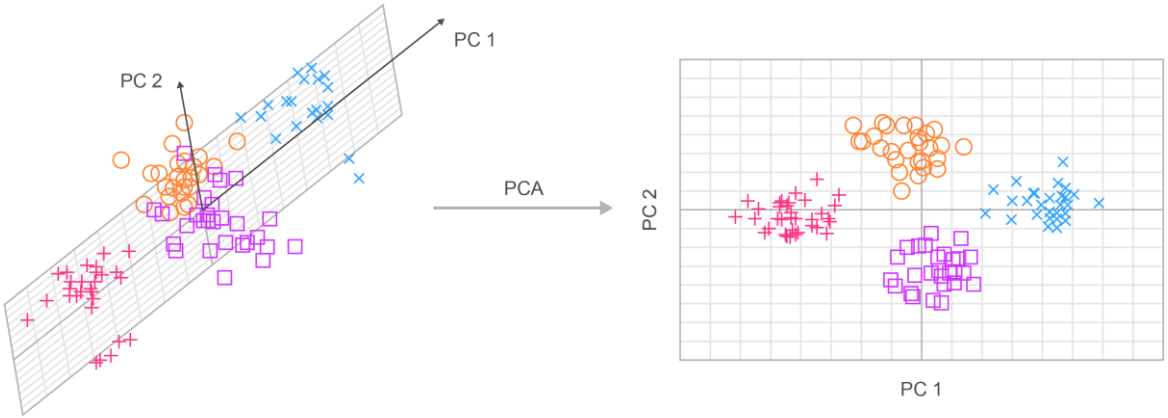
FIAP

O PROBLEMA DA ALTA **DIMENSIONALIDADE...**

ANÁLISE **PCA**

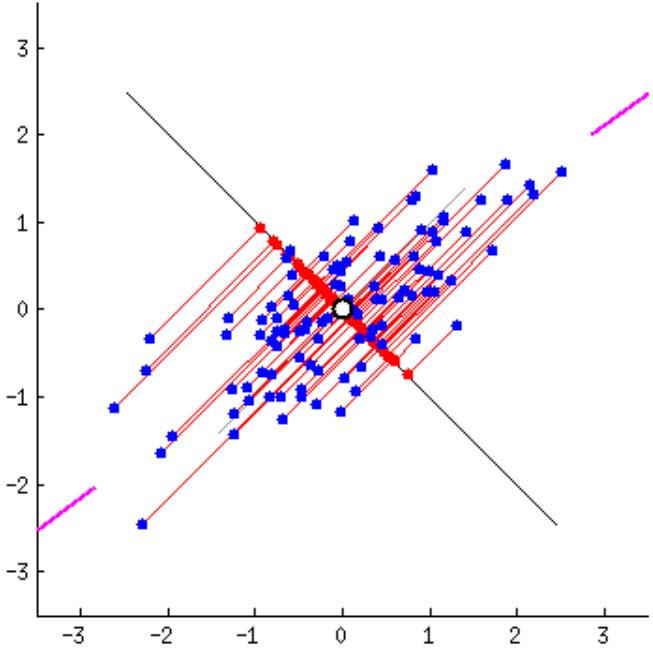
O objetivo da análise é encontrar um meio de condensar a informação contida em várias variáveis originais em um conjunto menor de variáveis estatísticas (componentes) com uma perda mínima de informação (Adaptado Operdata - 2019)



https://medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9

ANÁLISE **PCA**

Transformação Ortogonal dos dados:



ANÁLISE PCA – EXEMPLO
VÍRUS DA CLASSE CORONA

Dataset 1

Seq	Host
ATGTTTGTTTTGCTTGTTGCATATGCCTTGTTGCATATTGCTGGTT...	human
ATGTTTTTGATACTTTTAATTTCCTTACCAATGGCTTTTGCTGTTA...	human
ATGTTTATTTTCTTATTATTTCTTACTCTCACTAGAGGTAGTGACC...	human
ATGACGCCTTTAATTTACTTCTGGTTGTTCTTACCAGTACTTCTAA...	porcine
ATGAAGTCTTTAACTTACTTCTGGTTGTTCTTACCAGTACTTTCAA...	porcine
ATGCAGAGAGCTCTATTGATTATGACCTTACTTTGTCTCGTTCGAG...	porcine
ATGTTTTTGATACTTTTAATTTCCTTACCAACGACTTTTGCTGTTA...	bovine
ATGAAACTTTTTATAGTTTTTGTGCTCCTTTTTAGGGTGTGTTATT...	bat
ATGTTGGTGAAGTCACTGTTTTTAGTGACTCTTTTGTTTGCACTAT...	avian
ATGTTGGTAACACCTCTTTTATTAGTGACTCTTTTGTTTGCACTAT...	avian

730 rows

Fonte: Notas de aula Prof. Dr. Ronaldo Hashimoto – IME USP

ANÁLISE PCA – EXEMPLO
VÍRUS DA CLASSE CORONA

$$P_{xy} = \frac{f_{xy}}{f_x f_y}$$

Where:

- f_x is the frequence of nucleotide x,
- f_y is the frequence of nucleotide y,
- f_{xy} is the frequence of dinucleotide xy.

P_{AA}	P_{AT}	P_{AC}	P_{AG}	P_{TA}	P_{TT}	P_{TC}	P_{TG}	P_{CA}	P_{CC}	P_{CT}	P_{CG}	P_{GA}	P_{GT}	P_{GC}	P_{GG}

So, we will have measurements for 16 dinucleotides!!!

And, if we aggregate the frequence of mononucleotides T, C, and G.

f_A	f_T	f_C	f_G

So, we will have 19 measurements for each sequence!!!

ANÁLISE PCA – EXEMPLO
VÍRUS DA CLASSE CORONA

Feature Vector

Class

Sample

Dataset 1

	G	AG	GG	CG	TG	TT	Host
0	0.665225	-0.304268	-0.281132	1.510045	2.153277	1.823231	human
36	-2.466542	-0.859019	0.530849	-0.118405	-1.293848	-1.017817	human
72	-0.254466	-0.433647	-0.162090	-0.915341	-0.778621	-1.161352	human
108	-1.376046	-1.309138	-0.711194	0.430361	-0.279186	-0.342964	human
144	-0.989784	-0.582103	0.818058	-0.970409	1.390465	1.075697	human
180	-0.388242	-0.212588	-0.247272	0.758464	-0.598655	-0.412604	human
216	1.176985	-0.129859	-1.289674	0.330971	-0.855900	-0.671098	porcine
252	1.072384	-0.110824	-1.039439	-0.308360	-0.846077	-0.705007	porcine
288	1.135144	-0.065912	-0.983338	-0.327244	-0.835327	-0.692845	porcine
324	1.093304	-0.050562	-1.063914	-0.351032	-0.852434	-0.618089	porcine
360	-1.194676	1.171019	1.793185	0.959133	2.159648	1.867804	porcine
396	-0.022611	-0.617248	0.096755	-1.050849	-0.861156	-1.175692	bovine
432	-0.043884	-0.613164	0.126130	-1.517958	-0.737744	-1.247607	bovine
468	-0.212750	0.626935	-0.250648	0.604421	1.079619	1.205546	bat
504	-0.229375	-0.973427	-0.557187	0.780419	0.258101	-0.322454	bat
540	1.798511	-0.072512	-1.286707	1.121890	-0.186046	-0.351052	murine
576	-0.382002	1.768964	1.532858	0.815515	0.521813	1.266507	avian
612	-0.288020	1.730763	1.722878	0.087307	0.462897	1.034374	avian
648	-0.335804	1.652483	1.683257	0.165865	0.520066	0.872854	avian
684	-0.238121	1.729452	1.214314	0.326928	0.544972	1.038252	avian
720	-2.321768	1.091406	2.558835	0.117995	1.634372	2.089170	avian

