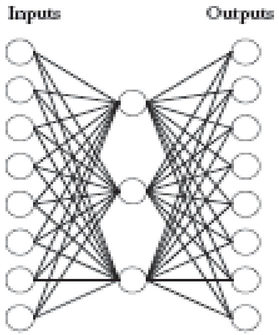
# Learning Hidden layer representation



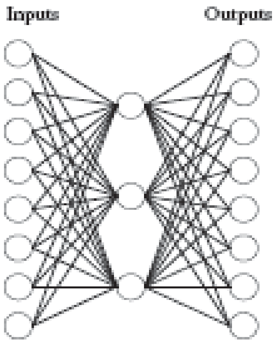
A target function:

| Input | | Output |
|---|---|---|
| 10000000 | → | 10000000 |
| 01000000 | → | 01000000 |
| 00100000 | → | 00100000 |
| 00010000 | → | 00010000 |
| 00001000 | → | 00001000 |
| 00000100 | → | 00000100 |
| 00000010 | → | 00000010 |
| 00000001 | → | 00000001 |

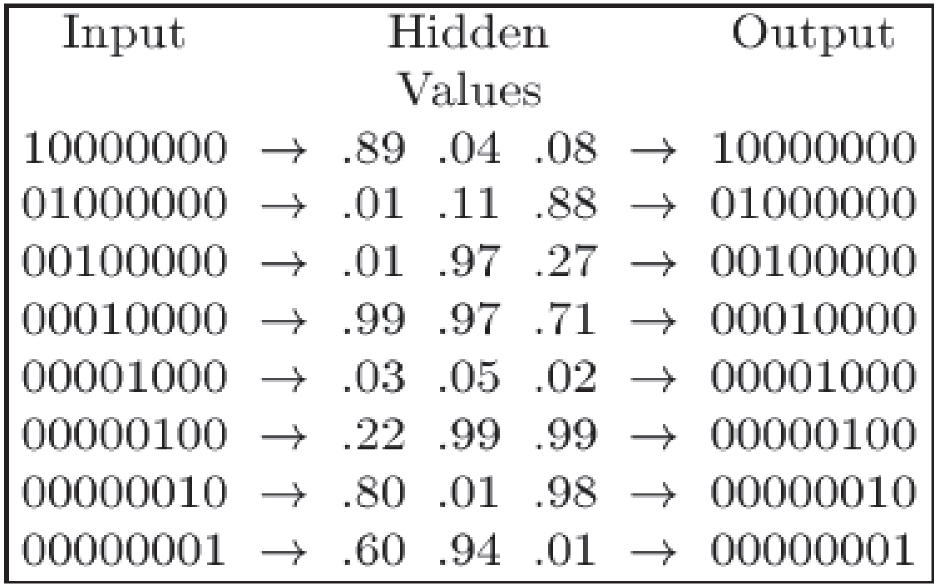
Can this be learned??

# Learning Hidden layer representation

A network:

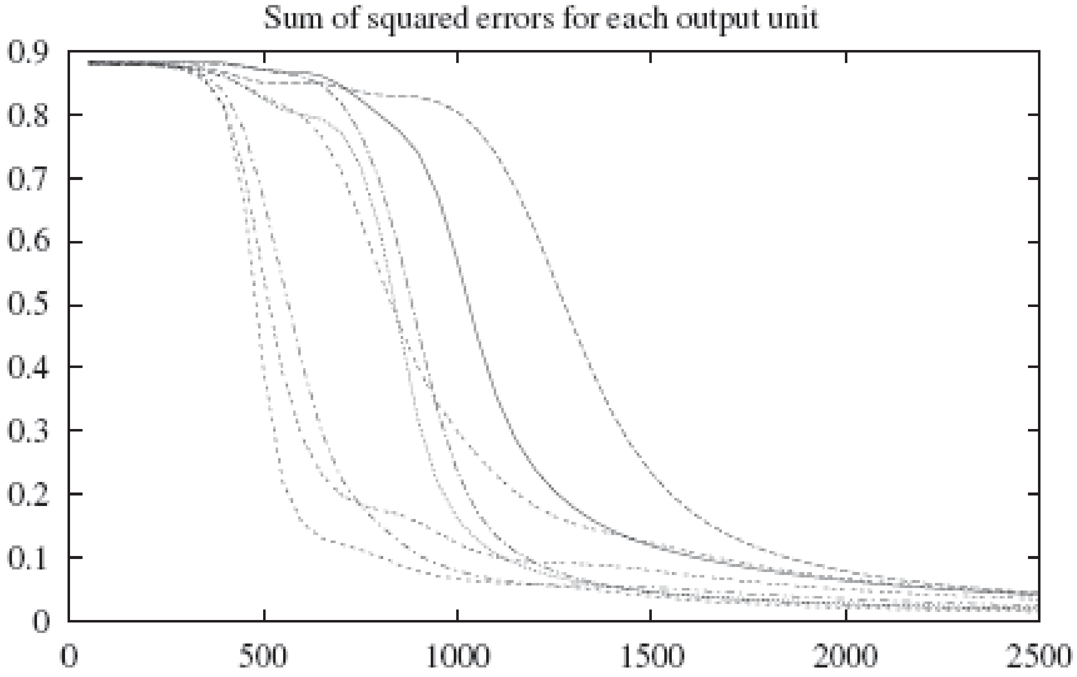


Learned hidden layer representation:

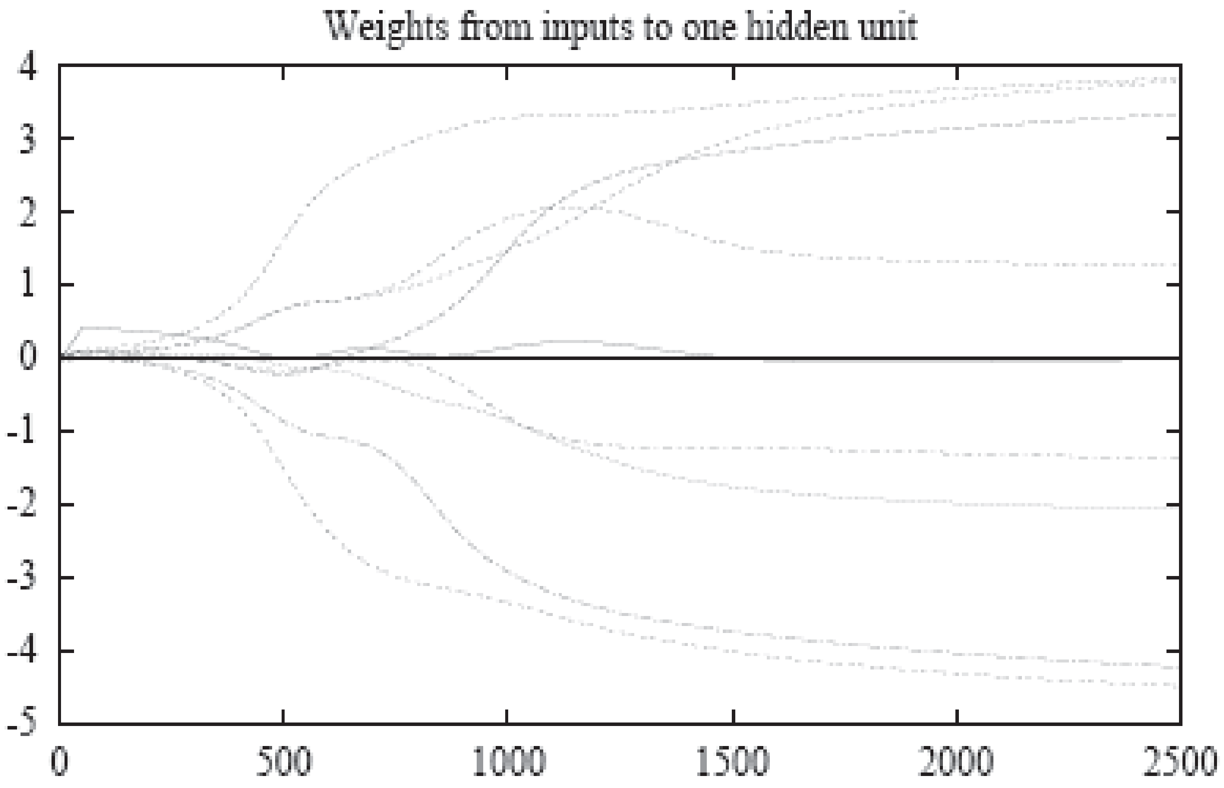


| Input | | Hidden<br>Values | | | | Output |
|---|---|---|---|---|---|---|
| 10000000 | → | .89 | .04 | .08 | → | 10000000 |
| 01000000 | → | .01 | .11 | .88 | → | 01000000 |
| 00100000 | → | .01 | .97 | .27 | → | 00100000 |
| 00010000 | → | .99 | .97 | .71 | → | 00010000 |
| 00001000 | → | .03 | .05 | .02 | → | 00001000 |
| 00000100 | → | .22 | .99 | .99 | → | 00000100 |
| 00000010 | → | .80 | .01 | .98 | → | 00000010 |
| 00000001 | → | .60 | .94 | .01 | → | 00000001 |

