# Machine Learning

## and having it deep and structured

Hung-yi Lee

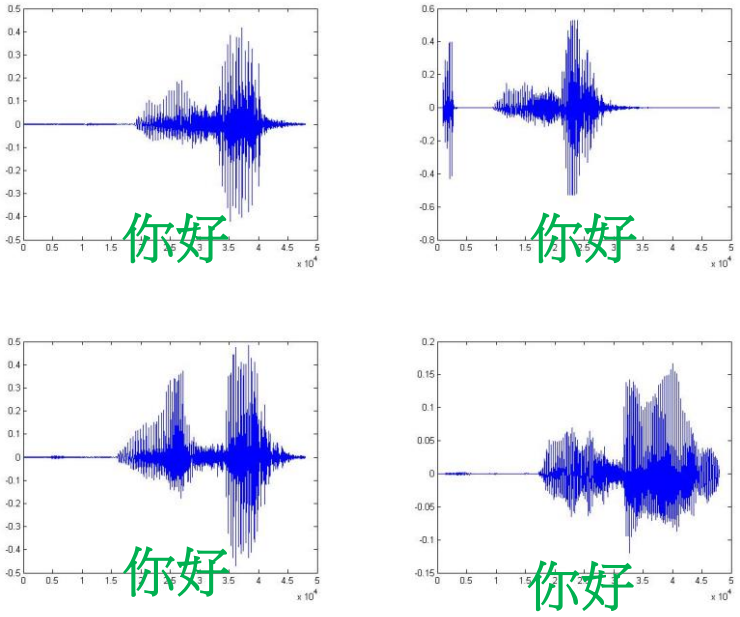
# What is Machine Learning?

# You know how to program …


- You can ask computers to do lots of things for you.
- However, computer can only do what you ask it to do.
- Computer can never solve the problem you can’t solve.

# Some tasks are very complex

- One day, you are asked to write a program for speech recognition.

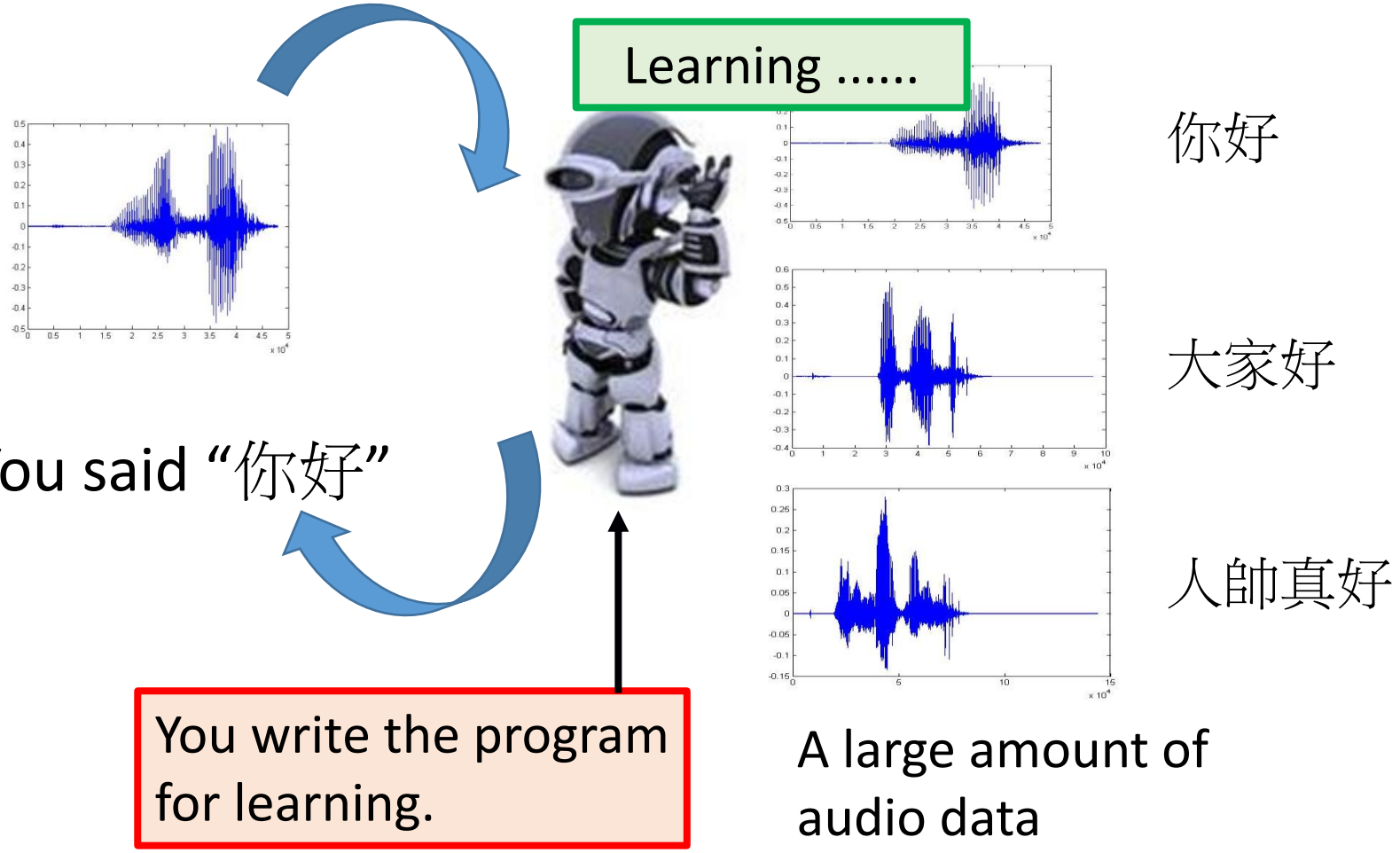



Find the common patterns from the left waveforms.

You quickly get lost in the exceptions and special cases.

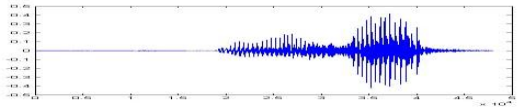
It seems impossible to write a program for speech recognition.

