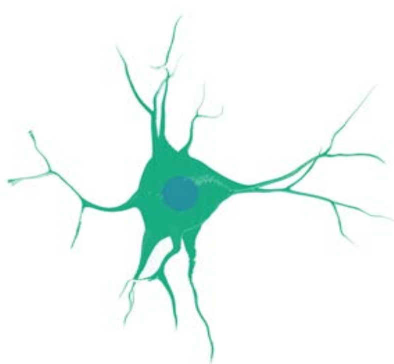
# Neural Networks

Cynthia Rudin

Duke Machine Learning

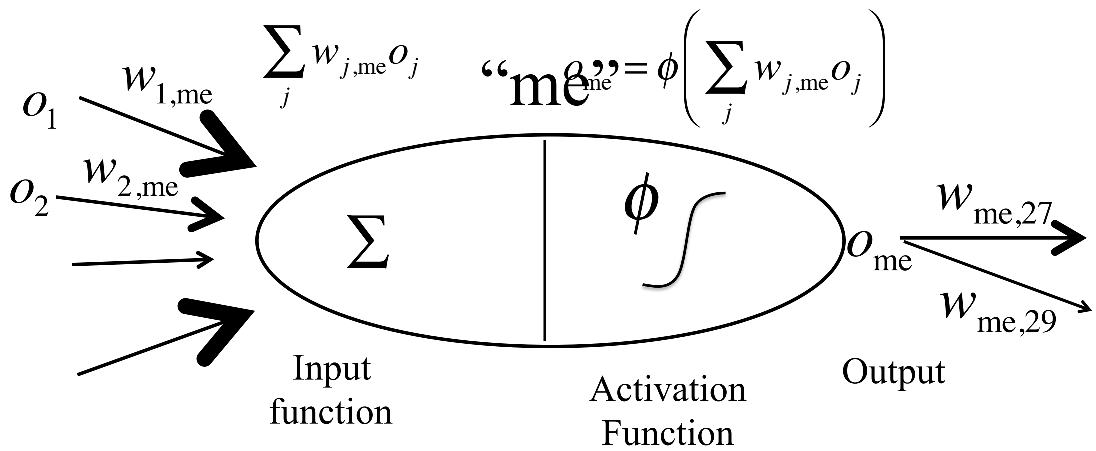
## Neurons

- $10^{11}$ neurons in a brain, $10^{14}$ synapses (connections).
- Signals are electrical potential spikes that travel through the network.

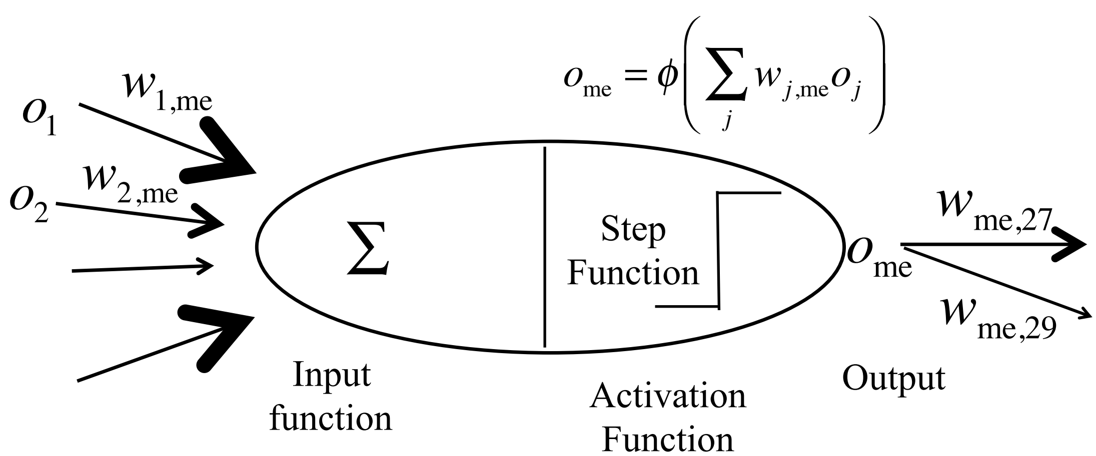


(Credit: Adapted from Russell and Norvig)

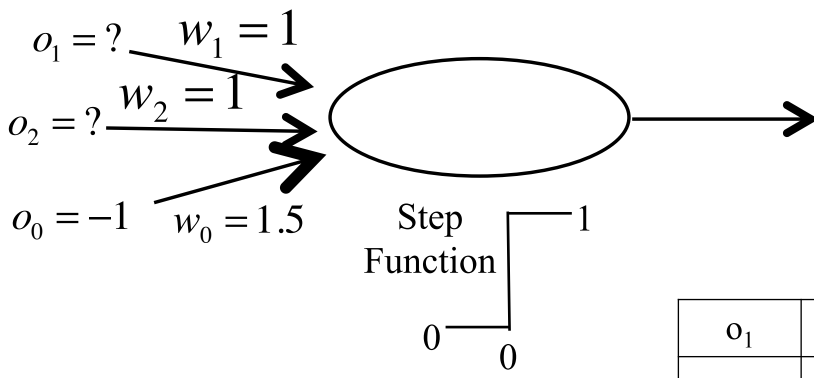
## McCulloch-Pitts "Neuron"




## McCulloch-Pitts "Neuron"

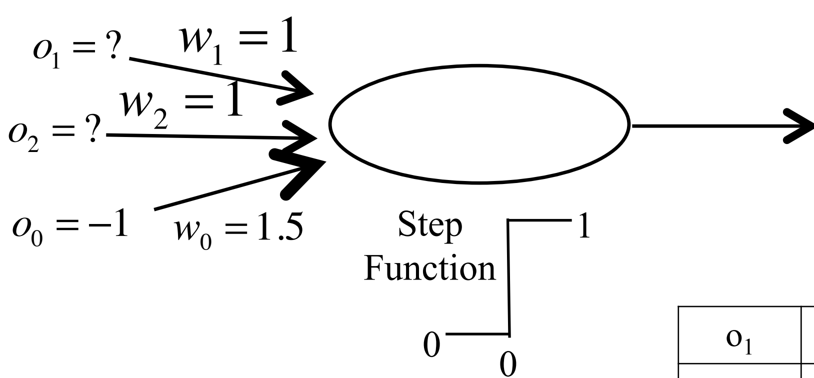


# McCulloch-Pitts “Neuron”




| $o_1$ | $o_2$ | output |
|---|---|---|
| | | |
| | | |
| | | |
| | | |

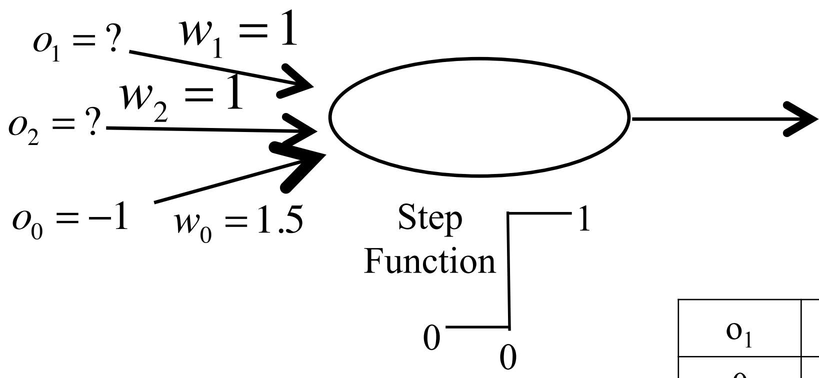
# McCulloch-Pitts “Neuron”




| $o_1$ | $o_2$ | output |
|---|---|---|
| 0 | 0 | |
| | | |
| | | |
| | | |

0+0-1.5 = -1.5

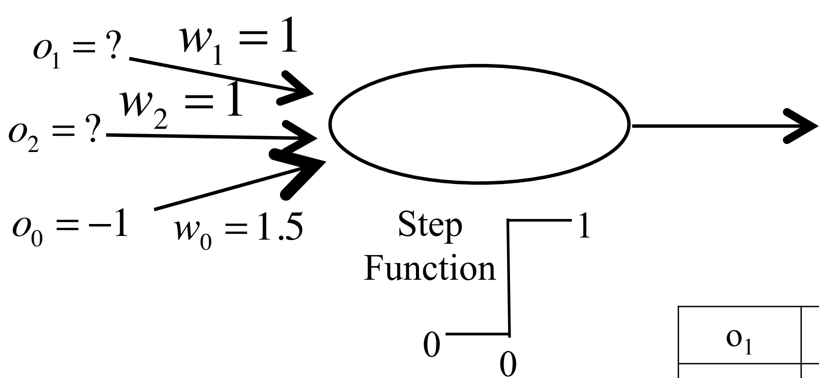
## McCulloch-Pitts "Neuron"




| $o_1$ | $o_2$ | output |
|---|---|---|
| 0 | 0 | 0 |
| | | |
| | | |
| | | |

0+0-1.5 = -1.5

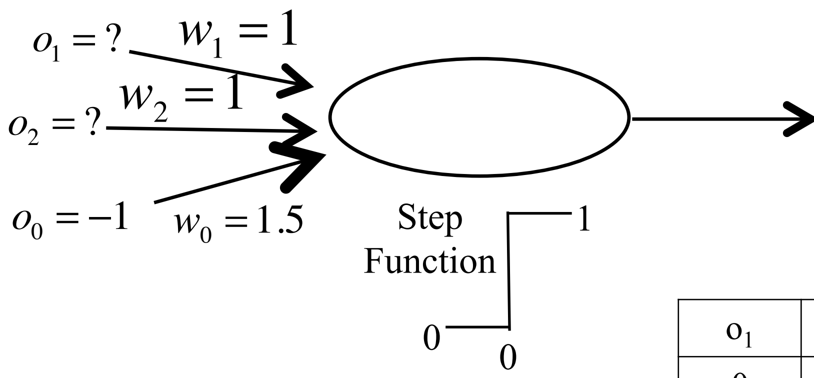
## McCulloch-Pitts "Neuron"




| $o_1$ | $o_2$ | output |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| | | |
| | | |

