Part A: Universal Approximators; Nodes and Layers

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner

Design Choices

Activation function

Loss function

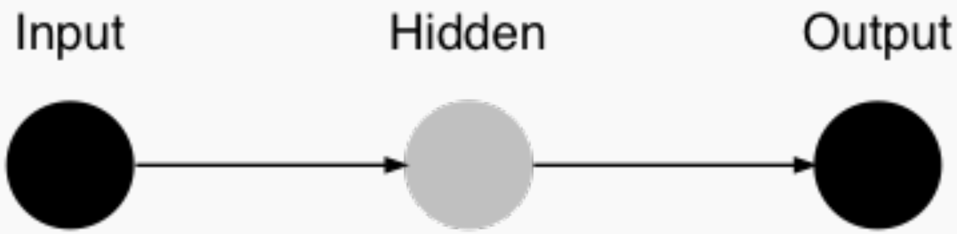
Output units

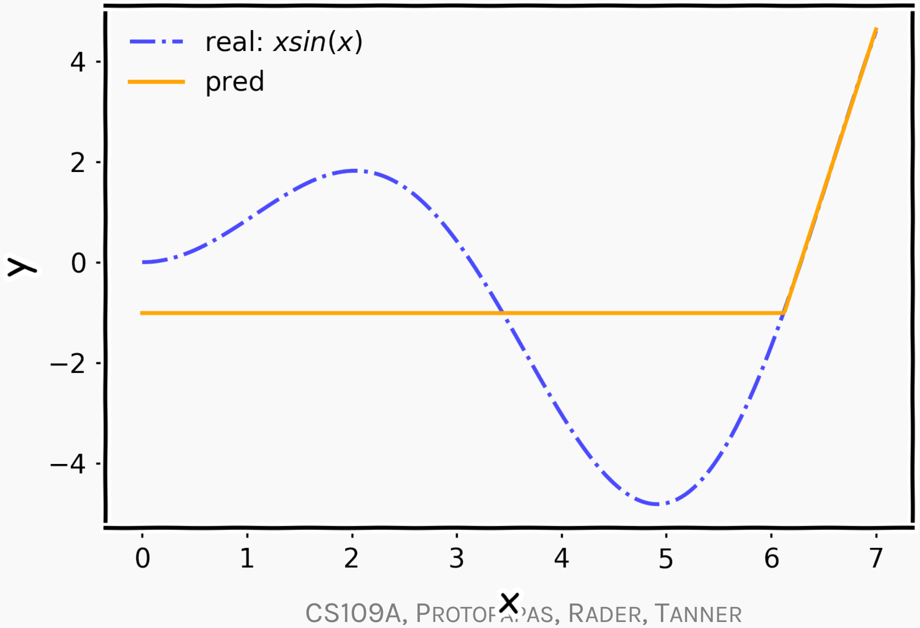
Architecture

Optimizer

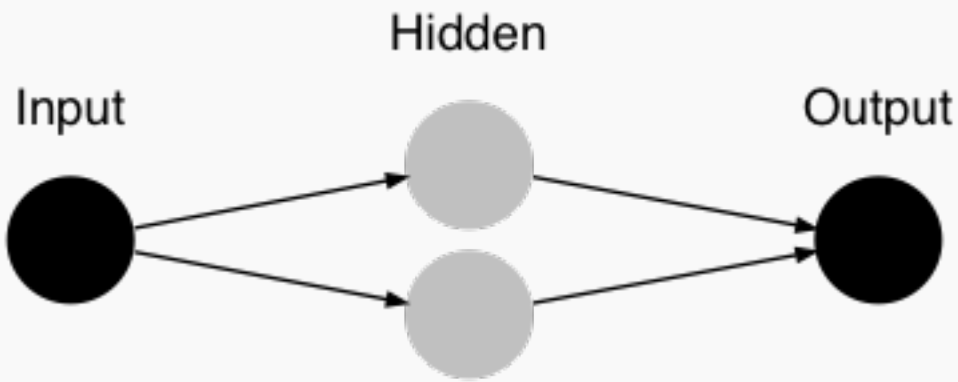
How to bully machine learning training

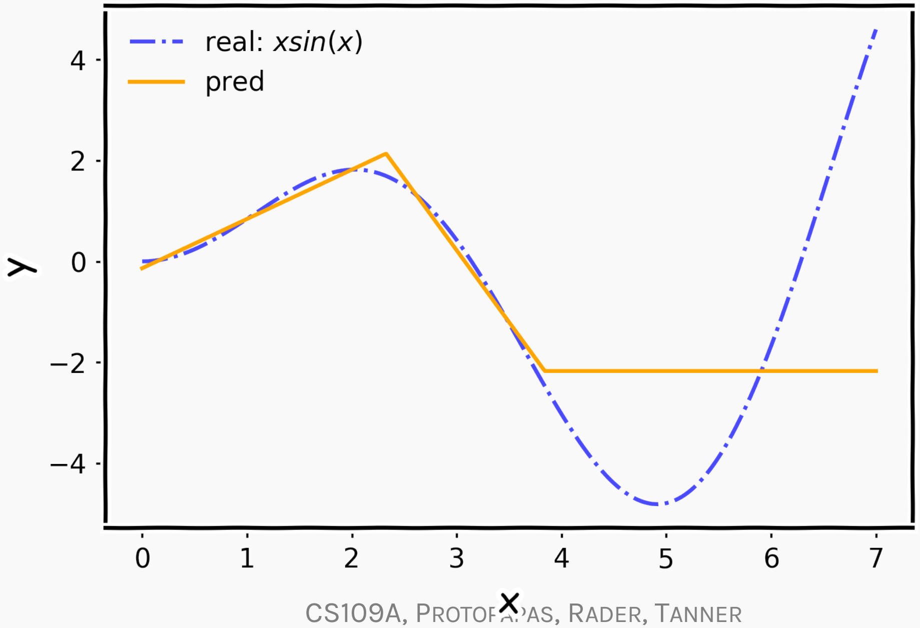
