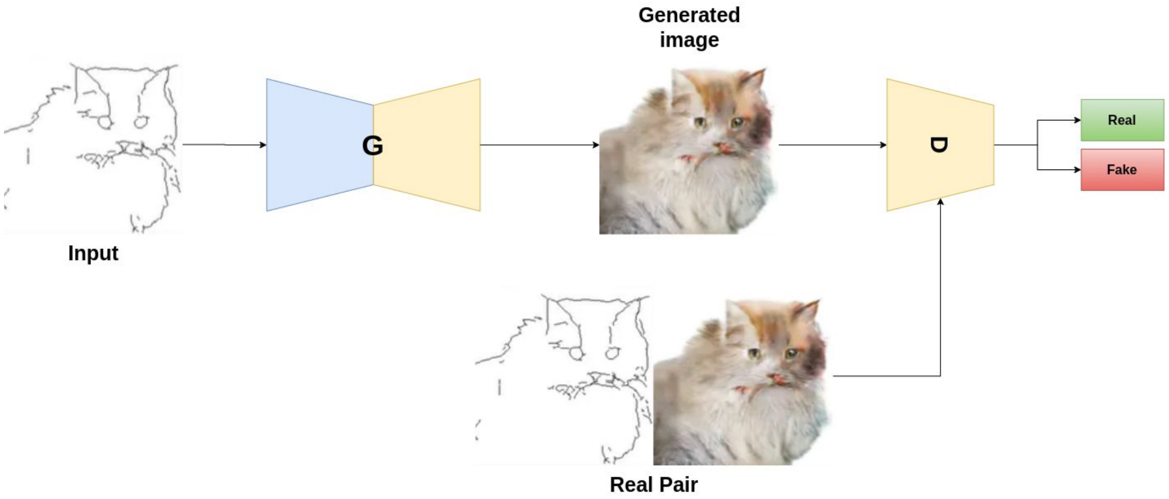
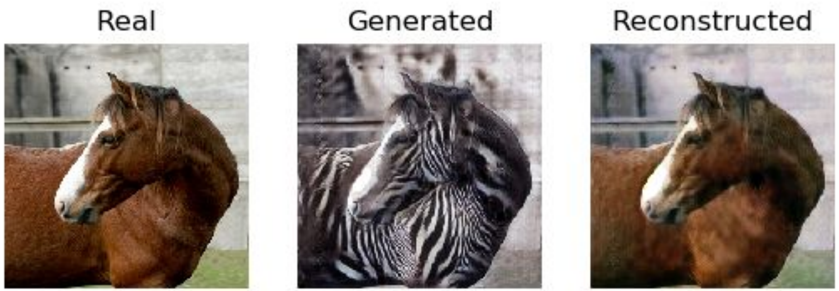


