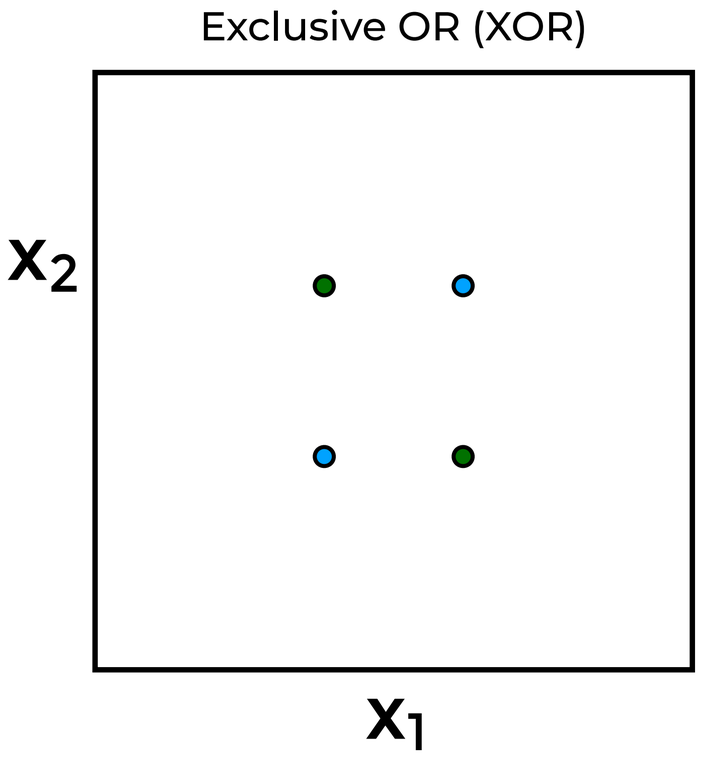
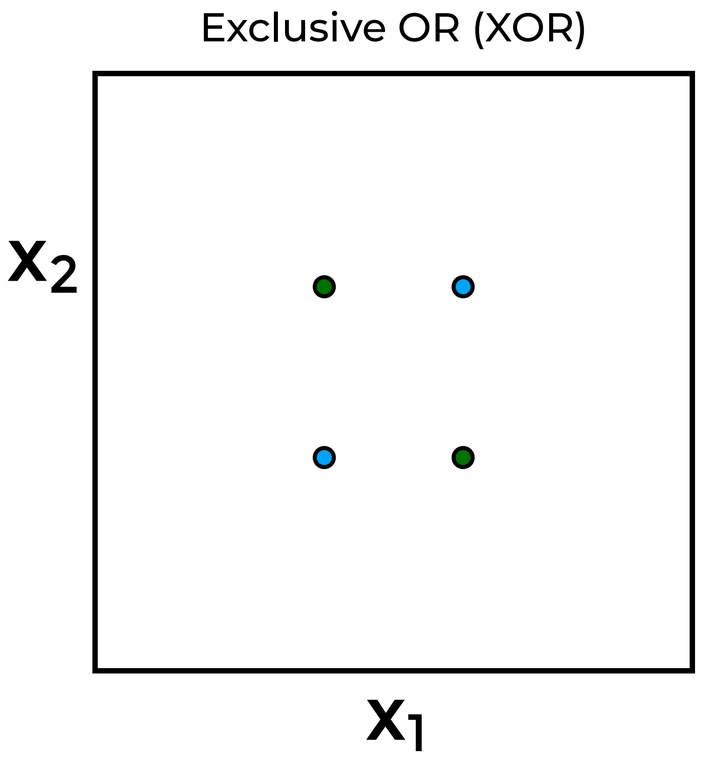
Machine Learning I
60629A

Summary
Neural Networks
— Week #5

What if data is not linearly separable?



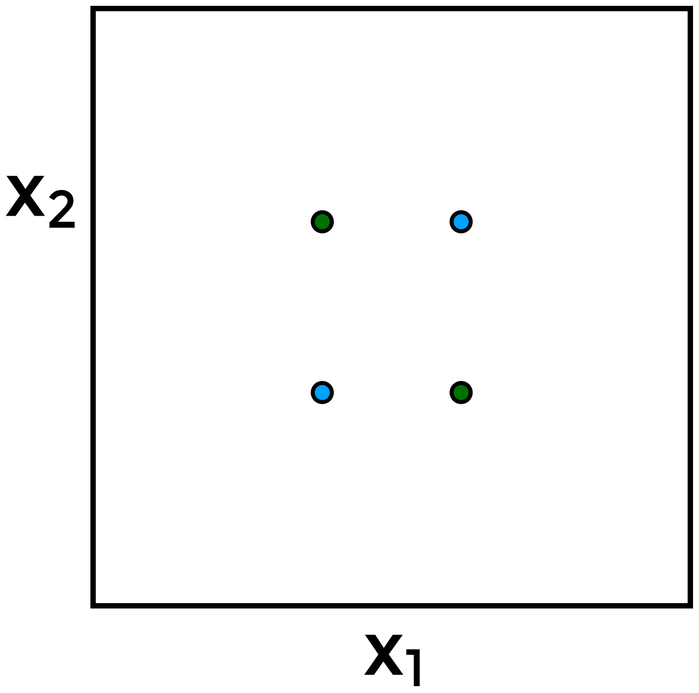
What if data is not linearly separable?