# Let the machine learn by itself



# Learning ≈ Looking for a Function

- Speech Recognition

$f($  $)=$ “你好”

- Handwritten Recognition

$f($  $)=$ “2”

- Weather forecast

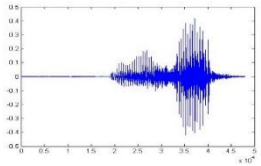
$f($ weather today $)=$ “sunny tomorrow”

- Play video games

$f($ Positions and number of enemies $)=$ “fire”

# Framework

$x$:



$\hat{y}$: “你好” “2”

label

Model

Hypothesis Function Set

$f_1, f_2 \cdots$

Training: Pick the best Function f*

“Best” Function

$f^*$

Testing:

$f^*(x') = y'$

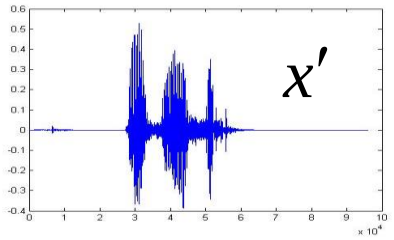
$y' =$ 大家好

Training Data

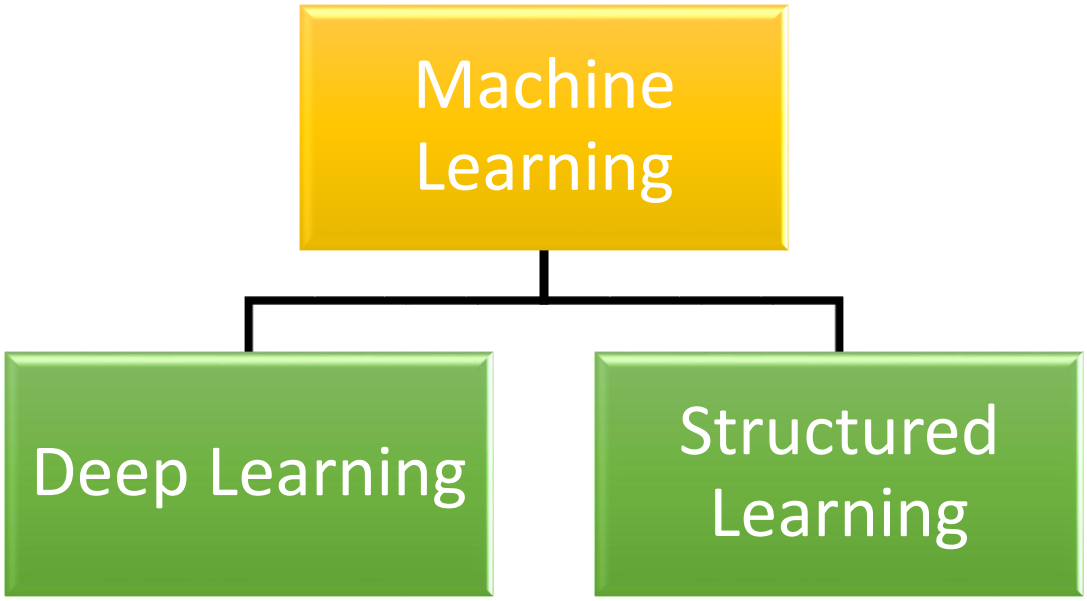
$x$ :function input

$\hat{y}$ :function output

$\{(x^1, \hat{y}^1), (x^2, \hat{y}^2), \ldots\}$



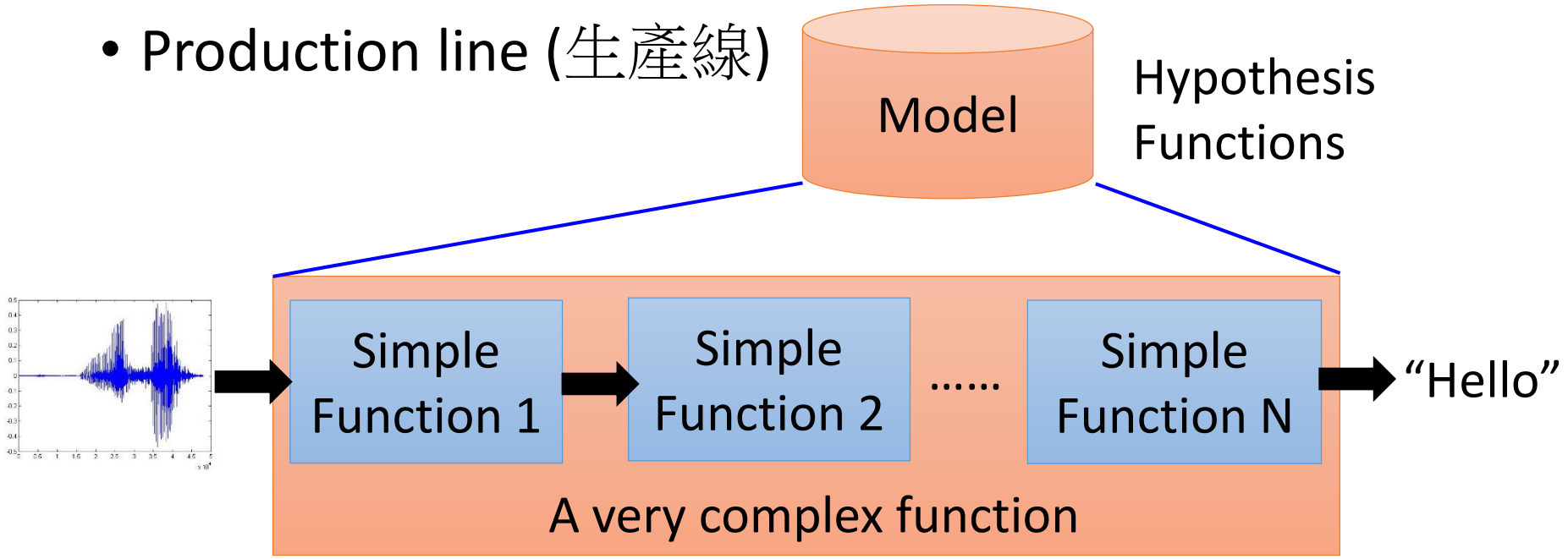
# This Course



# Deep Learning

# What is Deep Learning?

- Production line (生產線)

