CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Artificial Neural Networks 1

Morten Nielsen
Department of Health Technology,
DTU

Objectives

Input → Neural network → Output

- Neural network:
  - is a black box that no one can understand
  - over-predicts performance
  - Overfitting - many thousand parameters fitted on few data

HUNKAT

HUNKAT

HUMAN

NETtalk

(T. Sejnowski and C. Rosenberg, 1987)

M**a**ry h**a**d **a** little l**a**mb

Three of the **a**'s must be pronounced differently! Reading aloud is a *context sensitive* cognitive skill.

Weight matrices (PSSM)

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

- A weight matrix is given as

  $W_{ij} = \log(p_{ij}/q_j)$

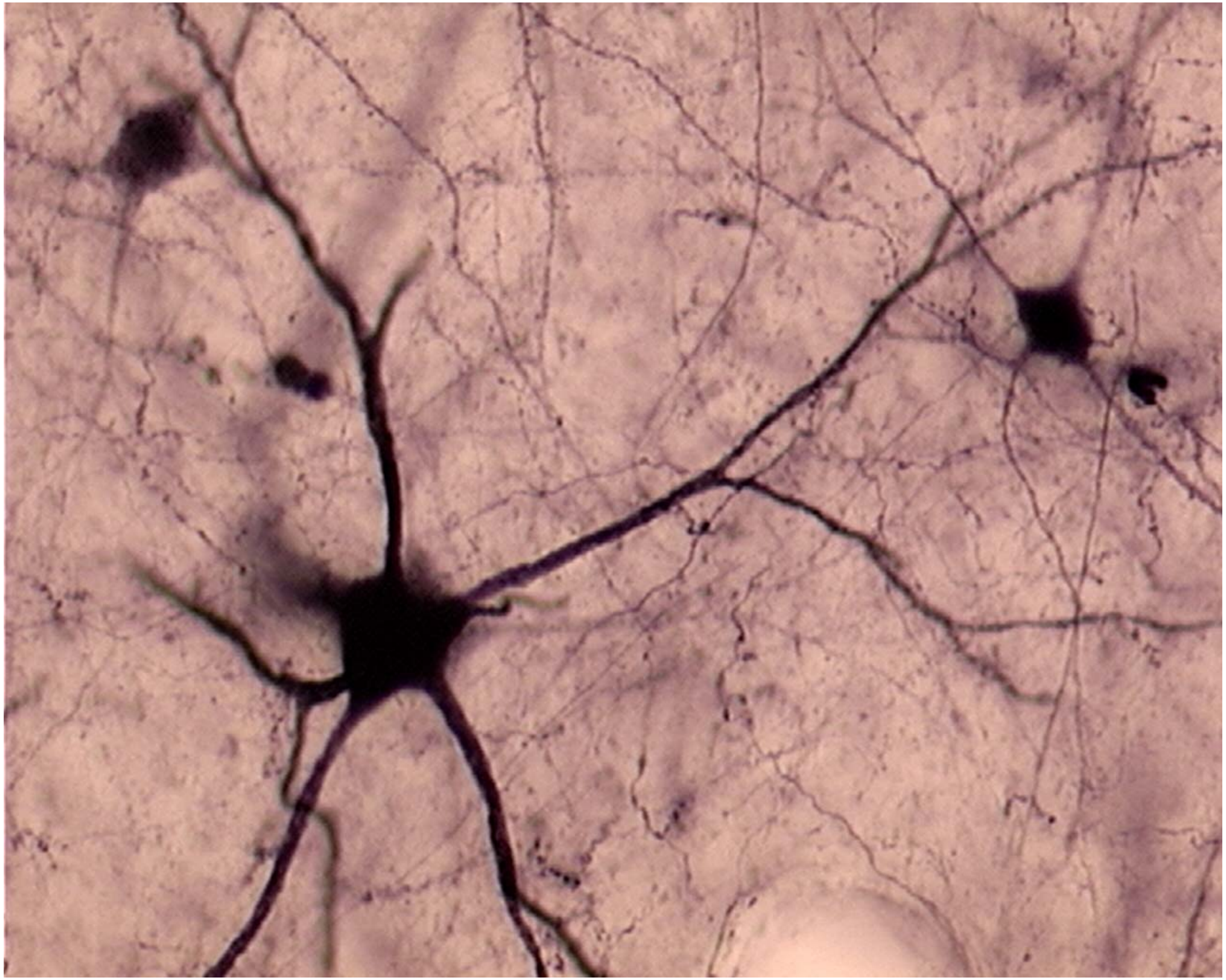
  - where i is a position in the motif, and j an amino acid. q_j is the background frequency for amino acid j.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.6	0.4	-3.5	-2.4	-0.4	-1.9	-2.7	0.3	-1.1	1.0	0.3	0.0	1.4	1.2	-2.7	1.4	-1.2	-2.0	1.1	0.7
2	-1.6	-6.6	-6.5	-5.4	-2.5	-4.0	-4.7	-3.7	-6.3	1.0	5.1	-3.7	3.1	-4.2	-4.3	-4.2	-0.2	-5.9	-3.8	0.4
3	0.2	-1.3	0.1	1.5	0.0	-1.8	-3.3	0.4	0.5	-1.0	0.3	-2.5	1.2	1.0	-0.1	-0.3	-0.5	3.4	1.6	0.0
4	-0.1	-0.1	-2.0	2.0	-1.6	0.5	0.8	2.0	-3.3	0.1	-1.7	-1.0	-2.2	-1.6	1.7	-0.6	-0.2	1.3	-6.8	-0.7
5	-1.6	-0.1	0.1	-2.2	-1.2	0.4	-0.5	1.9	1.2	-2.2	-0.5	-1.3	-2.2	1.7	1.2	-2.5	-0.1	1.7	1.5	1.0
6	-0.7	-1.4	-1.0	-2.3	1.1	-1.3	-1.4	-0.2	-1.0	1.8	0.8	-1.9	0.2	1.0	-0.4	-0.6	0.4	-0.5	-0.0	2.1
7	1.1	-3.8	-0.2	-1.3	1.3	-0.3	-1.3	-1.4	2.1	0.6	0.7	-5.0	1.1	0.9	1.3	-0.5	-0.9	2.9	-0.4	0.5
8	-2.2	1.0	-0.8	-2.9	-1.4	0.4	0.1	-0.4	0.2	-0.0	1.1	-0.5	-0.5	0.7	-0.3	0.8	0.8	-0.7	1.3	-1.1
9	-0.2	-3.5	-6.1	-4.5	0.7	-0.8	-2.5	-4.0	-2.6	0.9	2.8	-3.0	-1.8	-1.4	-6.2	-1.9	-1.6	-4.9	-1.6	4.5

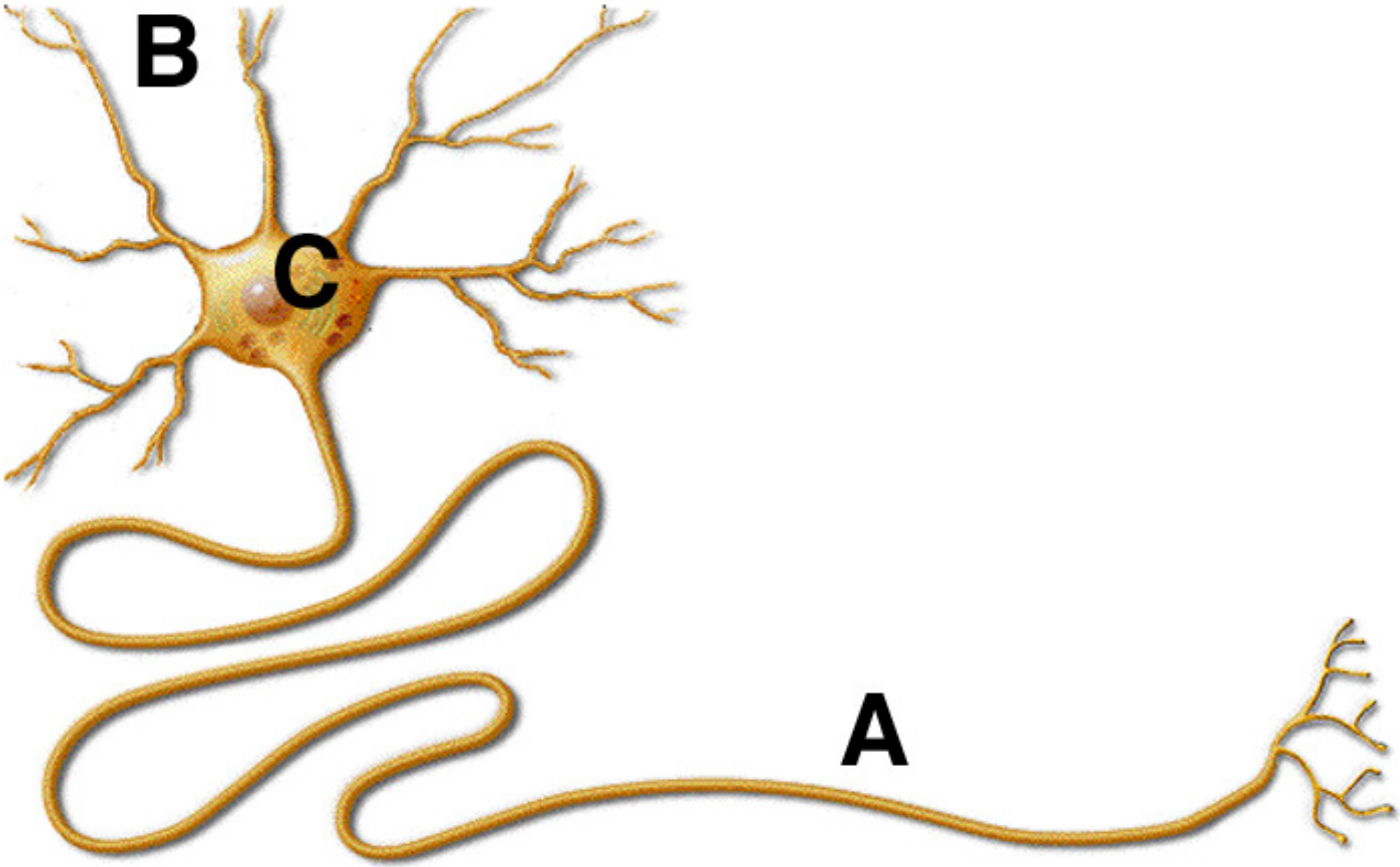
- W is a L x 20 matrix, L is motif length

SLLPAIVEL
YLIPAIVHI
TLWVDPYEV

Biological Neural network



Biological neuron structure



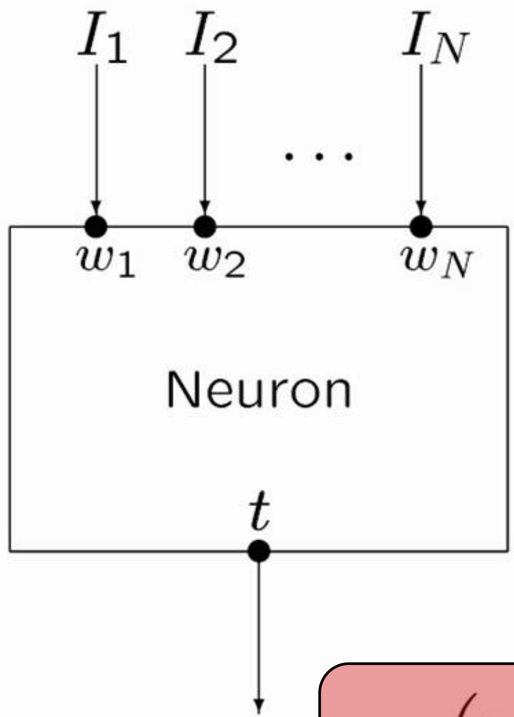
Artificial neuron

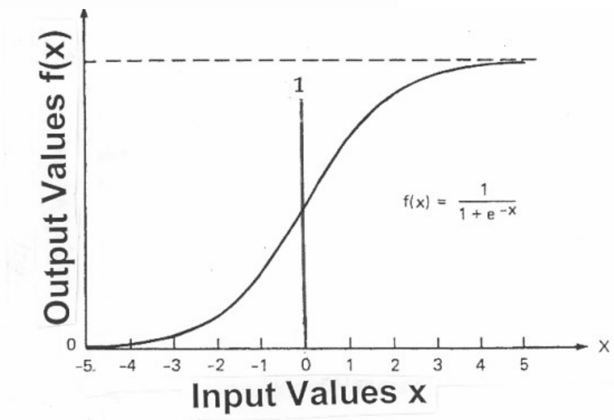
Input signals

Synaptic weights

Threshold

Output signal



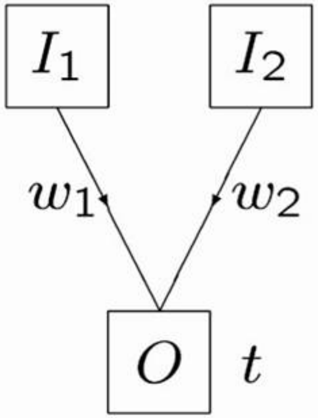


$$O = \sigma\left(\sum_{n=1}^{N} w_n I_n - t\right)$$

Transfer of biological principles to artificial neural network algorithms

- Non-linear relation between input and output
- Massively parallel information processing
- Data-driven construction of algorithms
- Ability to generalize to new data items

Linear separation by simple neural network

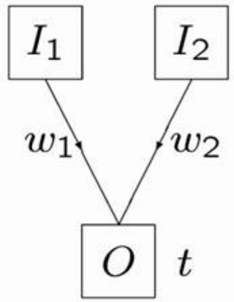


Two input features and one output.

$$O = \begin{cases} 1 & \text{for } w_1 I_1 + w_2 I_2 > t \\ 0 & \text{otherwise} \end{cases}$$

Similar to SMM, except for step function!

