elo7

Software
FORA DE SÉRIE
elo7

2019

elo7

Buyers
9 million
active buyers

elo7

Sellers
30 thousand
active sellers

2019

enterpreneurship
innovation
artisanship
creativity
technology
humanization
personalization
handcraft

Since maternity, Elo7 has been involved in the growth of many babies and children

1 out of 4 brides organize their wedding with Elo7

What it is like to be a data scientist at Elo7?

Creativity → Research + Product development
Study groups → Research + Product development
Autonomy → Research + Product development
Teamwork → Research + Product development

2019

Igor Bonadio, PhD
Computer Science
Lead Data Scientist @ Elo7

Cinthia Tanaka, PhD
Applied Math
Data Scientist @ Elo7

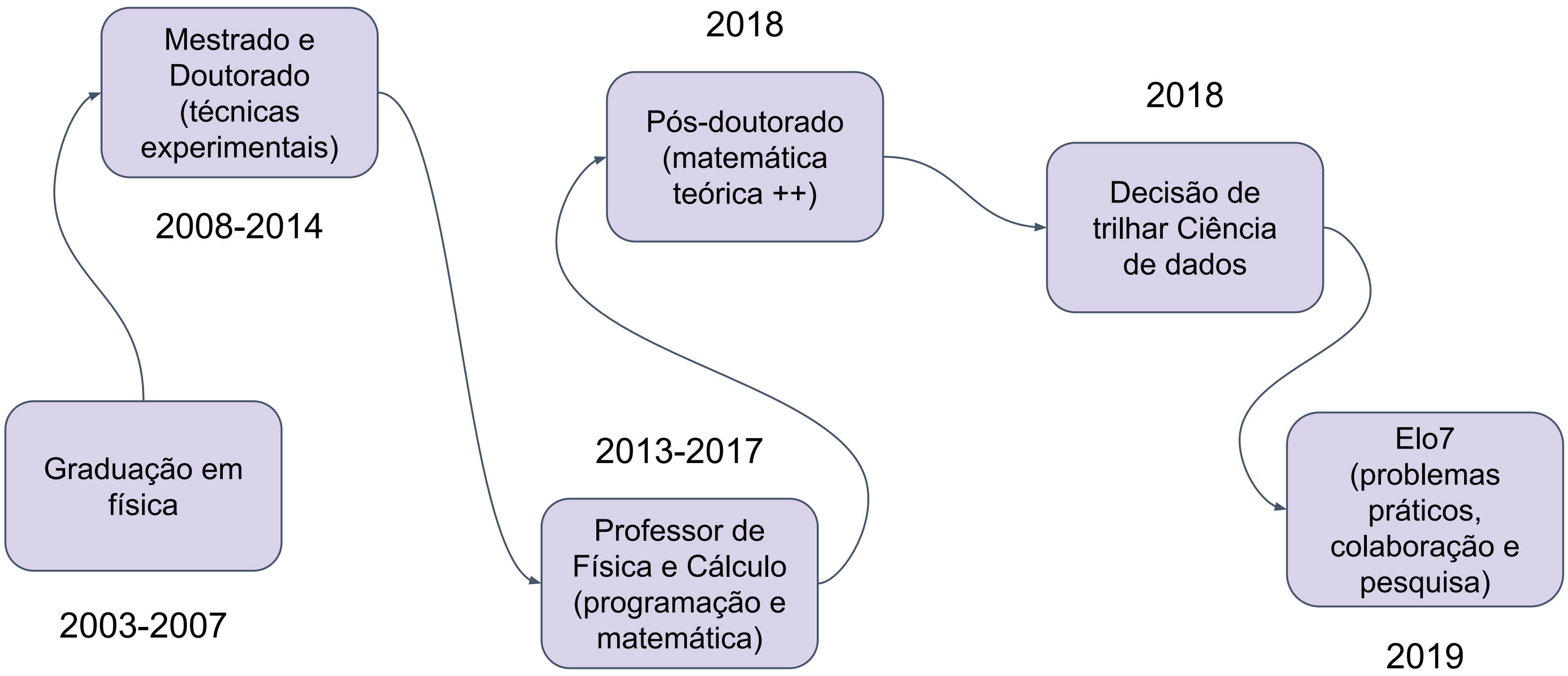
Samara Alves, MSc.
Statistics
Data Scientist @ Elo7

Junior A. Koch, PhD
Mathematical Physics
Data Scientist @ Elo7

Renato Cordeiro
Computer Science
ML Engineer @ Elo7

About me

elo7



2019

elo7

www.elo7.com.br

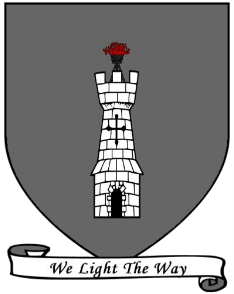
https://elo7.gupy.io/

Engenheir@ de Software Backend	São Paulo-SP	Full-time employee
Engenheiro de Software - DevOps	São Paulo-SP	Full-time employee
Engenheiro de Software Pleno (Nativo)	São Paulo-SP	Full-time employee
Engenheiro de Software Pleno (Nativo) - Remoto	São Paulo-SP	Full-time employee
Não encontrou uma vaga com o seu perfil? Cadastre-se aqui e faça parte do nosso banco!	São Paulo-SP	Talent pool

Modelos Geradores

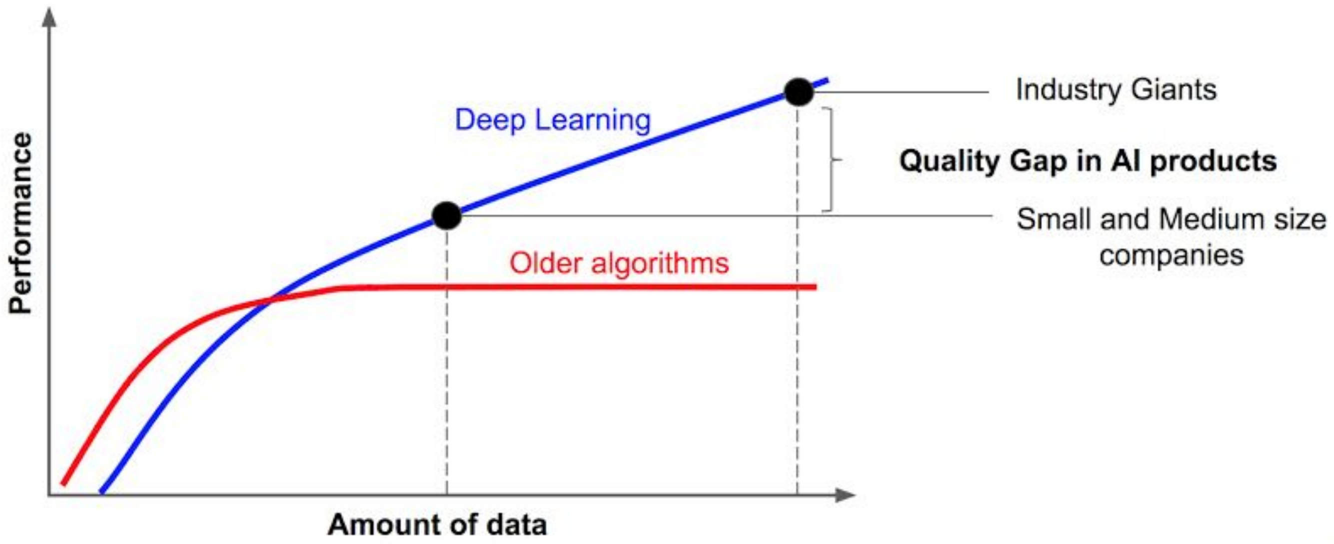
2019

Team Hightower
PhD. Junior A. Koch
Data Scientist @ Elo7



I need data, lots of data

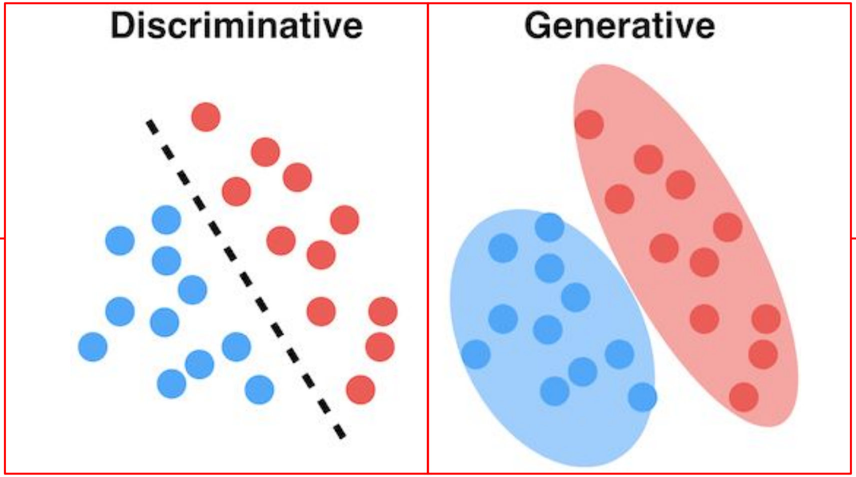
elo7



2019

O que é um modelo gerador?

elo7



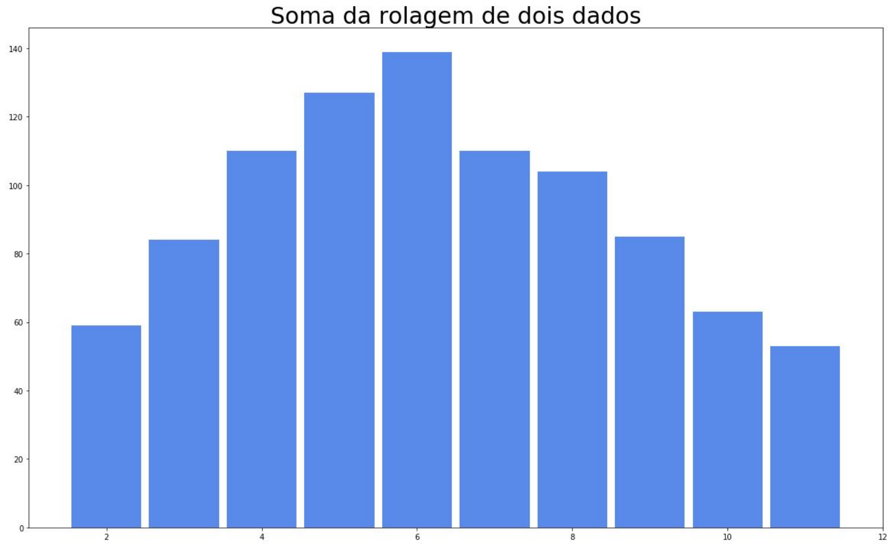
Machine learning tradicional: temos os dados e queremos uma função ou regra que categoriza os dados.

Abordagem de geradores: estamos preocupados em aprender a distribuição dos dados de cada umas das classes presentes.