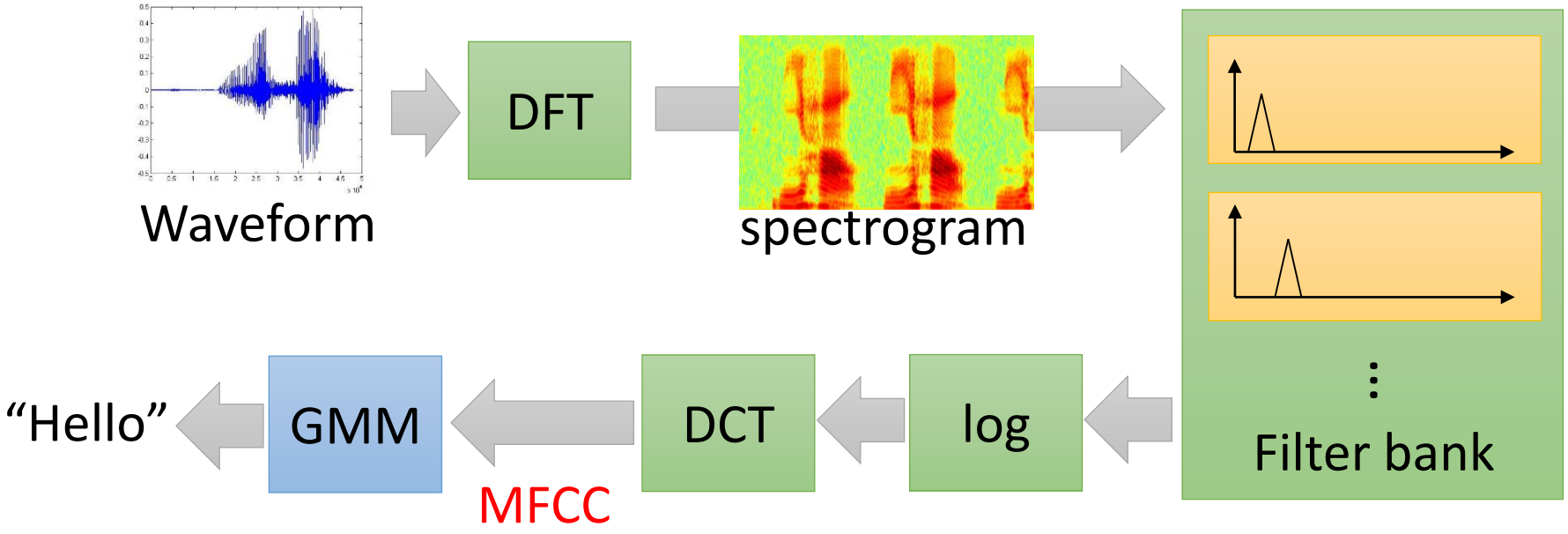



End-to-end training:

What each function should do is learned automatically

# Deep v.s. Shallow - Speech Recognition

- Shallow Approach



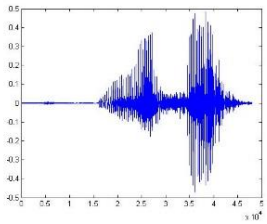

Each box is a simple function in the production line:

:hand-crafted :learned from data

# Deep v.s. Shallow - Speech Recognition

- Deep Learning

"Bye bye, MFCC"
- Deng Li in Interspeech 2014



$f_1$ → $f_2$ → $f_3$ → $f_4$ → $f_5$ → "Hello"