Number of nodes



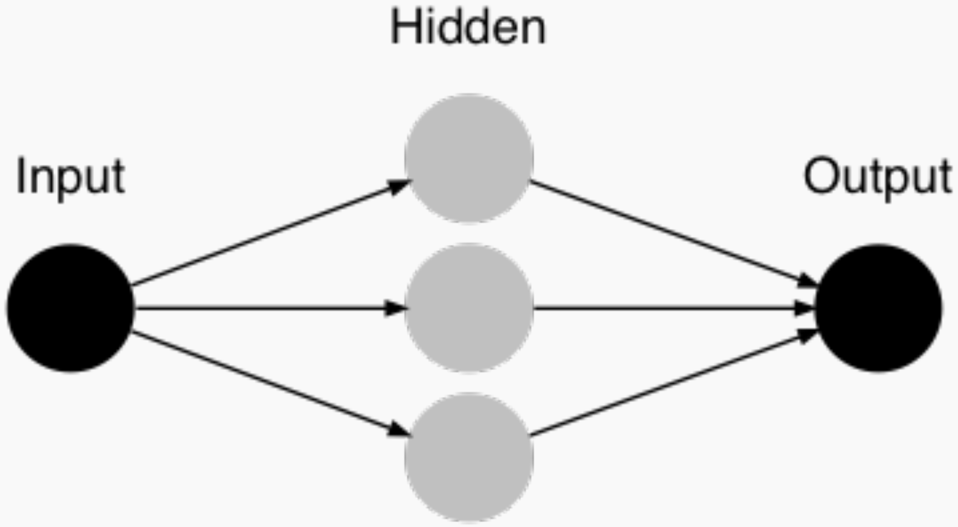


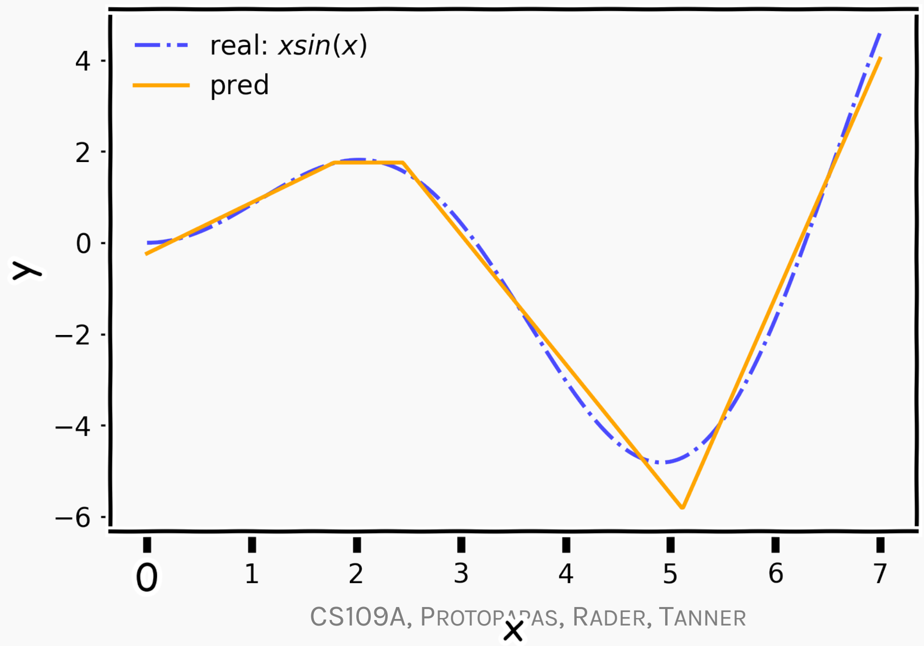
Number of nodes



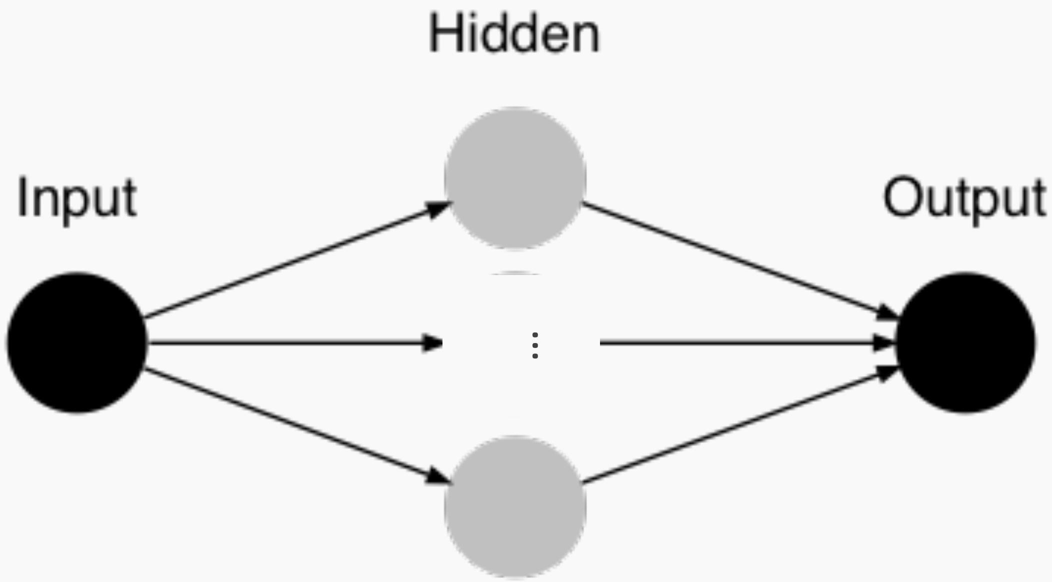


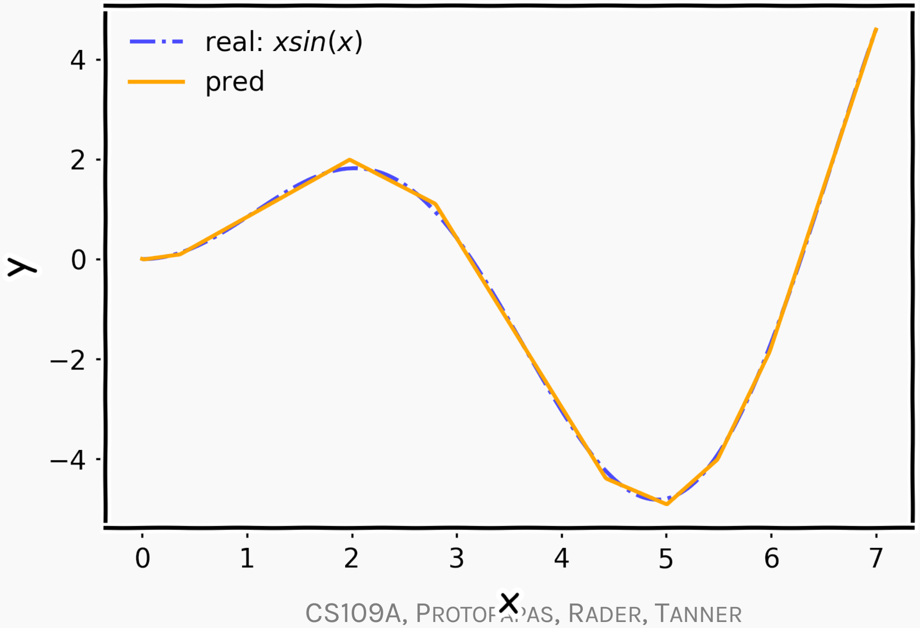
Number of nodes



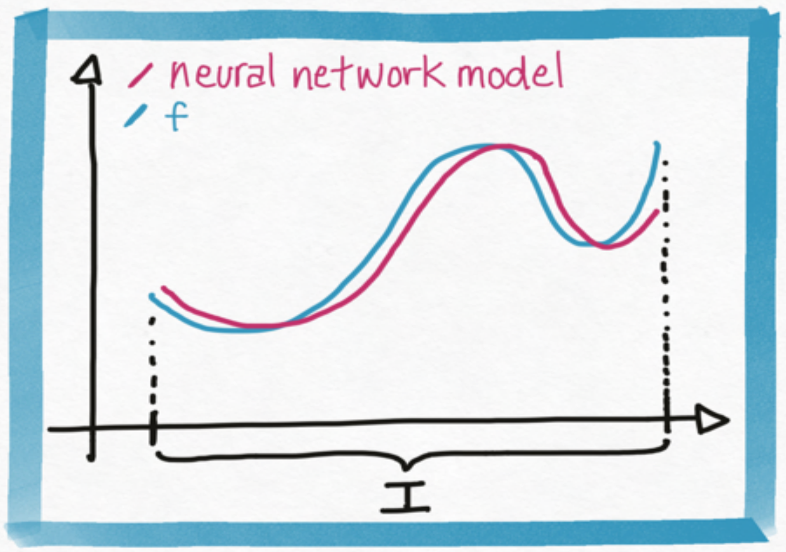


Number of nodes





Neural Networks as Universal Approximators



We have seen that neural networks can represent complex functions, but are there limitations on what a neural network can express?

