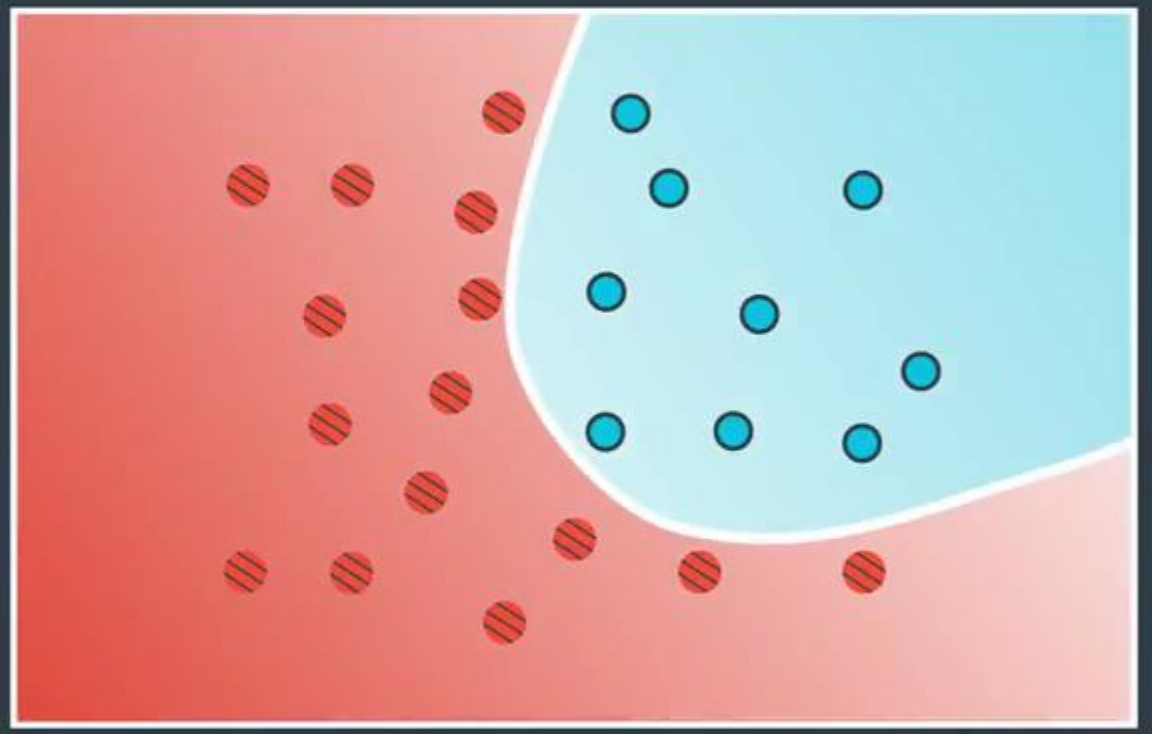
Neural Network

by Hassan Badawy

Acceptance at a University

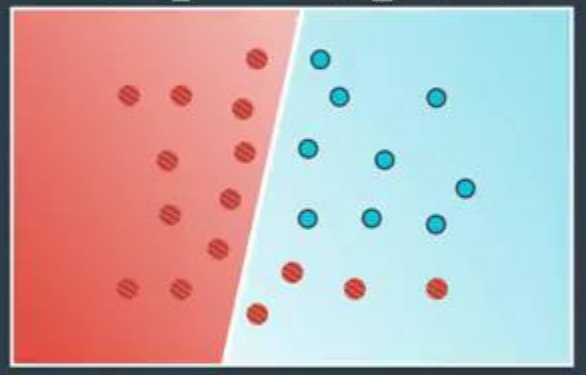
Acceptance at a University

Non-Linear Regions

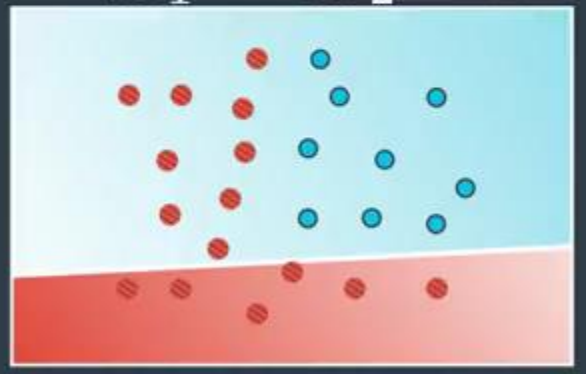


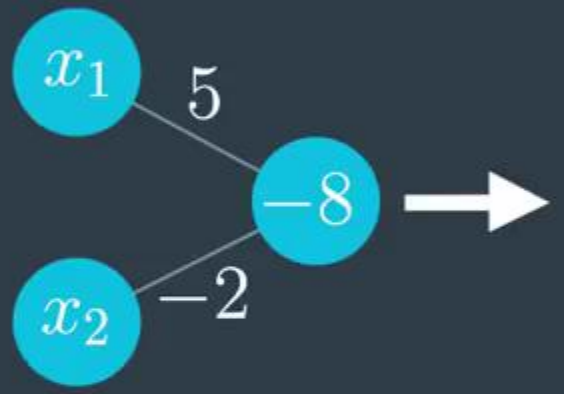
Neural Network

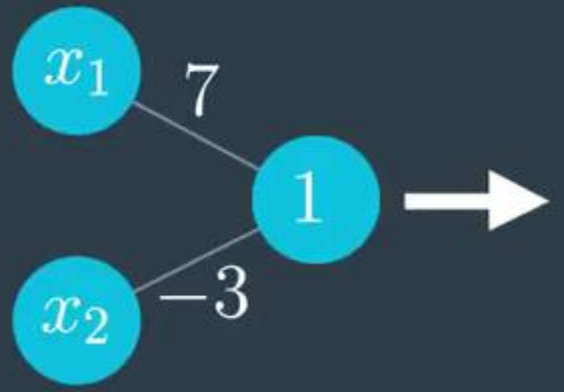
$5x_1 - 2x_2 + 8$



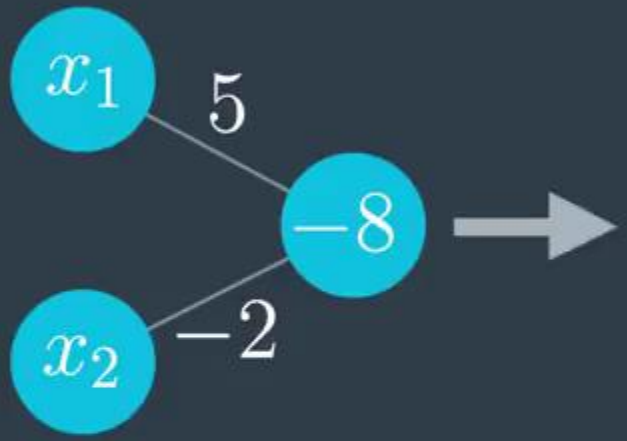
$7x_1 - 3x_2 - 1$

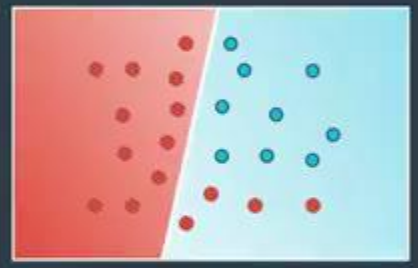


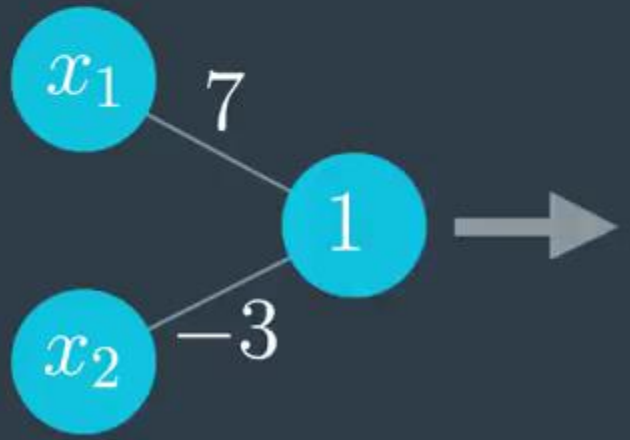


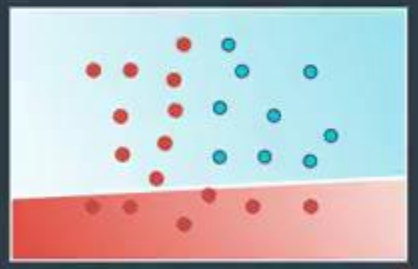


Neural Network

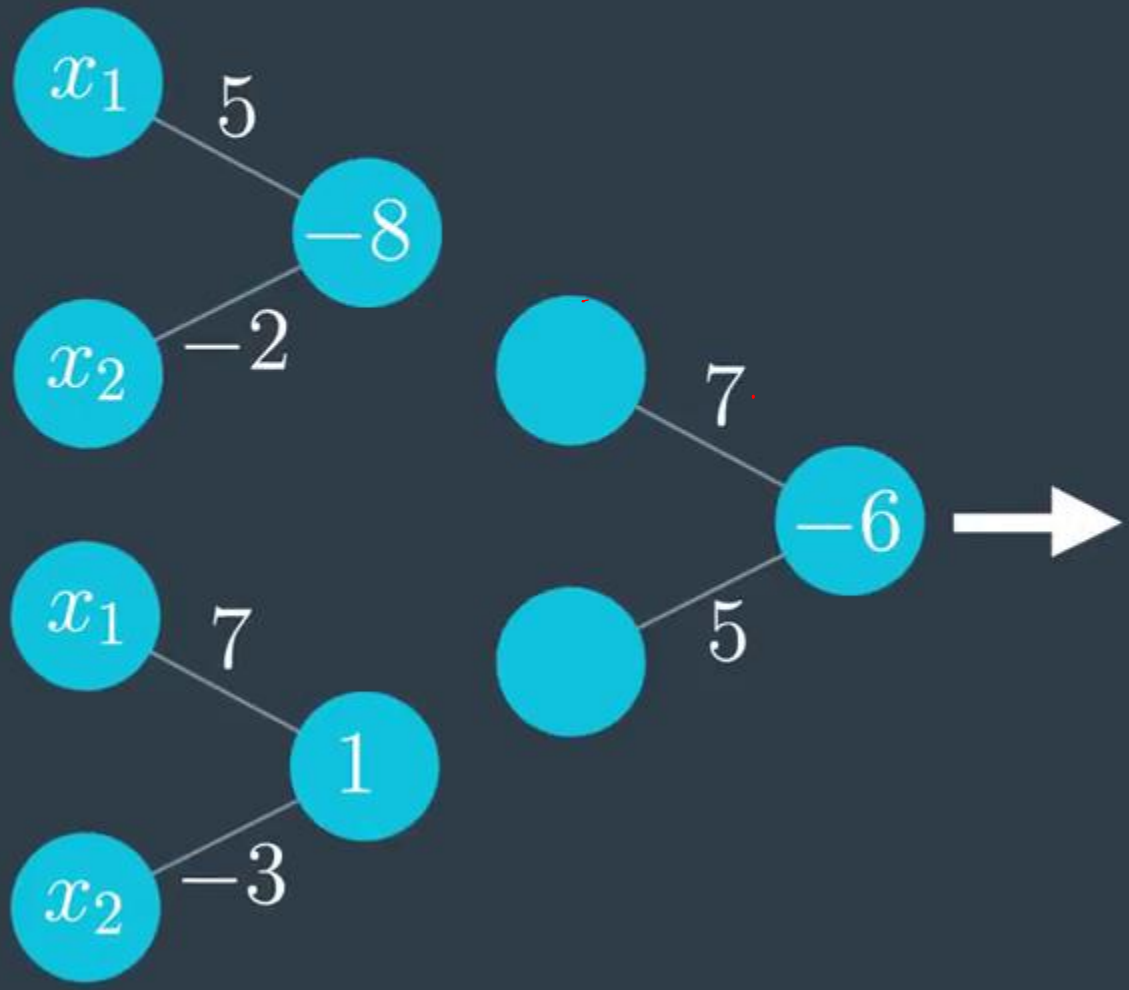


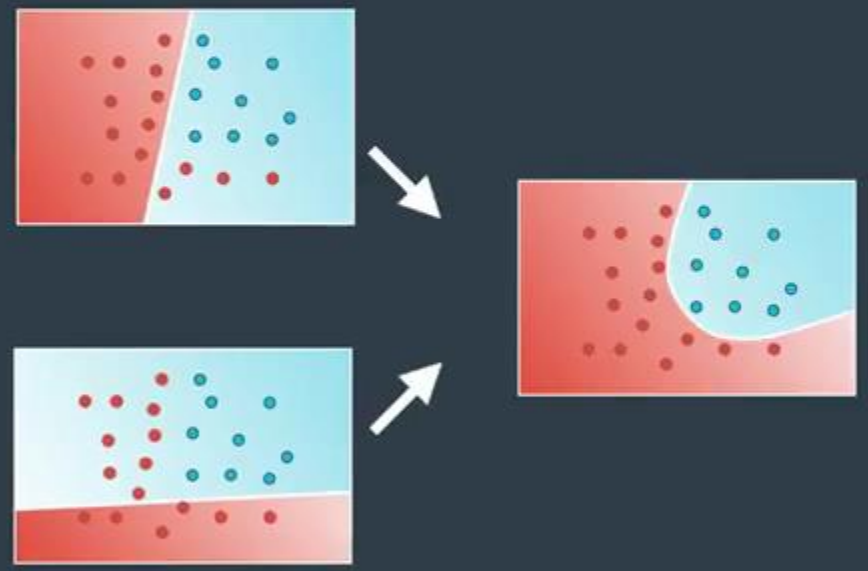




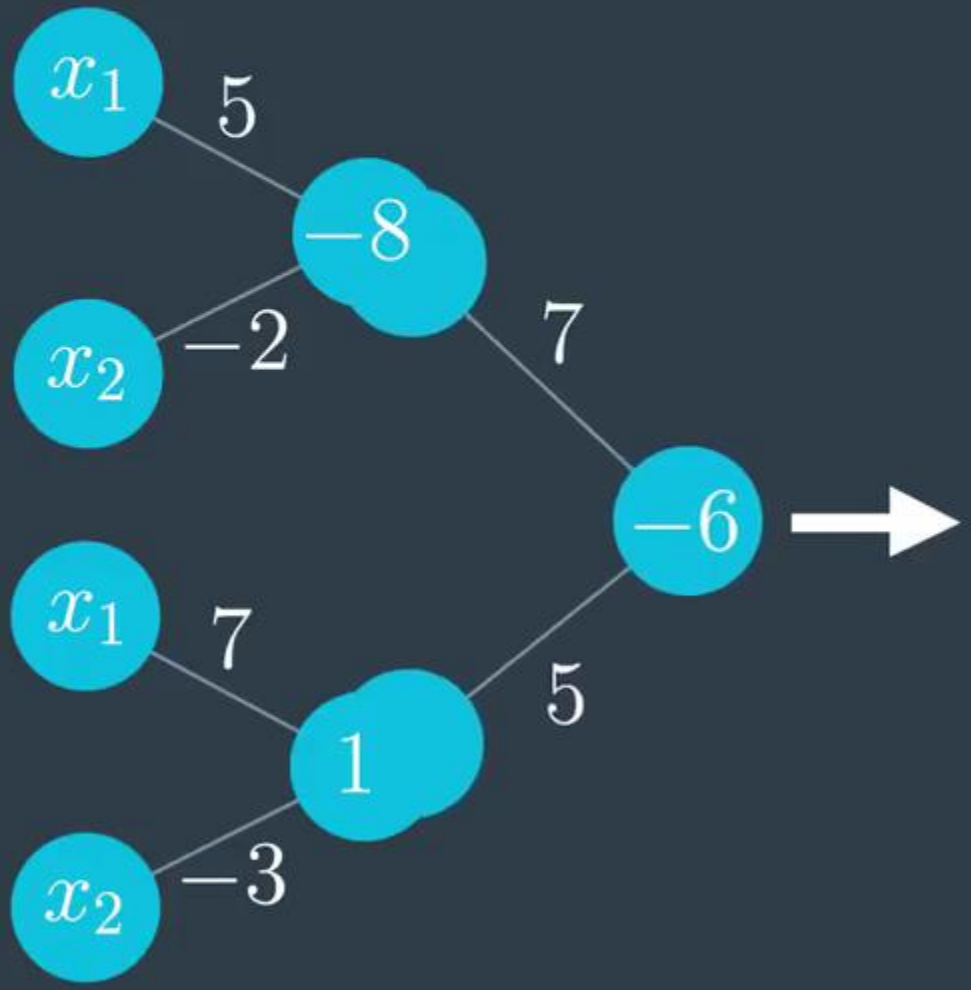


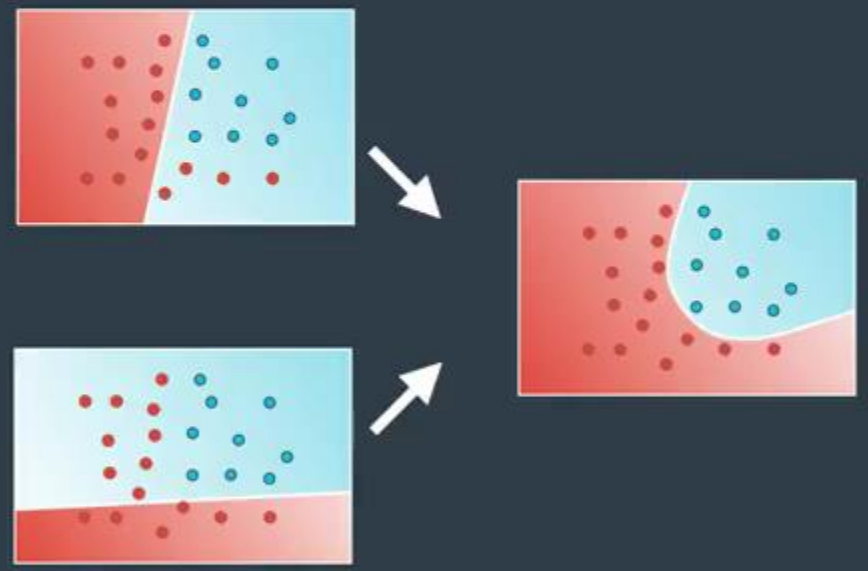
Neural Network



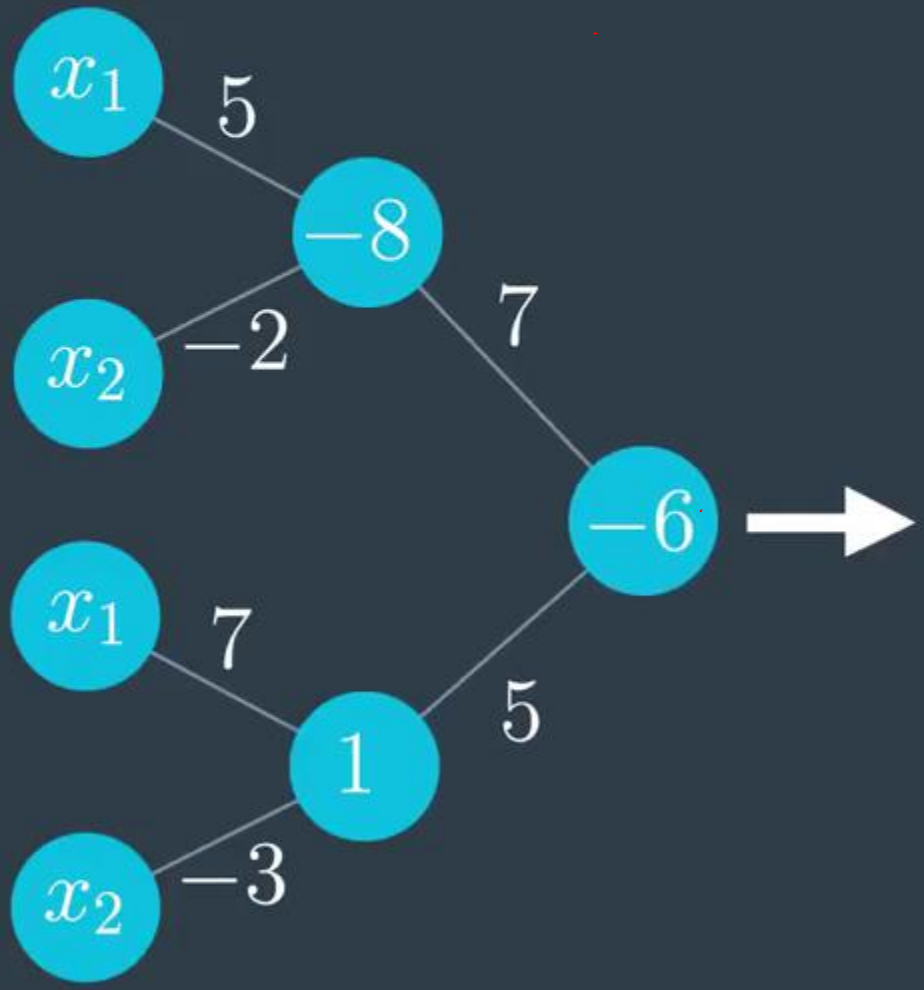


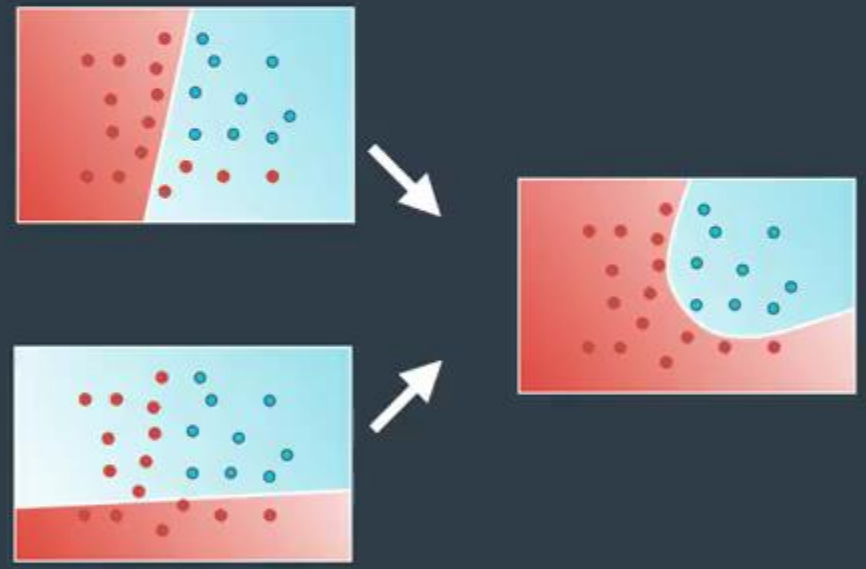
Neural Network



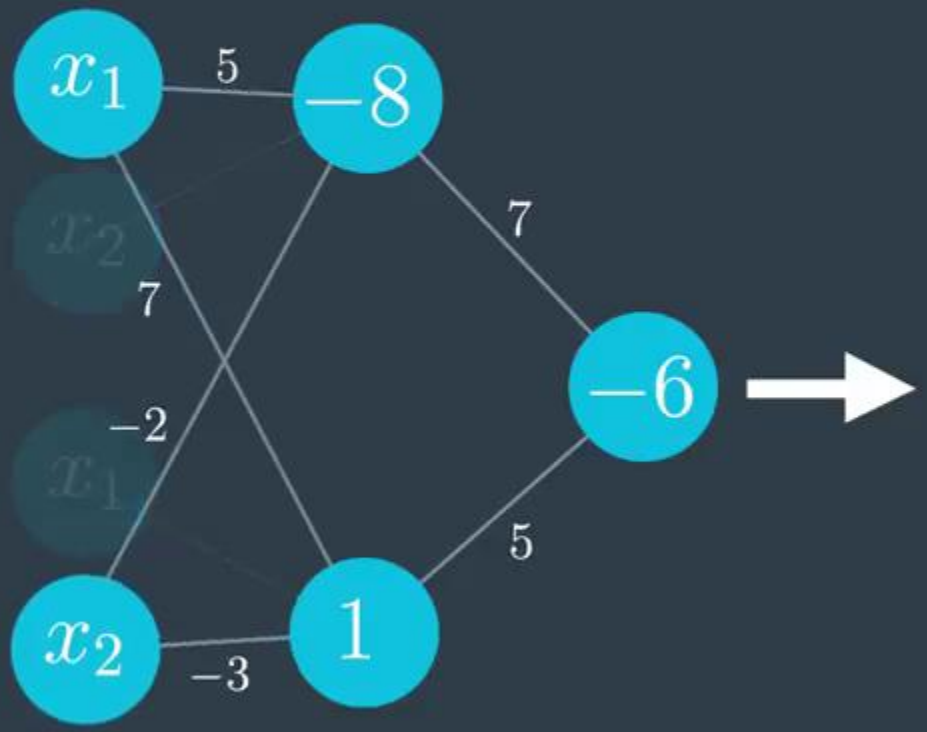


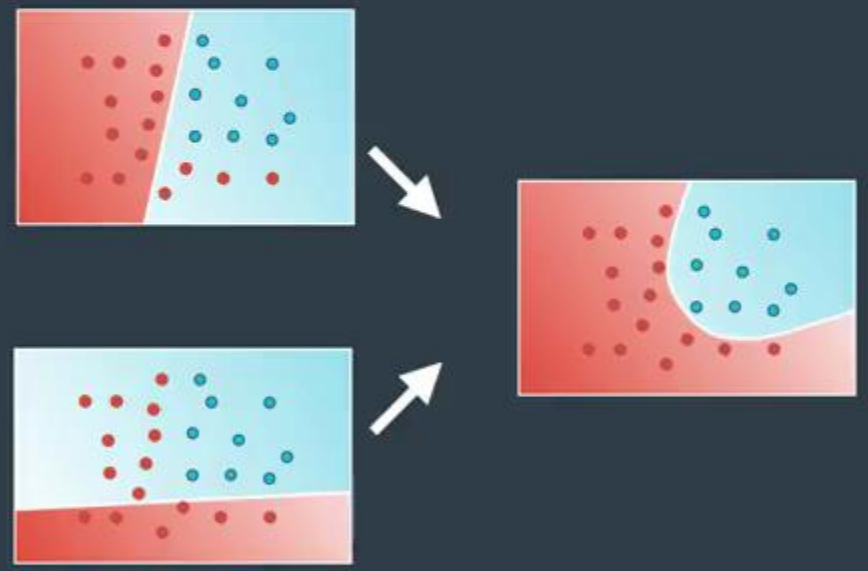
Neural Network



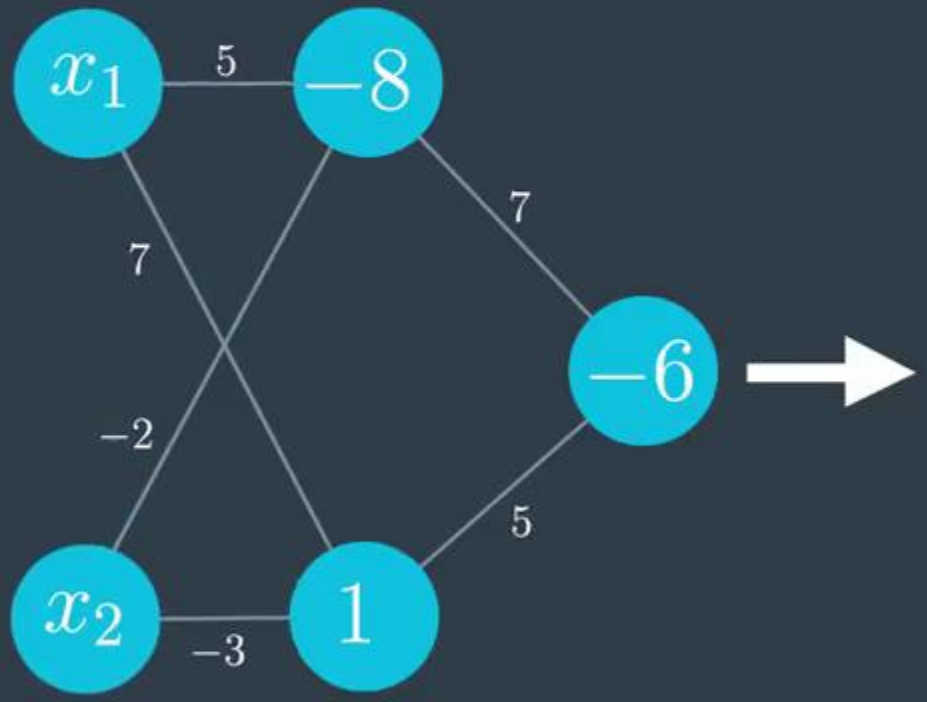


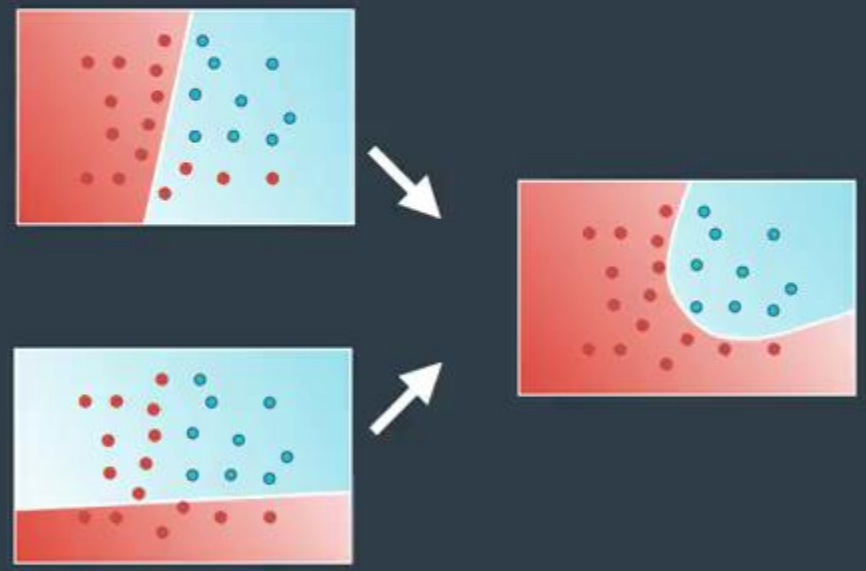
Neural Network



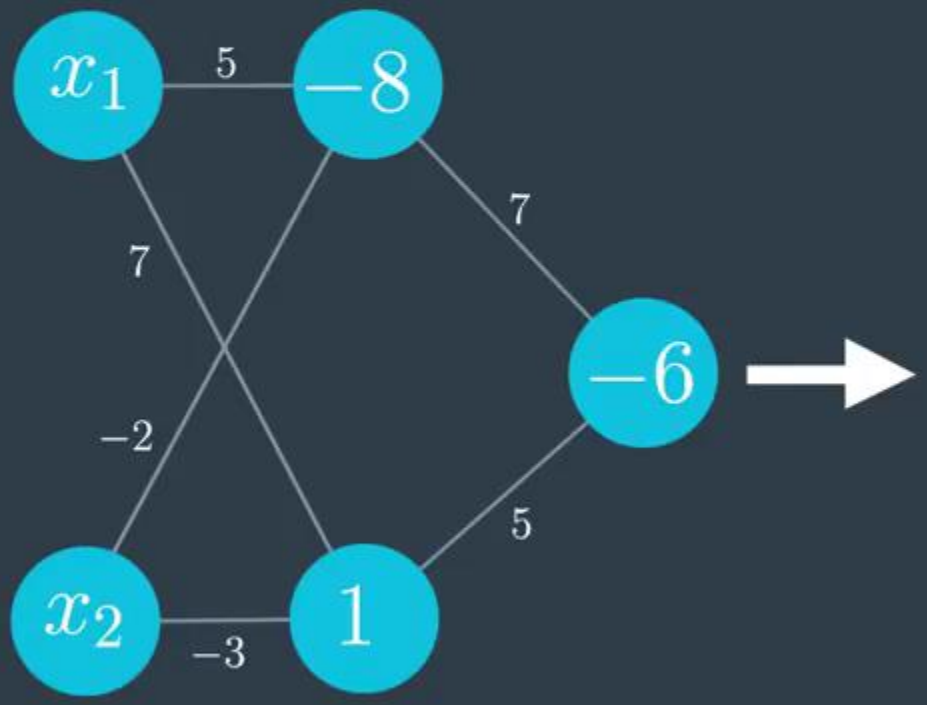


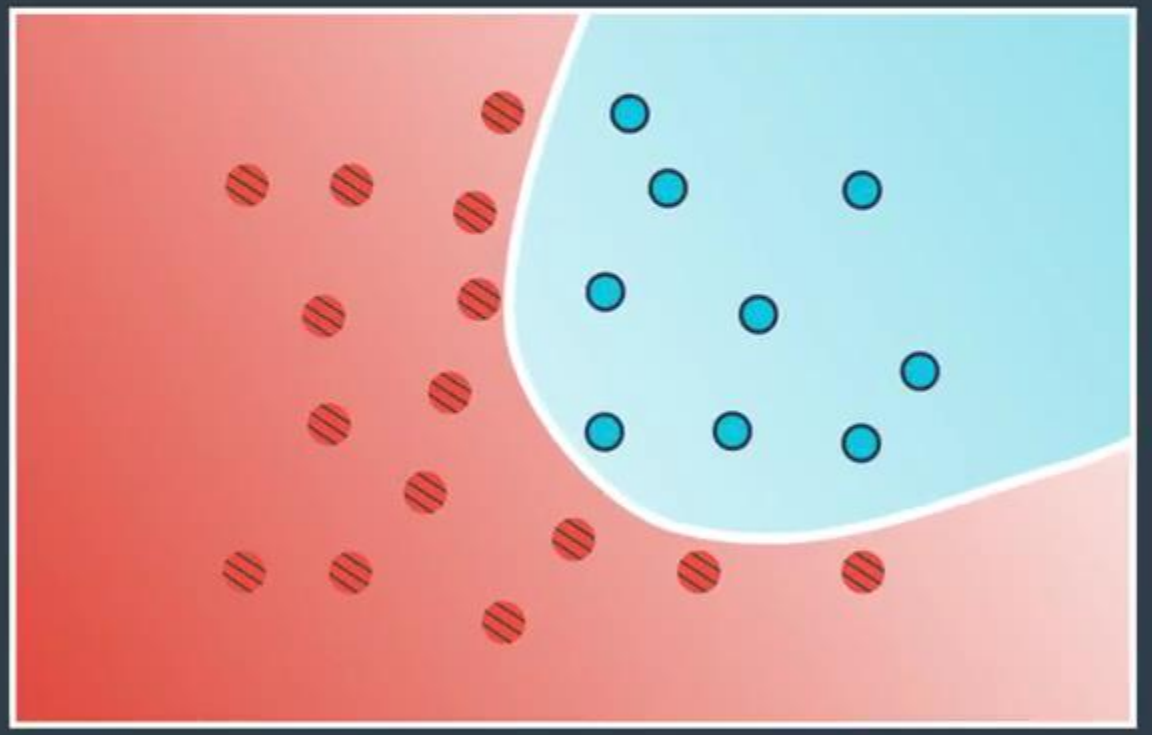