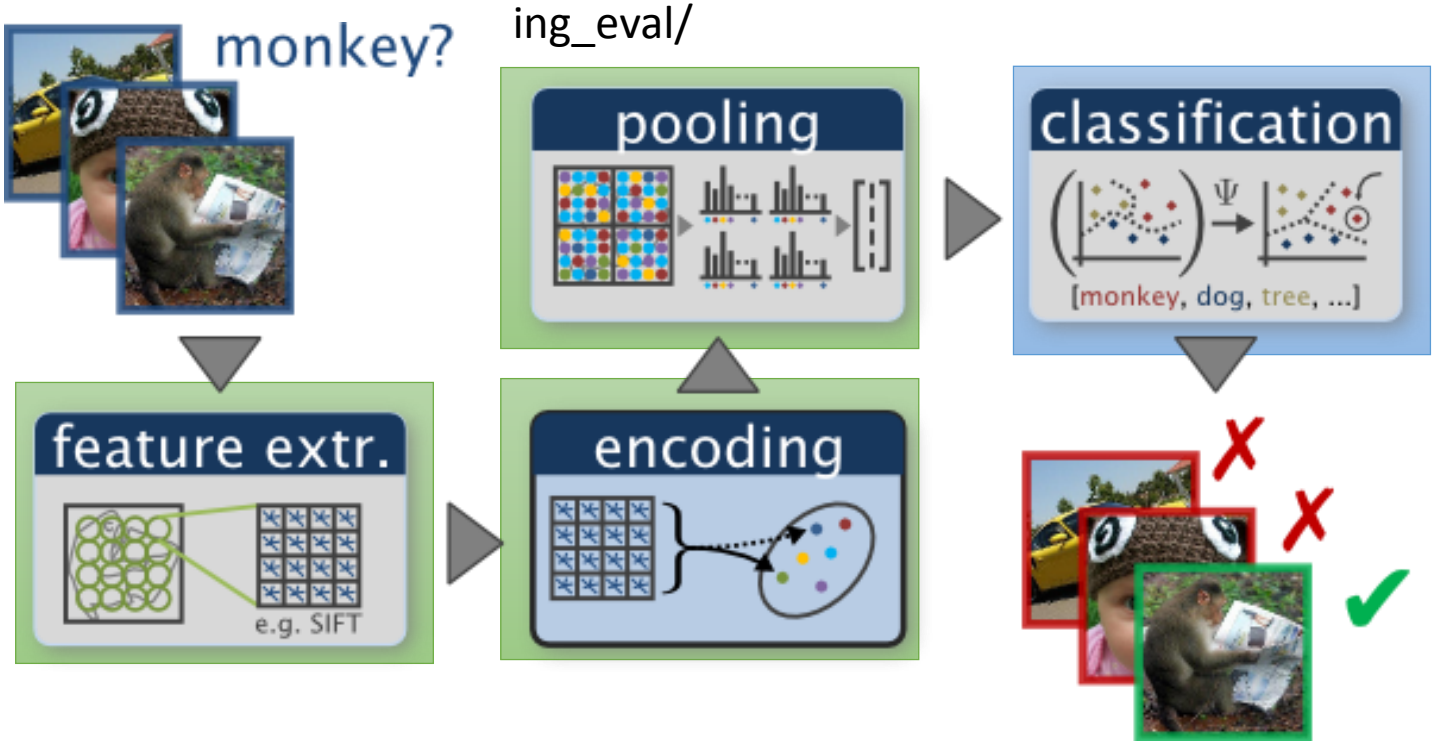
All functions are learned from data

Less engineering labor, but machine learns more

# Deep v.s. Shallow
# - Image Recognition

- Shallow Approach

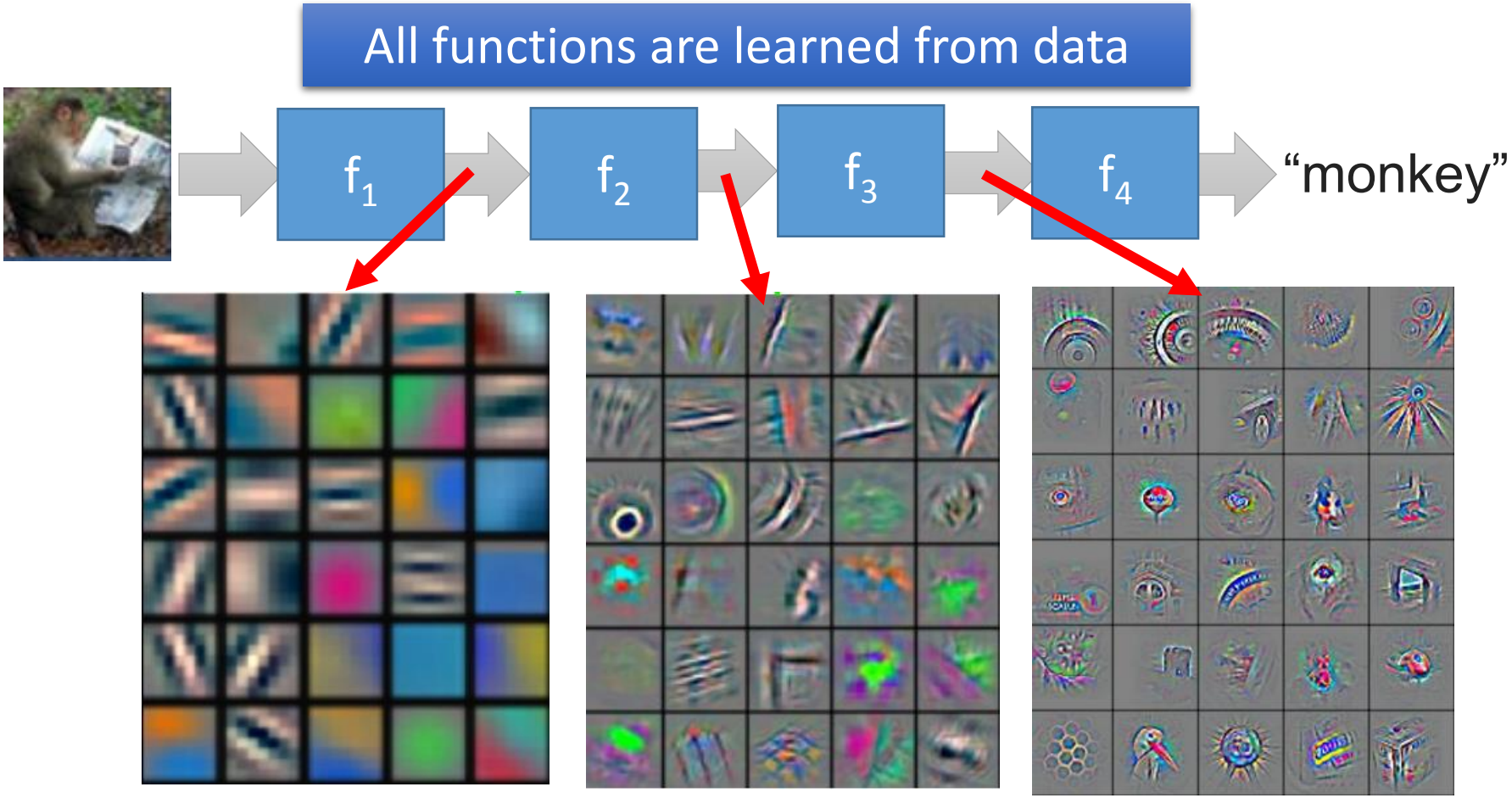
http://www.robots.ox.ac.uk/~vgg/research/encoding_eval/



# Deep v.s. Shallow - Image Recognition

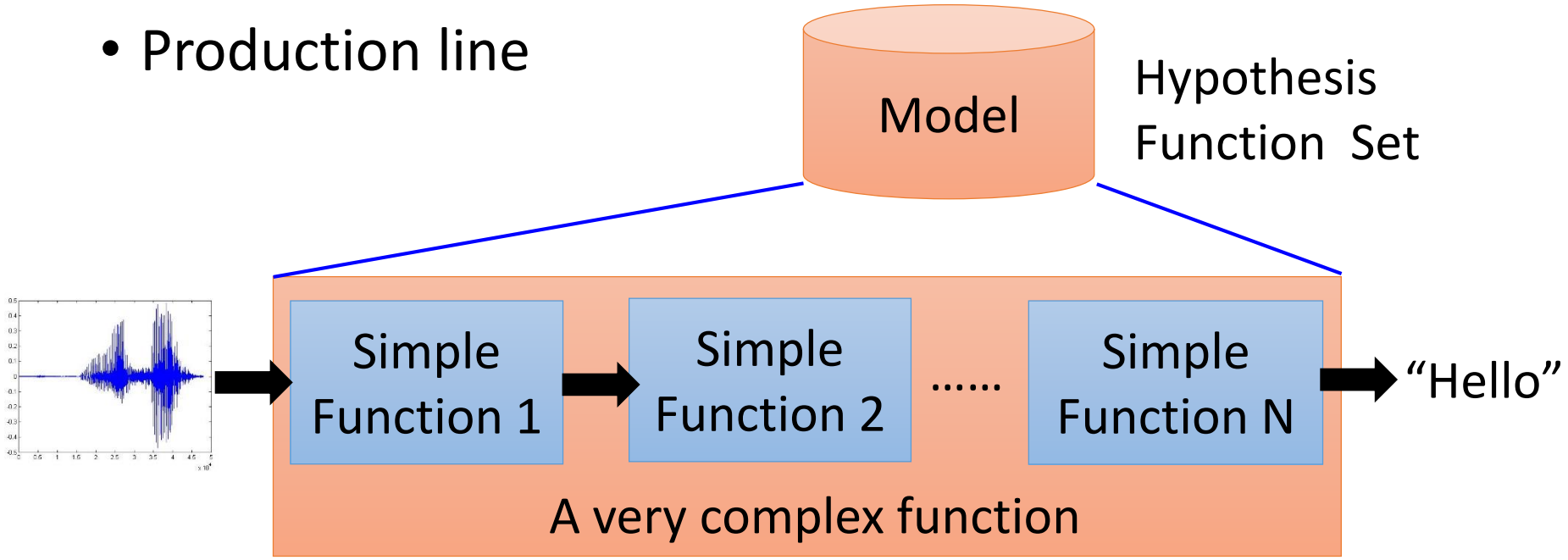
Reference: Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014* (pp. 818-833)