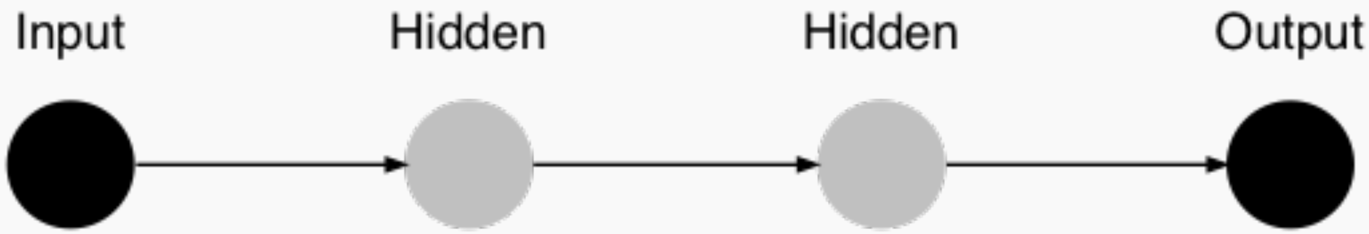
Theorem:

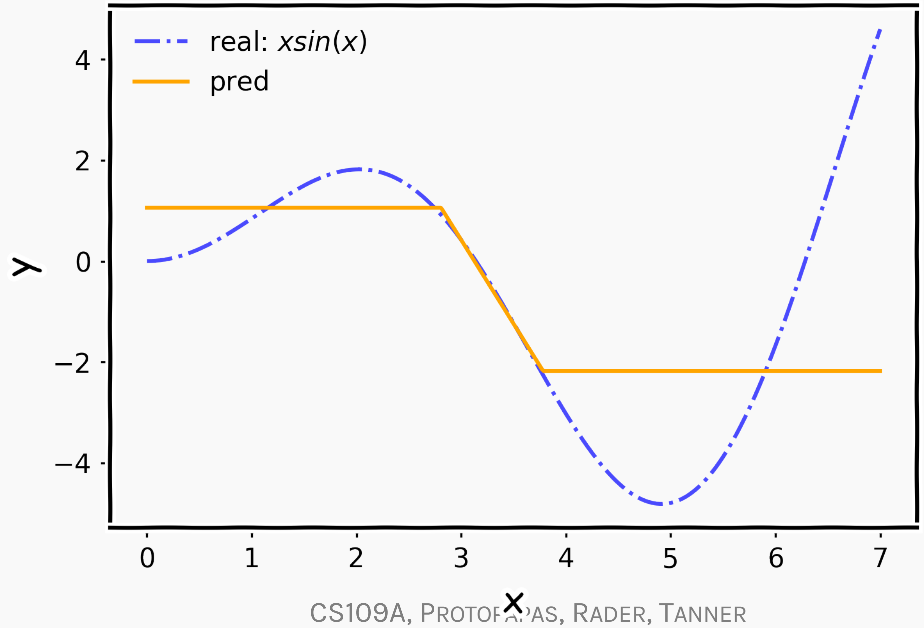
For any continuous function f defined on a bounded domain, we can find a neural network that approximates f with an arbitrary degree of accuracy.

One hidden layer is enough to represent an approximation of any function to an arbitrary degree of accuracy.