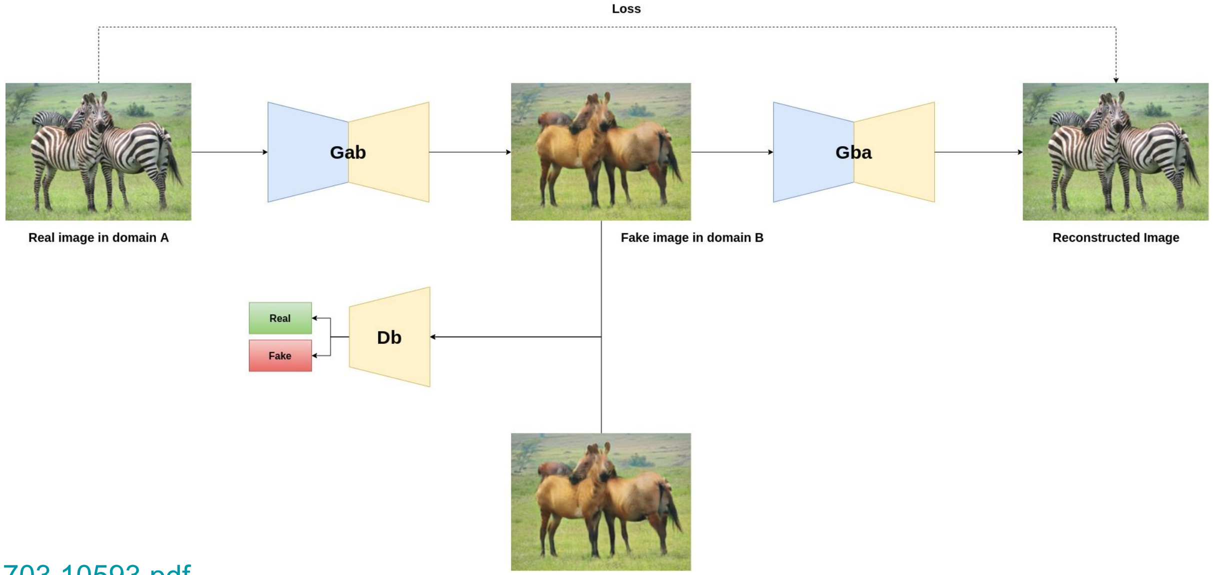
Image-to-image translation



https://arxiv.org/abs/1611.07004

CycleGAN

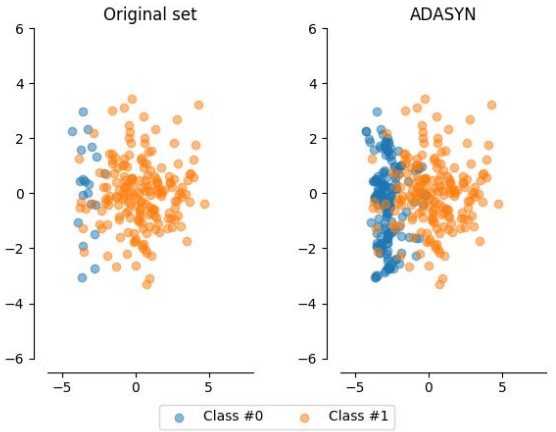




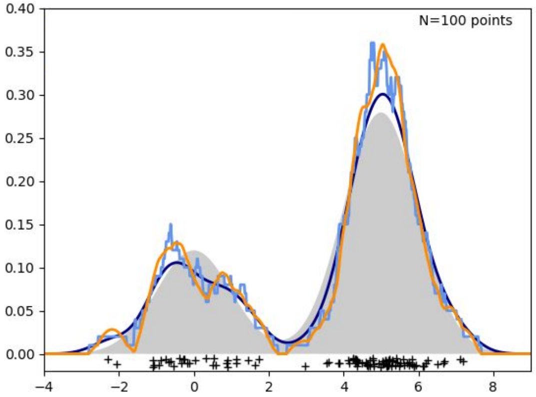
https://arxiv.org/pdf/1703.10593.pdf

But what about Tabular data?

What is Synthetic data?