- Deep Learning



# What is Deep Learning?

- Production line

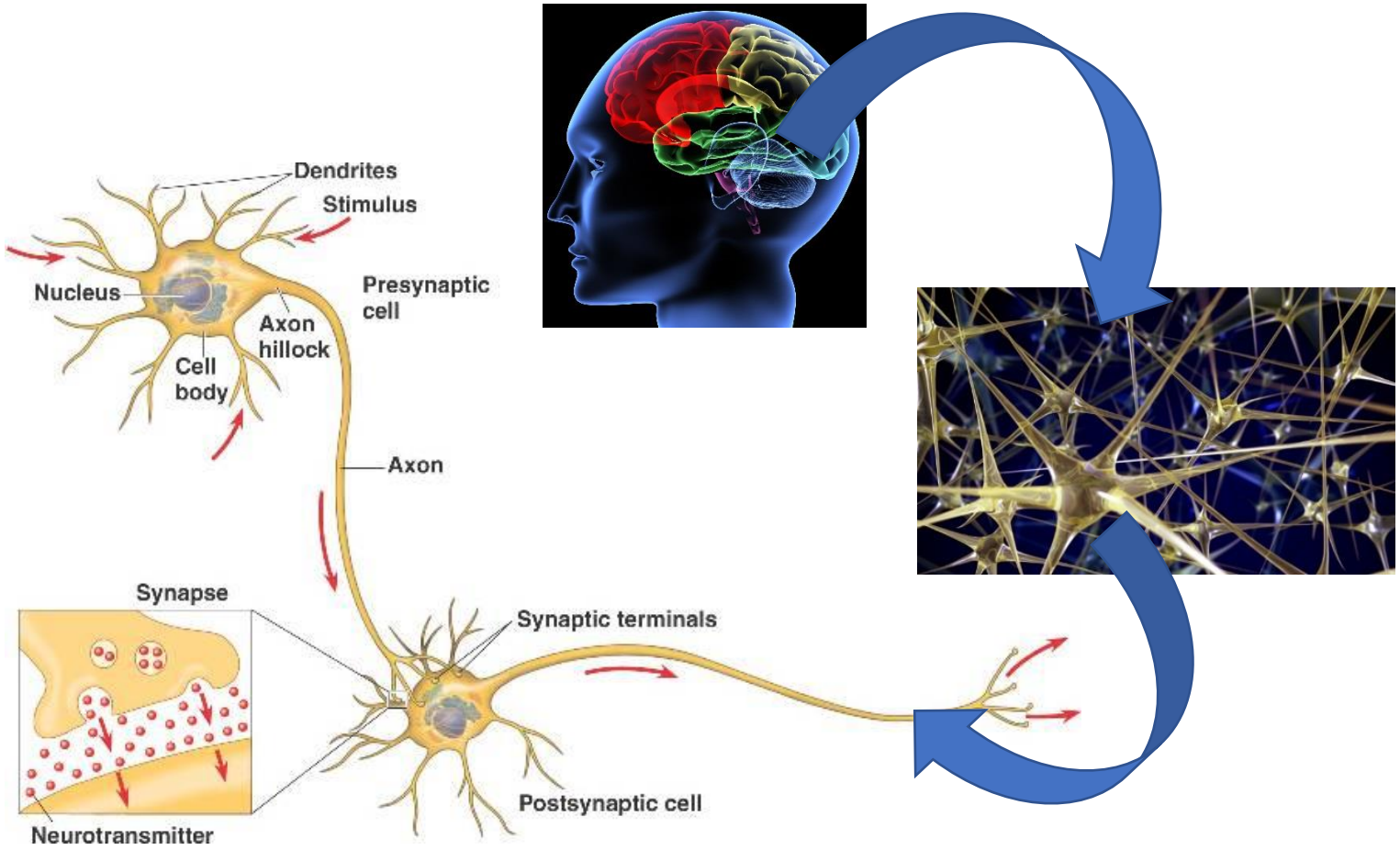



End-to-end training:

What each function should do is learned automatically

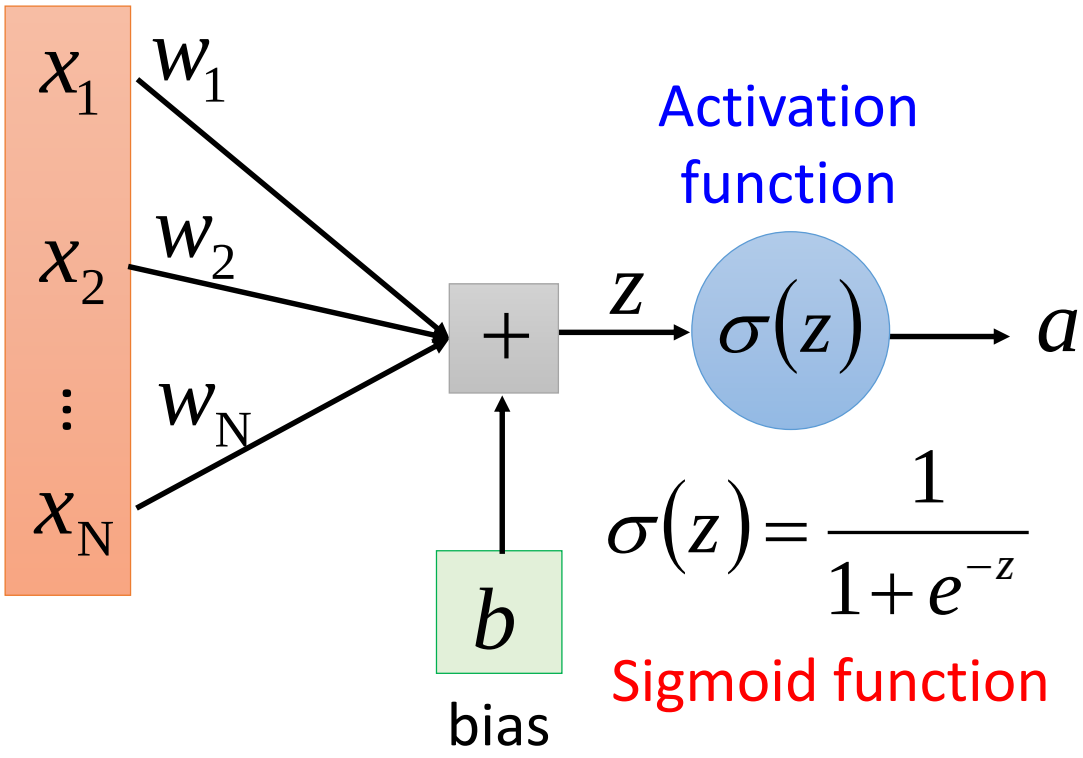
- Deep learning usually referred to neural network based approach