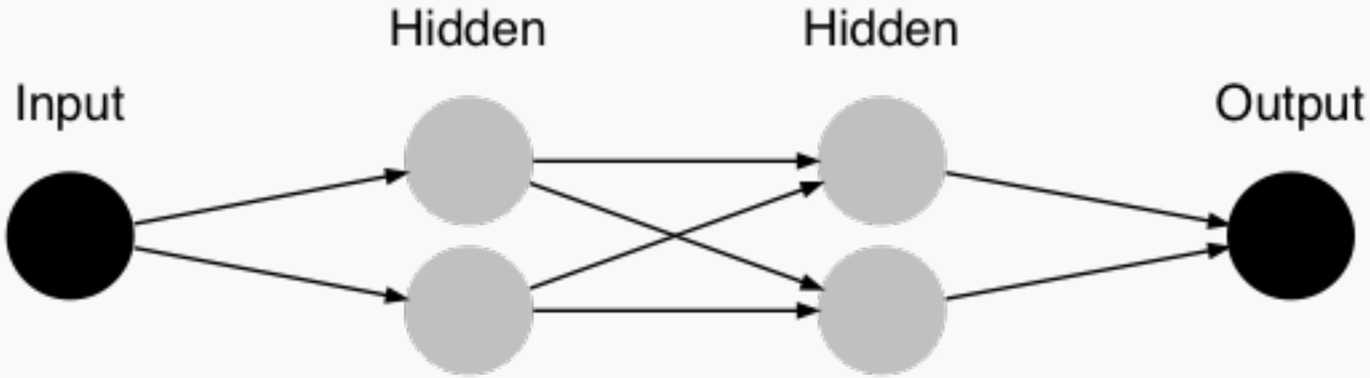
So why deeper?

