MBA USP ESALQ

Other Machine Learning Models III

João F. Serrajordia R. de Mello

You will need...

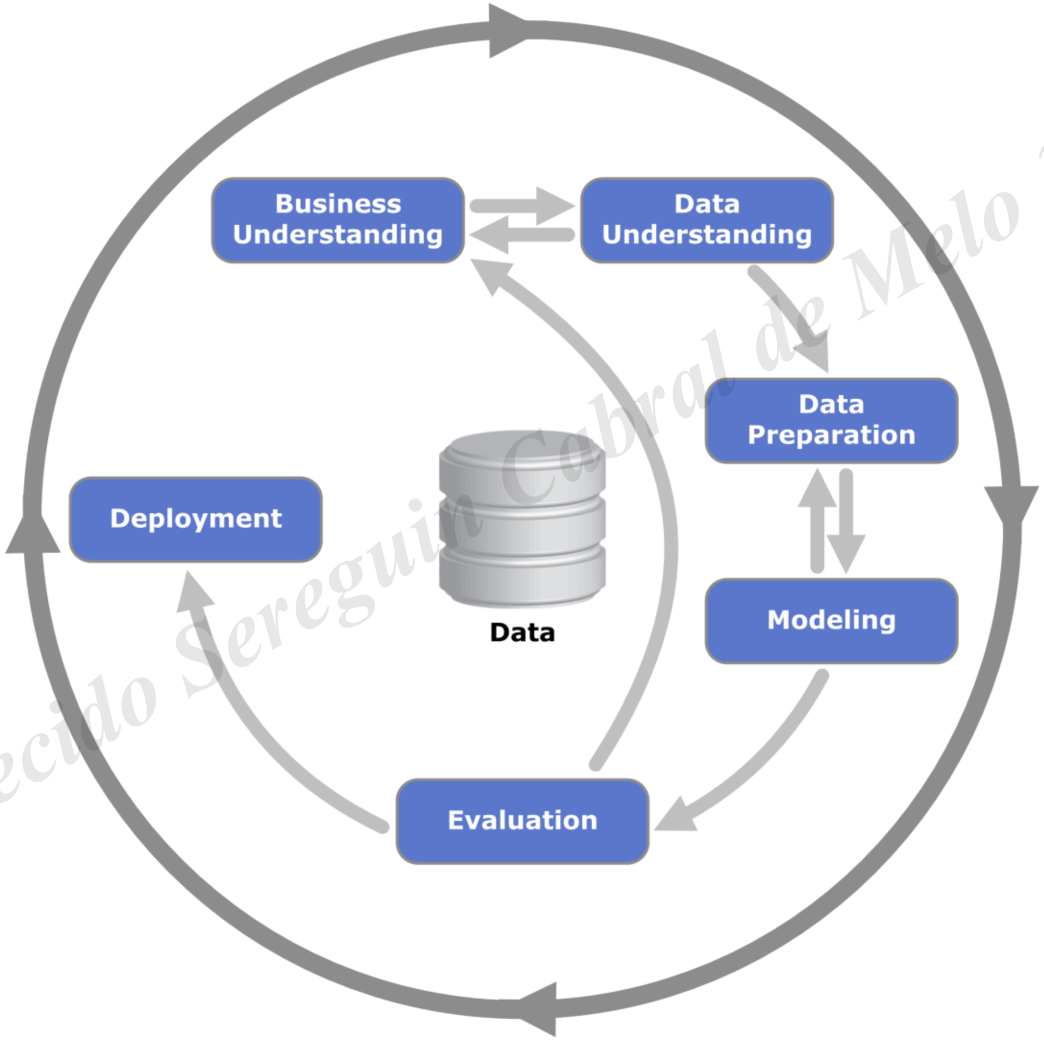
Preparations

- Open R
- Import libraries
- Something to take your notes

Agenda

- Review
- *Historic*
- *Basic ideas*
- *Uses*

CRISP-DM



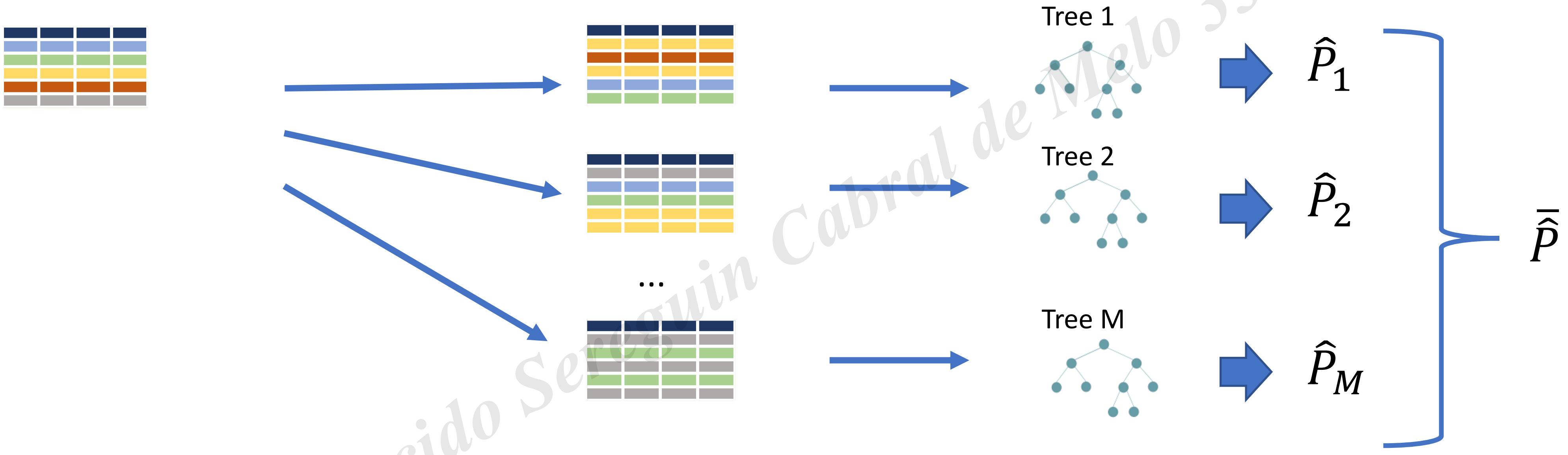
Source: https://www.the-modeling-agency.com/crisp-dm.pdf

Ensemble

An ensemble is any combination of existing models. The main types are:

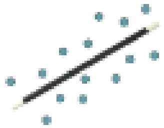
- *Bagging*
- *Boosting*
- *Stacking*

Bootstrap – aggregation (bagging)

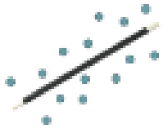


Bagging with trees is the famous *Random Forest*

ID	...	Y
1	...	1
2	...	0
...	...	...
N	...	0



Y	P	ERRO
1	75%	25%
0	20%	20%
...	...	...
0	40%	40%



ERRO	Δ	P	ERRO
25%	10%	85%	15%
-20%	-10%	10%	-10%
...	...	...	...
-40%	-15%	25%	-25%

ERRO	Δ	P	ERRO
15%	2%	87%	5%
-10%	-1%	9%	5%
...	...	...	...
-25%	-5%	20%	10%

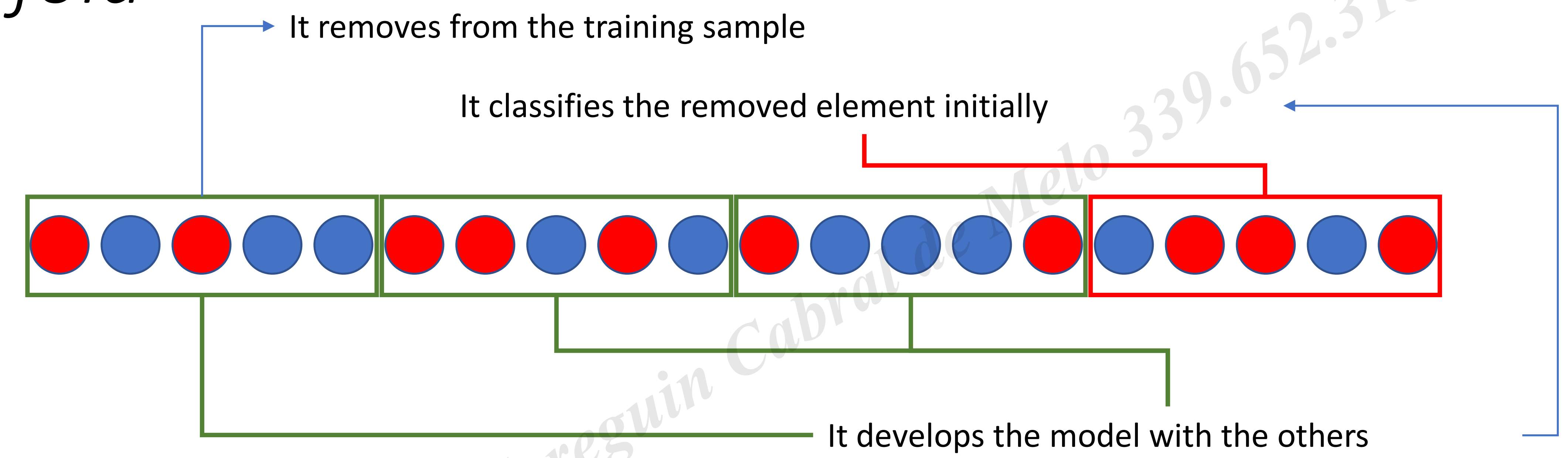
The response variable of an iteration is the 'error' of the previous one.

The response variable of an iteration is the 'error' of the previous one.

Boosting

- *Boosting* methods are sequential models that try to improve the error of the previous model

K-fold



- We divide the base into sub-samples k
- For each sub-sample:
    - We remove the sub-sample as validation
    - We train the model with the remaining observations
    - We use this model to classify the removed sub-sample
    - We evaluate the metrics of the model's performance
- We calculate the average of the metrics of the model's performance