2019

O que é um modelo gerador? Exemplo

elo7

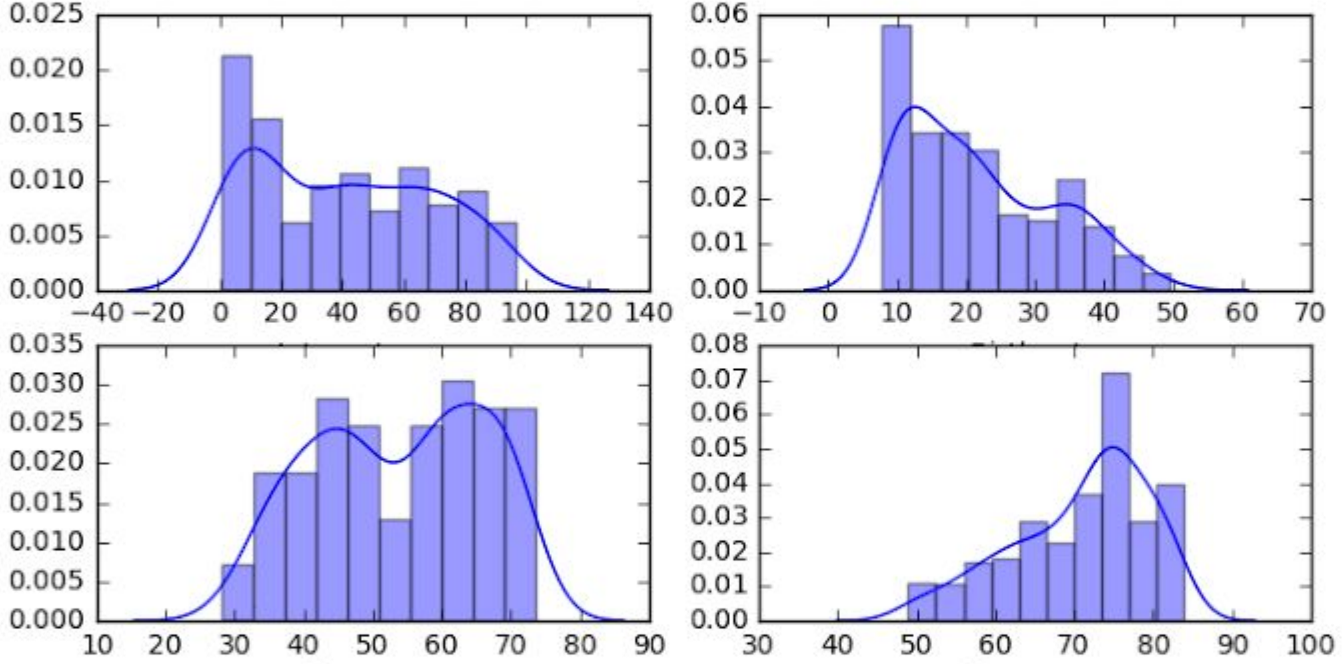


2019

Distribuição dos dados

Preço dos produtos

Preço do frete



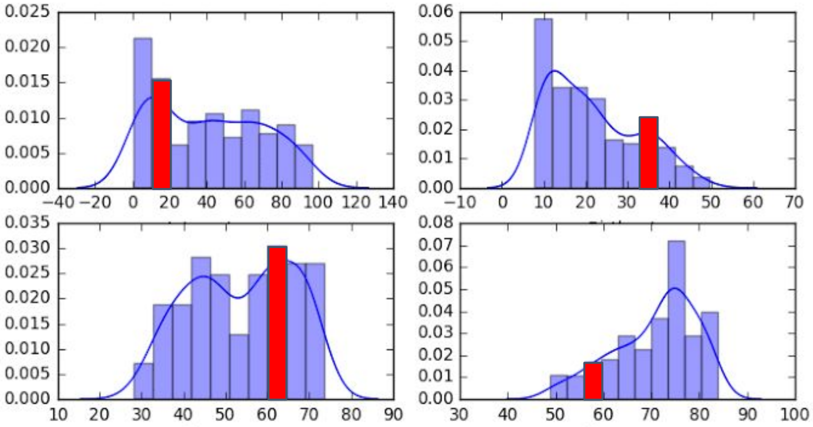
Idade dos compradores

Tempo mensal no site

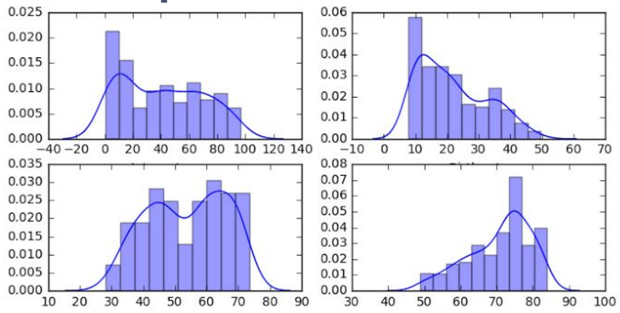
2019

Distribuição dos dados

Amostragem de novos dados



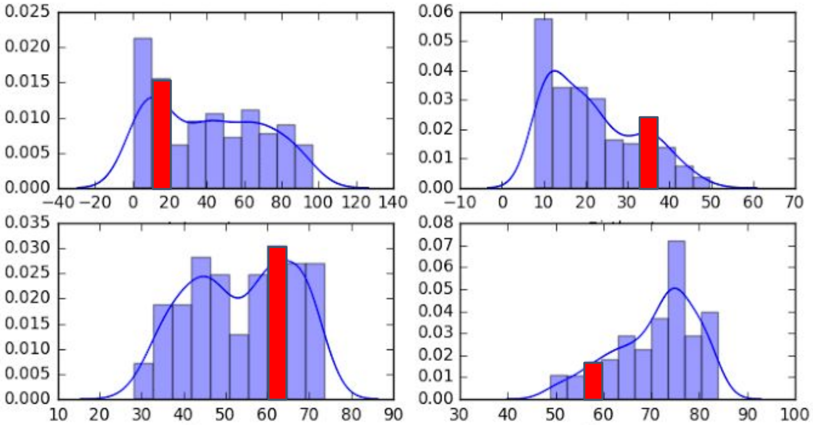
$$p_{data}(X) \Rightarrow x \sim p_{data}(X)$$



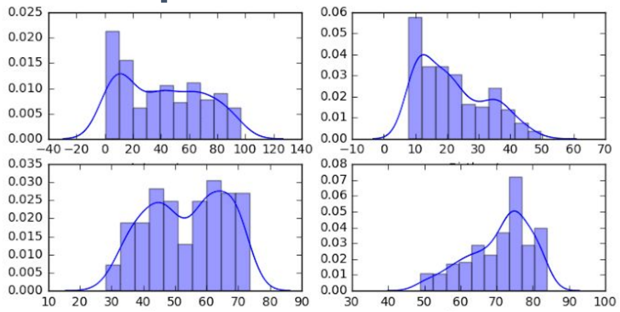
Dados originais

Distribuição dos dados

Amostragem de novos dados



$$p_{data}(X) \Rightarrow x \sim p_{data}(X, \boxed{Y})$$



Dados originais

Tipos de modelos geradores

elo7

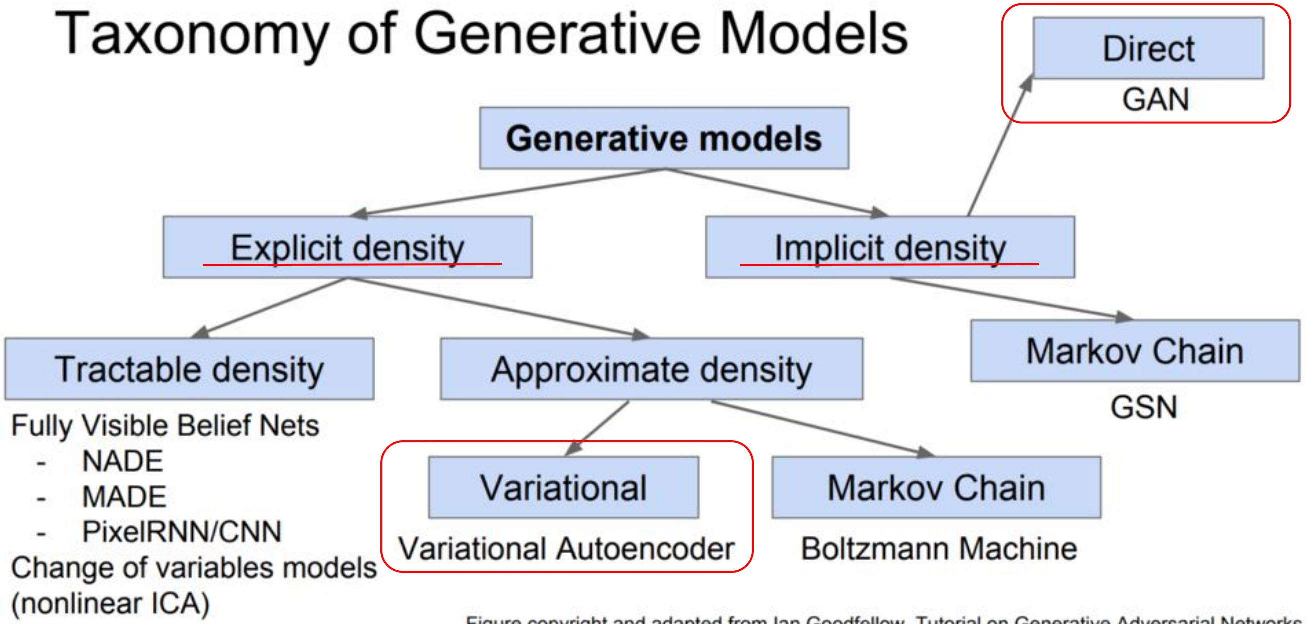
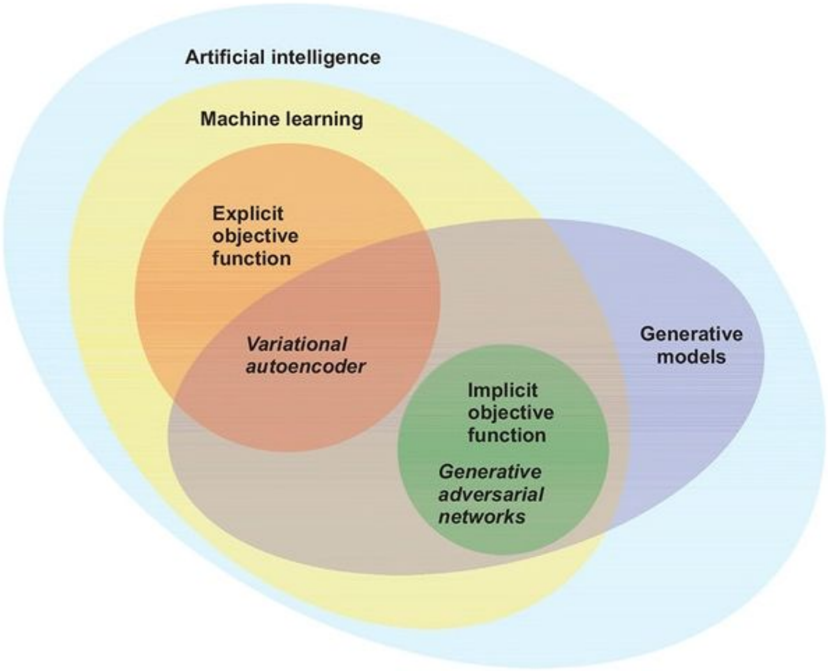


Figure copyright and adapted from Ian Goodfellow, Tutorial on Generative Adversarial Networks, 2017.