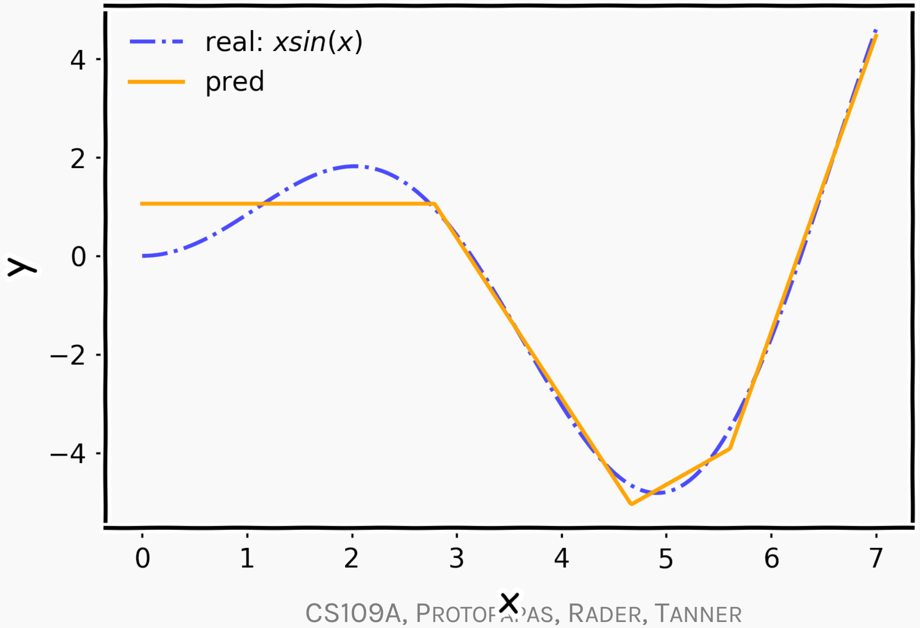
Layers


Input
Hidden
Hidden
Output

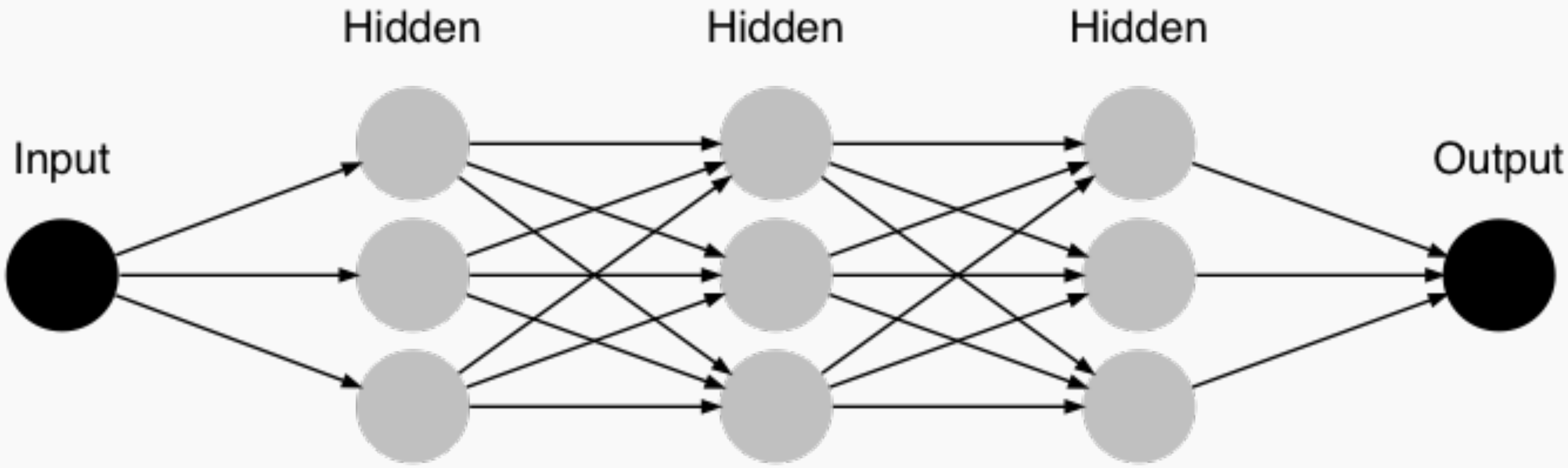

real: $x\sin(x)$
pred
y
x
4
2
0
−2
−4
0
1
2
3
4
5
6
7

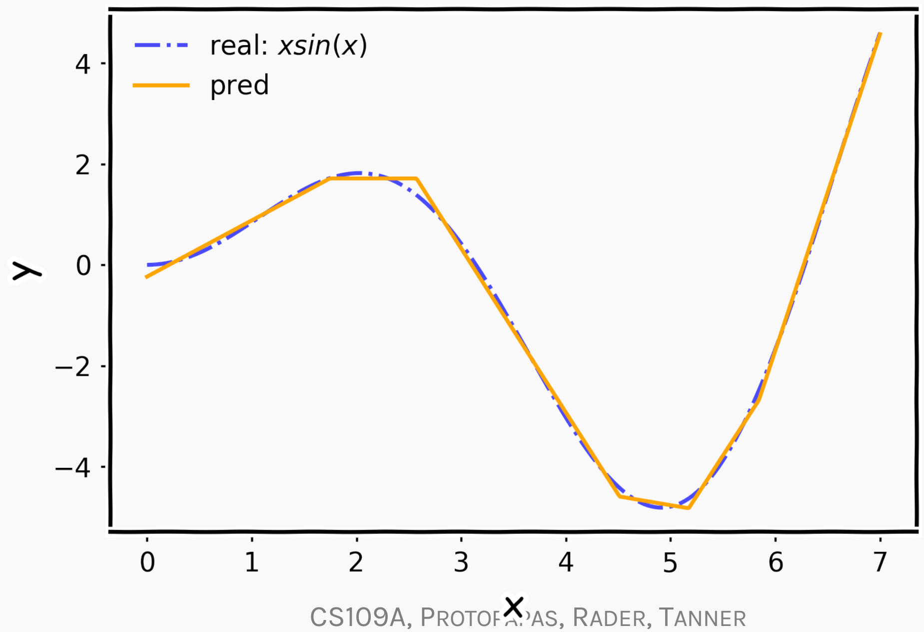
Layers


Hidden
Hidden
Input
Output


real: $xsin(x)$
pred
y
x
4
2
0
−2
−4
0
1
2
3
4
5
6
7

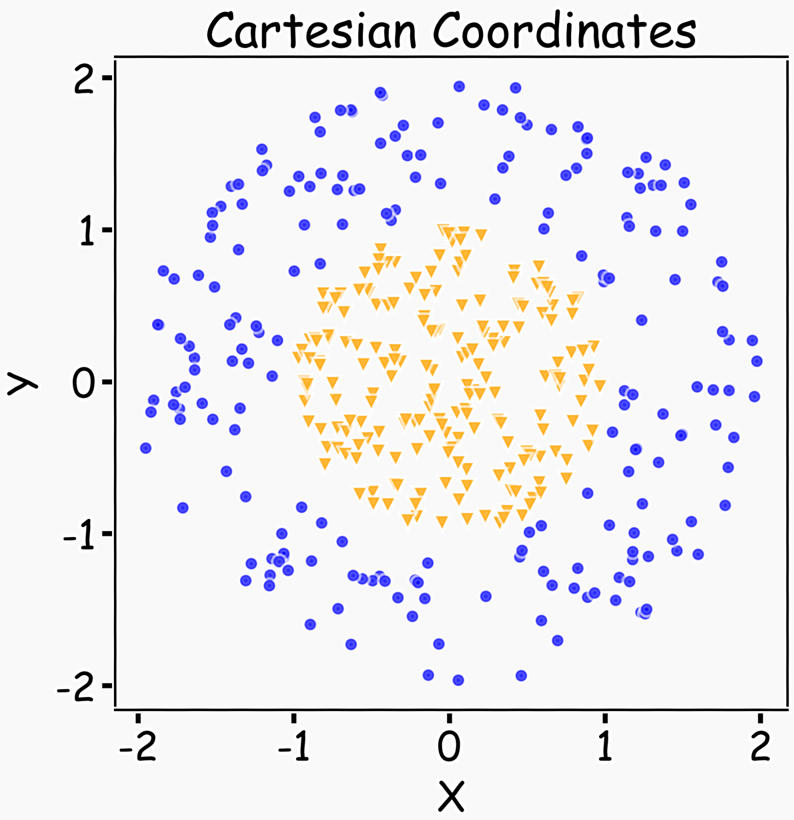
Layers

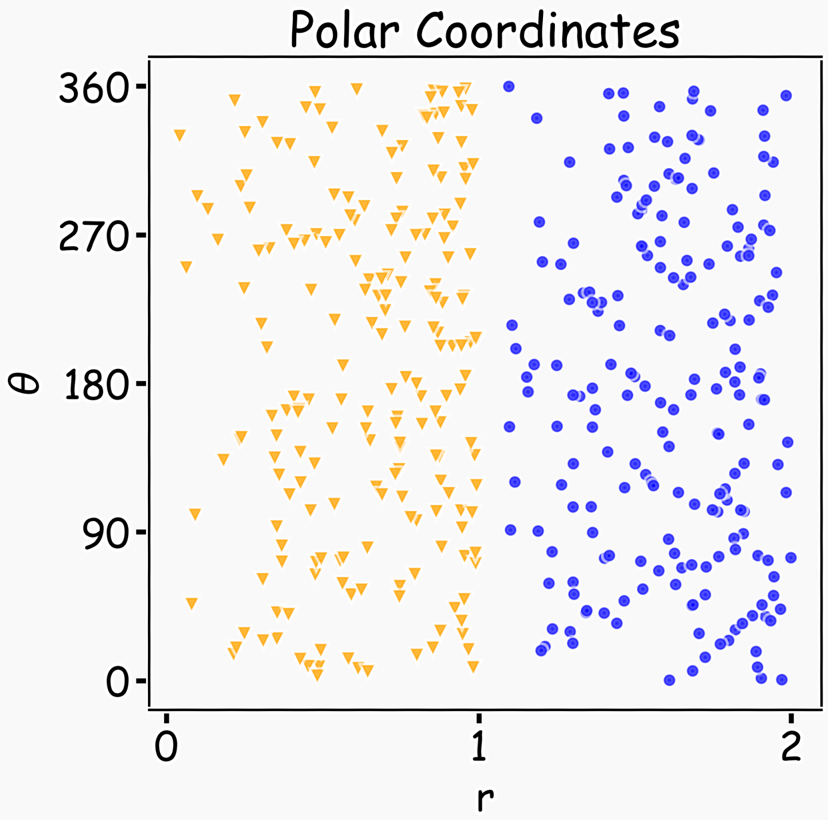