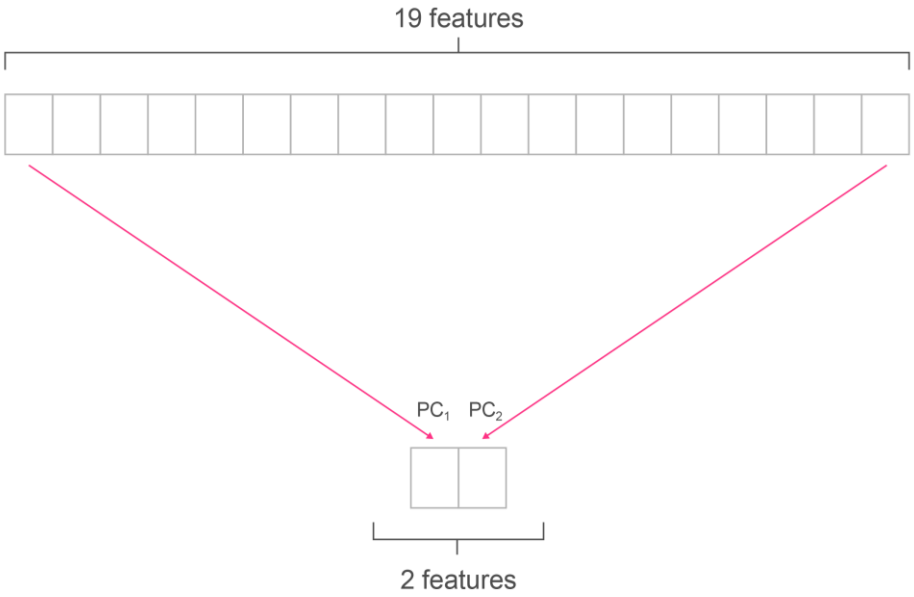
730 rows x 19 cols

n = 730 samples

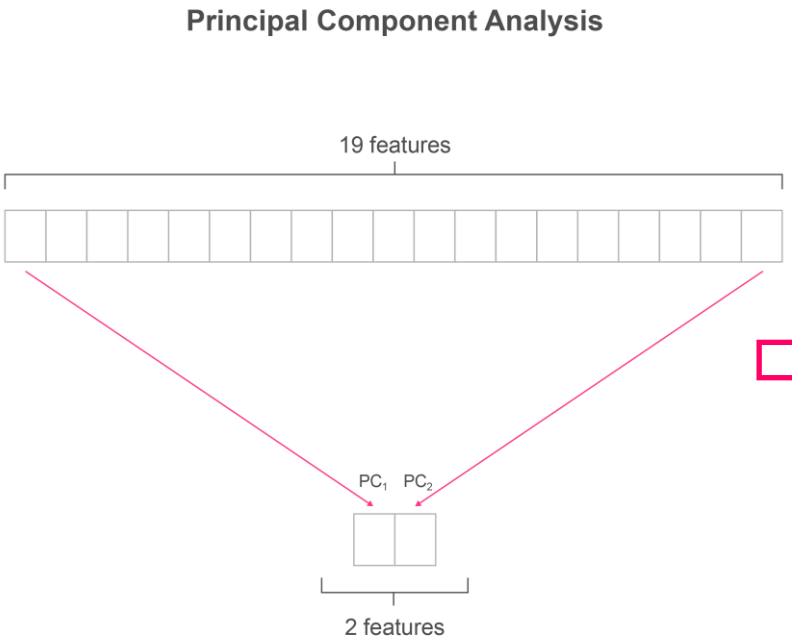
Dimension: d = 19 features

ANÁLISE PCA – EXEMPLO
VÍRUS DA CLASSE CORONA

Principal Component Analysis (não supervisionado)



ANÁLISE PCA – EXEMPLO
VÍRUS DA CLASSE CORONA

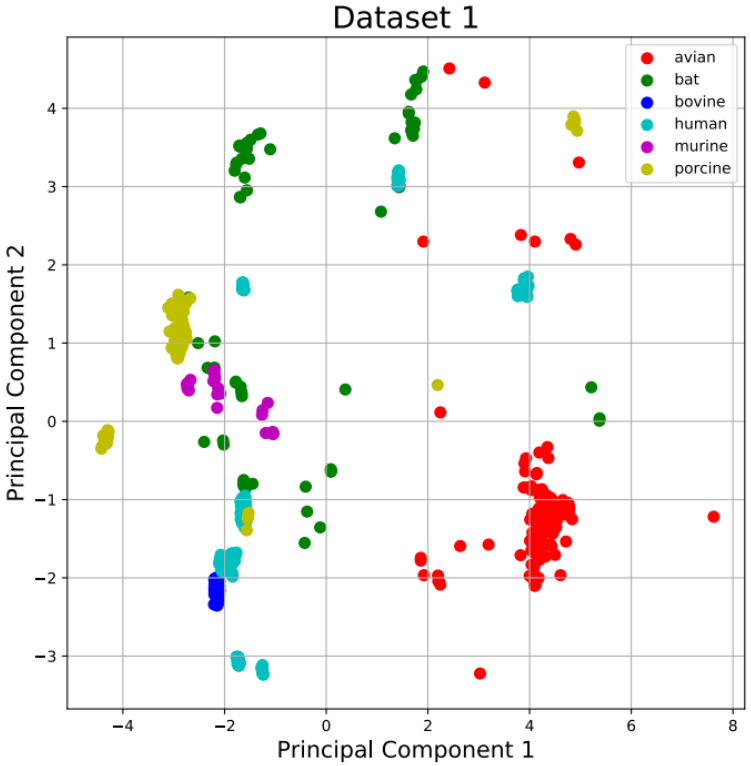


Dataset 1

	PC 1	PC 2	Host
0	3.775624	1.613269	human
36	-1.254264	-3.114602	human
72	-1.845260	-1.941322	human
108	-1.657282	-0.982642	human
144	1.436796	3.099578	human
180	-1.643563	1.725470	human
216	-2.941701	1.510518	porcine
252	-2.935069	1.062321	porcine
288	-2.914471	0.908152	porcine
324	-2.912731	0.903012	porcine
360	4.935531	3.713215	porcine
396	-2.124629	-2.173583	bovine
432	-2.178364	-2.104020	bovine
468	1.622693	3.955135	bat
504	-1.699718	2.866013	bat
540	-1.049189	-0.162940	murine
576	4.694584	-1.157702	avian
612	4.440957	-1.328387	avian
648	4.228678	-0.941589	avian
684	4.290554	-1.297679	avian
720	4.907956	2.258447	avian

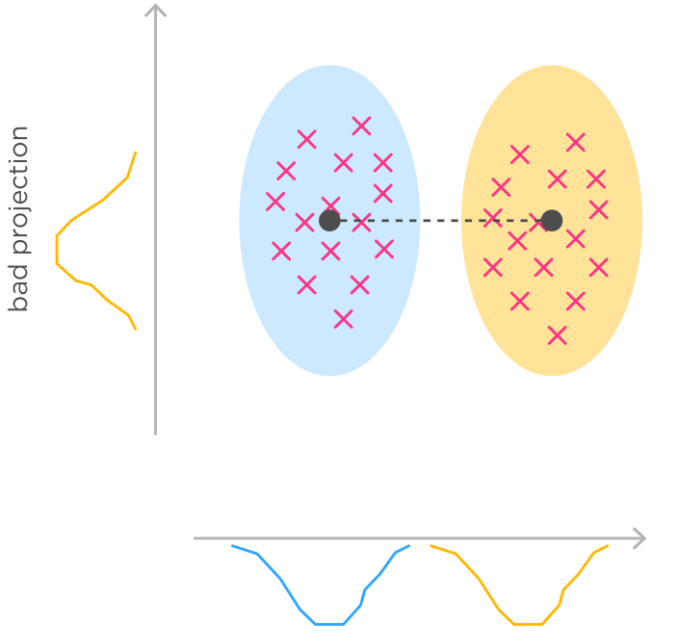
730 rows x 2 cols

ANÁLISE PCA – EXEMPLO
VÍRUS DA CLASSE CORONA



REDUÇÃO DE **DIMENSIONALIDADE**

- Análise do Discriminante Linear (LDA) (supervisionado)