2019

Tipos de modelos geradores

elo7



2019

Modelos Geradores são difíceis

elo7

Pipeline de aprendizado comum no estudo de machine learning:

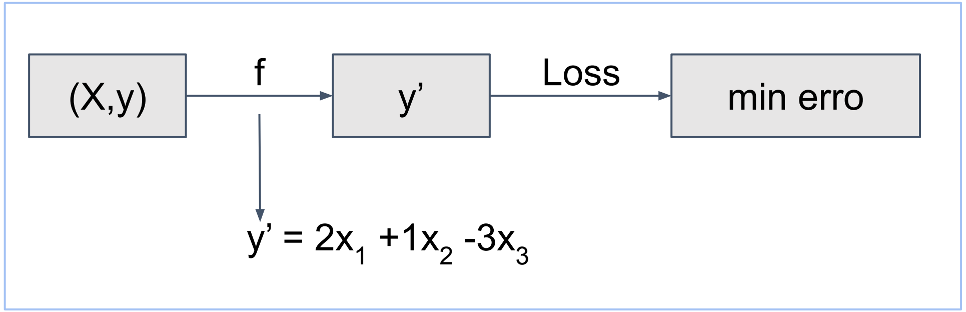
1. Basics
    1. Supervised,unsupervised,reinforcement
    2. Bias-variance trade-off
    3. Overfitting, underfitting
2. Gradient descent:
3. Linear discriminant analysis (LDA)
4. Principal Component Analysis(PCA)
5. Learning Vector Quantization (LVQ)
6. Regularization methods
7. Kernel smoothing methods
8. Ensemble learning
9. Ordinary least squares
10. ...

Por que (na minha opinião)?

1. Exige mais modelagem
2. É difícil criar uma biblioteca standard
3. Cada aplicação apresenta suas próprias peculiaridades

2019

Associação e causa



Podemos dizer que **SE** aumentarmos x_1 e x_2 **então** y' vai aumentar também?

Lei zero de inferência causal:

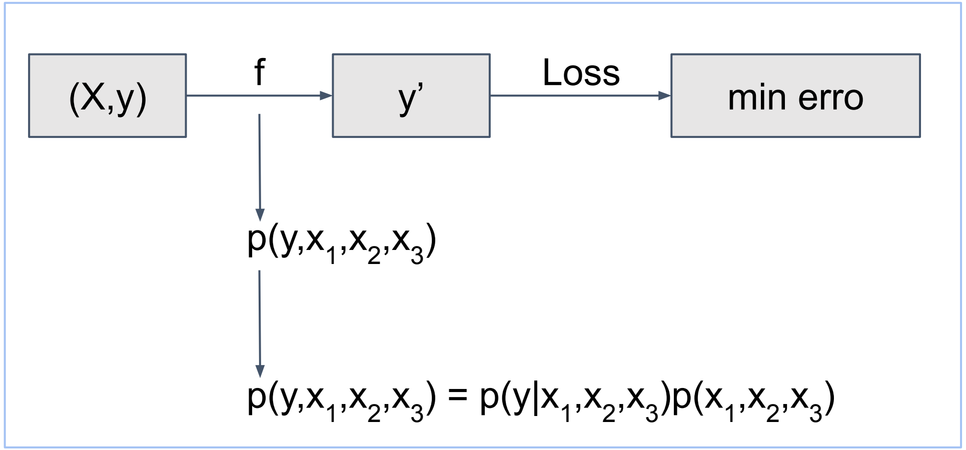
Associação e causa são diferentes.

Em ML tradicional apenas aprendemos uma função que de adequa aos dados e isso não quer dizer que podemos inferir que as features escolhidas são a causa da variável de saída.

2019

Como fazer inferência causal?

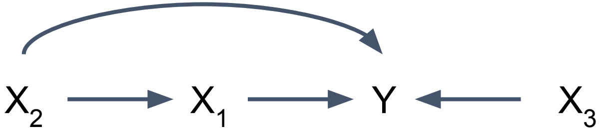
elo7



Precisamos conhecer muito bem o problema para escolher boas features e então aprender distribuições condicionais para o grafo causal escolhido.

Nota: podemos fazer tudo isso e perceber que não foi suficiente.

Exemplo de gráfico causal:

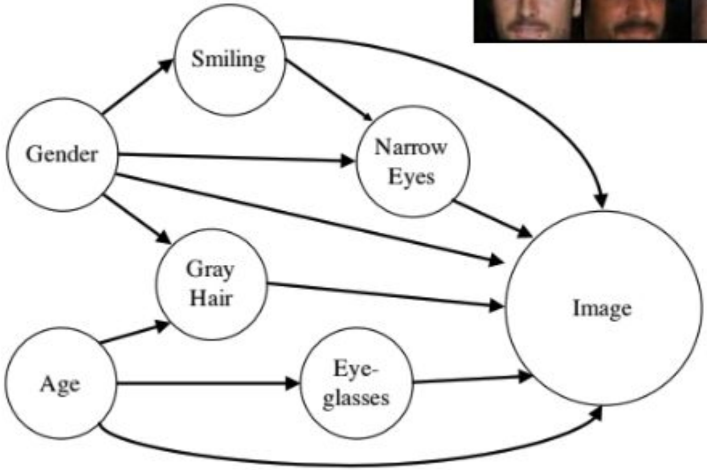


Wasserman, Larry. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

2019

Dependência entre características dos dados

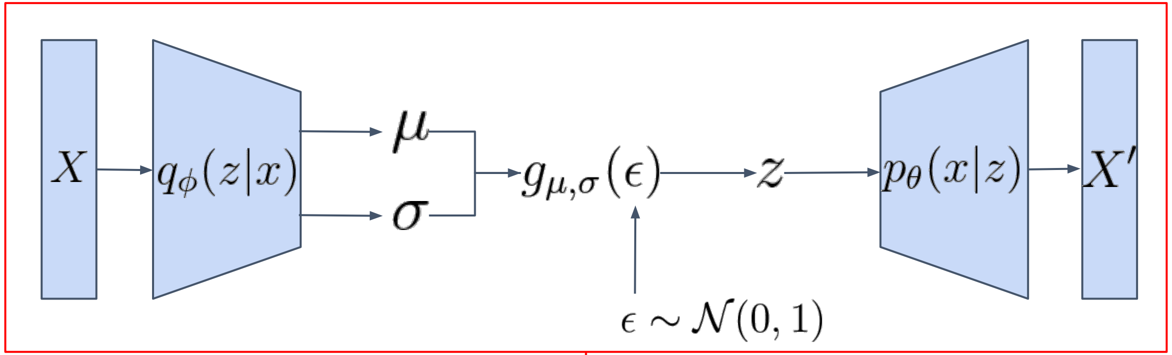
elo7



2019

Kocaoglu, Murat, et al. "Causalgan: Learning causal implicit generative models with adversarial training." *arXiv preprint arXiv:1709.02023* (2017).

Variational Autoencoder



$$Loss = \mathbb{E}\left[\log p(x|z)\right] - KL\left[q(z|x)||p(z)\right]$$

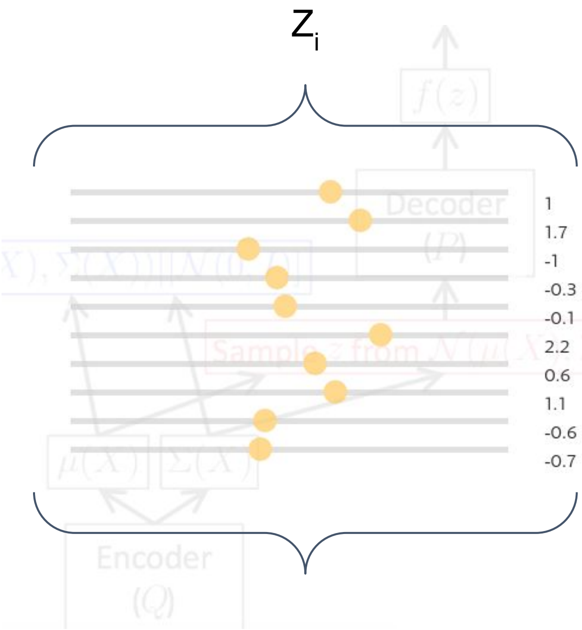
$$min \begin{cases} \mathbb{E}\left[\log p(x|z)\right] \rightarrow \text{cross entropy}(x, x') \\ KL\left[q(z|x)||p(z)\right] \rightarrow \dfrac{1}{2}\sum_n \left(e^\sigma + \mu^2 - 1 - \sigma\right) \end{cases}$$

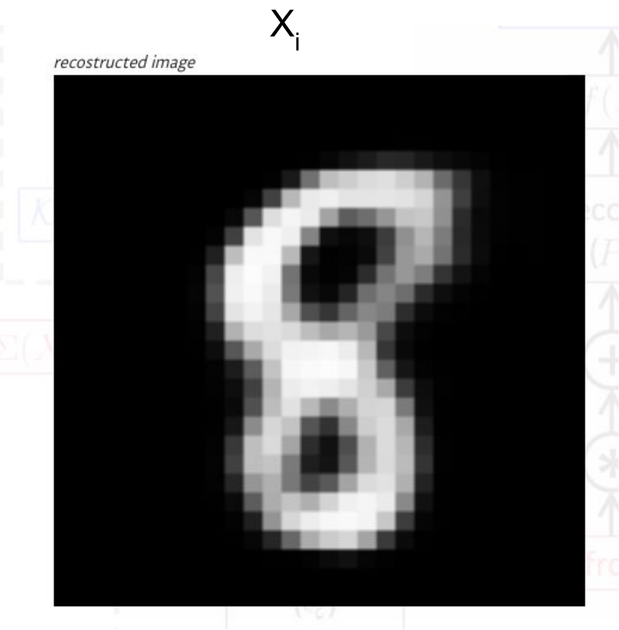
Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).