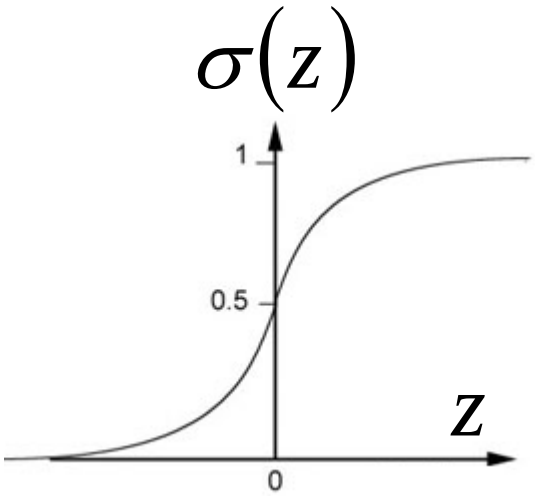
# Inspired from Human Brains



# A Neuron for Machine

Each neuron is a very simple function



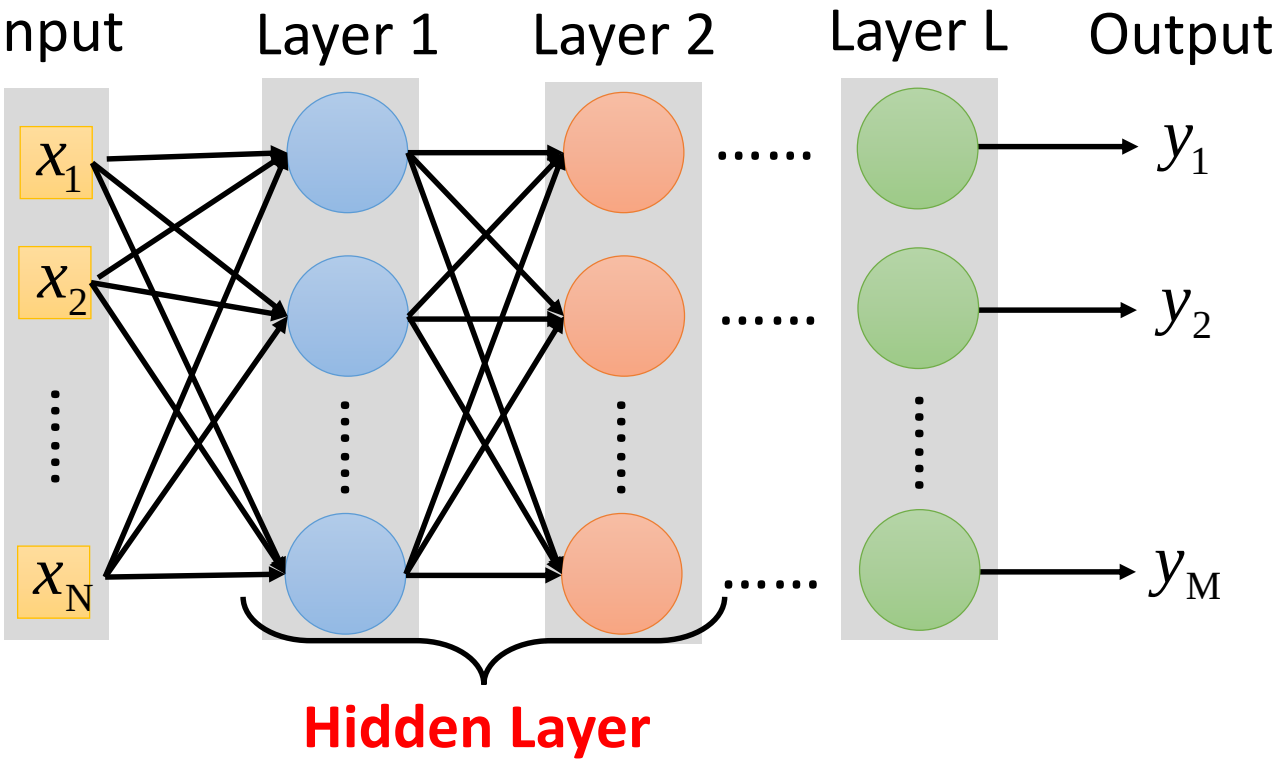
# Deep Learning

A neural network is a complex function:

$$f : R^N \rightarrow R^M$$

- Cascading the neurons to form a neural network.

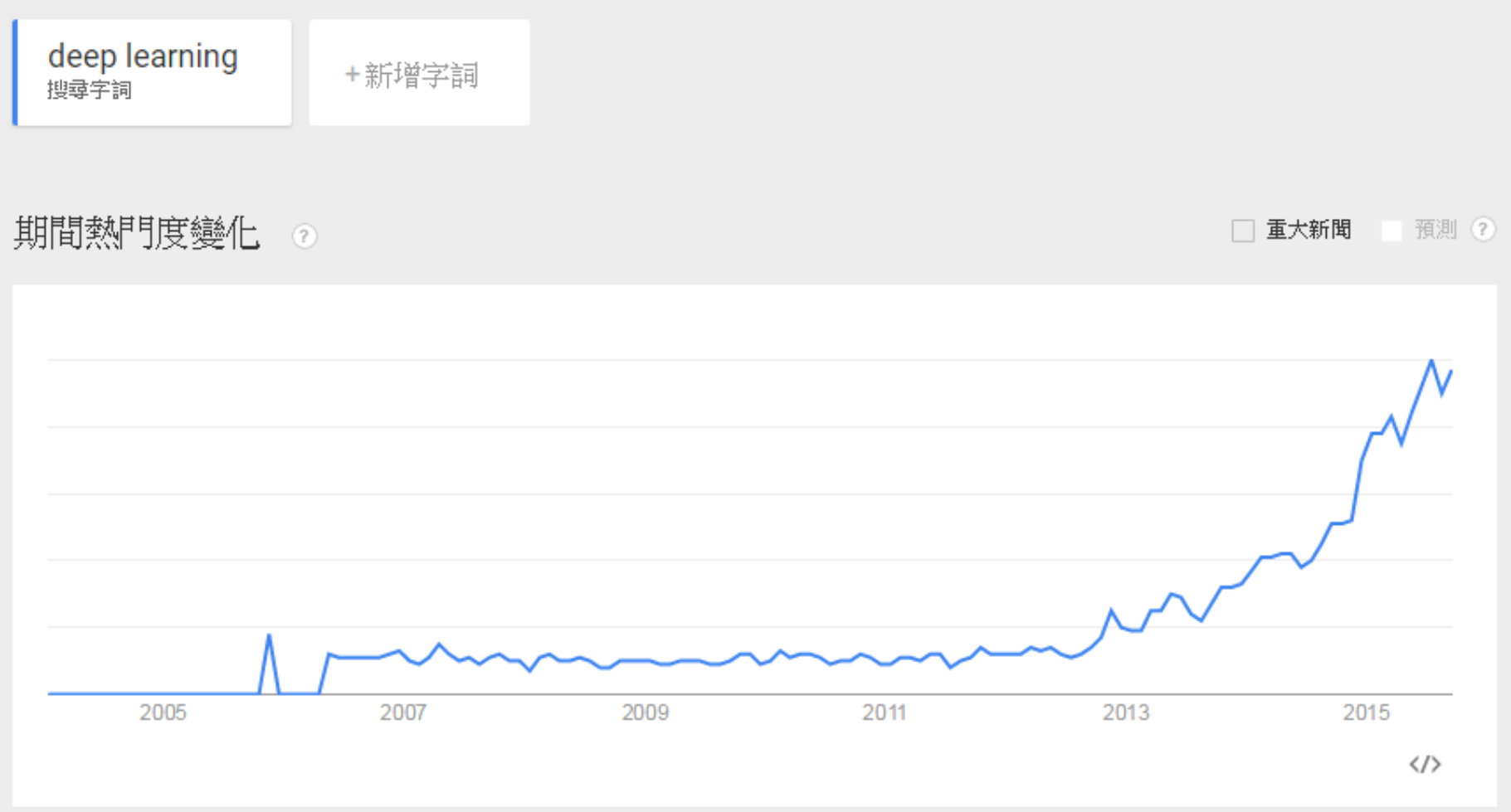
  Each layer is a simple function in the production line.