Hidden
Hidden
Hidden
Input
Output


real: $xsin(x)$
pred
4
2
0
−2
−4
y
0
1
2
3
4
5
6
7
x

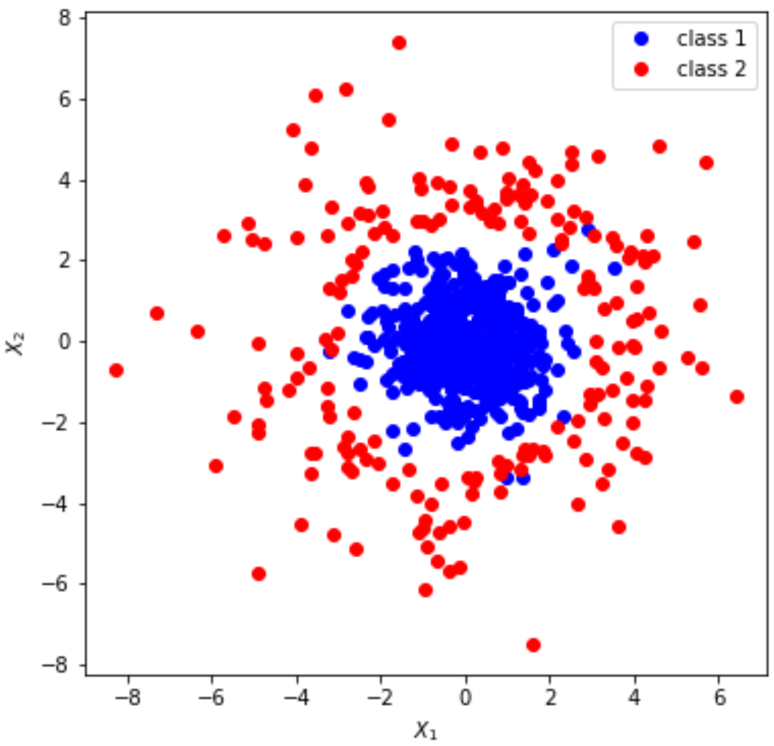
Why layers?

Representation matters!

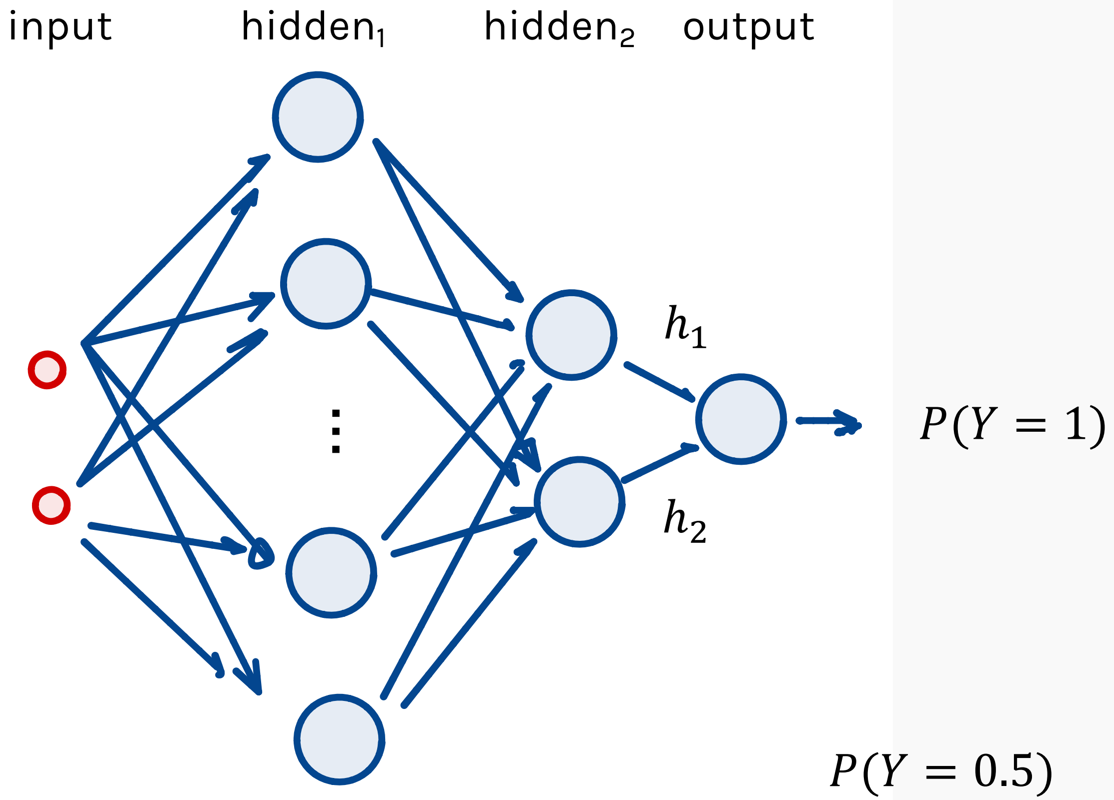




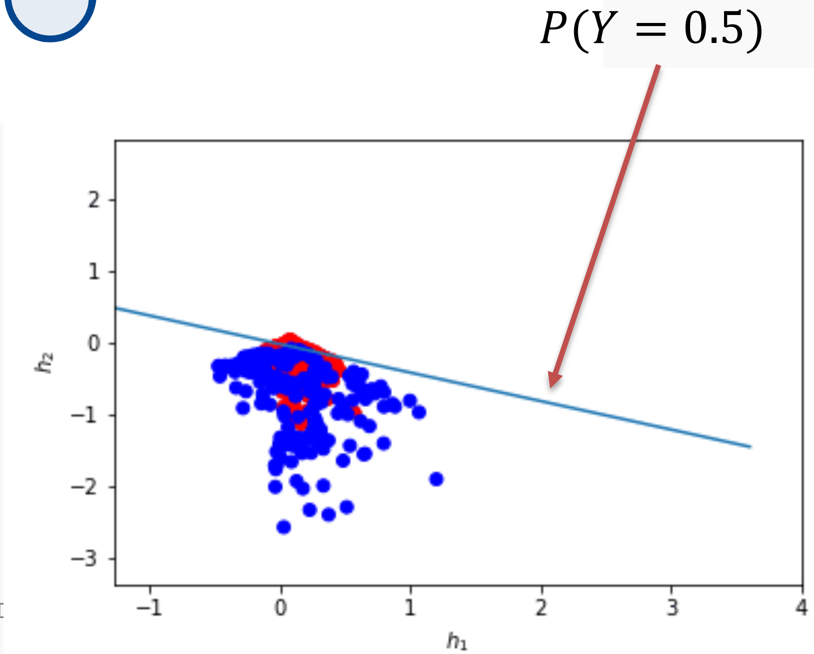
Neural networks can **learn useful representations** for the problem. This is another reason why they can be so powerful!