1+0-1.5 = -0.5

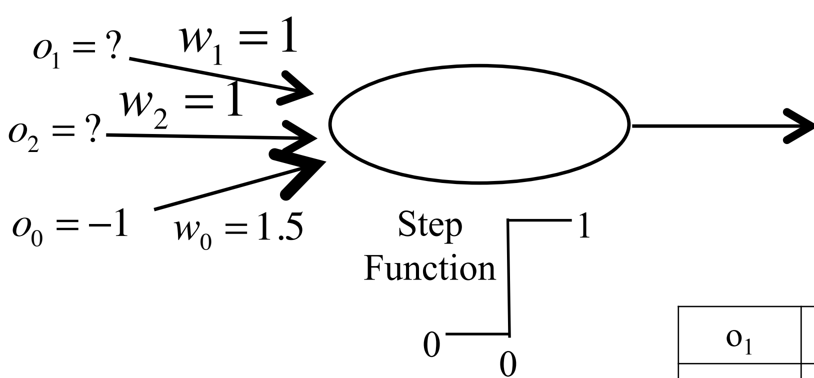
# McCulloch-Pitts “Neuron”




| $o_1$ | $o_2$ | output |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| | | |

0+1-1.5 = -0.5

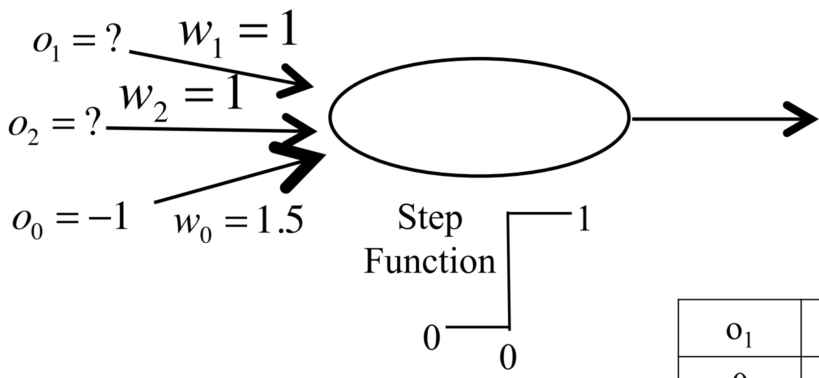
# McCulloch-Pitts “Neuron”




| $o_1$ | $o_2$ | output |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 1 | |

1+1-1.5 = 0.5

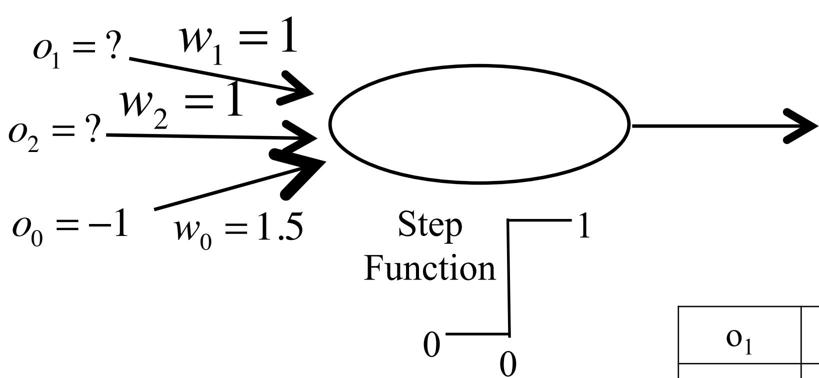
# McCulloch-Pitts “Neuron”




| $o_1$ | $o_2$ | output |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 1 | 1 |

1+1-1.5 = 0.5

# McCulloch-Pitts “Neuron”



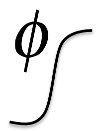

| $o_1$ | $o_2$ | output |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 1 | 1 |

This neuron computes the function “and.”

There are “or” and “not” neurons too.

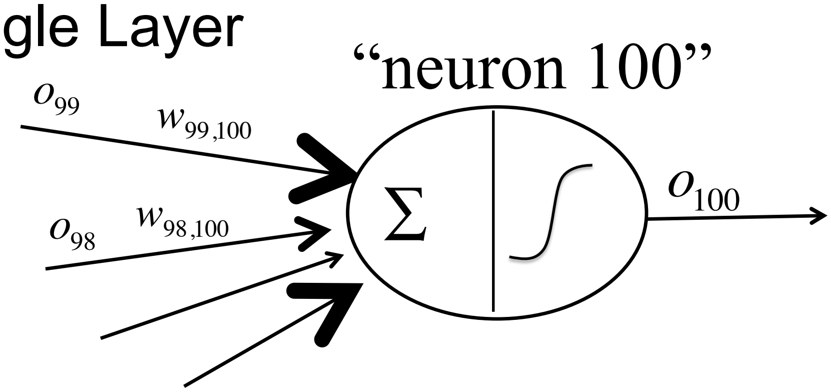
# McCulloch-Pitts "Neuron"

$$\phi\left(\sum_j w_{j,\mathrm{me}} a_j\right) = 1/(1+e^{-x}) \quad \text{“Sigmoid”}$$




Activation
Function

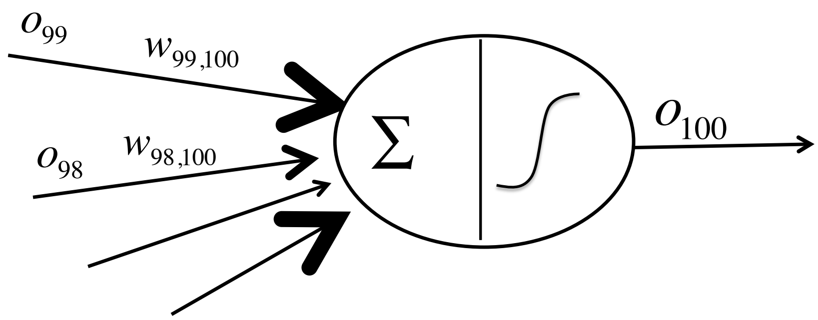
# Single Layer




Error on one observation

$$E = \frac{1}{2}(y - o_{100})^2$$

## Single Layer




Error on one observation

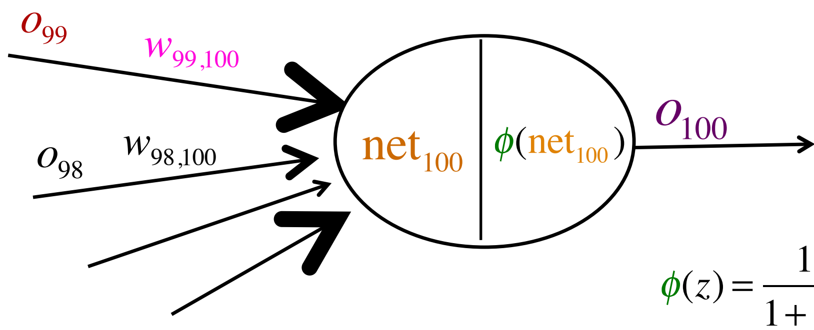
$$E = \frac{1}{2}(y - o_{100})^2$$

- In a brain, the synapses strengthen and weaken in order to learn.
- Say the same thing happens here.
- How should we set the weights in order to learn (reduce the error)?
- Minimize E with respect to the weights.

## Backpropagation

- An algorithm that trains the weights of a neural network
- Requires us to propagate information backwards through the network, then forwards, then backwards, then forwards, etc.
- Propagate backwards = chain rule from calculus.

## Single Layer




Error on one observation

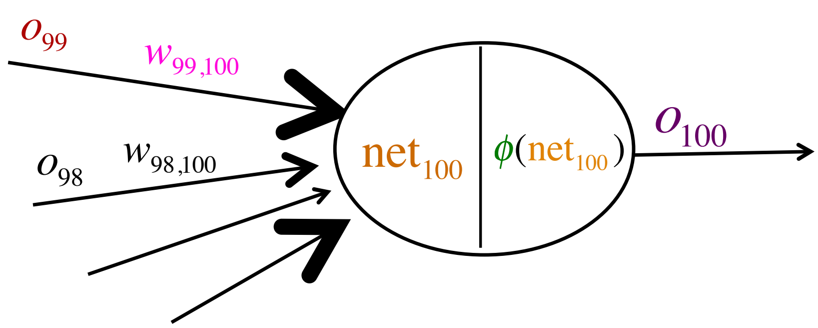
$$E = \frac{1}{2}(y - o_{100})^2$$

$$\phi(z) = \frac{1}{1+e^{z}}$$

$$\phi'(z) = \frac{d\phi(z)}{dz} = \phi(z)(1-\phi(z))$$

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \text{net}_{100}} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

## Single Layer




Error on one observation

$$E = \frac{1}{2}(y - o_{100})^2$$

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \text{net}_{100}} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