# Ups and downs of Deep Learning

- 1960s: Perceptron (single layer neural network)
- 1969: Perceptron has limitation
- 1980s: Multi-layer perceptron
  - Do not have significant difference from DNN today
- 1986: Backpropagation
  - Usually more than 3 hidden layers is not helpful
- 1989: 1 hidden layer is "good enough", why deep?
- 2006: RBM initialization (breakthrough)
- 2009: GPU
- 2011: Start to be popular in speech recognition
- 2012: win ILSVRC competition (image)

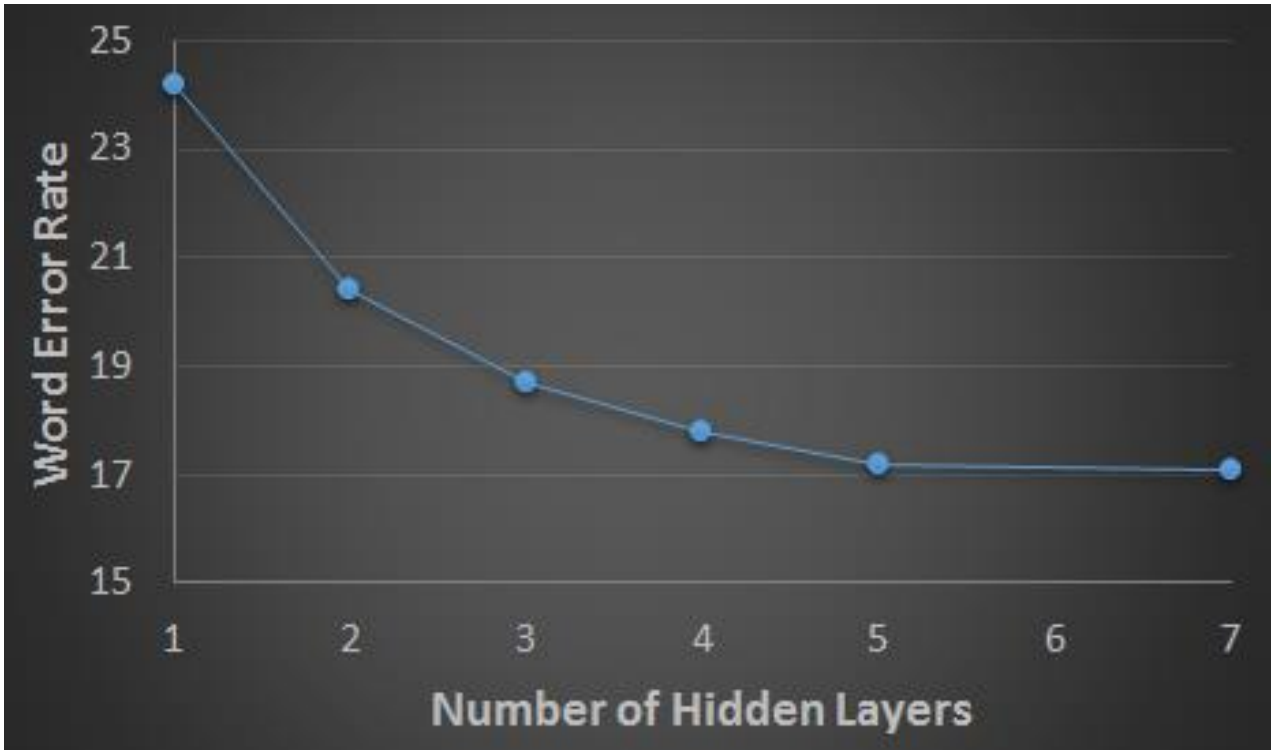
# Become very Popular



# Why Deep Learning?

Deeper is Better.