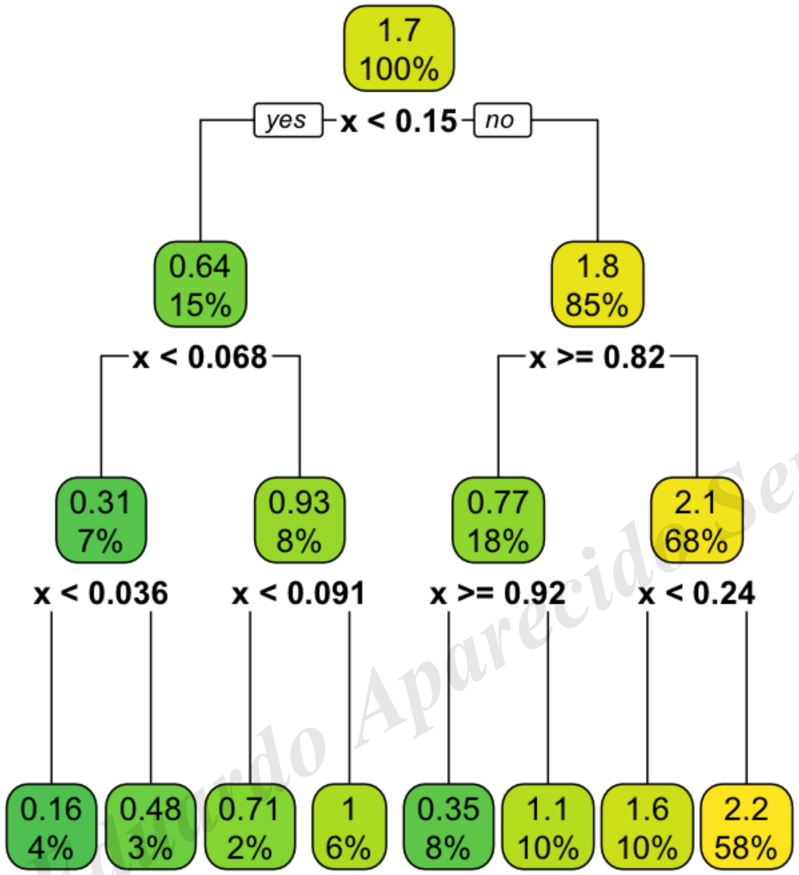
Regression trees

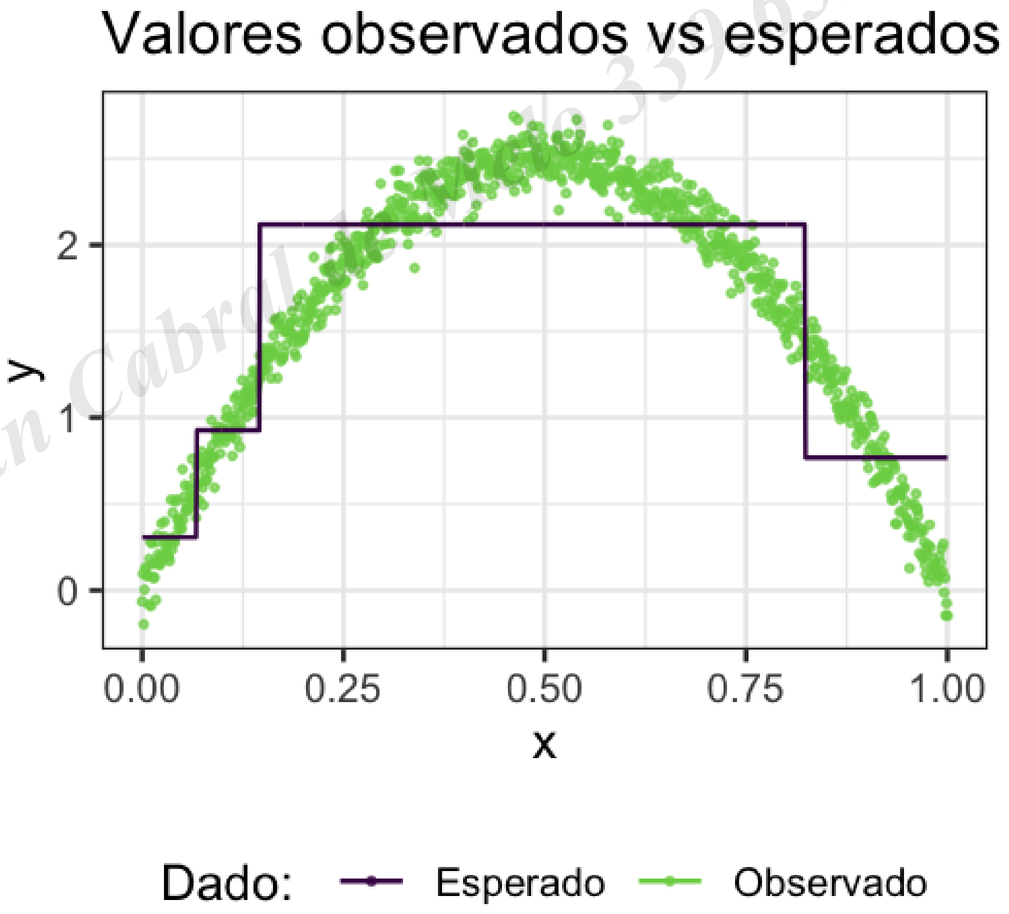
They are very similar to classification trees

The criterion of impurity is what changes.

$$SQE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Regression trees





Predictive and classification problems

What is the efficacy of a vaccine?

Will the customer pay the loan?

How much oil is in the well?

Will the customer buy my product?

What is the person doing?

How green is this vehicle?

Classification

Algorithms classification

Supervised

- Regression
- GLM
- GLMM
- Support vector machines
- Naive Bayes
- K-nearest neighbors
- Neural Networks
- Decision Trees

Unsupervised

- K-Means
- Hierarchical methods
- Gaussian Mixture
- DBScan
- Mini-Batch-K-Means

We are here!

Algorithms classification

Continuous response

- Regression
- GLM
- GLMM
- Support vector machines
- K-nearest neighbors
- Neural Networks
- Regression Trees

Discrete response

- Logistic Regression
- Classification trees
- Neural Networks
- GLM
- GLMM

We are here!

Algorithms classification

Machine Learning Methods

- Decision Trees
- Bagging
- Boosting
- K-NN
- Neural Networks
- Support Vector Machines

Machine Learning Statistics Methods

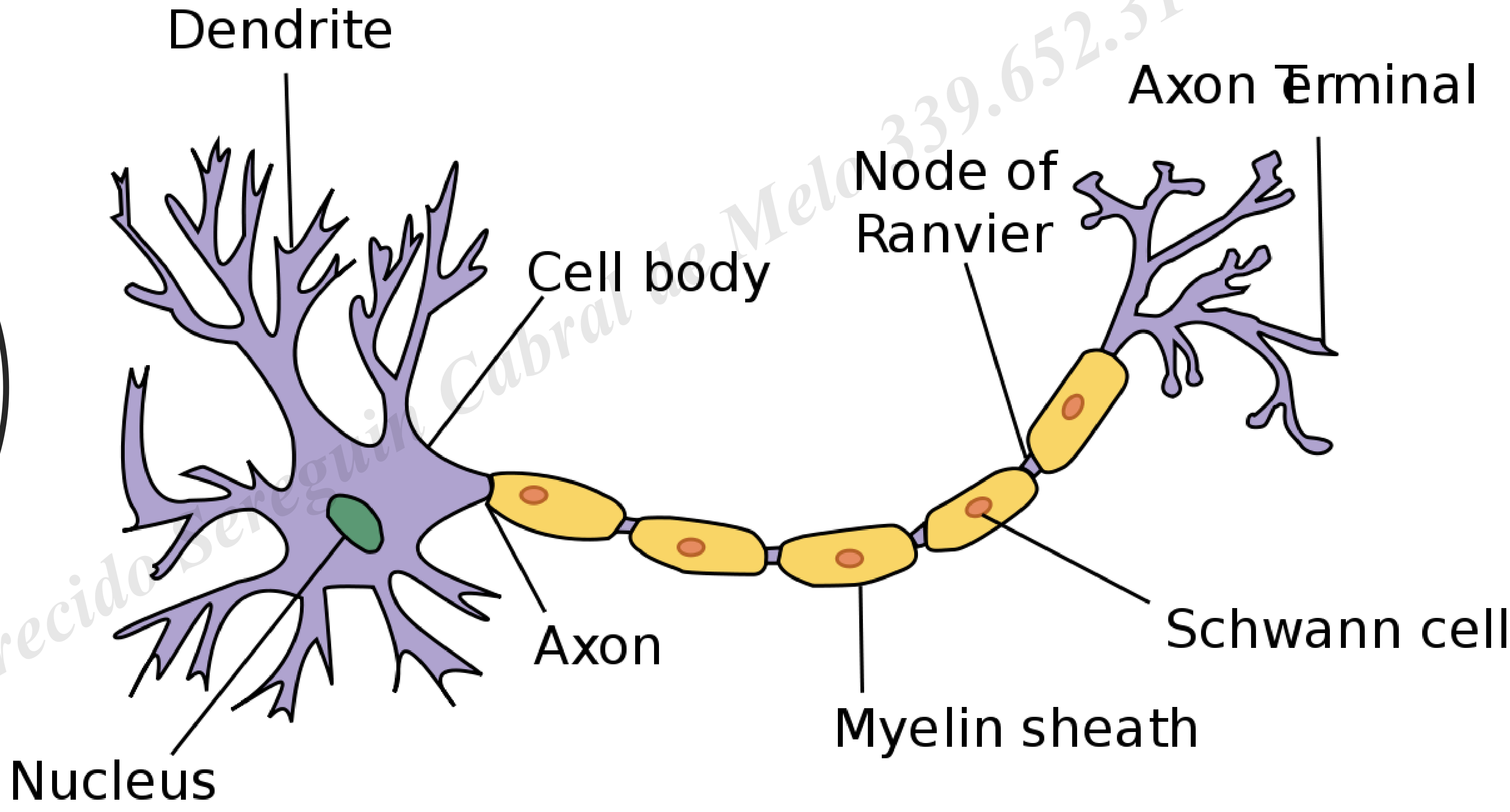
- Regression
- GLM
- GLMM
- ANOVA

We are here!