Neural Network



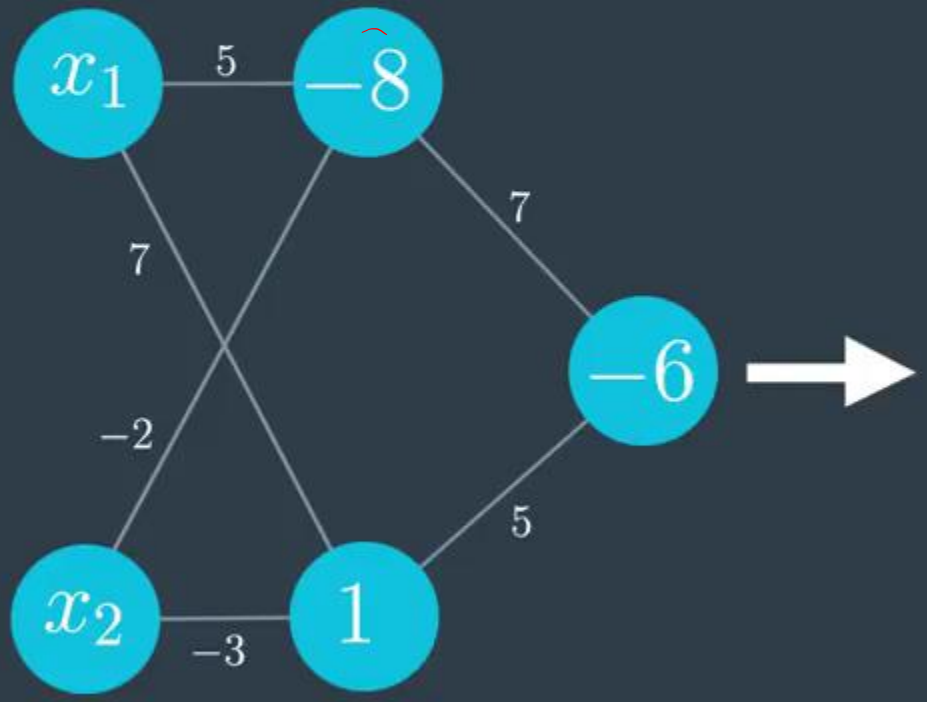


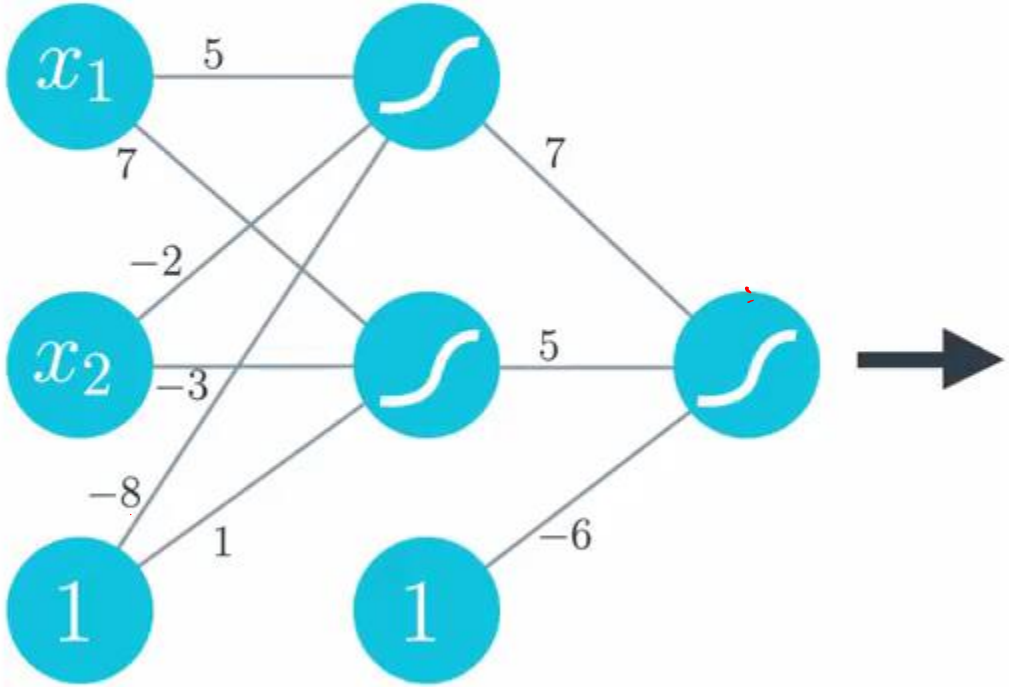
Neural Network



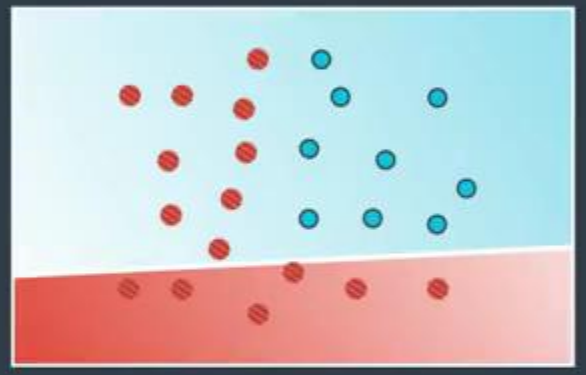


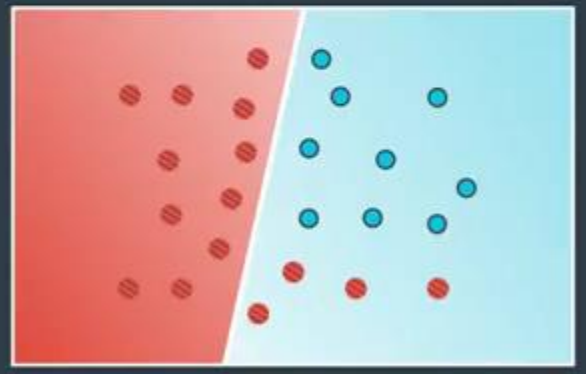
Neural Network

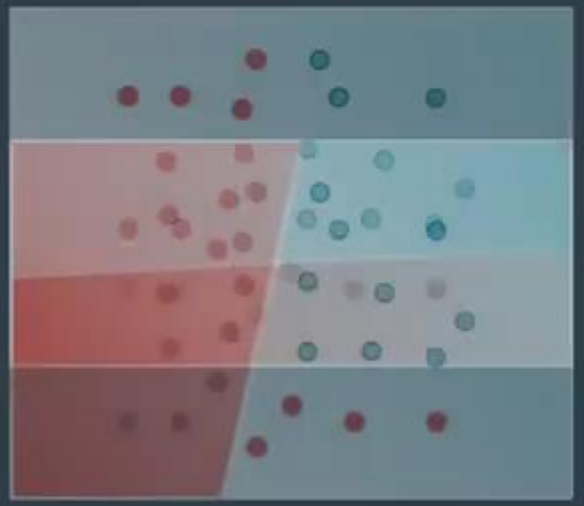




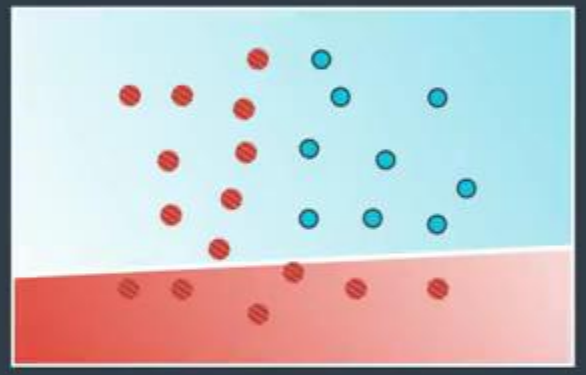
Combining Regions

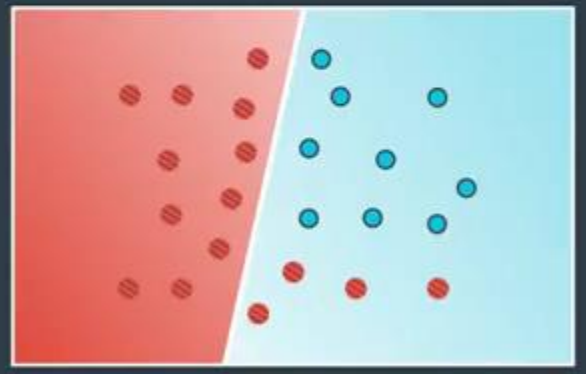


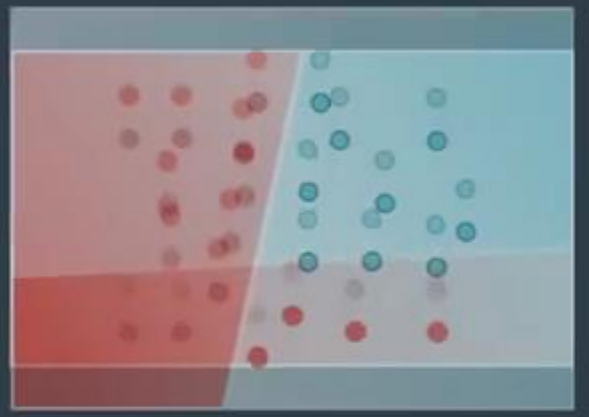




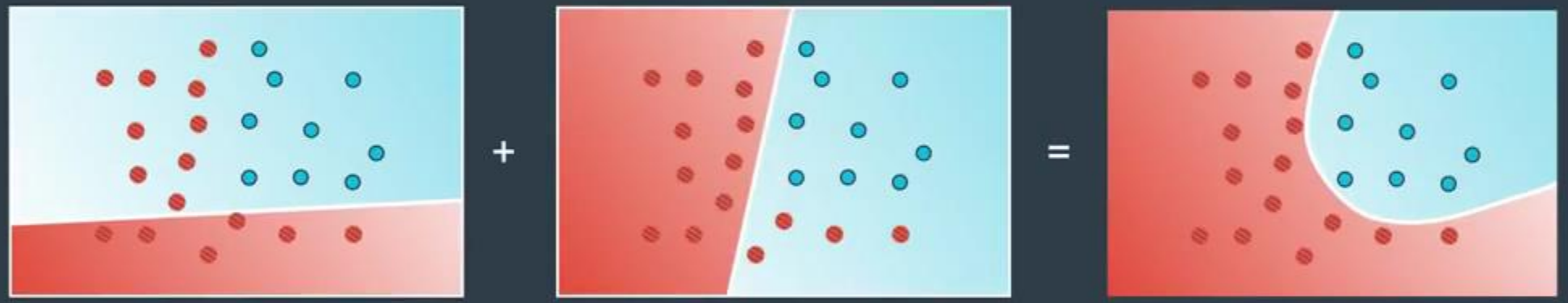
Combining Regions



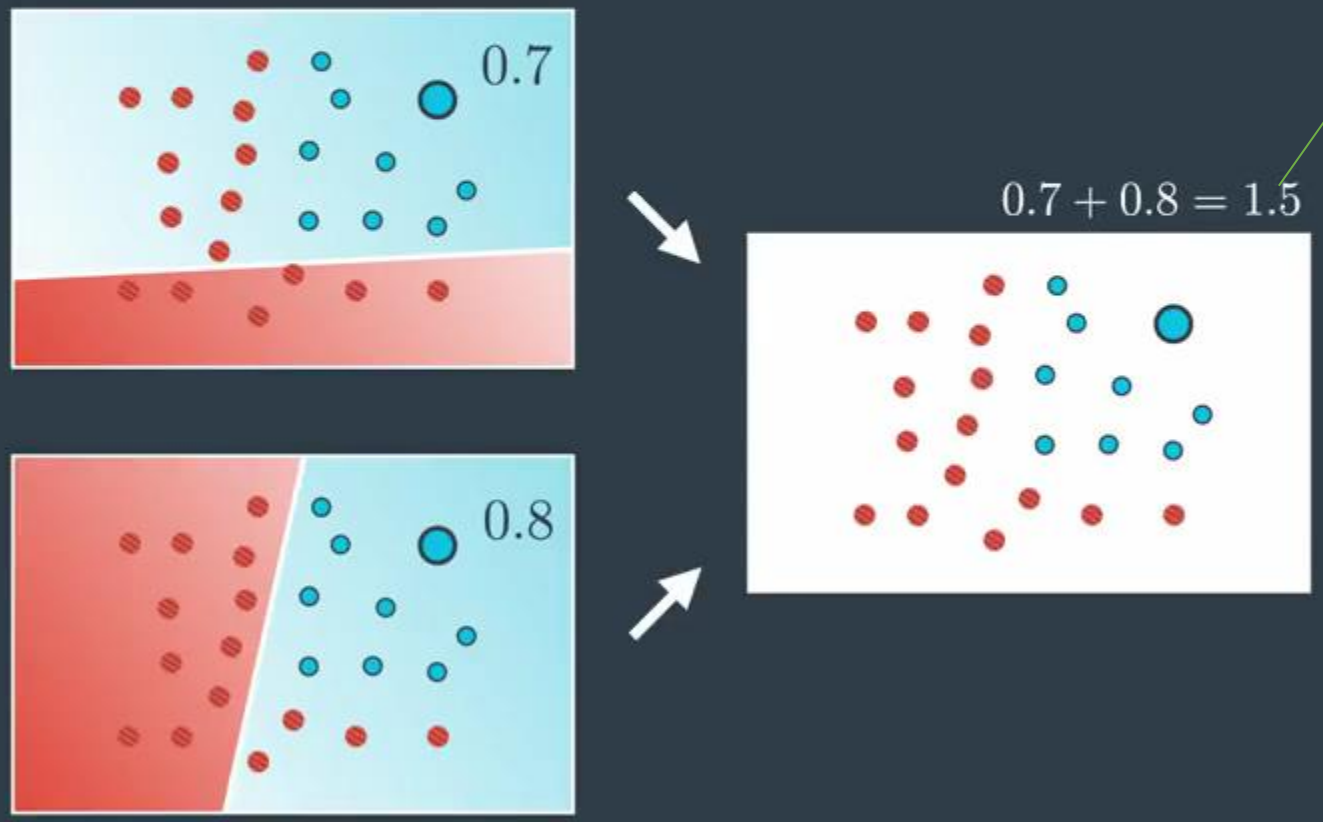




Combining Regions

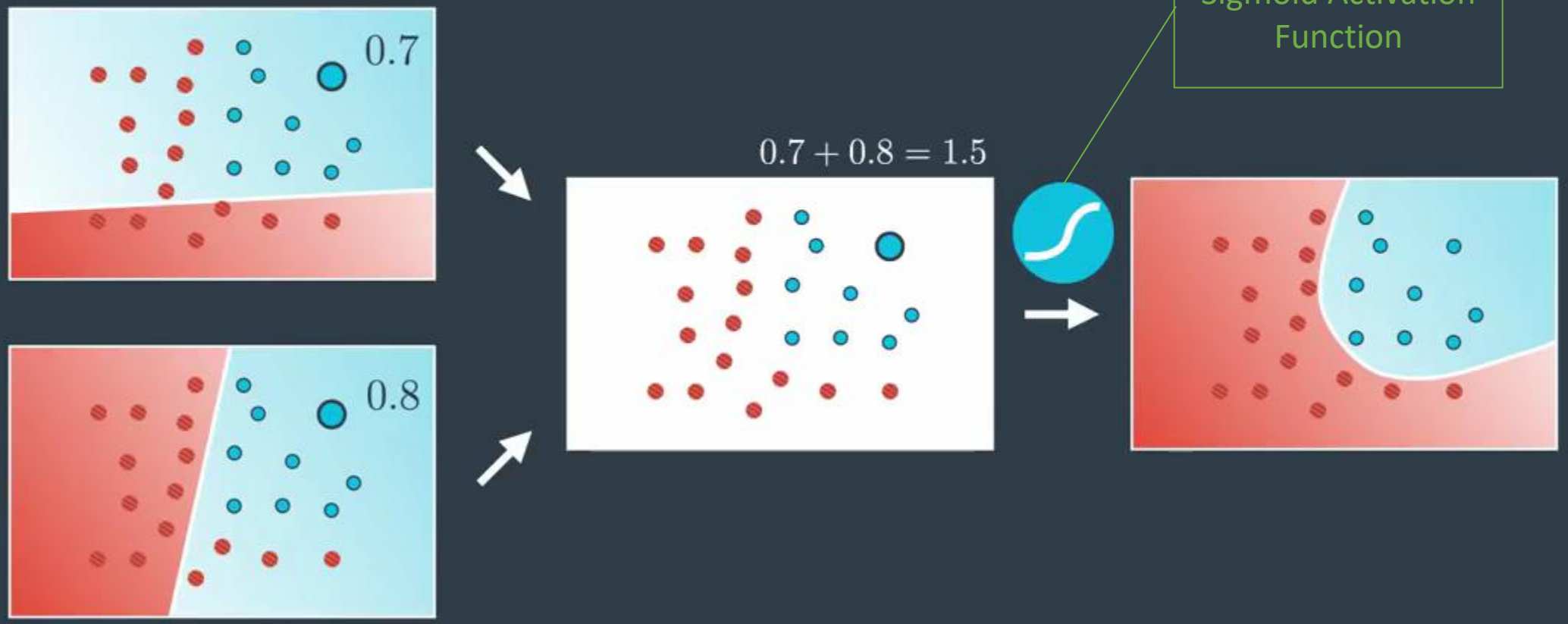


Neural Network

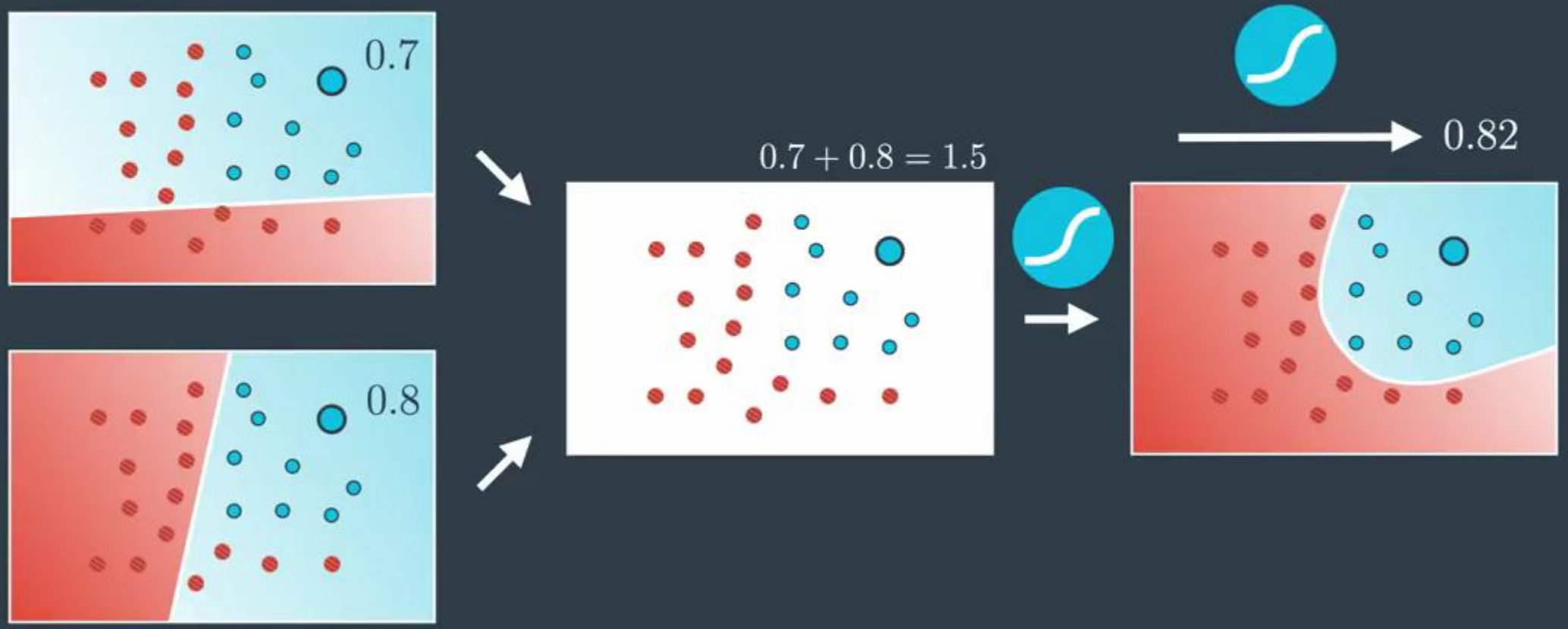


It's bigger than 1
We have to do something

Neural Network

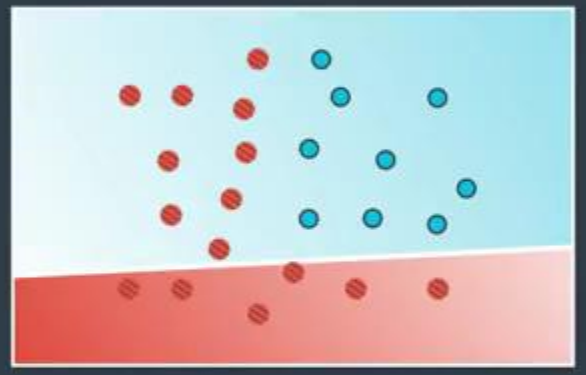


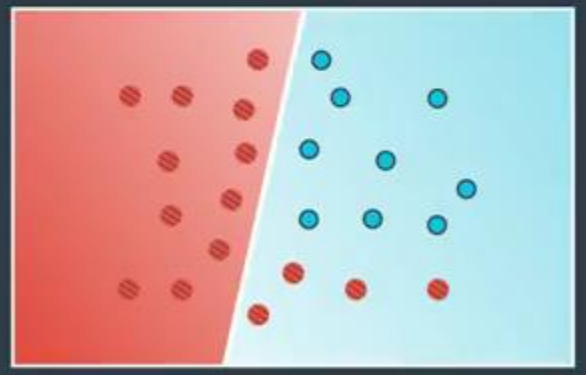
Neural Network



Combining Regions

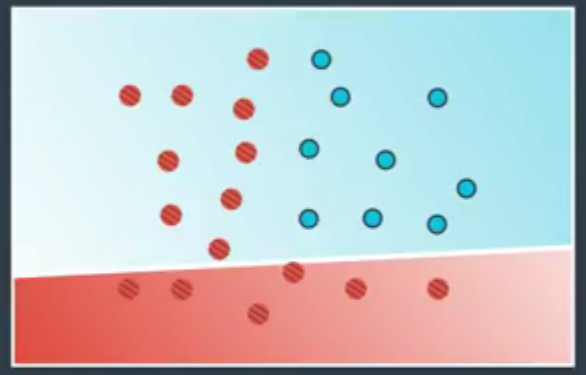
What Weights can do?

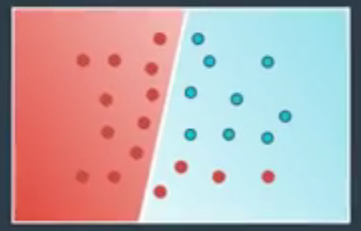




Combining Regions

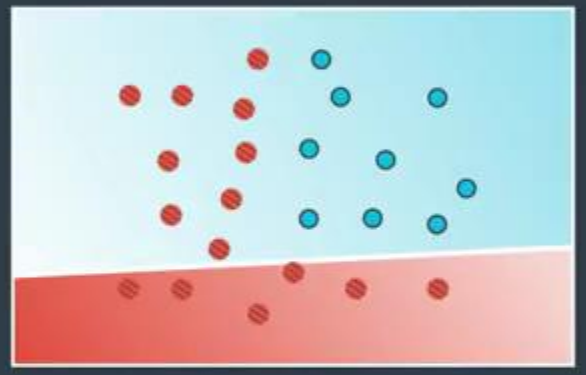
What Weights can do?

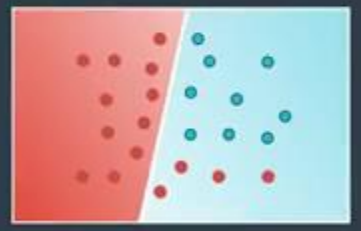


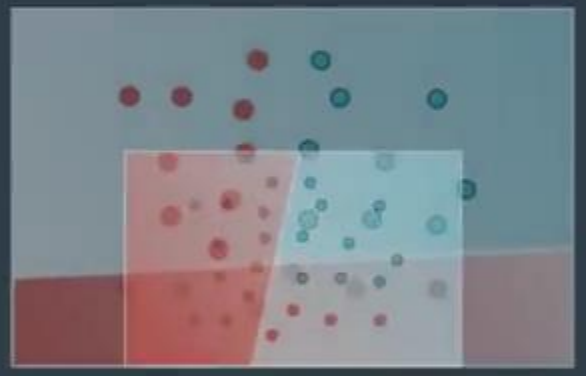


Combining Regions

What Weights can do?