2019

Variational Autoencoder

elo7





2019

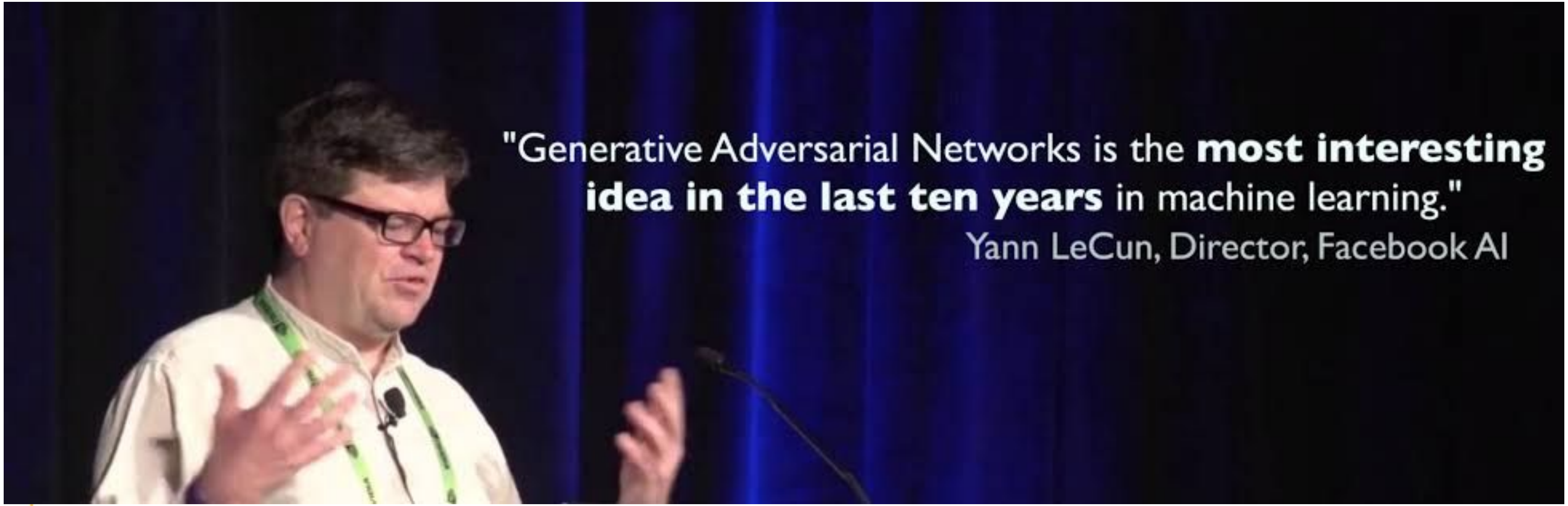
GANs e Treino Adversarial

2019

Team Hightower
PhD. Junior A. Koch
Data Scientist @ Elo7

Por que GANs?

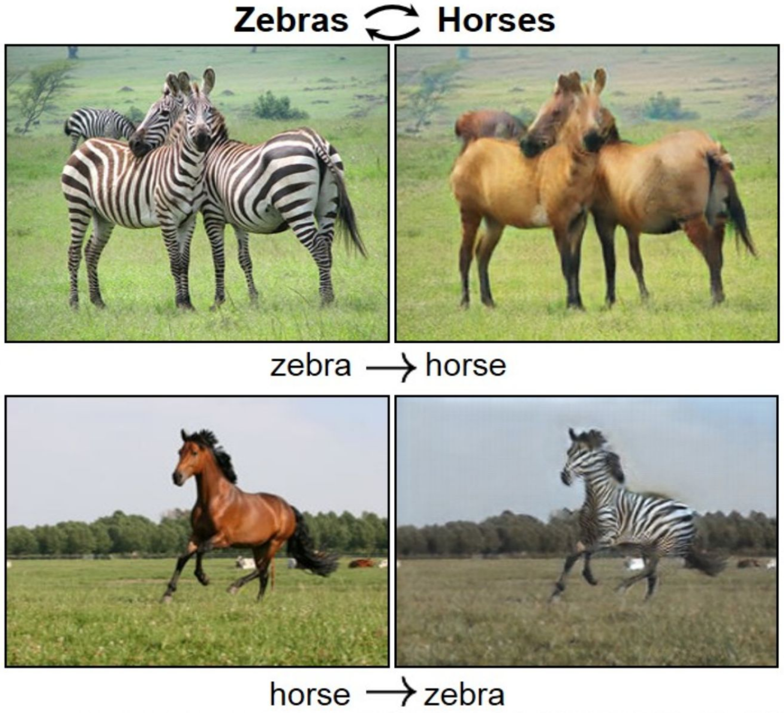
elo7



Aplicações: CycleGAN

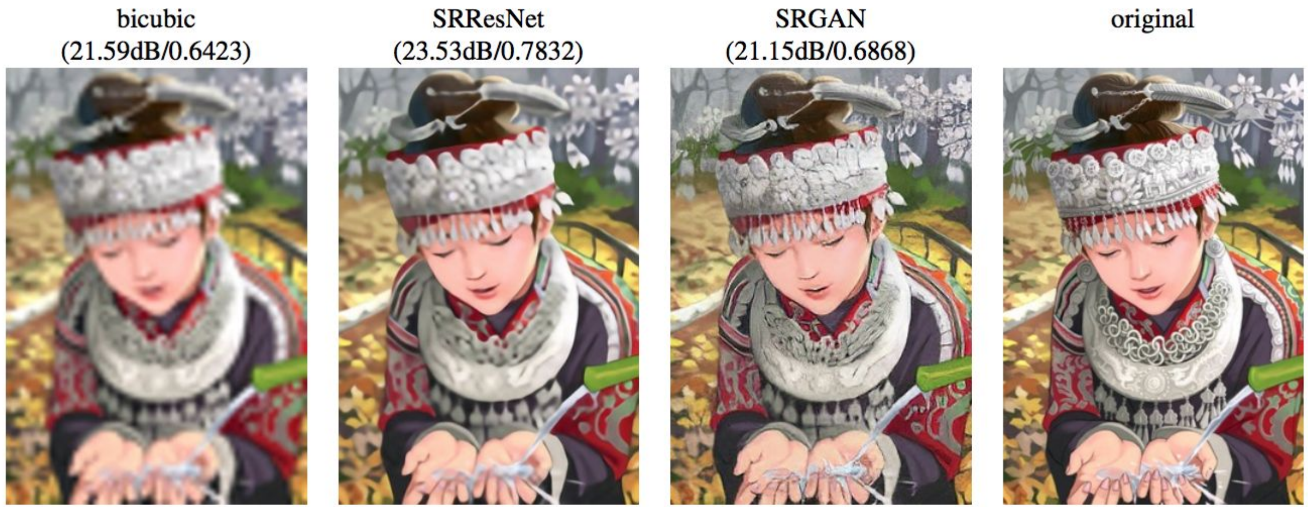
Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

Text description	this large yellow blossom has numerous yellow stamen and hundreds of very thin yellow petals.	this white flower has rounded petals and a yellow orange stamen.	this flower is purple in color, and has petals that are very skinny.	this flower is yellow in color, and has petals that are rounded and curled around the center.	the flower has pointed pale pink petals and several white anthers.	the flower has yellow petals with a yellow stigma and green pedicel.	this flower has petals that are yellow and has red dots
Stage-2 images							
Stage-2 with cycle consistency							
Generated Captions	this flower has petals that are yellow and very thin	this flower has petals that are white with yellow lines	this flower has petals that are pink and has purple dots	this flower has petals that are orange and has yellow edges	this flower has petals that are pink with white shading	this flower has petals that are yellow and has yellow stamen	this flower has petals that are red and has yellow tips



Aplicações: SRGAN

elo7



Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

2019

Aplicações: InfoGAN

elo7

Chen, Xi, et al. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets." *Advances in neural information processing systems*. 2016.

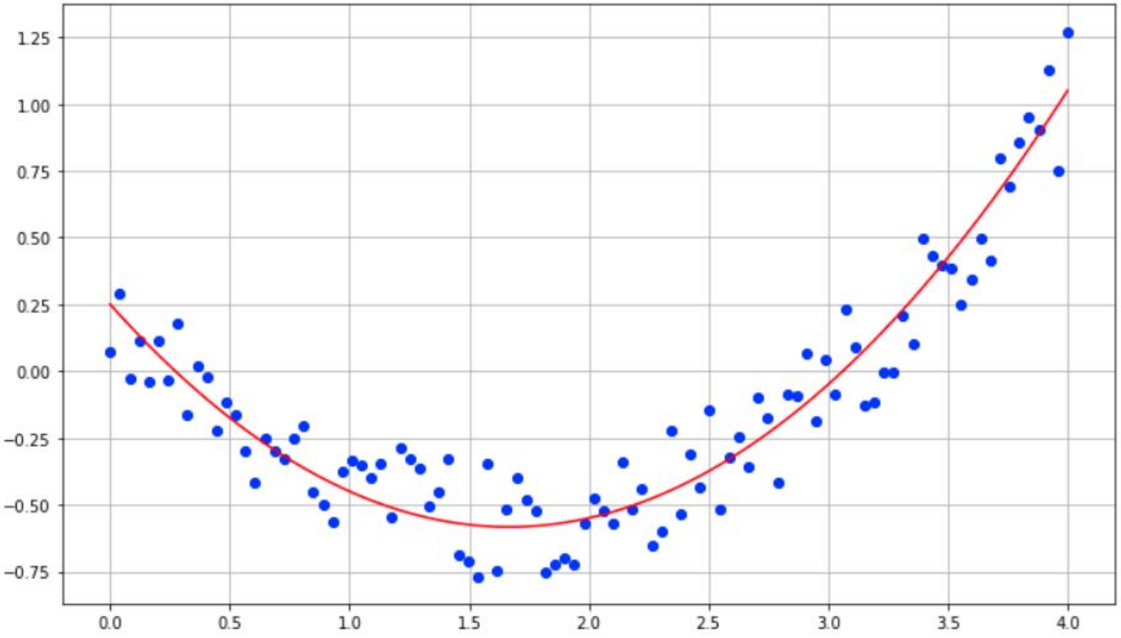
(a) Azimuth (pose) (b) Elevation

(c) Lighting (d) Wide or Narrow

2019

Dados reais ou falsos?

elo7



2019

Dados reais ou falsos?

elo7

Nome	Entrada 1	Saída 1	Entrada 2	Saída 2
John The Armless	09:03	11:47	13:05	18:15
John The Armless	08:40	12:03	12:54	17:50
John The Armless	09:05	11:50	13:15	19:12
John The Armless	10:20	11:55	13:04	18:39

Nome	Entrada 1	Saída 1	Entrada 2	Saída 2
John The Armless	09:00	12:00	13:00	18:00
John The Armless	09:00	12:00	13:00	18:00
John The Armless	09:00	12:00	13:00	18:00
John The Armless	09:00	12:00	13:00	18:00