Artificial Neural Networks

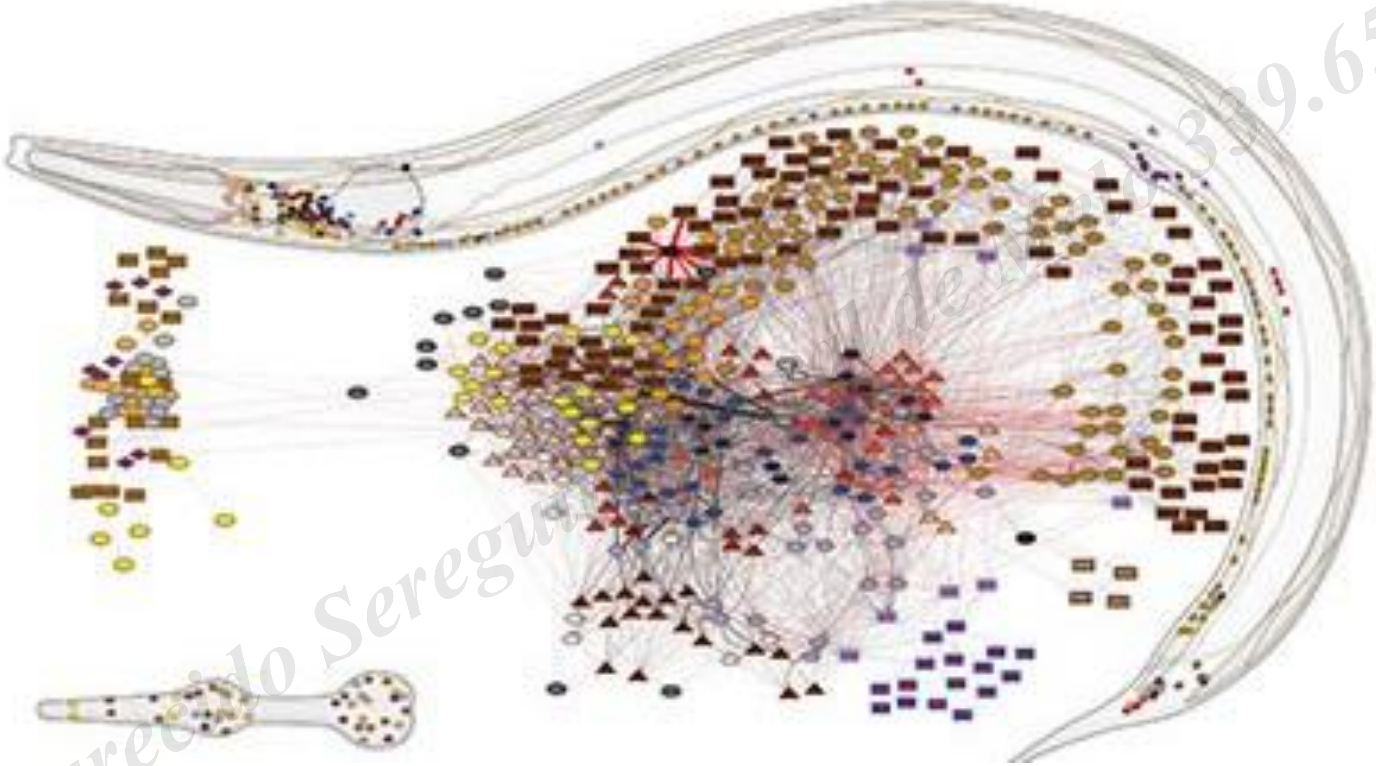
System

Neurons



https://en.wikipedia.org/wiki/Myelin

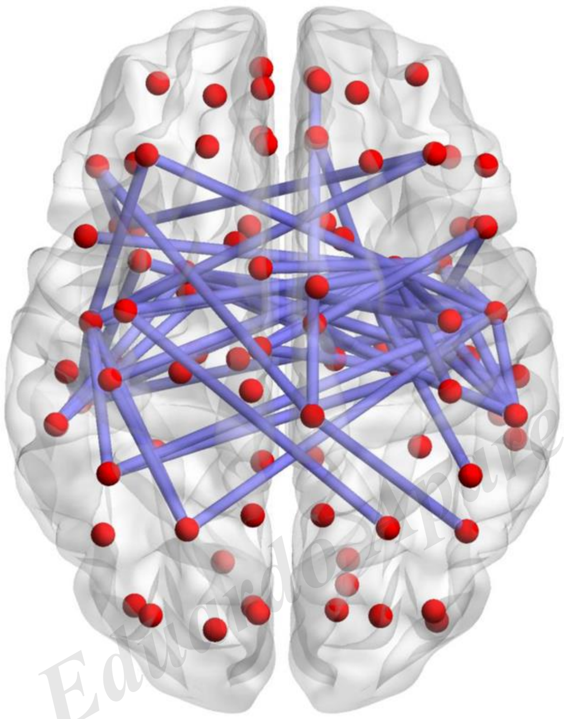
Biological Example

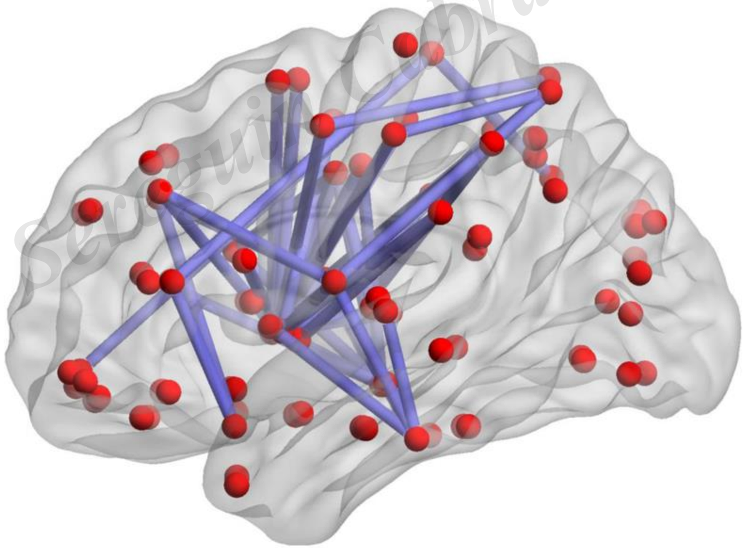


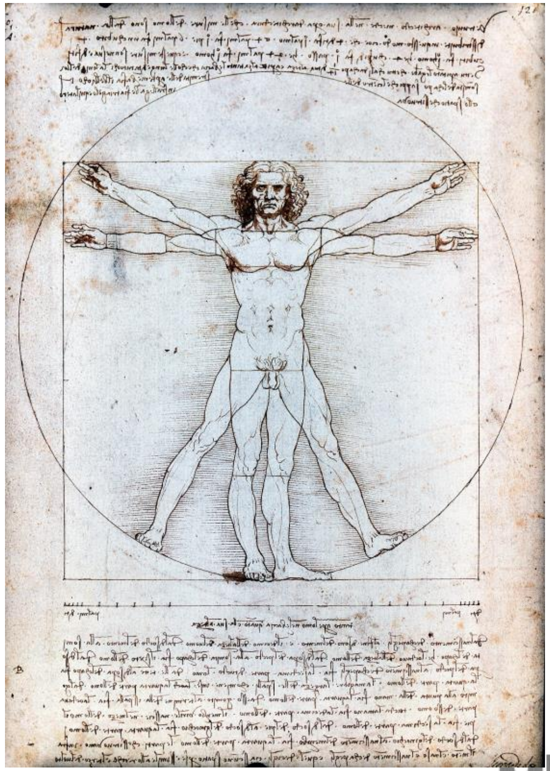
- Roundworm: 302 neurons

Human Neural Network

- *Homo sapiens*: 100.000.000.000 neurons





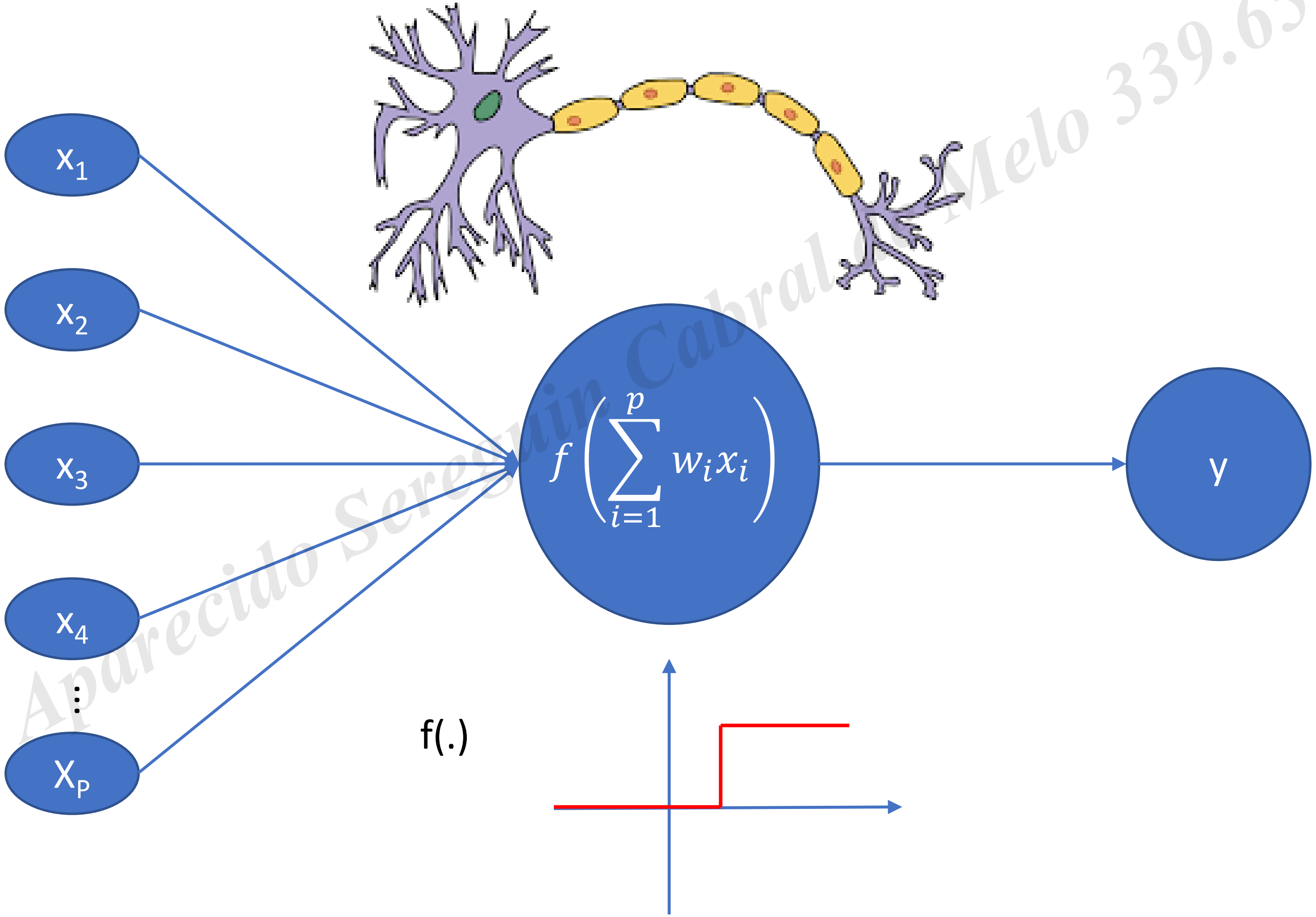


Where do they live?

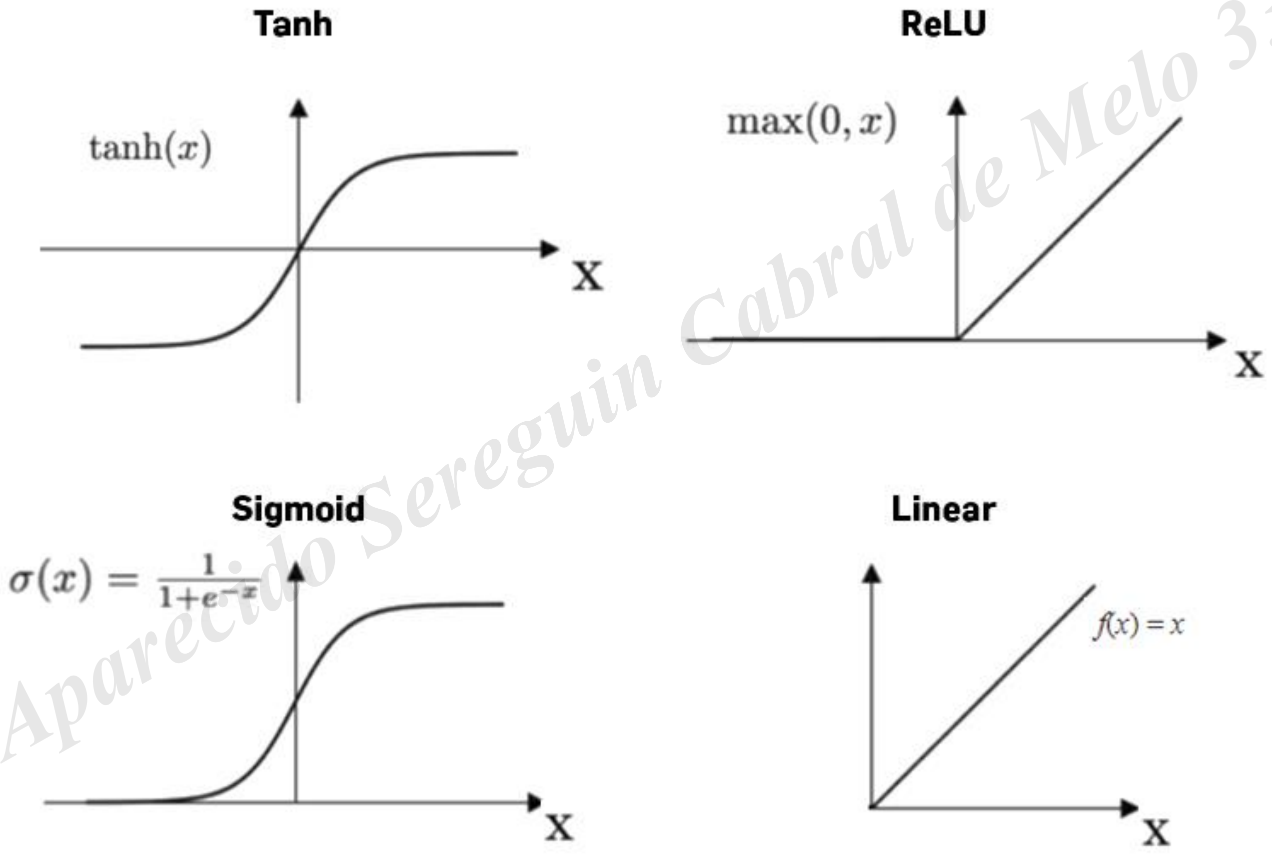
  

Artificial Neural Networks have been very successful in problems with little structured data such as images, audios, texts, and videos.