## Single Layer

Error on one observation

$$E = \frac{1}{2}(y - o_{100})^2$$

$$\frac{dE}{do_{100}} = \frac{1}{2}2(y - o_{100})(-1) = -(y - o_{100})$$

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \text{net}_{100}} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

$$-(y - o_{100})$$

## Single Layer

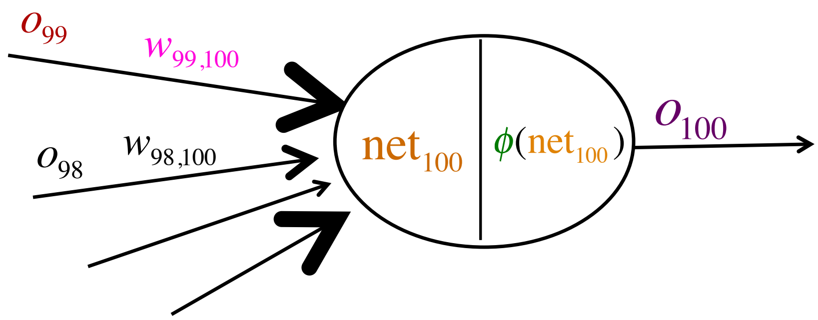
Error on one observation

$$E = \frac{1}{2}(y - o_{100})^2$$

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \text{net}_{100}} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

$$-(y - o_{100})$$

## Single Layer




Error on one observation

$$E = \frac{1}{2}(y - o_{100})^2$$

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \text{net}_{100}} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

$$-(y - o_{100})$$

## Single Layer

$\phi(\text{net}_{100})$ $o_{100}$

$$\frac{do_{100}}{d\ \text{net}_{100}} = \frac{d\phi(\text{net}_{100})}{d\ \text{net}_{100}} = \phi'(\text{net}_{100}) = \phi(\text{net}_{100})(1 - \phi(\text{net}_{100})) = o_{100}(1 - o_{100})$$

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \text{net}_{100}} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

$$-(y - o_{100})$$

# Single Layer

$\phi(\mathrm{net}_{100})$ $o_{100}$

$$\frac{do_{100}}{d\ \mathrm{net}_{100}} = \frac{d\phi(\mathrm{net}_{100})}{d\ \mathrm{net}_{100}} = \phi'(\mathrm{net}_{100}) = \phi(\mathrm{net}_{100})(1-\phi(\mathrm{net}_{100})) = o_{100}(1-o_{100})$$

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \mathrm{net}_{100}} \frac{d\ \mathrm{net}_{100}}{dw_{99,100}}$$

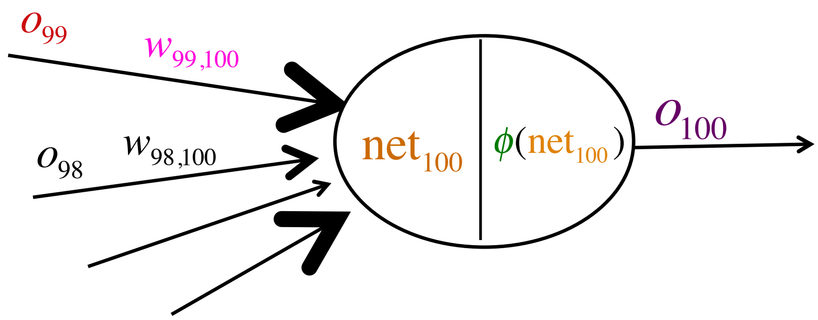
$-(y-o_{100})$ $o_{100}(1-o_{100})$

# Single Layer

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \mathrm{net}_{100}} \frac{d\ \mathrm{net}_{100}}{dw_{99,100}}$$

$-(y-o_{100})$ $o_{100}(1-o_{100})$

# Single Layer



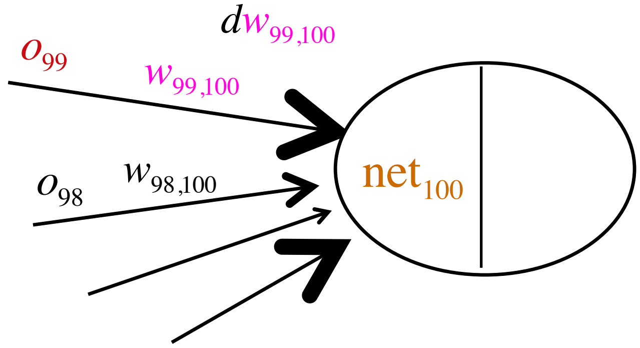

Error on one observation

$$E = \frac{1}{2}(y - o_{100})^2$$

$$\frac{dE}{dw_{99,100}} = \underbrace{\frac{dE}{do_{100}}}_{-(y-o_{100})} \underbrace{\frac{do_{100}}{d\ \text{net}_{100}}}_{o_{100}(1-o_{100})} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

---

$$\frac{d\ \text{net}_{100}}{dw_{99,100}} = \frac{d\ (w_{99,100}o_{99} + w_{98,100}o_{98} + w_{97,100}o_{97} + \ldots)}{dw_{99,100}} = o_{99}$$




$$\frac{dE}{dw_{99,100}} = \underbrace{\frac{dE}{do_{100}}}_{-(y-o_{100})} \underbrace{\frac{do_{100}}{d\ \text{net}_{100}}}_{o_{100}(1-o_{100})} \underbrace{\frac{d\ \text{net}_{100}}{dw_{99,100}}}_{o_{99}}$$

$$\frac{dE}{dw_{99,100}} = \frac{dE}{do_{100}} \frac{do_{100}}{d\ \text{net}_{100}} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

$o_{99}$

$-(y - o_{100})$ $\qquad$ $o_{100}(1 - o_{100})$

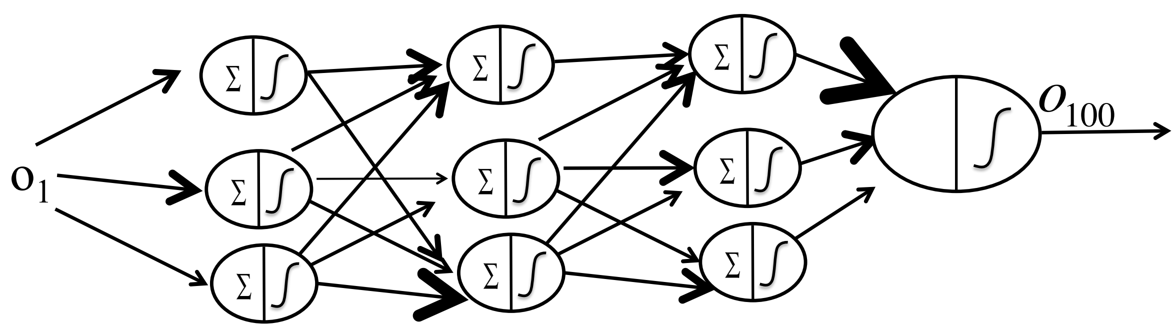
We will need this later – it depends only on node 100

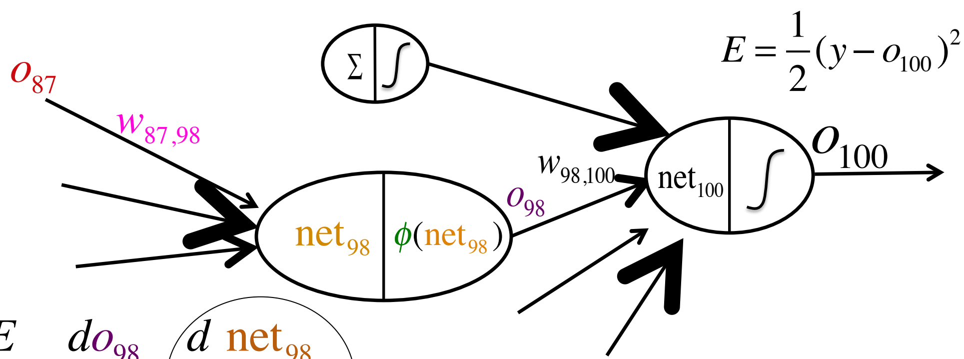
$$\frac{dE}{dw_{99,100}} = \overbrace{\frac{dE}{do_{100}} \frac{do_{100}}{d\ \text{net}_{100}}}^{\delta_{100}} \frac{d\ \text{net}_{100}}{dw_{99,100}}$$

$$= (-(y - o_{100}))o_{100}(1 - o_{100})o_{99}$$

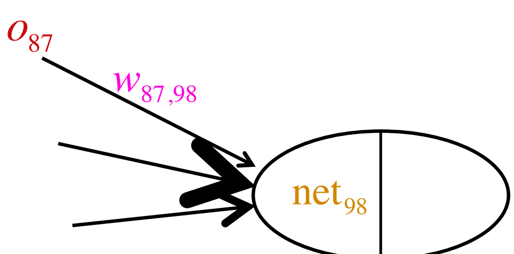
# Backpropagation

- Go one layer deeper.