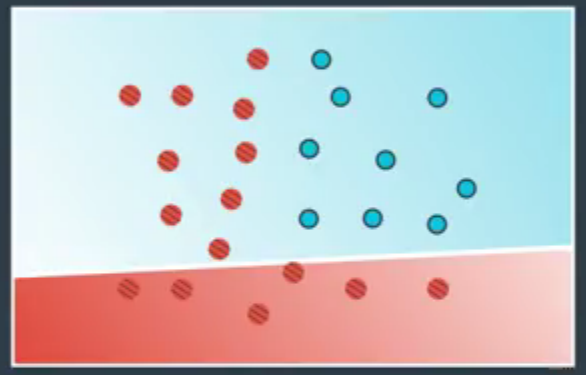


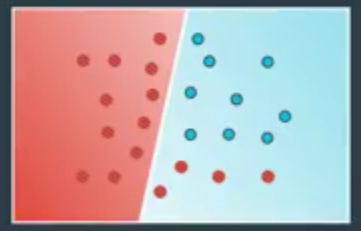


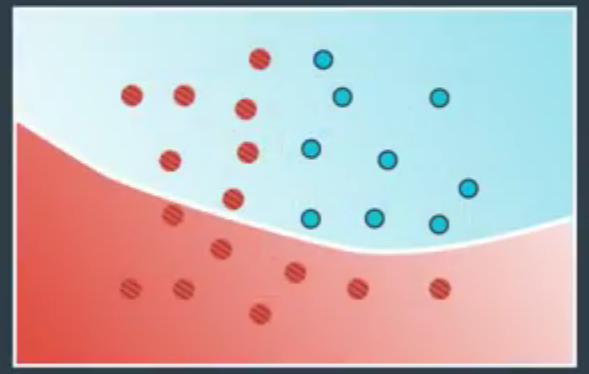


Combining Regions

What Weights can do?

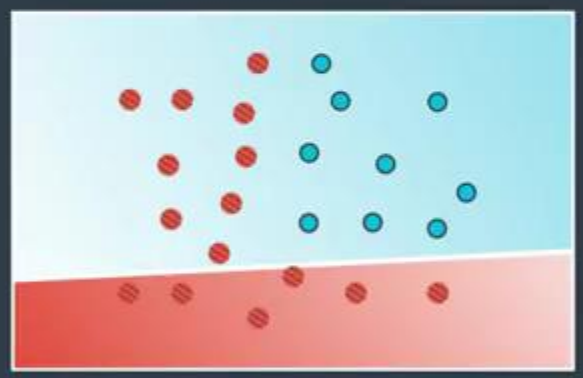






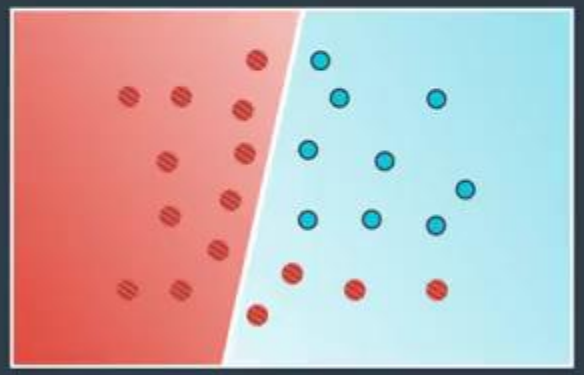
Combining Regions

What Weights can do?

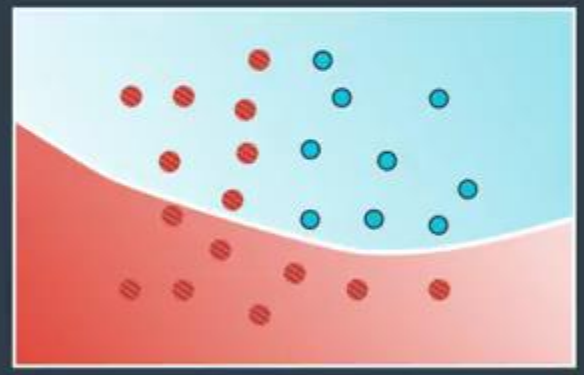


7

+

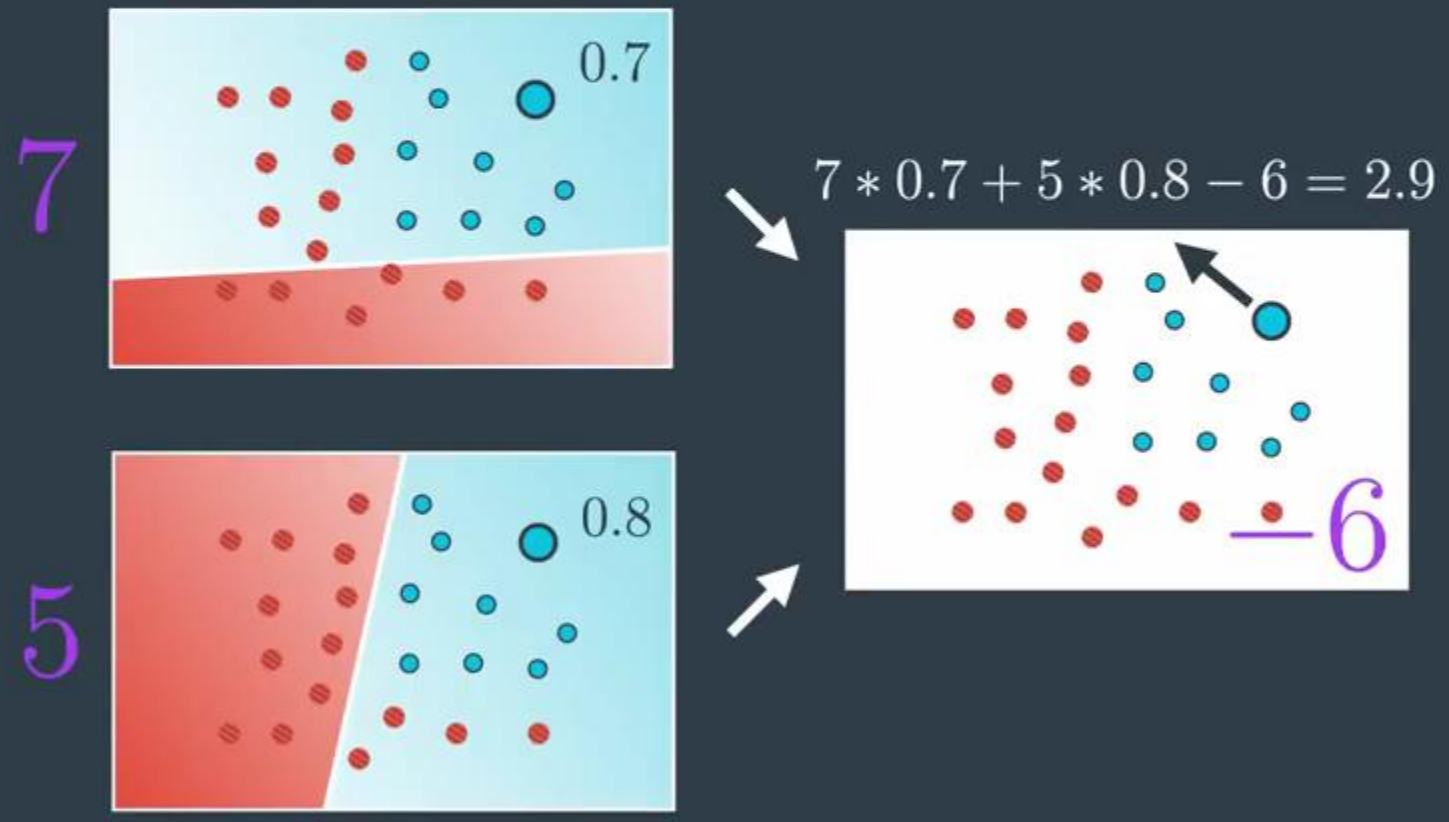


=



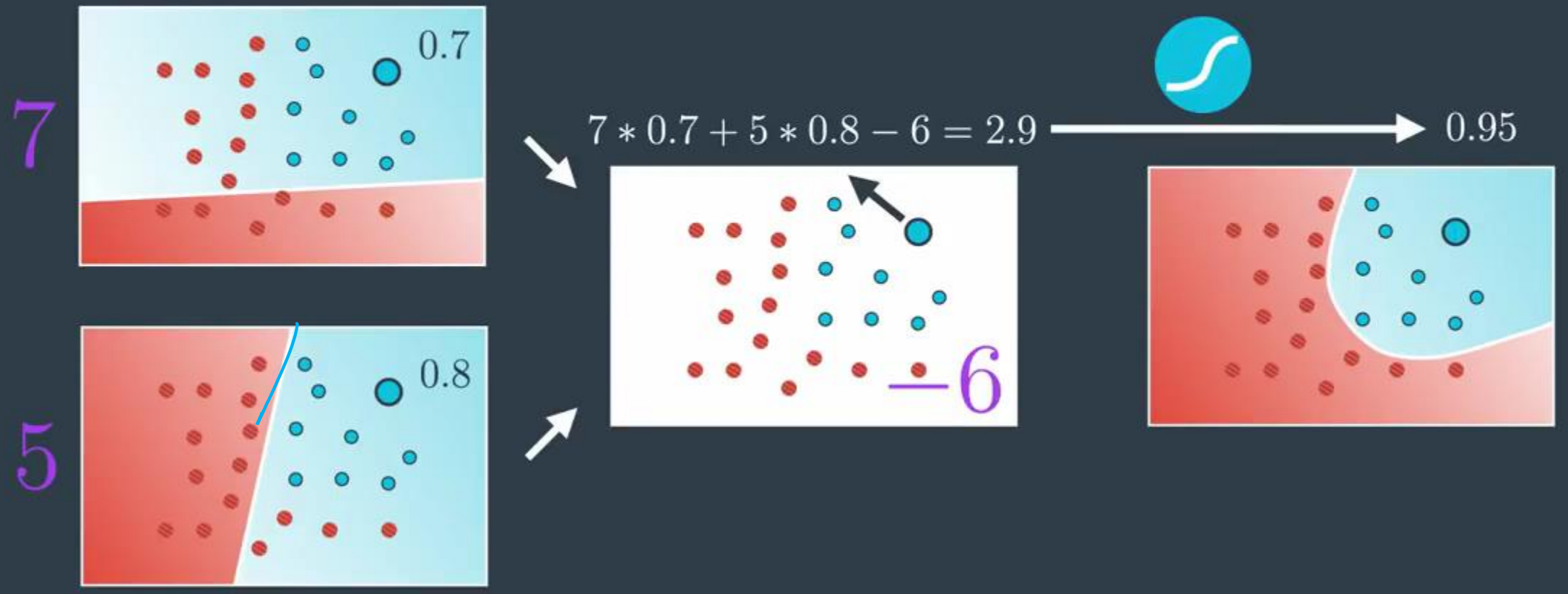
Neural Network

What Weights can do?



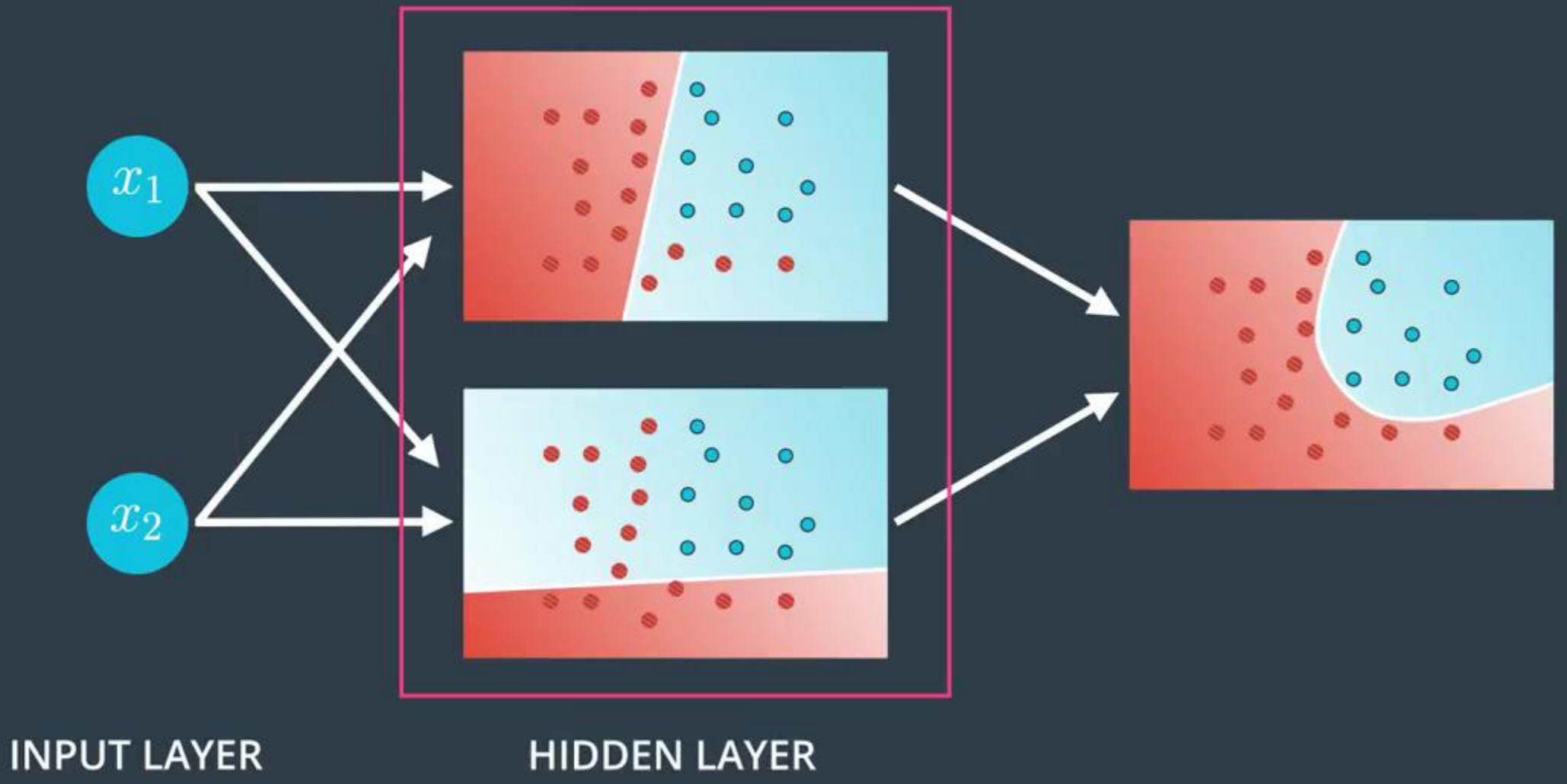
Neural Network

What Weights can do?



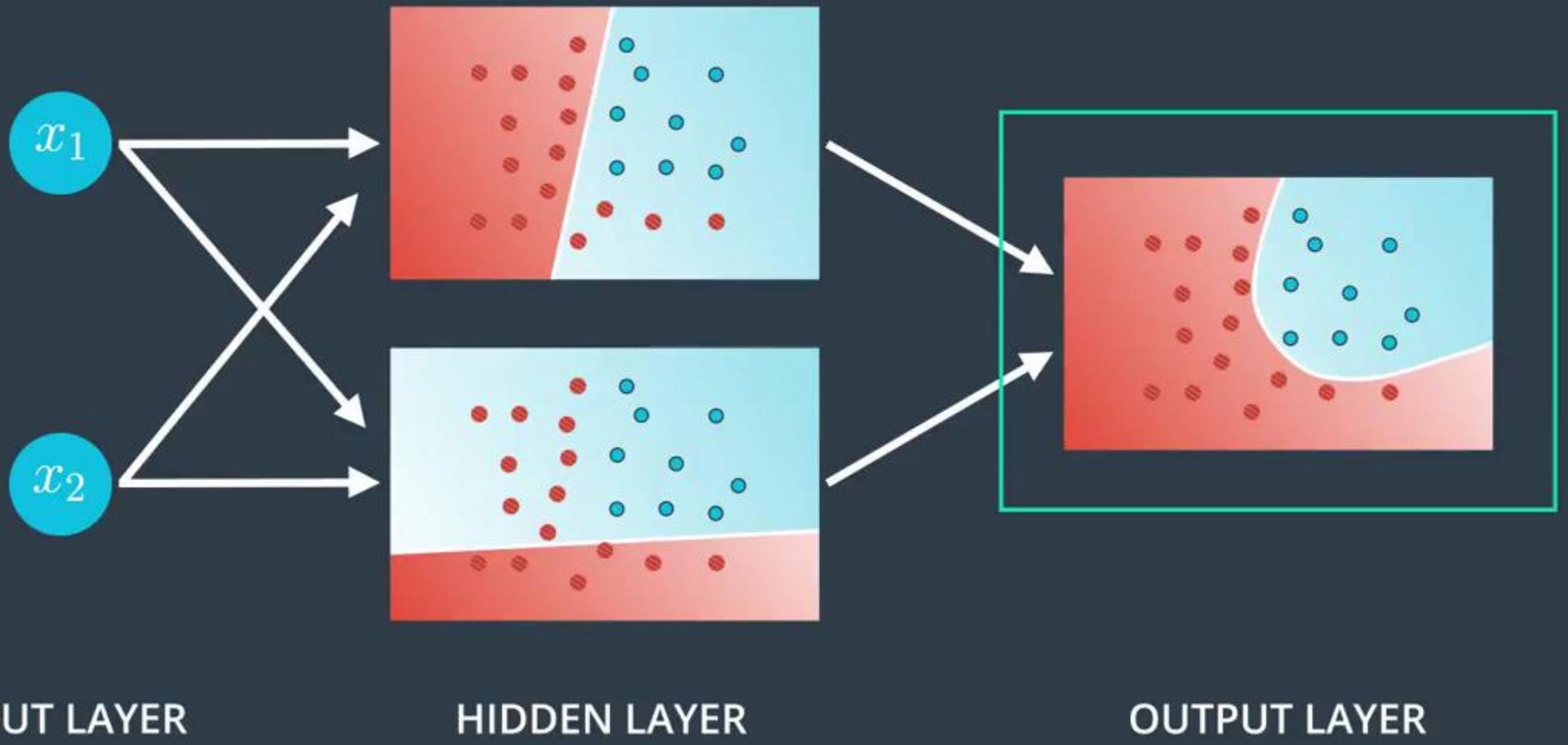
What Layers can do?

Neural Network

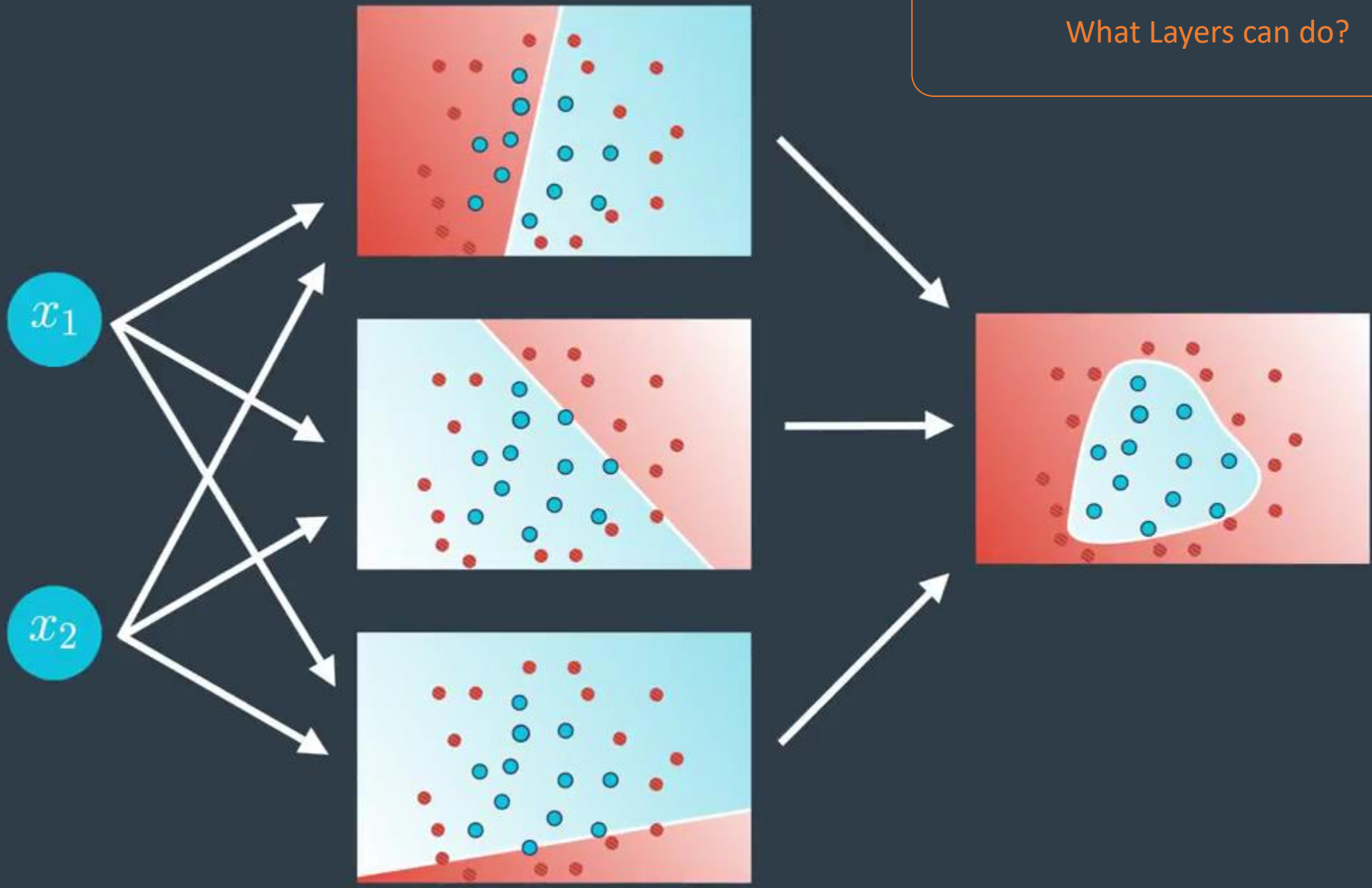


What Layers can do?

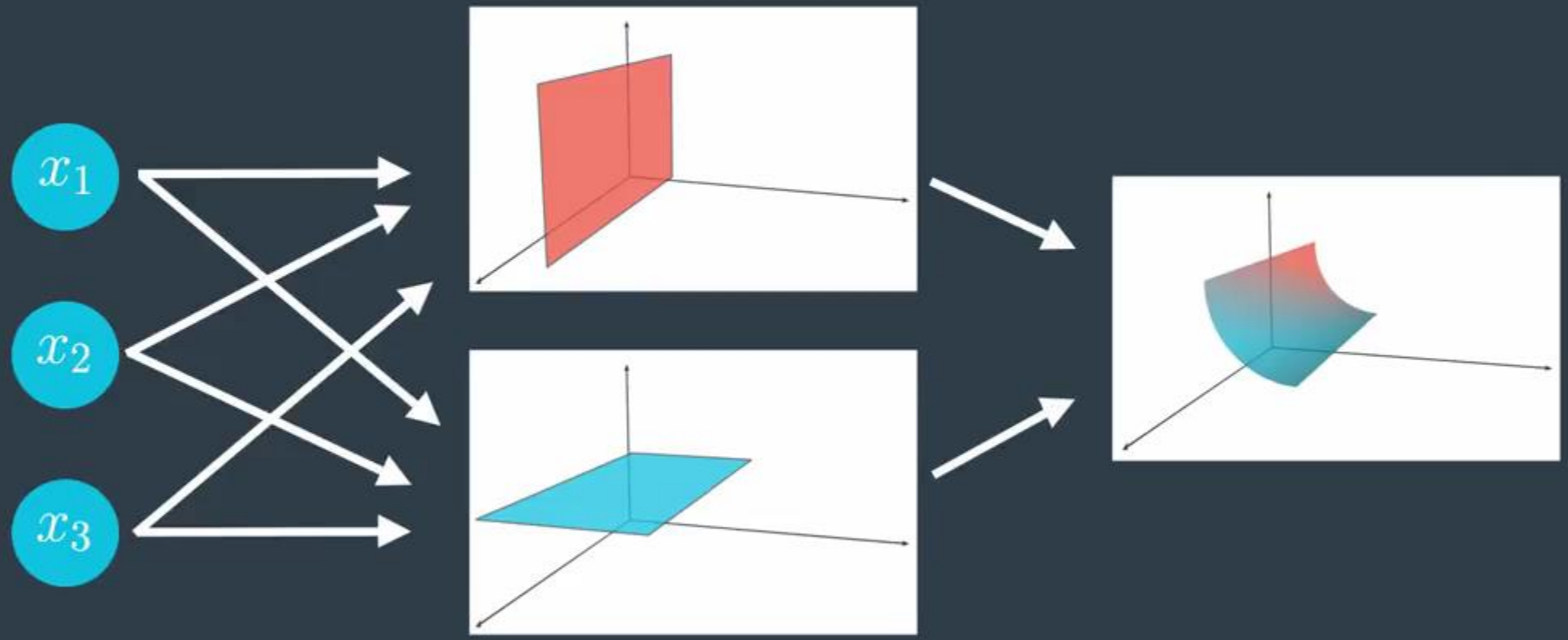
Neural Network



What Layers can do?

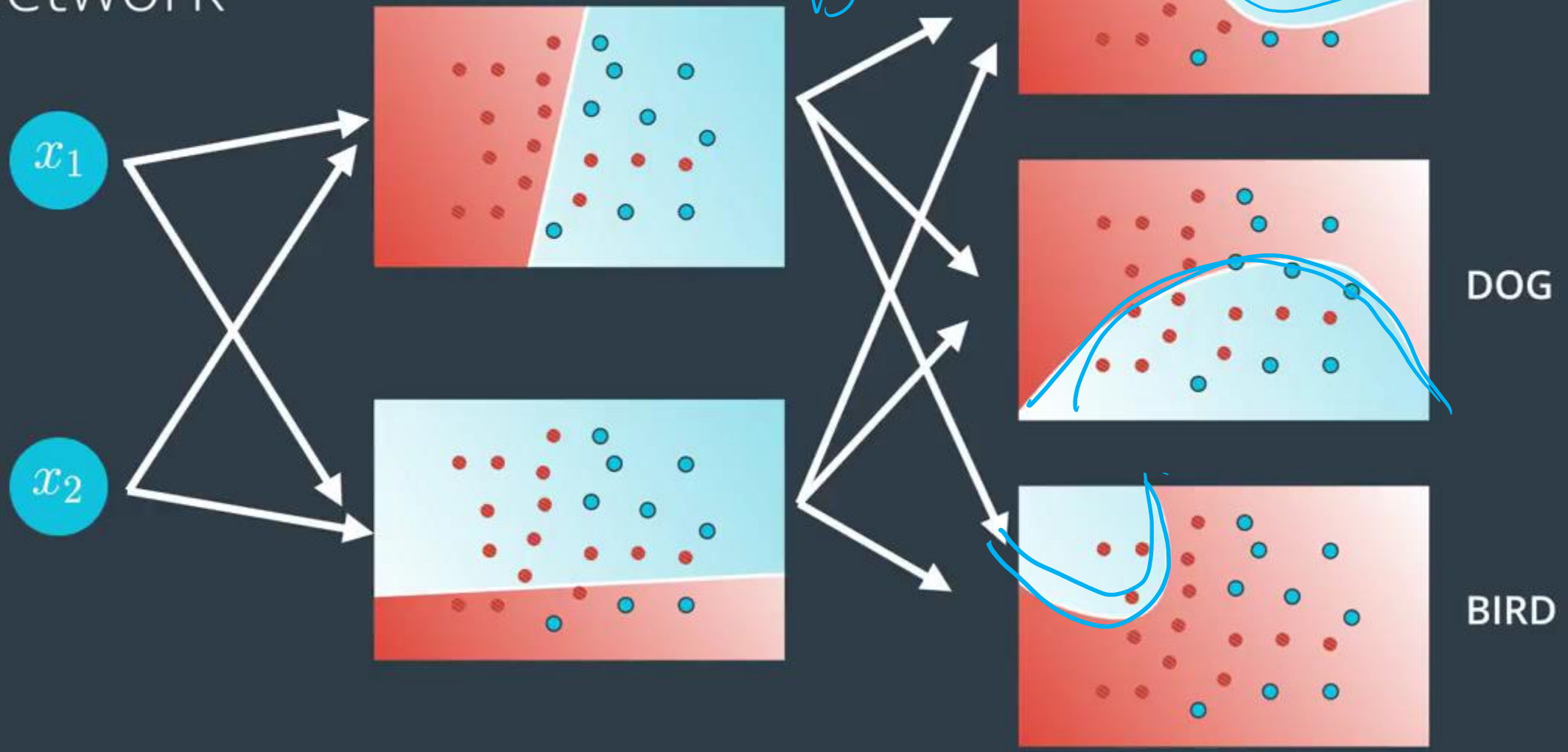


What Layers can do?