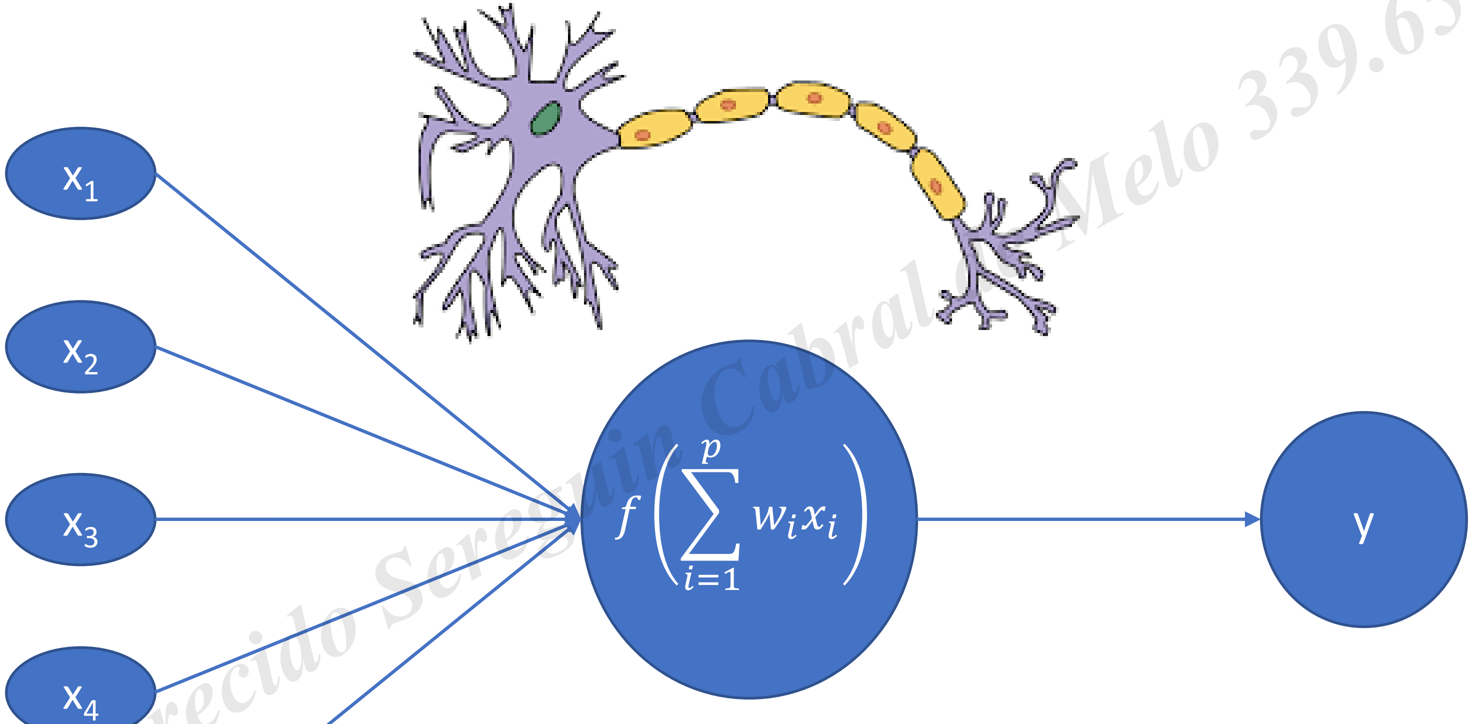
McCulloch-Pitts Neurons



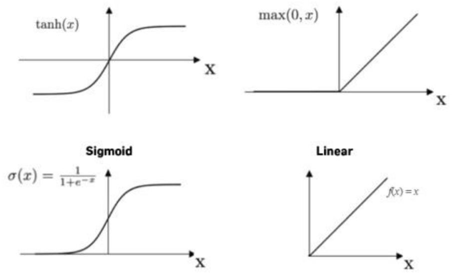
Activation Functions



Perceptron



x1
x2
x3
x4
⋮
Xp
$f\left(\sum_{i=1}^{p} w_i x_i\right)$
y

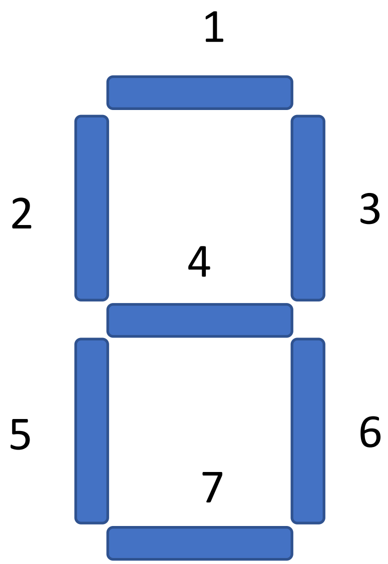


tanh(x)
x
max(0, x)
x
Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$
x
Linear
f(x) = x
x

OCR – Optical Character Recognition

Let's think about a very simple version of the problem. Digits of an old clock have a very simple structure.

OCR – Optical Character Recognition



There are 7 basic regions, which can be active or inactive, and they define a digit.

For example, if only regions 1, 3 and 6 are activated, we have the number 7.

Rosenblatt's Perceptron

- The Rosenblatt's Perceptron (~1950-1960) has this idea, but only with a general purpose
- it was built to perform OCF (optical character recognition)
- For this, it maps regions of an image as "activate" and "inactive"
- Each unit is a McCullogh-Pitt's neuron

FIG. 1 — Organization of a biological brain. (Red areas indicate active cells, responding to the letter X.)

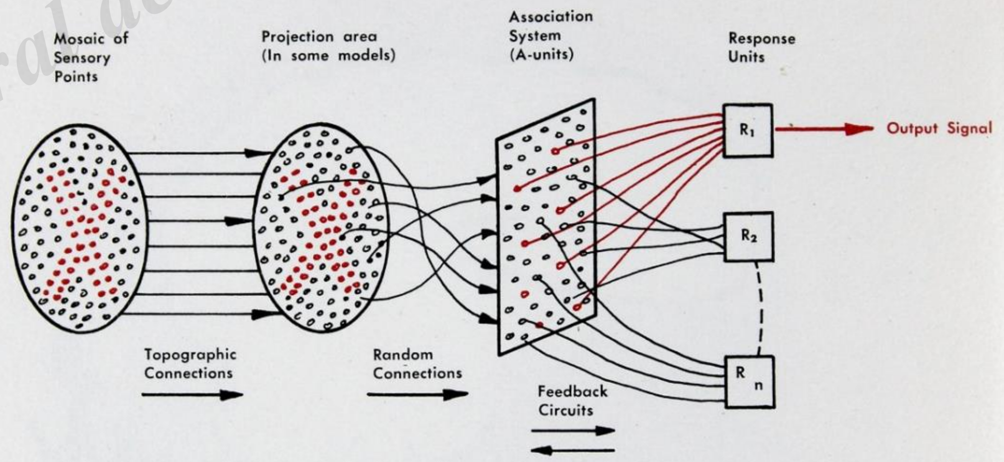
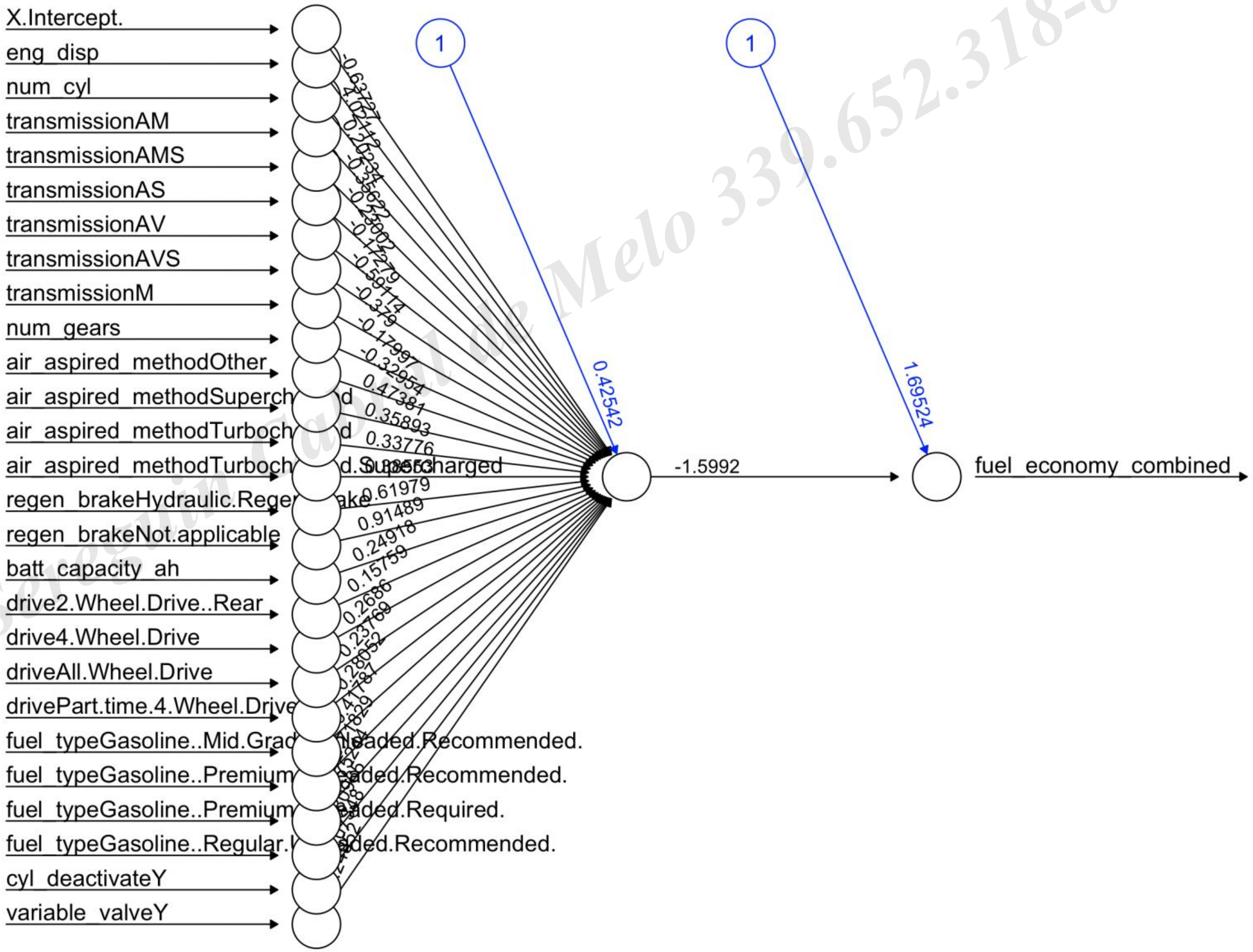


FIG. 2 — Organization of a perceptron.