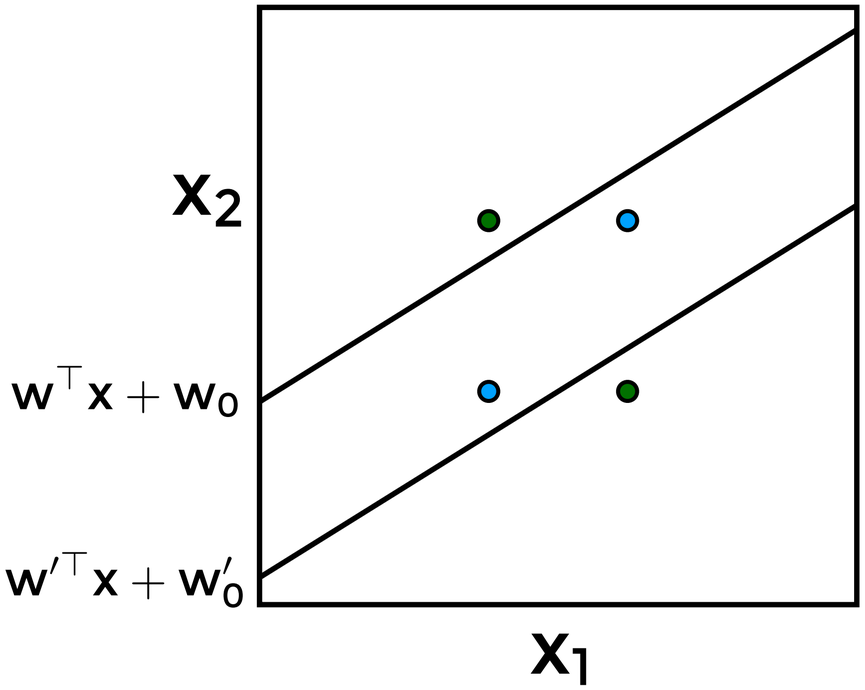
Use the joint decision of several linear classifier?

What if data is not linearly separable?

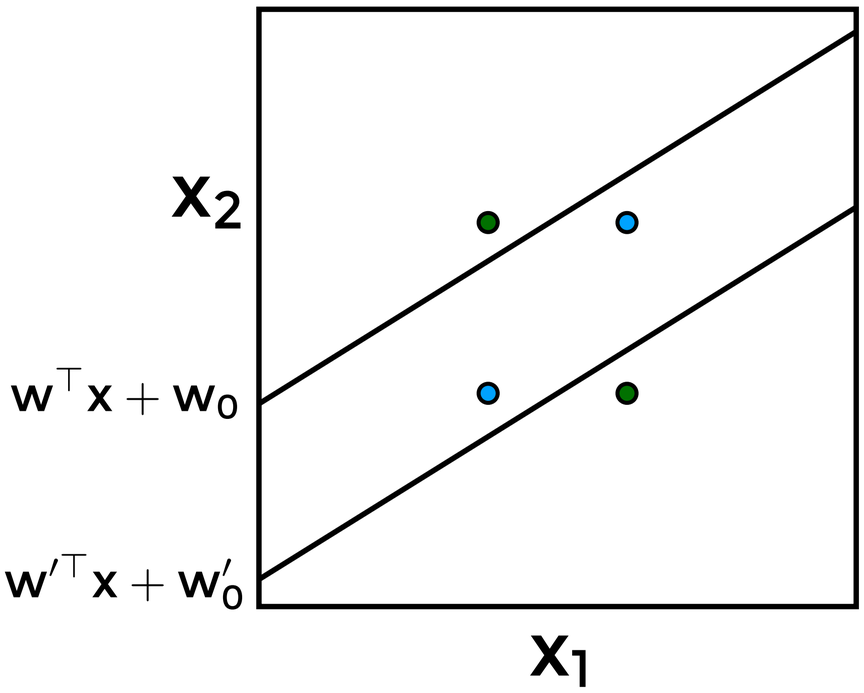
Exclusive OR (XOR)



Use the joint decision of several linear classifier?