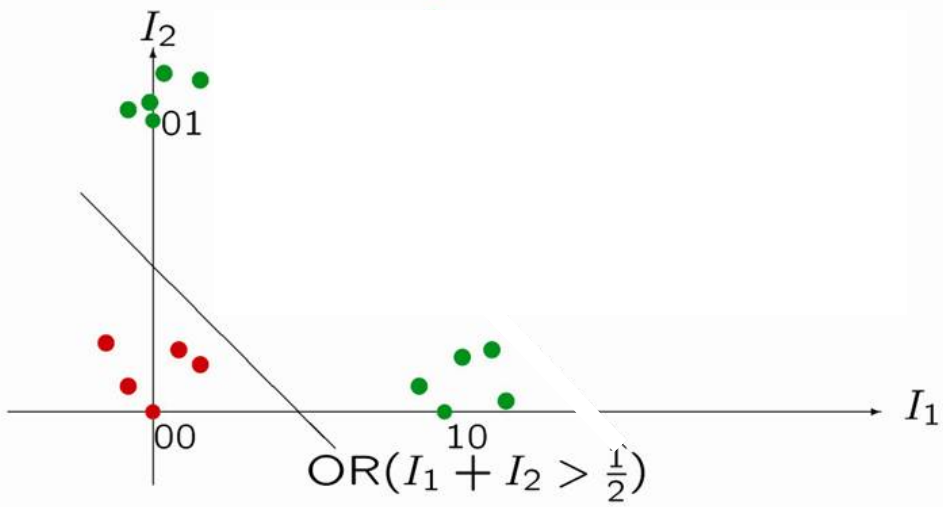
Linear separation by simple neural network



Two input features and one output.

$$O = \begin{cases} 1 & \text{for } w_1 I_1 + w_2 I_2 > t \\ 0 & \text{otherwise} \end{cases}$$

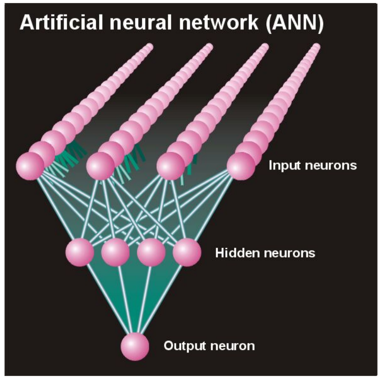
Equation $w_1 I_1 + w_2 I_2 = t$ is straight line in $I_1 I_2$-plane:



Higher order correlations

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

- The effect on the binding affinity of having a given amino acid at one position can be influenced by the amino acids at other positions in the peptide (sequence correlations).
  - Two adjacent amino acids may for example compete for the space in a pocket in the MHC molecule.
- Artificial neural networks (ANN) are ideally suited to take such correlations into account



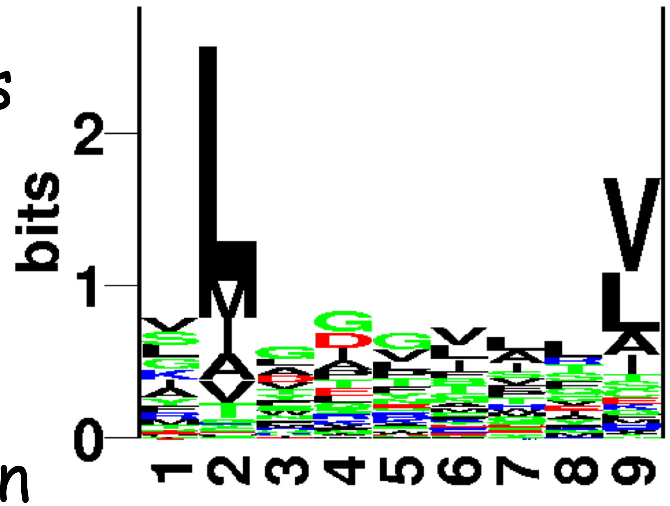
MHC peptide binding

SLLPAIVEL	YLLPAIVHI	TLWVDPYEV	GLVPFLVSV	KLLEPVLLL	LLDVPTAAV	LLDVPTAAV	LLDVPTAAV
LLDVPTAAV	VLFRGGPRG	MVDGTLLLL	YMNGTMSQV	MLLSVPLLL	SLLGLLVEV	ALLPPINIL	TLIKIQHTL
HLIDYLVTS	ILAPPVVKL	ALFPQLVIL	GILGFVFTL	STNRQSGRQ	GLDVLTAKV	RILGAVAKV	QVCERIPTI
ILFGHENRV	ILMEHIHKL	ILDQKINEV	SLAGGIIGV	LLIENVASL	FLLWATAEA	SLPDFGISY	KKREEAPSL
LERPGGNEI	ALSNLEVKL	ALNELLQHV	DLERKVESL	FLGENISNF	ALSDHHIYL	GLSEFTEYL	STAPPAHGV
PLDGEYFTL	GVLVGVALI	RTLDKVLEV	HLSTAFARV	RLDSYVRSL	YMNGTMSQV	GILGFVFTL	ILKEPVHGV
ILGFVFTLT	LLFGYPVYV	GLSPTVWLS	WLSLLVPFV	FLPSDFFPS	CLGGLLTMV	FIAGNSAYE	KLGEFYNQM
KLVALGINA	DLMGYIPLV	RLVTLKDIV	MLLAVLYCL	AAGIGILTV	YLEPGPVTA	LLDGTATLR	ITDQVPFSV
KTWGQYWQV	TITDQVPFS	AFHHVAREL	YLNKIQNSL	MMRKLAILS	AIMDKNIIL	IMDKNIILK	SMVGNWAKV
SLLAPGAKQ	KIFGSLAFL	ELVSEFSRM	KLTPLCVTL	VLYRYGSFS	YIGEVLVSV	CINGVCWTV	VMNILLQYV
ILTVILGVL	KVLEYVIKV	FLWGPRALV	GLSRYVARL	FLLTRILTI	HLGNVKYLV	GIAGGLALL	GLQDCTMLV
TGAPVTYST	VIYQYMDDL	VLPDVFIRC	VLPDVFIRC	AVGIGIAVV	LVVLGLLAV	ALGLGLLPV	GIGIGVLAA
GAGIGVAVL	IAGIGILAI	LIVIGILIL	LAGIGLIAA	VDGIGILTI	GAGIGVLTA	AAGIGIIQI	QAGIGILLA
KARDPHSGH	KACDPHSGH	ACDPHSGHF	SLYNTVATL	RGPGRAFVT	NLVPMVATV	GLHCYEQLV	PLKQHFQIV
AVFDRKSDA	LLDFVRFMG	VLVKSPNHV	GLAPPQHLI	LLGRNSFEV	PLTFGWCYK	VLEWRFDSR	TLNAWVKVV
GLCTLVAML	FIDSYICQV	IISAVVGIL	VMAGVGSPY	LLWTLVVLL	SVRDRLARL	LLMDCSGSI	CLTSTVQLV
VLHDDLLEA	LMWITQCFL	SLLMWITQC	QLSLLMWIT	LLGATCMFV	RLTRFLSRV	YMDGTMSQV	FLTPKKLQC
ISNDVCAQV	VKTDGNPPE	SVYDFFVWL	FLYGALLLA	VLFSSDFRI	LMWAKIGPV	SLLLELEEV	SLSRFSWGA
YTAFTIPSI	RLMKQDFSV	RLPRIFCSC	FLWGPRAYA	RLLQETELV	SLFEGIDFY	SLDQSVVEL	RLNMFTPYI
NMFTPYIGV	LMIIPLINV	TLFIGSHVV	SLVIVTTFV	VLQWASLAV	ILAKFLHWL	STAPPHVNV	LLLLTVLTV
VVLGVVFGI	ILHNGAYSL	MIMVKCWMI	MLGTHTMEV	MLGTHTMEV	SLADTNSLA	LLWAARPRL	GVALQTMKQ
GLYDGMEHL	KMVELVHFL	YLQLVFGIE	MLMAQEALA	LMAQEALAF	VYDGREHTV	YLSGANLNL	RMFPNAPYL
EAAGIGILT	TLDSQVMSL	STPPPGTRV	KVAELVHFL	IMIGVLVGV	ALCRWGLLL	LLFAGVQCQ	VLLCESTAV
YLSTAFARV	YLLEMLWRL	SLDDYNHLV	RTLDKVLEV	GLPVEYLQV	KLIANNTRV	FIYAGSLSA	KLVANNTRL
FLDEFMEGV	ALQPGTALL	VLDGLDVLL	SLYSFPEPE	ALYVDSLFF	SLLQHLIGL	ELTLGEFLK	MINAYLDKL
AAGIGILTV	FLPSDFFPS	SVRDRLARL	SLREWLLRI	LLSAWILTA	AAGIGILTV	AVPDEIPPL	FAYDGKDYI
AAGIGILTV	FLPSDFFPS	AAGIGILTV	FLPSDFFPS	AAGIGILTV	FLWGPRALV	ETVSEQSNV	ITLWQRPLV

Mutual information

- How is mutual information calculated?
- Information content was calculated as
  - Gives information in a single position

$$I = \sum_a p_a \log(\frac{p_a}{q_a})$$



- Similar relation for mutual information
  - Gives mutual information between two positions

$$I = \sum_{a,b} p_{ab} \log(\frac{p_{ab}}{p_a \cdot p_b})$$

Mutual information. Example

Knowing that you have G at P_1 allows you to make an educated guess on what you will find at P_6.
$P(V_6) = 4/10$. $P(V_6|G_1) = 1.0$!

$$I = \sum_{a,b} p_{ab} \log\left(\frac{p_{ab}}{p_a \cdot p_b}\right)$$

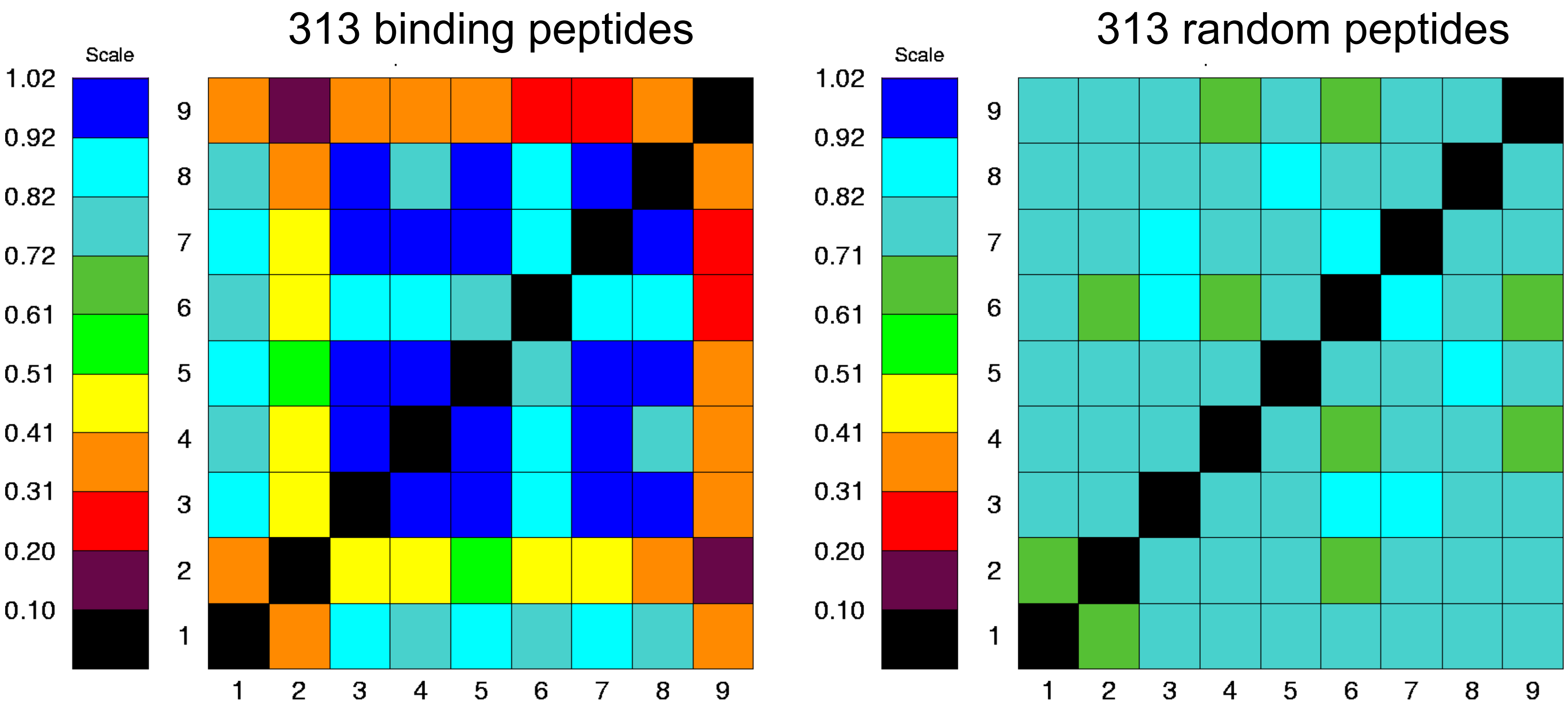
$P(G_1) = 2/10 = 0.2$, ..
$P(V_6) = 4/10 = 0.4$,..
$P(G_1,V_6) = 2/10 = 0.2$,
$P(G_1)*P(V_6) = 8/100 = 0.0.8$

$\log(0.2/0.08) > 0$

P1 P6

```
ALWGFFPVA
ILKEPVHGV
ILGFVFTLT
LLFGYPVYV
GLSPTVWLS
YMNGTMSQV
GILGFVFTL
WLSLLVPFV
FLPSDFFPS
WVPLELRDE
```

Mutual information

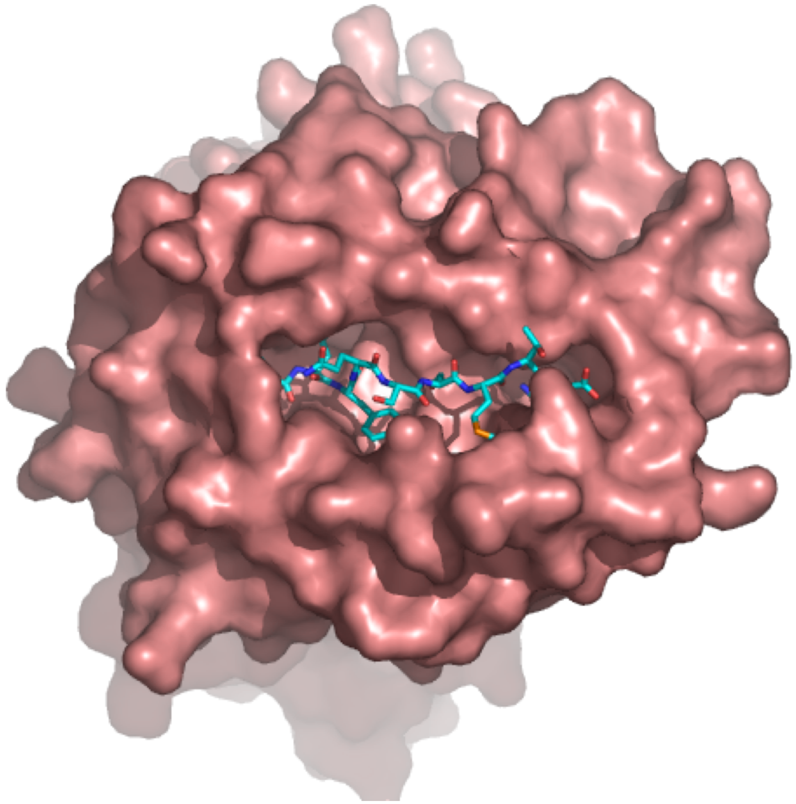


Higher order sequence correlations

- Neural networks can learn higher order correlations!
  - What does this mean?

