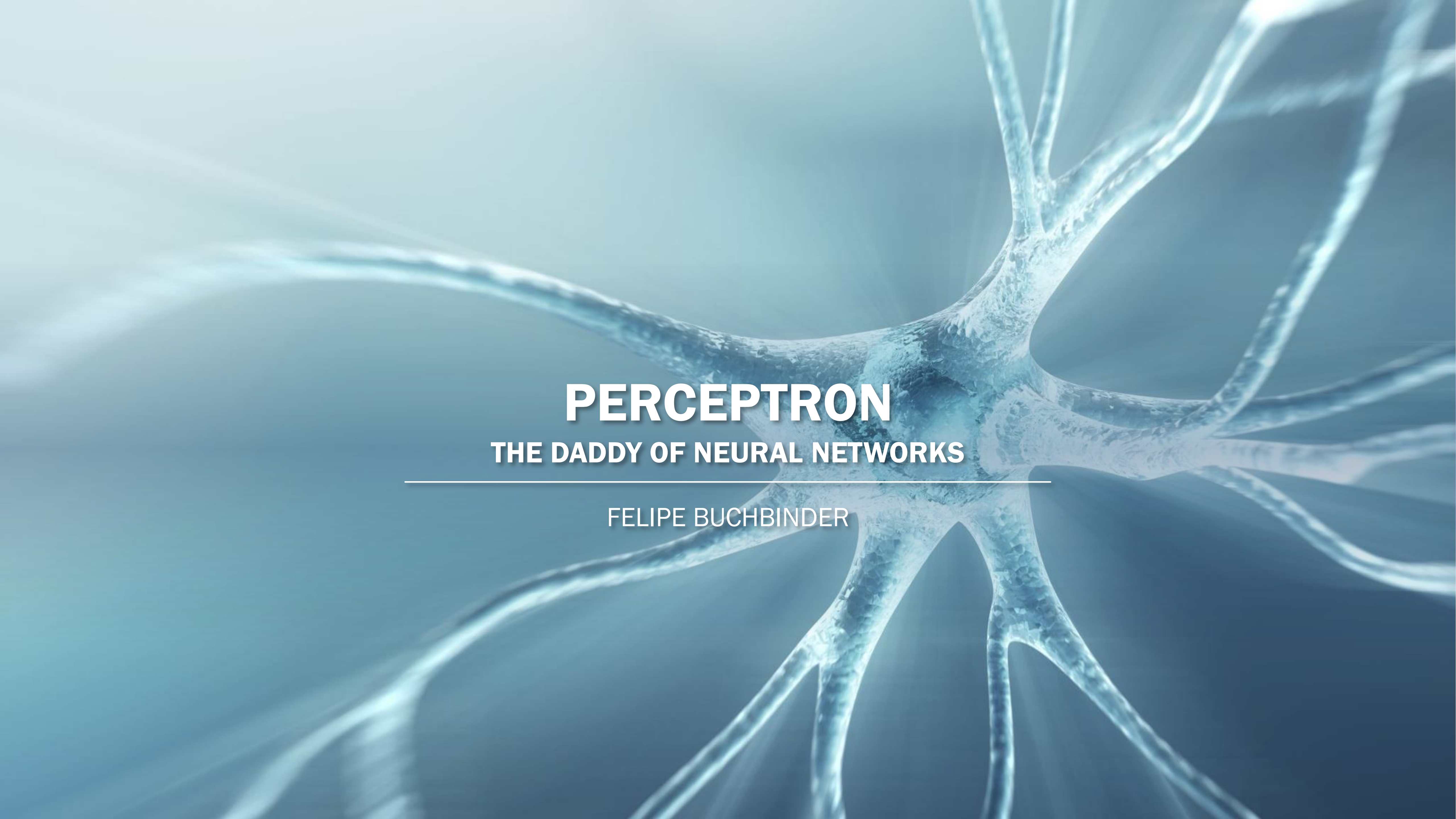
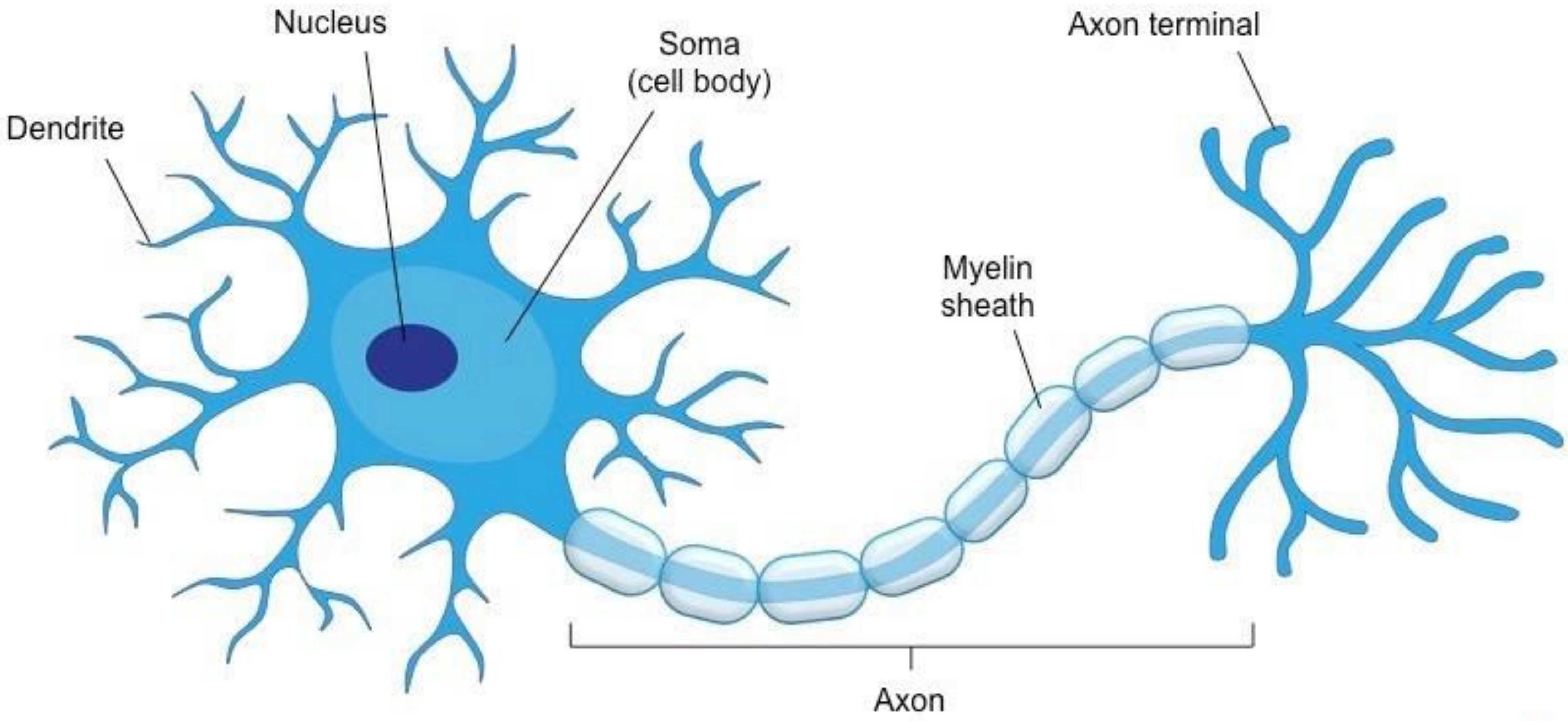
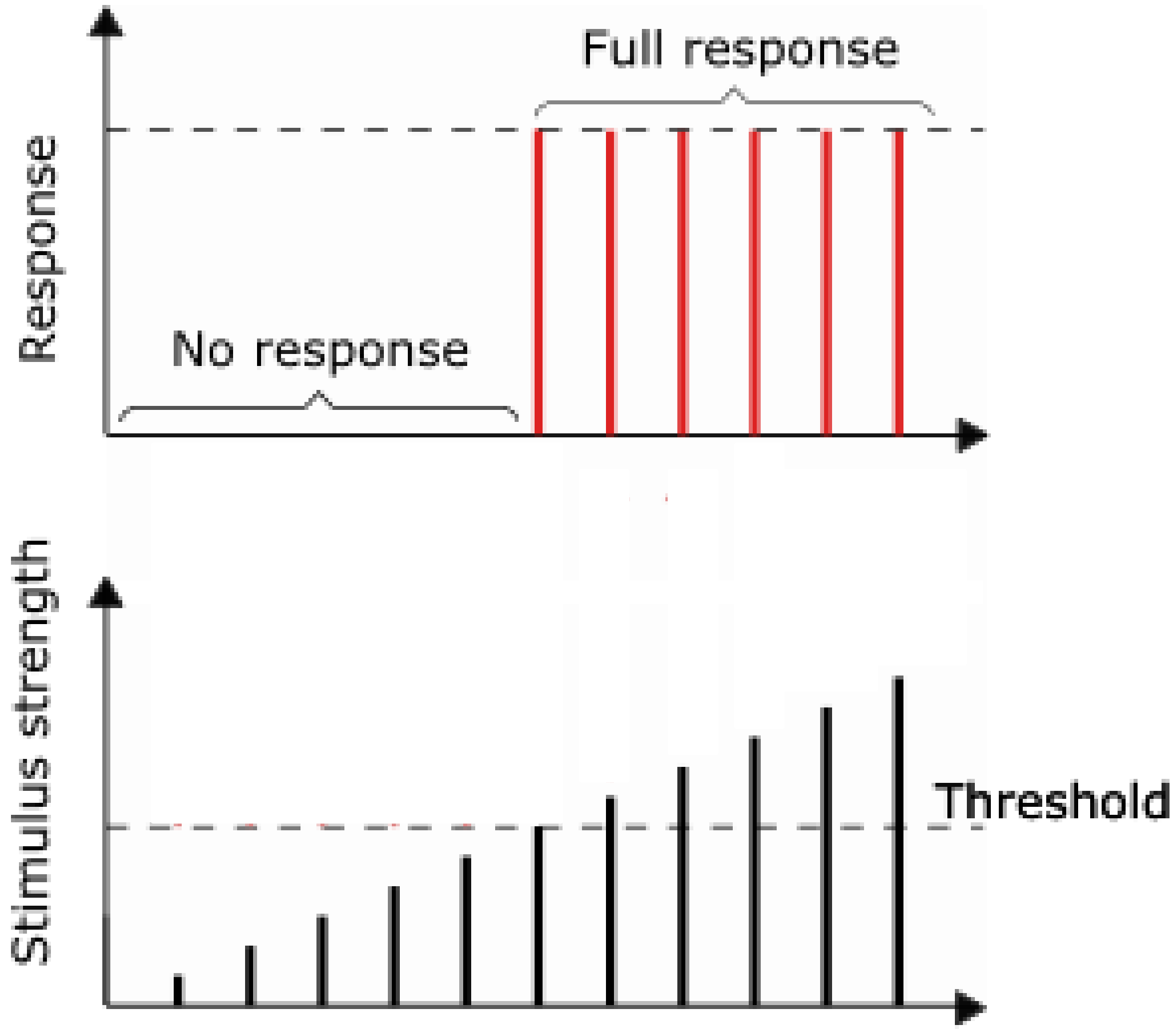