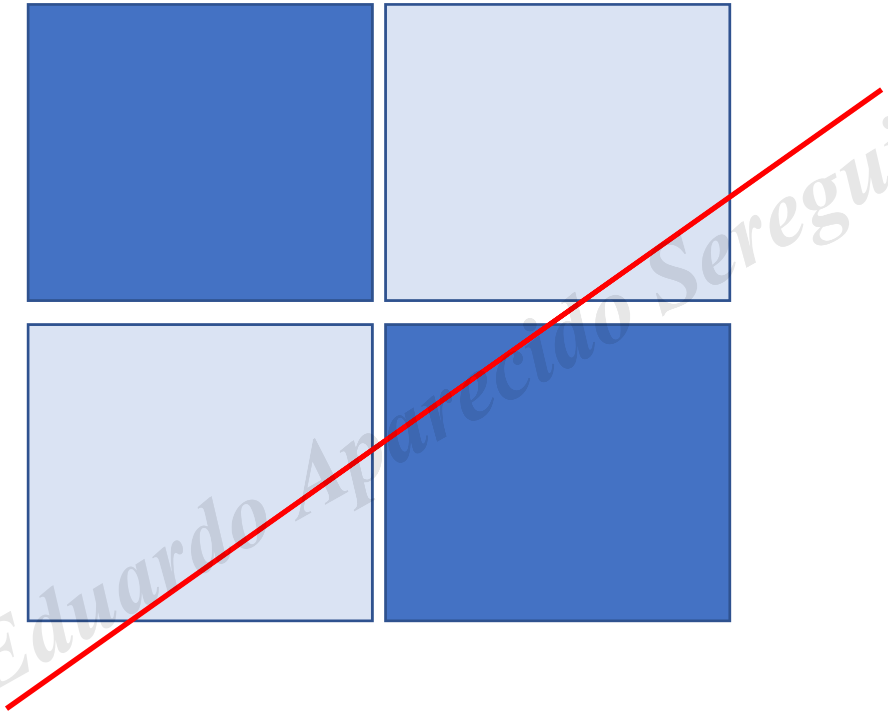
Linear Perceptron

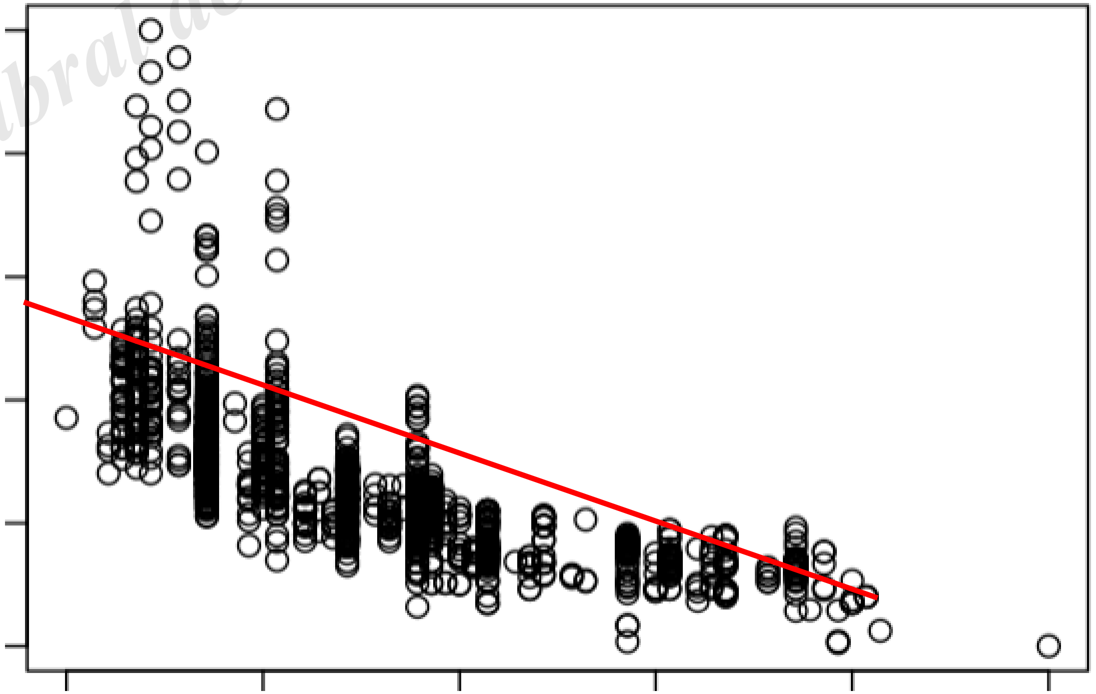
- It has the same structure as a linear regression with the activation function indicated.



Limitations of linear perceptron

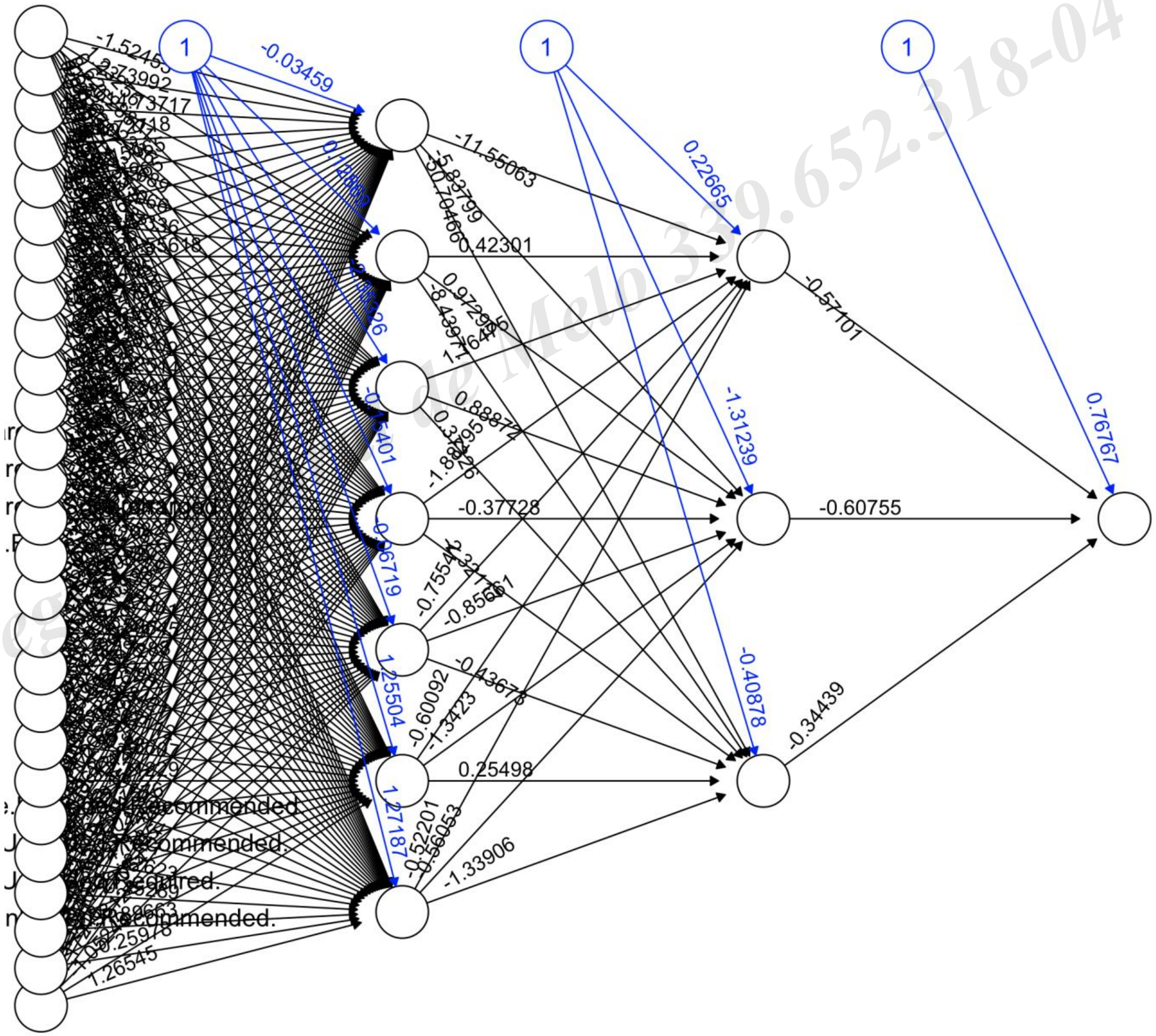
- Linear perceptron only captures linear standards





Multilayer Perceptron

- It has intermediate "hidden" layers
- It captures nonlinear standards
- It can make use of the parallel processing of GPUs
- It is not “interpretable” as regression



Loss Functions

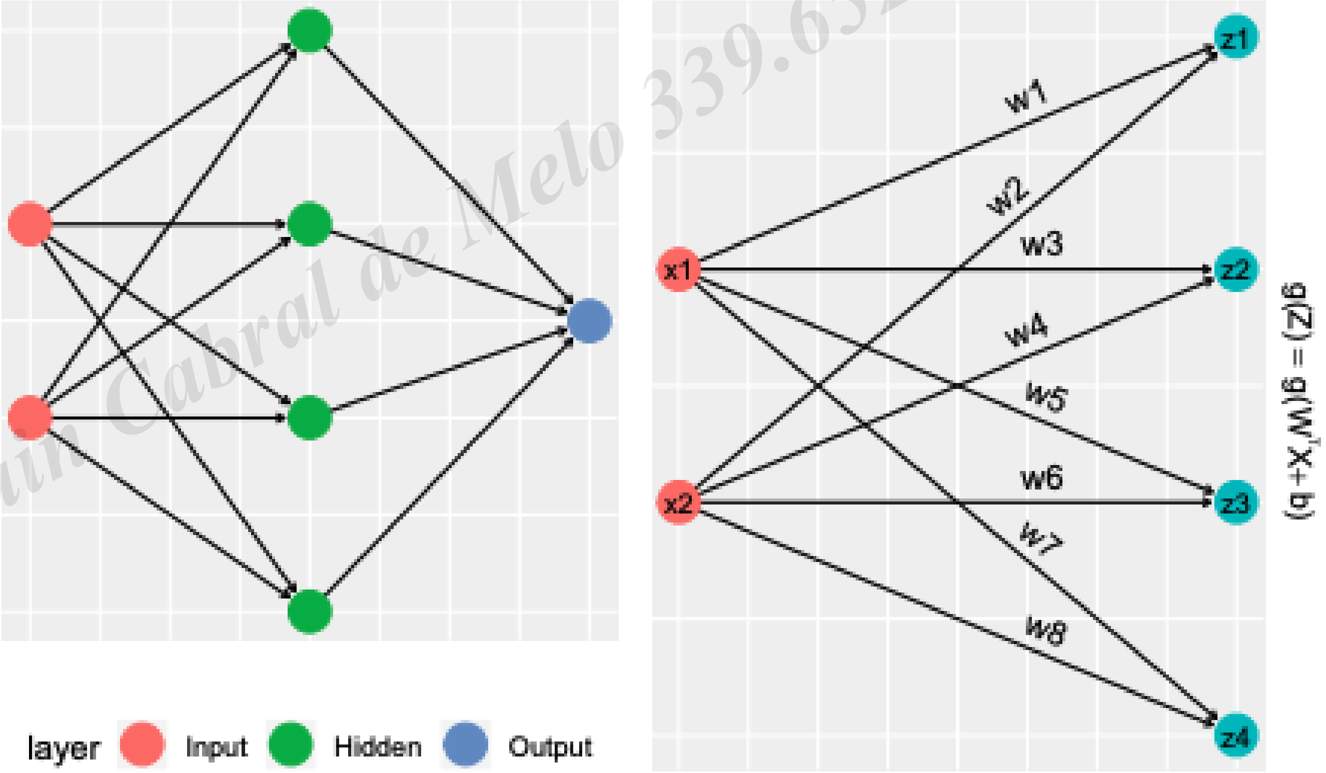
Continuous Variables
SQE

$$SQE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Binary Variables
Cross-Entropy