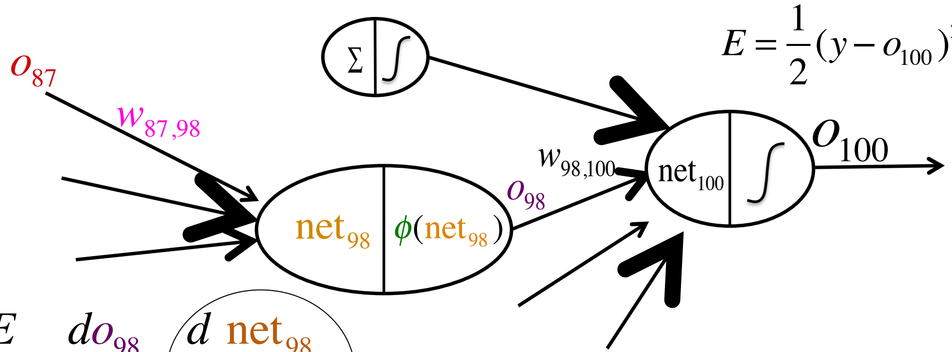

$$\frac{dE}{dw_{87,98}} = \frac{dE}{do_{98}} \frac{do_{98}}{d\ \text{net}_{98}} \frac{d\ \text{net}_{98}}{dw_{87,98}}$$




$$\frac{dE}{dw_{87,98}} = \frac{dE}{do_{98}} \frac{do_{98}}{d\ \text{net}_{98}} \frac{d\ \text{net}_{98}}{dw_{87,98}}$$

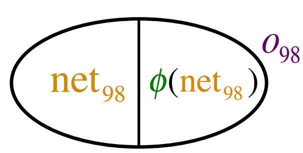
$$\frac{d\ \text{net}_{98}}{dw_{87,98}} = \frac{d\ (w_{87,98}o_{87} + w_{86,98}o_{86} + w_{85,98}o_{85} + \ldots)}{dw_{87,98}} = o_{87}$$

$$E = \frac{1}{2}(y - o_{100})^2$$




$$\frac{dE}{dw_{87,98}} = \frac{dE}{do_{98}} \frac{do_{98}}{d\ \mathrm{net}_{98}} \frac{d\ \mathrm{net}_{98}}{dw_{87,98}}$$

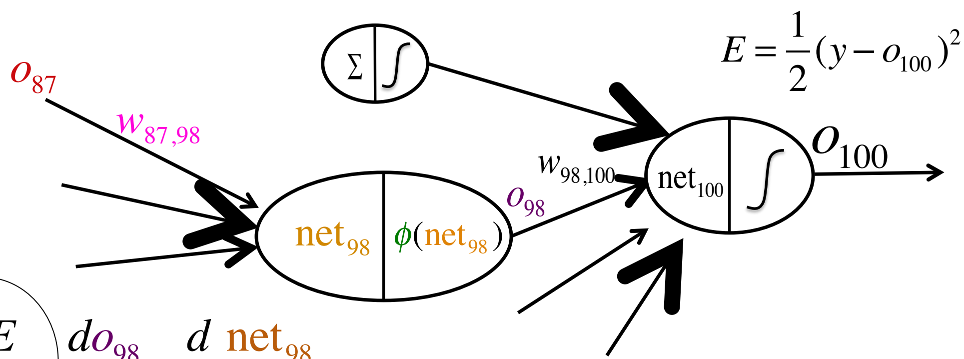
$o_{87}$




$$\frac{dE}{dw_{87,98}} = \frac{dE}{do_{98}} \frac{do_{98}}{d\ \mathrm{net}_{98}} \frac{d\ \mathrm{net}_{98}}{dw_{87,98}}$$

$o_{87}$

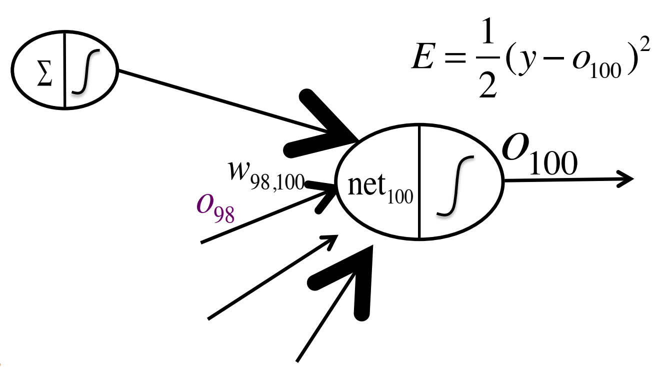
$$\frac{do_{98}}{d\ \mathrm{net}_{98}} = \frac{d\phi(\mathrm{net}_{98})}{d\ \mathrm{net}_{98}} = \phi'(\mathrm{net}_{98}) = \phi(\mathrm{net}_{98})(1 - \phi(\mathrm{net}_{98})) = o_{98}(1 - o_{98})$$




$$\frac{dE}{dw_{87,98}} = \frac{dE}{do_{98}} \frac{do_{98}}{d\ \text{net}_{98}} \frac{d\ \text{net}_{98}}{dw_{87,98}}$$

$o_{87}$

$o_{98}(1 - o_{98})$

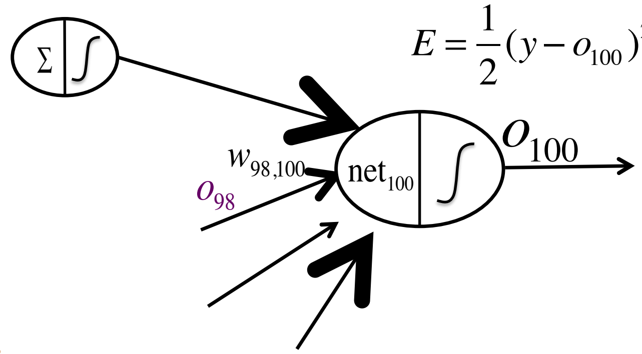



$$\frac{dE}{dw_{87,98}} = \frac{dE}{do_{98}} \frac{do_{98}}{d\ \text{net}_{98}} \frac{d\ \text{net}_{98}}{dw_{87,98}}$$

$o_{87}$

$o_{98}(1 - o_{98})$

$$E=\frac{1}{2}(y-o_{100})^2$$




$$\frac{dE}{do_{98}}=\frac{dE}{d\text{net}_{100}}\frac{d\text{net}_{100}}{do_{98}}$$

$\delta_{100}$

$$\frac{dE}{dw_{87,98}}=\frac{dE}{do_{98}}\frac{do_{98}}{d\ \text{net}_{98}}\frac{d\ \text{net}_{98}}{dw_{87,98}}$$

$o_{87}$

$o_{98}(1-o_{98})$

$$\frac{d\text{net}_{100}}{do_{98}}=\frac{d(w_{99,100}o_{99}+w_{98,100}o_{98}+\ldots)}{do_{98}}=w_{98,100}$$

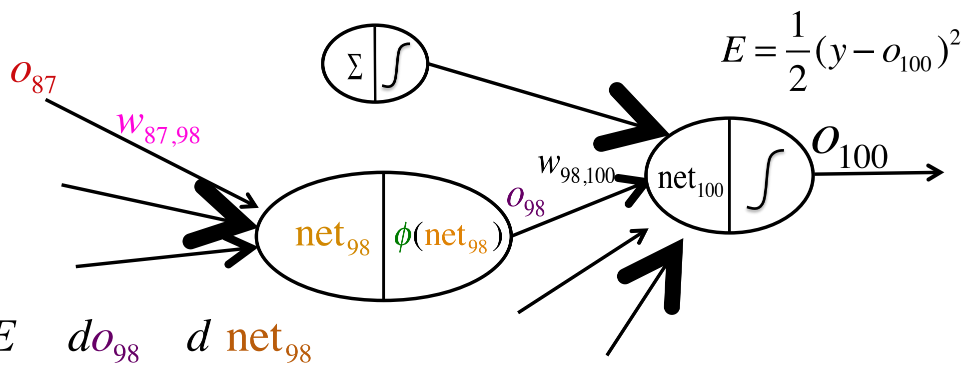
$$\frac{dE}{do_{98}}=\frac{dE}{d\text{net}_{100}}\frac{d\text{net}_{100}}{do_{98}}$$

$\delta_{100}$

$$\frac{dE}{dw_{87,98}}=\frac{dE}{do_{98}}\frac{do_{98}}{d\ \text{net}_{98}}\frac{d\ \text{net}_{98}}{dw_{87,98}}$$

$o_{87}$

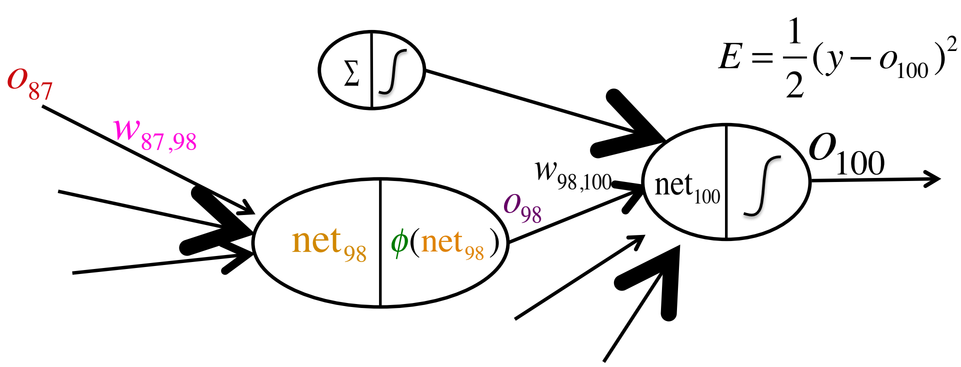
$o_{98}(1-o_{98})$




$$\frac{dE}{dw_{87,98}} = \frac{dE}{do_{98}} \frac{do_{98}}{d\ \text{net}_{98}} \frac{d\ \text{net}_{98}}{dw_{87,98}}$$