PERCEPTRON

THE DADDY OF NEURAL NETWORKS

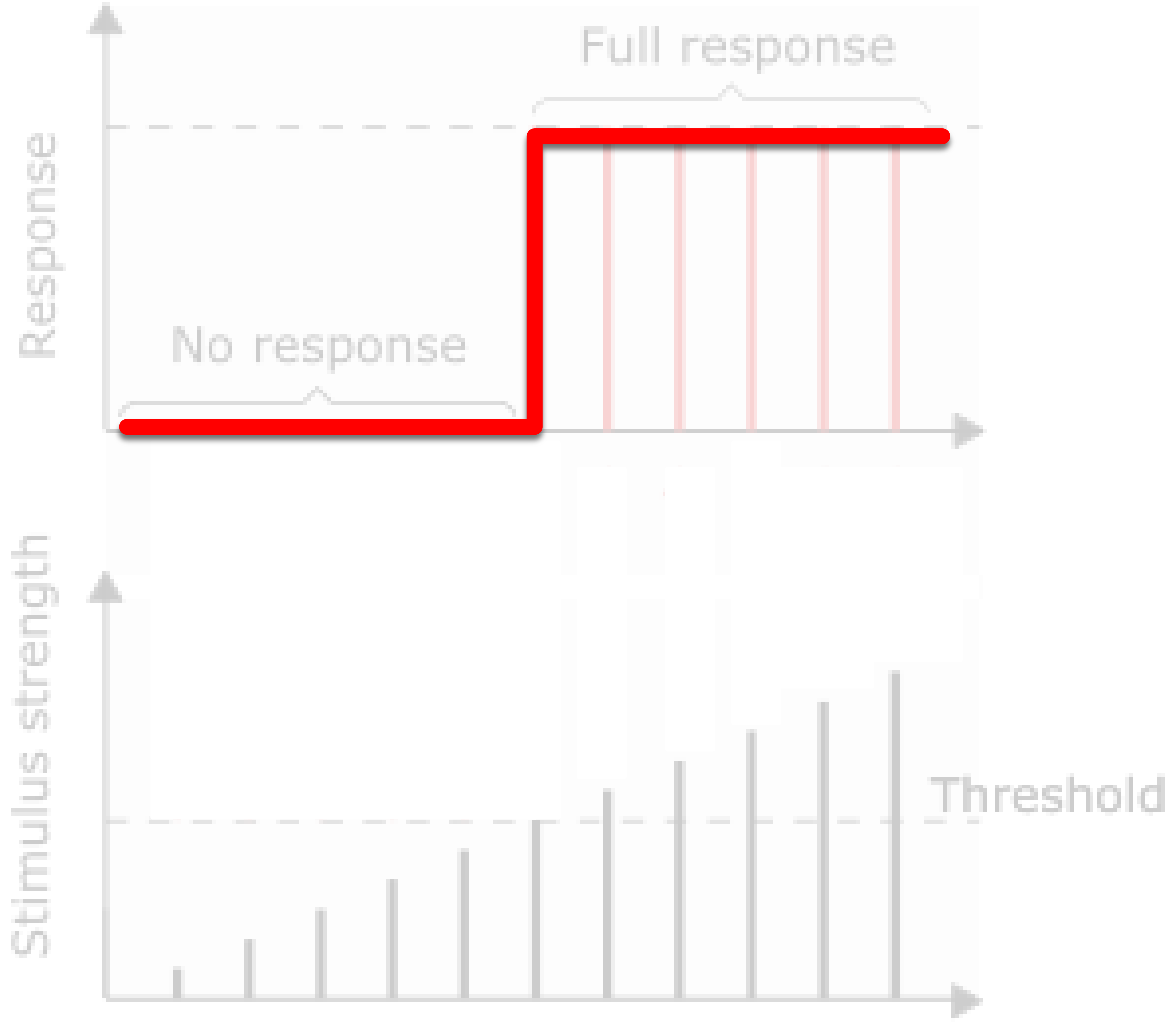
FELIPE BUCHBINDER


Nucleus
Soma
(cell body)
Axon terminal
Dendrite
Myelin
sheath
Axon



ALL-OR-NOTHING LAW OF NEURONAL ACTIVATION

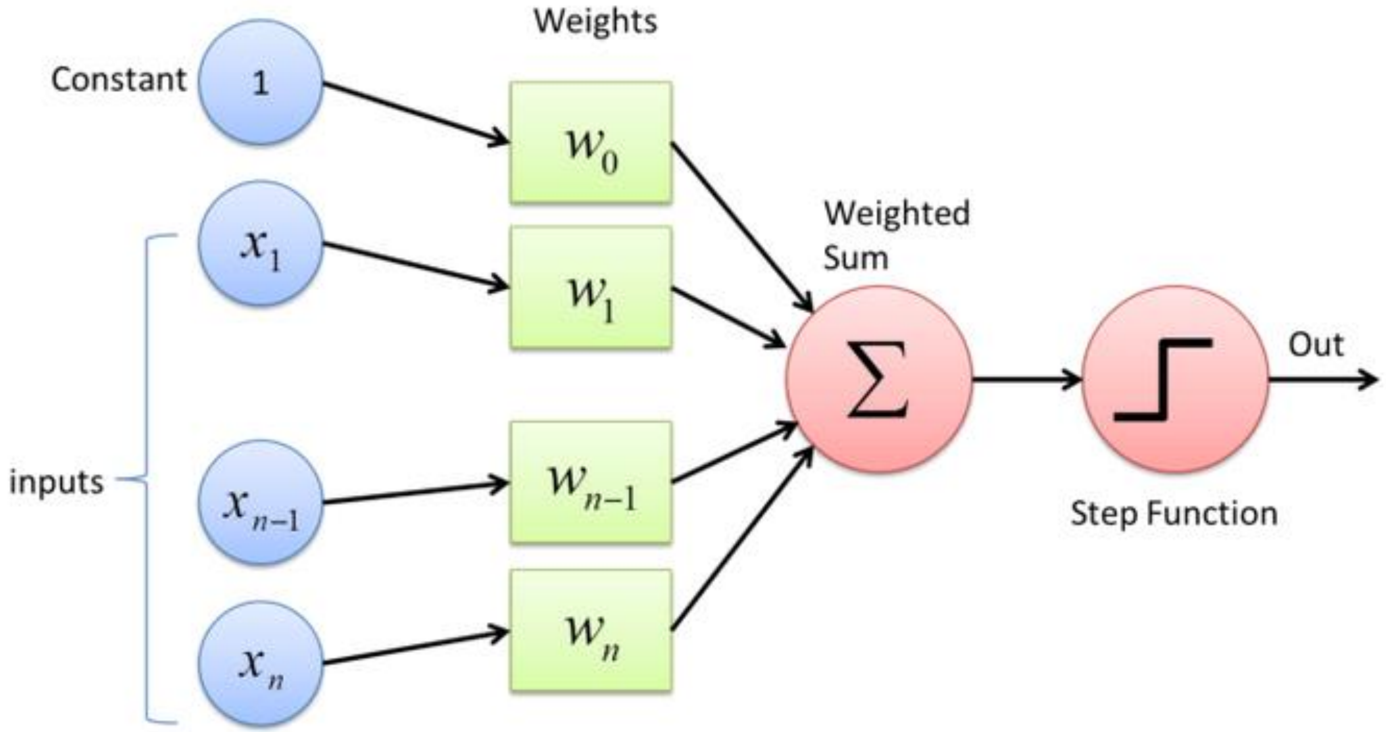
(WHY IS IT, THEN, THAT WE SOMETIMES FEEL MORE PAIN OR LESS PAIN?)