Say that the peptide needs one and only one large amino acid in the positions P3 and P4 to fill the binding cleft

How would you formulate this to test if a peptide can bind?



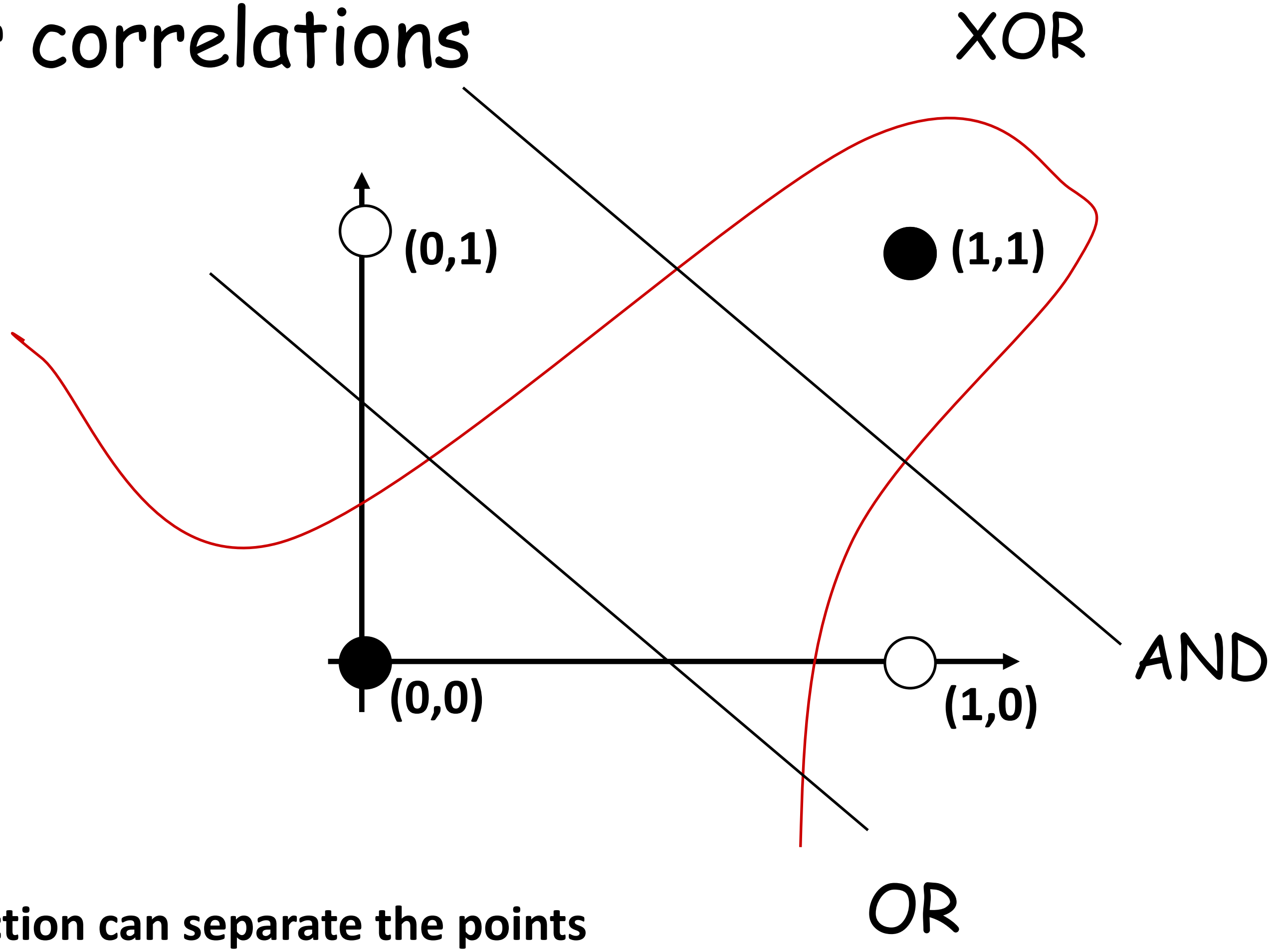
S S => 0
L S => 1
S L => 1
L L => 0

=> **XOR function**

Neural networks

- Neural networks can learn higher order correlations

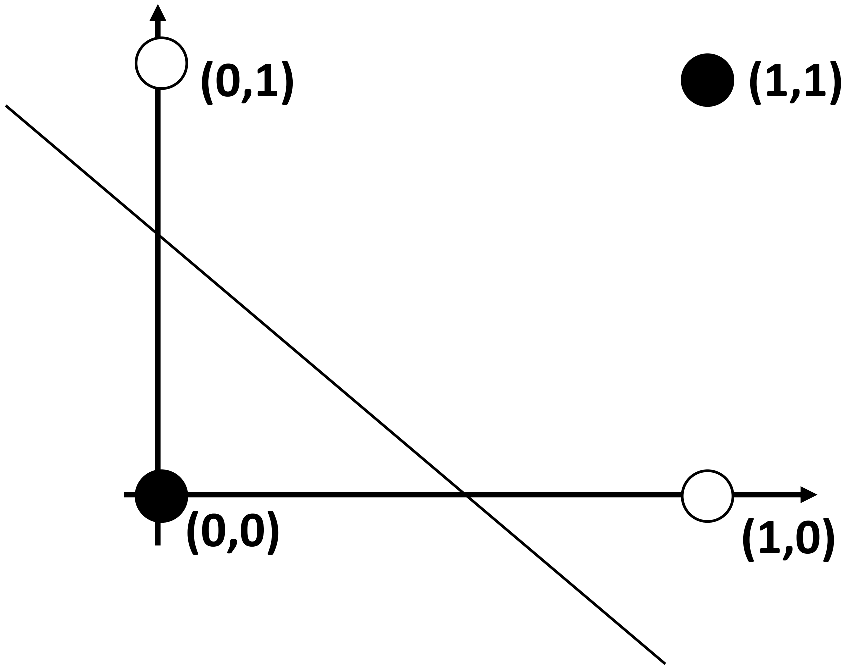
XOR function:
0 0 => 0
1 0 => 1
0 1 => 1
1 1 => 0