$o_{87}$

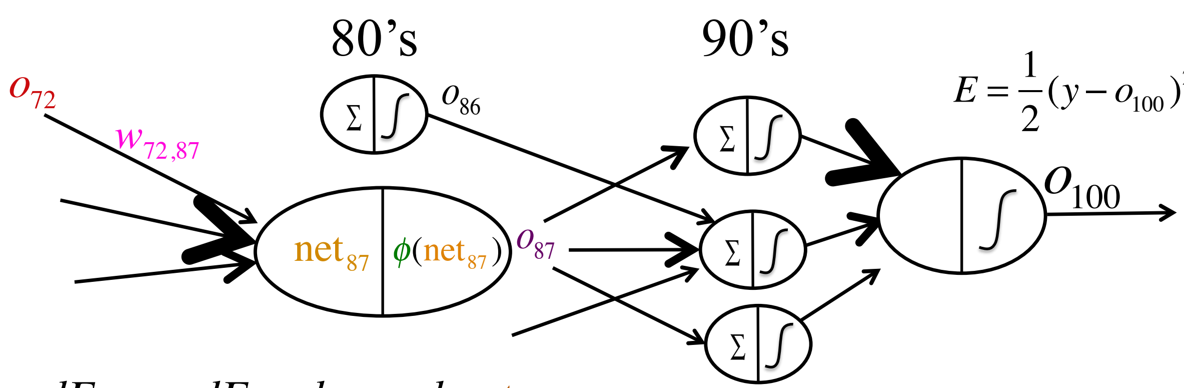
$\delta_{100} w_{98,100}$ $\quad$ $o_{98}(1 - o_{98})$



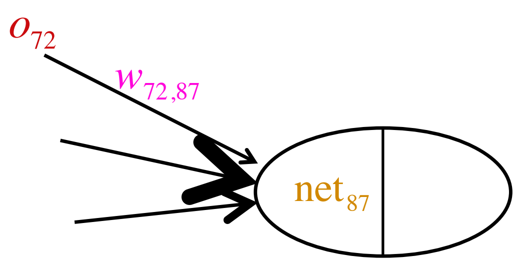

$$\frac{dE}{dw_{87,98}} = \delta_{100} w_{98,100} o_{98}(1 - o_{98}) o_{87}$$

# Backpropagation

- Go even one layer deeper.
- Third time is a charm.

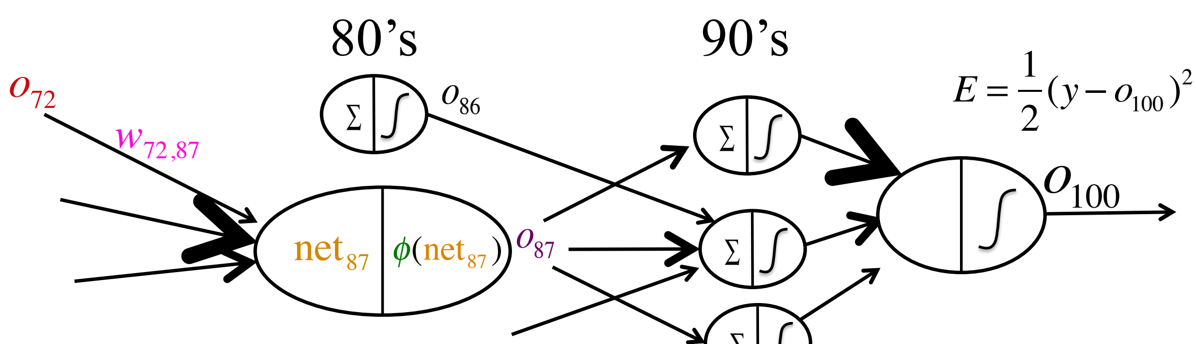



$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \frac{do_{87}}{d\ \mathrm{net}_{87}} \frac{d\ \mathrm{net}_{87}}{dw_{72,87}}$$



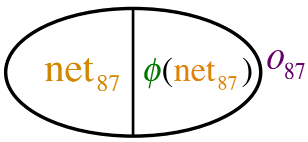

$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \frac{do_{87}}{d\ \text{net}_{87}} \frac{d\ \text{net}_{87}}{dw_{72,87}}$$

$o_{72}$




$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \frac{do_{87}}{d\ \text{net}_{87}} \frac{d\ \text{net}_{87}}{dw_{72,87}}$$

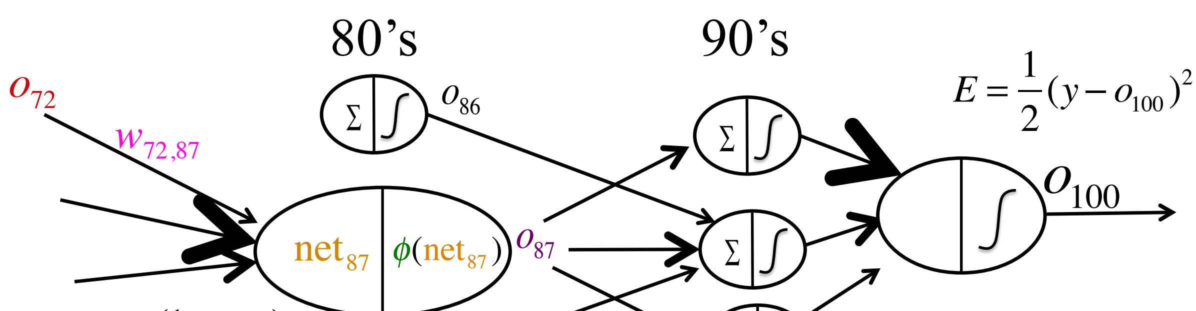
$o_{72}$




$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \frac{do_{87}}{d\ \text{net}_{87}} \frac{d\ \text{net}_{87}}{dw_{72,87}}$$

$o_{72}$

$$\frac{do_{87}}{d\ \text{net}_{87}} = \frac{d\phi(\text{net}_{87})}{d\ \text{net}_{87}} = \phi'(\text{net}_{87}) = \phi(\text{net}_{87})(1-\phi(\text{net}_{87})) = o_{87}(1-o_{87})$$



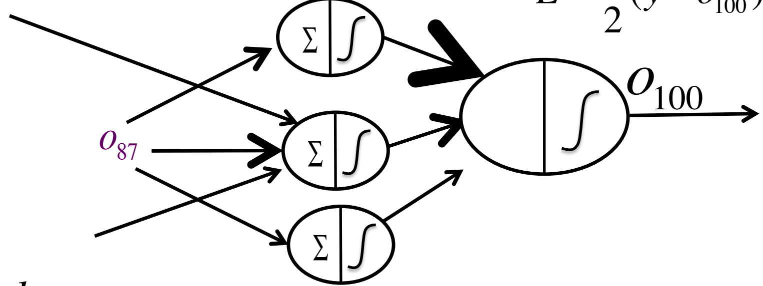

$o_{87}(1-o_{87})$

$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \frac{do_{87}}{d\ \text{net}_{87}} \frac{d\ \text{net}_{87}}{dw_{72,87}}$$

$o_{72}$

# 90's

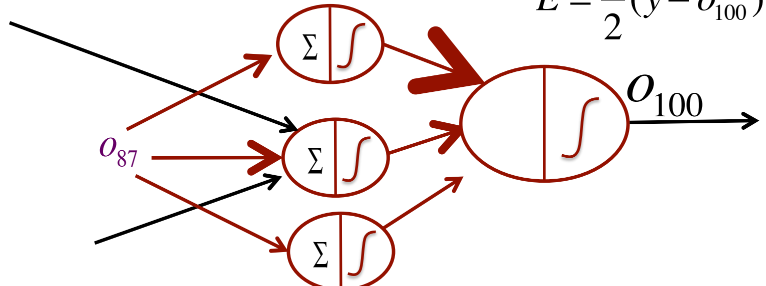
$$E = \frac{1}{2}(y - o_{100})^2$$




$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \overbrace{\frac{do_{87}}{d\ \mathrm{net}_{87}}}^{o_{87}(1-o_{87})} \underbrace{\frac{d\ \mathrm{net}_{87}}{dw_{72,87}}}_{o_{72}}$$

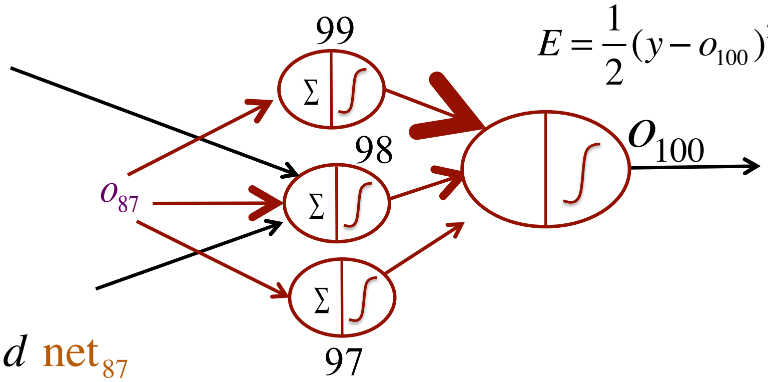
# 90's

$$E = \frac{1}{2}(y - o_{100})^2$$




$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \overbrace{\frac{do_{87}}{d\ \mathrm{net}_{87}}}^{o_{87}(1-o_{87})} \underbrace{\frac{d\ \mathrm{net}_{87}}{dw_{72,87}}}_{o_{72}}$$