class 1
class 2
X_1
X_2

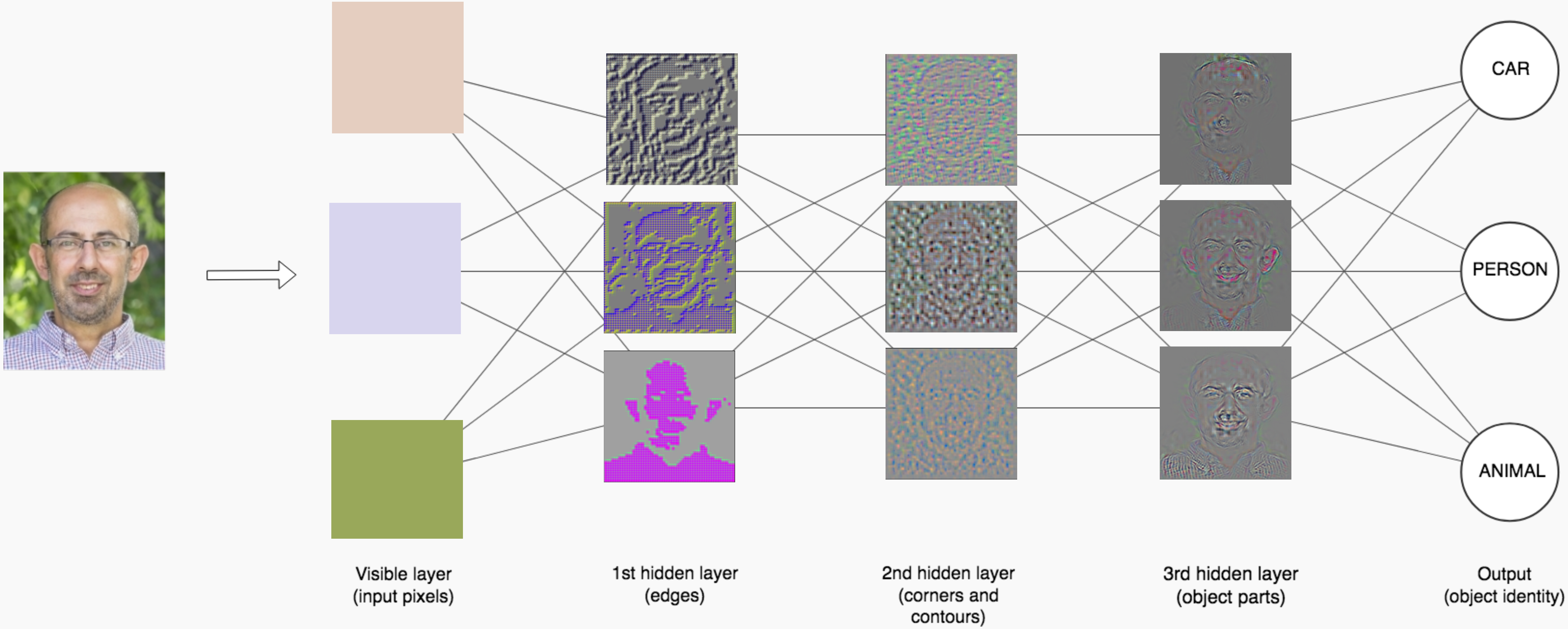


input
hidden$_1$
hidden$_2$
output
h_1
h_2
$P(Y = 1)$

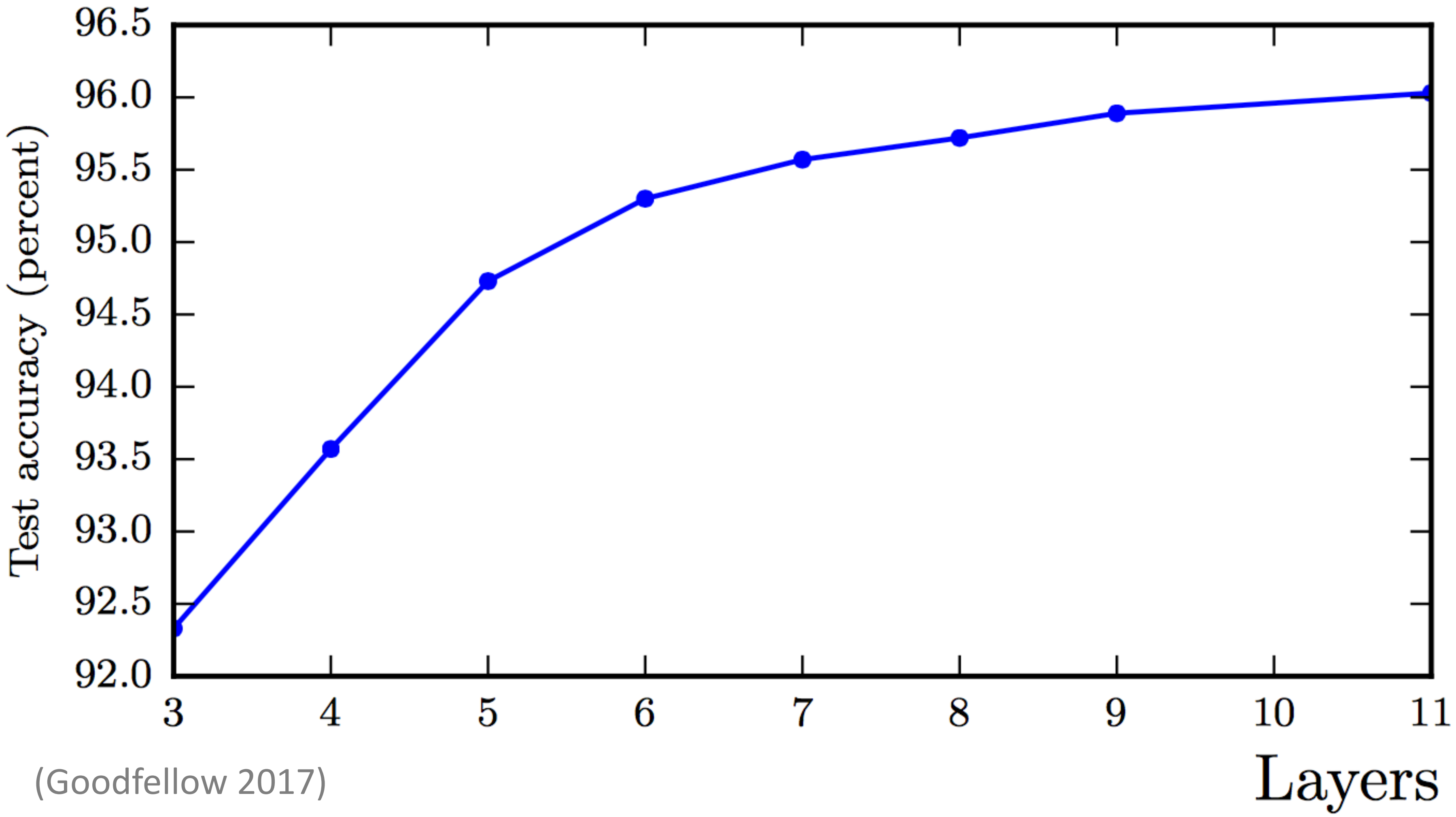


$P(Y = 0.5)$
h_1
h_2

Depth = Repeated Compositions



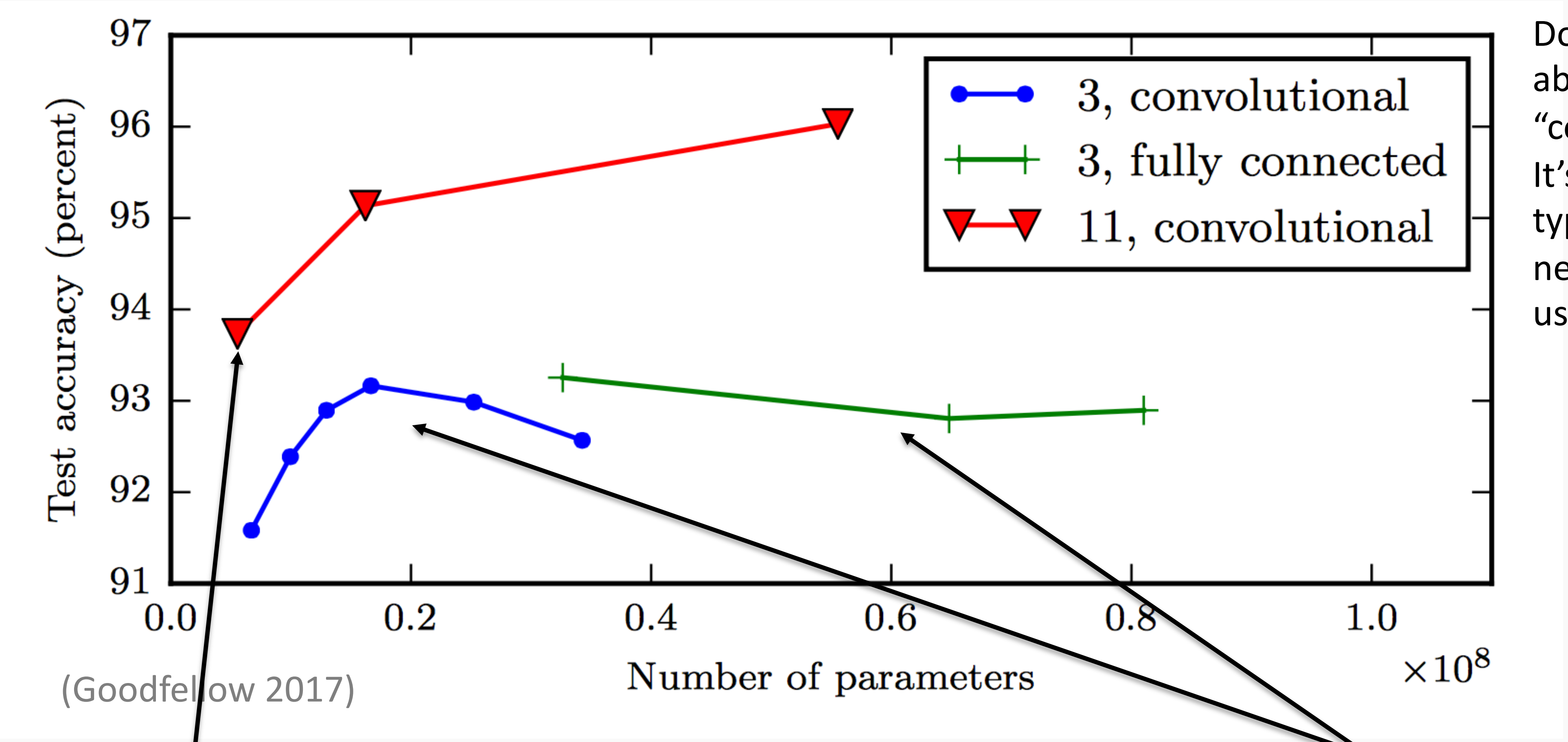
Better Generalization with Depth



(Goodfellow 2017)

Shallow Nets Overfit More

Depth helps, and it's not just because of more parameters