No linear function can separate the points

Error estimates

XOR	Predict	Error
0 0 => 0	0	0
1 0 => 1	1	0
0 1 => 1	1	0
1 1 => 0	1	1

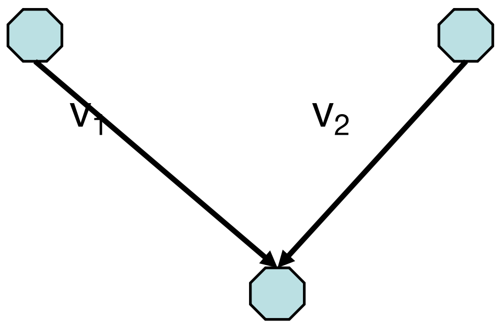


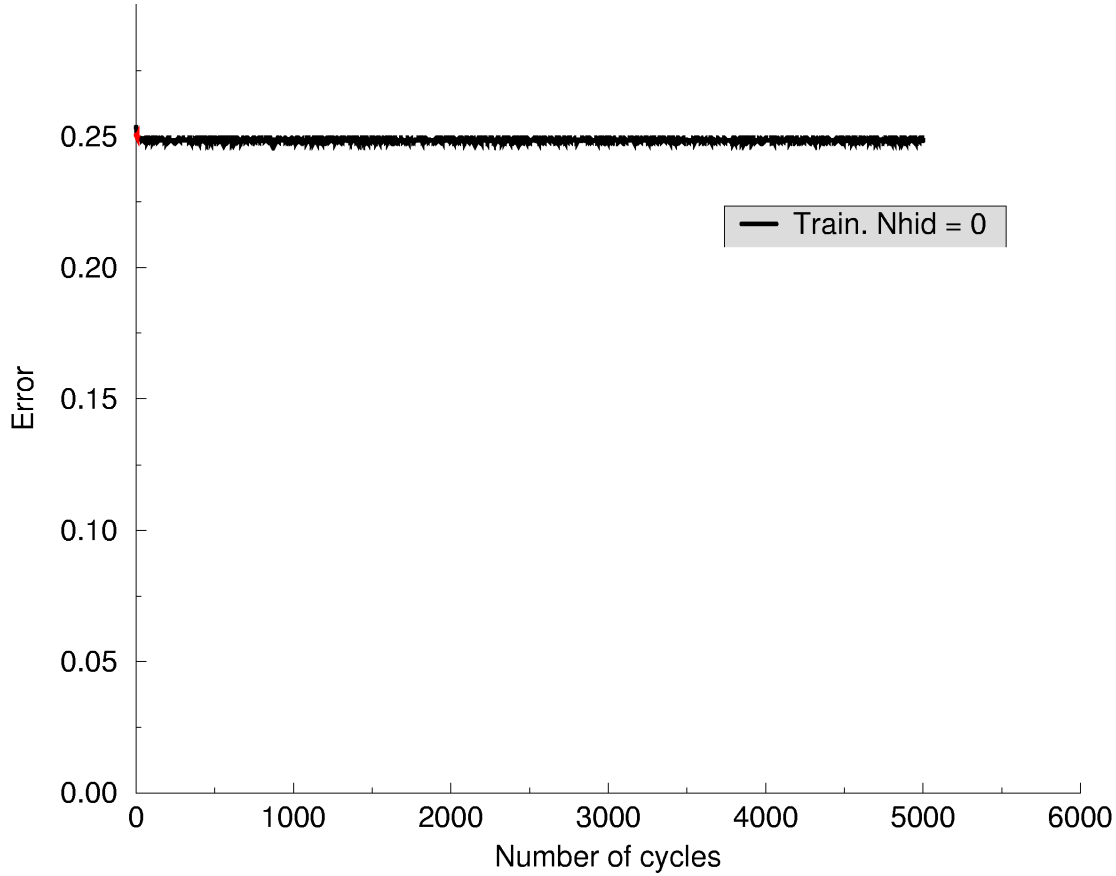
Mean error: 1/4

Neural networks

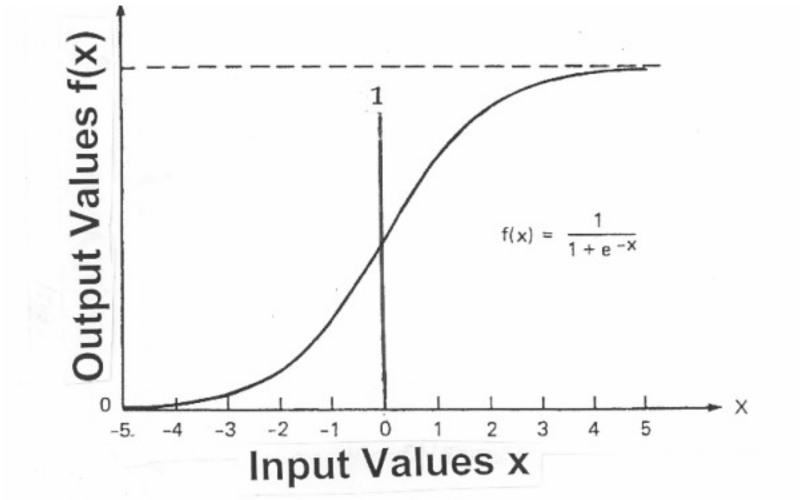
Linear function

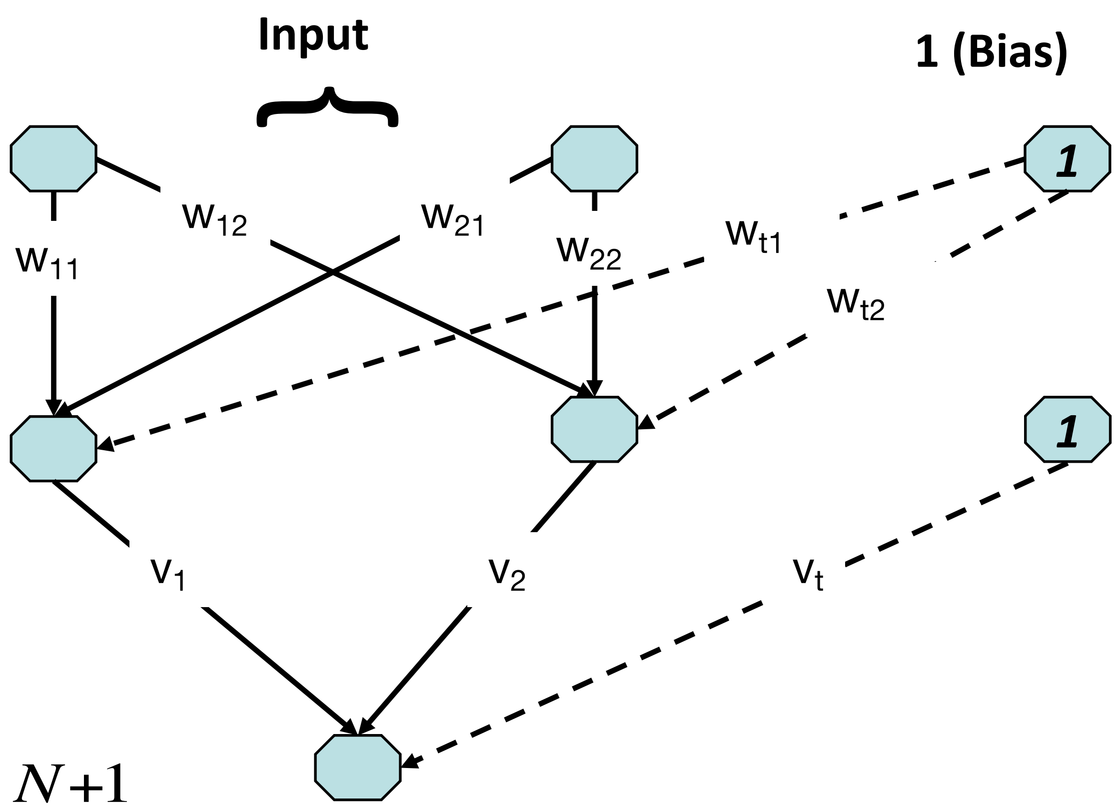
$$y = x_1 \cdot v_1 + x_2 \cdot v_2$$





Neural networks with a hidden layer



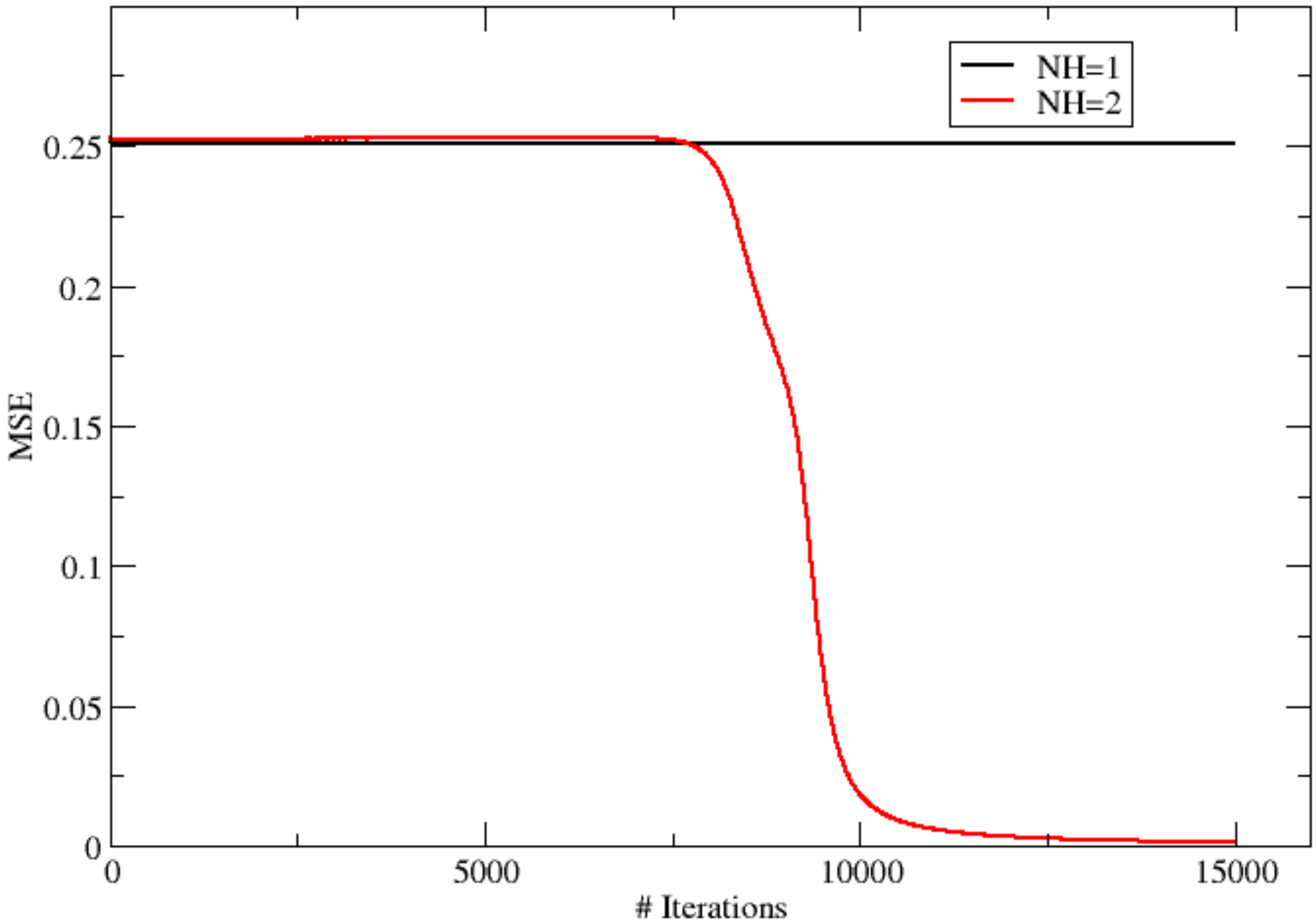


$$O = \frac{1}{1+\exp(-o)}$$

$$o = \sum_{i=1}^{N} x_i \cdot w_i + t = \sum_{i=1}^{N+1} x_i \cdot w_i$$

$$x_N = 1$$

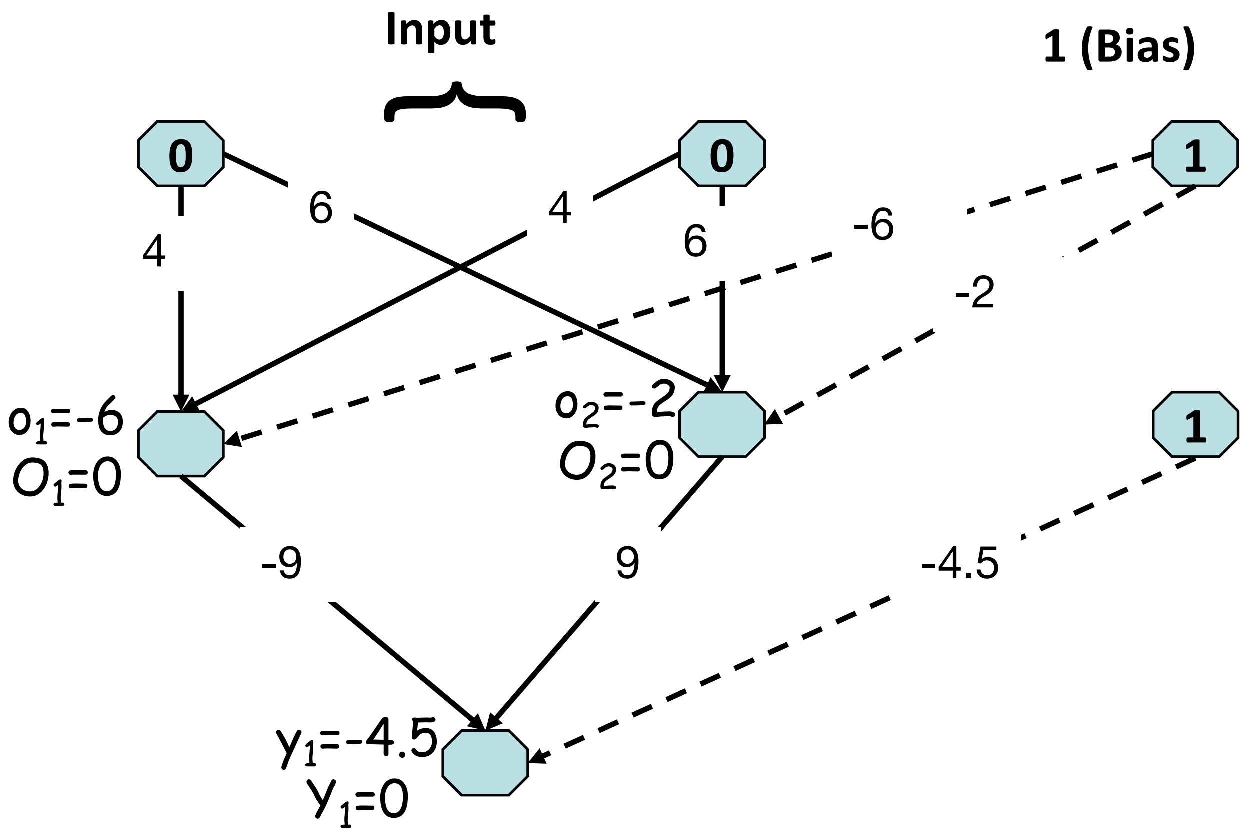
Neural networks



How does it work?
Ex. Input is (0 0)

$$O = \frac{1}{1+\exp(-o)}$$

$$o = \sum x_i \cdot w_i$$



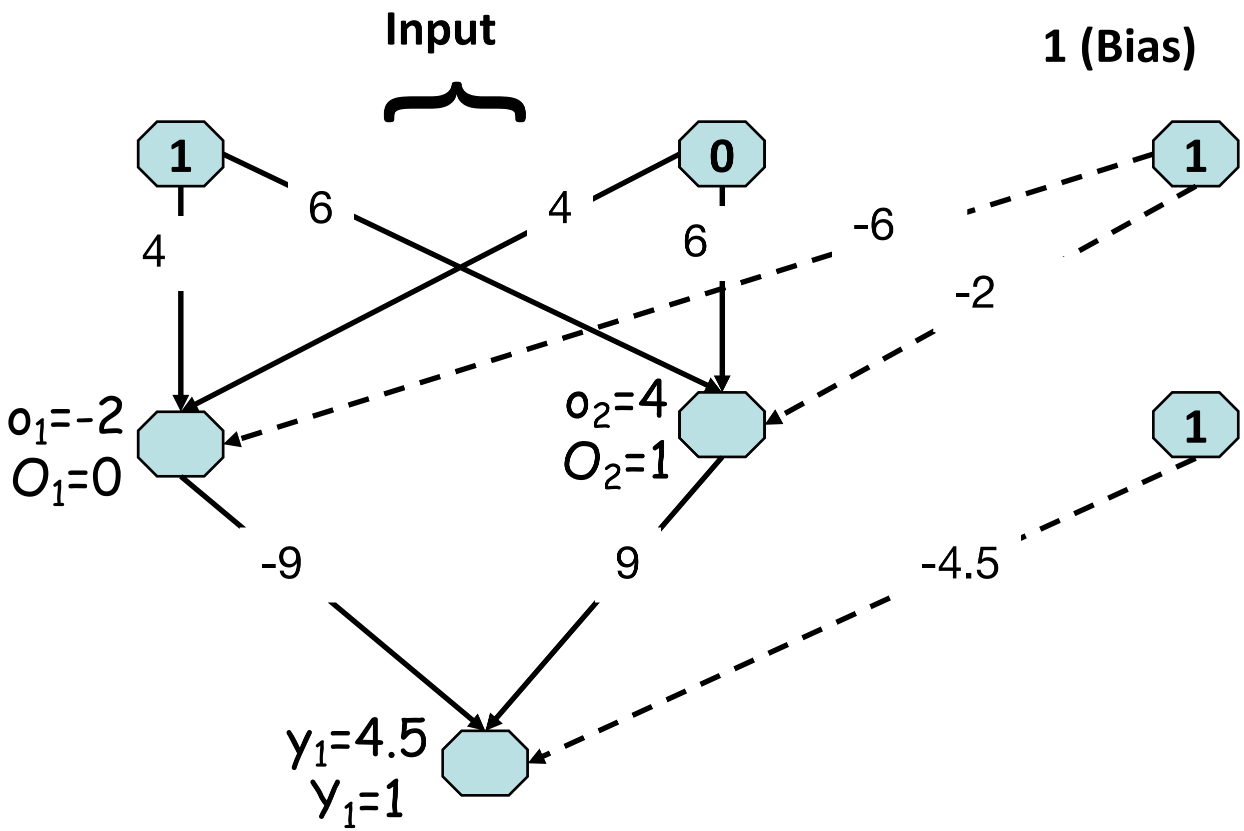
Neural networks. How does it work?