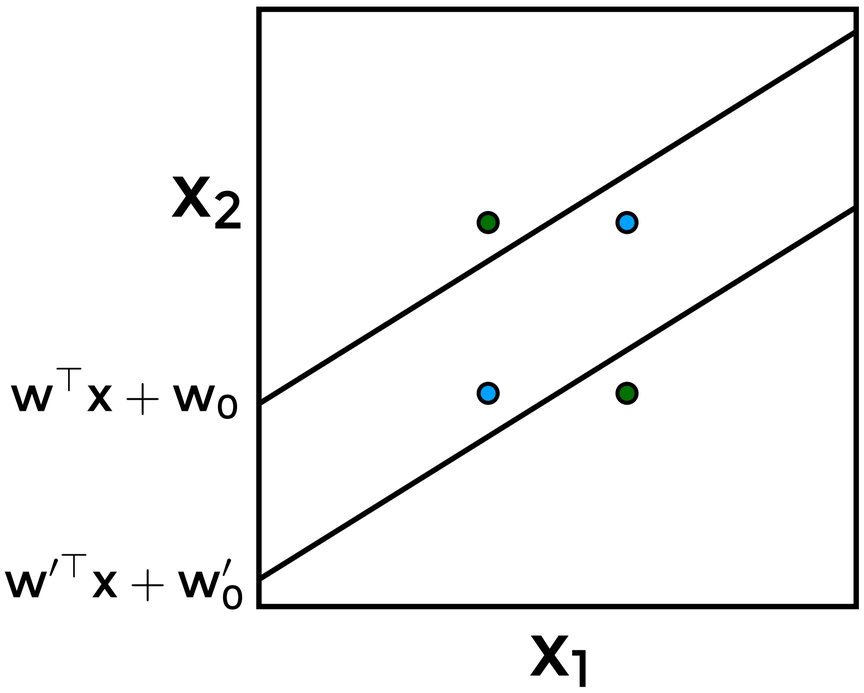
Combining models



$$f(x): \begin{cases} (w^\top x + w_0) > 0 \implies \\ (w^\top x + w_0) < 0 \implies \end{cases}$$

$$f'(x): \begin{cases} (w'^\top x + w'_0) > 0 \implies \\ (w'^\top x + w'_0) < 0 \implies \end{cases}$$

Combining models

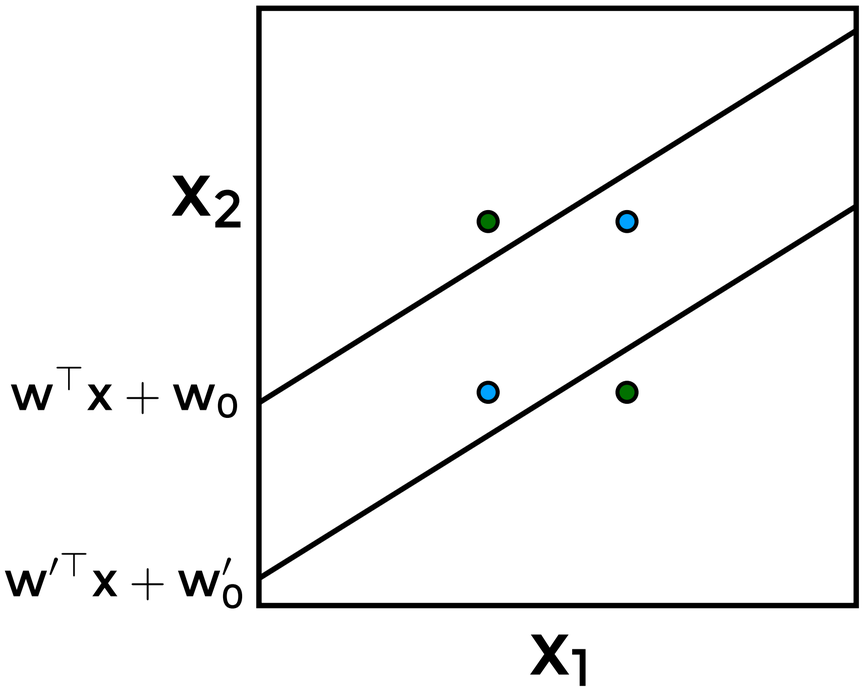


$f(x)$: $(w^\top x + w_0) > 0 \implies$ ●
$(w^\top x + w_0) < 0 \implies$ ●

$f'(x)$: $(w'^\top x + w'_0) > 0 \implies$ ●
$(w'^\top x + w'_0) < 0 \implies$ ●

$f(x) =$ ● and $f'(x) =$ ● $\implies$ ●
$f(x) =$ ● and $f'(x) =$ ● $\implies$ ●
$f(x) =$ ● and $f'(x) =$ ● $\implies$ ●