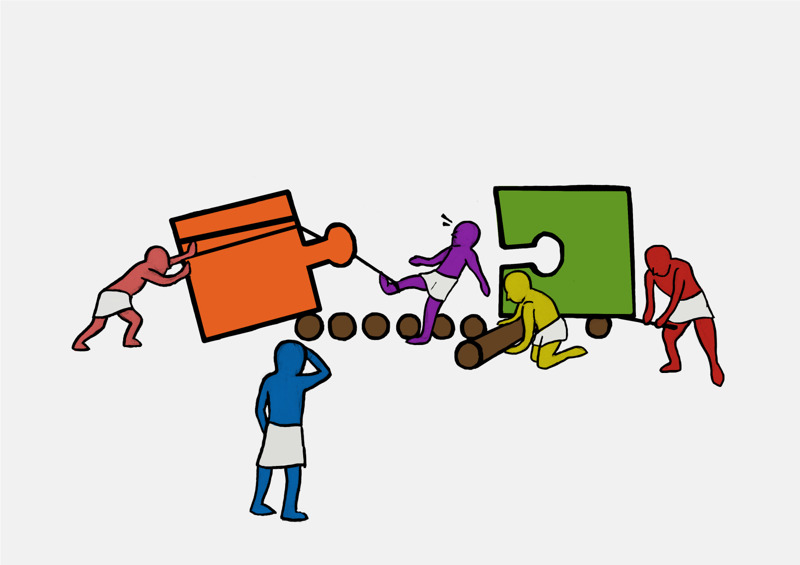


(Goodfellow 2017)

Don't worry about this word "convolutional". It's just a special type of neural network, often used for images.

The **11-layer net** generalizes better on the test set when controlling for number of parameters.

The 3-layer nets perform worse on the test set, even with similar number of total parameters.



Classifier using Keras on Iris data