$$L = y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)$$

Artificial Neural Networks



Deep learning with R - Abhijit Ghatak, ed. Springer, 2019

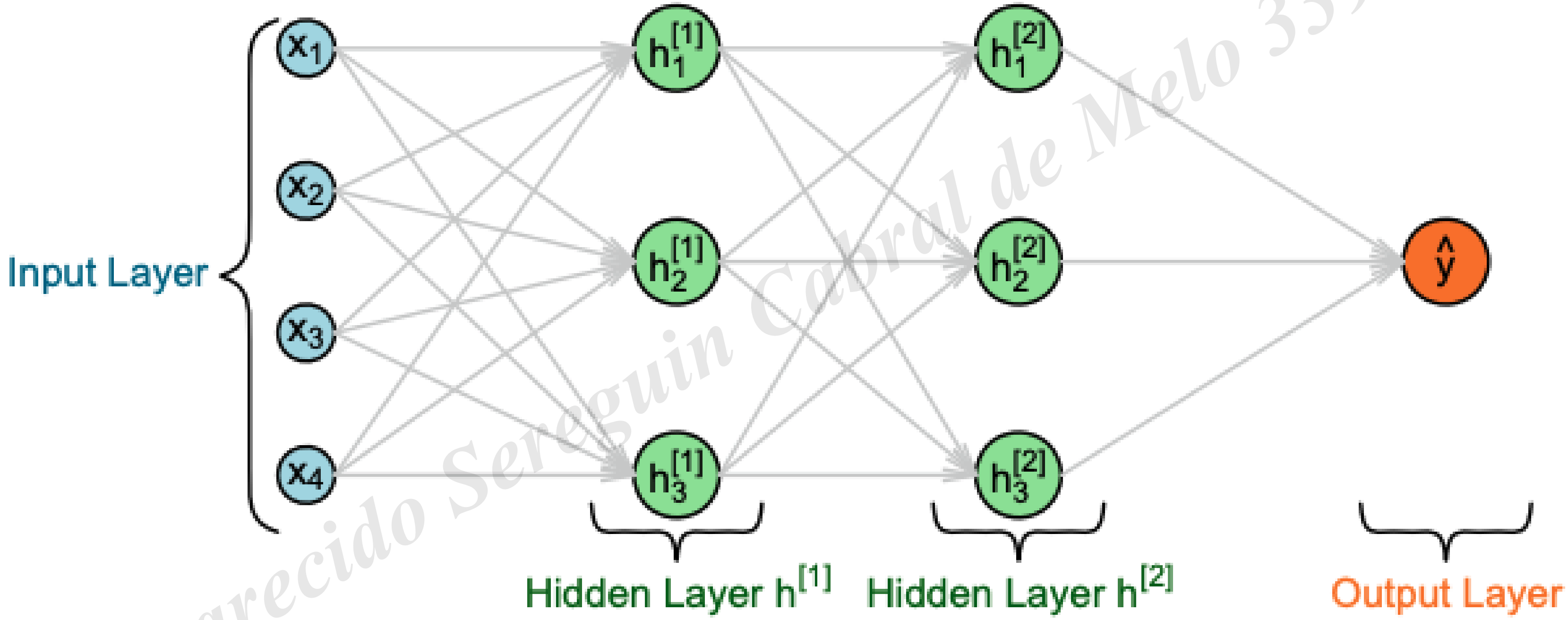
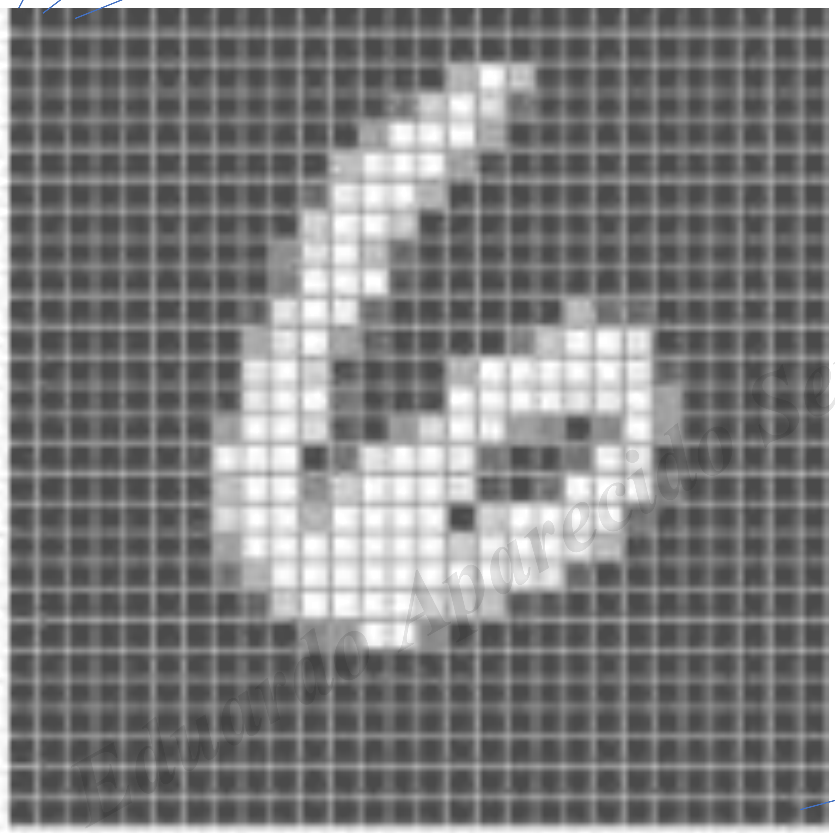


Fig. 2.3 A representation of a neural network with four input features, two hidden layers with three nodes each, and an output layer

Deep learning with R - Abhijit Ghatak, ed. Springer, 2019

Initial treatment of data

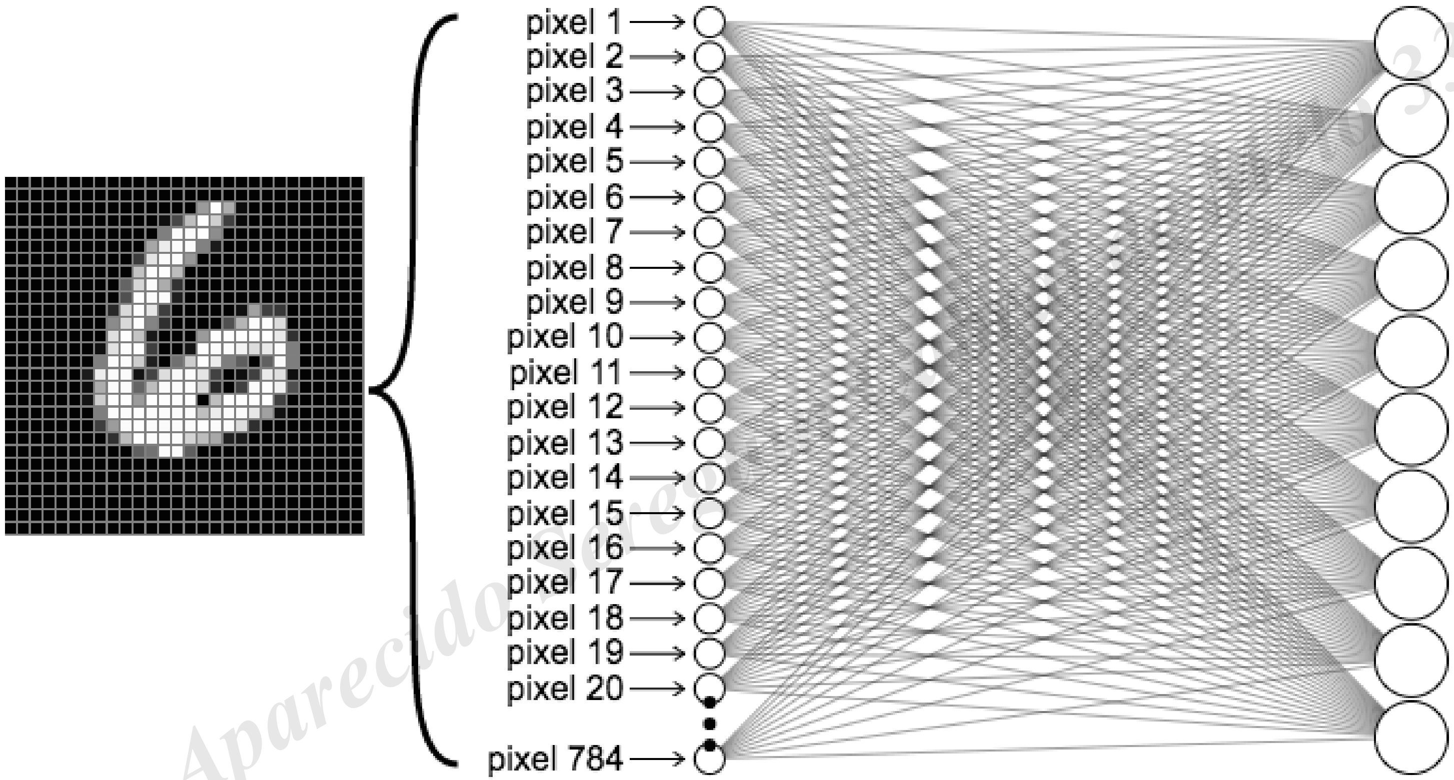


60.000 of these

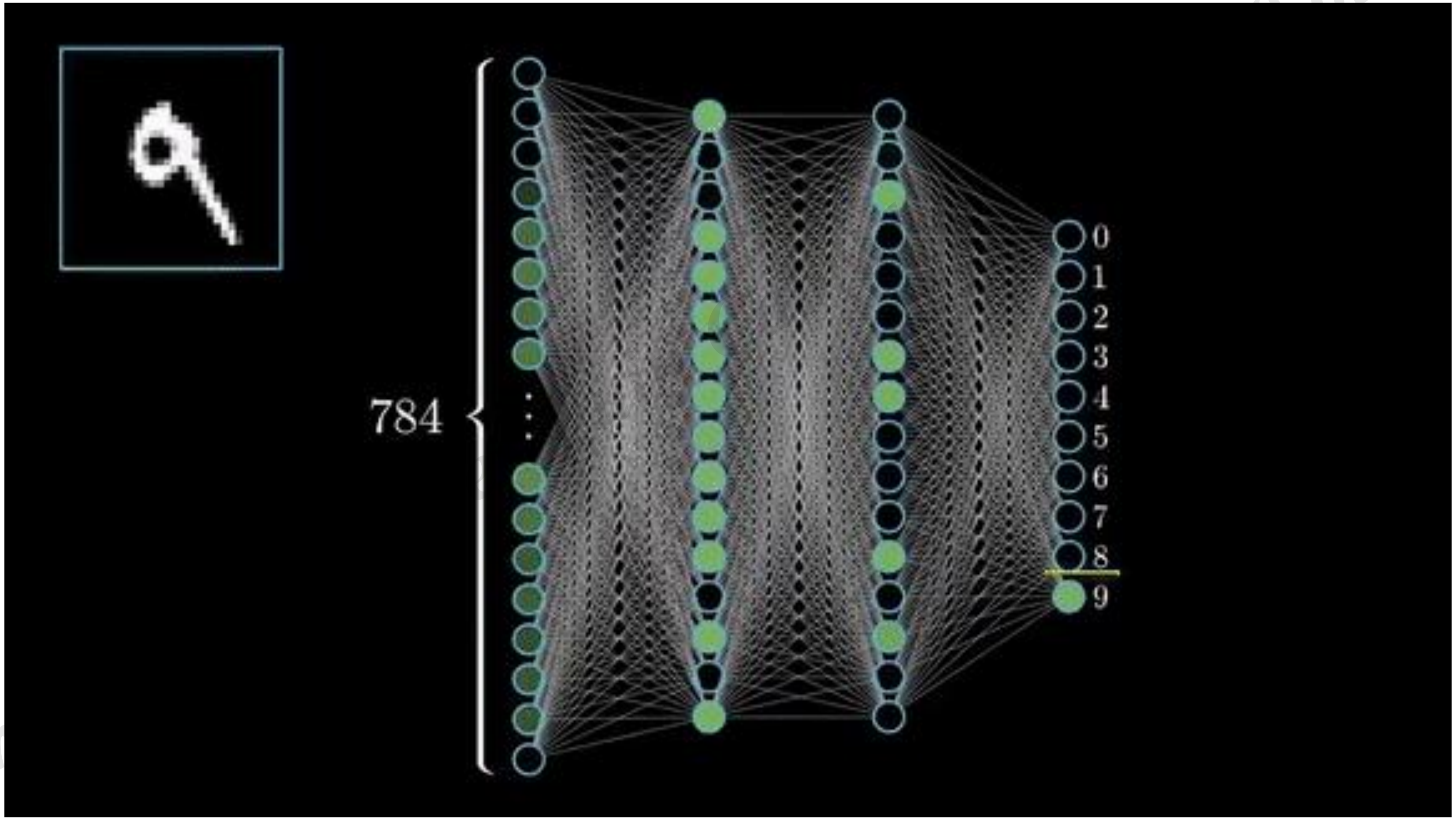
Turn into a table like this

	Pixel 1	Pixel 2	Pixel 3	...	Pixel 784	Label
Image 1	0	0	0...		0	4
Image 2	0	0	0...		0	3
Image 3	0	0	0...		0	9
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Image 60.000	0	0	0...		0	5

Neural Network on MNIST



We have 784 x 10 = 7.840 parameters with only one layer!