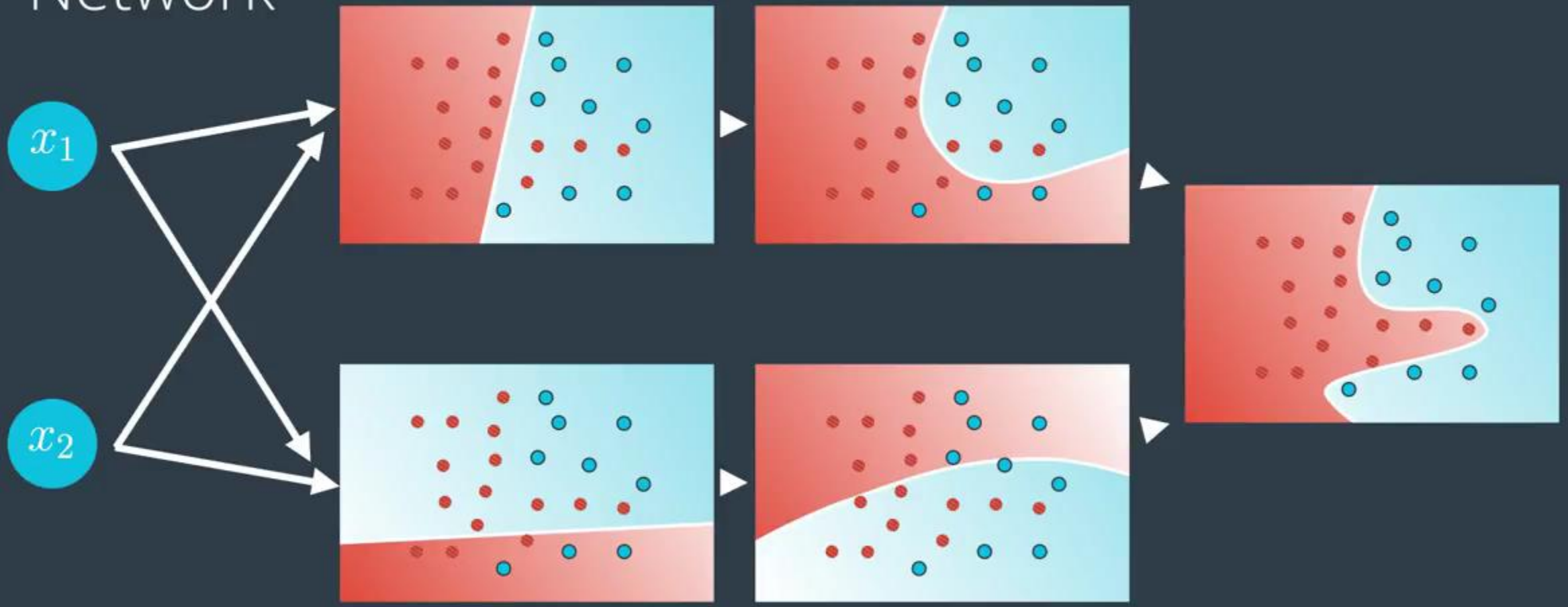
Deep Neural Network

Multi-Class Classification



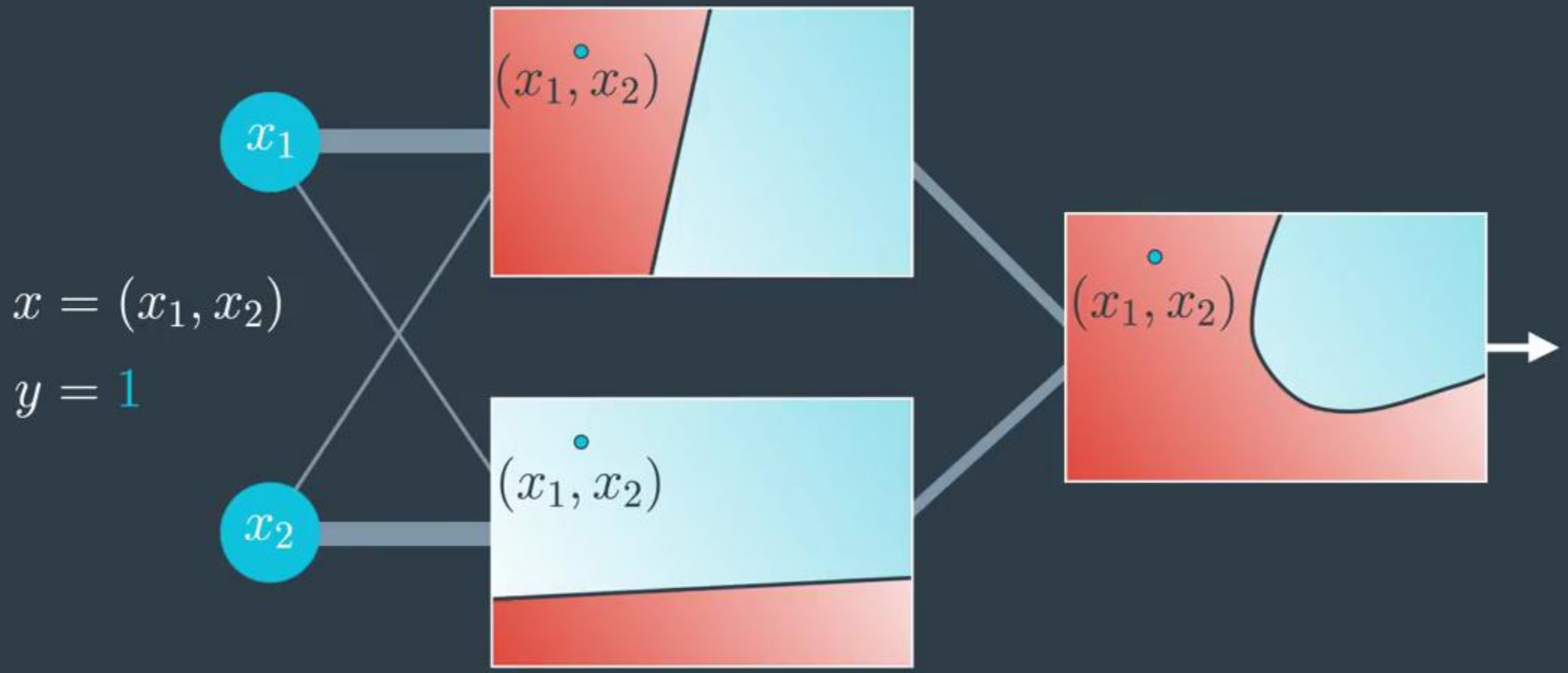
Deep Neural Network

Multi-Layer non-linear Classification

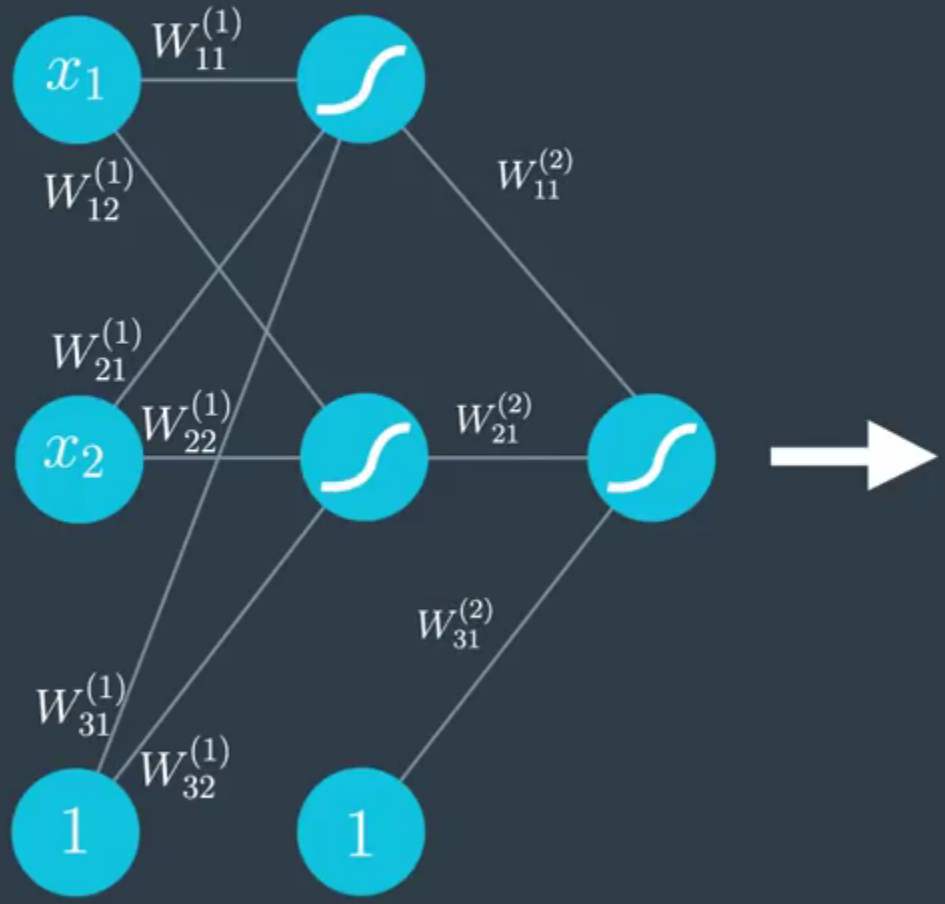


Feed-Forward NN

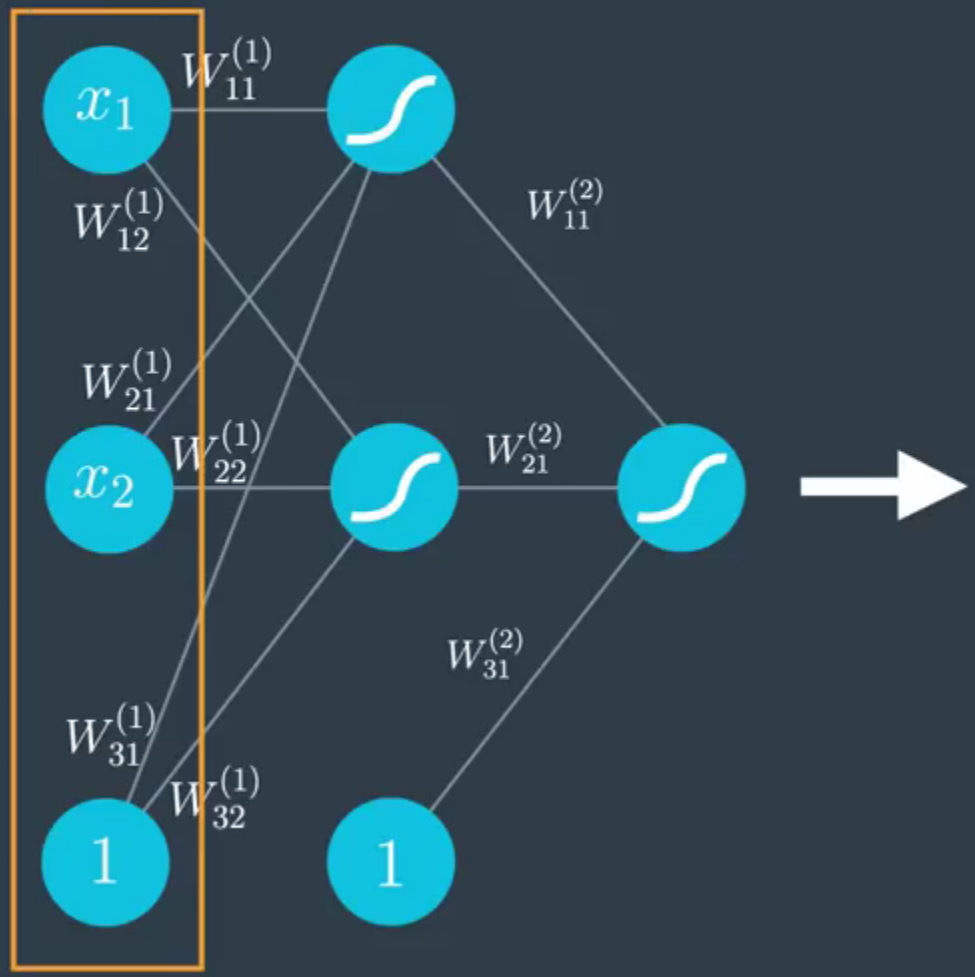
Neural Network



Feedforward

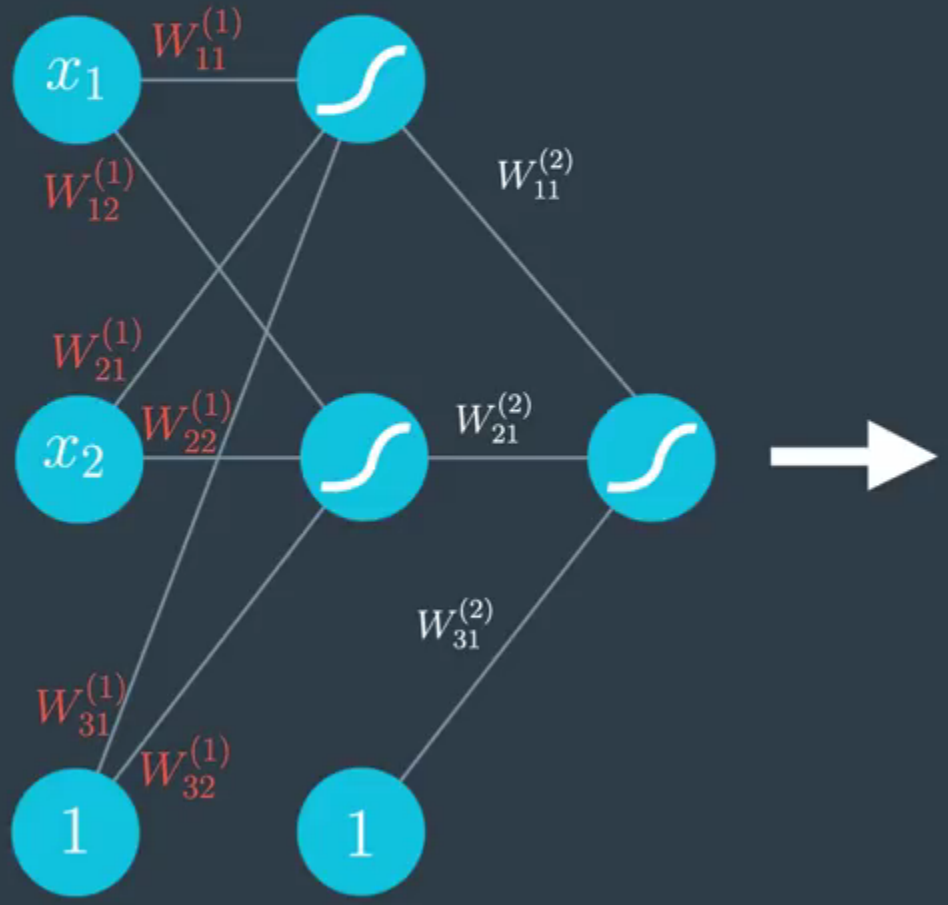


Feedforward



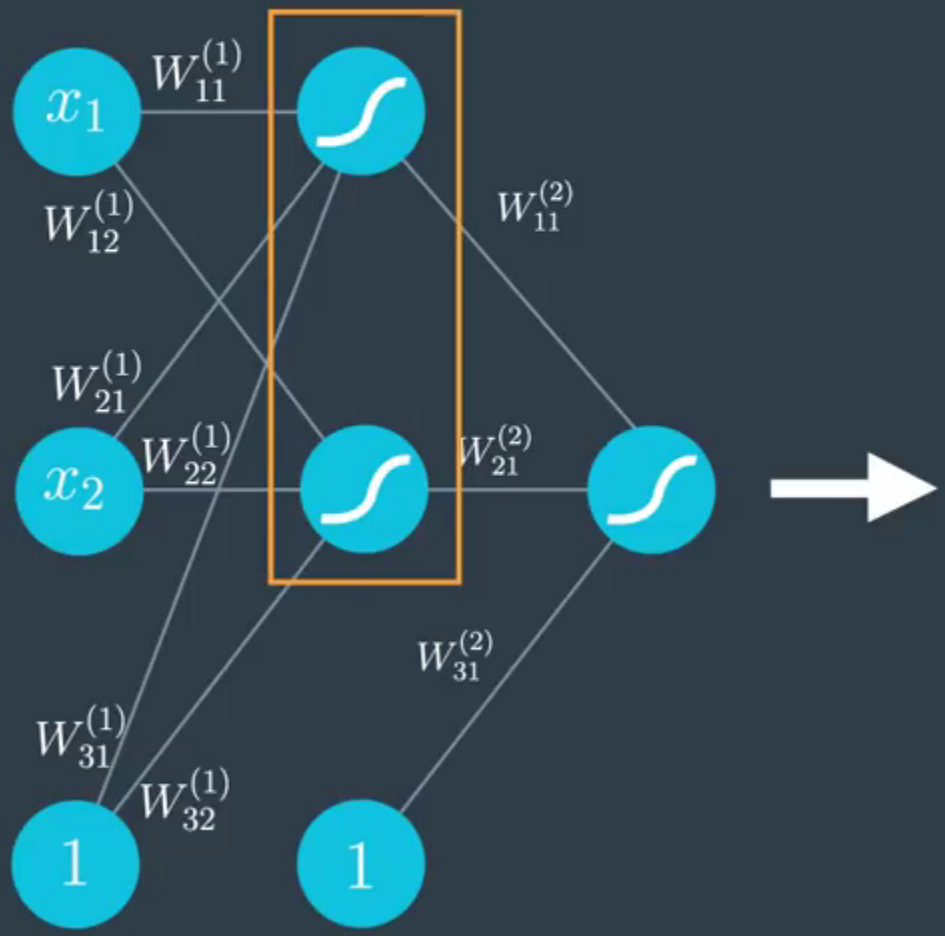
$$\begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

Feedforward



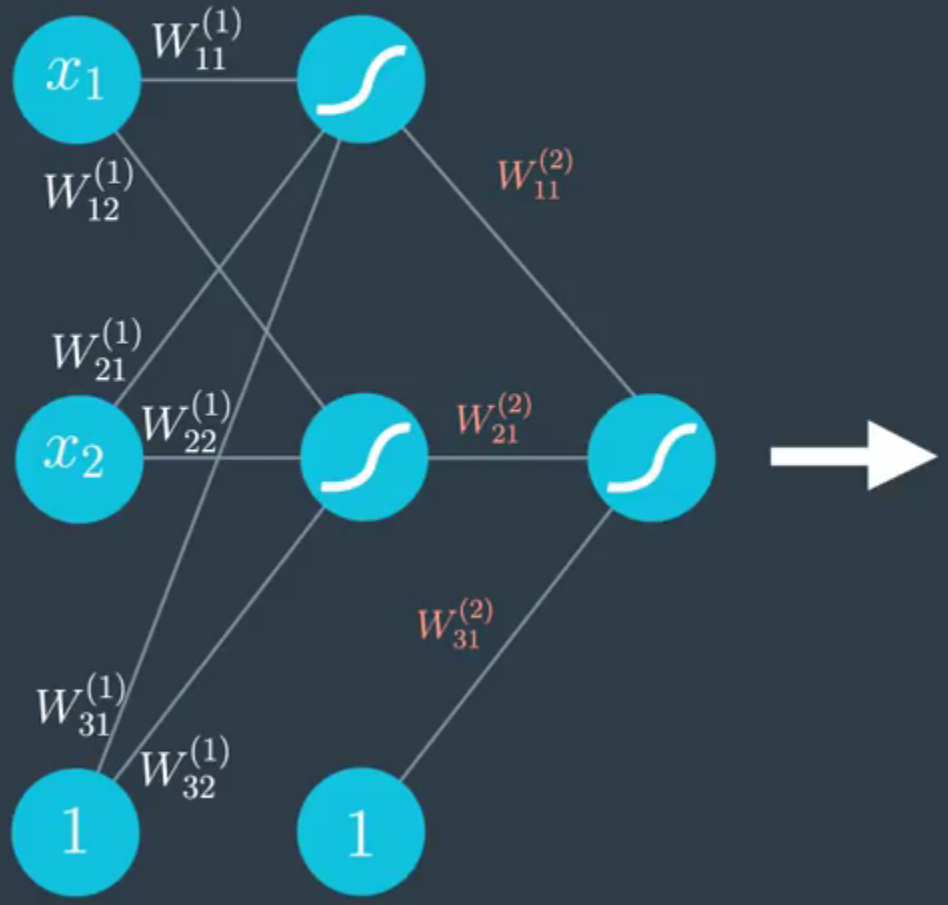
$$\begin{pmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} \\ W_{31}^{(1)} & W_{32}^{(1)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

Feedforward



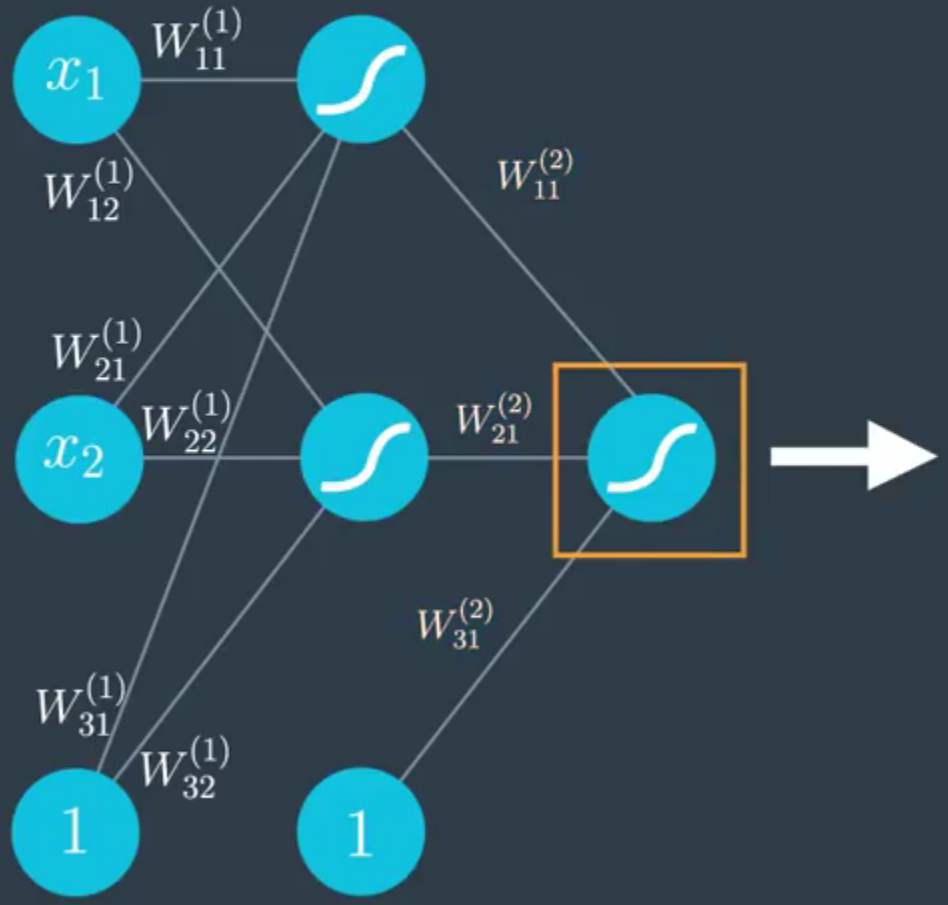
$$\sigma \begin{pmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} \\ W_{31}^{(1)} & W_{32}^{(1)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

Feedforward



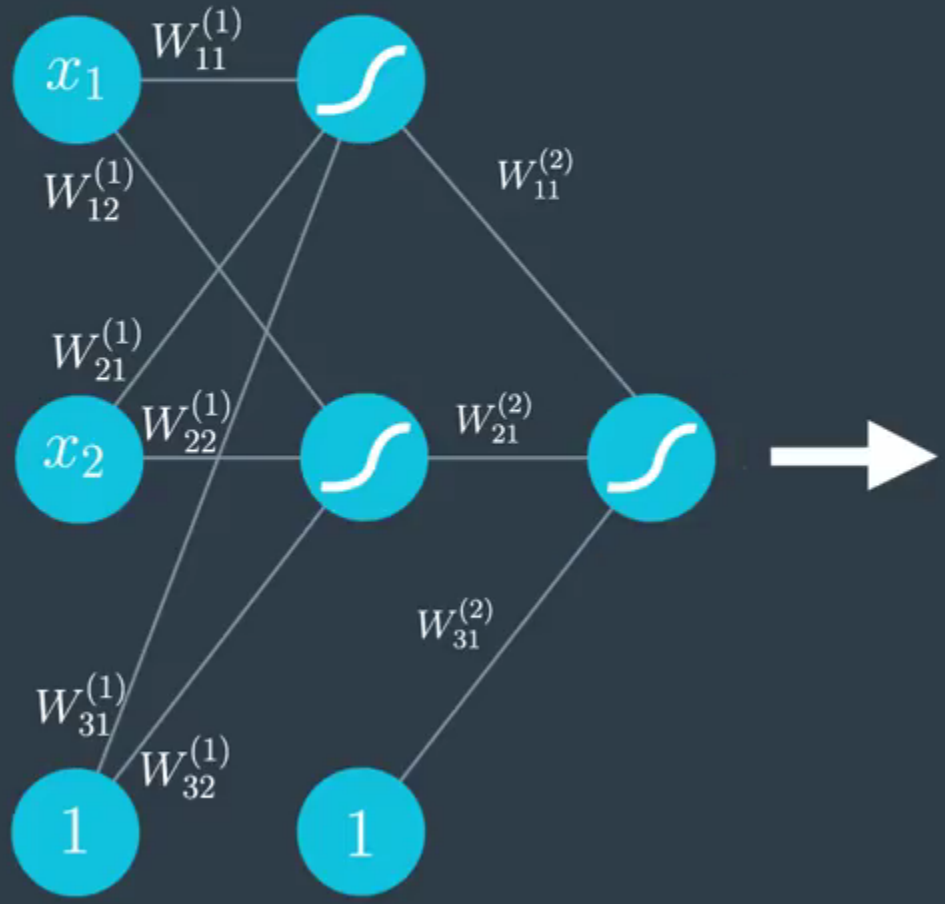
$$\begin{pmatrix} W_{11}^{(2)} \\ W_{21}^{(2)} \\ W_{31}^{(2)} \end{pmatrix} \sigma \begin{pmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} \\ W_{31}^{(1)} & W_{32}^{(1)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

Feedforward



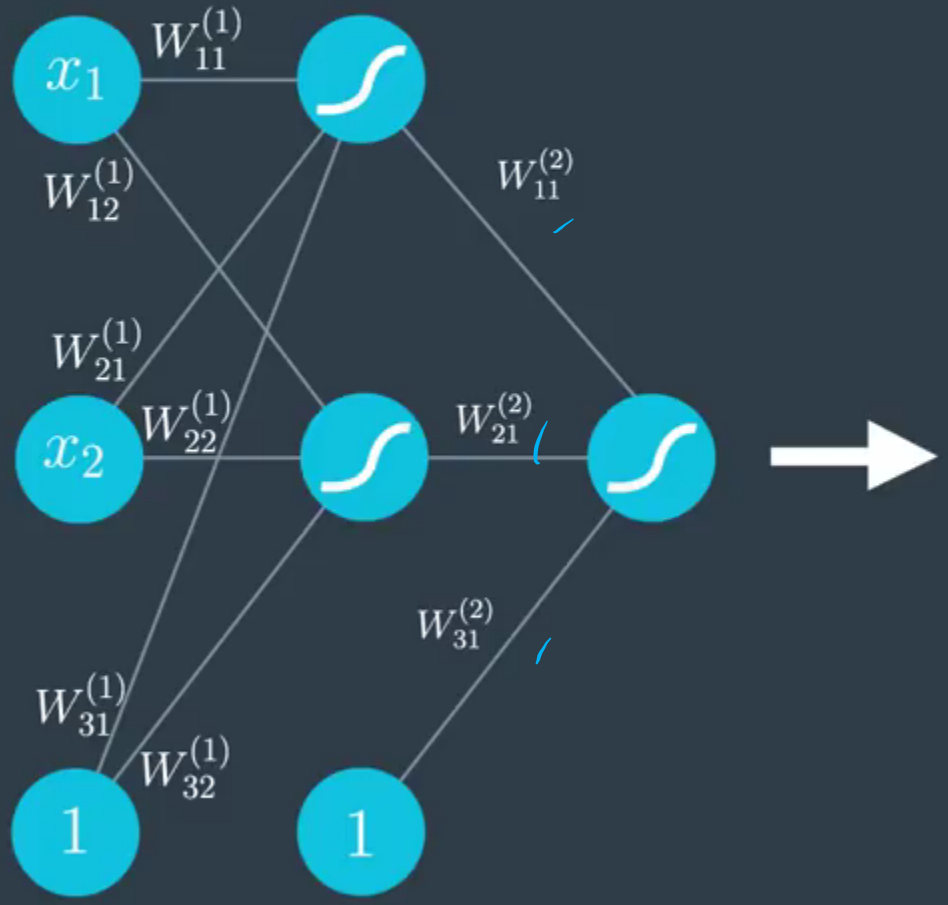
$$\sigma\begin{pmatrix} W_{11}^{(2)} \\ W_{21}^{(2)} \\ W_{31}^{(2)} \end{pmatrix} \sigma\begin{pmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} \\ W_{31}^{(1)} & W_{32}^{(1)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

Feedforward



$$\hat{y} = \sigma \begin{pmatrix} W_{11}^{(2)} \\ W_{21}^{(2)} \\ W_{31}^{(2)} \end{pmatrix} \sigma \begin{pmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} \\ W_{31}^{(1)} & W_{32}^{(1)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