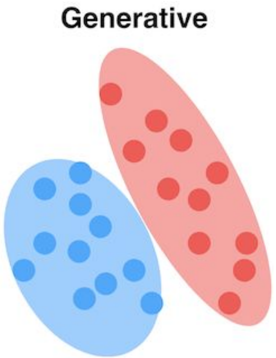
2019

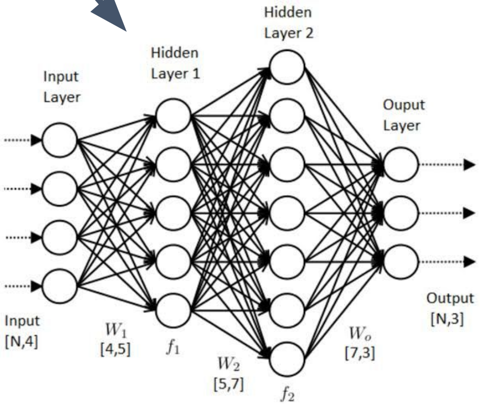
Real ou Falso?

elo7

2019

Generative Adversarial Networks



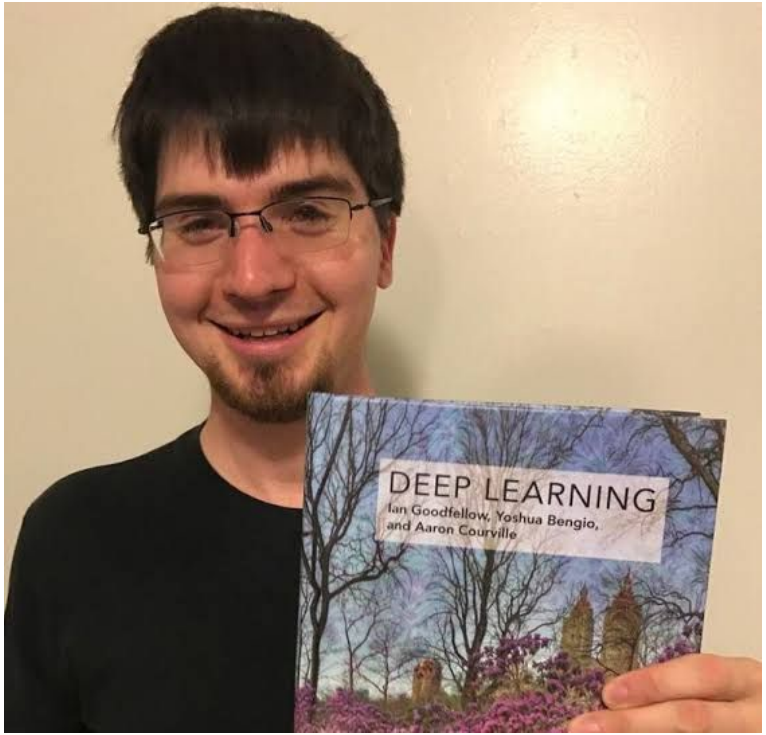


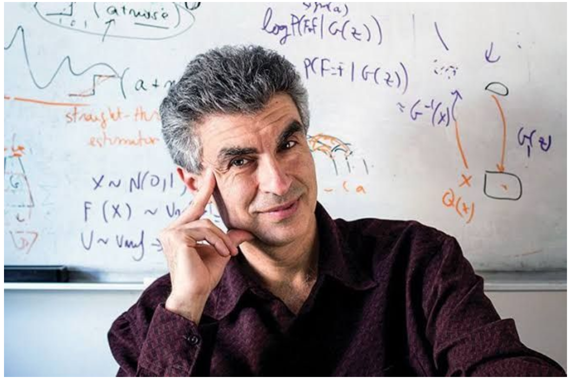
2019

The GANfather - Ian Goodfellow, 2014

elo7

2019



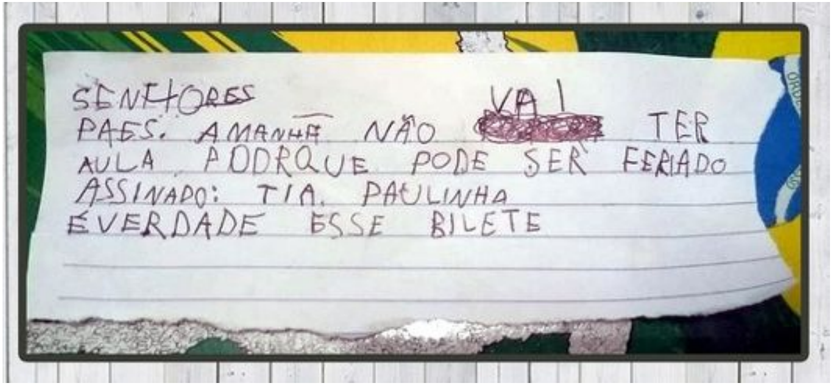


Yoshua Bengio

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Let`s play a game

elo7

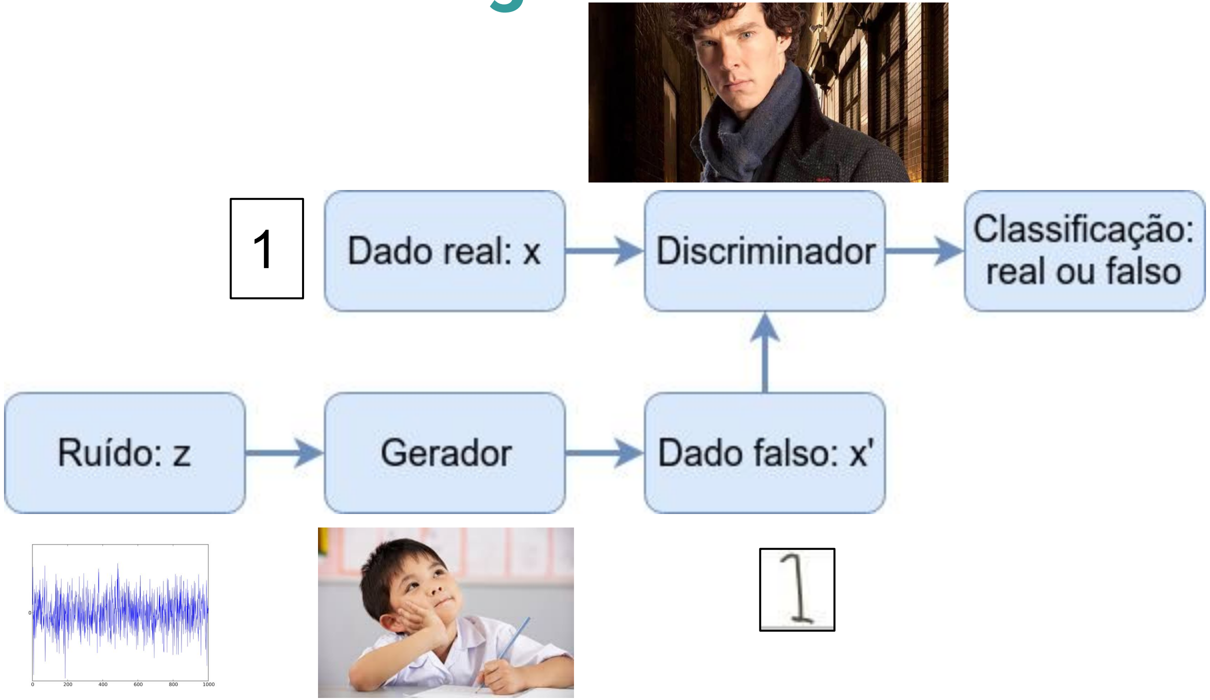


Investigador - Discriminador

Falsificador - Gerador

Adversarial Training

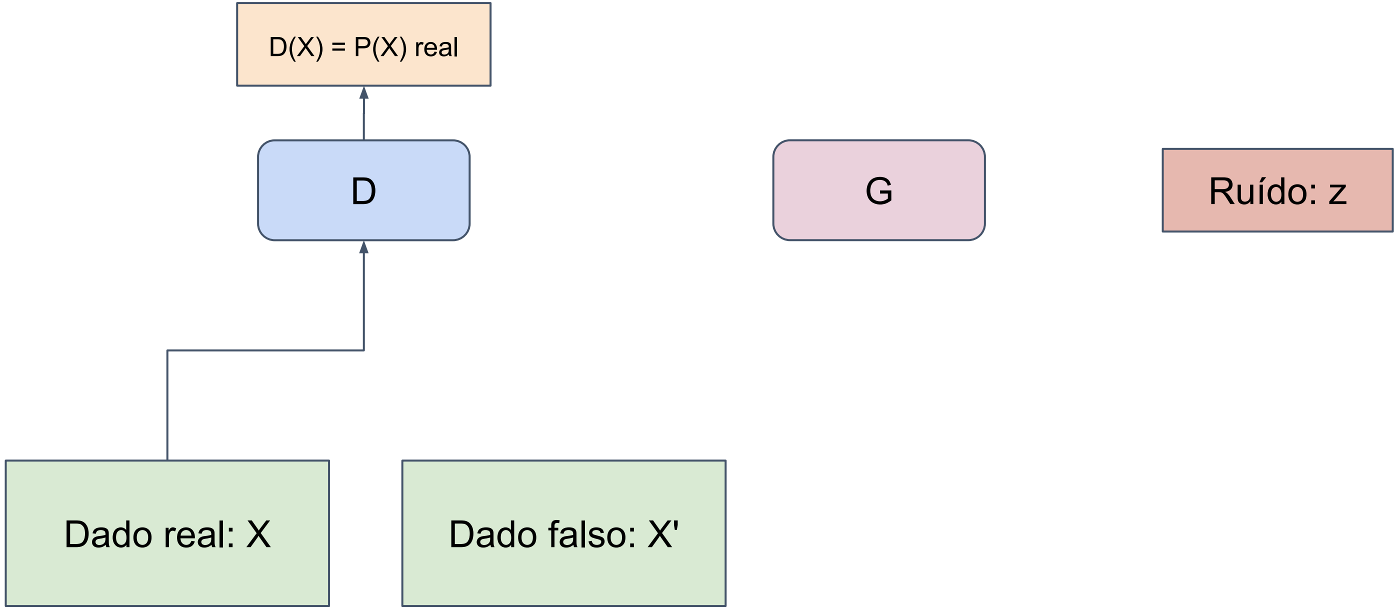
elo7



2019

Adversarial Training

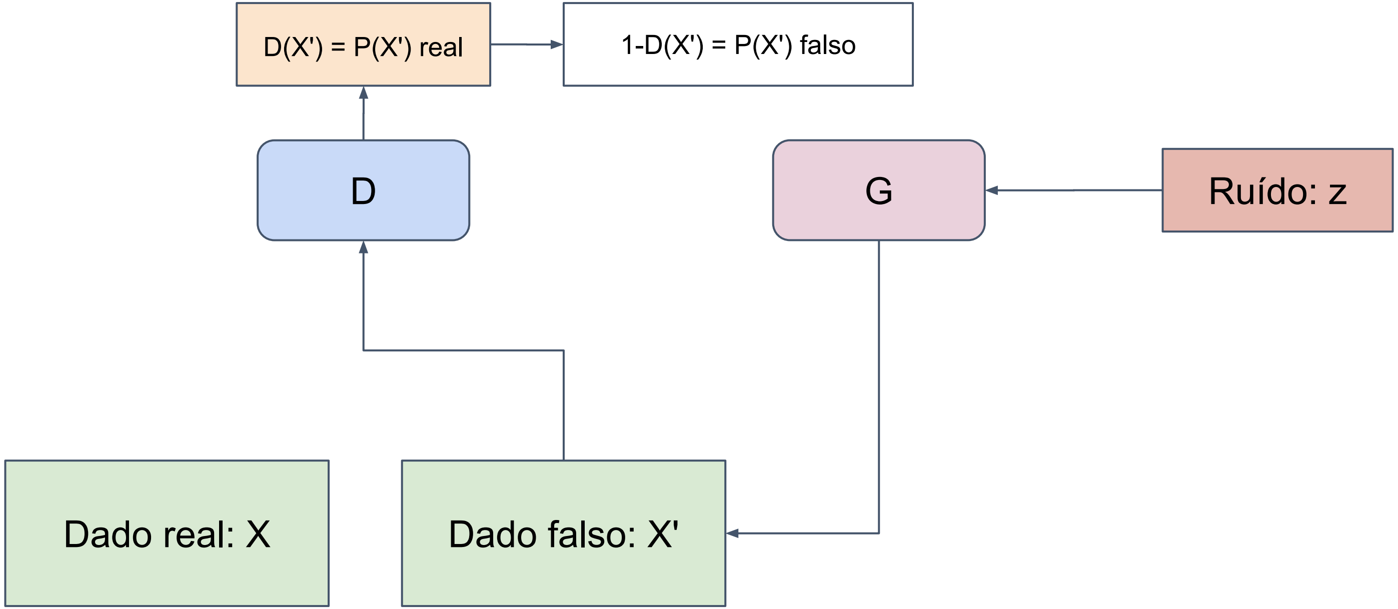
Etapa 1: dados reais



Adversarial Training

elo7

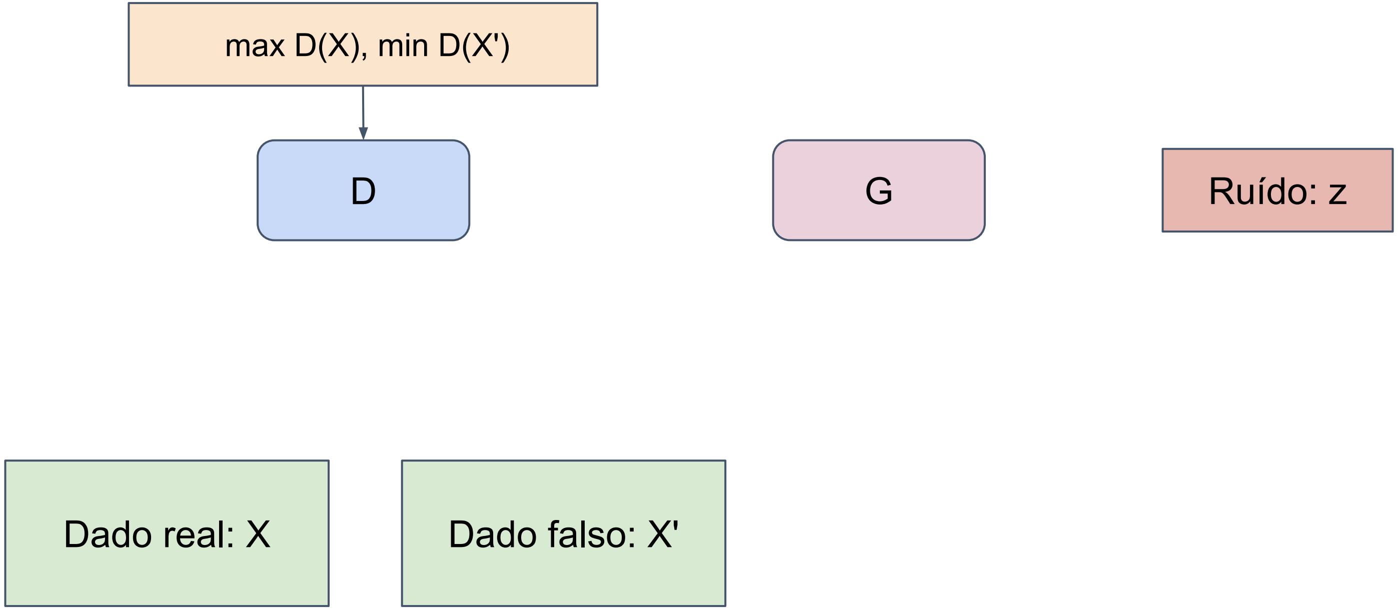
Etapa 2: dados falsos



2019

Adversarial Training

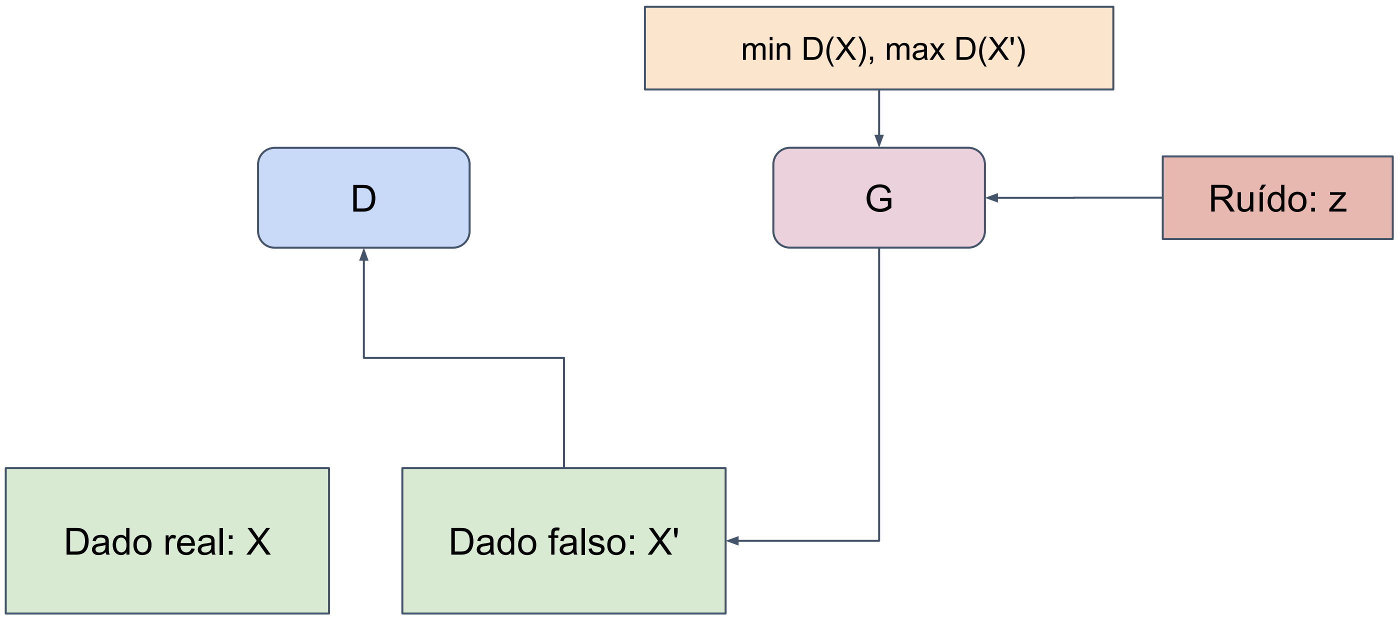
Etapa 3: treino do discriminador



Adversarial Training

elo7

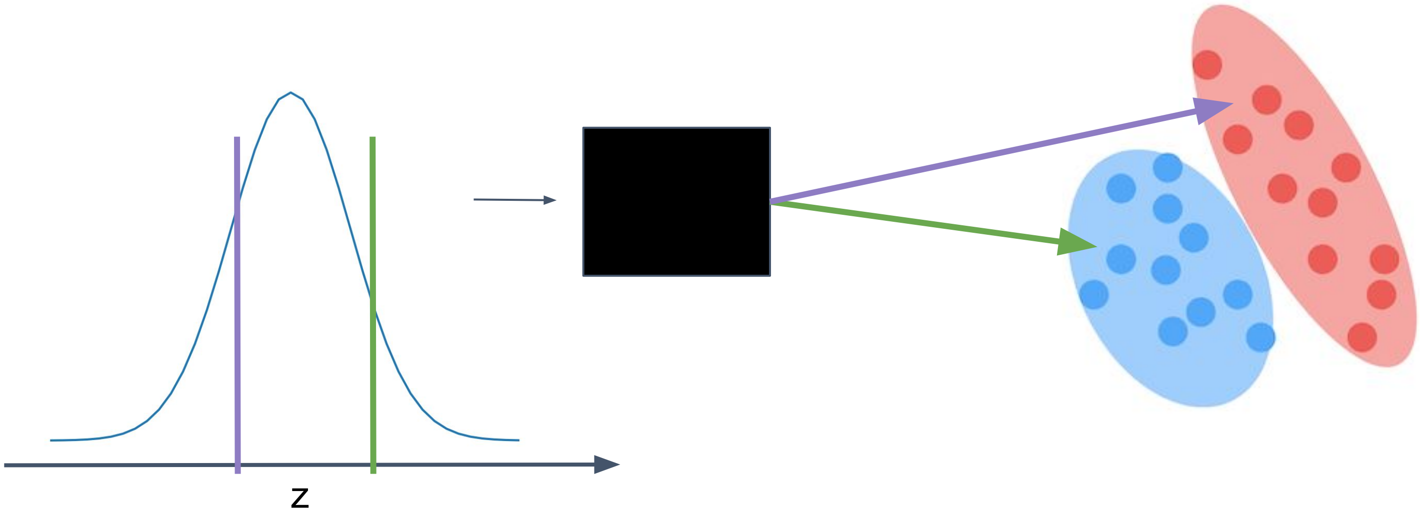
Etapa 4: treino do gerador



2019

Adversarial Training

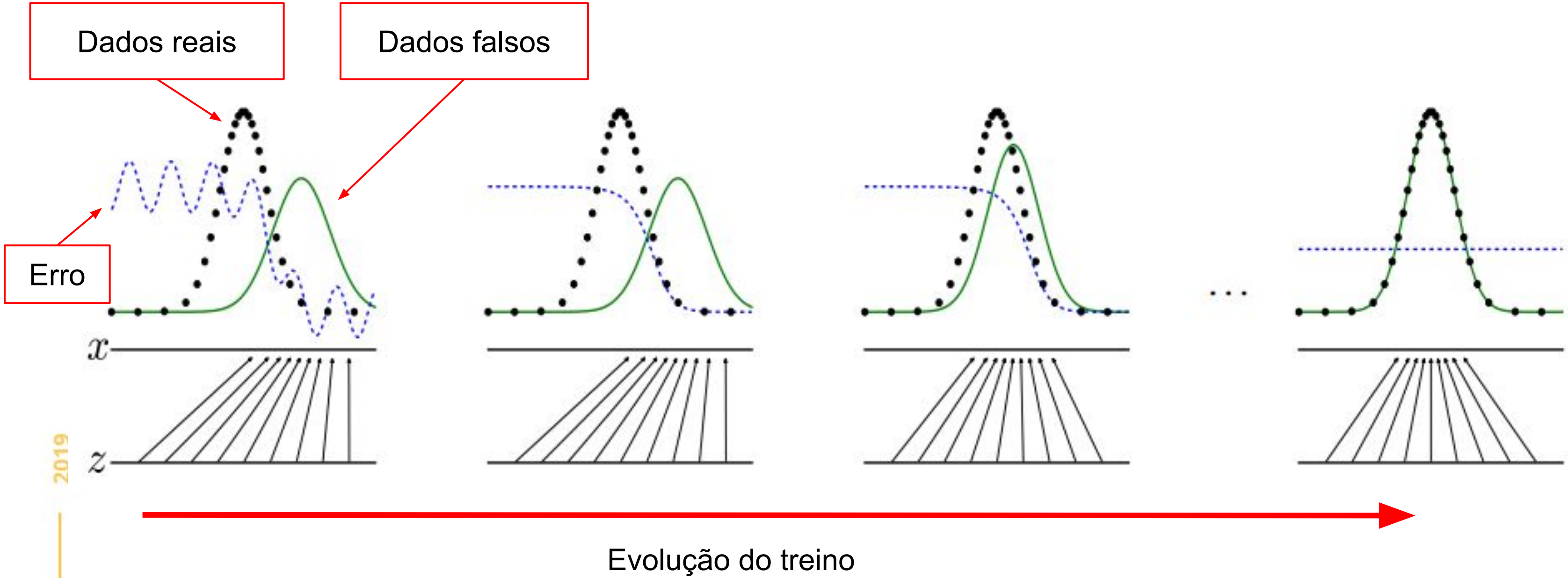
elo7



2019

Adversarial Training

elo7