$$E = \frac{1}{2}(y - o_{100})^2$$



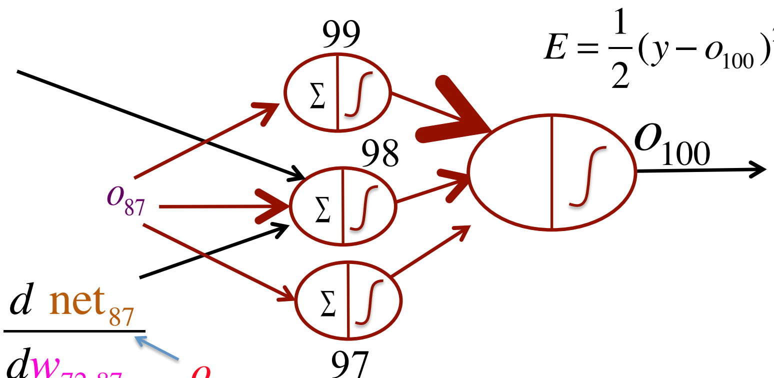

$$o_{87}(1 - o_{87})$$

$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \frac{do_{87}}{d\ \mathrm{net}_{87}} \frac{d\ \mathrm{net}_{87}}{dw_{72,87}}$$

$$o_{72}$$

$$\frac{dE}{do_{87}} = \frac{dE}{d\ \mathrm{net}_{99}} \frac{d\ \mathrm{net}_{99}}{do_{87}} + \frac{dE}{d\ \mathrm{net}_{98}} \frac{d\ \mathrm{net}_{98}}{do_{87}} + \frac{dE}{d\ \mathrm{net}_{97}} \frac{d\ \mathrm{net}_{97}}{do_{87}} + \ldots$$

$$E = \frac{1}{2}(y - o_{100})^2$$




$$o_{87}(1 - o_{87})$$

$$\frac{dE}{dw_{72,87}} = \frac{dE}{do_{87}} \frac{do_{87}}{d\ \mathrm{net}_{87}} \frac{d\ \mathrm{net}_{87}}{dw_{72,87}}$$

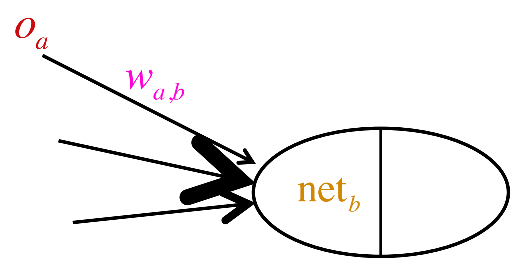
$$o_{72}$$

$$\frac{dE}{do_{87}} = \frac{dE}{d\ \mathrm{net}_{99}} \frac{d\ \mathrm{net}_{99}}{do_{87}} + \frac{dE}{d\ \mathrm{net}_{98}} \frac{d\ \mathrm{net}_{98}}{do_{87}} + \frac{dE}{d\ \mathrm{net}_{97}} \frac{d\ \mathrm{net}_{97}}{do_{87}} + \ldots$$

$$= \delta_{99} w_{87,99} + \delta_{98} w_{87,98} + \delta_{97} w_{87,97} + \ldots = \sum_{\ell \in L} \delta_{87} w_{87,\ell}$$

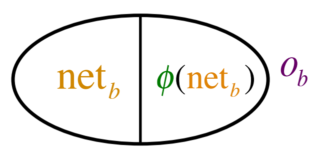
$$\frac{dE}{dw_{a,b}} = \frac{dE}{do_b}\frac{do_b}{d\ \mathrm{net}_b}\frac{d\ \mathrm{net}_b}{dw_{a,b}}$$

$$= \frac{dE}{do_b}\frac{do_b}{d\ \mathrm{net}_b} o_a$$




$$\frac{dE}{dw_{a,b}} = \frac{dE}{do_b}\frac{do_b}{d\ \mathrm{net}_b}\frac{d\ \mathrm{net}_b}{dw_{a,b}}$$

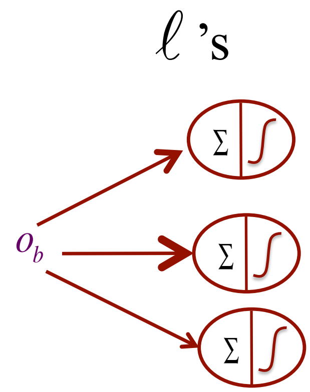
$$= \frac{dE}{do_b} o_b(1-o_b) o_a$$

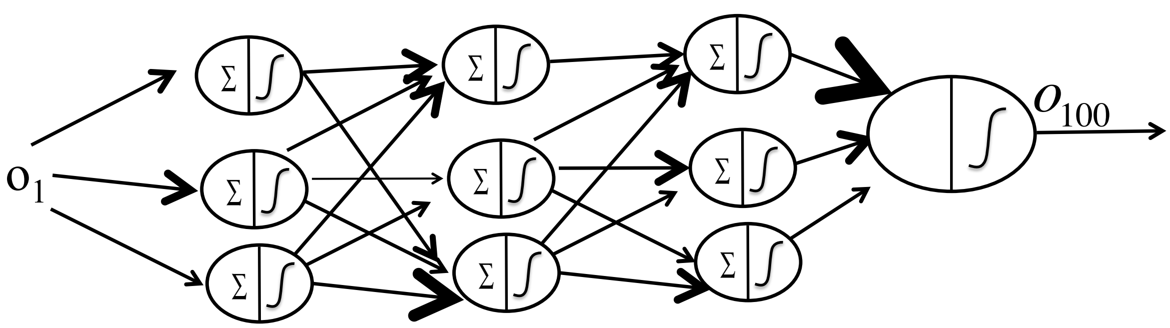


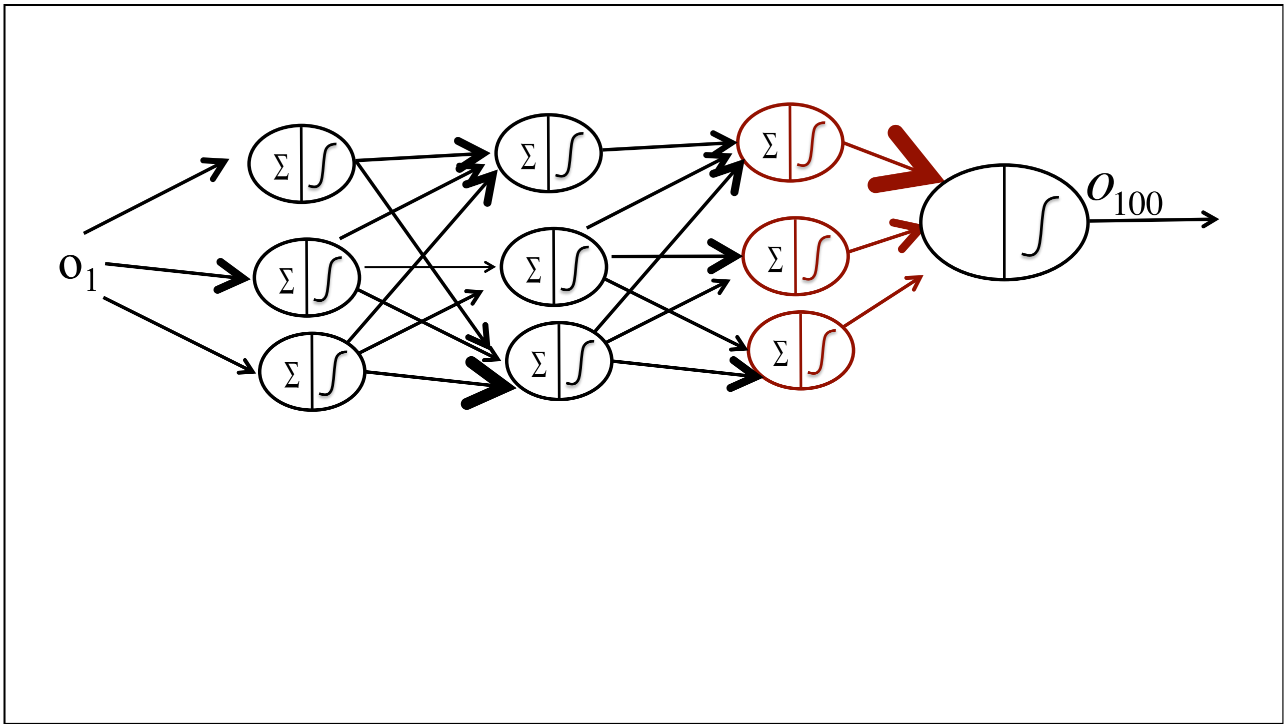
$$\frac{dE}{dw_{a,b}} = \frac{dE}{do_b}\,\frac{do_b}{d\ \text{net}_b}\,\frac{d\ \text{net}_b}{dw_{a,b}}$$