Oversampling methods

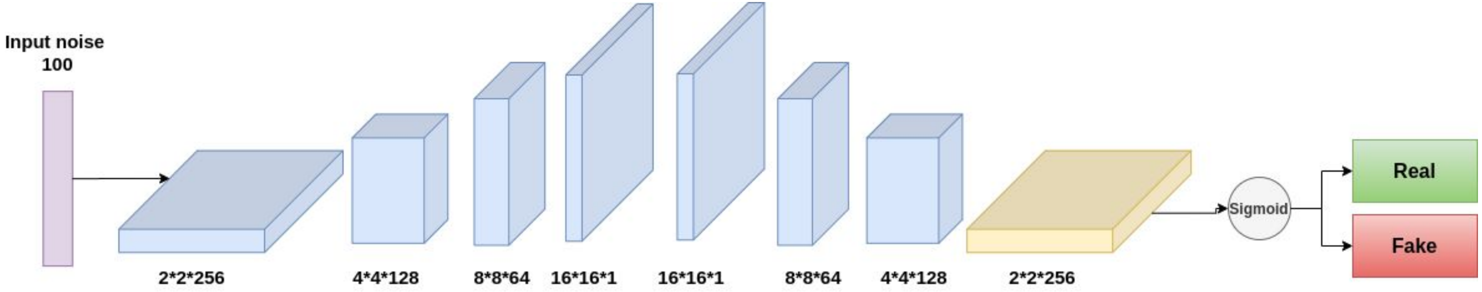


Multivariate statistical methods

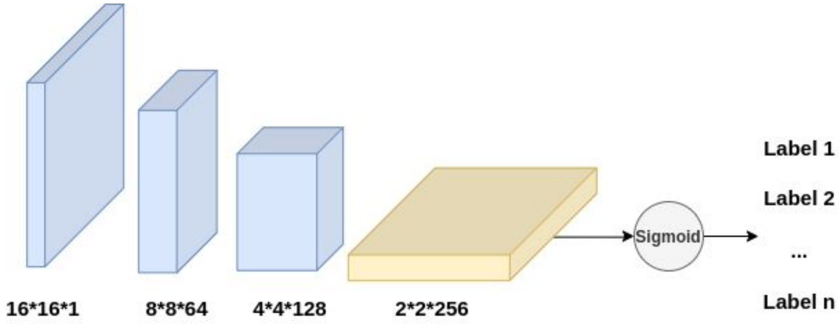
Agent-based simulation

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

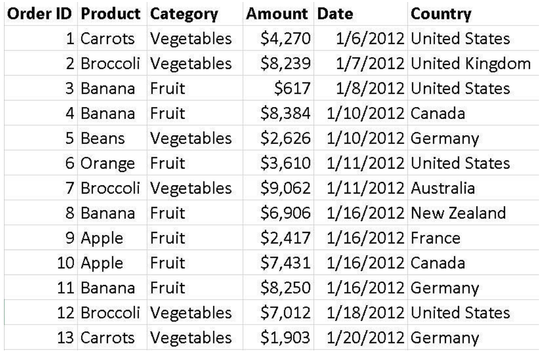
$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Wasserstein loss

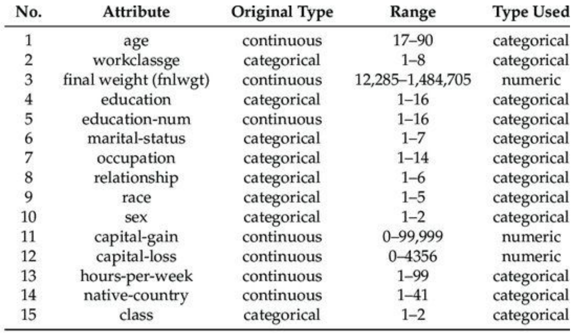
$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma\in\Pi(\mathbb{P}_r,\mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\,\|x-y\|\,\big]$$

Challenges

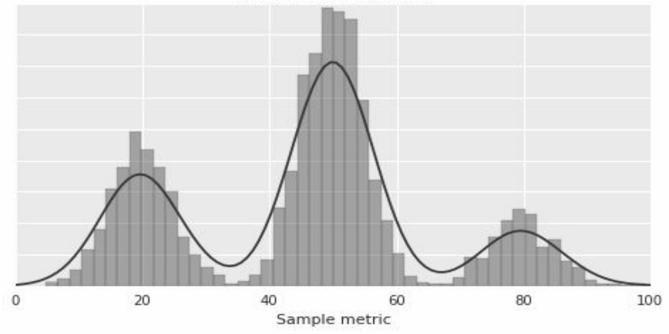
Tabular data particular challenges

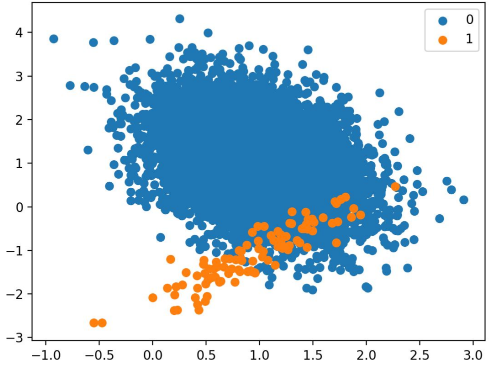


Order ID	Product	Category	Amount	Date	Country
1	Carrots	Vegetables	$4,270	1/6/2012	United States
2	Broccoli	Vegetables	$8,239	1/7/2012	United Kingdom
3	Banana	Fruit	$617	1/8/2012	United States
4	Banana	Fruit	$8,384	1/10/2012	Canada
5	Beans	Vegetables	$2,626	1/10/2012	Germany
6	Orange	Fruit	$3,610	1/11/2012	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

GitHub

The GAN Playground

Thank you!

We help adopters of AI to **improve** and **generate high quality** data so they can become the tomorrow's **industry leaders**

Fabiana Clemente

fabiana.clemente@ydata.ai

@fab_clemente

fabiana.clemente