T-SNE: T-Distributed Stochastic **Neighbour Embedding**

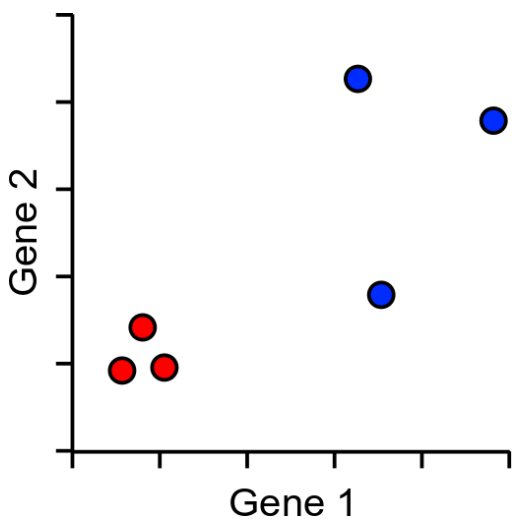
O objetivo t-SNE é endereçar alguns problemas do próprio PCA, tais como:

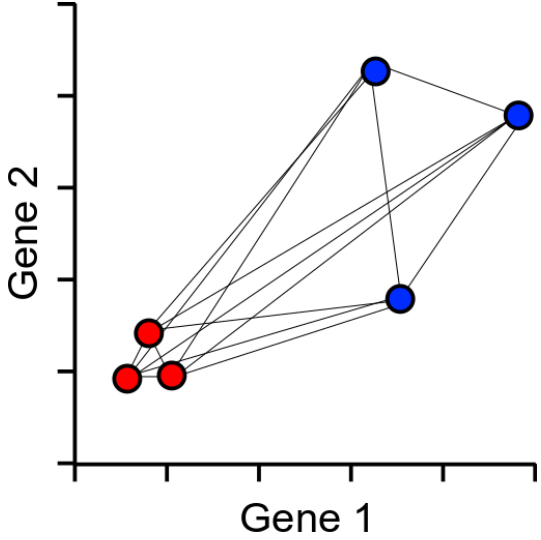
- Escala não linear para representar mudanças em diferentes níveis de dimensionalidade;
- Separação ideal em 2 dimensões;
- Não-supervisionado (originalmente)
- Distorção do espaço para melhorar visualização

T-SNE: T-Distributed Stochastic **Neighbour Embedding**

Como Funciona?

Com base na tabela all-vs-all de distâncias de célula a célula em pares



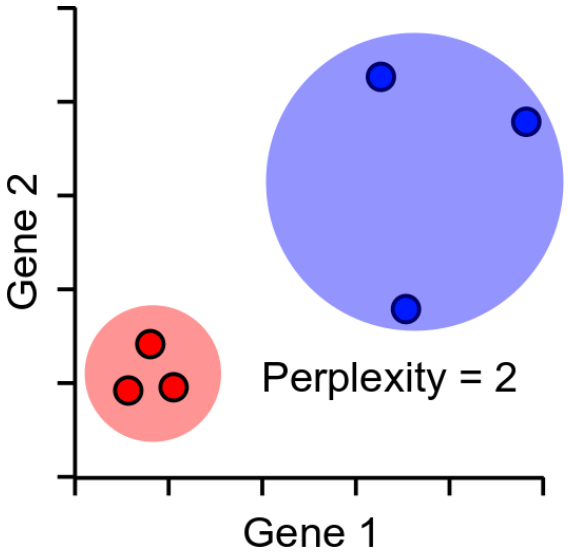


	0	10	10	295	158	153
	9	0	1	217	227	213
	1	8	0	154	225	238
	205	189	260	0	23	45
	248	227	246	44	0	54
	233	176	184	41	36	0

T-SNE: T-Distributed Stochastic **Neighbour Embedding**

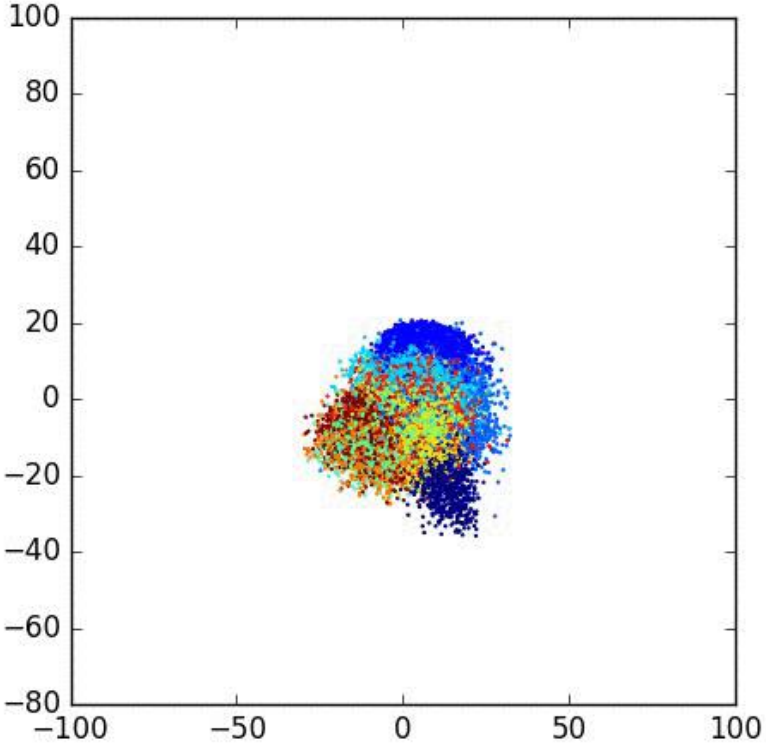
Perplexidade = número esperado de vizinhos dentro de um cluster

Distâncias dimensionadas em relação aos vizinhos de perplexidade

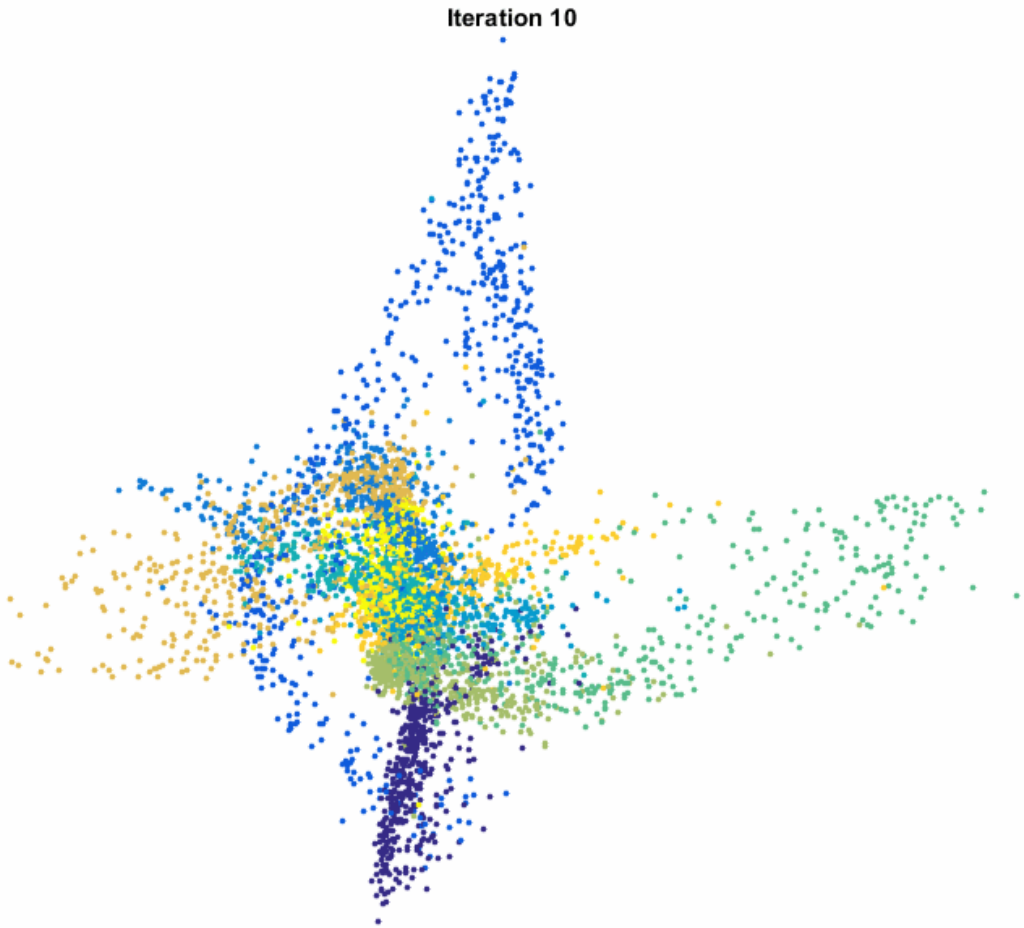


	0	4	6	586	657	836
	4	0	4	815	527	776
	9	3	0	752	656	732
	31	28	29	0	4	7
	31	24	25	4	0	7
	40	37	32	8	8	0