Combining models



$f(x)$: $(w^\top x + w_0) > 0 \implies$ ●
$(w^\top x + w_0) < 0 \implies$ ●

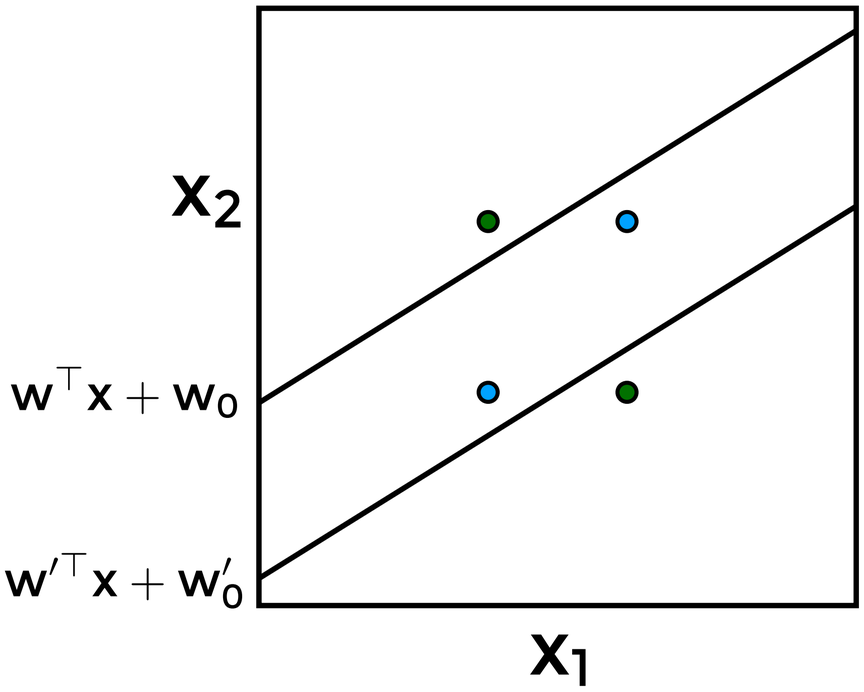
$f'(x)$: $(w'^\top x + w'_0) > 0 \implies$ ●
$(w'^\top x + w'_0) < 0 \implies$ ●

1. Evaluate each model

$f(x) =$ ● and $f'(x) =$ ● $\implies$ ●
$f(x) =$ ● and $f'(x) =$ ● $\implies$ ●
$f(x) =$ ● and $f'(x) =$ ● $\implies$ ●

2. Combine the output of models

Combining models



$$f(x): \begin{cases} (w^\top x + w_0) > 0 \implies \bullet \\ (w^\top x + w_0) < 0 \implies \bullet \end{cases}$$

$$f'(x): \begin{cases} (w'^\top x + w'_0) > 0 \implies \bullet \\ (w'^\top x + w'_0) < 0 \implies \bullet \end{cases}$$

1. Evaluate each model

$f(x) = \bullet$ and $f'(x) = \bullet \implies \bullet$

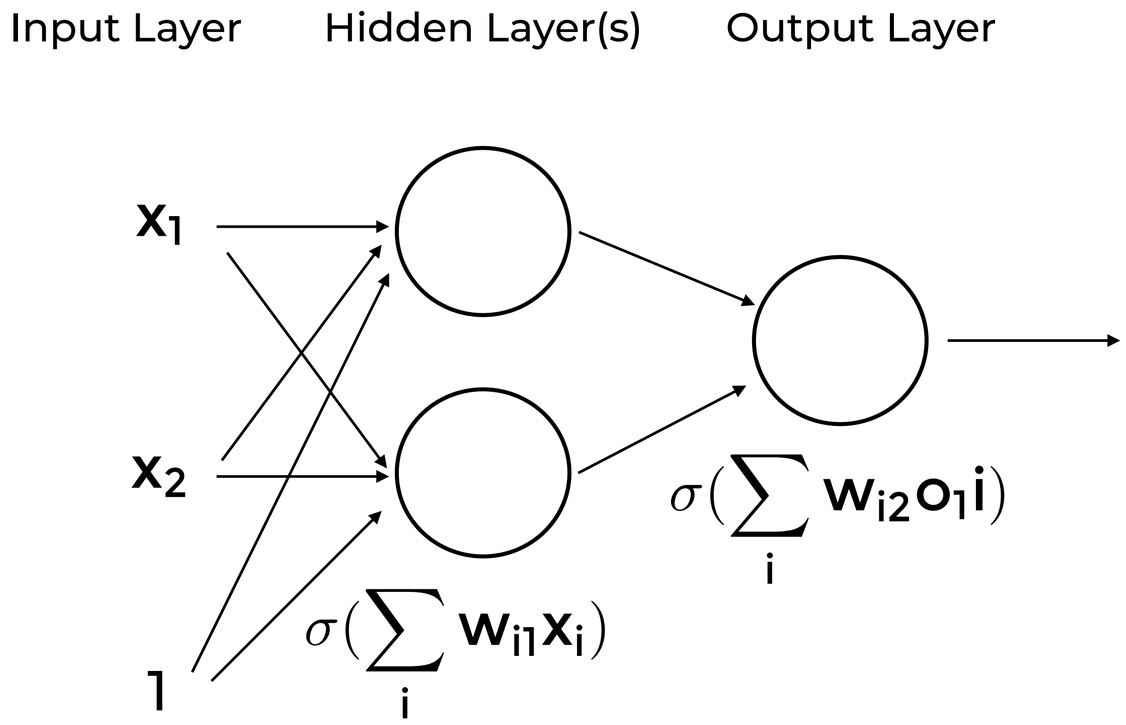
$f(x) = \bullet$ and $f'(x) = \bullet \implies \bullet$