3blue1brown - https://www.youtube.com/watch?v=aircAruvnKk

Gradient Descent

It is the most popular algorithm to train artificial neural networks since it presents some characteristics:

- It can change the estimates with small subsets of points to each iteration (in the limit of 1 only point)
- It does not depend on the inversion of the matrix
- It works with a very large database
- It can be processed in parallel with GPU
- It allows to interrupt the algorithm to a certain point, or to continue later or in another similar problem (*transfer learning*)

Gradient Descent in Networks

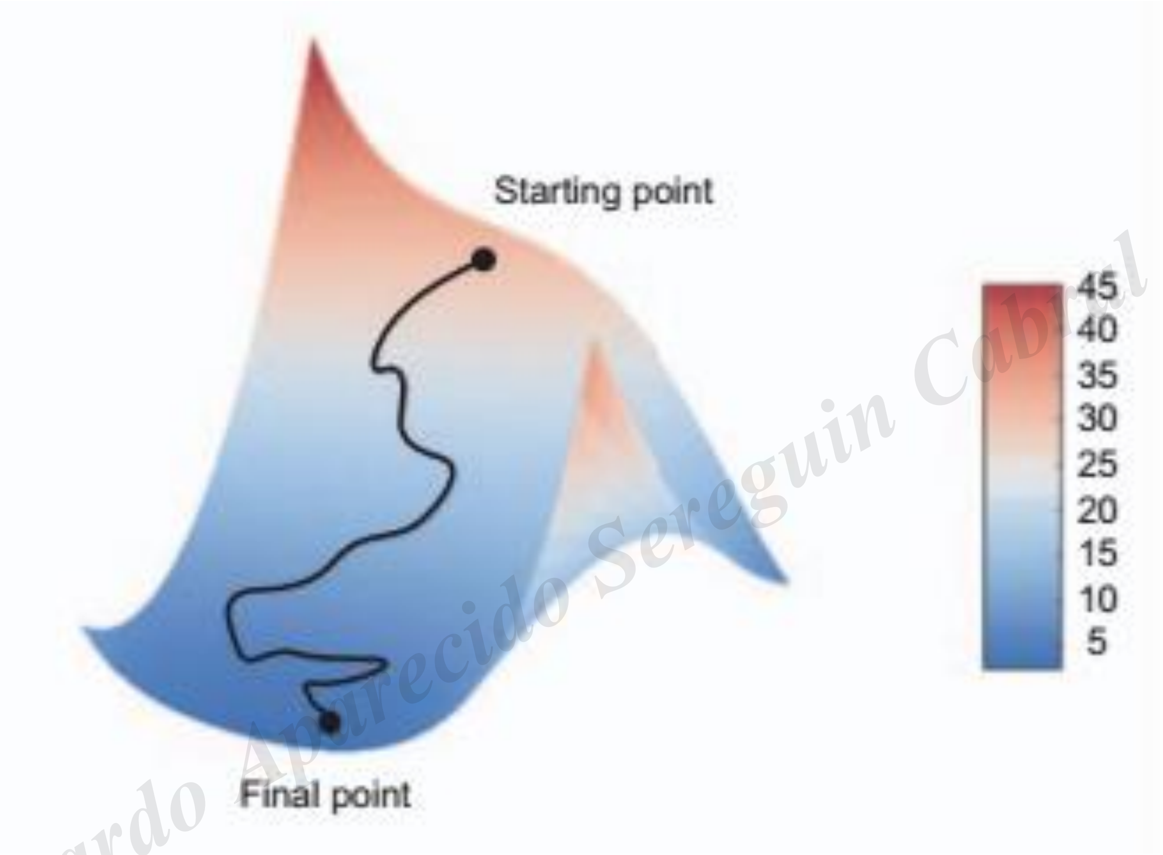


Figure 2.12 **Gradient descent down a 2D loss surface (two learnable parameters)**

Deep learning with python – François Chollet

Gradient descent

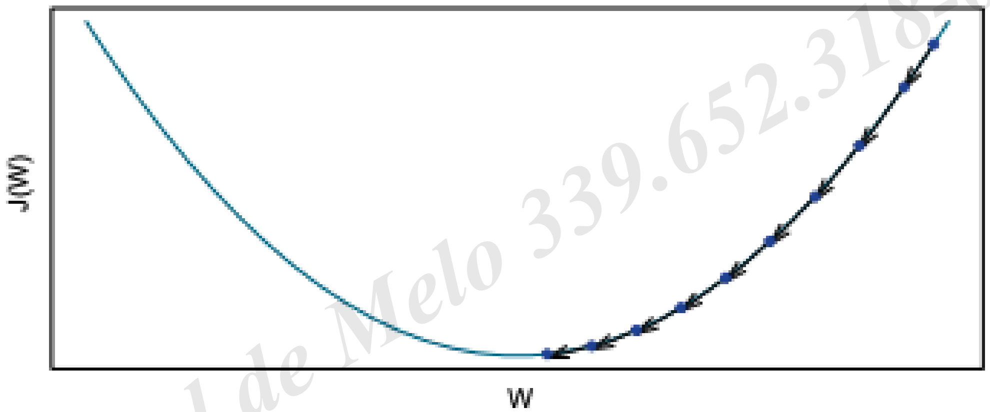


Fig. 1.4 Gradient descent: Rolling down to the minima by updating the weights by the gradient of the loss function

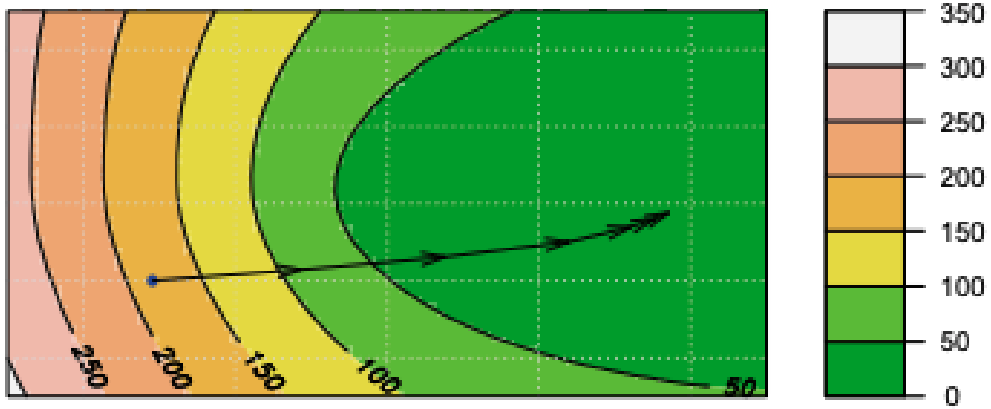
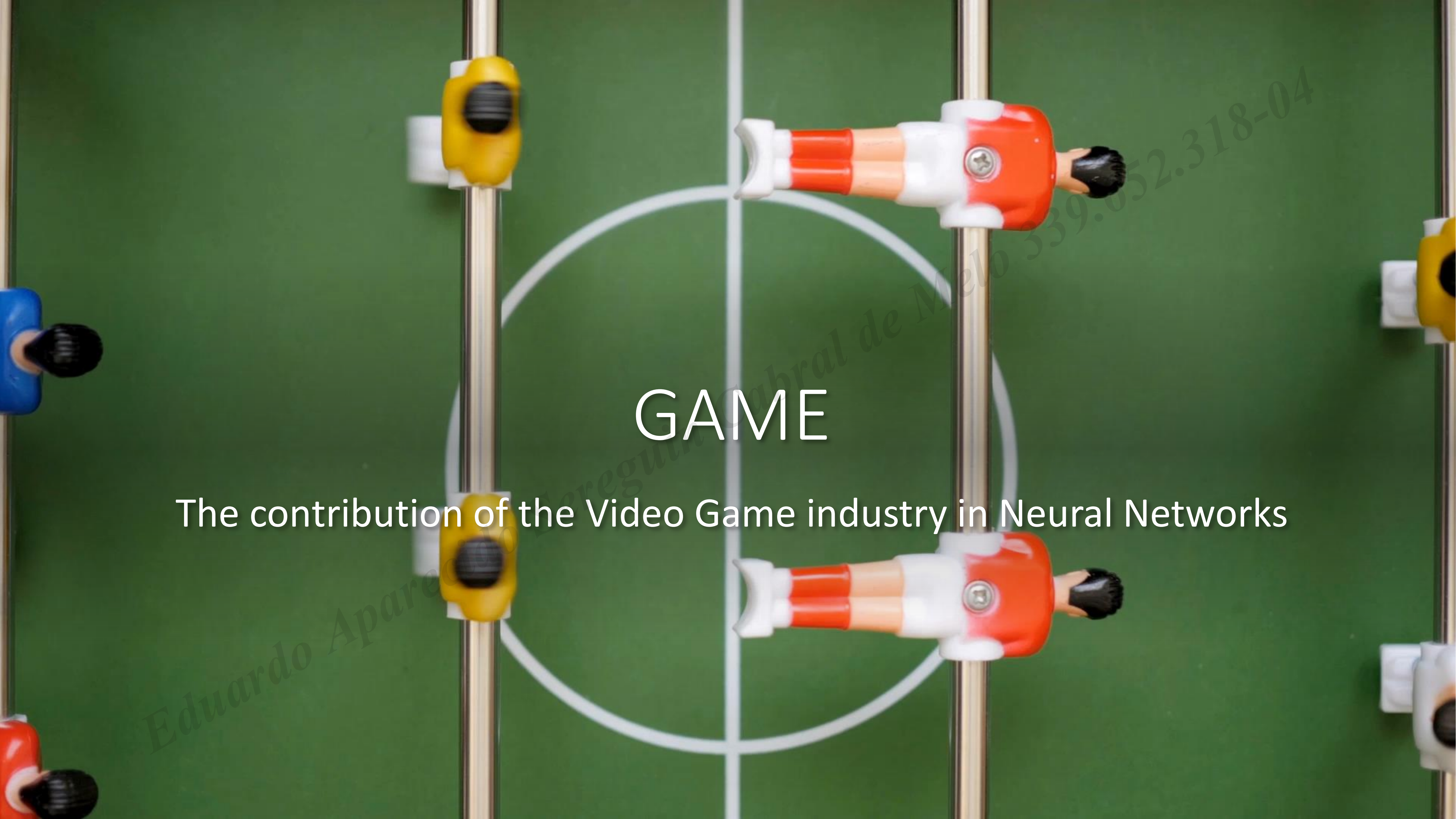


Fig. 1.5 A contour plot showing the cost contours of a sigmoid activation neural network and the cost minimization steps using the gradient descent optimization function

Vehicle consumption prediction

- Engine's size
- Fuel.
- Number of cylinders
- Brand
- Power of the engine
- Traction