2019

https://cs.stanford.edu/people/karpathy/gan/

Loss - MinMax Game

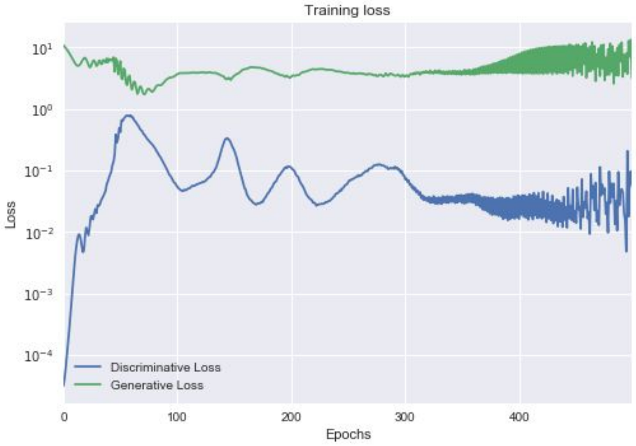
elo7

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

D: tenta maximizar a probabilidade de X ser real e G(z) ser falso

G: tenta minimizar a probabilidade de X real e G(z) ser falso

```
for epoch in epochs:
  prob_X_real = D(X)
  prob_X_fake = D(G(z))
  loss_D = max{log(prob_X_real) + log(1-prob_X_fake)}
  loss_G = min{log(prob_X_real) + log(1-prob_X_fake)}
```



2019

Adversarial Training

elo7

O primeiro desafio: instabilidade

$J_D - J_G = 0$

$$
\begin{aligned}
C(G) &= \max_D V(G, D) \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}}[\log(1 - D_G^*(G(\boldsymbol{z})))] \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim p_g}[\log(1 - D_G^*(\boldsymbol{x}))] \\
&= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x} \sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right]
\end{aligned}
$$

$$\text{No equilíbrio: } p_{data}(x) = \frac{1}{2} \text{ e } p_g(x) = \frac{1}{2}$$

2019

Loss - Non Saturating GAN (NSGAN)