$f(x) = \bullet$ and $f'(x) = \bullet \implies \bullet$

2. Combine the output of models

$$f''(x) = \text{threshold}(w''^\top \begin{bmatrix} f(x) \\ f'(x) \end{bmatrix} + w''_0)$$

Feed-forward neural network



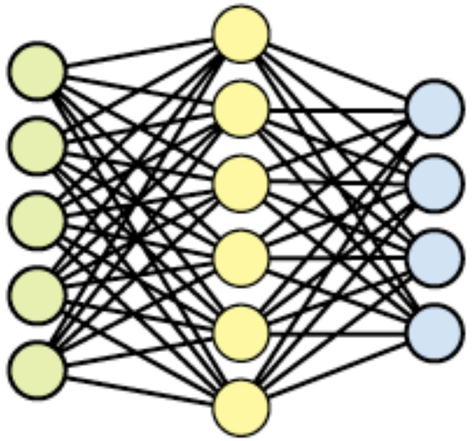
- Each arrow denotes a connection
  - A signal associated with a weight
- Each node is the weighted sum of its input followed by a non-linear activation
- Connections go left to right
  - No connections within a layer
  - No backward connections (recurrent)

Gradient descent

- No closed-form formula
- Repeat the following steps (for t=0,1,2,... until convergence):
  1. Calculate a gradient ∇w_{ij}^{t}
  2. Apply the update $w_{ij}^{t+1} = w_{ij}^{t} - \alpha \nabla w_{ij}^{t}$
- Stochastic gradient descent
  - One example at a time
- Batch gradient descent
  - All examples at a time

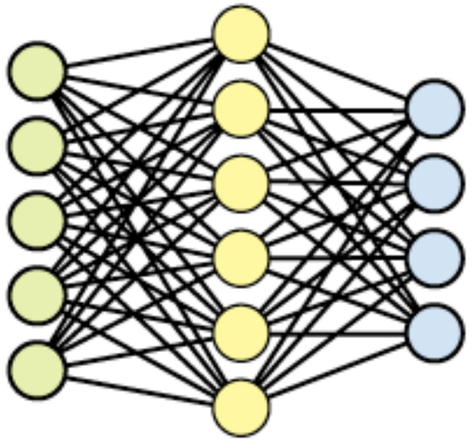
From Neural Networks to **Deep** Neural Networks