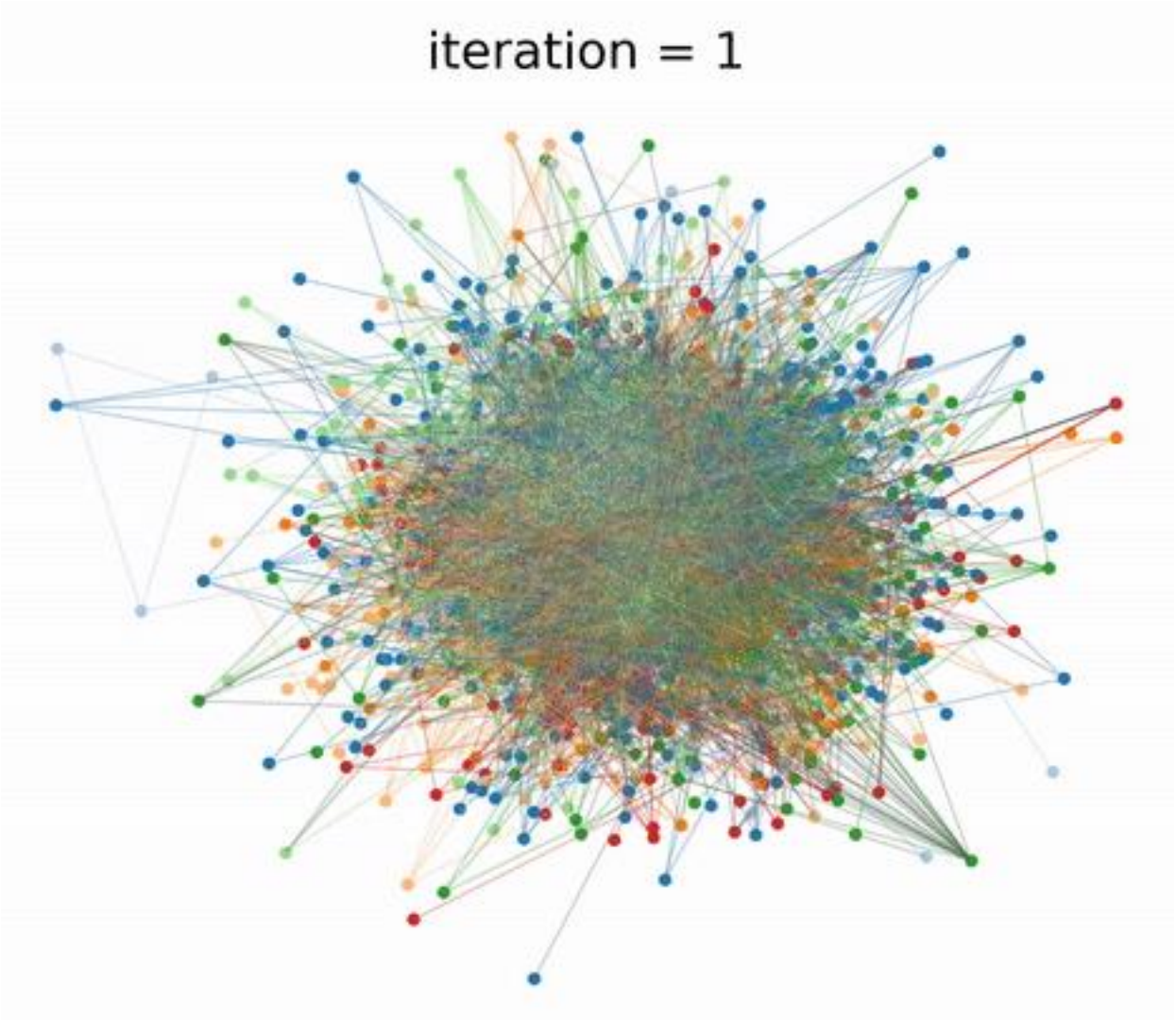
T-SNE: T-Distributed Stochastic **Neighbour Embedding**



T-SNE



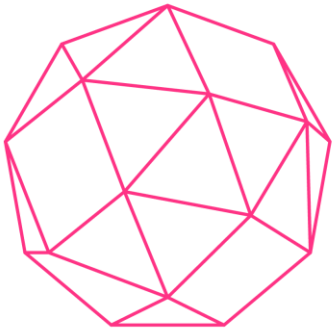
T-SNE



COMO FAZER REDUÇÃO DA DIMENSIONALIDADE ONDE **A AMOSTRA DO NOSSO PROBLEMA É UMA IMAGEM?**

AUTOENCODERS

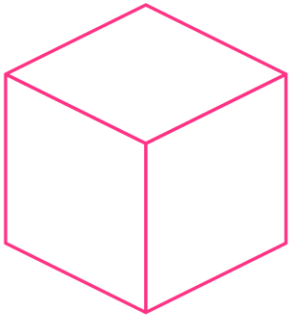
Multidimensional Data



Data represented best

Slow performance, High Precision

Low Dimensional Data

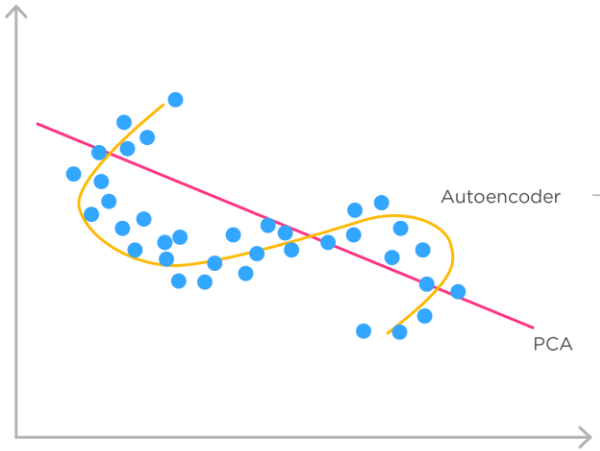


Reduce Precision

High Performance

AUTOENCODERS

Linear vs nonlinear dimensionality reduction



Non-linear Transformations

Non-linear activation function and multiple layers.

Convolutional Layers

An autoencoder doesn't have to learn dense layers.

Higher Efficiency

More efficient to learn several layers with an autoencoder.

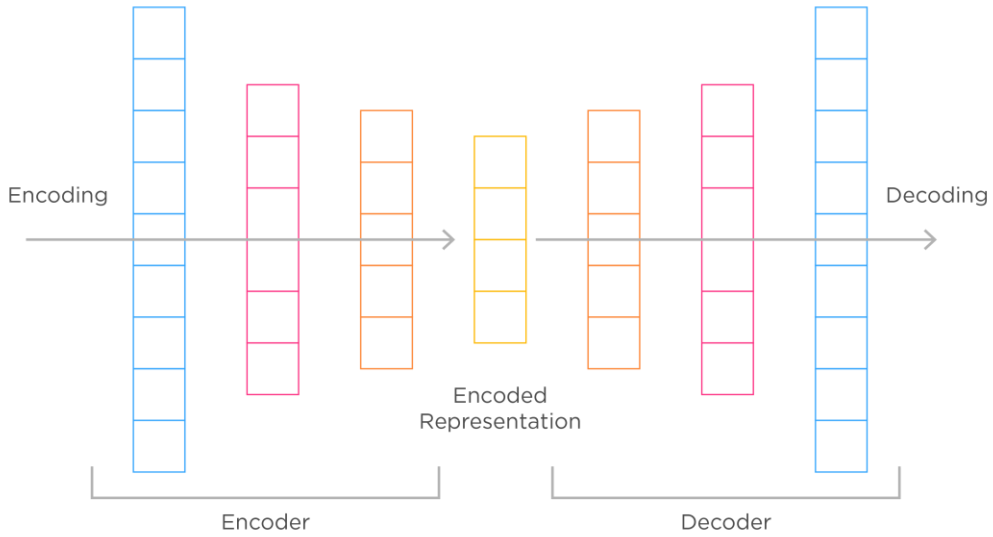
Multiple Transformation

It gives a representation as the output of each layer.