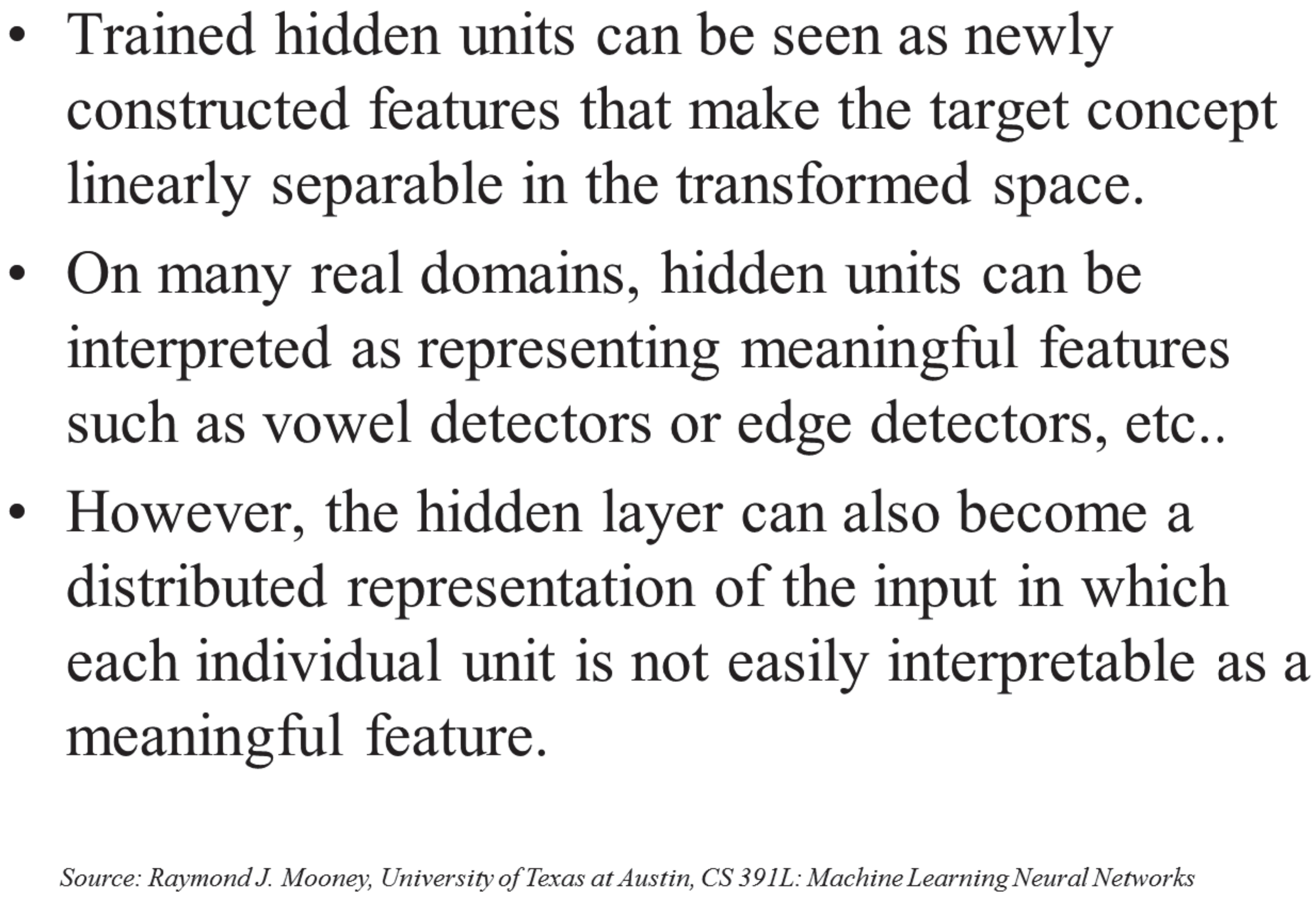
# Sum of squared errors for each output unit



# Weights from inputs to one hidden unit



# Hidden Unit Representations



- Trained hidden units can be seen as newly constructed features that make the target concept linearly separable in the transformed space.
- On many real domains, hidden units can be interpreted as representing meaningful features such as vowel detectors or edge detectors, etc..
- However, the hidden layer can also become a distributed representation of the input in which each individual unit is not easily interpretable as a meaningful feature.

*Source: Raymond J. Mooney, University of Texas at Austin, CS 391L: Machine Learning Neural Networks*