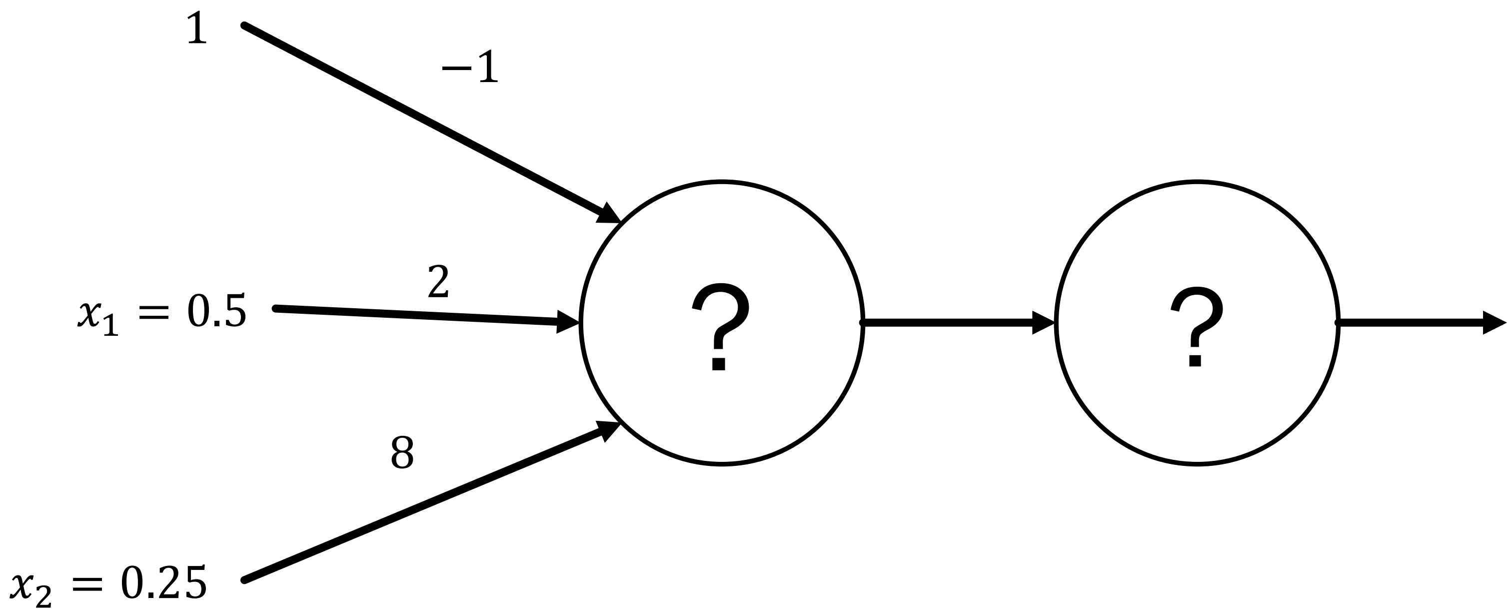
ALL-OR-NOTHING LAW OF NEURONAL ACTIVATION

(WHY IS IT, THEN, THAT WE SOMETIMES FEEL MORE PAIN OR LESS PAIN?)