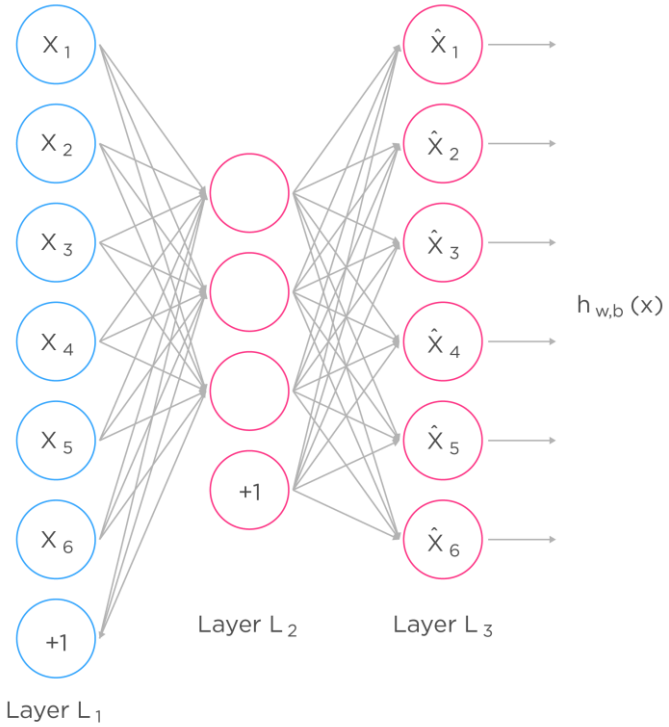
AUTOENCODERS

AUTOENCODERS

An **autoencoder** neural network is an unsupervised Machine learning algorithm that applies backpropagation, setting the target values to be equal to the inputs.



AUTOENCODERS - **DEFINIÇÃO**



Key Facts about Autoencoders

- It is an unsupervised ML algorithm similar to PCA.
- It minimizes the same objective function as PCA.
- It is a neural network.
- The neural network's target output is its input.

AUTOENCODERS

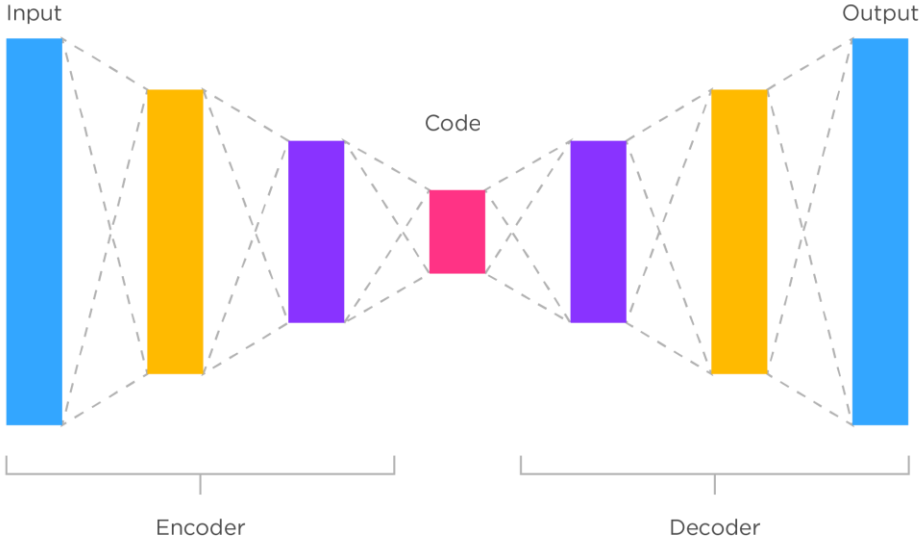
Componentes:

Components of Autoencoders

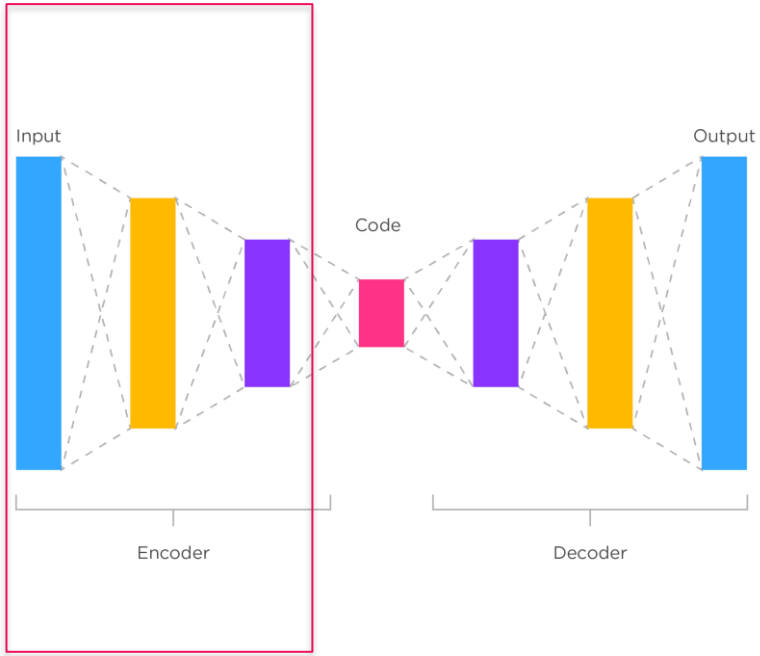
01 Encoder

02 Code

03 Decoder



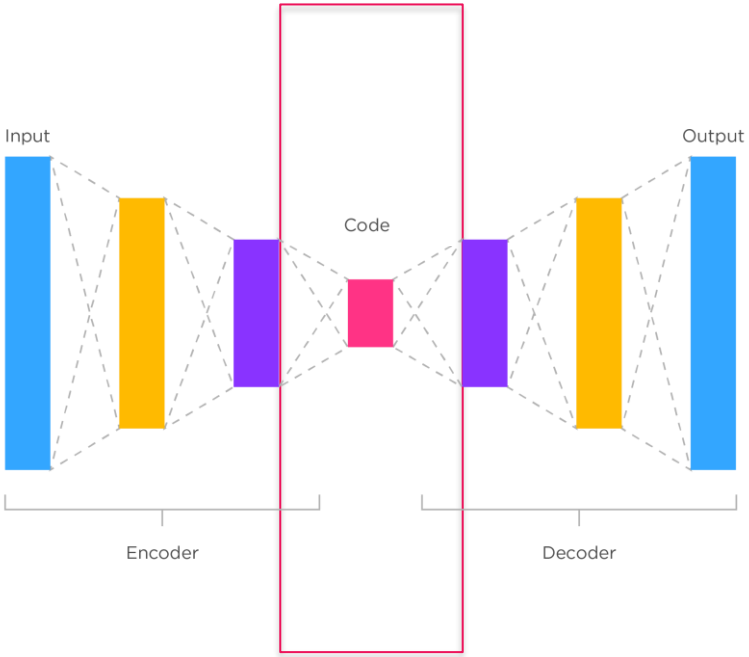
AUTOENCODERS



Encoder

This is the part of the networks that compresses the input into a latent space representation.