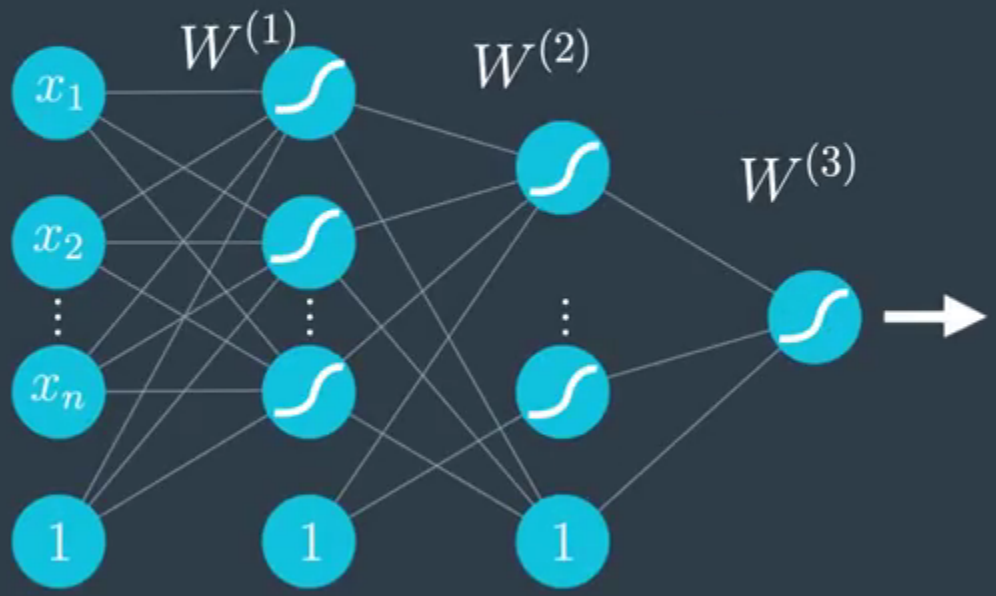
Feedforward



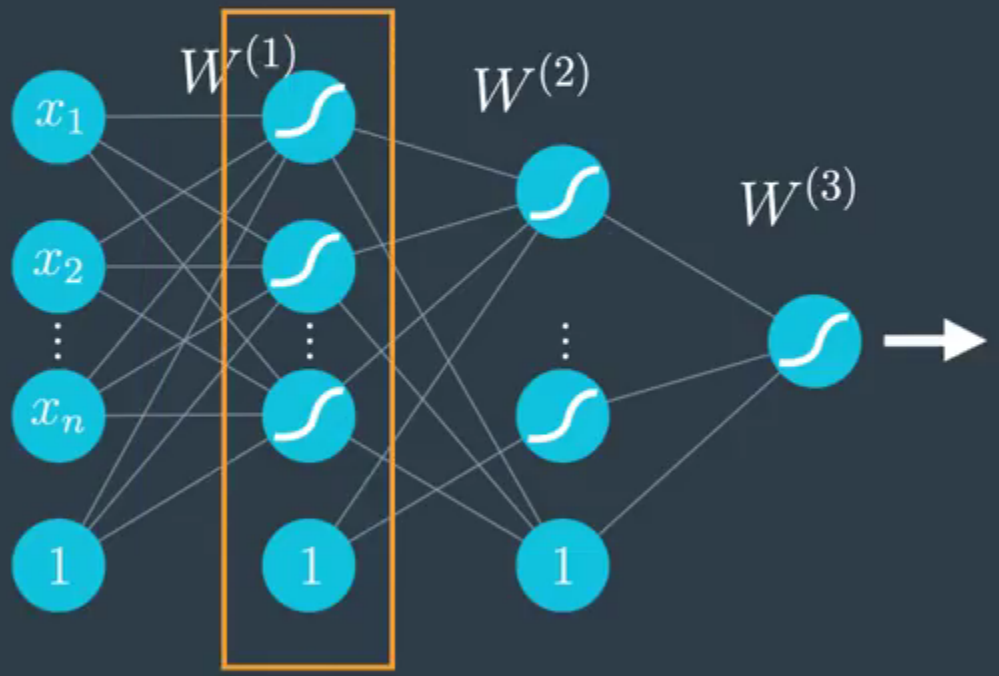
$$\hat{y} = \sigma \begin{pmatrix} W_{11}^{(2)} \\ W_{21}^{(2)} \\ W_{31}^{(2)} \end{pmatrix} \sigma \begin{pmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} \\ W_{31}^{(1)} & W_{32}^{(1)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}$$

$$\hat{y} = \sigma \circ W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

Multi-layer Perceptron



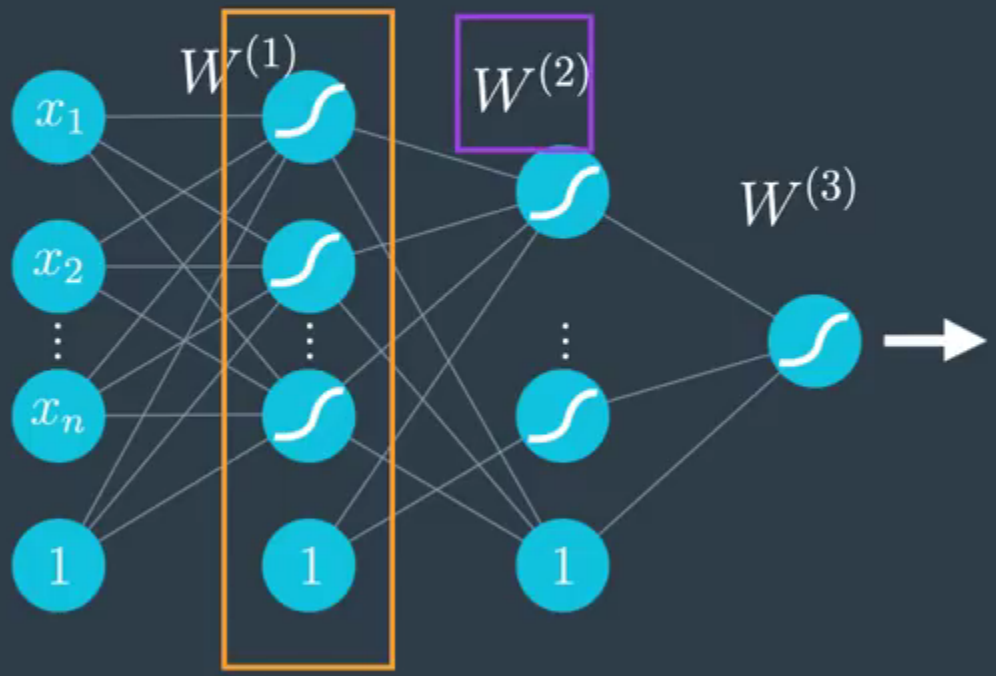
Multi-layer Perceptron



PREDICTION

$$\hat{y} = \sigma \circ W^{(1)}(x)$$

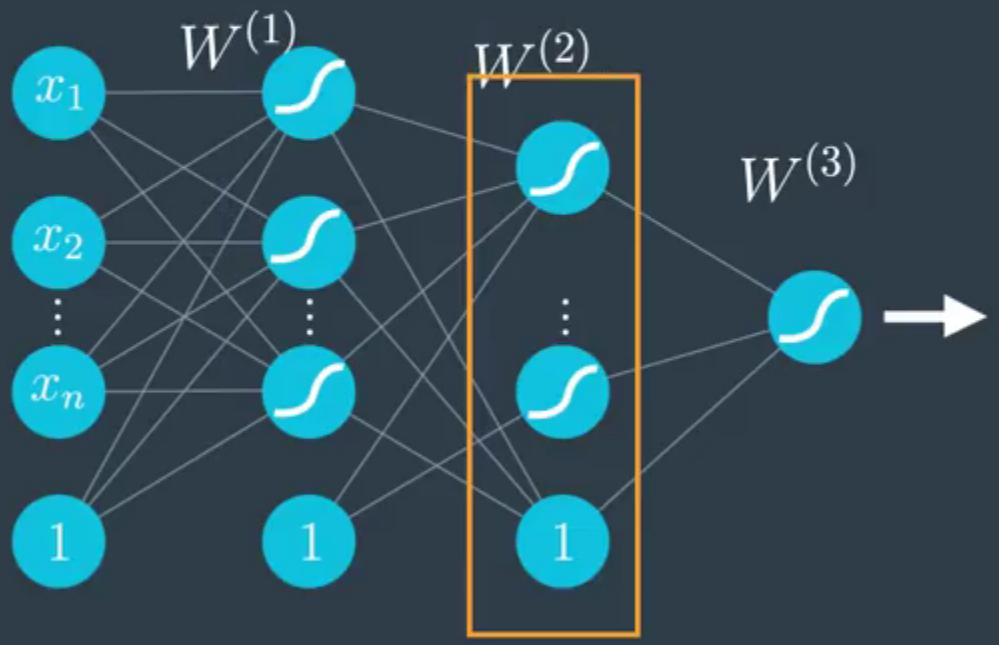
Multi-layer Perceptron



PREDICTION

$$\hat{y} = W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

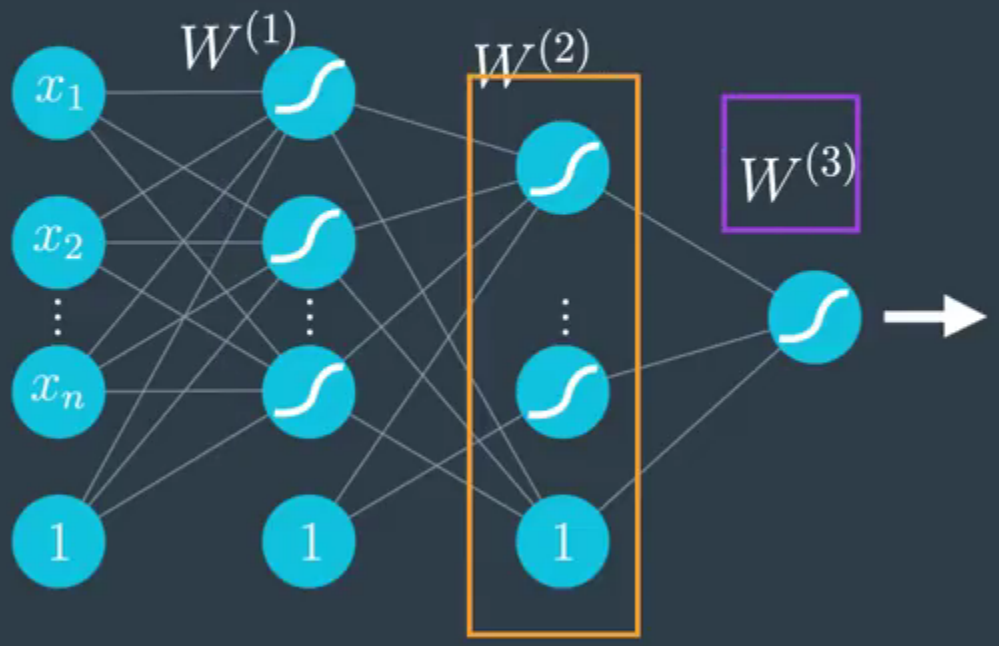
Multi-layer Perceptron



PREDICTION

$$\hat{y} = \sigma \circ W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

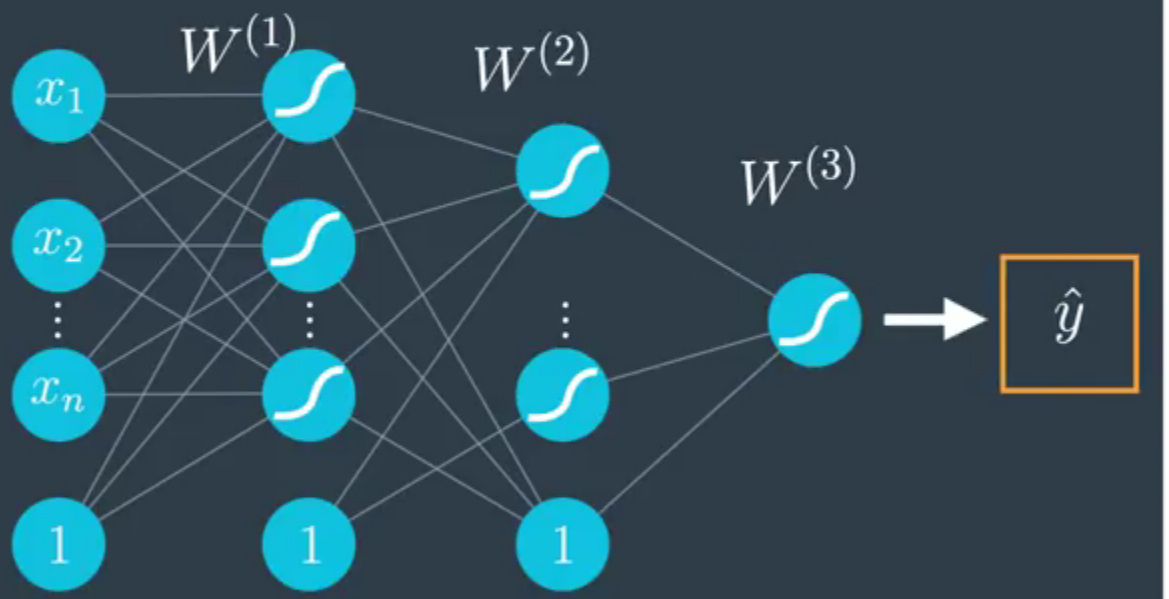
Multi-layer Perceptron



PREDICTION

$$\hat{y} = \sigma \circ W^{(3)} \circ \sigma \circ W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

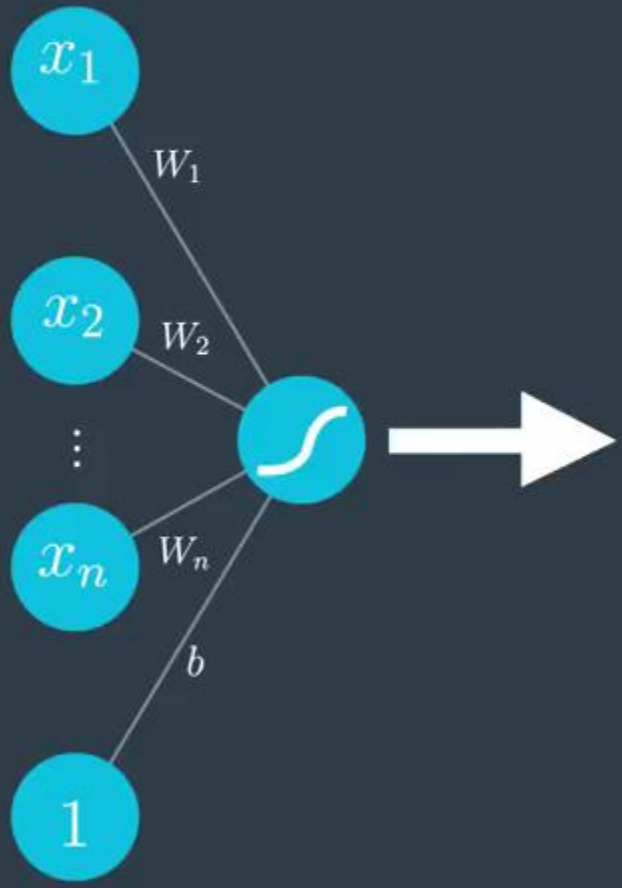
Multi-layer Perceptron



PREDICTION

$$\hat{y} = \sigma \circ W^{(3)} \circ \sigma \circ W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

Perceptron

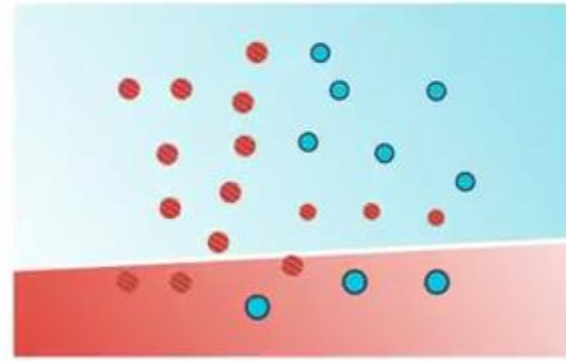


PREDICTION

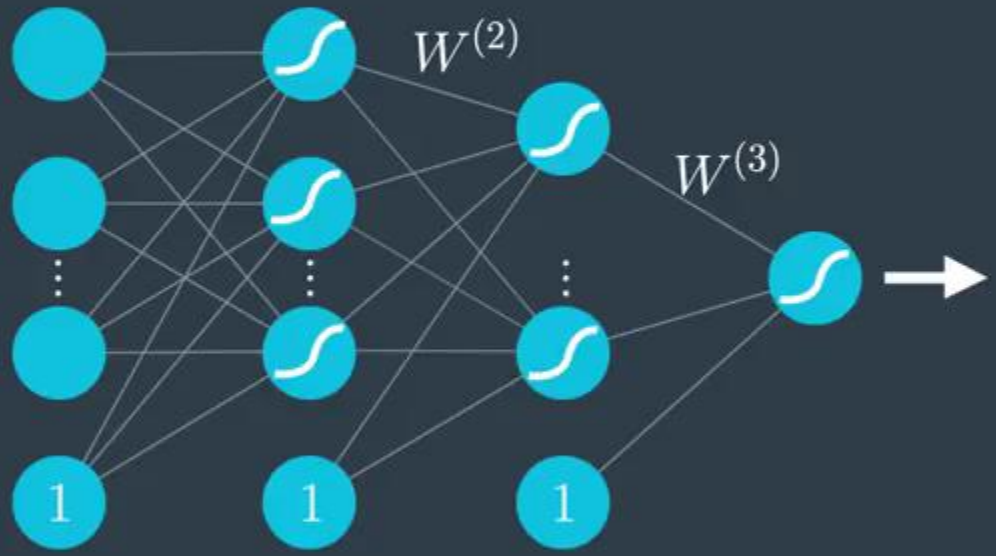
$$\hat{y} = \sigma(Wx + b)$$

ERROR FUNCTION

$$E(W) = -\frac{1}{m}\sum_{i=1}^{m} y_i ln(\hat{y}_i) + (1 - y_i) ln(1 - \hat{y}_i)$$



Multi-layer Perceptron

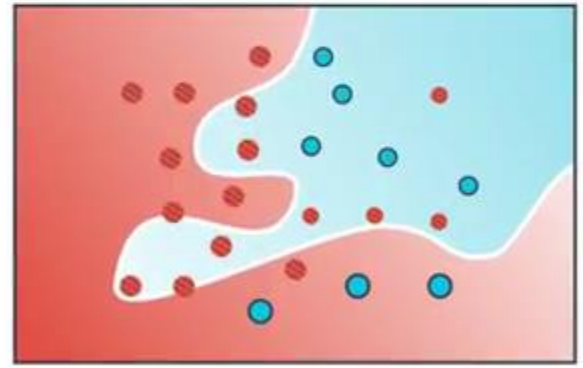


PREDICTION

$$\hat{y} = \sigma \circ W^{(3)} \circ \sigma \circ W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

ERROR FUNCTION

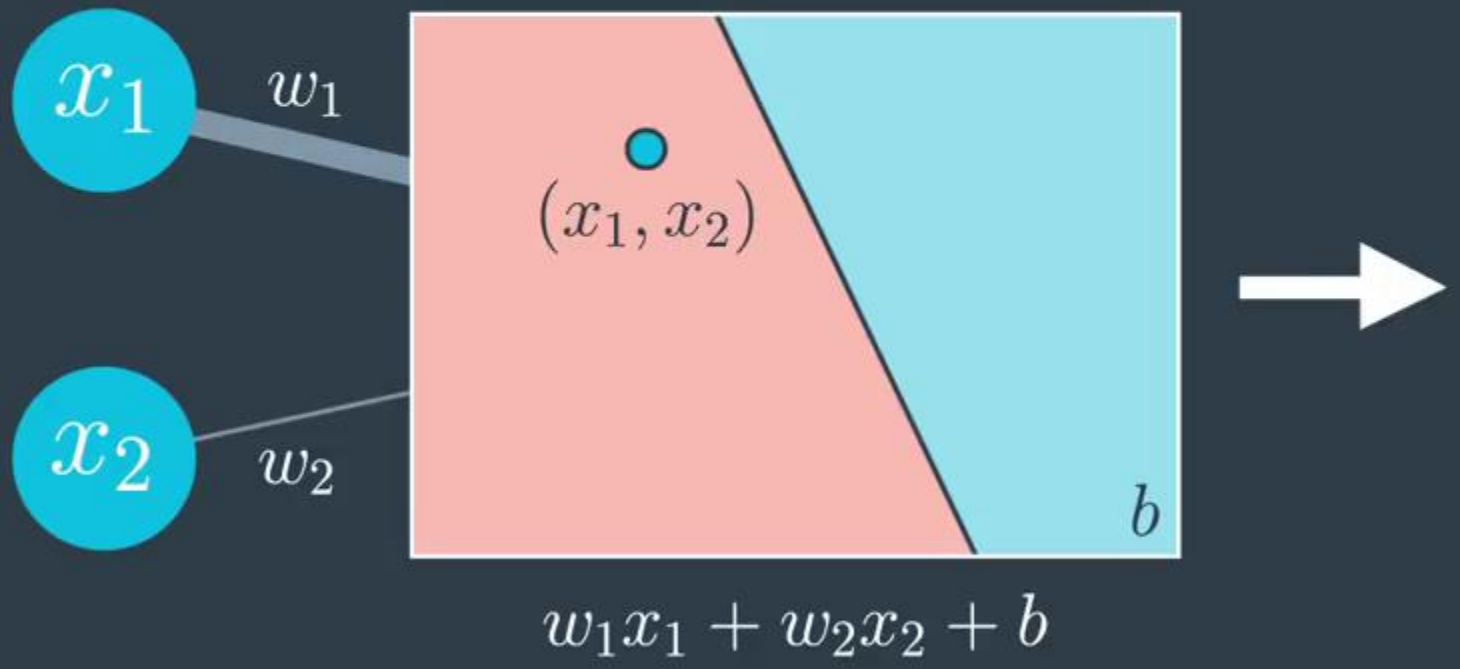
$$E(W) = -\frac{1}{m}\sum_{i=1}^{m} y_i ln(\hat{y}_i) + (1 - y_i) ln(1 - \hat{y}_i)$$



Backpropagation

$x = (x_1, x_2)$

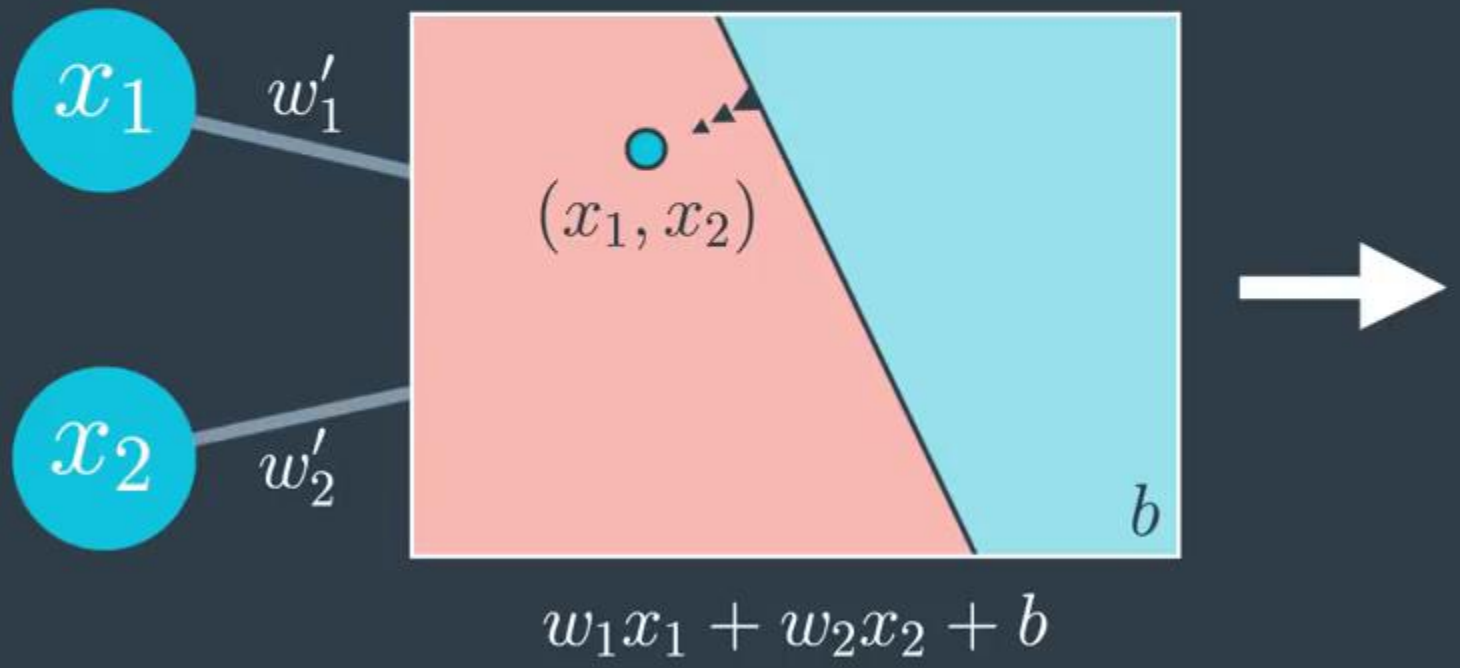
$y = 1$



Backpropagation

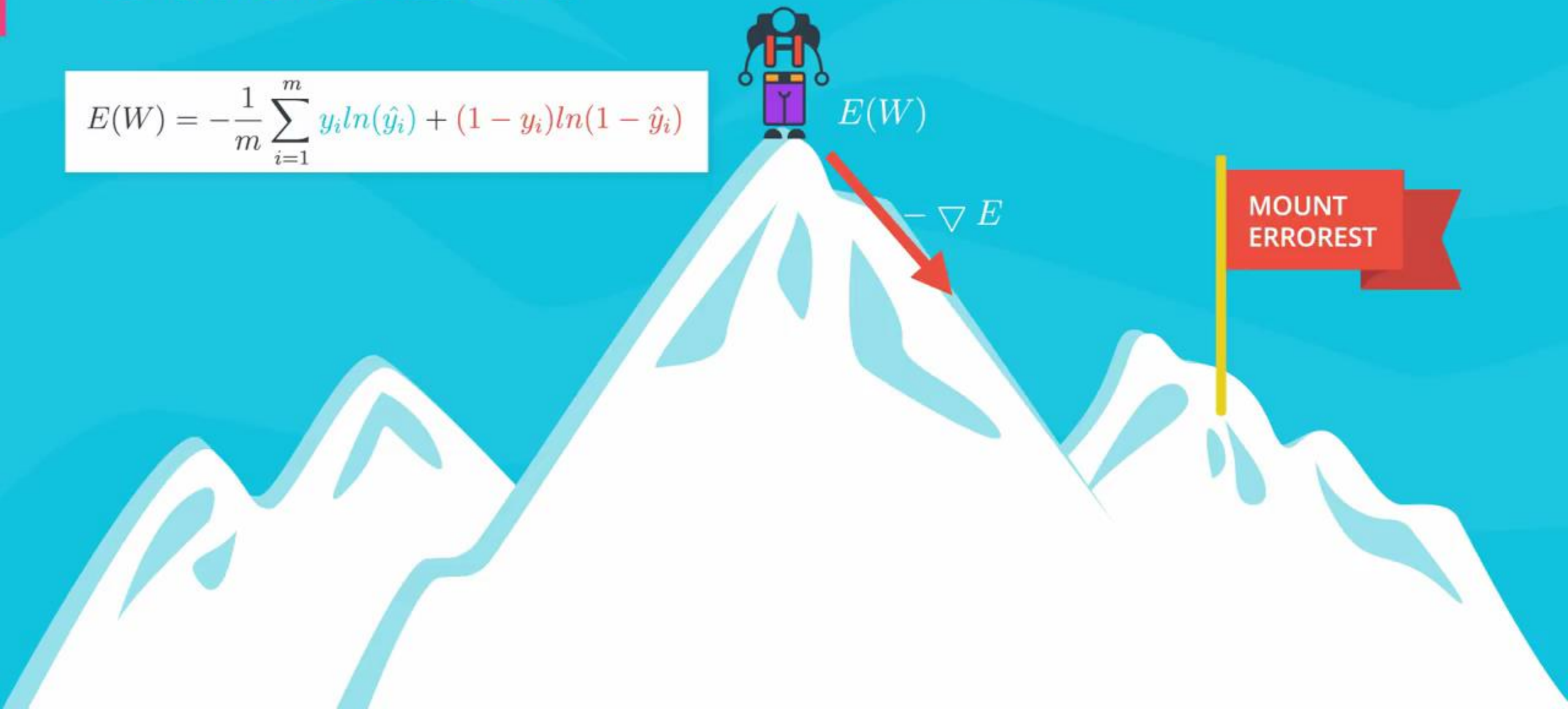
$x = (x_1, x_2)$

$y = 1$



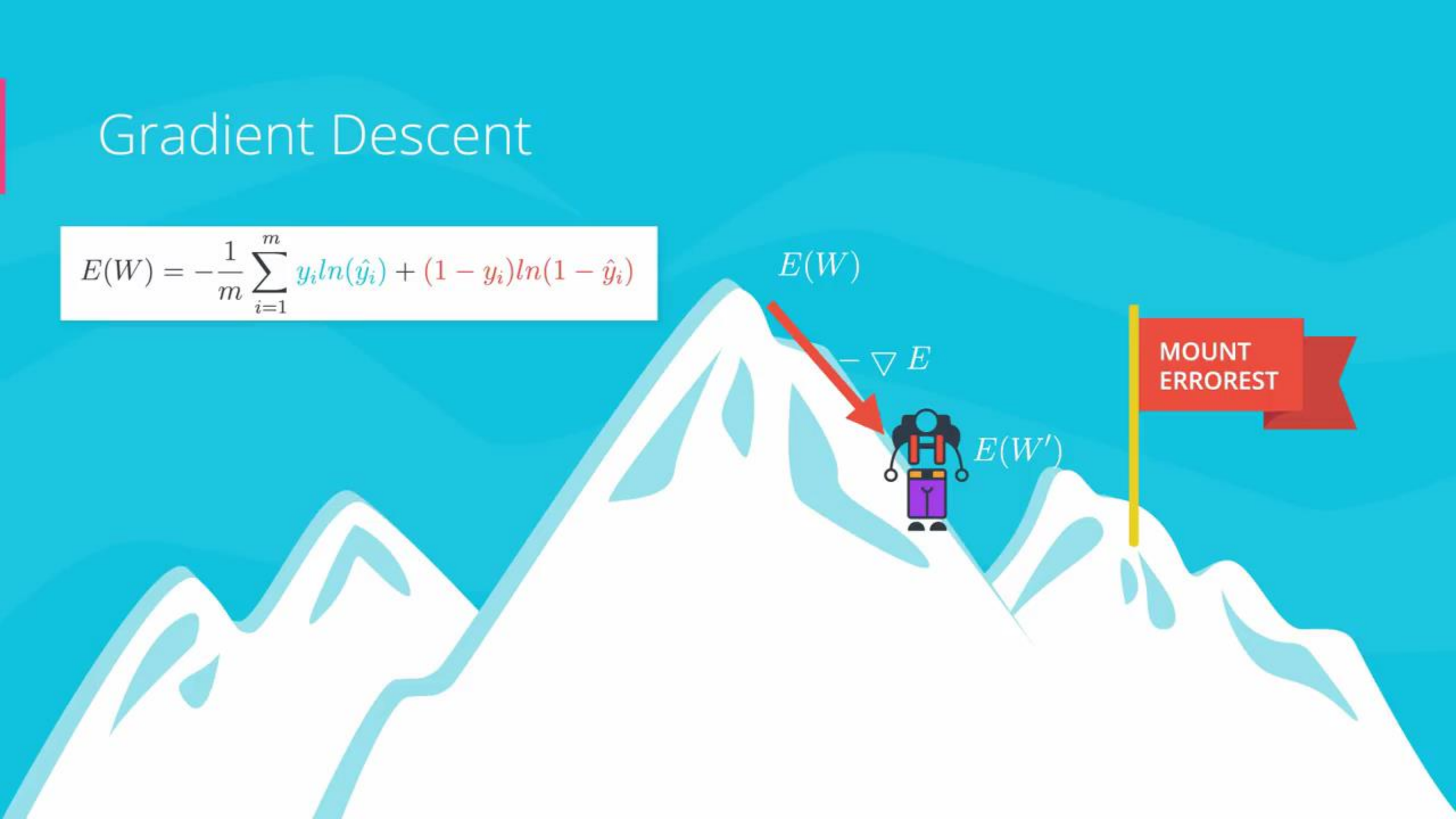
Gradient Descent

$$E(W) = -\frac{1}{m}\sum_{i=1}^{m} y_i ln(\hat{y}_i) + (1-y_i) ln(1-\hat{y}_i)$$