$$max_D \, V(D,G) = \mathbb{E}_{x \sim P_{data}} [\log D(x)] + \mathbb{E}_{z \sim noise} [\log (1 - D(G(z)))]$$

$$max_G \, V(D,G) = \mathbb{E}_{z \sim noise} [\log D(G(z))]$$

D: tenta maximizar a probabilidade de X ser real e G(z) ser falso

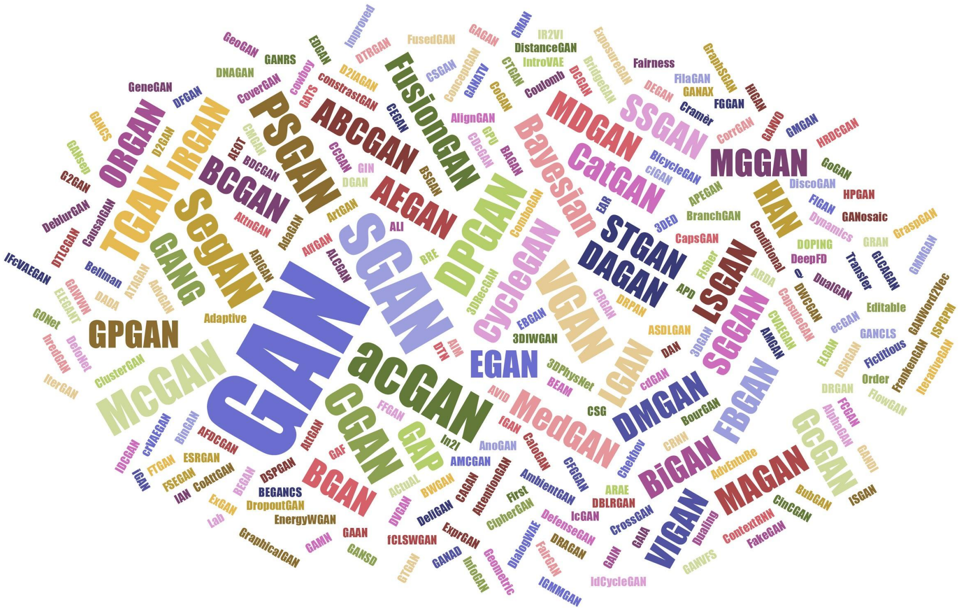
G: tenta maximizar a probabilidade de G(z) ser real

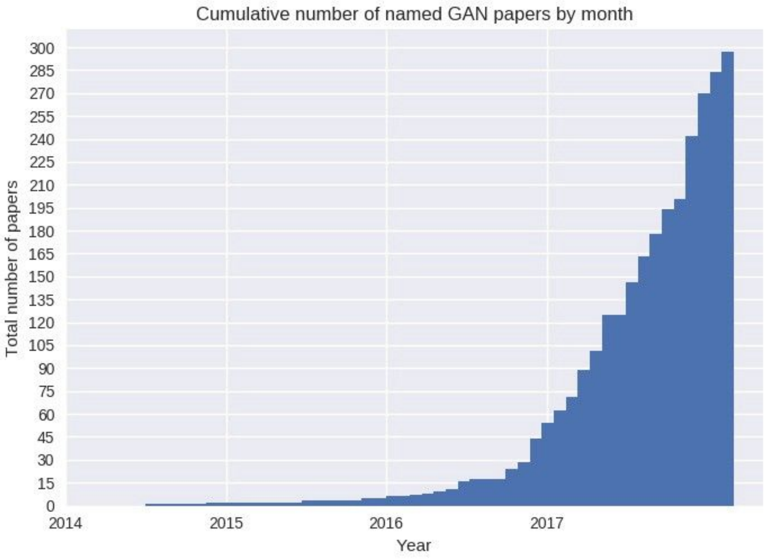
```
for epoch in epochs:
  prob_X_real = D(X)
  prob_X_fake = D(G(z))
  loss_D = max{log(prob_X_real) + log(1-prob_X_fake)}
  loss_G = max{log(prob_X_fake)}
```

2019

GAN Zoo

elo7


GAN
SGAN
acGAN
McGAN
CGAN
VGAN
DPGAN
PSGAN
FusionGAN
ABCGAN
AEGAN
BCGAN
SegGAN
TGAN
IRGAN
ORGAN
GANG
GPGAN
EGAN
CycleGAN
Bayesian
MDGAN
SSGAN
CatGAN
MGGAN
STGAN
DAGAN
LSGAN
SGGAN
HAN
MedGAN
DMGAN
FBGAN
BiGAN
MAGAN
VIGAN
GcGAN
LGAN
BGAN
GAP


Cumulative number of named GAN papers by month
Total number of papers
Year
0
15
30
45
60
75
90
105
120
135
150
165
180
195
210
225
240
255
270
285
300
2014
2015
2016
2017

Mão na massa...

elo7

2019

Conditional GAN

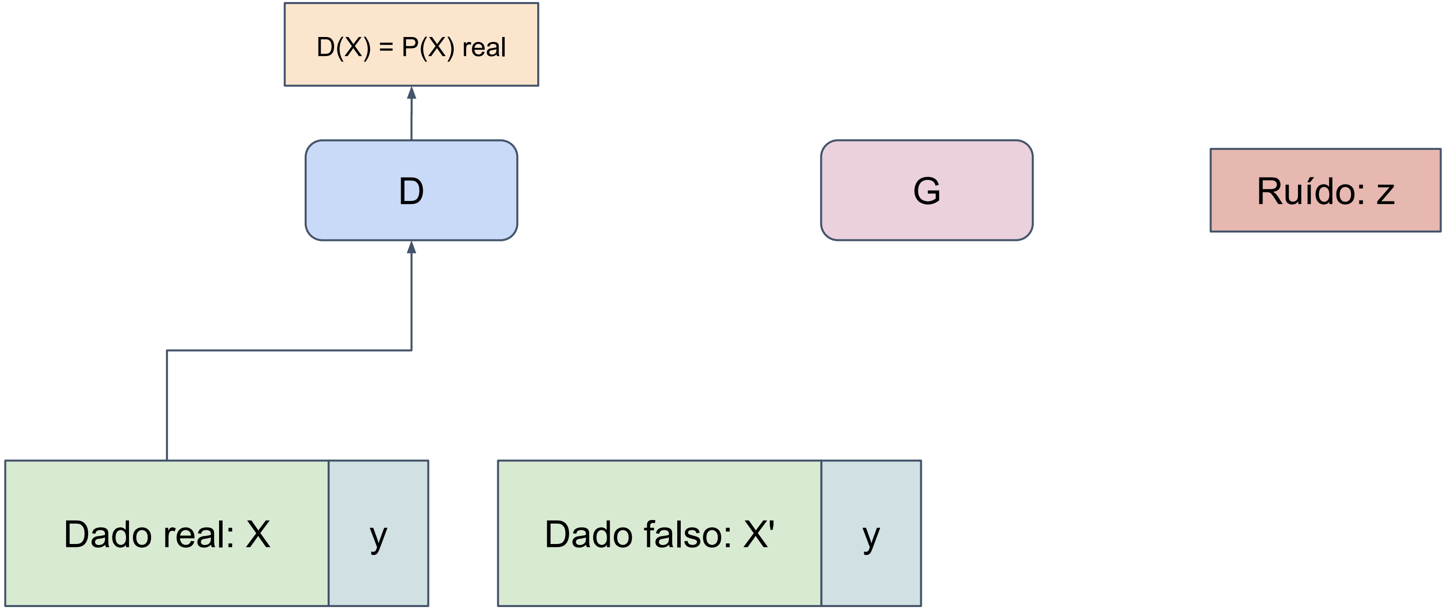
Team Hightower
PhD. Junior A. Koch
Data Scientist @ Elo7