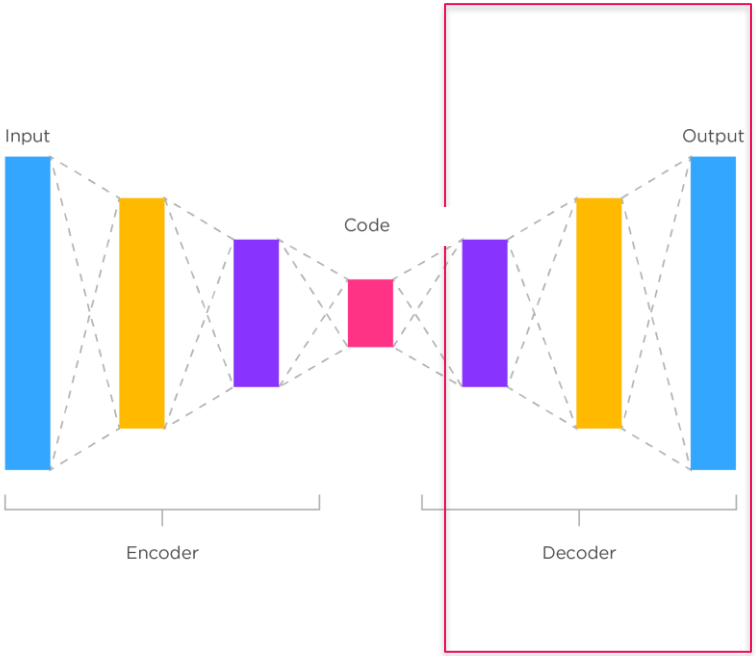
AUTOENCODERS



Code

This is the part of the network represents the compressed input that is fed to the decoder

AUTOENCODERS



Decoder

This part aims to reconstruct the input from the latent space representation.

PROPRIEDADES **DOS AUTOENCODERS**

01 Unsupervised

Autoencoders are considered an unsupervised learning technique since they don't need explicit lables to train on

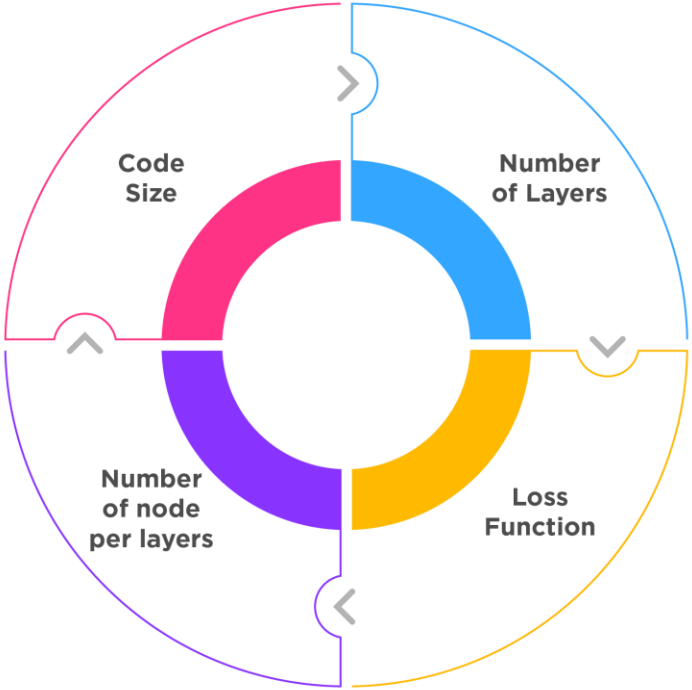
02 Data-specific

Autoencoders are only able to meaningfully compress data similar to what they have been trained on

03 Lossy

The output of the autoencoder will not be exactly the same as the input, it will be a close but degraded representation