Hand out

Neural networks (1 0 && 0 1)

$$O = \frac{1}{1+\exp(-o)}$$

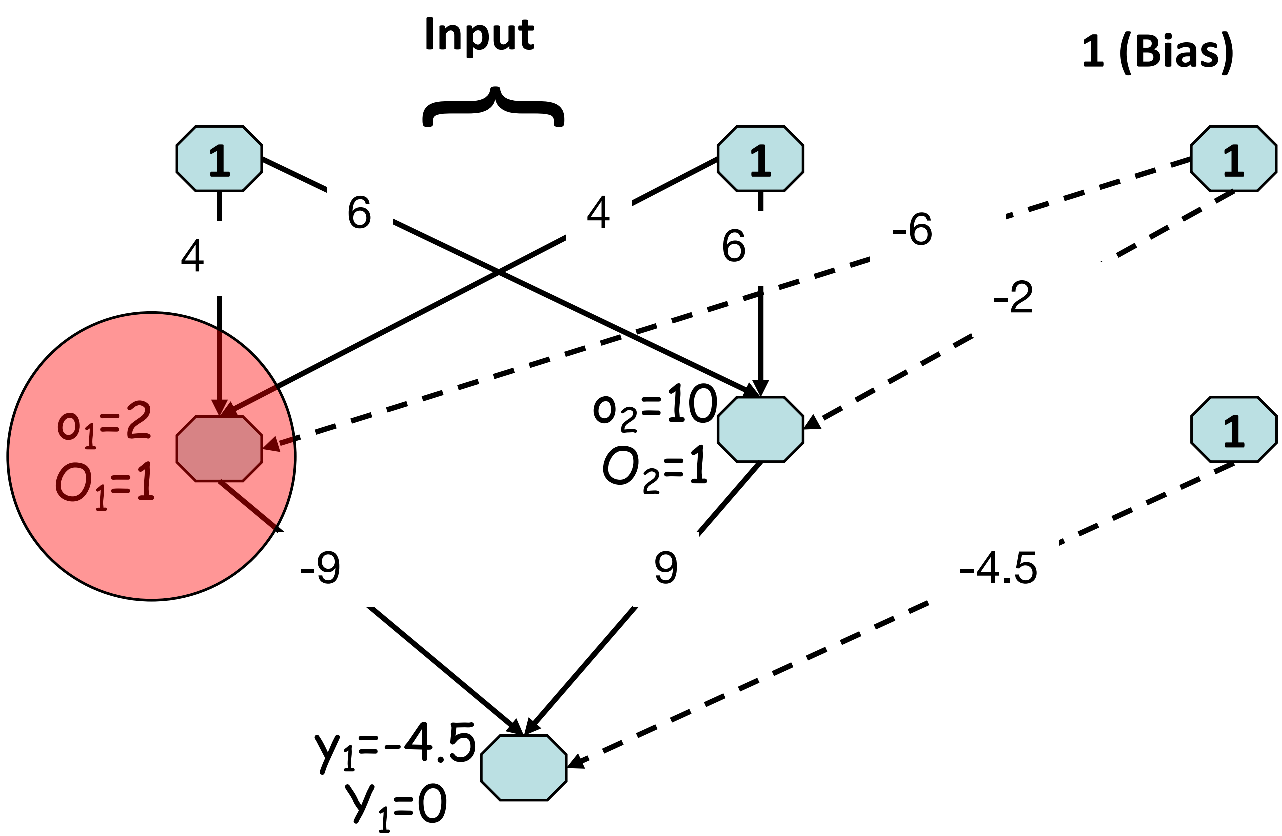
$$o = \sum x_i \cdot w_i$$



Neural networks (1 1)

$$O = \frac{1}{1+\exp(-o)}$$

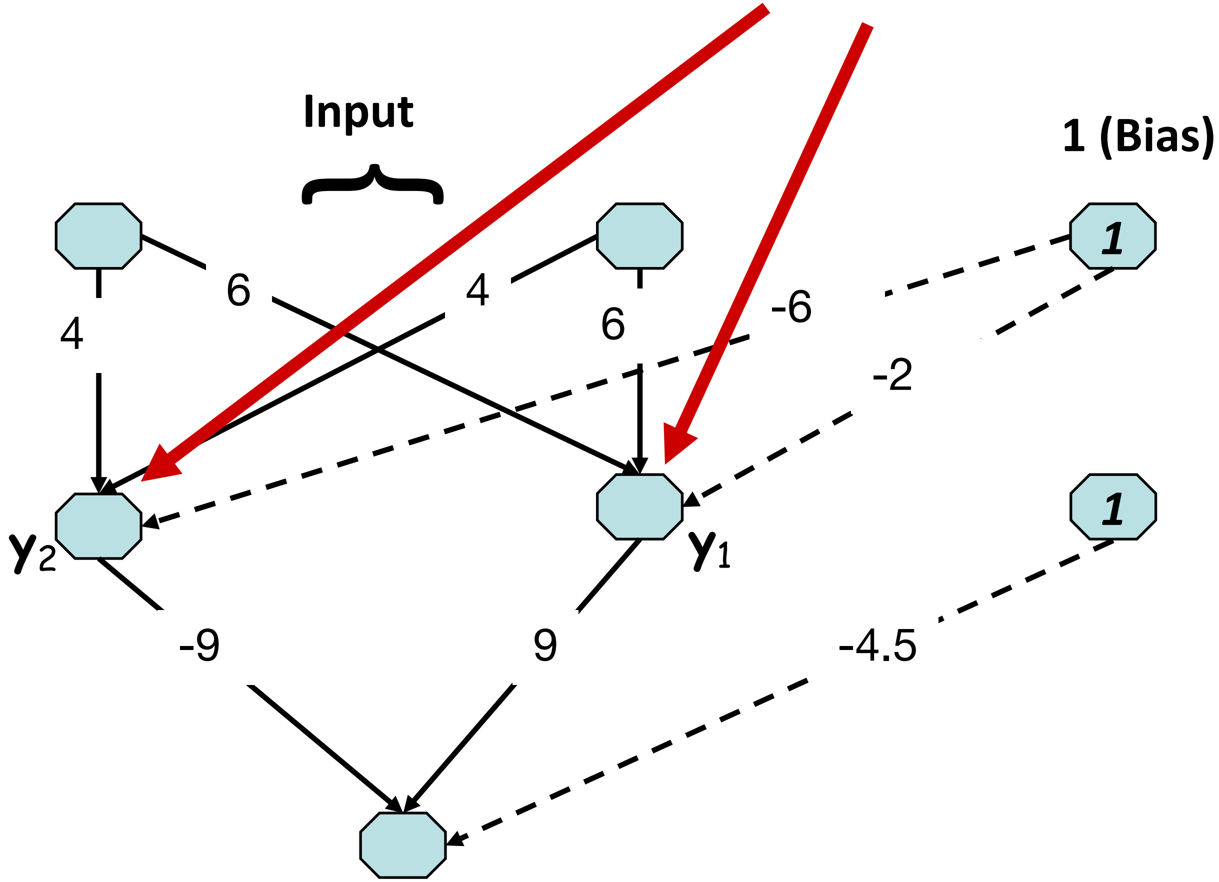
$$o = \sum x_i \cdot w_i$$



What is going on?

$$\boxed{f_{XOR}(x_1,x_2) = -2 \cdot x_1 \cdot x_2 + (x_1 + x_2) = -y_2 + y_1}$$

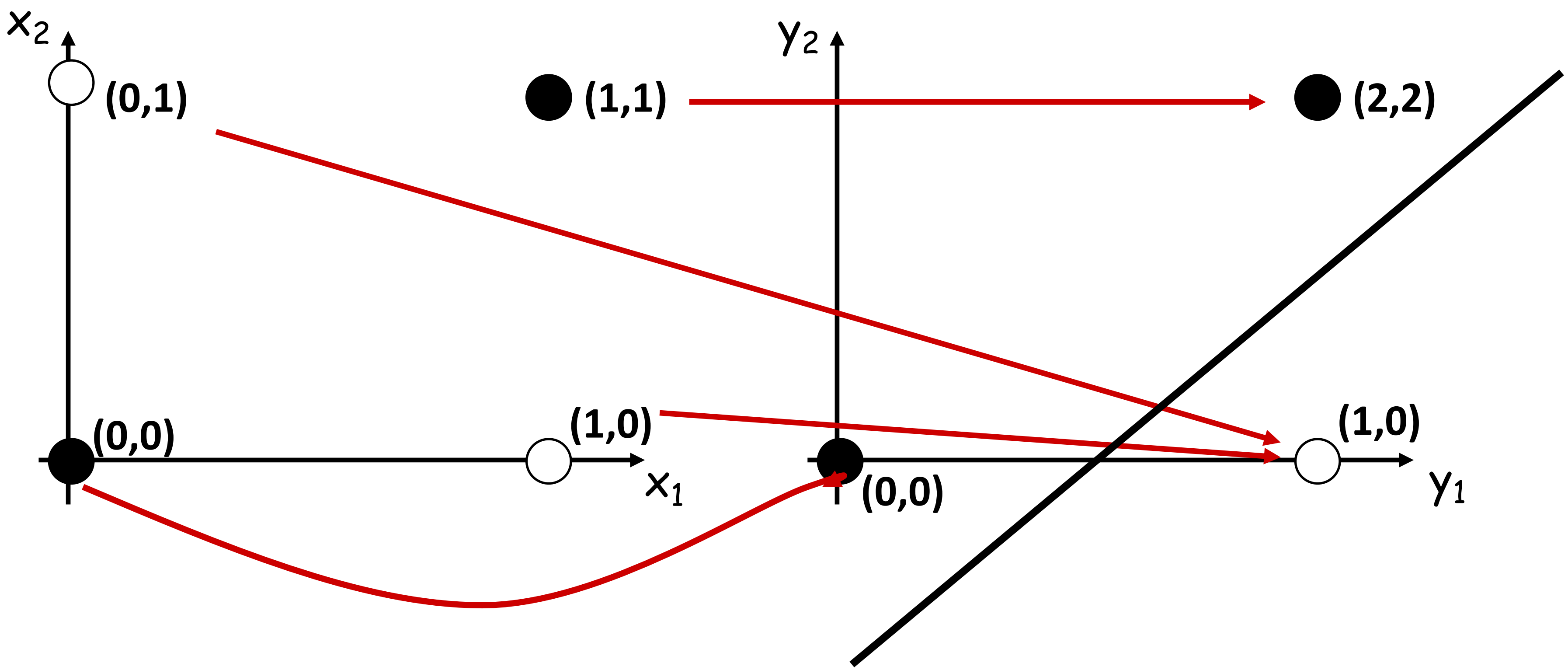
XOR function:
0 0 => 0
1 0 => 1
0 1 => 1
1 1 => 0



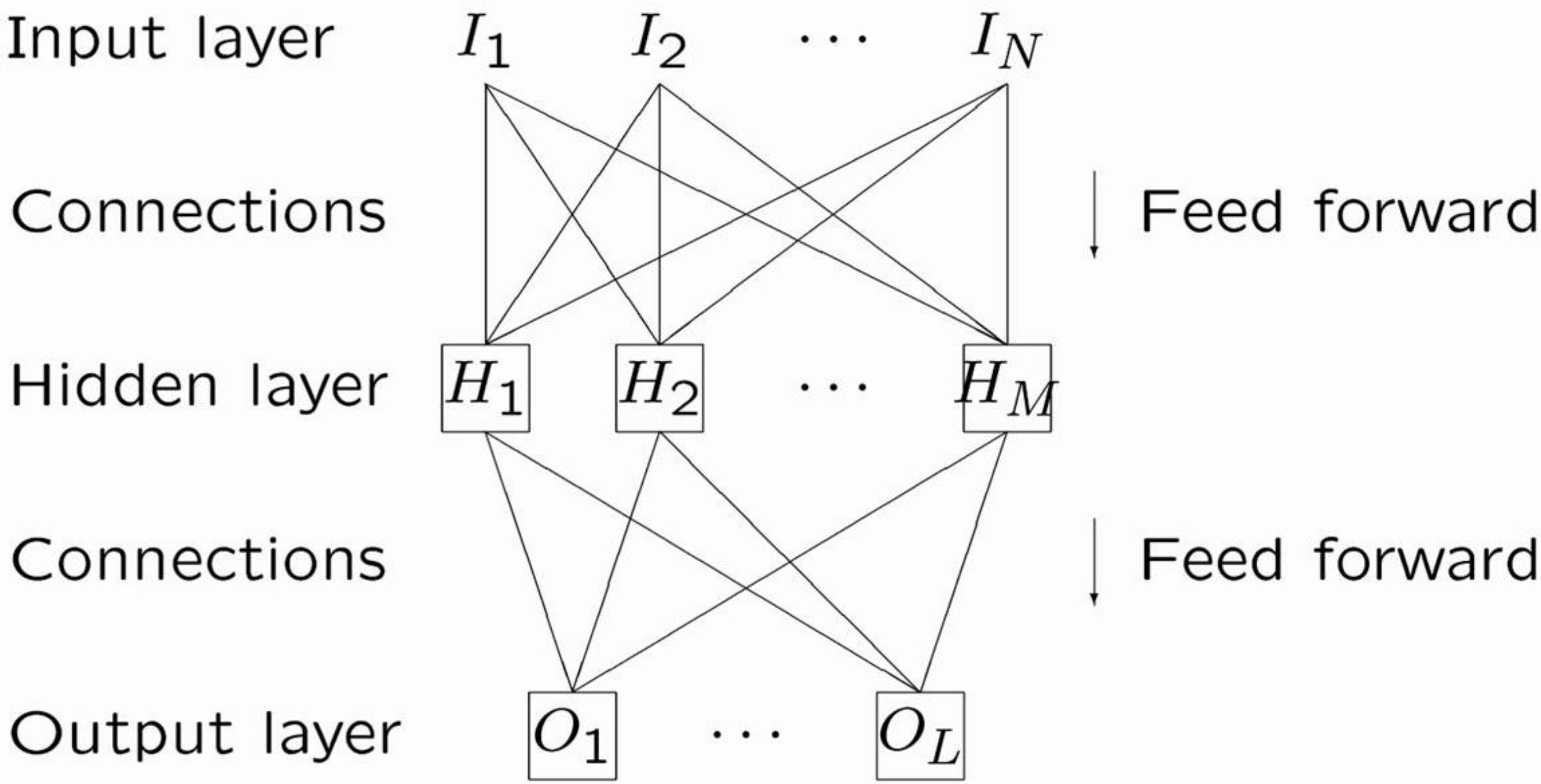
What is going on?