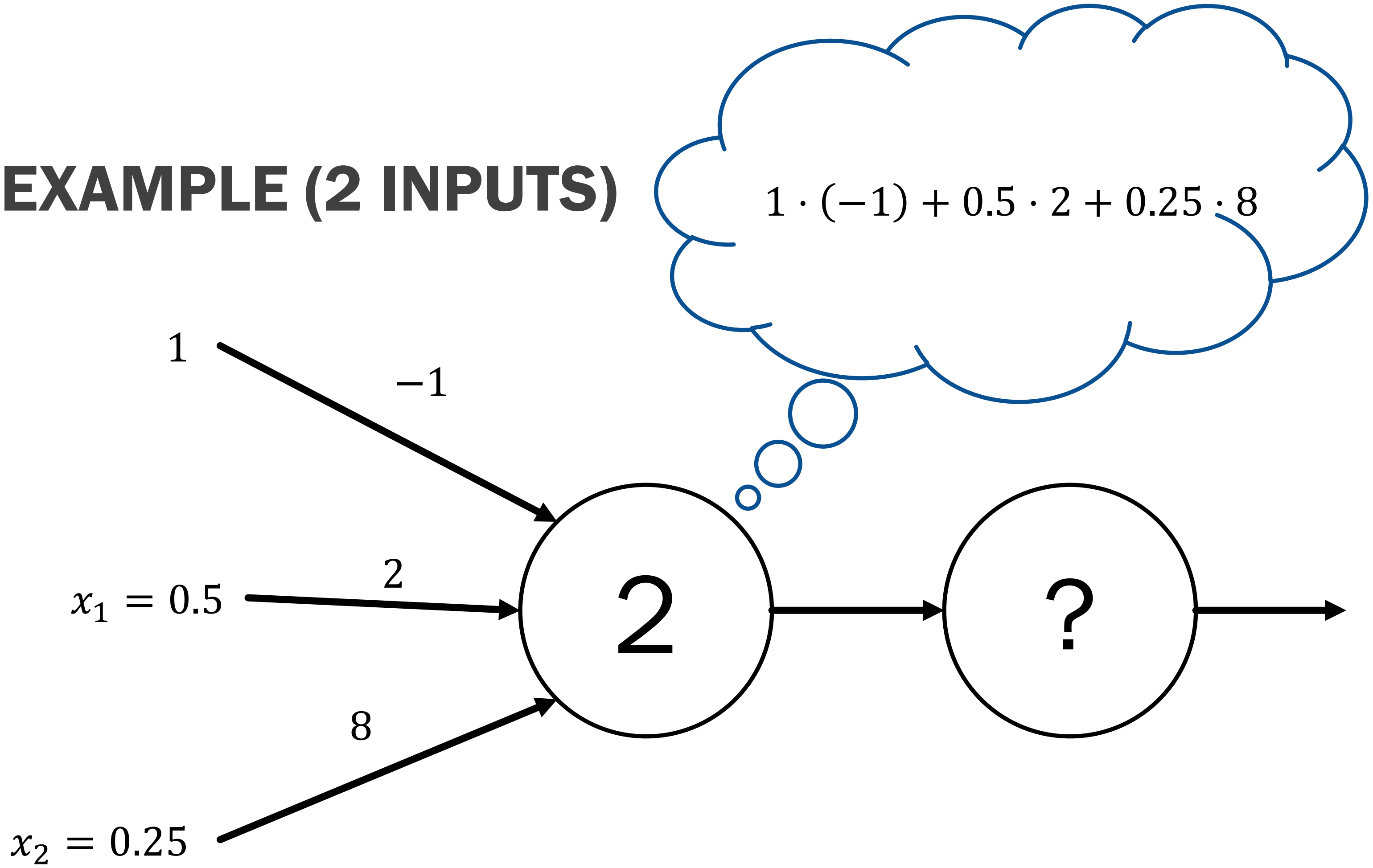
THE PERCEPTRON



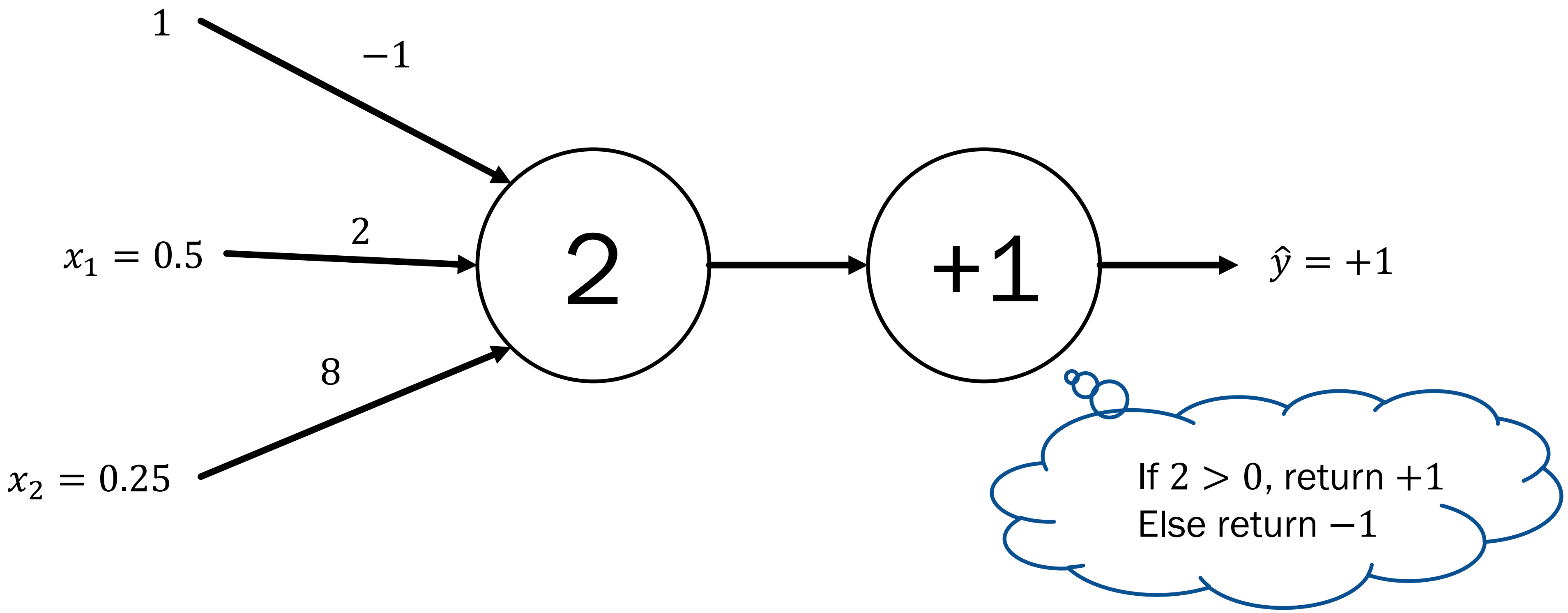
SIMPLE EXAMPLE (2 INPUTS)



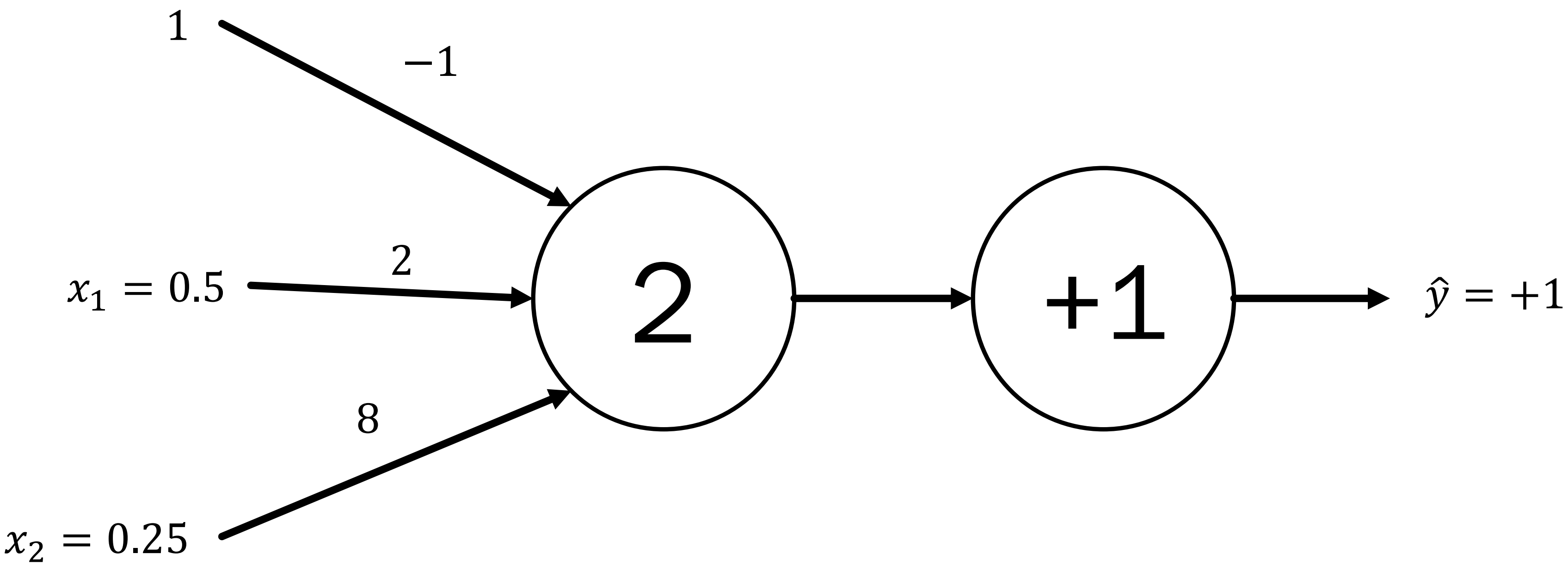
SIMPLE EXAMPLE (2 INPUTS)



SIMPLE EXAMPLE (2 INPUTS)



SIMPLE EXAMPLE (2 INPUTS)

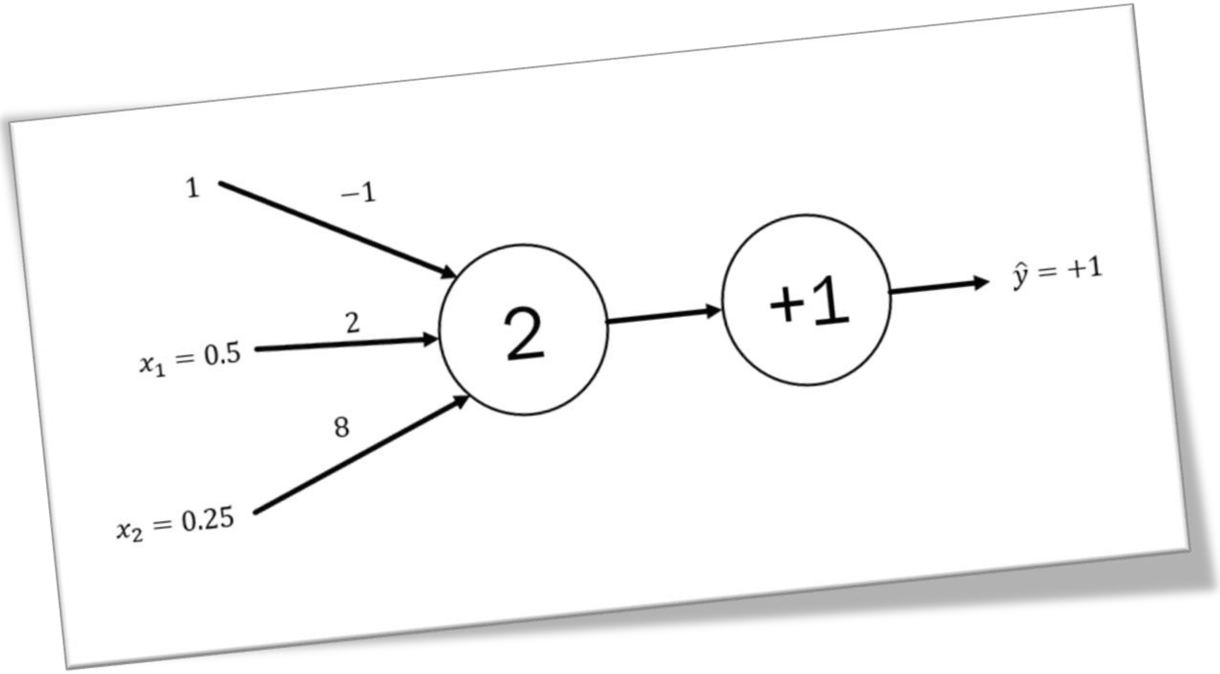


OUR PERCEPTRON'S PREDICTION CAN BE WRITTEN IN A SINGLE LINE

$$\hat{y} = \mathrm{sign}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)$$

OUR PERCEPTRON'S PREDICTION CAN BE WRITTEN IN A SINGLE LINE

$$\hat{y} = \text{sign}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)$$



$$\begin{aligned}\hat{y} &= \text{sign}\left((-1)\cdot 1 + 2\cdot 0.5 + 8\cdot 0.25\right)\\ &= \text{sign}(2)\\ &= +1\end{aligned}$$

PERCEPTRON'S CAN HAVE DIFFERENT ACTIVATION FUNCTIONS

$$\hat{y} = \mathrm{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

Heaviside (step) function

$$\hat{y} = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

Sigmoid function

$$\hat{y} = \tanh(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

Hyperbolic tangent function

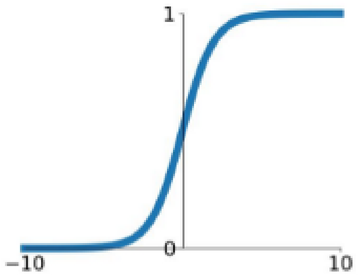
$$\hat{y} = \mathrm{ReLU}(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