Gradient Descent

$$E(W) = -\frac{1}{m}\sum_{i=1}^{m} y_i ln(\hat{y}_i) + (1 - y_i) ln(1 - \hat{y}_i)$$

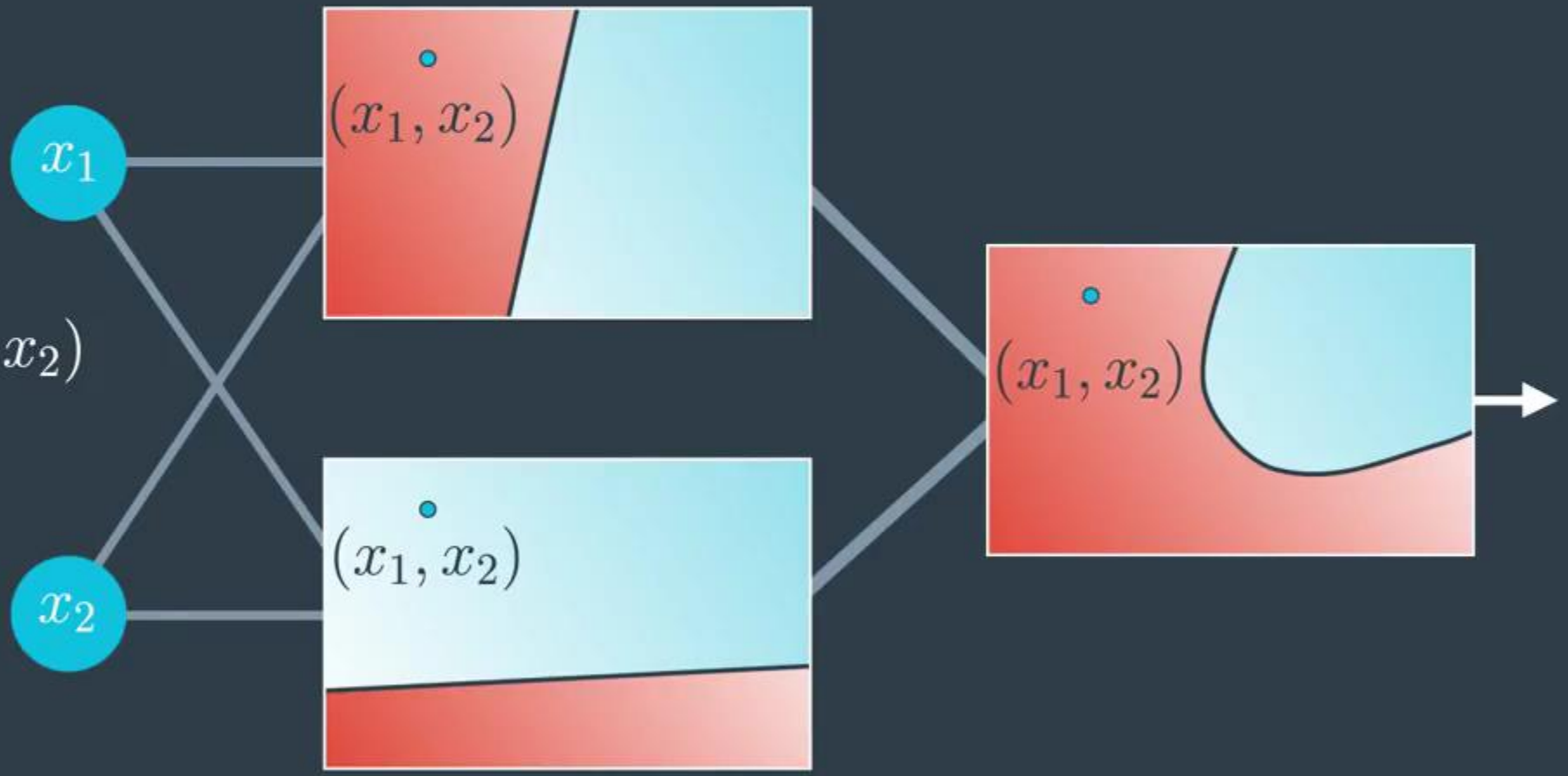


What about Multi-Layer Perceptrons?

FeedForward

$x = (x_1, x_2)$

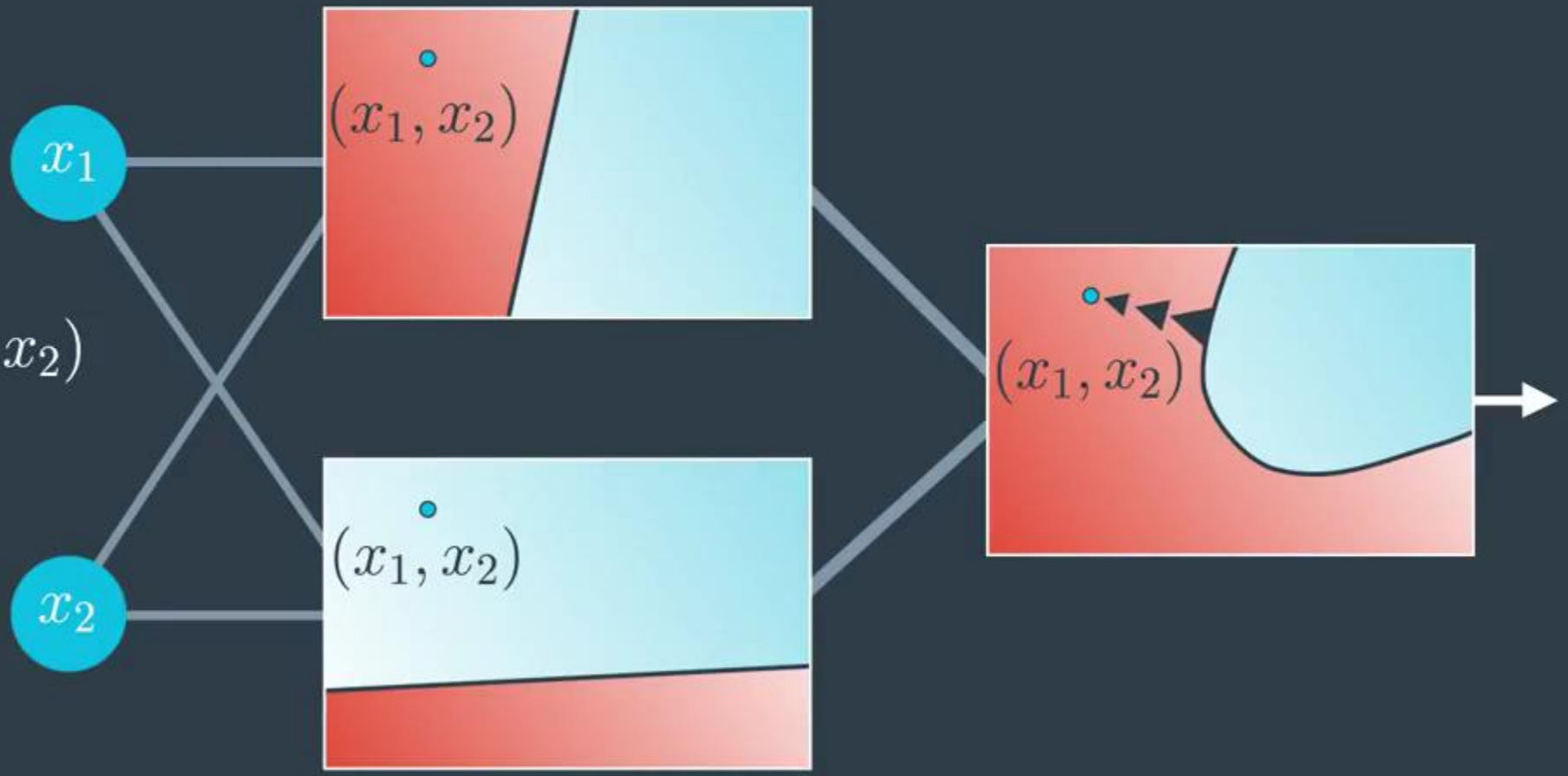
$y = 1$



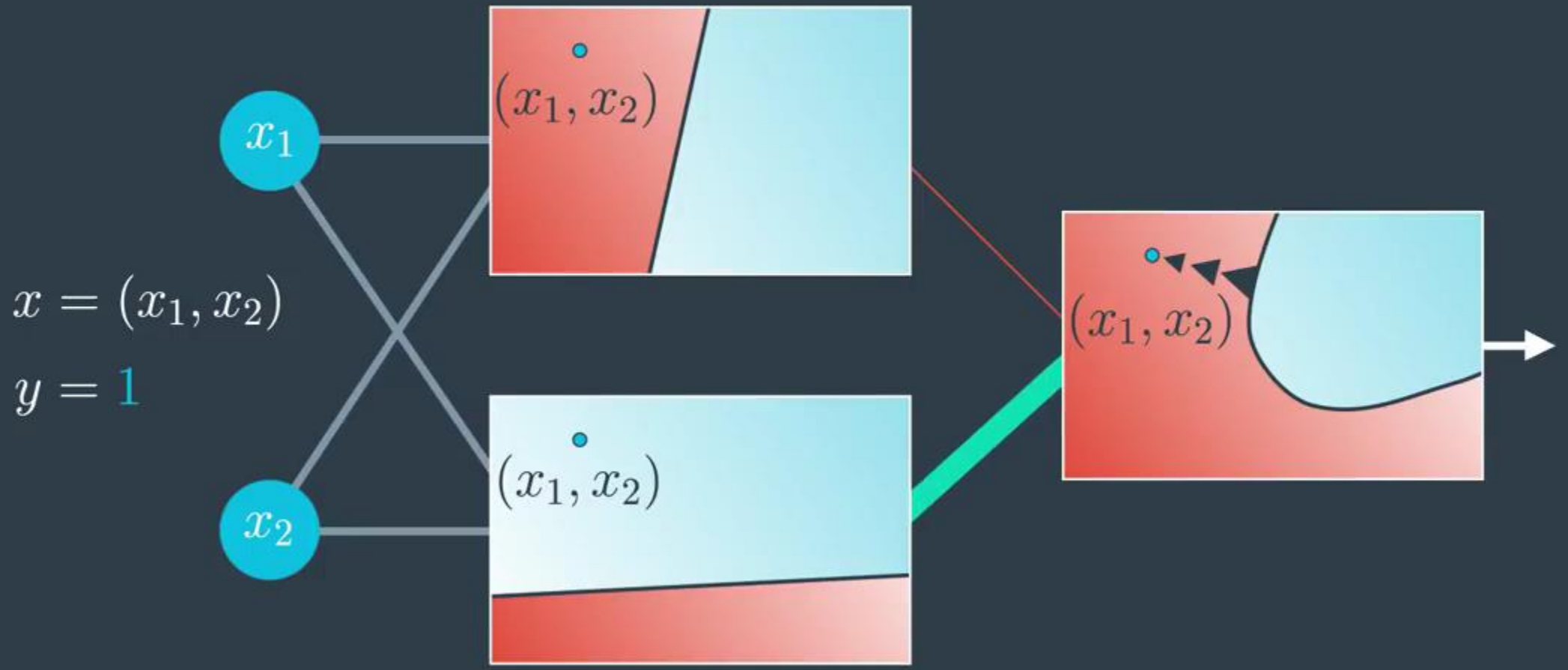
Backpropagation

$x = (x_1, x_2)$

$y = 1$



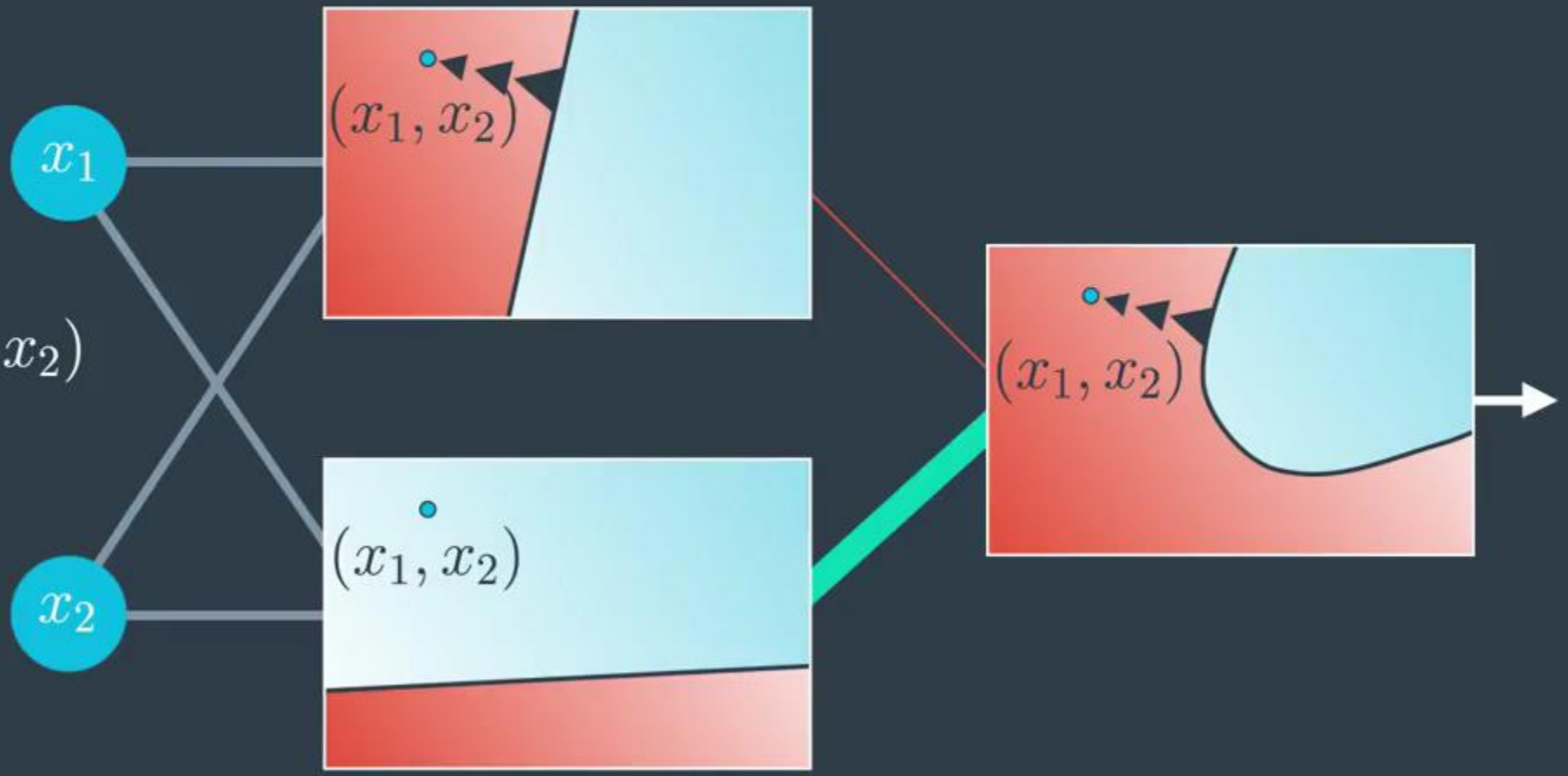
Backpropagation



Backpropagation

$x = (x_1, x_2)$

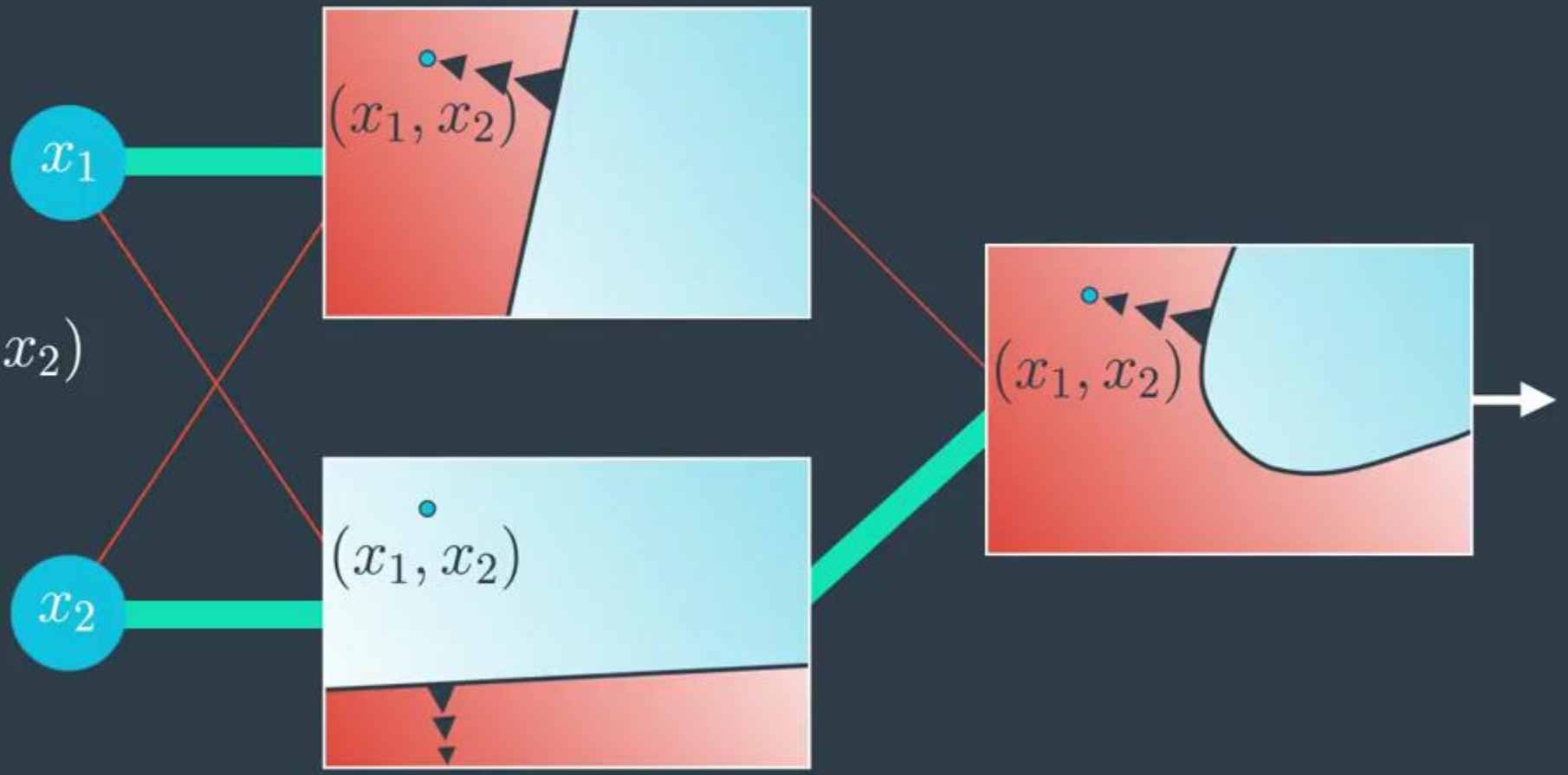
$y = 1$



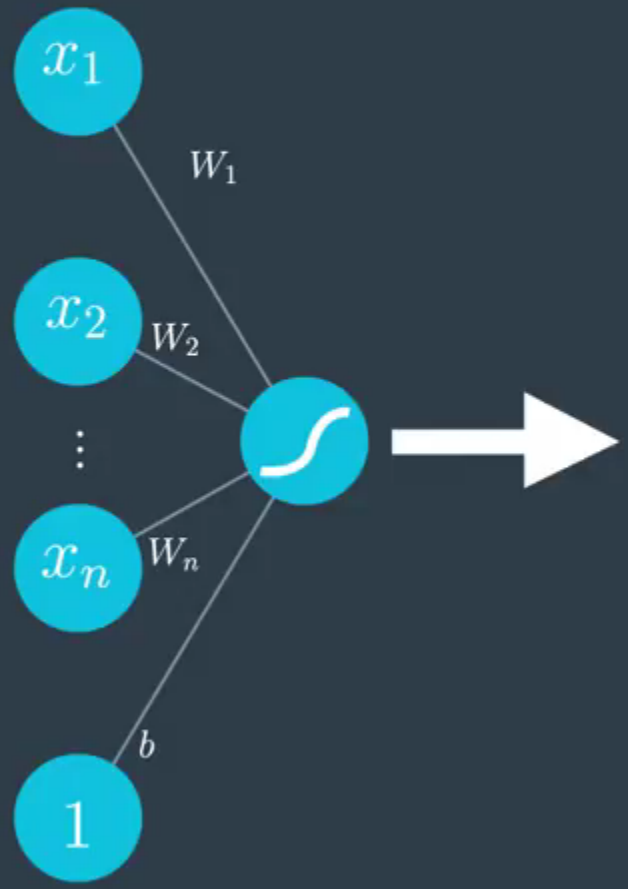
Backpropagation

$x = (x_1, x_2)$

$y = 1$



Perceptron



PREDICTION

$$\hat{y} = \sigma(Wx + b)$$

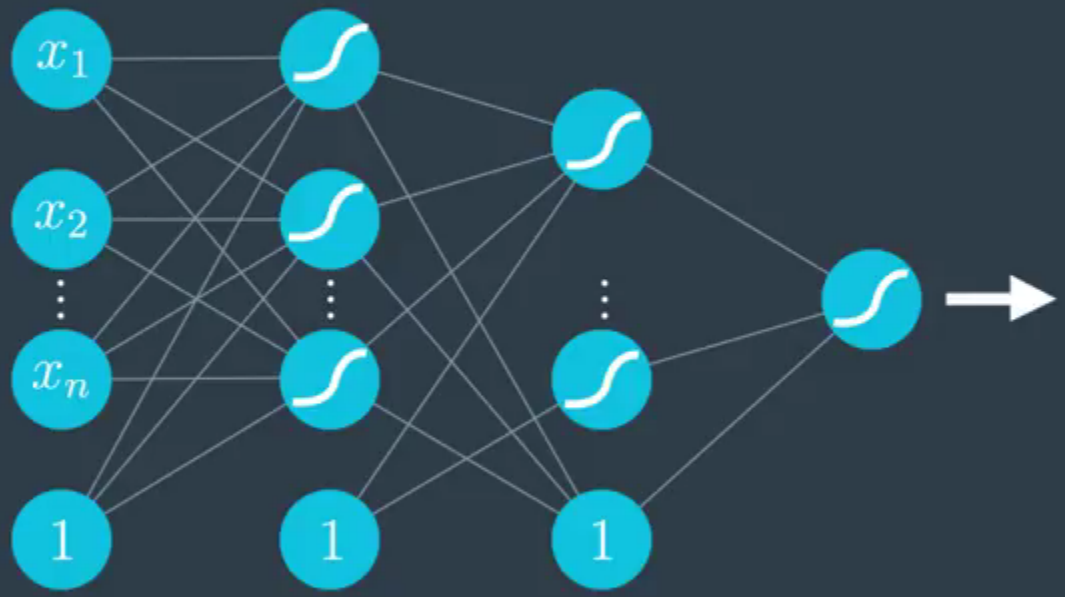
ERROR FUNCTION

$$E(W) = -\frac{1}{m}\sum_{i=1}^{m} y_i ln(\hat{y}_i) + (1 - y_i) ln(1 - \hat{y}_i)$$

GRADIENT OF THE ERROR FUNCTION

$$\nabla E = (\frac{\partial E}{\partial w_1}, ..., \frac{\partial E}{\partial w_n}, \frac{\partial E}{\partial b})$$

Multi-layer Perceptron



PREDICTION

$$\hat{y} = \sigma W^{(3)} \circ \sigma W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

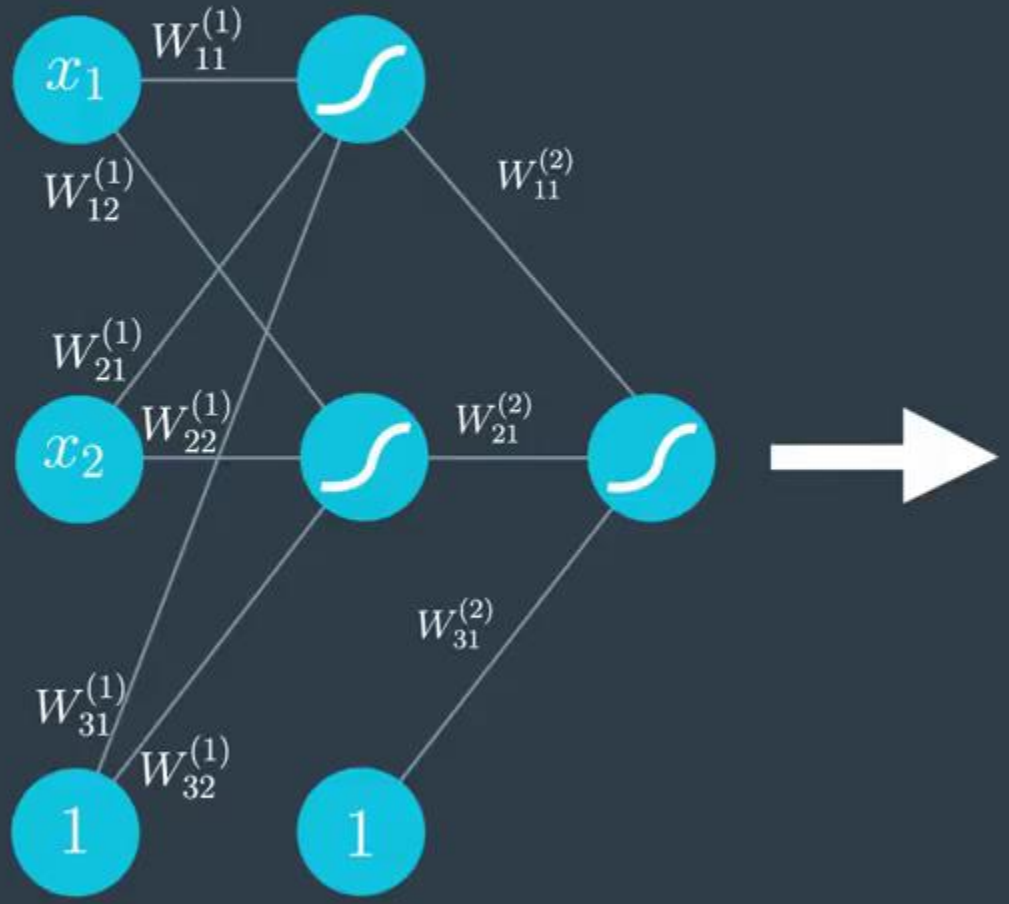
ERROR FUNCTION

$$E(W) = -\frac{1}{m} \sum_{i=1}^{m} y_i ln(\hat{y}_i) + (1 - y_i) ln(1 - \hat{y}_i)$$

GRADIENT OF THE ERROR FUNCTION

$$\bigtriangledown E = (..., \frac{\partial E}{\partial w_j^{(i)}}, ...)$$

Backpropagation



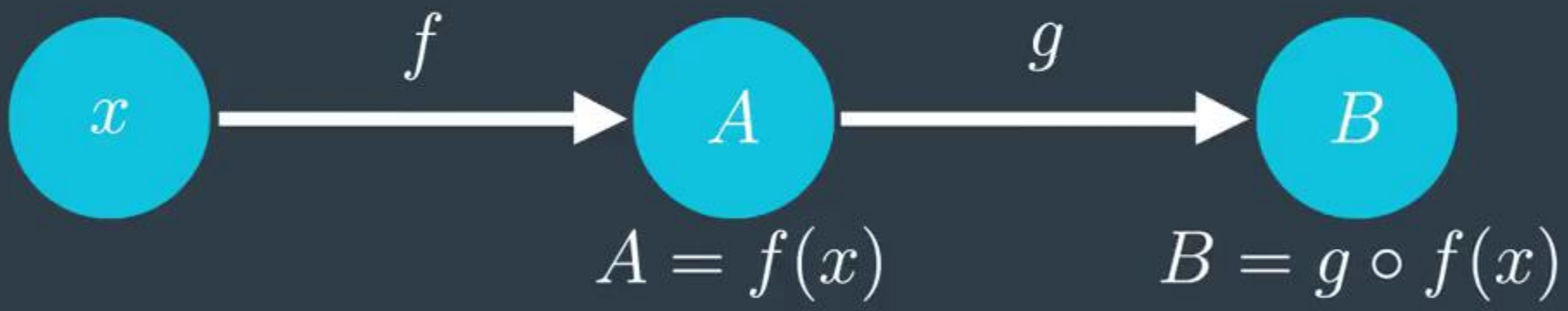
$$\hat{y} = \sigma W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

$$W^{(1)} = \begin{pmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{22}^{(1)} \\ W_{31}^{(1)} & W_{32}^{(1)} \end{pmatrix} \quad W^{(2)} = \begin{pmatrix} W_{11}^{(2)} \\ W_{21}^{(2)} \\ W_{31}^{(2)} \end{pmatrix}$$