PROPRIEDADES **DOS AUTOENCODERS**



Code Size

Smaller size results in more compression

Number of Layers

The autoencoder can have manu layers

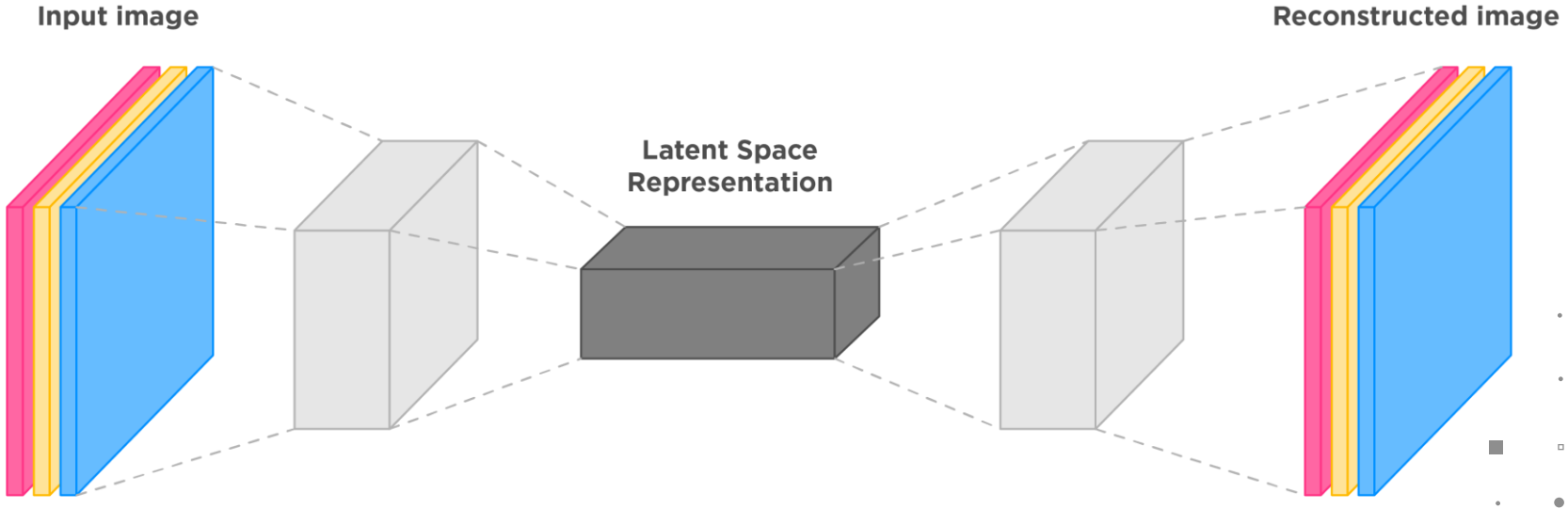
Loss Function

Mean squared error or binary cross entropy

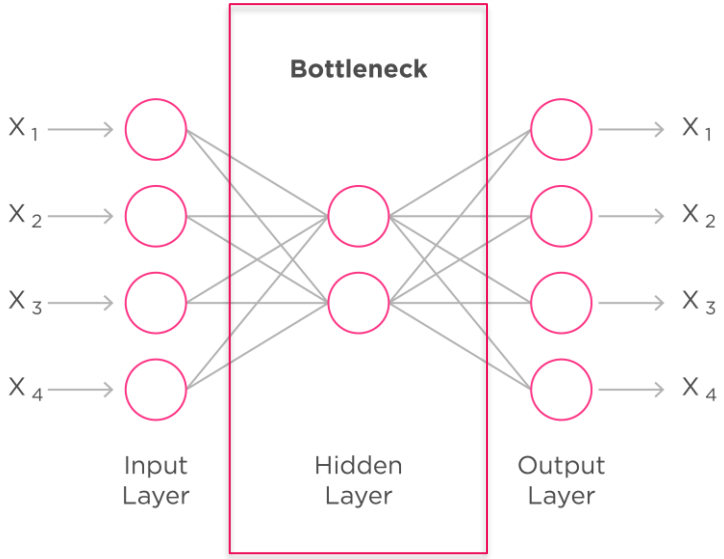
Number of node per layers

Stacked autoencoders look like a sandwich

ARQUITETURA **DOS AUTOENCODERS**



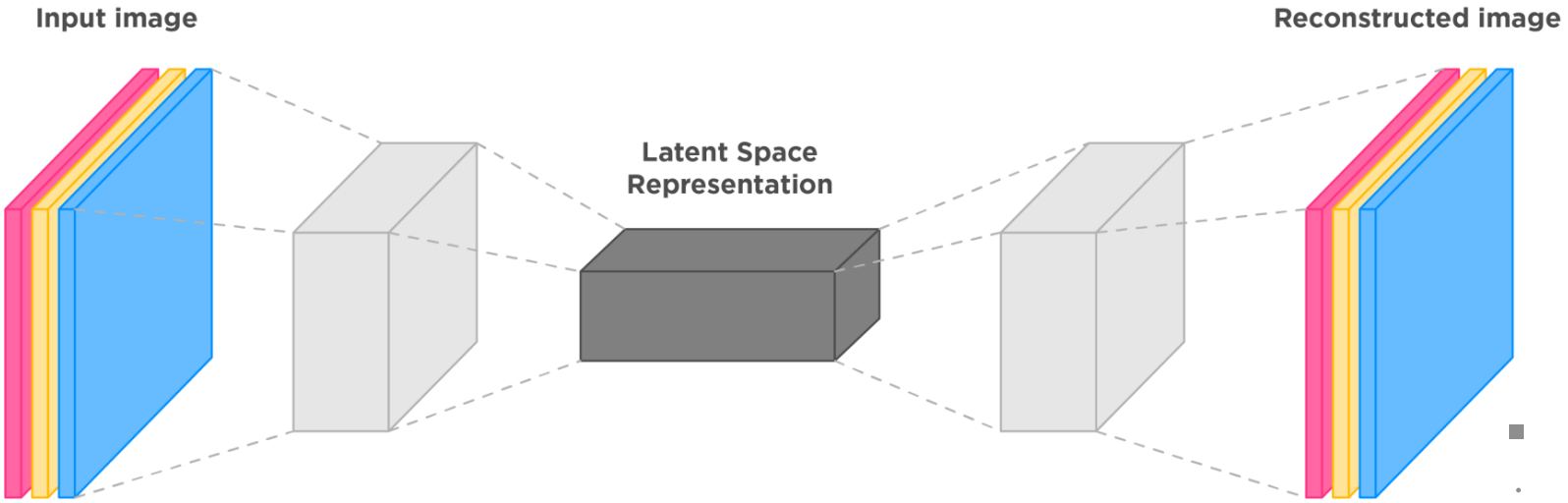
ARQUITETURA **DOS AUTOENCODERS**



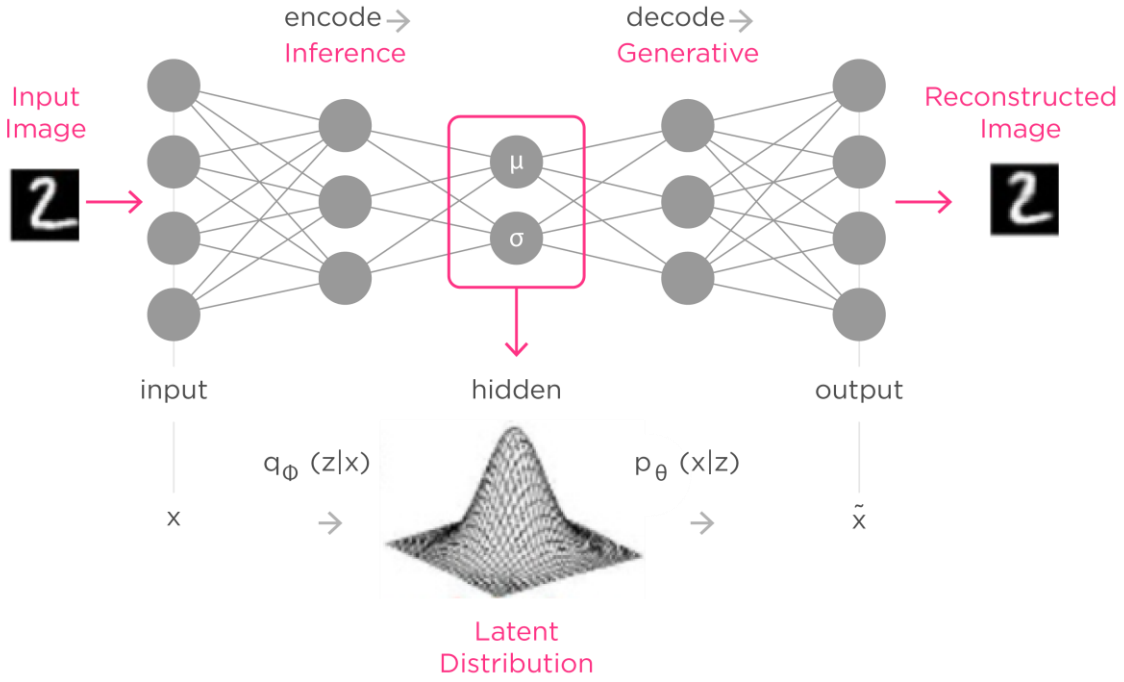
Bottleneck approach is an approach to for deciding which aspects of observed data are relevant information and what aspects can be thrown away.

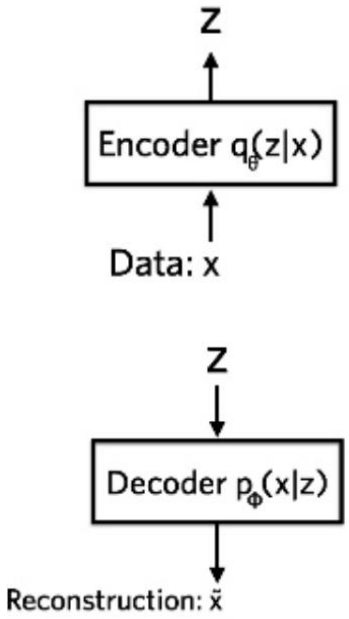
- Compactness of representation, measured as the compressibility.
- Representation retains about some behaviourally relevant variables.