Rectified Linear Unit

Activation Functions

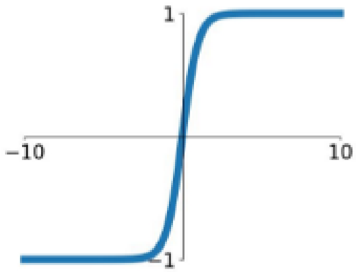
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



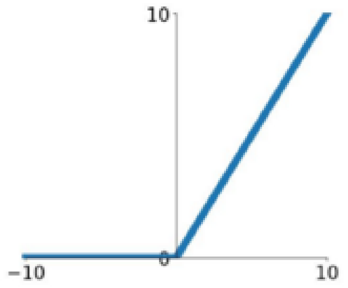
tanh

$\tanh(x)$



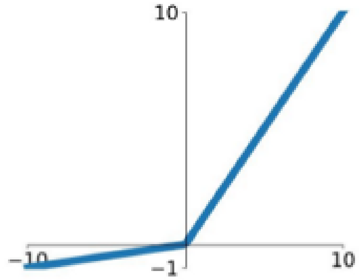
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

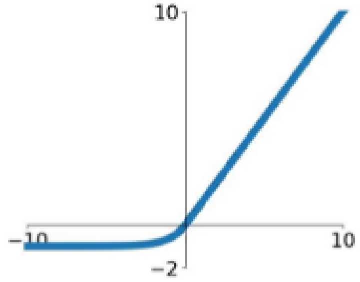


Maxout

$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



HOW CAN THE PERCEPTRON LEARN WHICH WEIGHTS TO USE?

CHOOSE WEIGHTS TO MINIMIZE SOME LOSS FUNCTION

PERCEPTRON'S (ORIGINAL) LEARNING RULE

$$\mathbf{w_n} = \mathbf{w_{n-1}} + \eta y \mathbf{x}$$

! Use only when prediction is wrong

PERCEPTRON'S (ORIGINAL) LEARNING RULE

$$\mathbf{w_n} = \mathbf{w_{n-1}} + \eta y \mathbf{x}$$

! Use only when prediction is wrong

We can get some valuable insights about this rule
if we write it a little bit differently…

PERCEPTRON'S (ORIGINAL) LEARNING RULE

$$\mathbf{w_n} = \mathbf{w_{n-1}} + \eta(\mathrm{y} - \hat{y})\mathbf{x}$$

What happens to the weights ($\mathbf{w}$) if the perceptron overestimates/underestimates the true value of y?

IT GETS EASIER TO INTERPRET IF YOU WRITE IT LIKE THIS:

$$\mathbf{w_n} = \mathbf{w_{n-1}} + \eta(\mathrm{y} - \hat{y})\mathbf{x}$$

What happens to the weights ($\mathbf{w}$) if the perceptron overestimates/underestimates the true value of y?

Perceptron Convergence Theorem:
If the data is linearly separable, then a perceptron is guaranteed to converge in a finite number of steps

PERCEPTRON'S (ORIGINAL) LEARNING RULE

$$\mathbf{w_n} = \mathbf{w_{n-1}} + \eta(y - \hat{y})\mathbf{x}$$

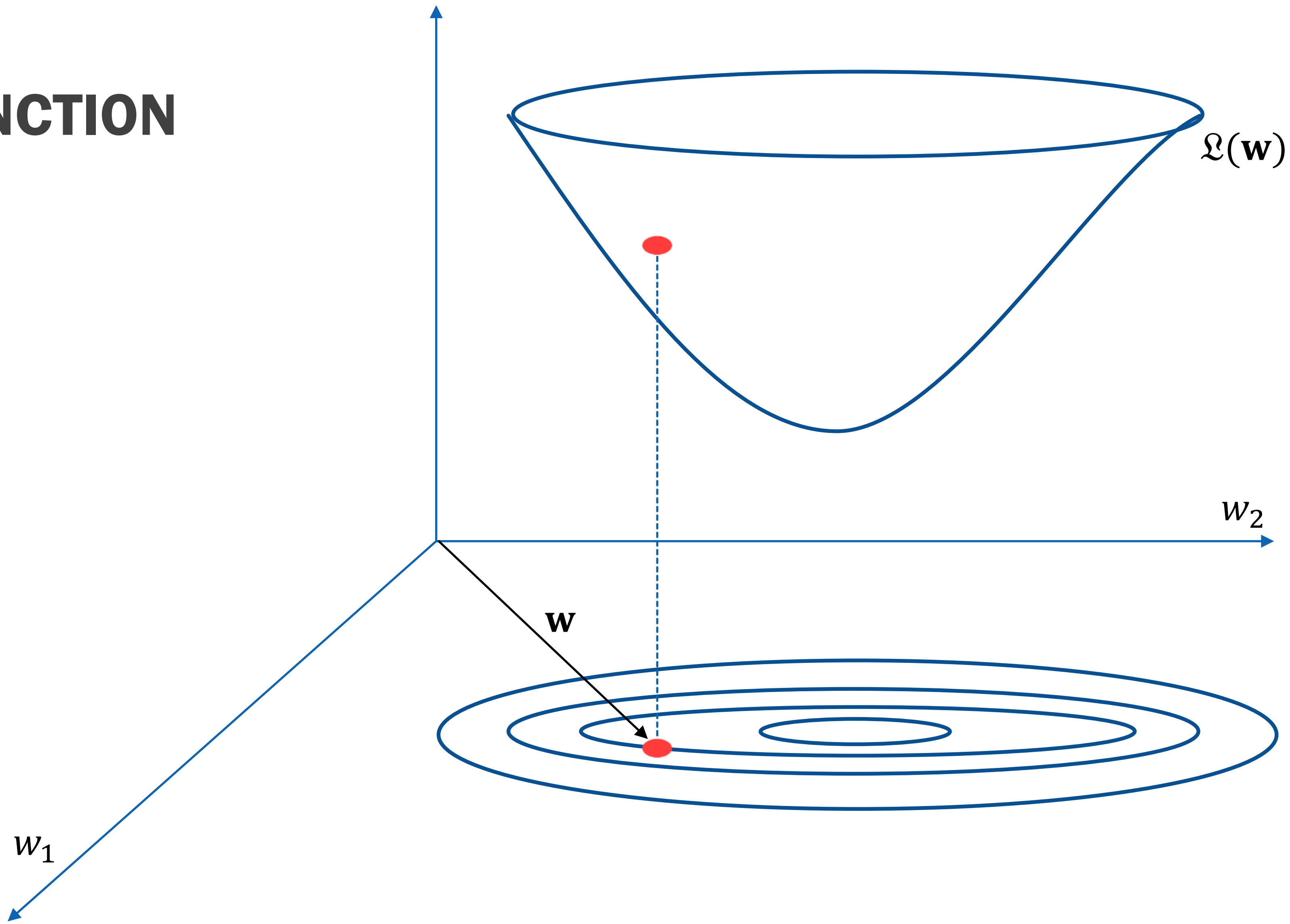
There might be multiple solutions!

Some solutions might not be good

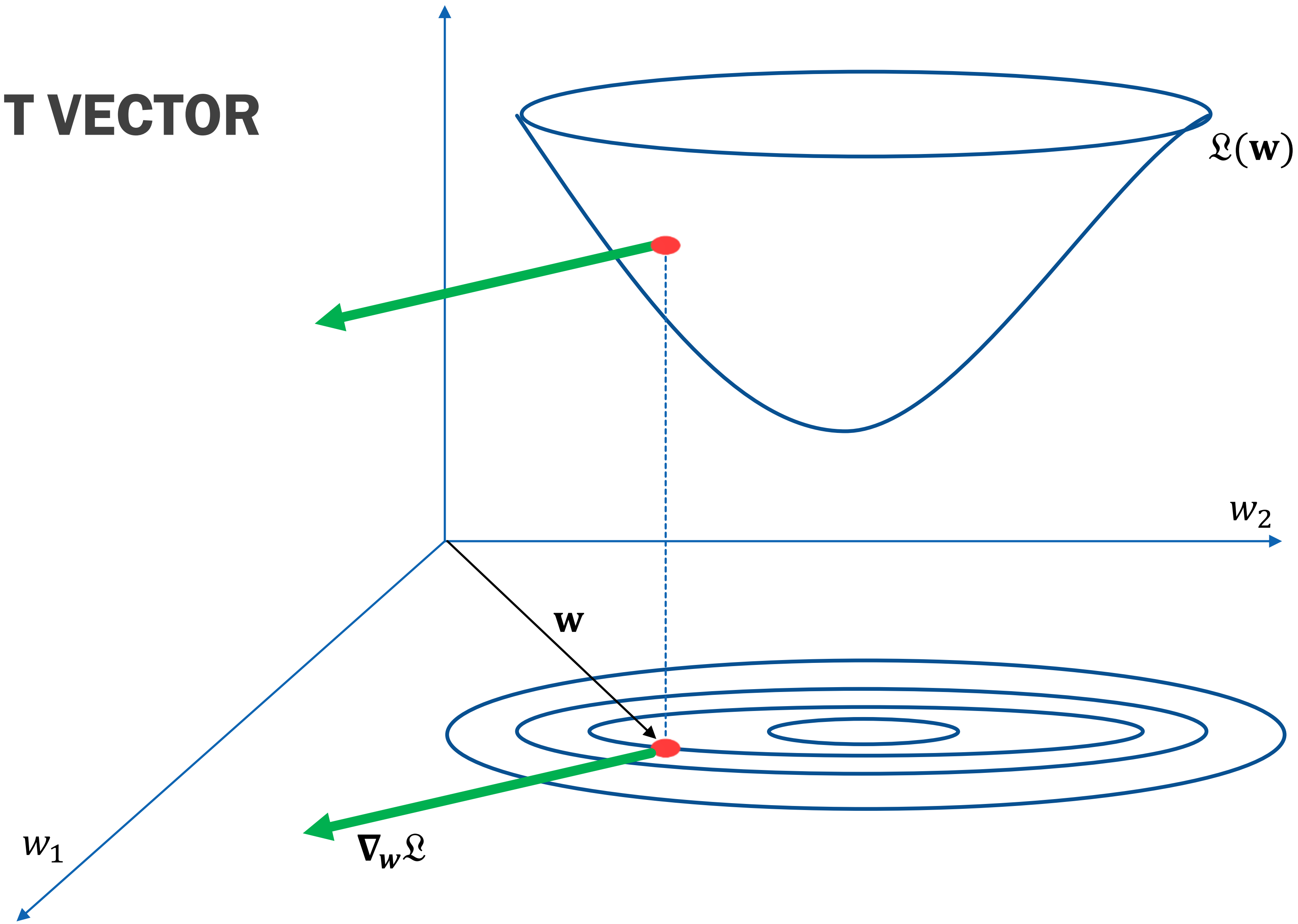
Ever heard of SVM? It's equivalent to a perceptron that gives the optimal solution!