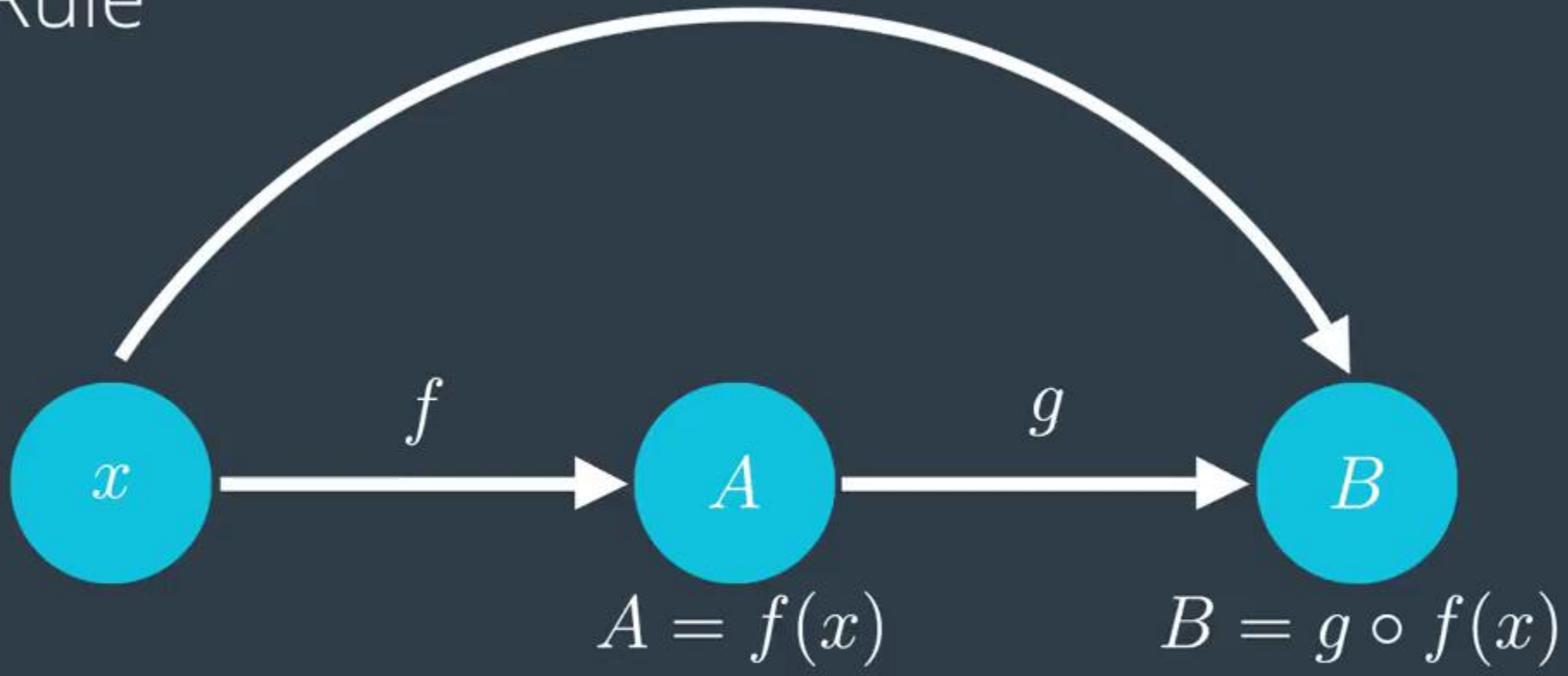
$$\nabla E = \begin{pmatrix} \frac{\partial E}{\partial W_{11}^{(1)}} & \frac{\partial E}{\partial W_{12}^{(1)}} & \frac{\partial E}{\partial W_{11}^{(2)}} \\ \frac{\partial E}{\partial W_{21}^{(1)}} & \frac{\partial E}{\partial W_{22}^{(1)}} & \frac{\partial E}{\partial W_{21}^{(2)}} \\ \frac{\partial E}{\partial W_{31}^{(1)}} & \frac{\partial E}{\partial W_{32}^{(1)}} & \frac{\partial E}{\partial W_{31}^{(2)}} \end{pmatrix}$$

$$W_{ij}^{\prime(k)} \leftarrow W_{ij}^{(k)} - \alpha \frac{\partial E}{\partial W_{ij}^{(k)}}$$

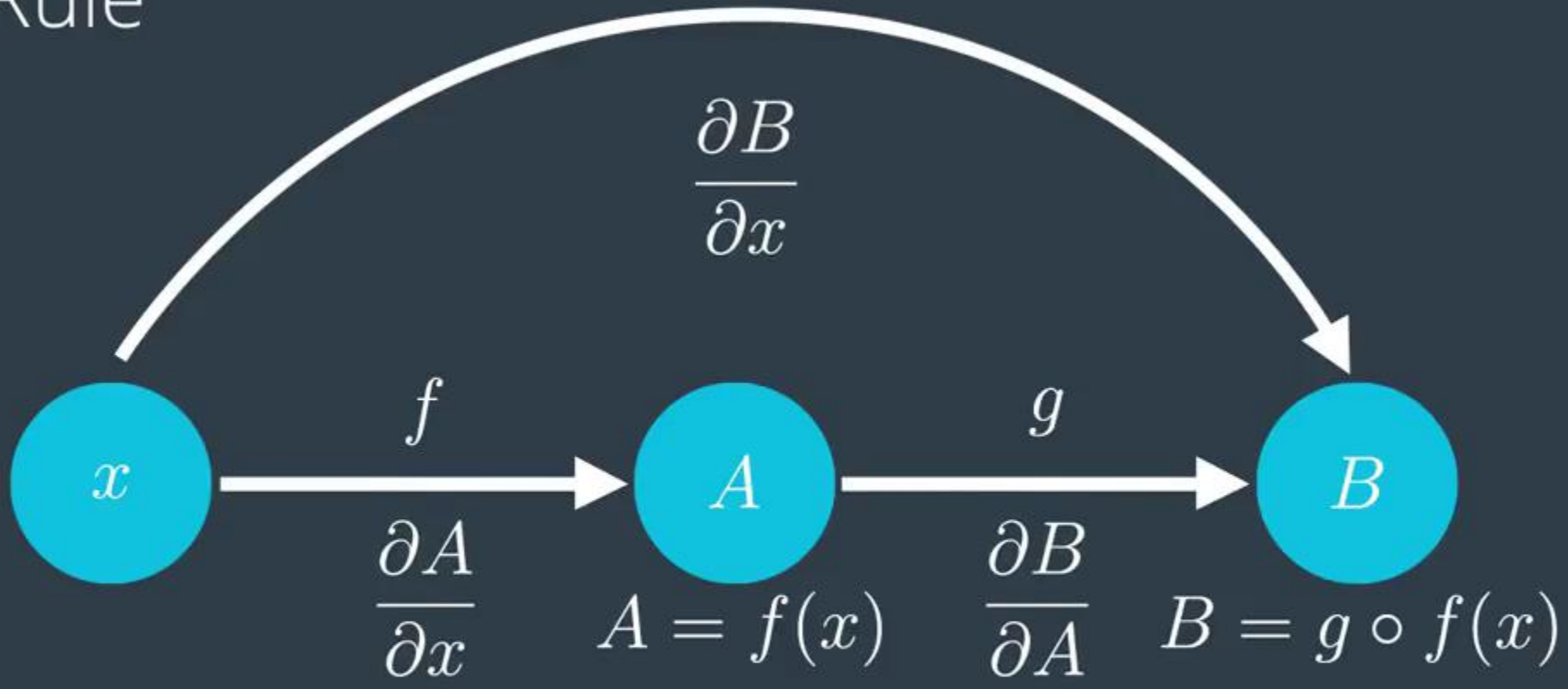
Chain Rule



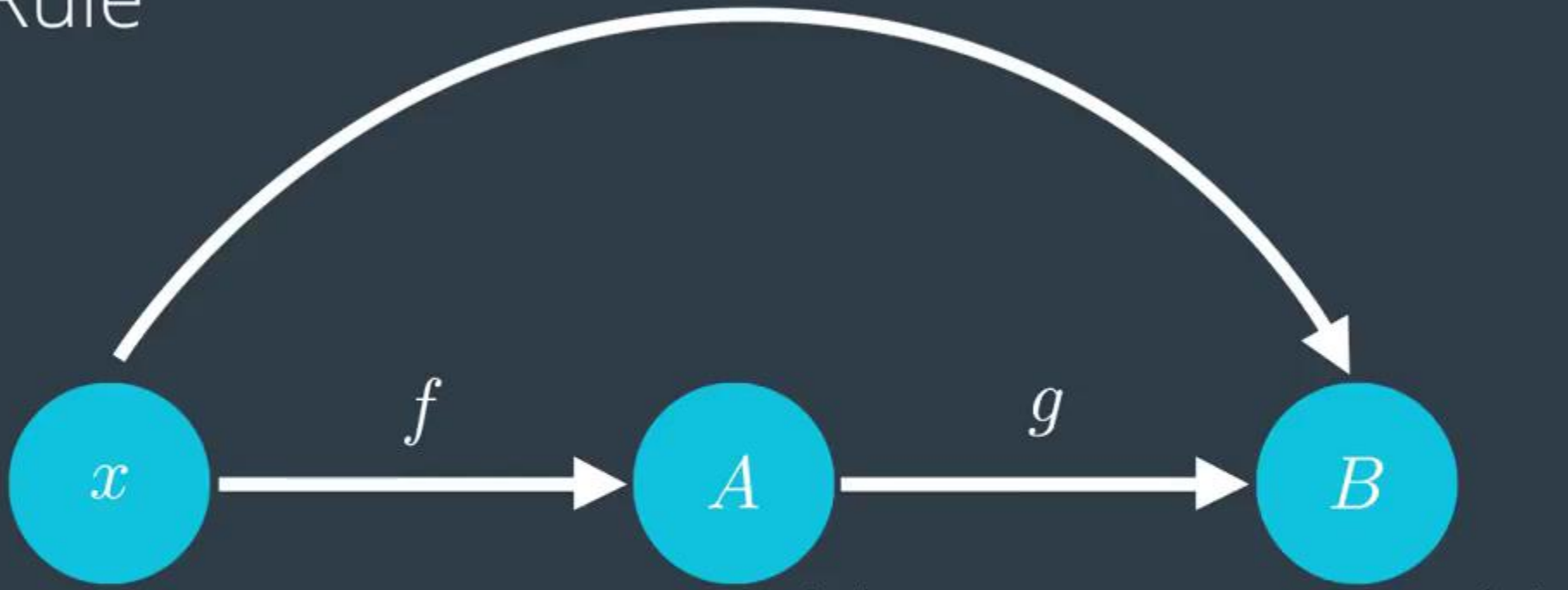
Chain Rule



Chain Rule

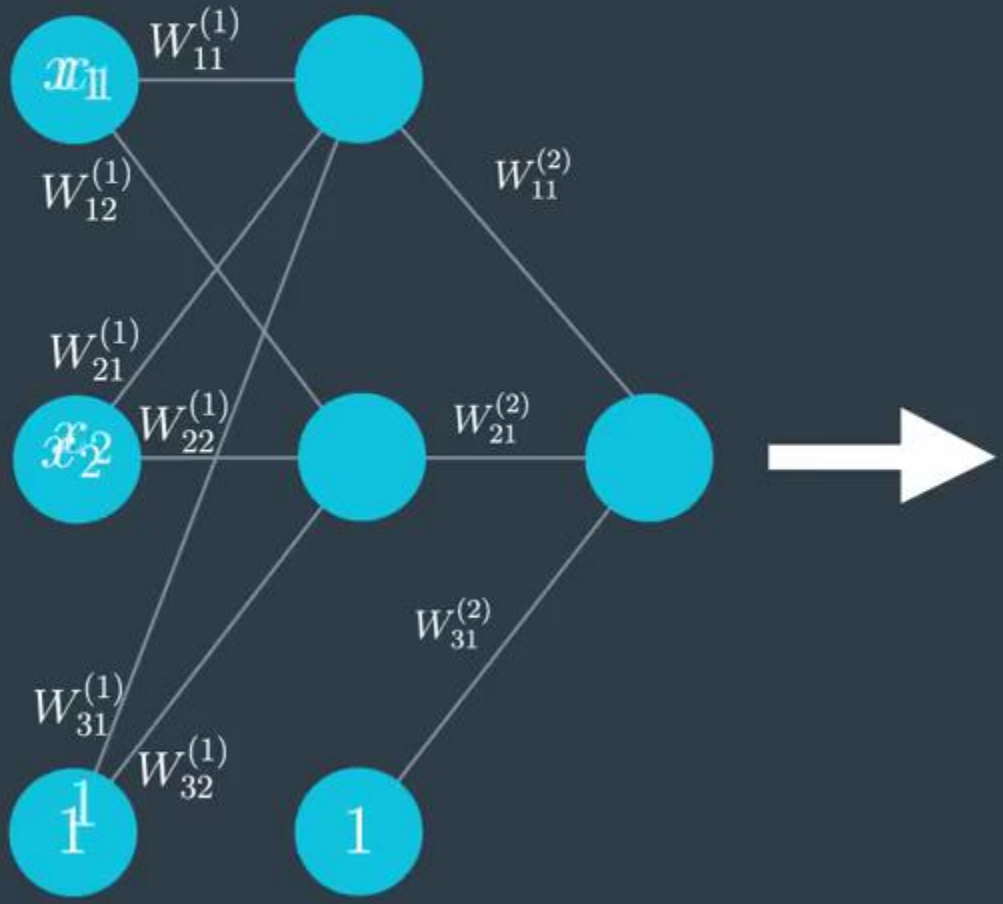

$\frac{\partial B}{\partial x}$
x
f
A
g
B
$\frac{\partial A}{\partial x}$
$A = f(x)$
$\frac{\partial B}{\partial A}$
$B = g \circ f(x)$

Chain Rule



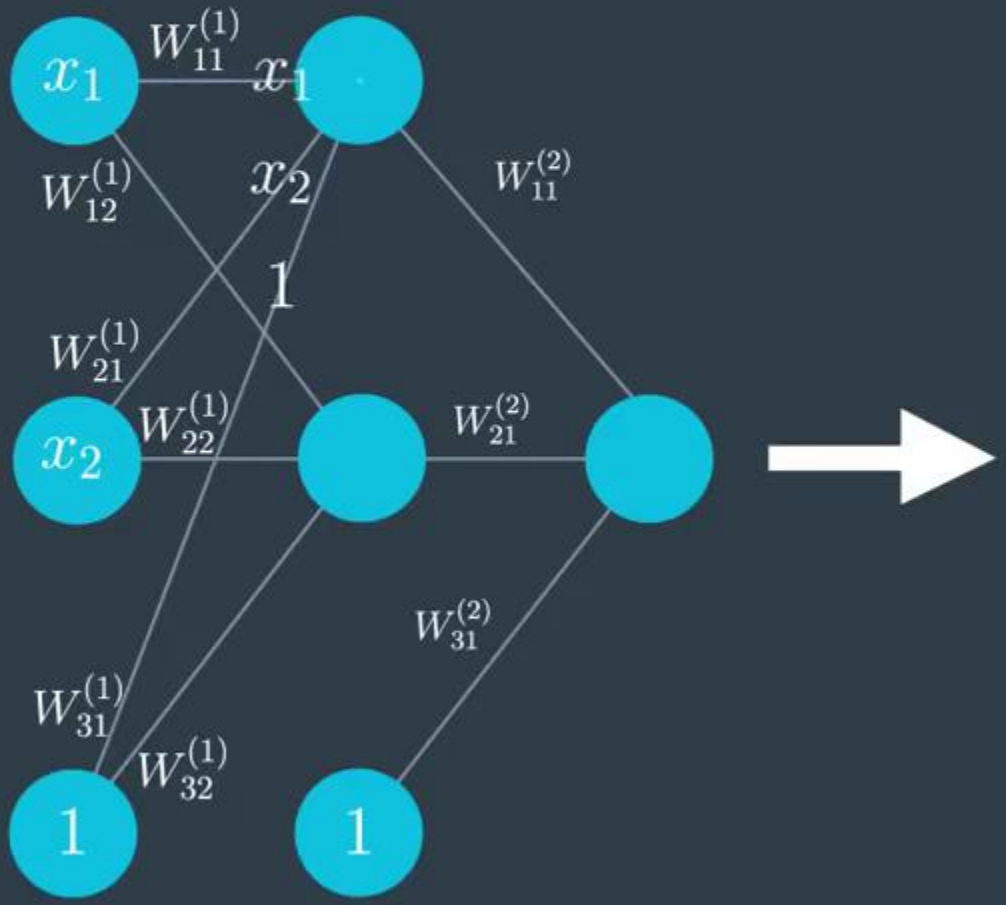
$$\frac{\partial B}{\partial x} = \frac{\partial B}{\partial A}\frac{\partial A}{\partial x}$$

Feedforward



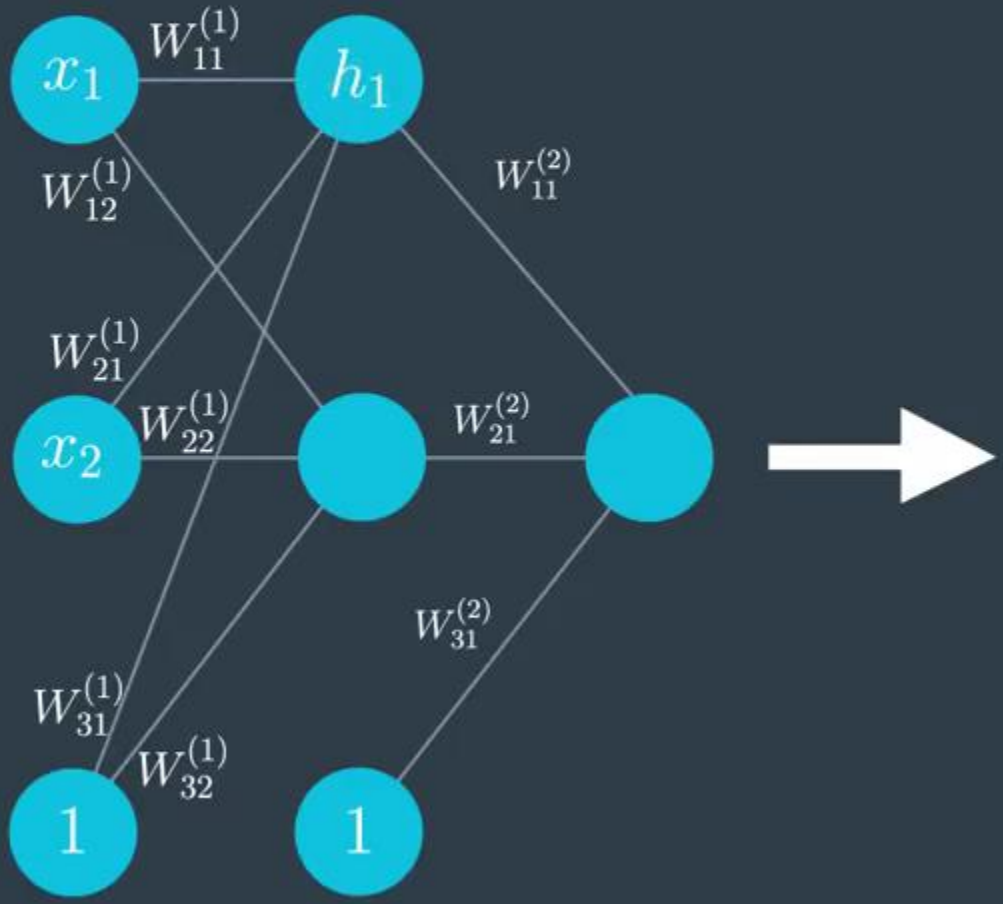
$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

Feedforward



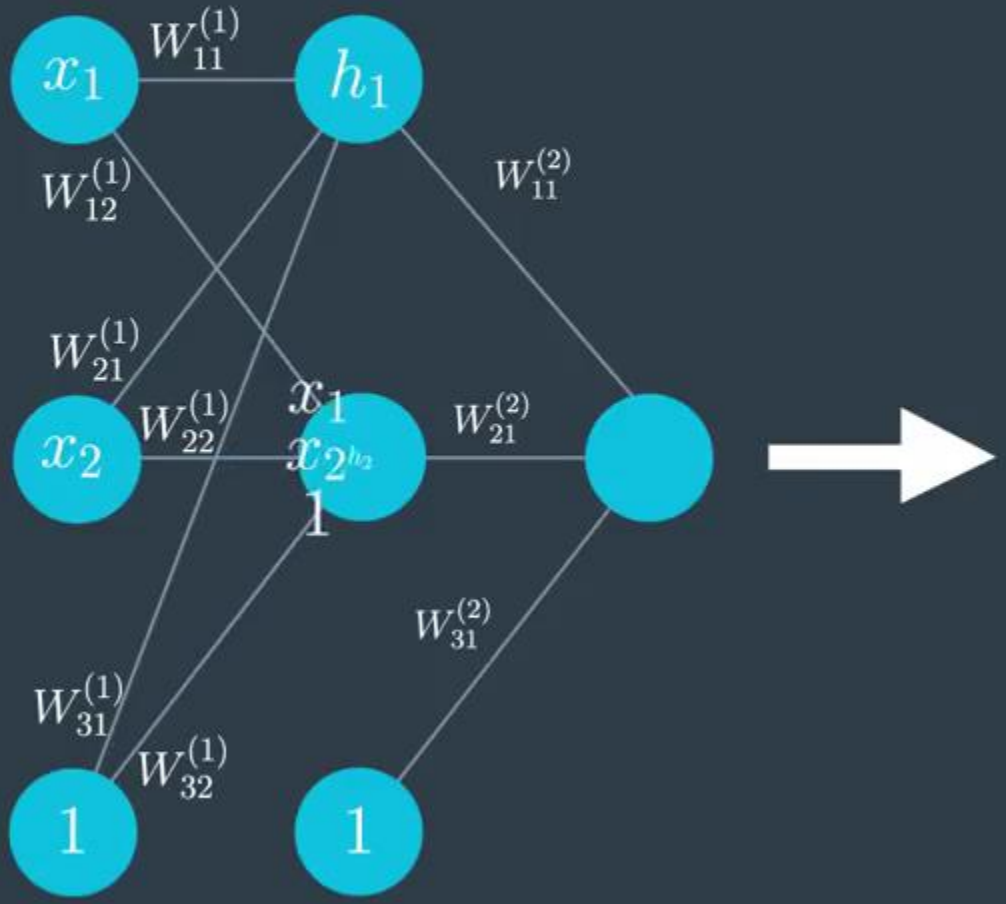
$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