A neural Network

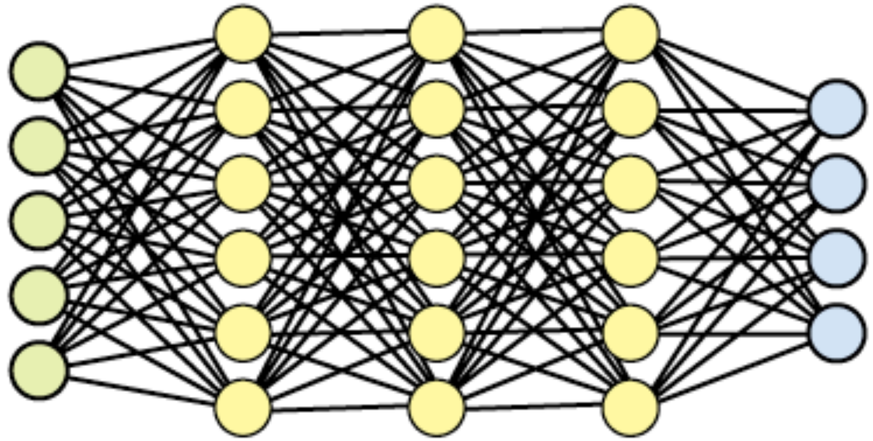


From Neural Networks to **Deep Neural Networks**

A neural Network

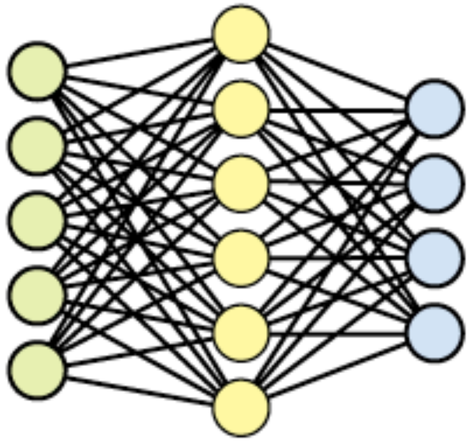


A deep neural Network

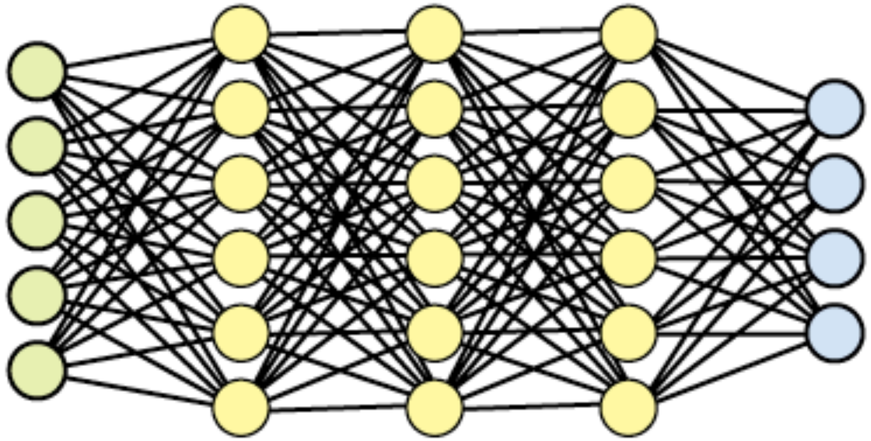


From Neural Networks to Deep Neural Networks

A neural Network



A deep neural Network

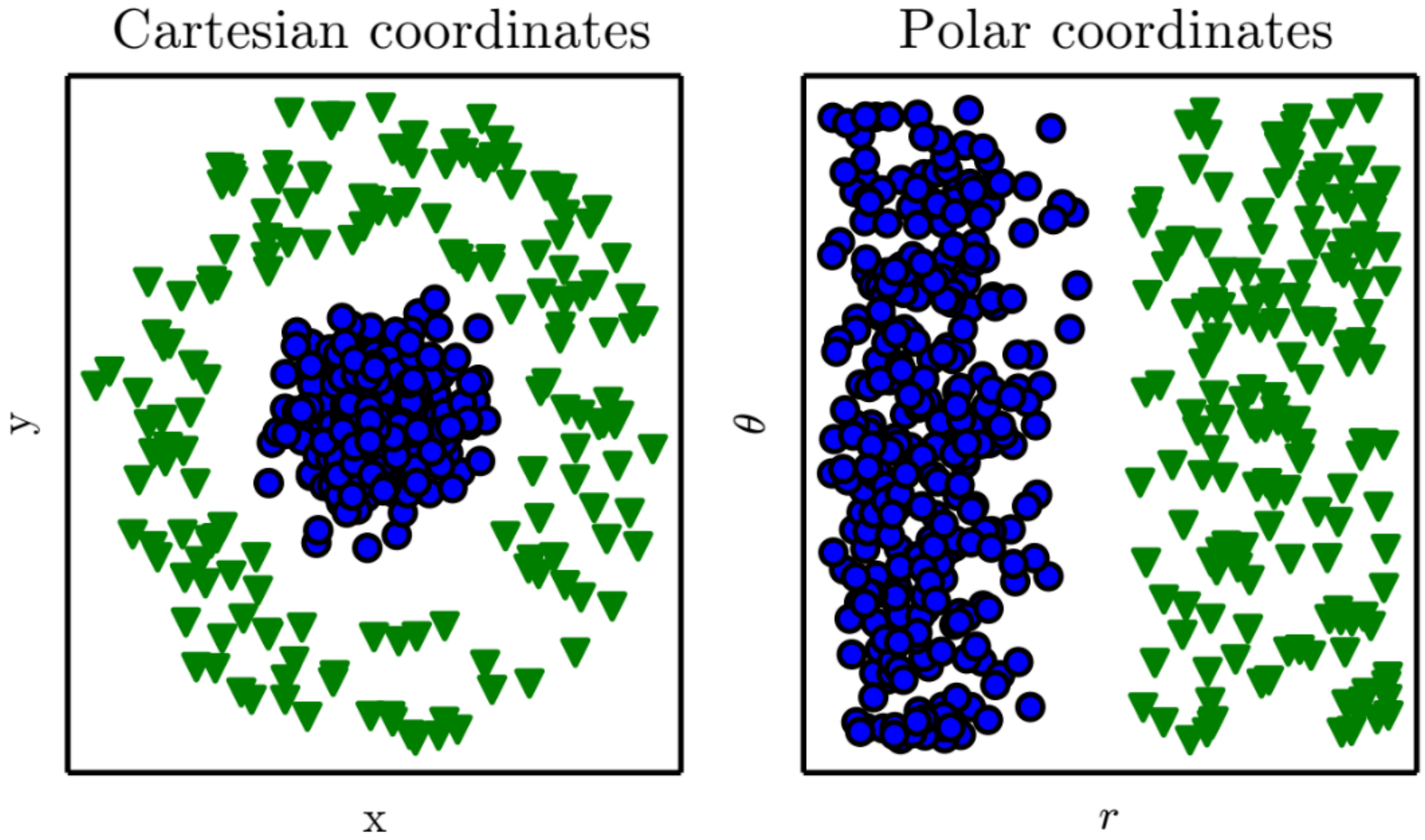