ARQUITETURA **DOS AUTOENCODERS**

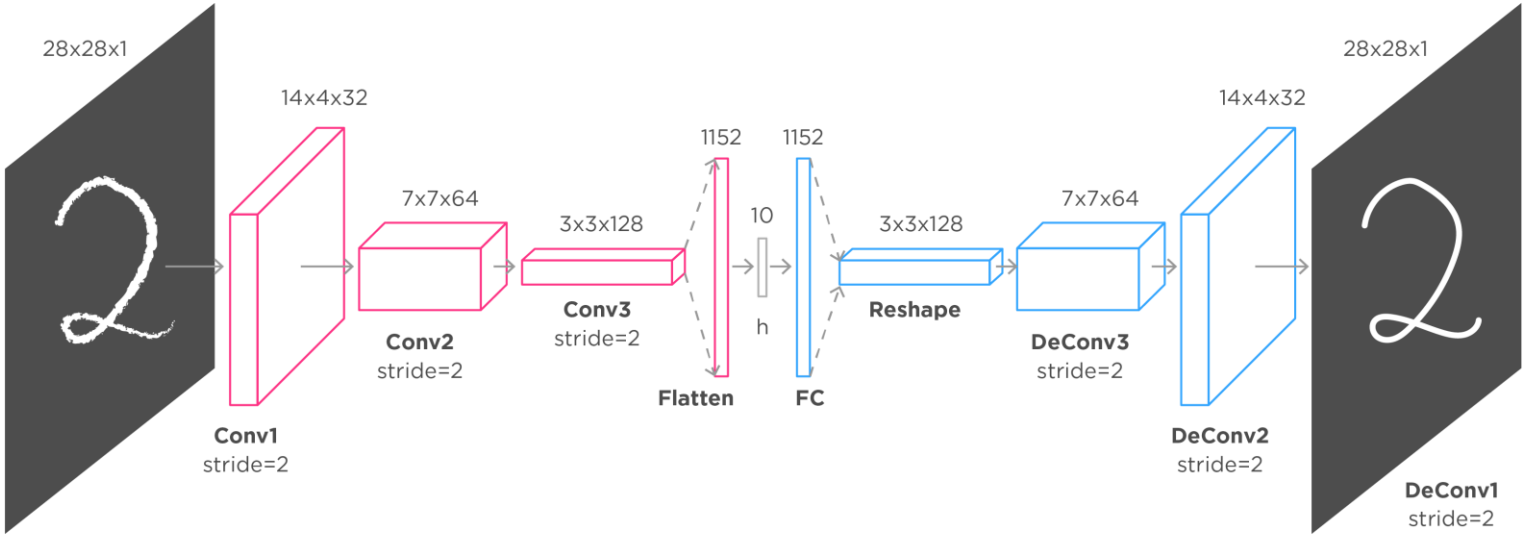


ARQUITETURA **DOS AUTOENCODERS**

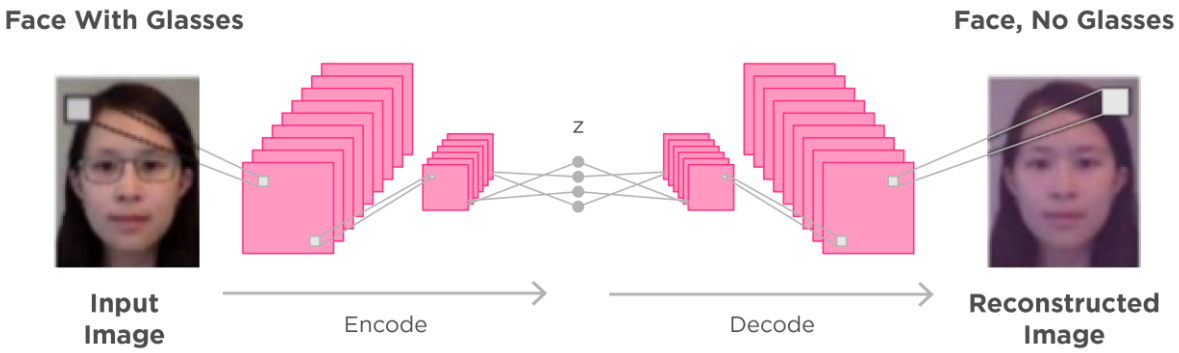




AUTOENCODERS **CONVOLUCIONAIS**



AUTOENCODERS **CONVOLUCIONAIS**

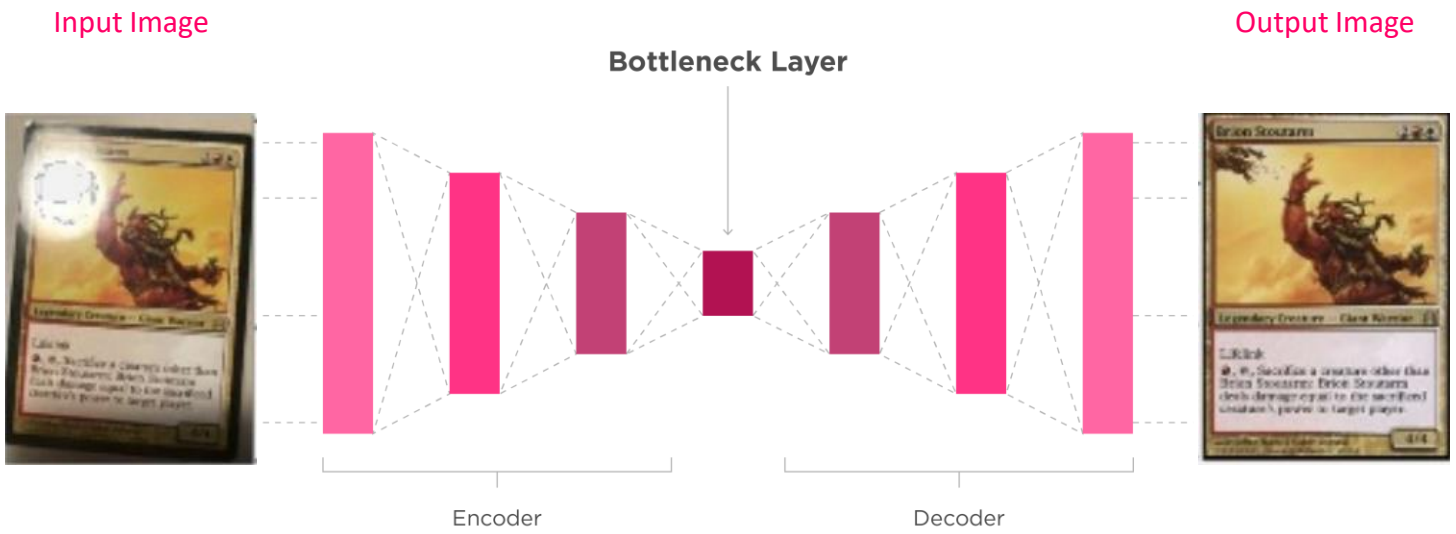


Learns to **remove noise** or **reconstruct** missing parts

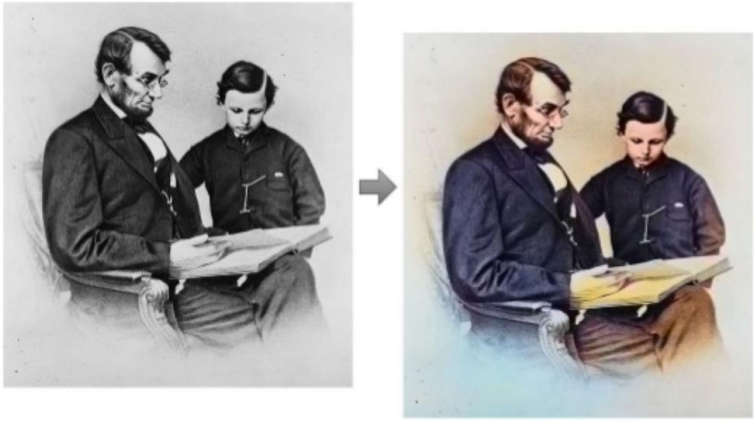
Noisy Version is converted to clean version

The network fills the gaps in the images

AUTOENCODERS **CONVOLUCIONAIS**

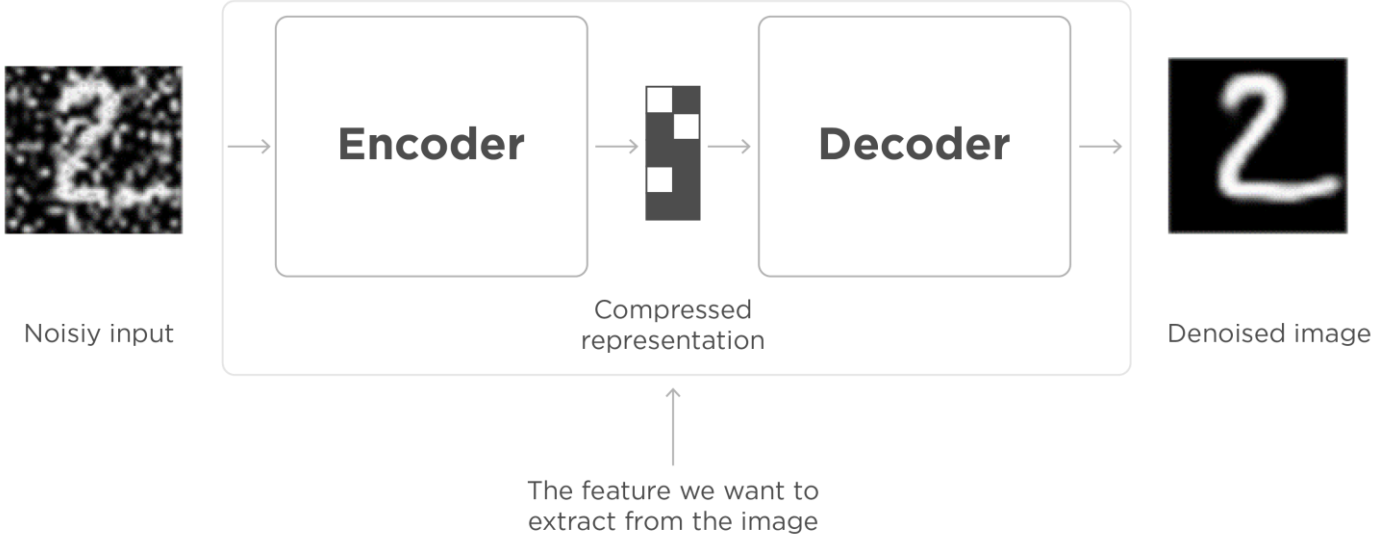


AUTOENCODERS **CONVOLUCIONAIS**



- Maps **circles** and **squares** from na image to the same image but with Colors
- Purple is formed sometimes because of **blend** of colors, where network hesitates between circle or square

AUTOENCODERS **CONVOLUCIONAIS**



AUTOENCODERS

Demonstração simplificada de Autoencoders com dataset MNIST

AUTOENCODERS

Demonstração remoção de ruídos dataset MNIST

Autoencoders com

OBRIGADO

FIAP

FIAP