Feedforward



$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

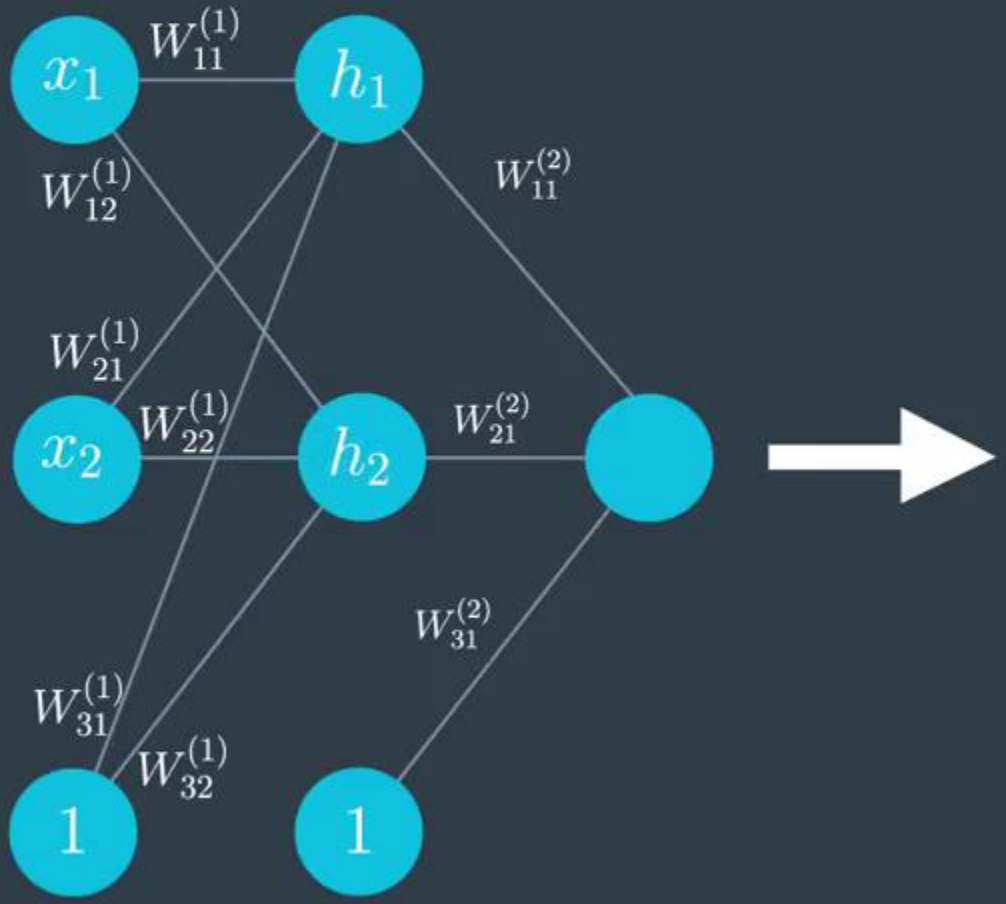
Feedforward



$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

$$h_2 = W_{12}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{32}^{(1)}$$

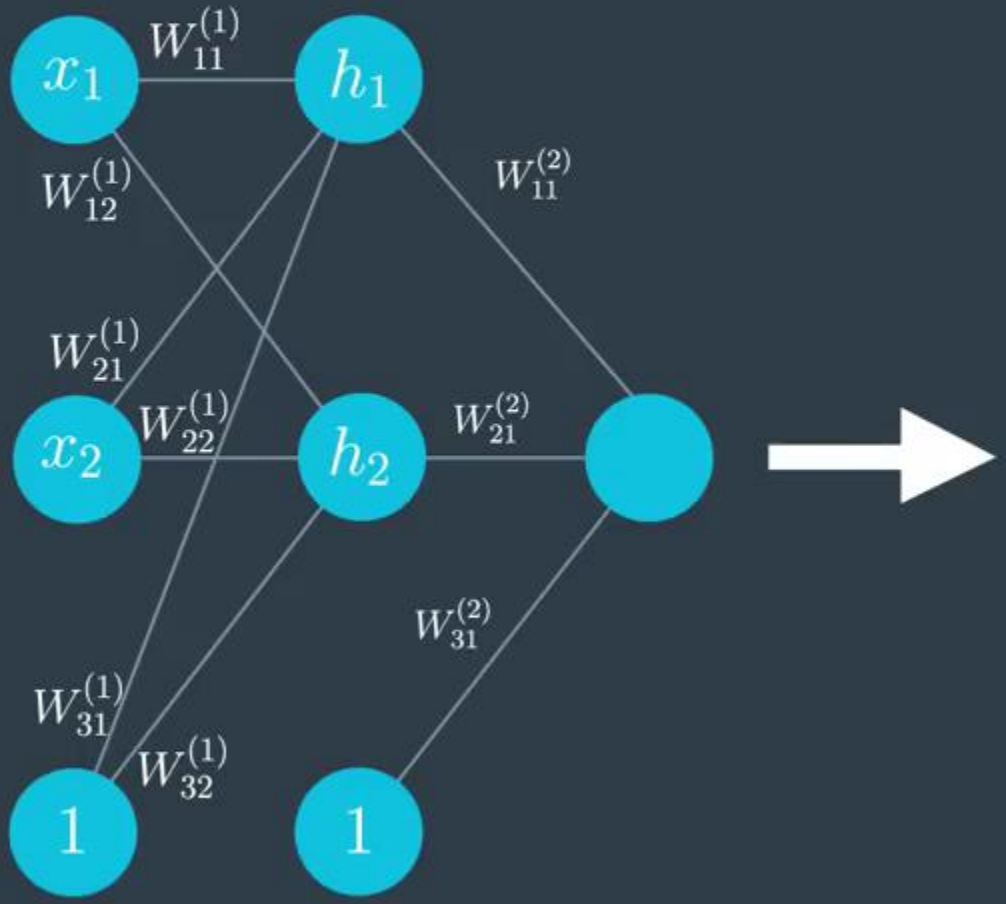
Feedforward



$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

$$h_2 = W_{12}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{32}^{(1)}$$

Feedforward

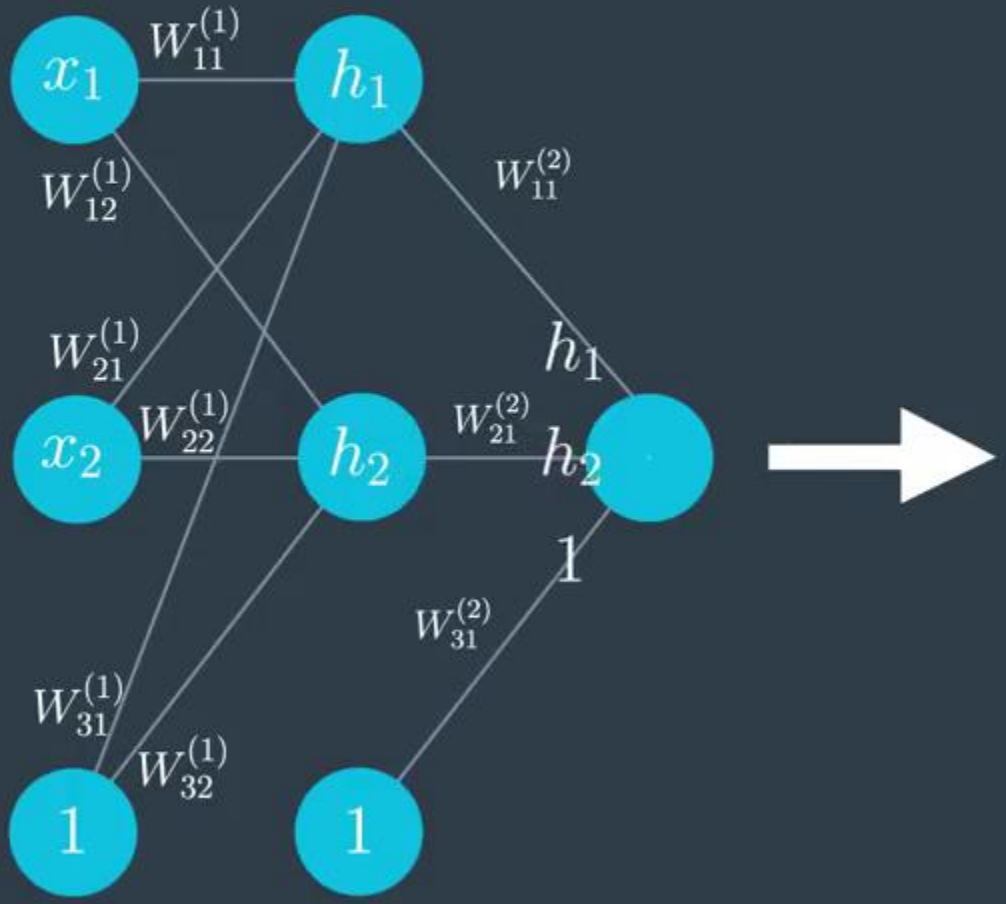


$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

$$h_2 = W_{12}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{32}^{(1)}$$

$$h = W_{11}^{(2)} \sigma(h_1) + W_{21}^{(2)} \sigma(h_2) + W_{31}^{(2)}$$

Feedforward

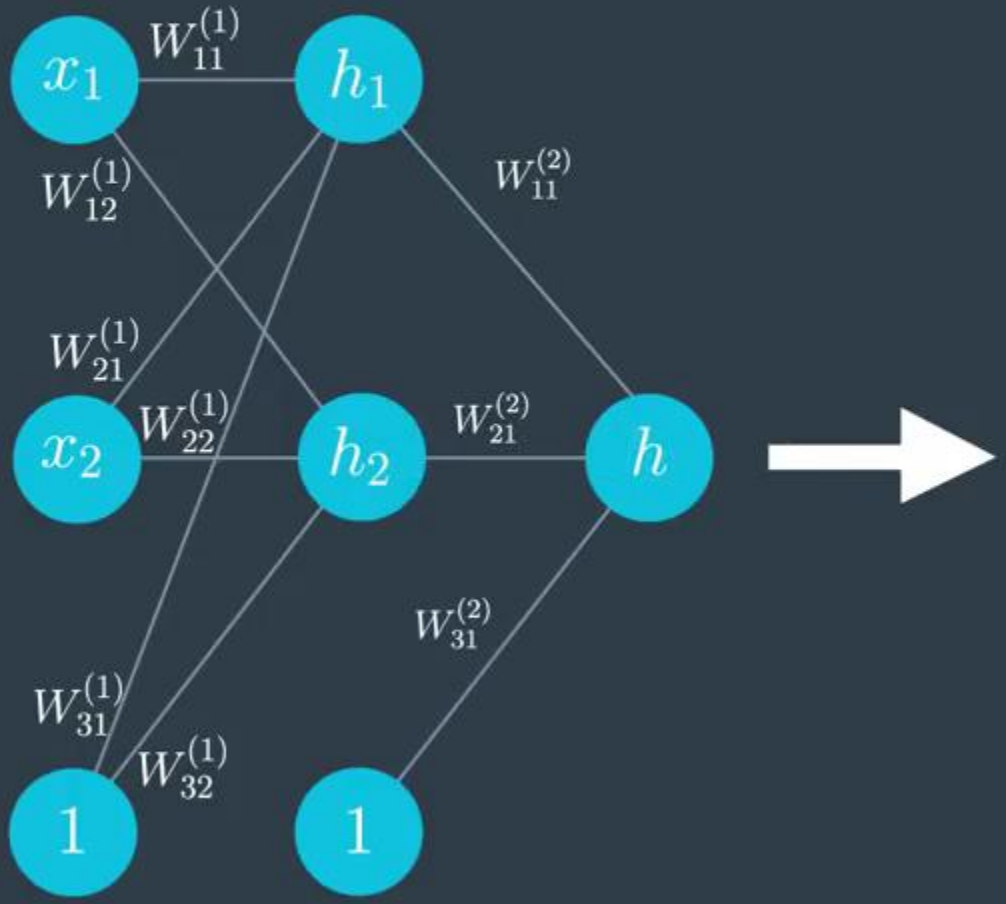


$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

$$h_2 = W_{12}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{32}^{(1)}$$

$$h = W_{11}^{(2)} \sigma(h_1) + W_{21}^{(2)} \sigma(h_2) + W_{31}^{(2)}$$

Feedforward

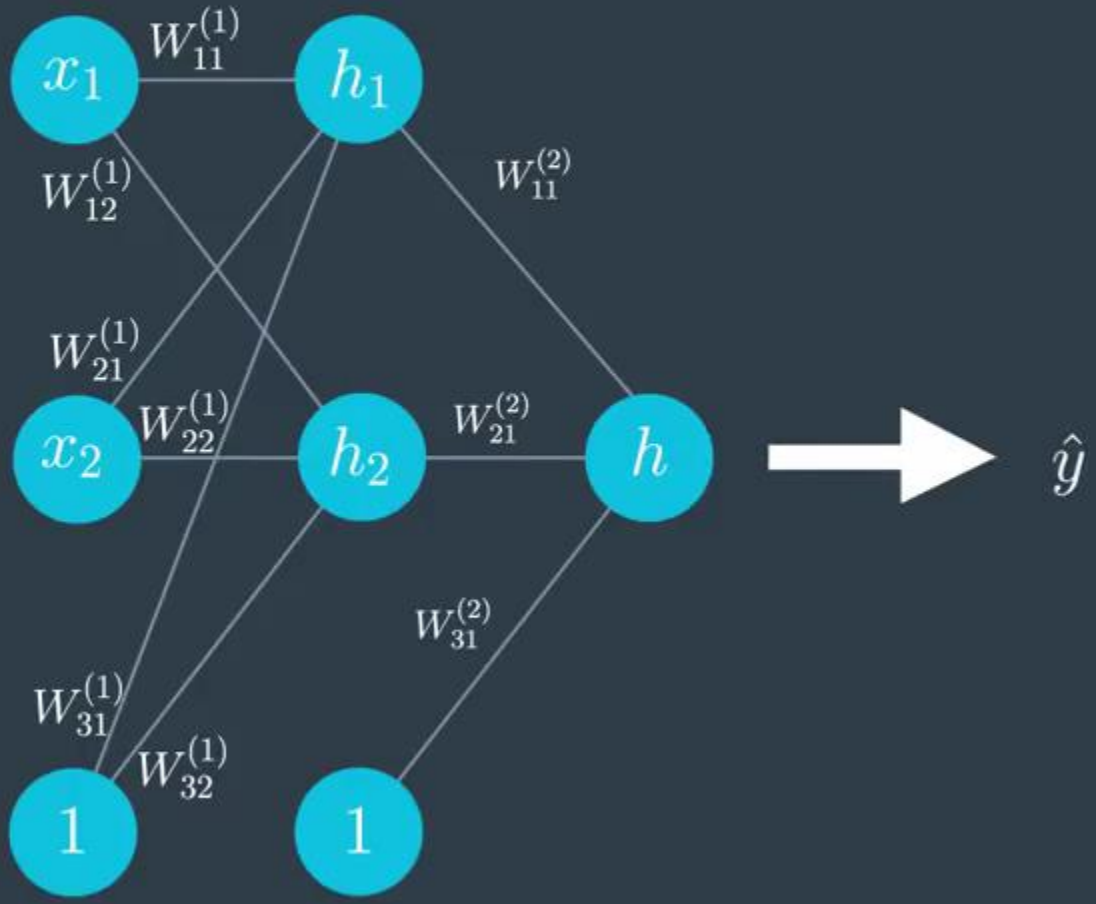


$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

$$h_2 = W_{12}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{32}^{(1)}$$

$$h = W_{11}^{(2)} \sigma(h_1) + W_{21}^{(2)} \sigma(h_2) + W_{31}^{(2)}$$

Feedforward



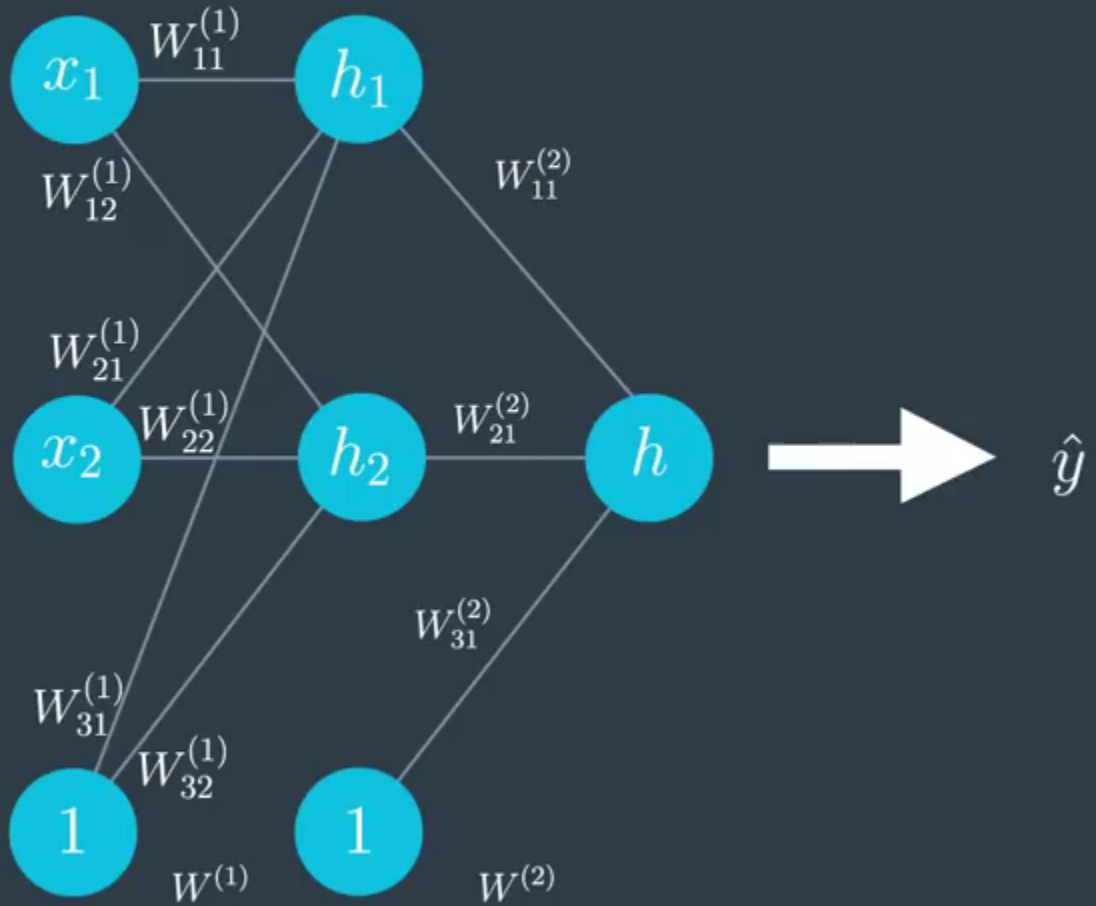
$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

$$h_2 = W_{12}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{32}^{(1)}$$

$$h = W_{11}^{(2)} \sigma(h_1) + W_{21}^{(2)} \sigma(h_2) + W_{31}^{(2)}$$

$$\hat{y} = \sigma(h)$$

Feedforward



$$h_1 = W_{11}^{(1)} x_1 + W_{21}^{(1)} x_2 + W_{31}^{(1)}$$

$$h_2 = W_{12}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{32}^{(1)}$$

$$h = W_{11}^{(2)} \sigma(h_1) + W_{21}^{(2)} \sigma(h_2) + W_{31}^{(2)}$$

$$\hat{y} = \sigma(h)$$

$$\hat{y} = \sigma \circ W^{(2)} \circ \sigma \circ W^{(1)}(x)$$

Sigmoid Derivative

$$
\begin{aligned}
\sigma'(x) &= \frac{\partial}{\partial x}\frac{1}{1+e^{-x}} \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\
&= \sigma(x)(1-\sigma(x)) \\
&= \sigma(h)(1-\sigma(h))
\end{aligned}
$$

$$\frac{dh}{dz} = h(1-h)$$

Log loss derivative

$$L(h, y) = -[\, y \log(h) + (1-y) \log(1-h) \,]$$

$$\frac{d(C \log(x))}{dx} = \frac{C}{x}$$

$$\therefore L(h, y) = -\frac{y}{h} - \frac{1-y}{1-h}$$

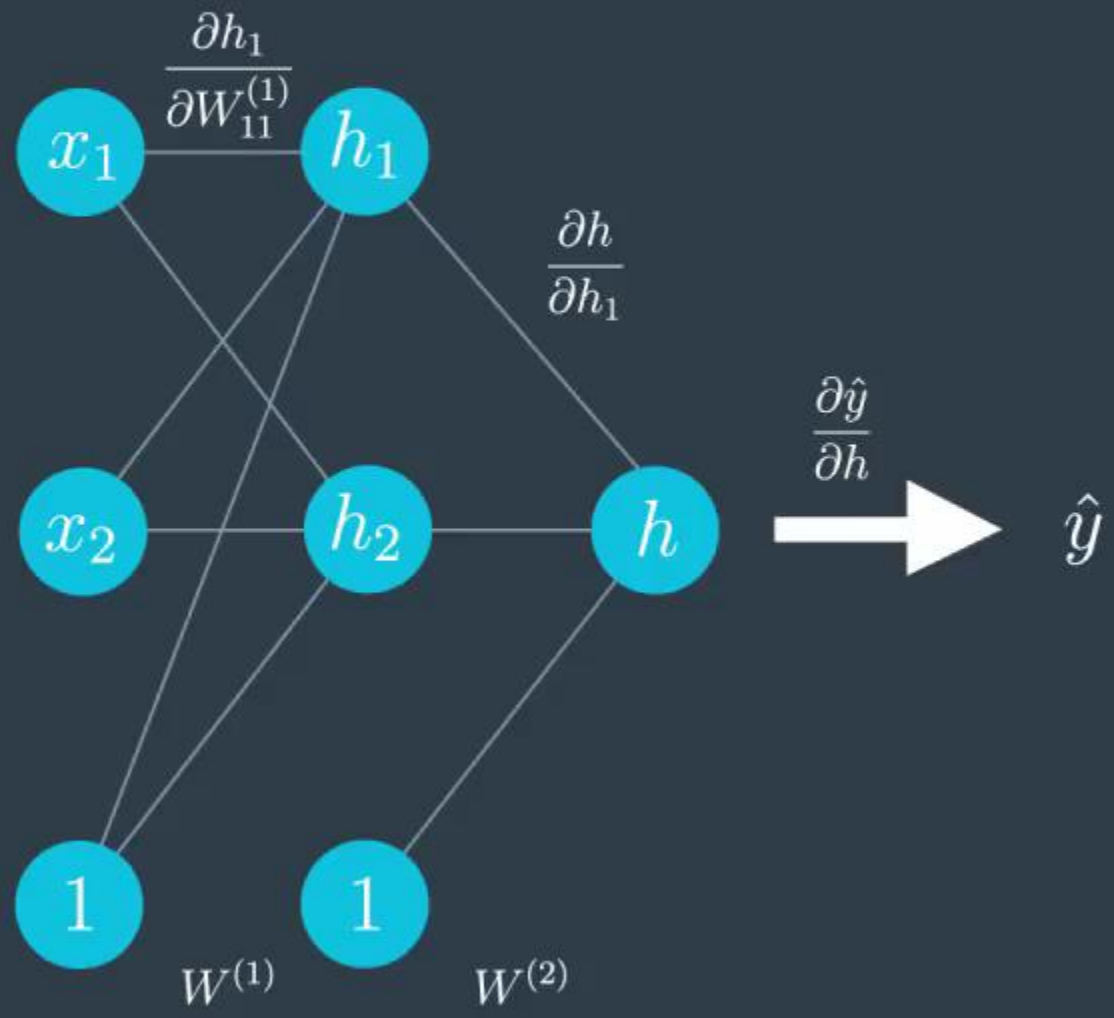
By using Chain Rule

$$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial h} \cdot \frac{\partial h}{\partial z}$$

$$\frac{\partial L}{\partial z} = \left(-\frac{y}{h} - \frac{1-y}{1-h}\right)\left(h(1-h)\right)$$

$$= -y(1-h) - (1-y)h$$

BACKPROPAGATION

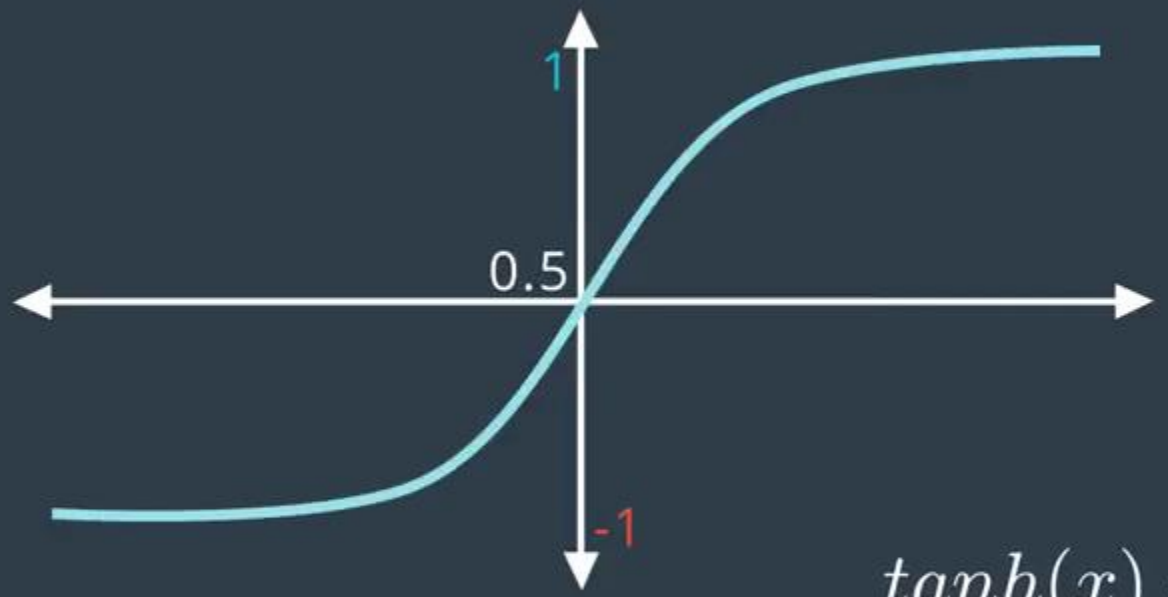


$$\frac{\partial E}{\partial W_{11}^{(1)}} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h} \frac{\partial h}{\partial h_1} \frac{\partial h_1}{\partial W_{11}^{(1)}}$$

TINY

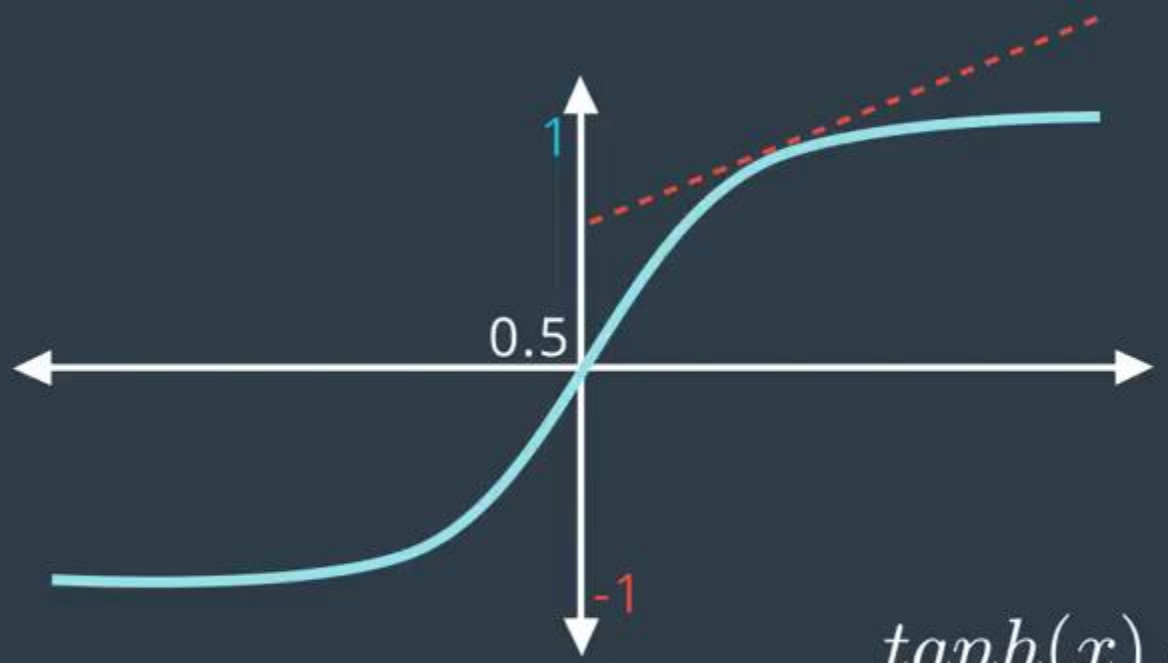
SMALL

HYPERBOLIC TANGENT FUNCTION



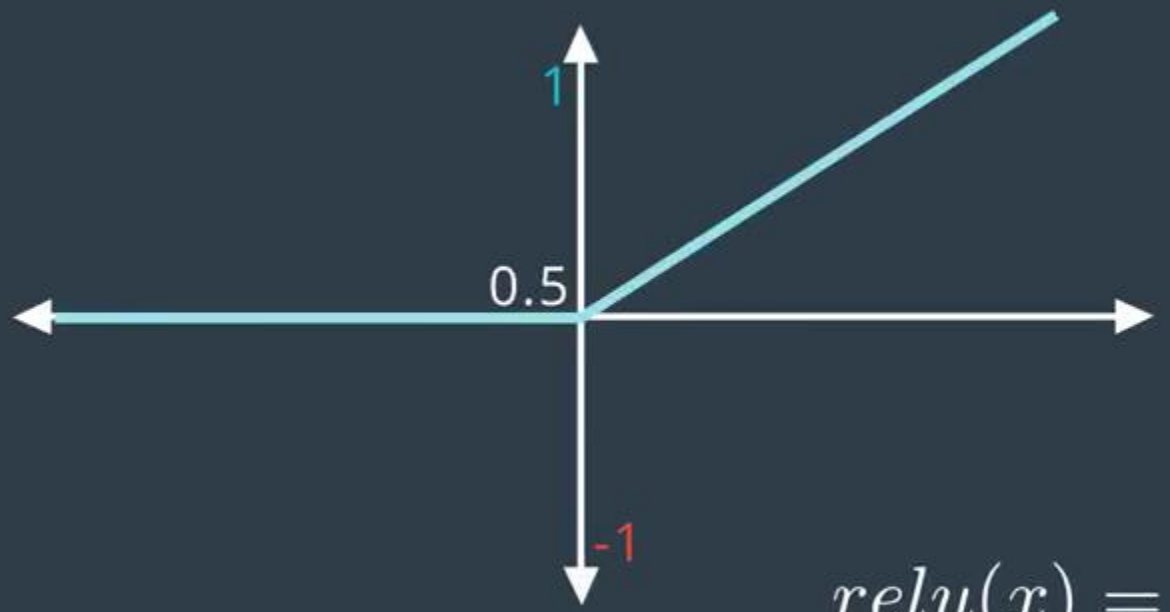
$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

HYPERBOLIC TANGENT FUNCTION



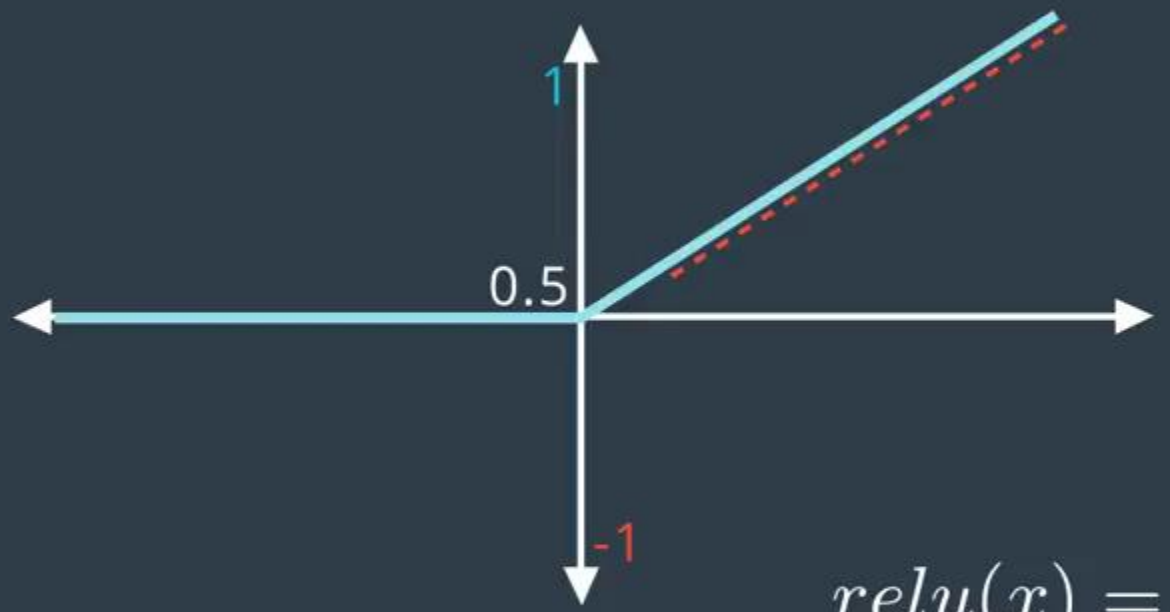
$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

RECTIFIED LINEAR UNIT (ReLU)



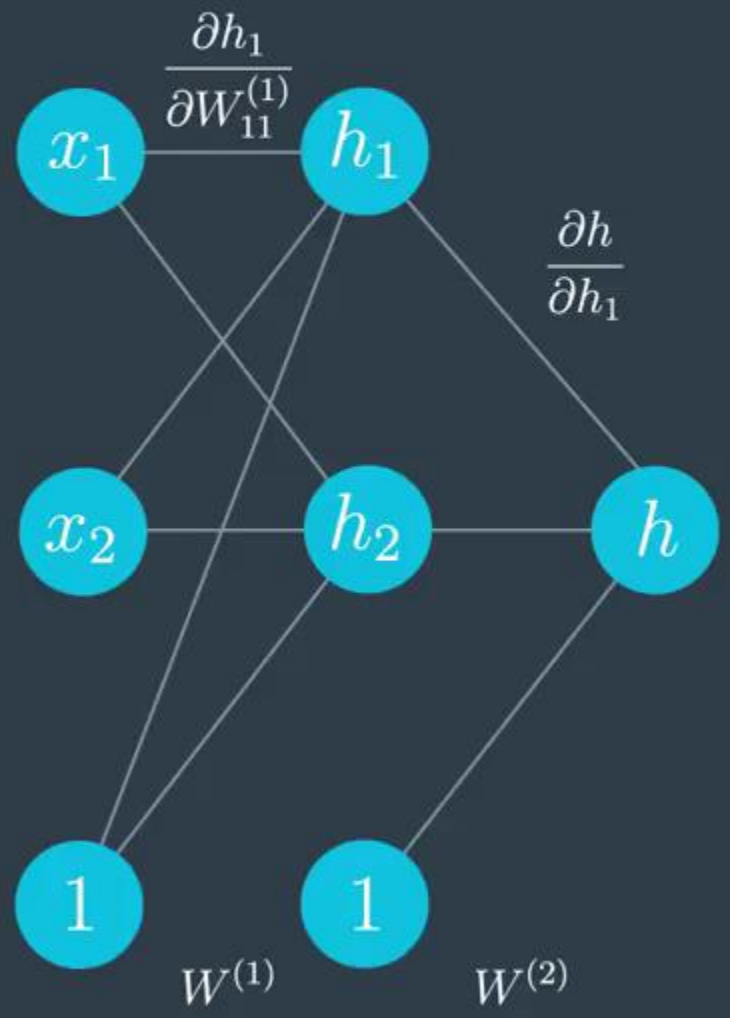
$$relu(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

RECTIFIED LINEAR UNIT (ReLU)



$$relu(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

BACKPROPAGATION



$$\boxed{\frac{\partial E}{\partial W_{11}^{(1)}}} = \boxed{\frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h} \frac{\partial h}{\partial h_1} \frac{\partial h_1}{\partial W_{11}^{(1)}}}$$

OK!

OK