Modern deep learning provides a powerful framework for supervised learning.
By adding more layers and more units within a layer, a deep network can represent functions of increasing complexity.

Deep Learning — Part II, p.163
http://www.deeplearningbook.org/contents/part_practical.html

Another View of deep learning

- **Representations are important**



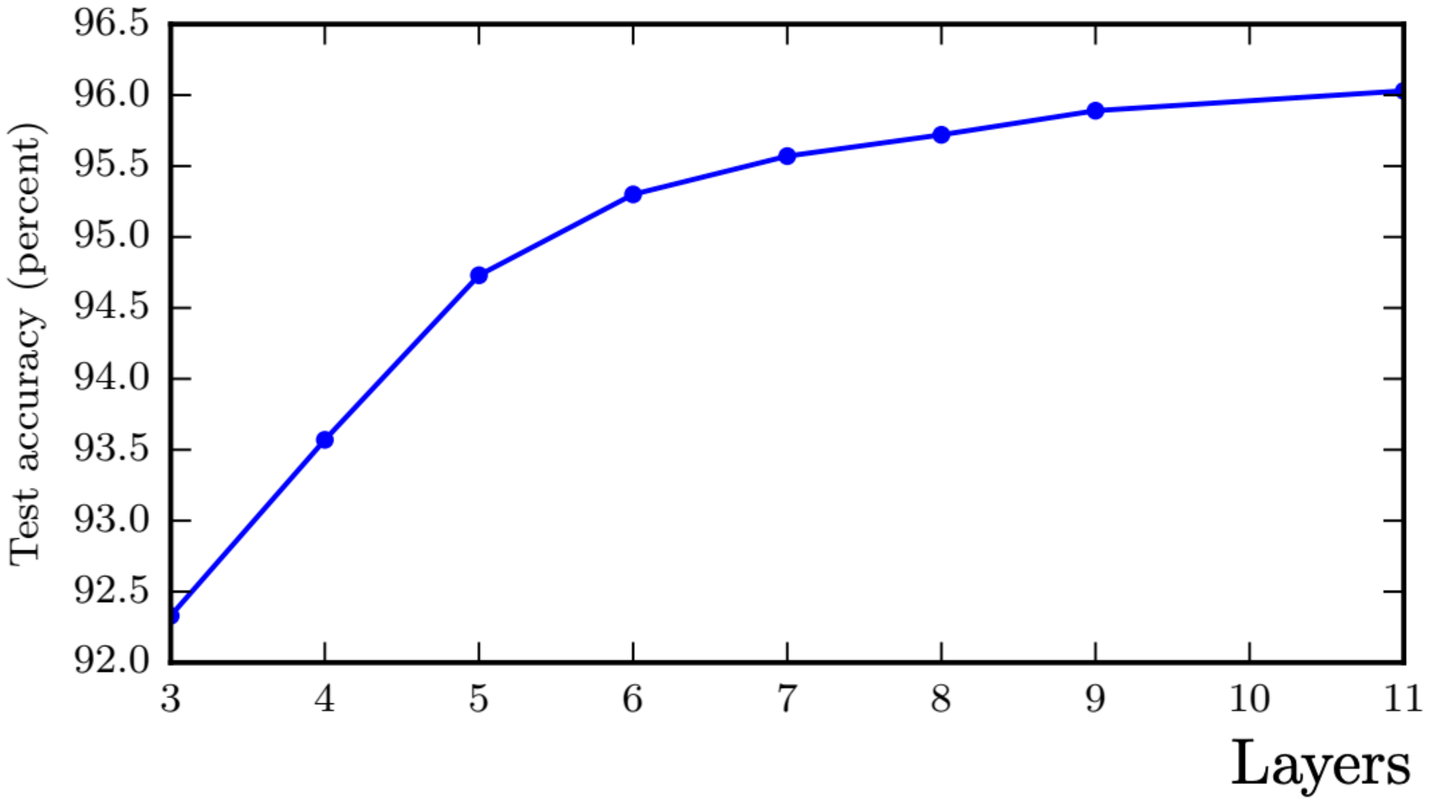
[From: Honglak Lee (and Graham Taylor)]