Perceptron Convergence Theorem:
If the data is linearly separable, then a perceptron is guaranteed to converge in a finite number of steps

LOSS FUNCTION



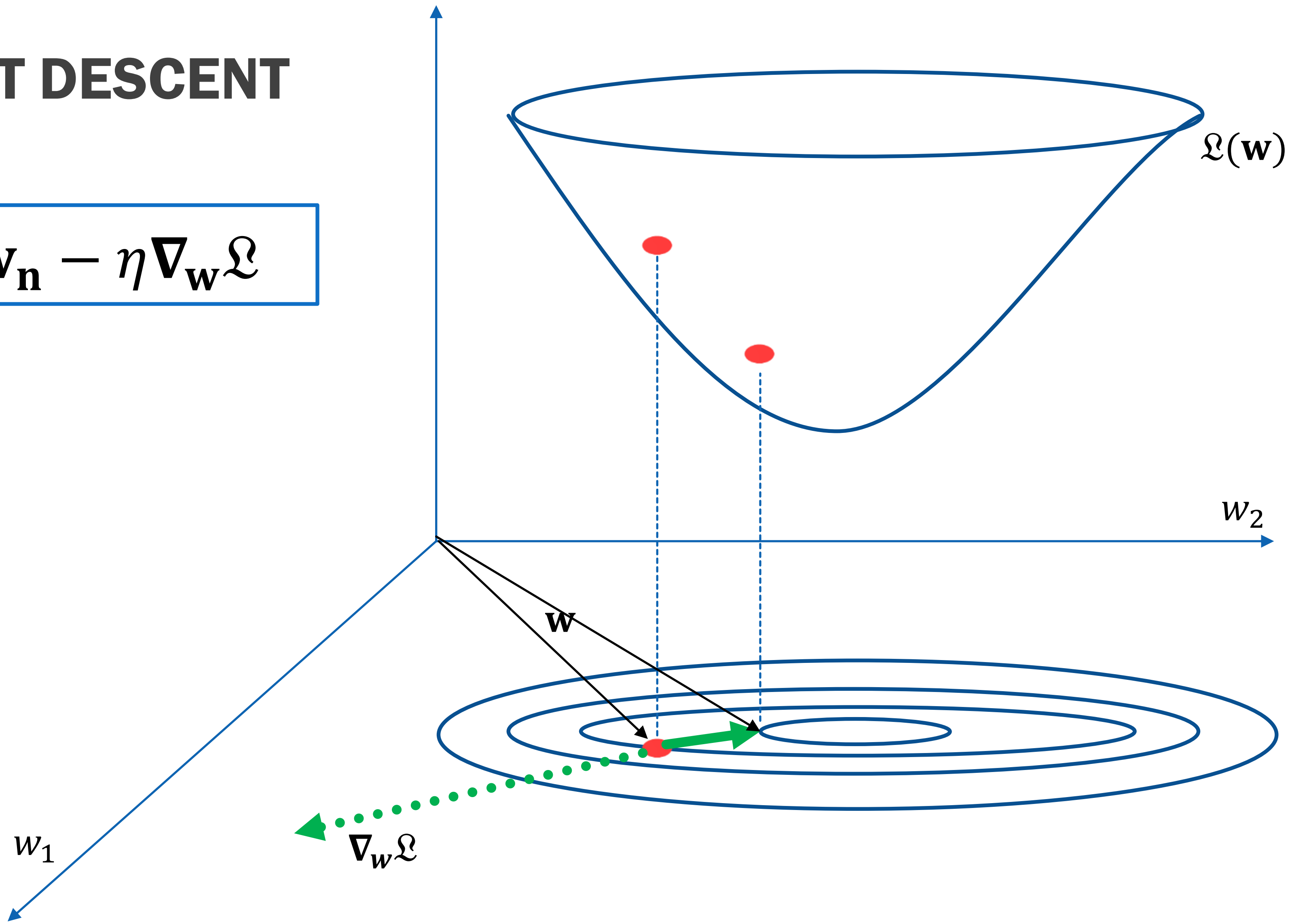
GRADIENT VECTOR



GRADIENT DESCENT

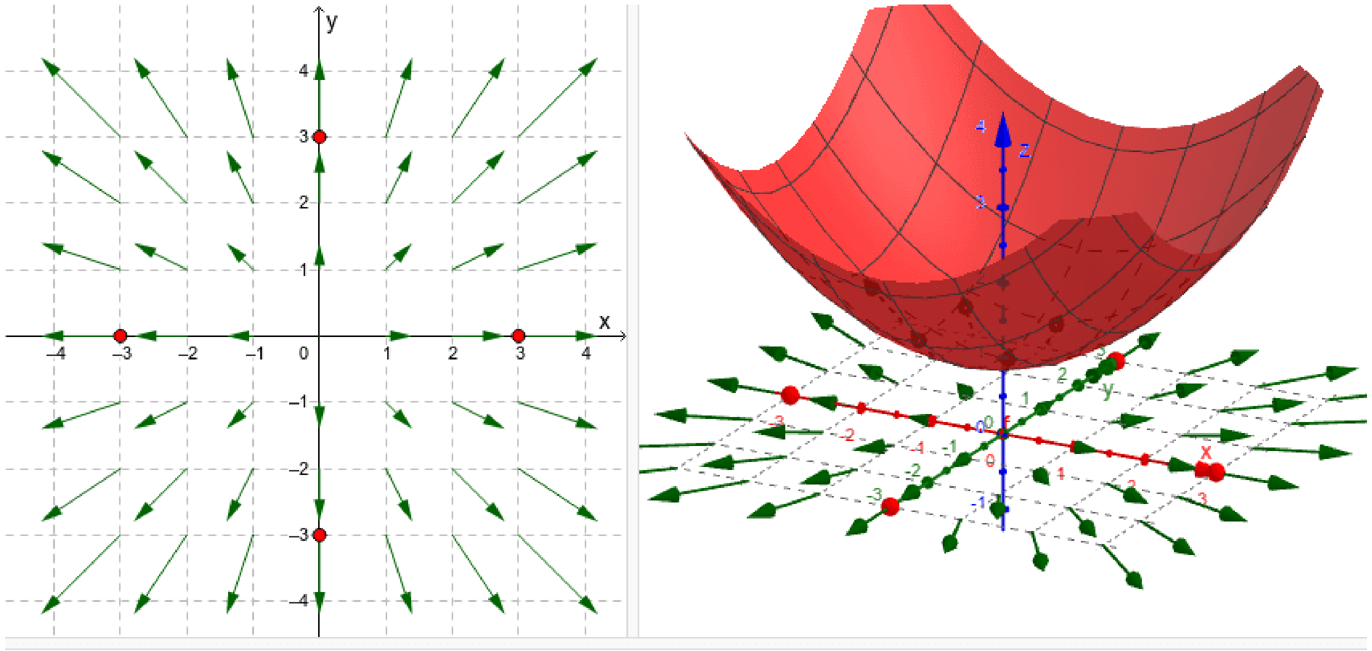
(WITH 1 POINT)