2019

CGAN

elo7

Etapa 1: dados reais

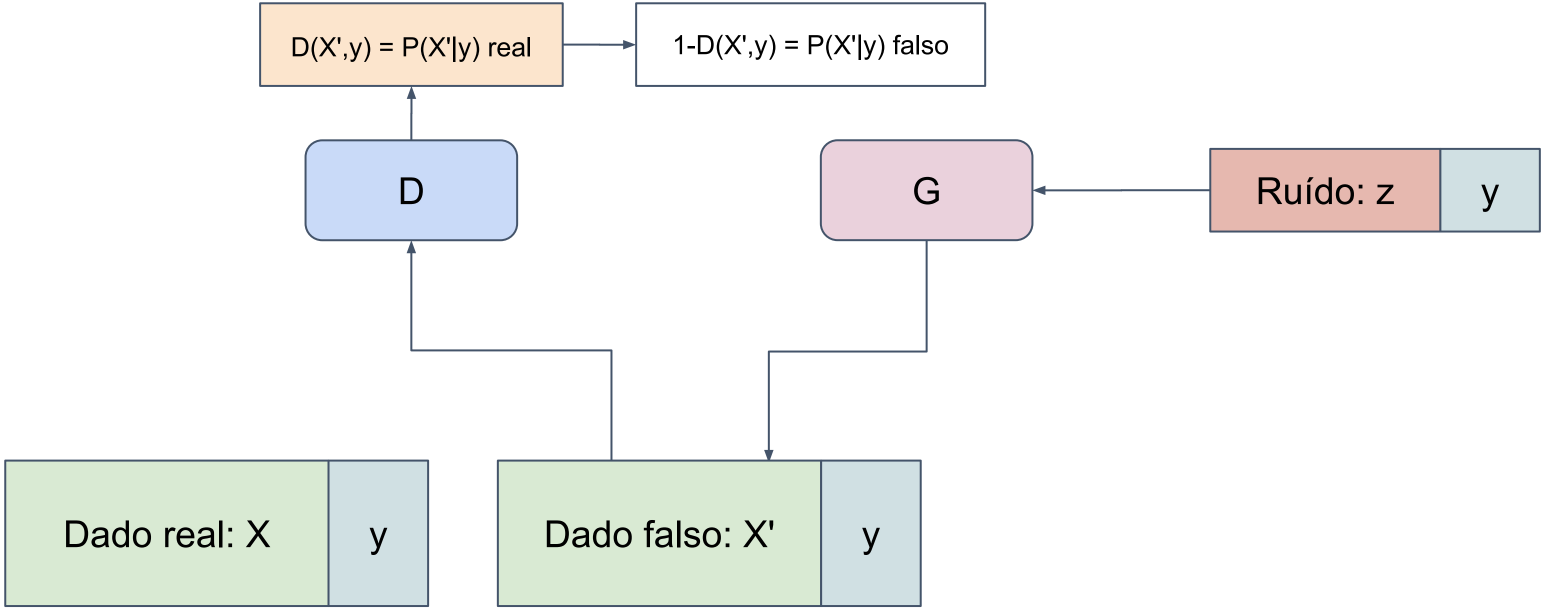


2019

Adversarial Training

elo7

Etapa 2: dados falsos

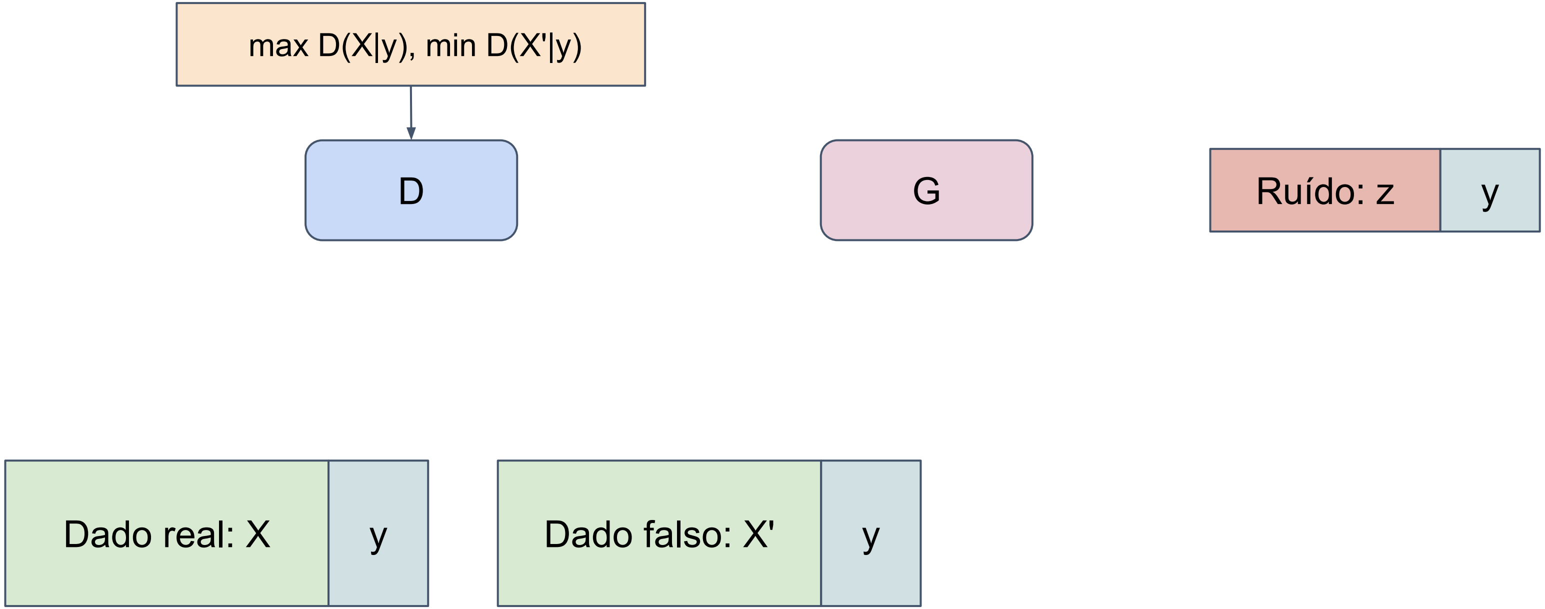


2019

Adversarial Training

elo7

Etapa 3: treino do discriminador

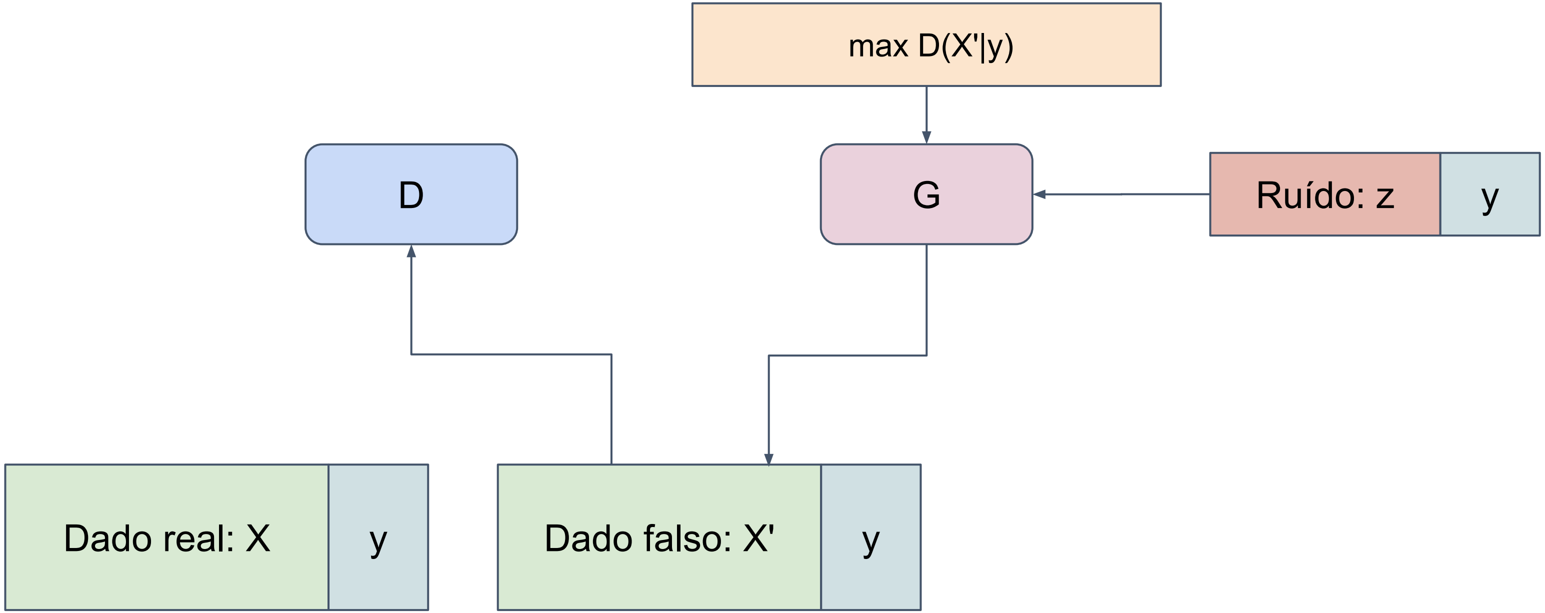


2019

Adversarial Training

elo7

Etapa 4: treino do gerador



2019

Loss - MinMax Game

elo7

$$max_D \, V(D,G) = \mathbb{E}_{x \sim P_{data}} [\log D(x|y)] + \mathbb{E}_{z \sim noise} [\log (1 - D (G(z|y)) |y)]$$
$$max_G \, V(D,G) = \mathbb{E}_{z \sim noise} [\log D (G(z|y)|y)]$$

D: tenta maximizar a probabilidade de (X|y) ser real e (G(z|y)|y) ser falso

G: tenta maximizar a probabilidade de (G(z|y)|y) ser real

```
for epoch in epochs:
  prob_X_real = D(X,y)
  prob_X_fake = D(G(z,y),y)
  loss_D = max{log(prob_X_real) + log(1-prob_X_fake)}
  loss_G = max{log(prob_X_fake)}
```

2019

Vantagens do CGAN

elo7

Mode collapse

Maior controle

2019

Mão na massa...

elo7

2019