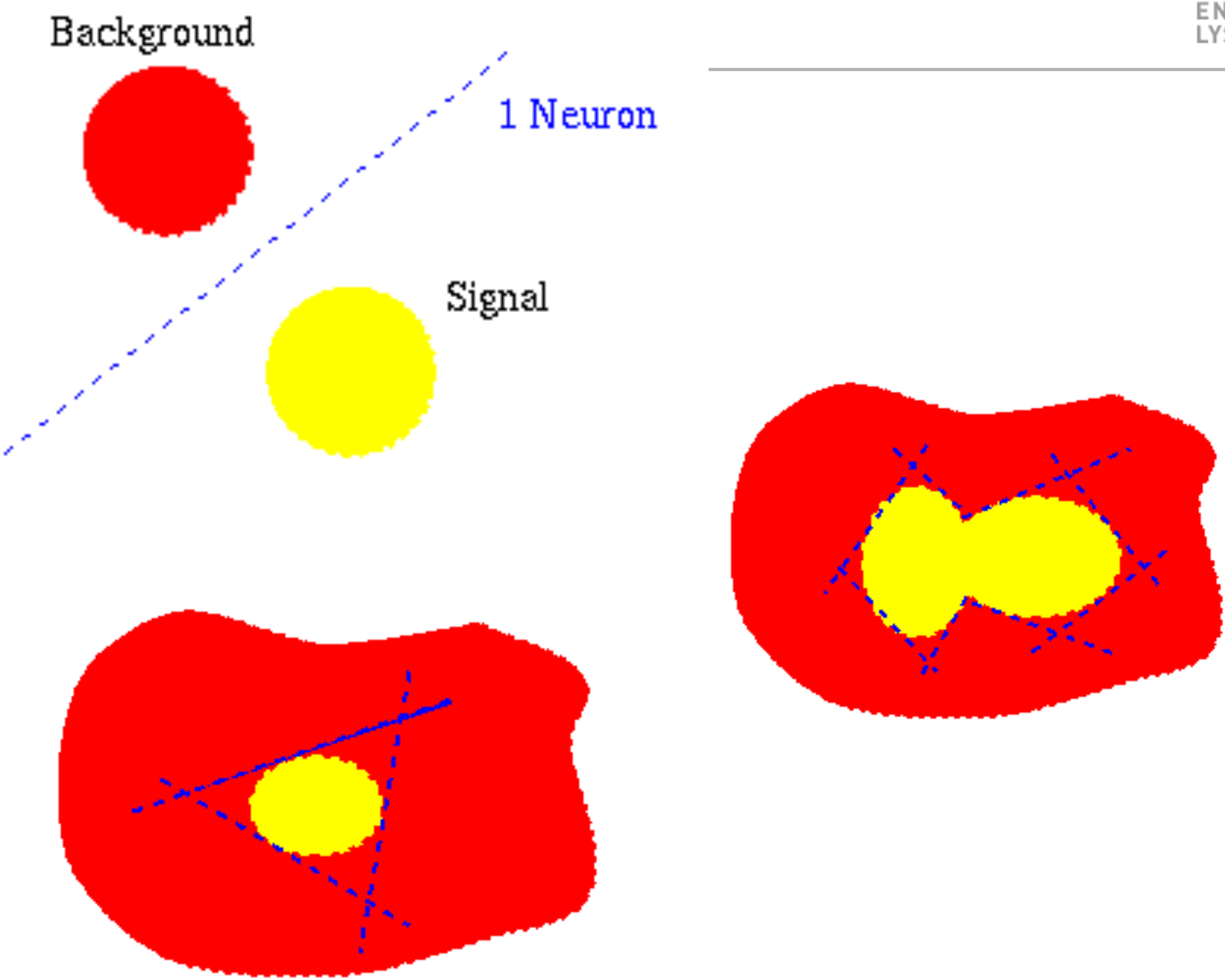
$$y_1 = x_1 + x_2$$

$$y_2 = 2 \cdot x_1 \cdot x_2$$



Network with more inputs and hidden units




Background
1 Neuron
Signal

Pattern Association

Pattern *association*.
Input is associated with output.
Classification, categorization, discrimination.

Goal: Find weights and thresholds.
Method: Training, not programming.

Training examples: I_j^α $(\alpha = 1, 2, \ldots; j = 1, 2, \ldots, N)$.

Desired targets: T_i^α $(\alpha = 1, 2, \ldots; i = 1, 2, \ldots, M)$.

Actual output: O_i^α $(\alpha = 1, 2, \ldots; i = 1, 2, \ldots, M)$.

Define quadratic error

$$E = \frac{1}{2} \sum_{\alpha, i} (O_i^\alpha - T_i^\alpha)^2$$

Measures least square deviation between desired result and actual output.

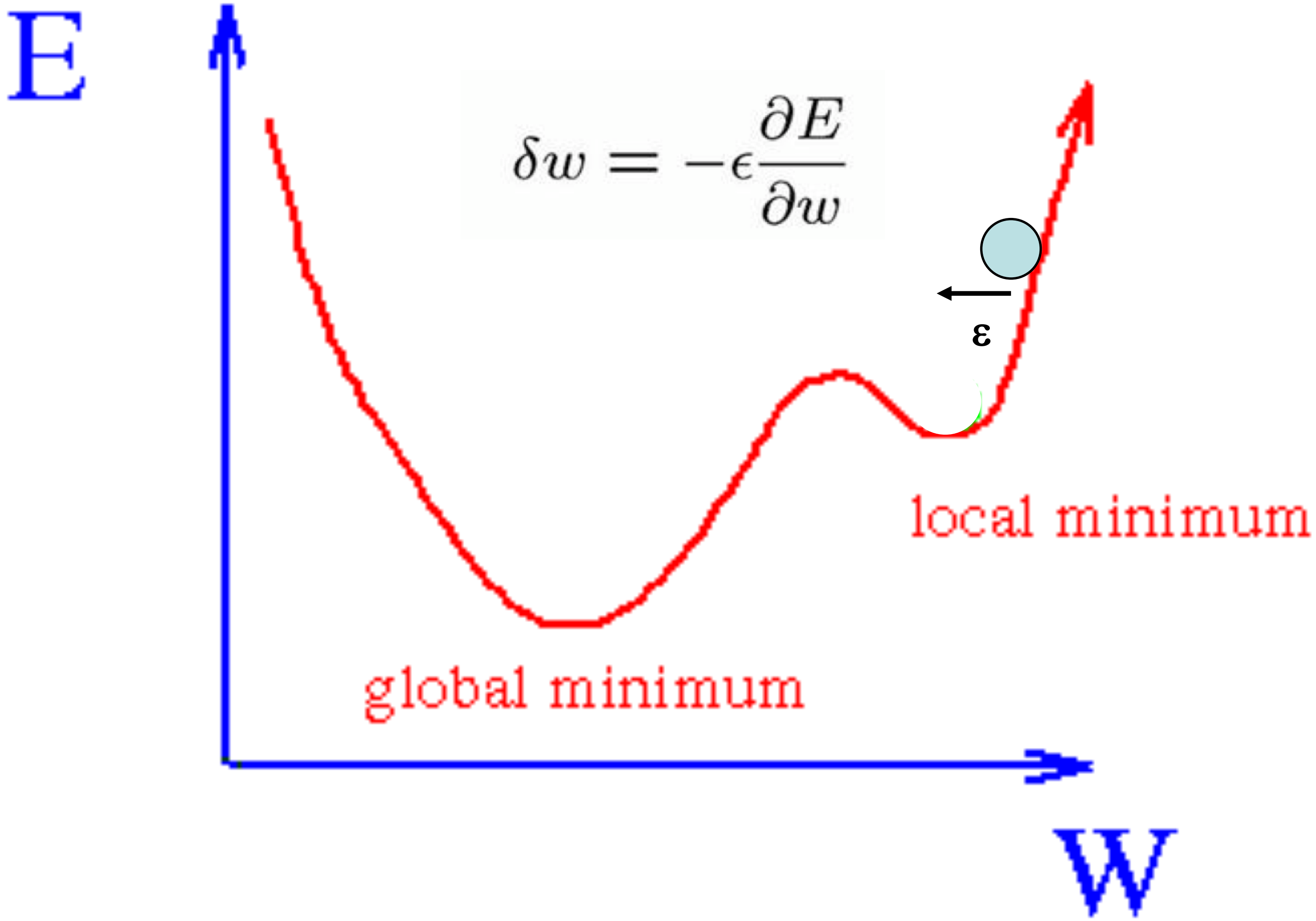
Minimize error by varying weights and thresholds.

$$\delta w = -\epsilon \frac{\partial E}{\partial w}$$

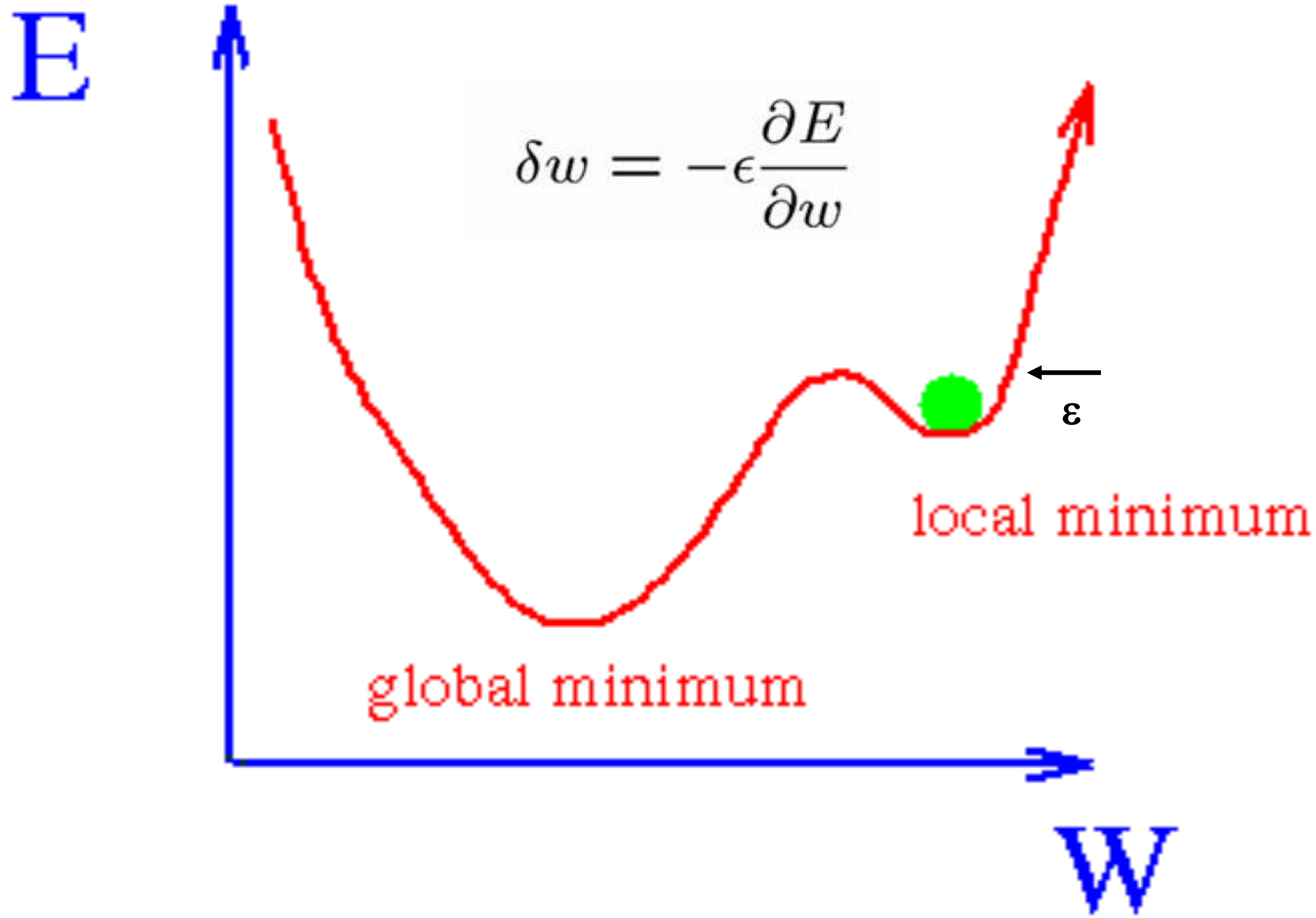
Gradient descent method.

Training and error reduction

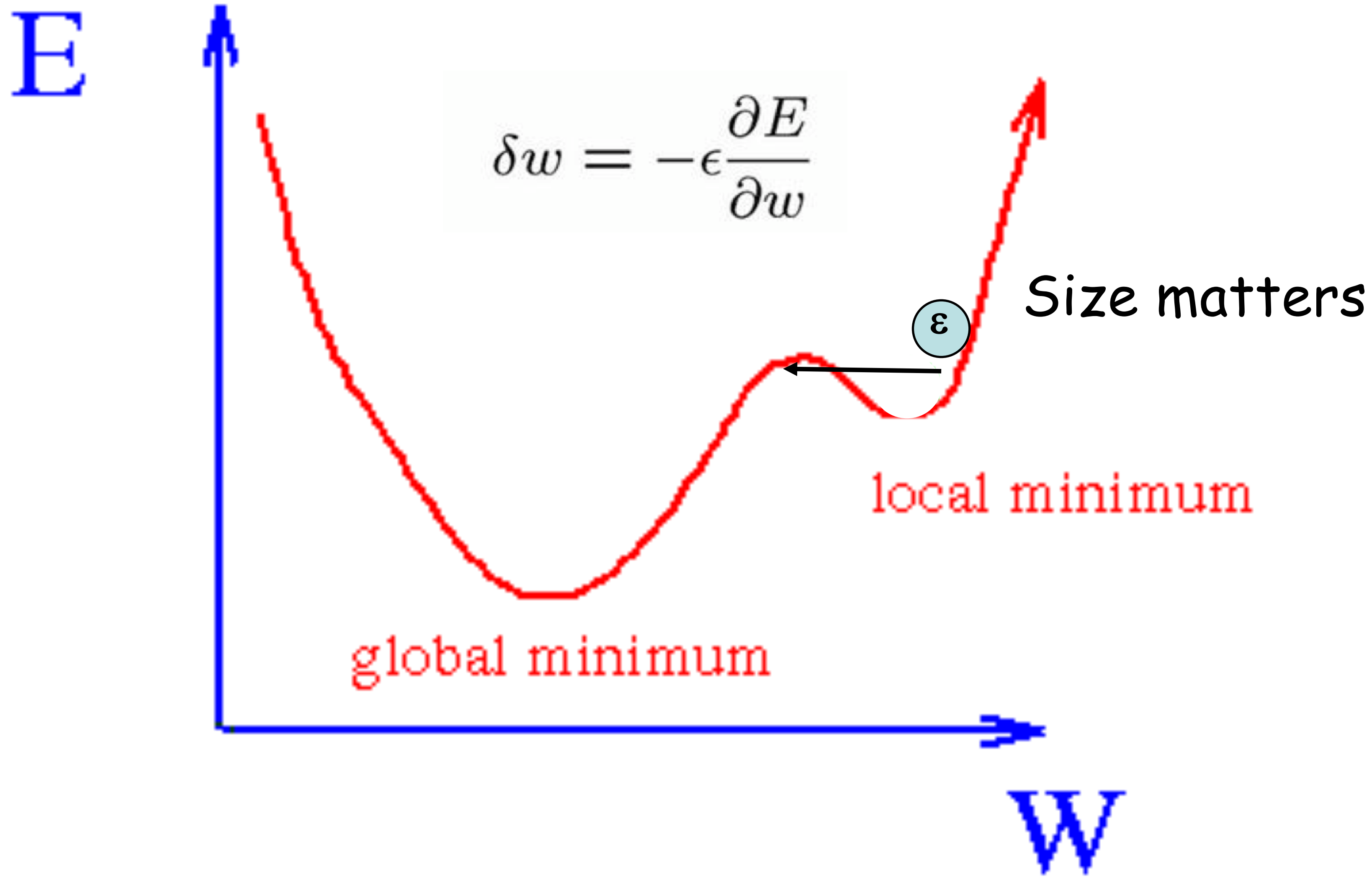
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