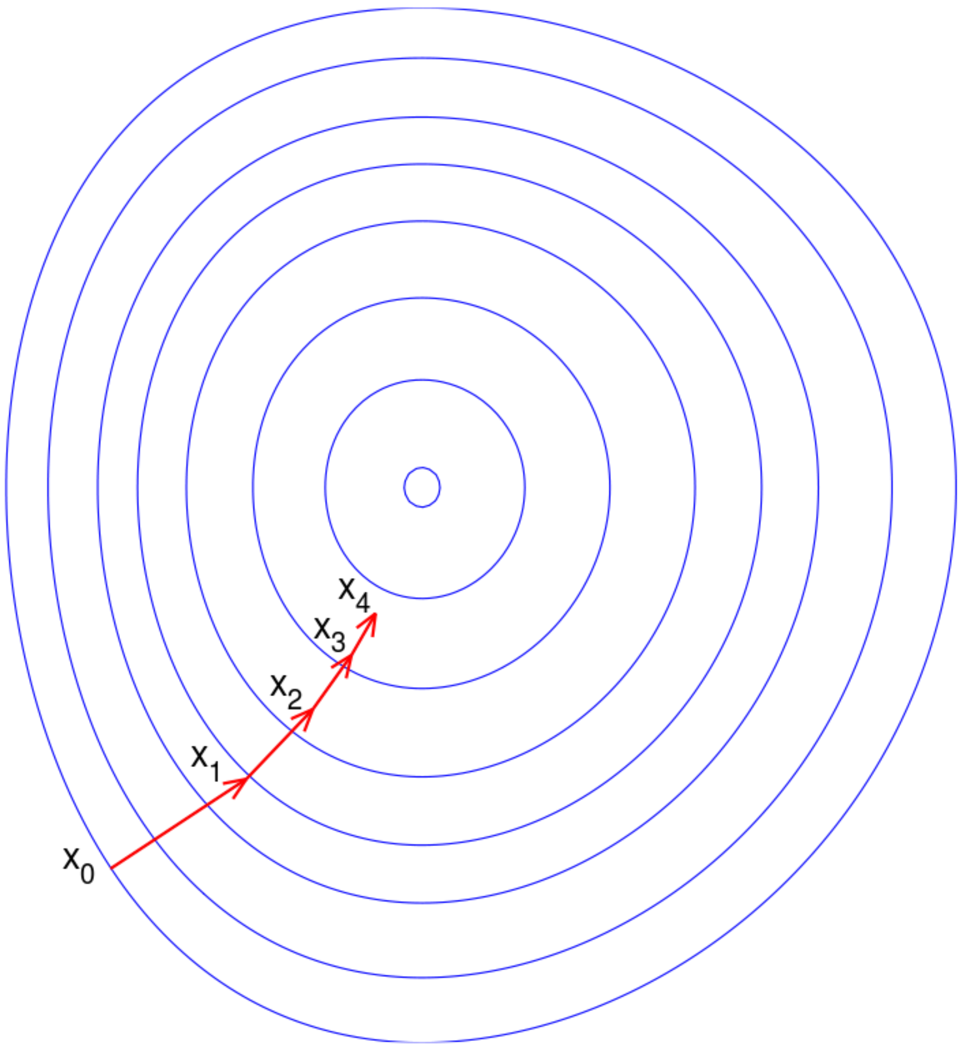
$$\mathbf{w_{n+1}} \leftarrow \mathbf{w_n} - \eta \nabla_{\mathbf{w}} \mathfrak{L}$$



GRADIENT DESCENT

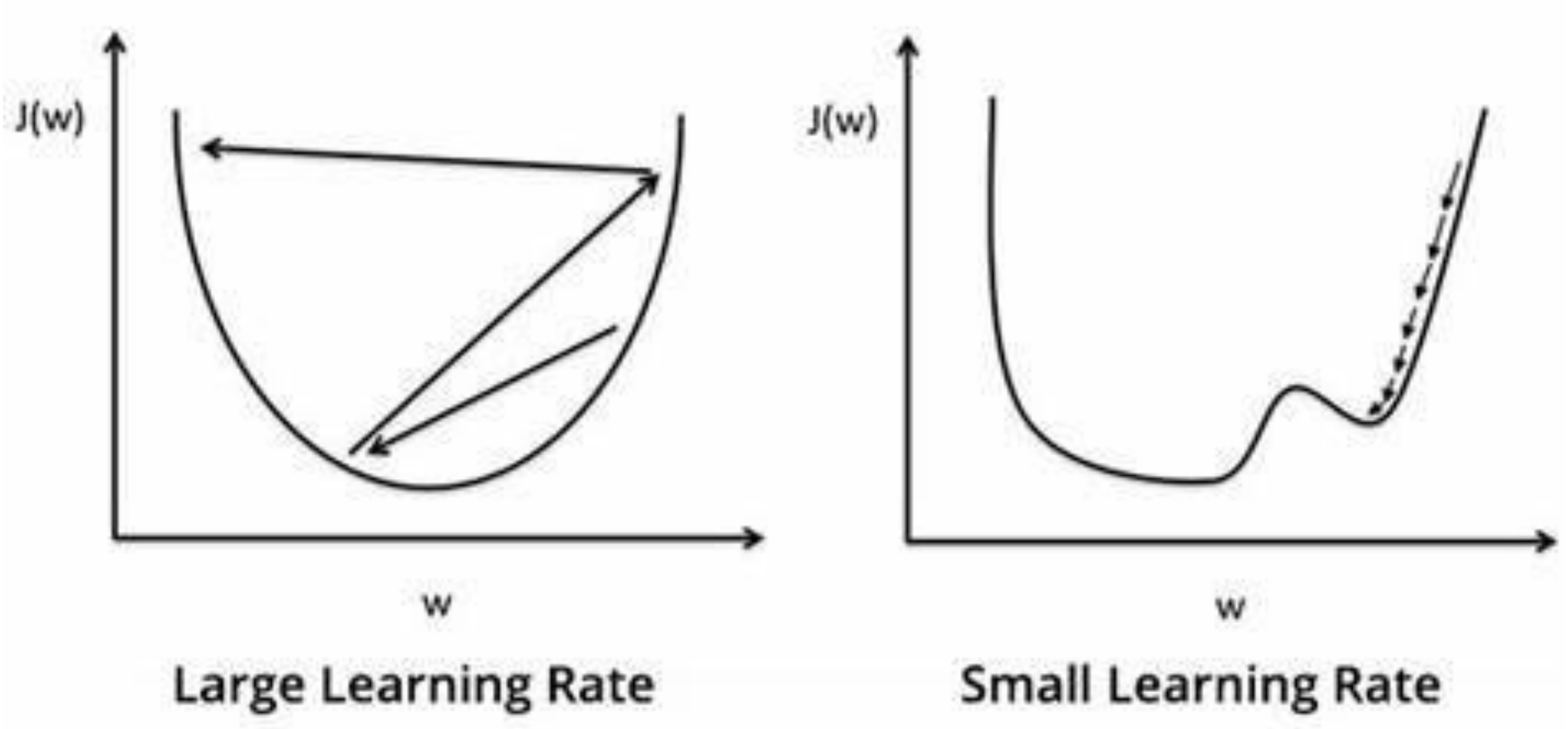
$$\mathbf{w_{n+1}} \leftarrow \mathbf{w_n} - \eta \sum_i \nabla_{\mathbf{w}} \mathfrak{L}$$





x0
x1
x2
x3
x4

EFFECT OF THE LEARNING RATE (η)



EXAMPLE: STEP ACTIVATION FUNCTION WITH HINGE LOSS

$$\mathfrak{L} = \max(0; 1 - y\hat{y})$$

$$\hat{y} = \operatorname{sign}\left(\mathbf{w}^{\mathbf{T}}\mathbf{x}\right)$$

$$\mathbf{w_{n+1}} = \mathbf{w_n} - \eta \boldsymbol{\nabla}_{\mathbf{w}} \boldsymbol{\mathfrak{L}}$$

$$\nabla_{\mathbf{w}} \mathfrak{L} = -y\mathbf{x}$$

$\therefore$

$$\mathbf{w_{n+1}} \leftarrow \mathbf{w_n} + \eta y \mathbf{x}$$

EXAMPLE: LOGISTIC ACTIVATION FUNCTION WITH HINGE LOSS

$$\mathfrak{L} = \max(0; 1 - y\hat{y})$$
$$\hat{y} = \sigma\left(\mathbf{w}^{\mathbf{T}}\mathbf{x}\right)$$
$$\mathbf{w}_{\mathbf{n+1}} = \mathbf{w}_{\mathbf{n}} - \eta \boldsymbol{\nabla}_{\mathbf{w}} \boldsymbol{\mathfrak{L}}$$

$$\nabla_{\mathbf{w}}\mathfrak{L} = -y\sigma\left(\mathbf{w}^{\mathbf{T}}\mathbf{x}\right)\left[\mathbf{1} - \sigma\left(\mathbf{w}^{\mathbf{T}}\mathbf{x}\right)\right]\mathbf{x}$$