Hyperparameters

1. Model specific
   - Activation functions (output & hidden), Network size
2. Optimisation Objective
   - Regularization, Early-stopping, Dropout
3. Optimization procedure
   - Momentum, Adaptive learning rates

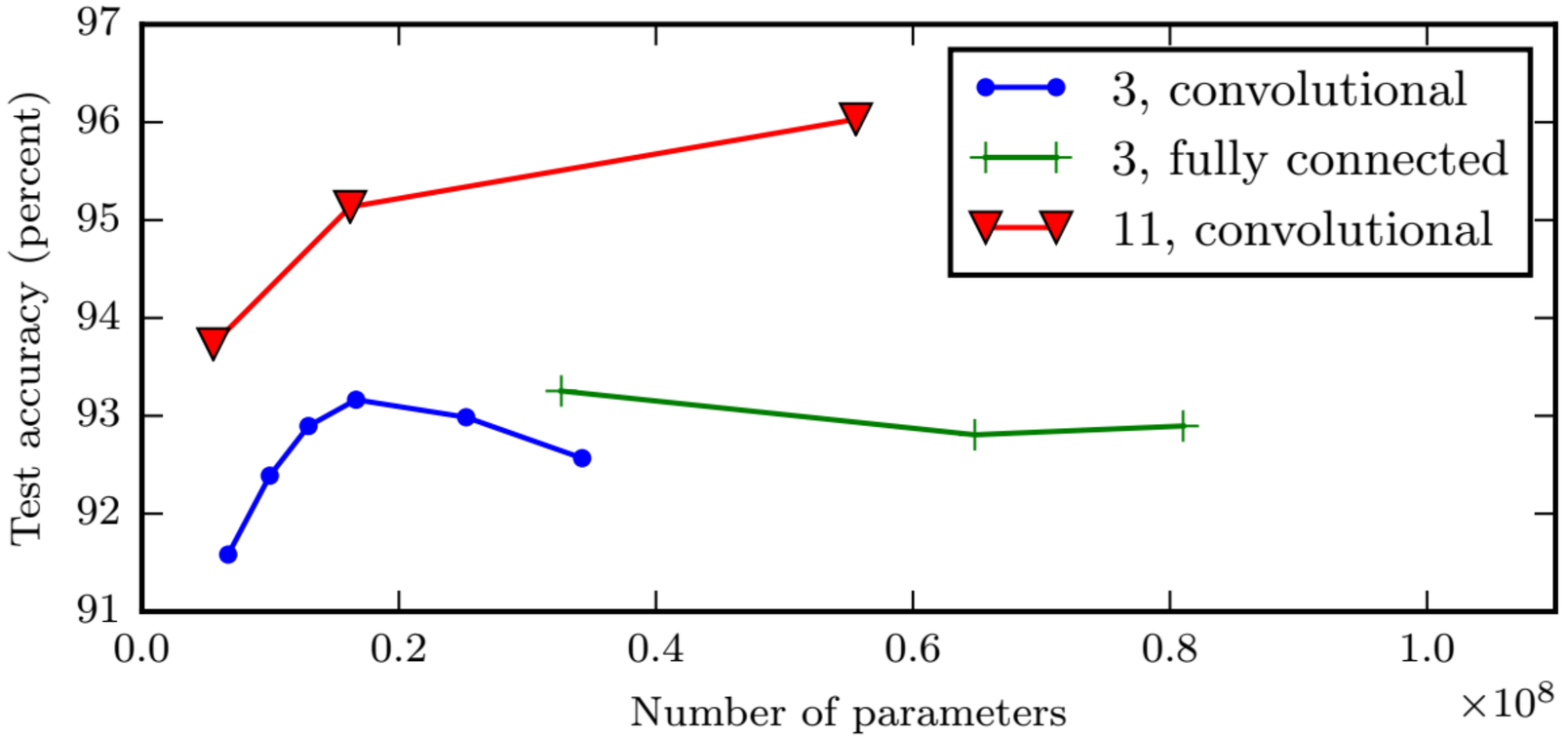
Wide or Deep?

Wide or Deep?



[Figure 6.6, Deep Learning, book]

Wide or Deep?



[Figure 6.7, Deep Learning, book]