https://www.fueleconomy.gov/feg/download.shtml

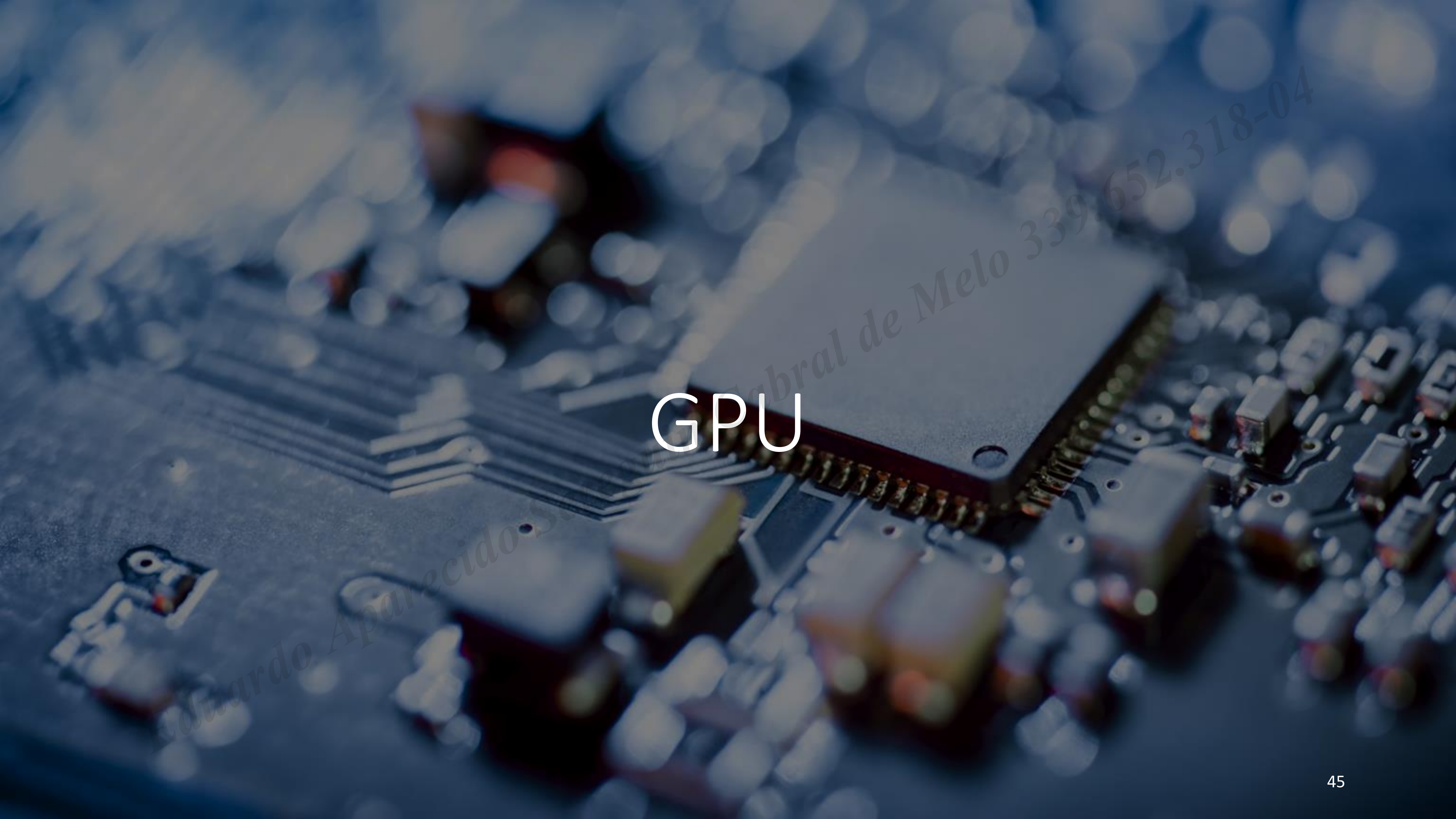


GAME

The contribution of the Video Game industry in Neural Networks

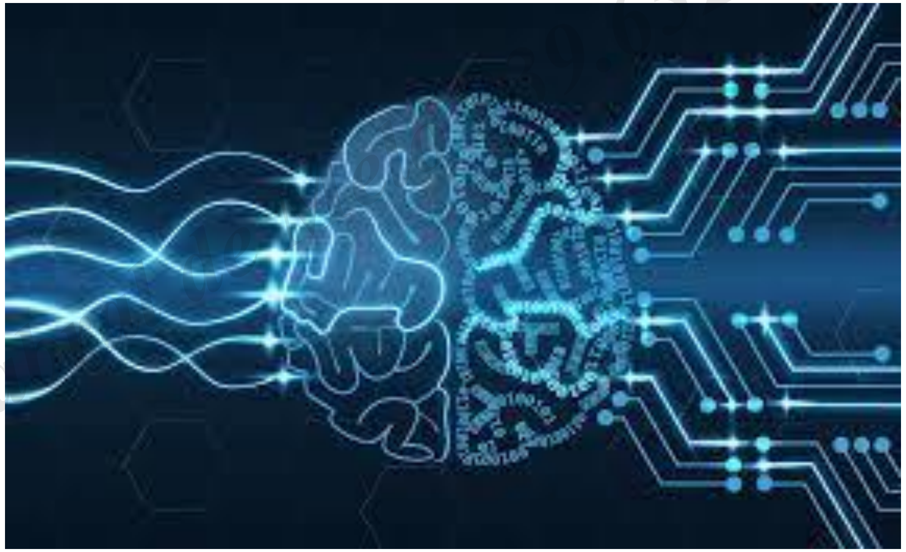
Processers

- Distance between transistors: 14 nm
- Strand of human hair: 80.000 nm
- Gold atom diameter: 0.3 nm



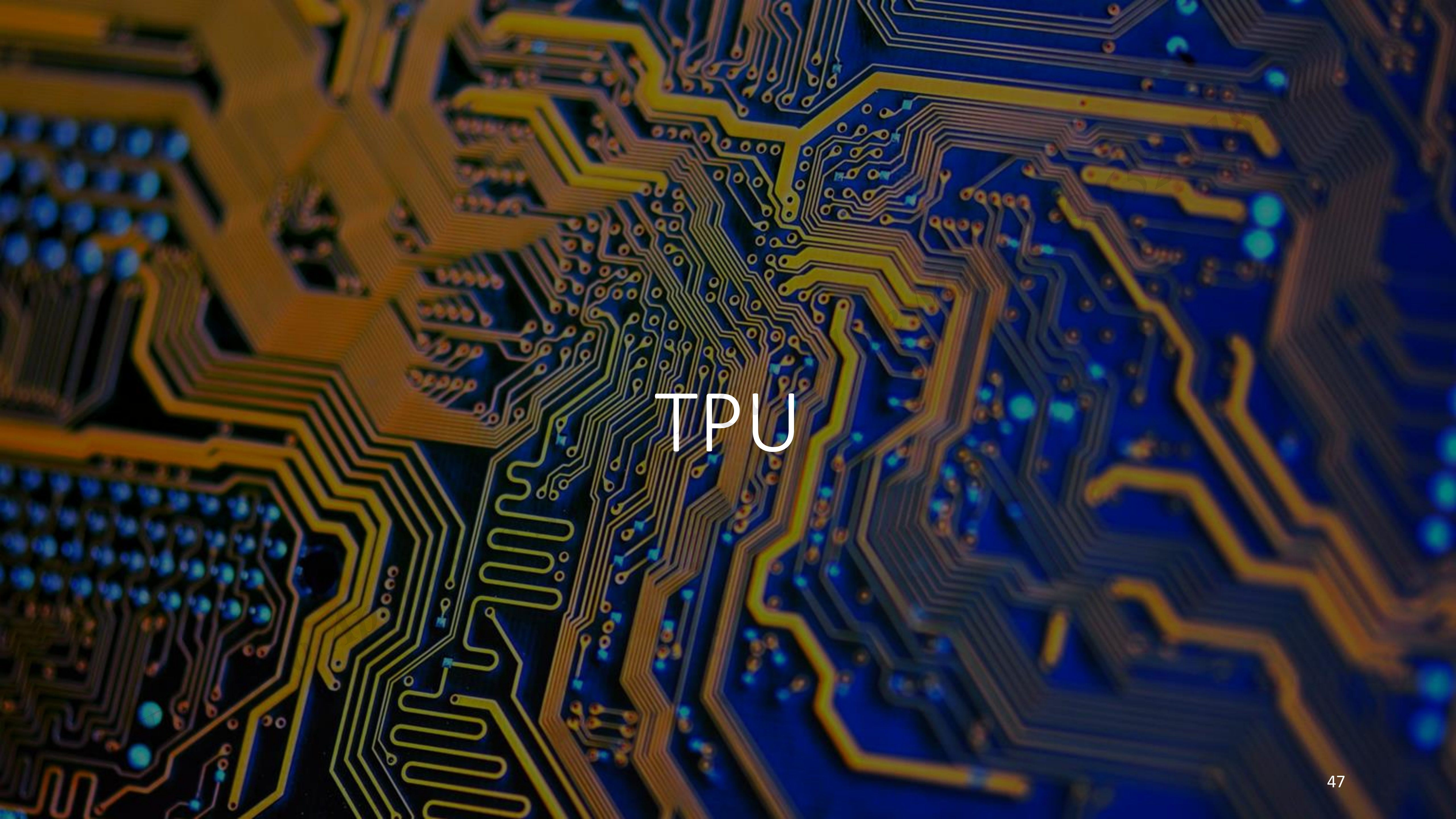
GPU

Processing with GPU



NVIDIA CUDA

AMD ROCm



TPU

L2 Regularization

Continuous Variables
SQE

$$SQE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum \beta_i^2$$

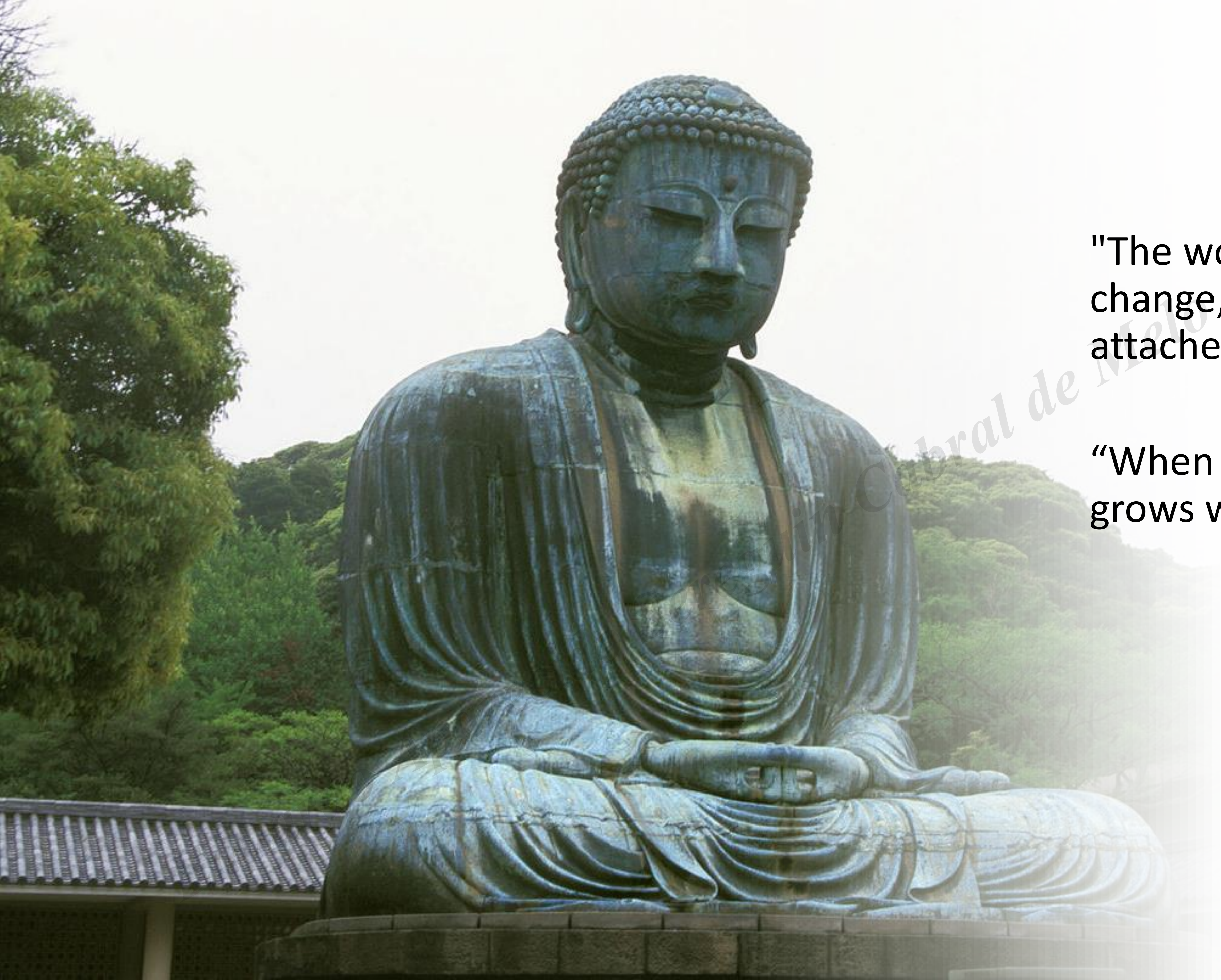
Binary Variables
Cross-Entropy

$$L = \sum y_i log(\hat{y}_i) + \lambda \sum \beta_i^2$$

Recognition of human activity with the smartphone

Conclusions

- Neural Networks are the introduction to Deep Learning (which is a very promising field)
- They are powerful and flexible
- They require special computational power (GPU / TPU)
- They are famous in less structured data (e.g. images, audios)



"The world is in constant change, so don't get too attached."

“When you grow, the world grows with you"

Sidarta Gautama

That's it for today ;)

linkedin.com/in/joao-serrajordia