$$= \left(\sum_{\ell \in L} \delta_\ell w_{b,\ell}\right) o_b(1-o_b)\,o_a$$

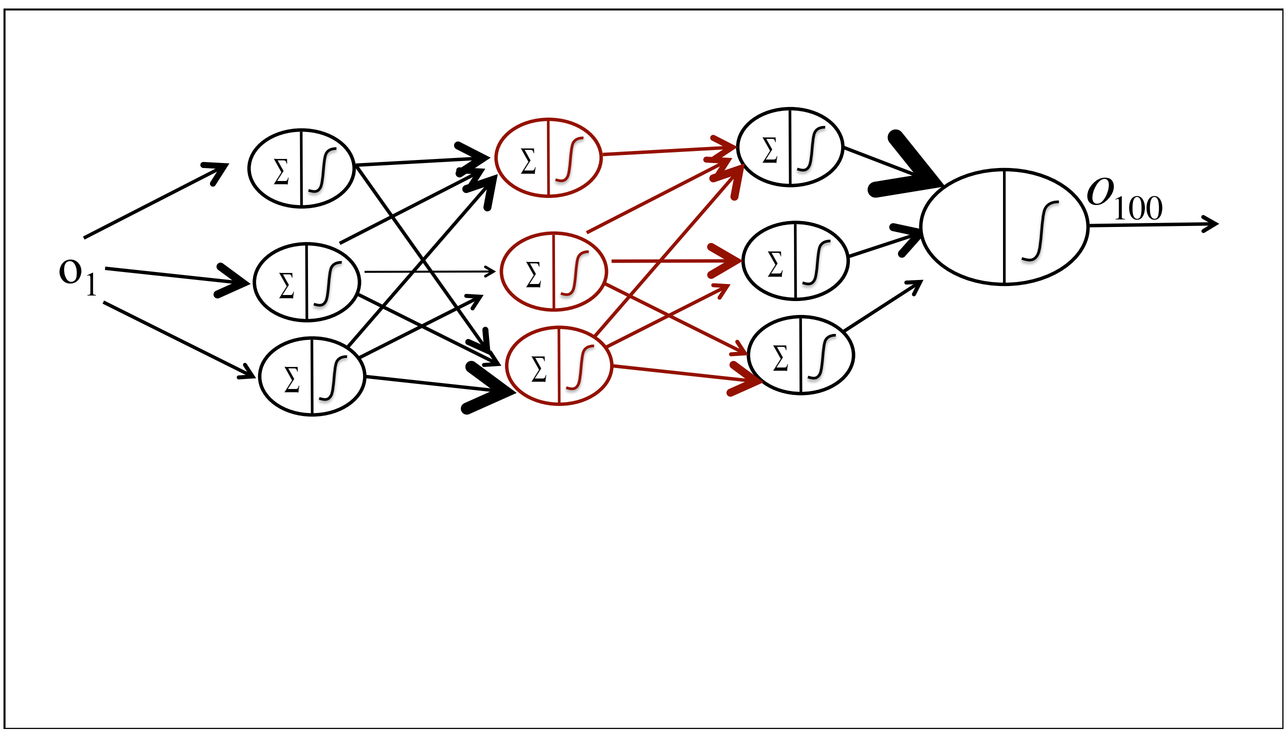
The L are downstream. We must have already computed all the $\delta_\ell$'s ahead of us to compute this.