Training and error reduction

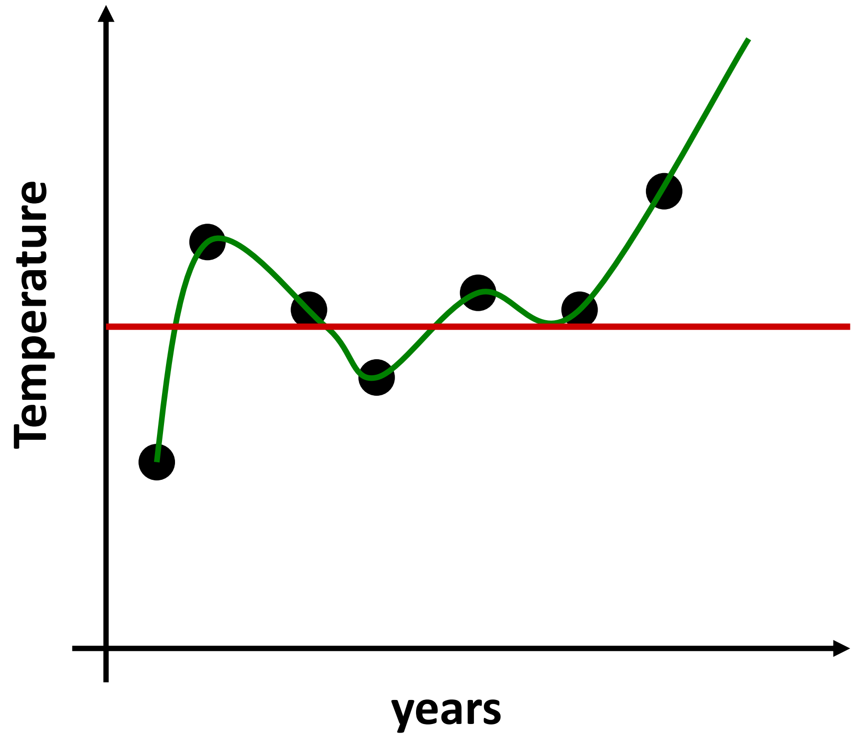


Training and error reduction



Neural network training

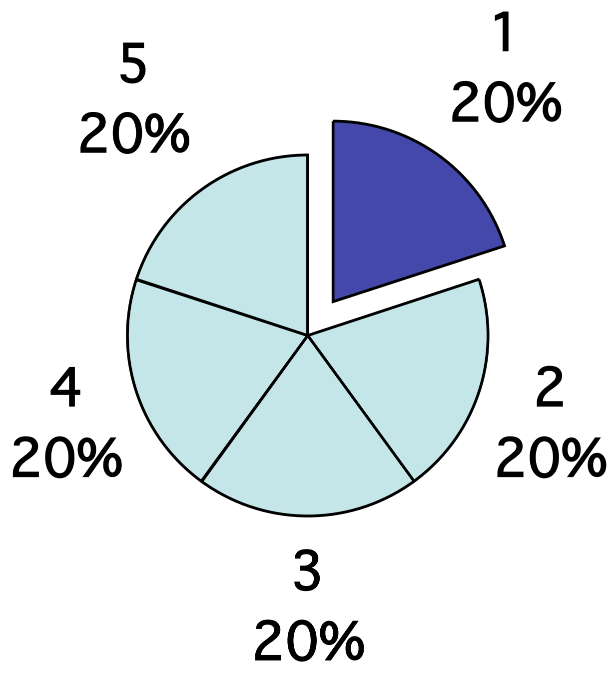
- A Network contains a very large set of parameters
  - A network with 5 hidden neurons predicting binding for 9meric peptides has 9x20x5=900 weights
  - 5 times as many weights as a matrix-based method
- Over fitting is a problem
- Stop training when test performance is optimal (use early stopping)



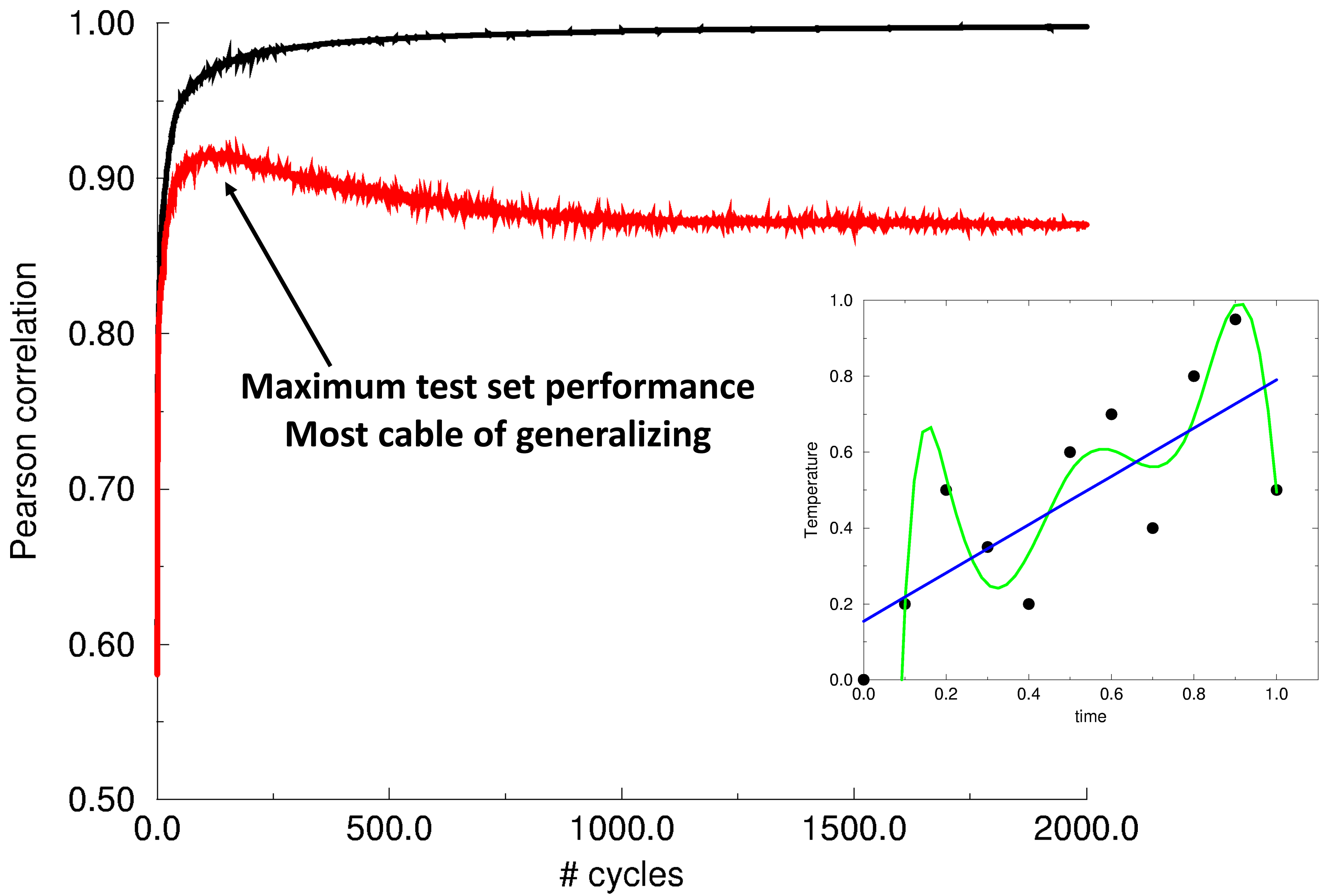
Neural network training. Cross validation

Cross validation

Train on 4/5 of data
Test on 1/5
=>
Produce 5 different neural networks each with a different prediction focus



Neural network training curve



Demo

Network training

- Encoding of sequence data
  - Sparse encoding
  - Blosum encoding
  - Sequence profile encoding

Sparse encoding

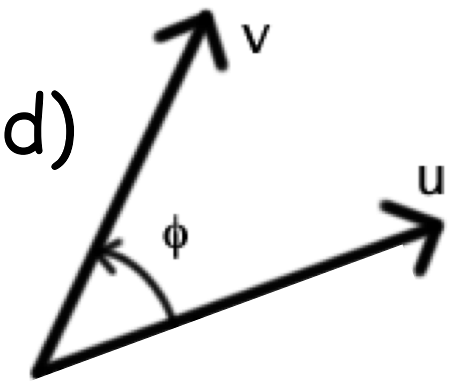
Inp Neuron	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
AAcid																				
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

BLOSUM encoding (Blosum50 matrix)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Sequence encoding (continued)

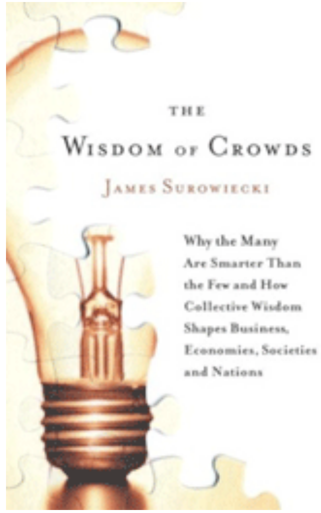
- Sparse encoding
  - `V:0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1`
  - `L:0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0`
  - V·L=0 (unrelated)
- Blosum encoding
  - `V: 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2  0 -3 -1 4`
  - `L:-1 -2 -3 -4 -1 -2 -3 -4 -3 2 4 -2 2  0 -3 -2 -1 -2 -1 1`
  - V·L = 0.88 (highly related)
  - V·R = -0.08 (close to unrelated)



The Wisdom of the Crowds

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

- The Wisdom of Crowds. Why the Many are Smarter than the Few. James Surowiecki

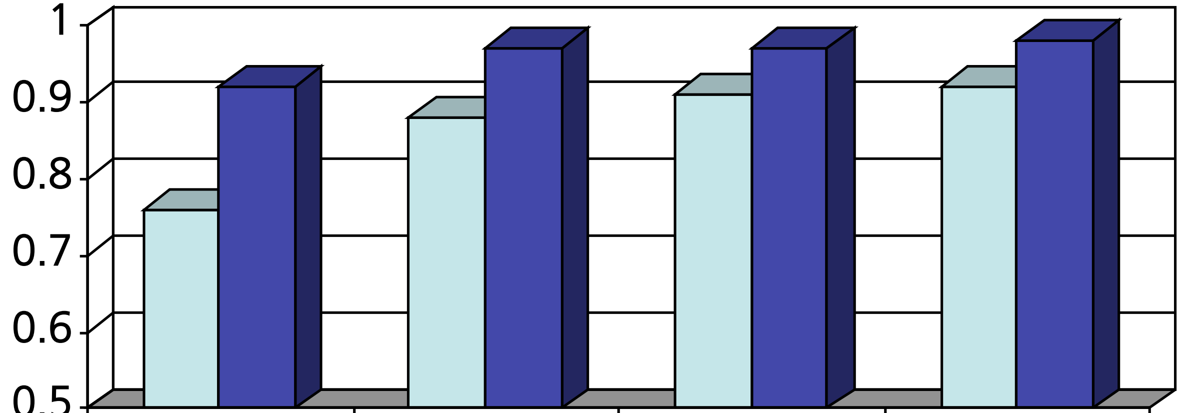


One day in the fall of 1906, the British scientist Fracis Galton left his home and headed for a country fair… *He believed that only a very few people had the characteristics necessary to keep societies healthy. He had devoted much of his career to measuring those characteristics, in fact, in order to prove that the vast majority of people did not have them.* … Galton came across a weight-judging competition…Eight hundred people tried their luck. They were a diverse lot, butchers, farmers, clerks and many other no-experts…The crowd had guessed … 1.197 pounds, the ox weighted *1.198*

Network ensembles

- No one single network with a particular architecture and sequence encoding scheme, will constantly perform the best
- Also for Neural network predictions will enlightened despotism fail
  - For some peptides, BLOSUM encoding with a four neuron hidden layer can best predict the peptide/MHC binding, for other peptides a sparse encoded network with zero hidden neurons performs the best
  - Wisdom of the Crowd
    - Never use just one neural network
    - Use Network ensembles