$\therefore$

$$\mathbf{w}_{\mathbf{n+1}} \leftarrow \mathbf{w}_{\mathbf{n}} + \eta y\sigma\left(\mathbf{w}^{\mathbf{T}}\mathbf{x}\right)\left[\mathbf{1} - \sigma\left(\mathbf{w}^{\mathbf{T}}\mathbf{x}\right)\right]\mathbf{x}$$

EXAMPLE: LOGISTIC ACTIVATION FUNCTION WITH HINGE LOSS

$$\mathfrak{L} = \max(0; 1 - y\hat{y})$$

$$\hat{y} = \sigma(\mathbf{w}^{\mathbf{T}}\mathbf{x})$$

$$\mathbf{w_{n+1}} = \mathbf{w_n} - \eta \nabla_{\mathbf{w}} \mathfrak{L}$$

This tells us about the importance of normalizing the inputs. Why?

$$\nabla_{\mathbf{w}} \mathfrak{L} = -y\sigma(\mathbf{w}^{\mathbf{T}}\mathbf{x})[\mathbf{1} - \sigma(\mathbf{w}^{\mathbf{T}}\mathbf{x})]\mathbf{x}$$

$\therefore$

$$\mathbf{w_{n+1}} \leftarrow \mathbf{w_n} + \eta y \sigma(\mathbf{w}^{\mathbf{T}}\mathbf{x})[\mathbf{1} - \sigma(\mathbf{w}^{\mathbf{T}}\mathbf{x})]\mathbf{x}$$

EXAMPLE: LOGISTIC ACTIVATION FUNCTION WITH HINGE LOSS

$$\mathfrak{L} = \max(0; 1 - y\hat{y})$$
$$\hat{y} = \sigma(\mathbf{w}^\mathbf{T}\mathbf{x})$$
$$\mathbf{w_{n+1}} = \mathbf{w_n} - \eta \nabla_\mathbf{w} \mathfrak{L}$$

This tells us about the importance of normalizing the inputs. Why?

$$\nabla_\mathbf{w} \mathfrak{L} = -y\sigma(\mathbf{w}^\mathbf{T}\mathbf{x})[1 - \sigma(\mathbf{w}^\mathbf{T}\mathbf{x})]\mathbf{x}$$

$\therefore$

$$\mathbf{w_{n+1}} \leftarrow \mathbf{w_n} + \eta y \sigma(\mathbf{w}^\mathbf{T}\mathbf{x})[\mathbf{1} - \sigma(\mathbf{w}^\mathbf{T}\mathbf{x})]\mathbf{x}$$

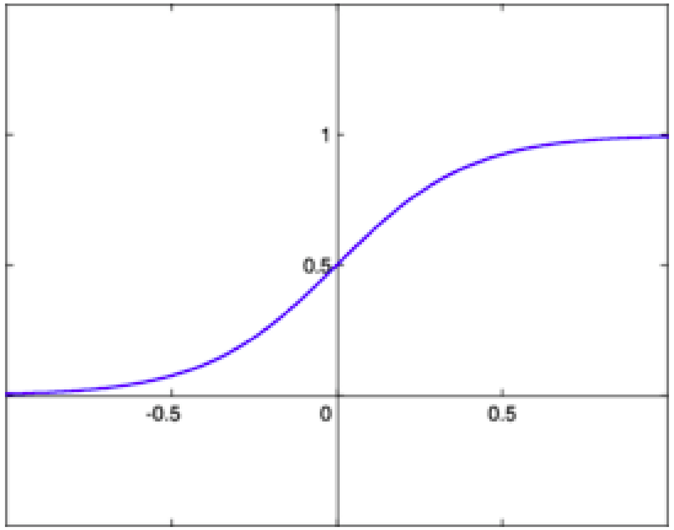
This also offers a nice illustration of the vanishing gradient problem. Why? – and what to do about it?

THE VANISHING GRADIENT PROBLEM

$$\mathbf{w_{n+1}} \leftarrow \mathbf{w_n} - \eta \sum_i \nabla_{\mathbf{w}} \mathfrak{L}$$

When $\nabla_{\mathbf{w}} \mathfrak{L} \rightarrow 0$, learning stops ($\mathbf{w_{n+1}} \approx \mathbf{w_n}$)

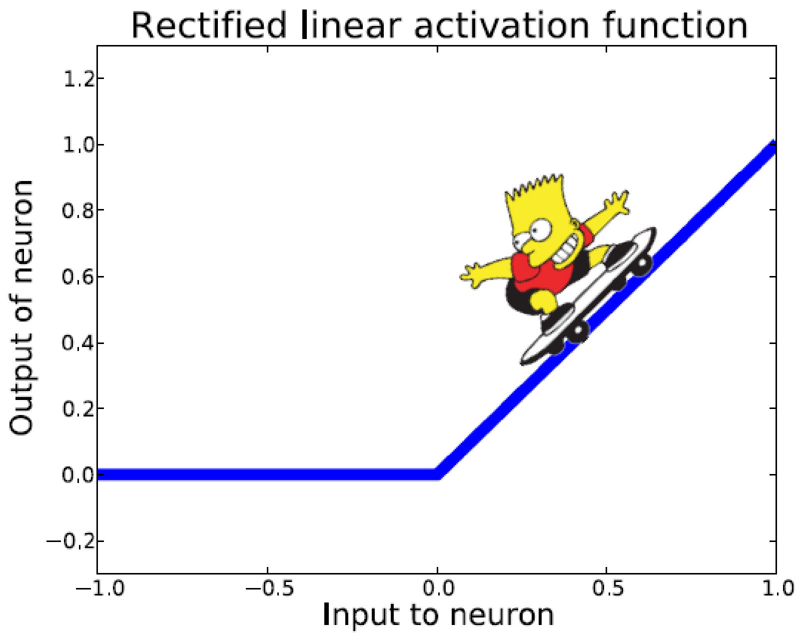
What can we do about it?



SOLUTIONS TO THE VANISHING GRADIENT PROBLEM

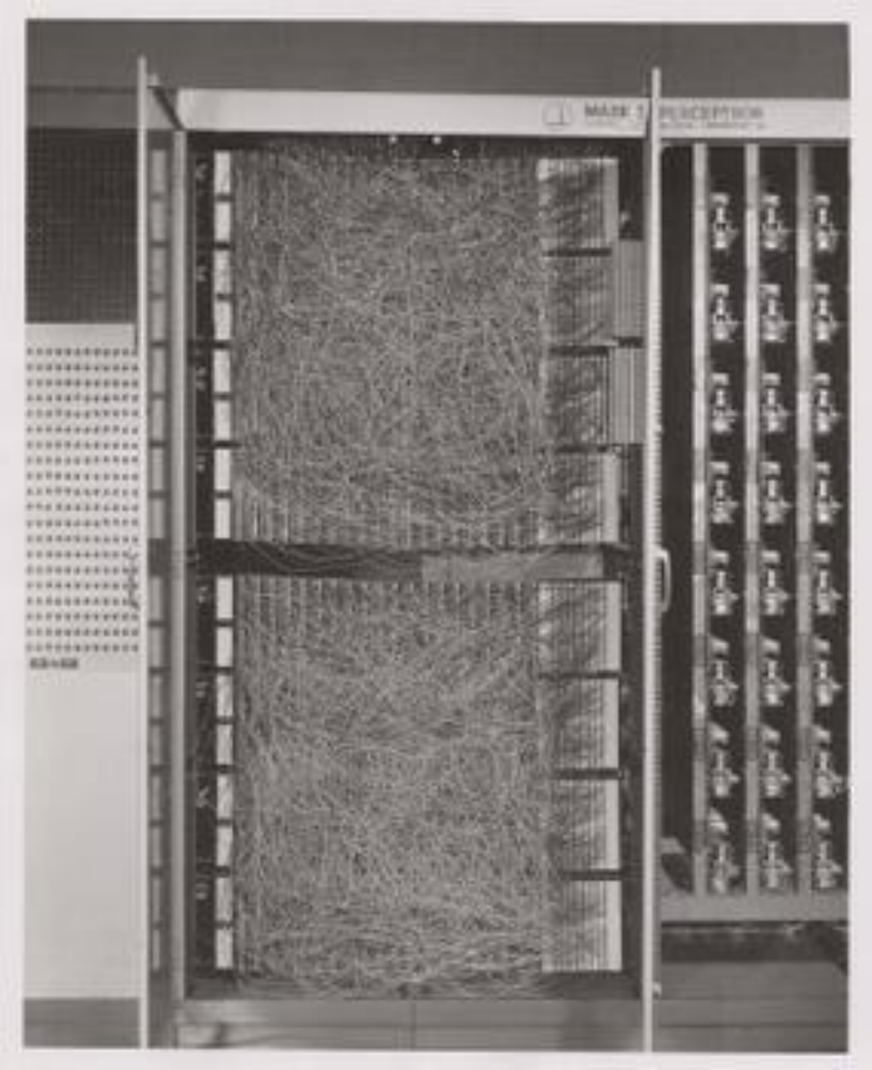
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Normalization



Changing the activation function

and other options we'll talk about later…



THE FIRST PERCEPTRON (IBM, 1958)

COMING UP NEXT:

NEURAL NETWORKS