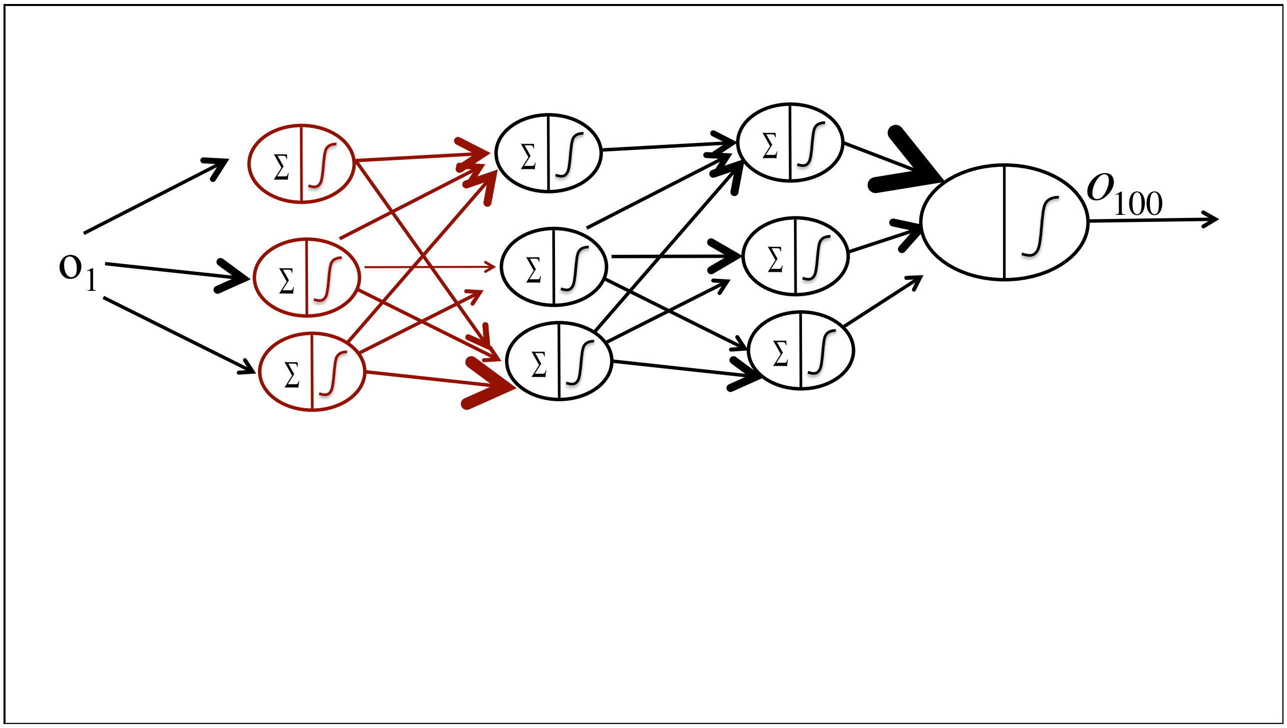





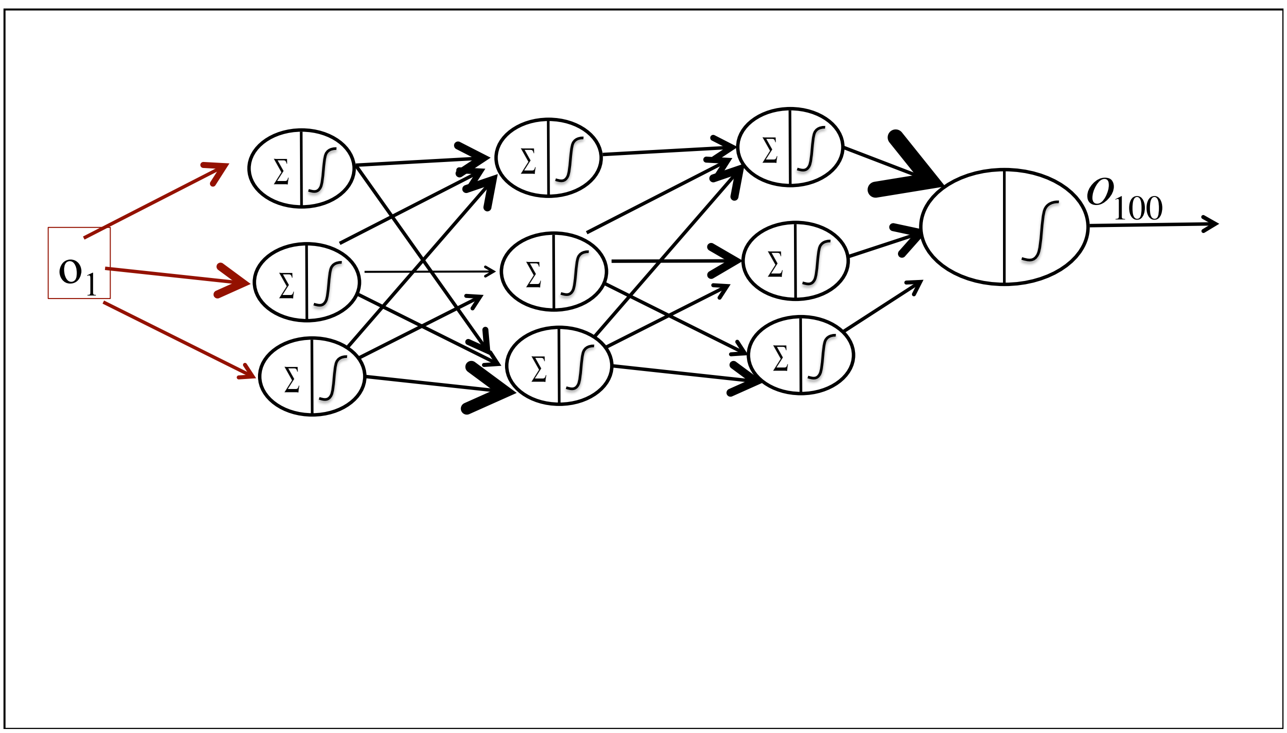
$o_1$
$o_{100}$

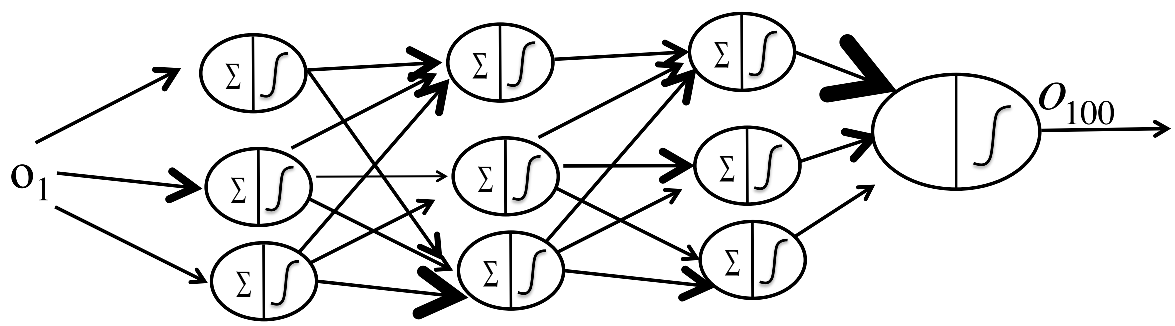
$o_1$
$o_{100}$



o1
o100

o1
o100

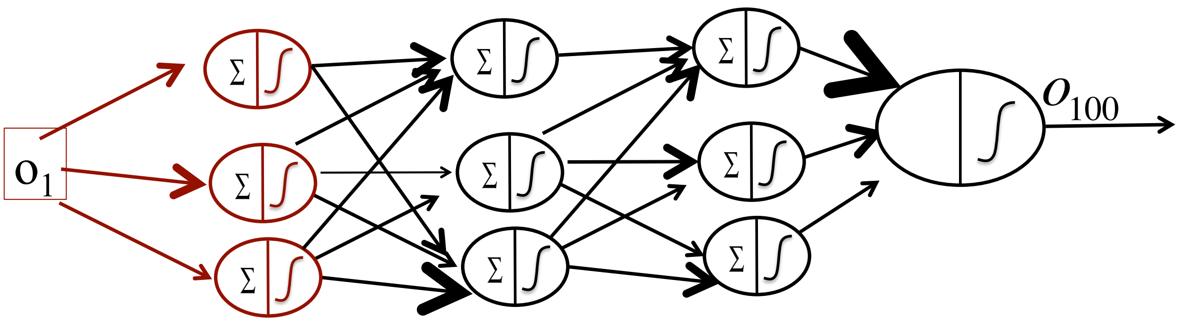



# Backpropagation

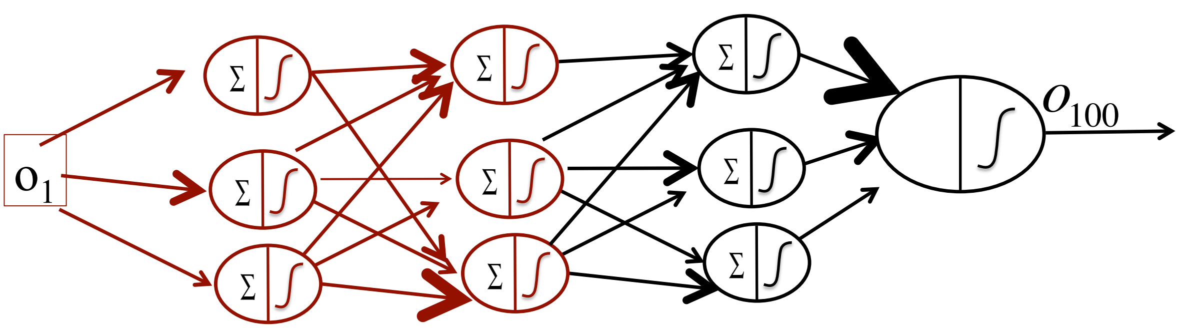
- Now we know how to compute $\frac{dE}{dw_{a,b}}$ for all of the $w_{a,b}$'s.
- Let's do gradient descent.

$$w_{a,b} \longleftarrow w_{a,b} - \alpha \frac{dE}{dw_{a,b}}$$

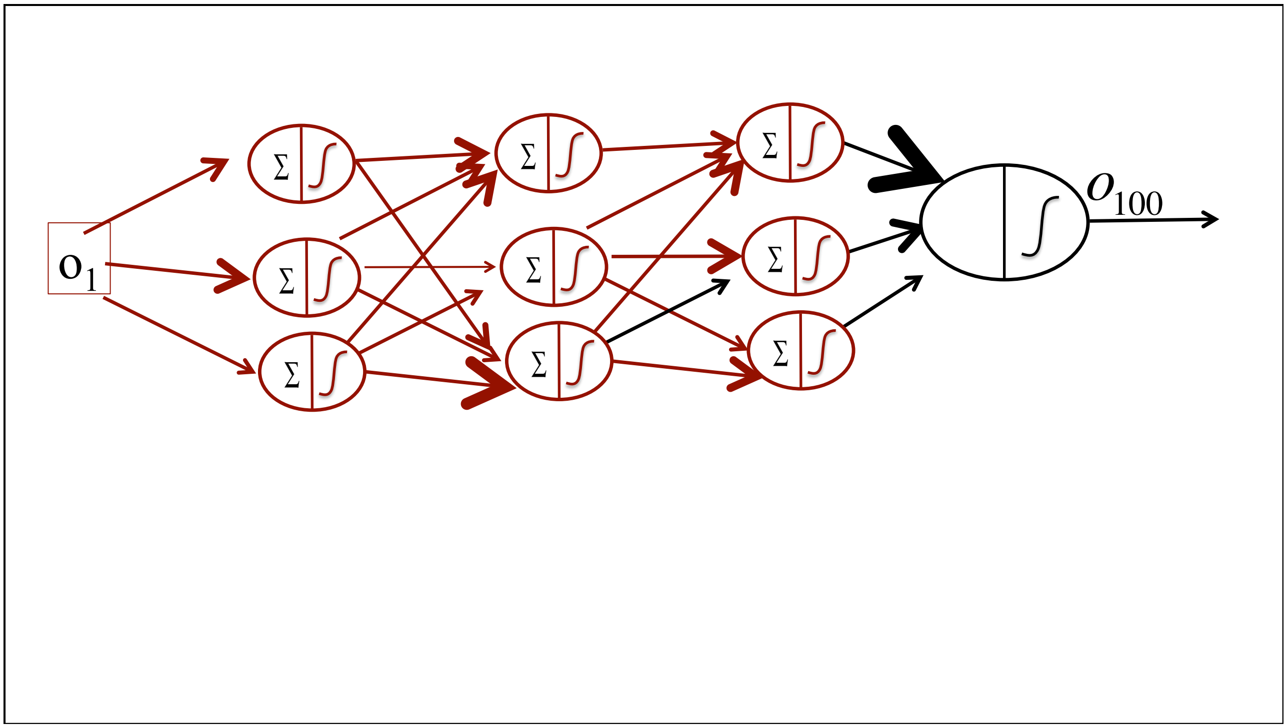
- α is between 0 and 1. Called the "learning rate".
- Now we know how to propagate errors back through the network.
- Remember how to go forward?



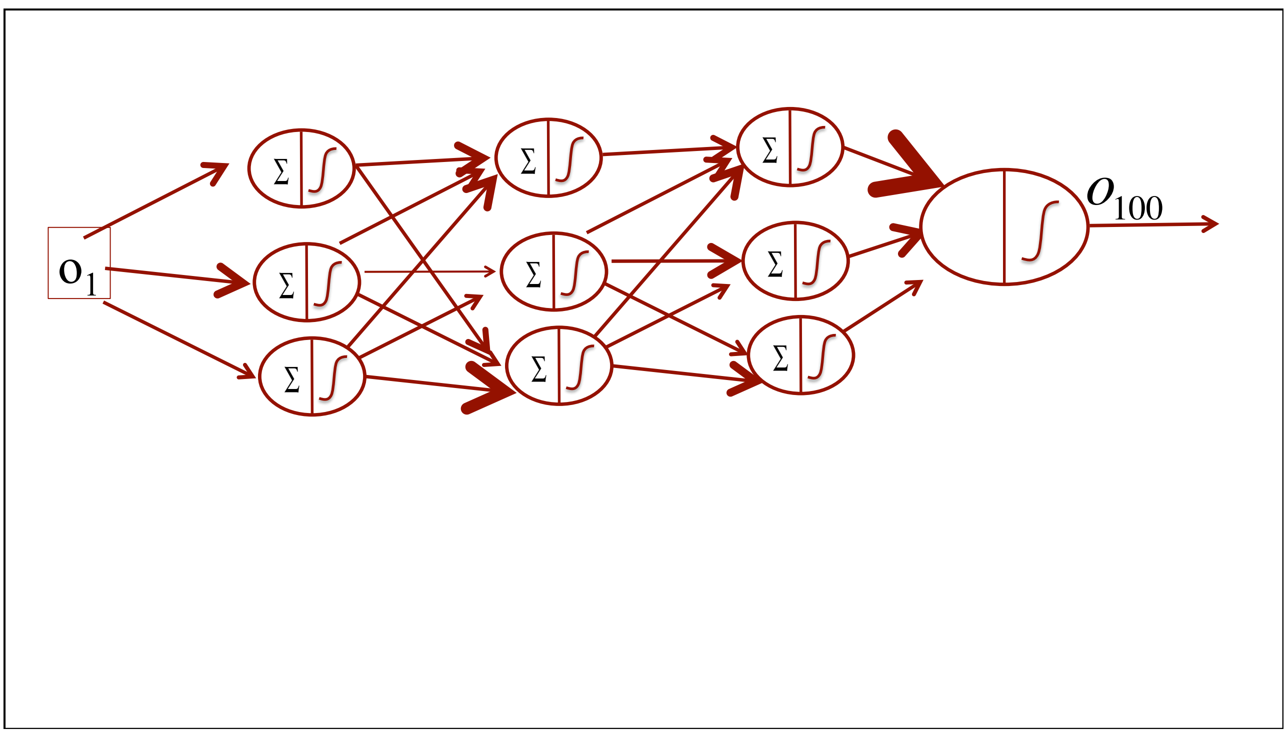
$o_1$
$o_{100}$

$o_1$
$o_{100}$



$o_1$
Σ
$o_{100}$

$o_1$
Σ
$o_{100}$

## Backpropagation

- Repeat going backwards (to calculate the gradients), adjusting the weights, and going forwards (to calculate the errors) over and over in order to learn.

## Neural networks

- Advantages:
  - highly expressive nonlinear models
  - have advances in computer vision and speech that other methods have not achieved
  - can capture latent structure within the hidden layers
- Disadvantages
  - can get stuck in local optima, could produce bad solutions
  - black box
  - lots of tuning parameters (e.g., the structure of the network)

Microsoft