Evaluation of prediction accuracy



	Motif	Sparse	BLOSUM	ENS
Pear	0.76	0.88	0.91	0.92
Aroc	0.92	0.97	0.97	0.98

ENS: Ensemble of neural networks trained using sparse, Blosum, and weight matrix sequence encoding

Applications of artificial neural networks

- Talk recognition
- Prediction of protein secondary structure
- Prediction of Signal peptides
- Post translation modifications
  - Glycosylation
  - Phosphorylation
- Proteasomal cleavage
- MHC:peptide binding

NETtalk

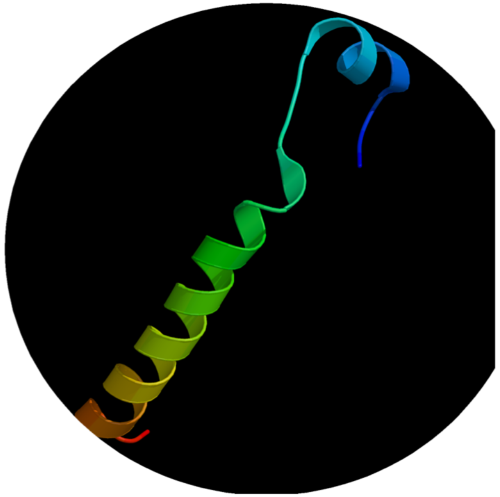
(T. Sejnowski and C. Rosenberg, 1987)

M**a**ry h**a**d **a** little l**a**mb

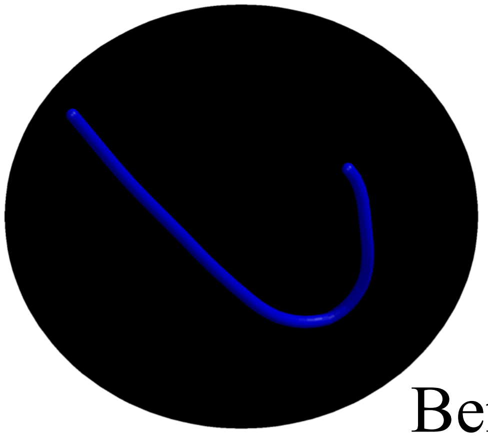
Three of the **a**'s must be pronounced differently! Reading aloud is a *context sensitive* cognitive skill.

Prediction of protein secondary structure

ß-strand



Helix

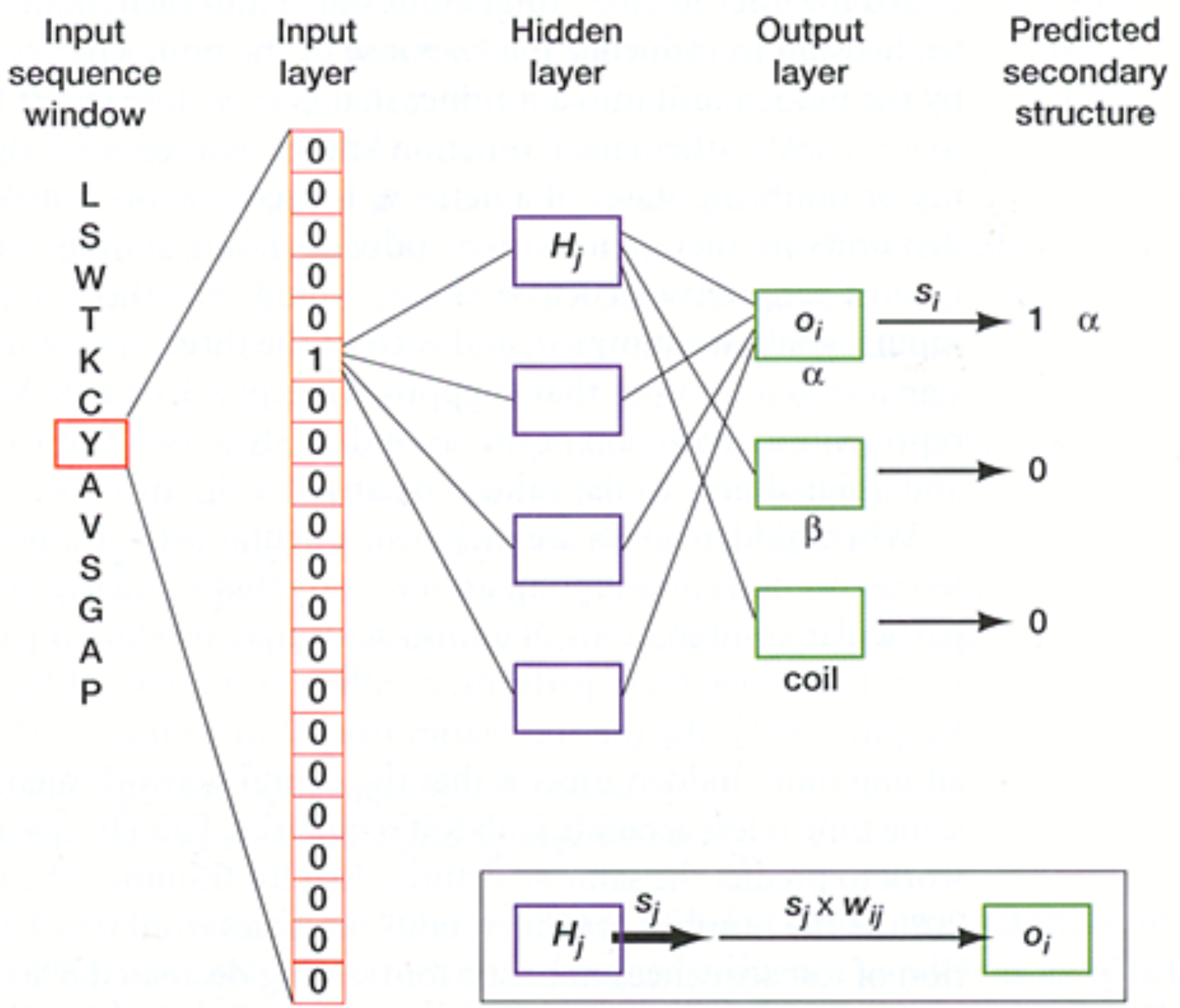


Bend

Turn

Sparse encoding of amino acid sequence windows

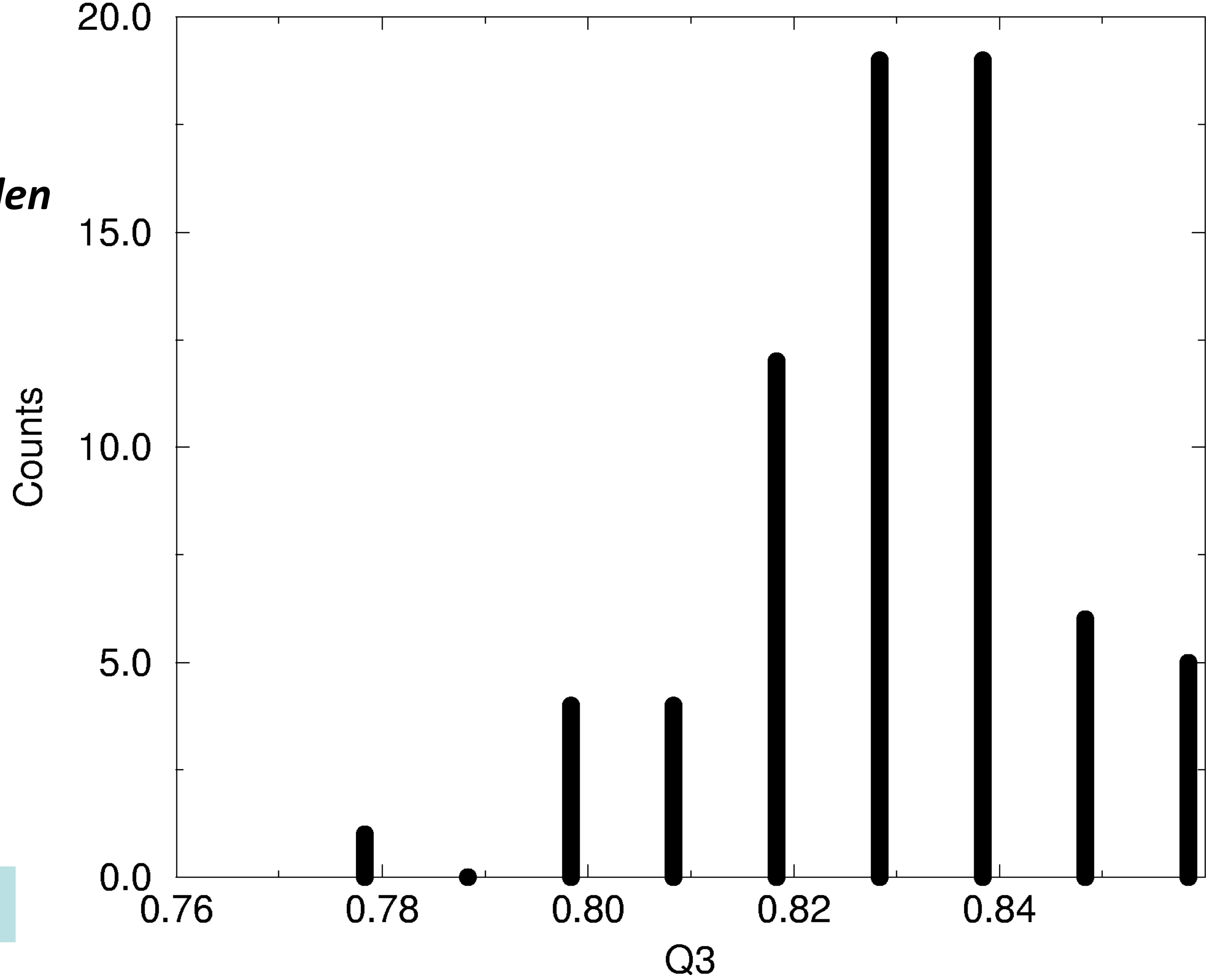
CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS



Why use network ensembles?

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

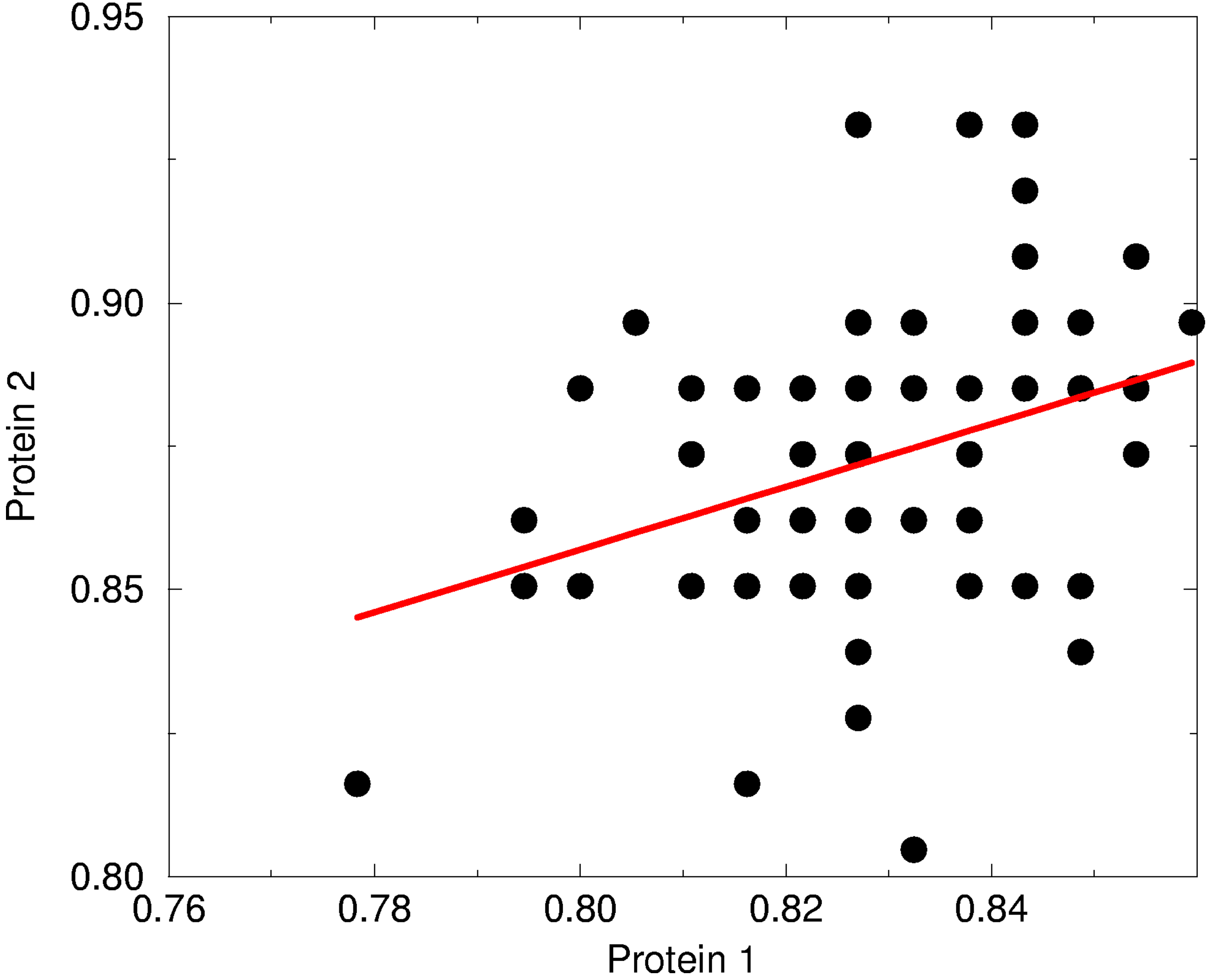
Network ensemble with 70 networks each trained with different data, number of hidden neurons, or initial weight configurations



Q3 is the overall accuracy

Why not select the best?

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS



What have we learned?

- Neural networks are not so bad as their reputation
- Neural networks can deal with higher order correlations
- Be careful when training a neural network
  - Over-fitting is an important issue
  - Always use cross validated training