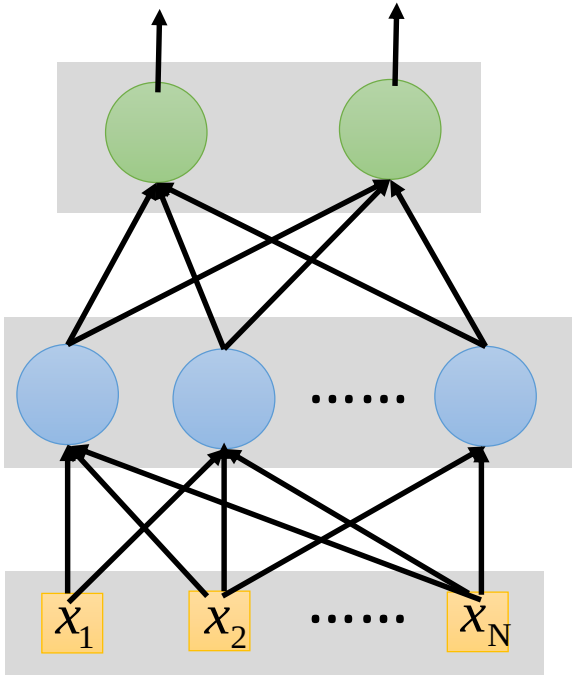
- Speech recognition



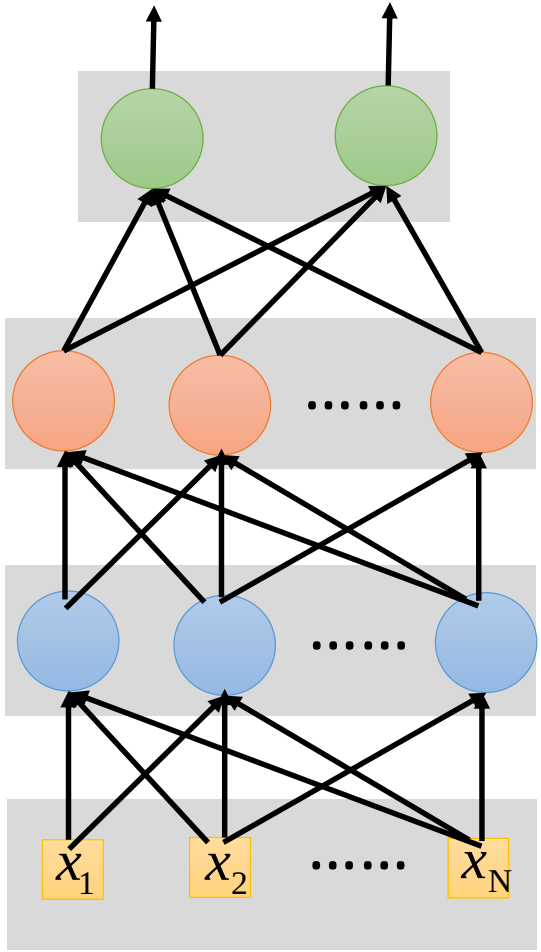

Seide, Frank, Gang Li, and Dong Yu. "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks." *Interspeech*. 2011.

# Why Deeper is Better?

Deep works better simply because it uses more parameters.




Shallow




Deep

# Universality Theorem

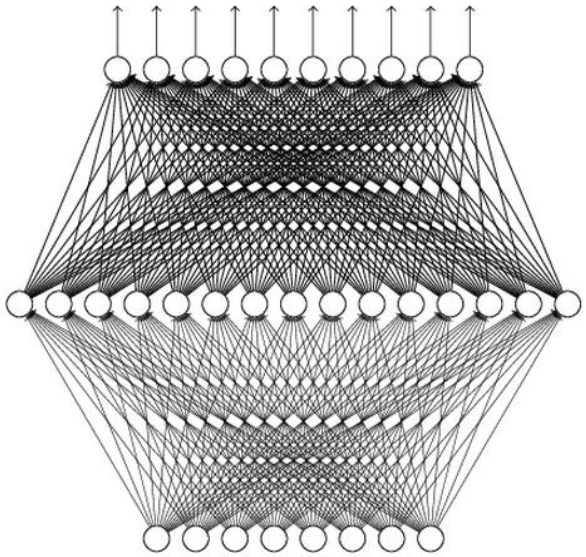
Any continuous function f

$$f : R^N \rightarrow R^M$$

Can be realized by a network with one hidden layer

(given **enough** hidden neurons)

What is the reason to be deep?



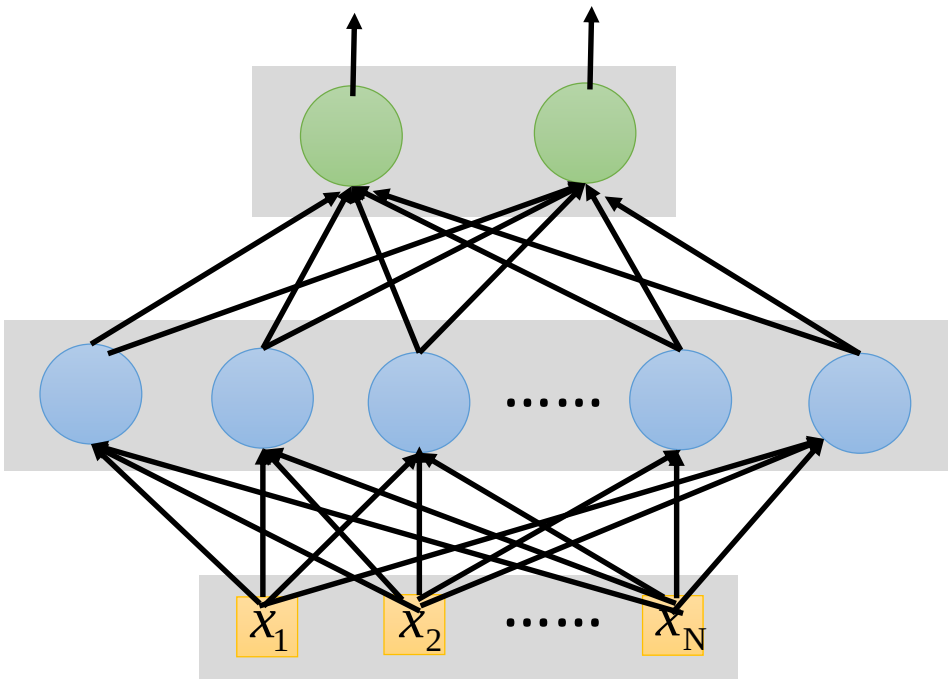
Reference:
http://neuralnetworksanddeeplearning.com/chap4.html

Why “Deep” neural network not “Fat” neural network?

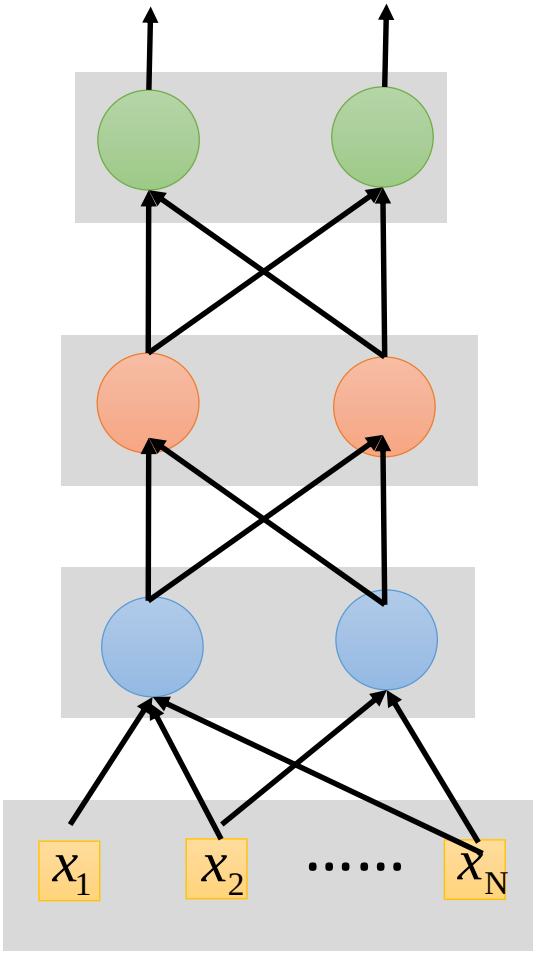
# Fat + Short v.s. Thin + Tall

If they have the same parameters,

Which one is better?




Shallow

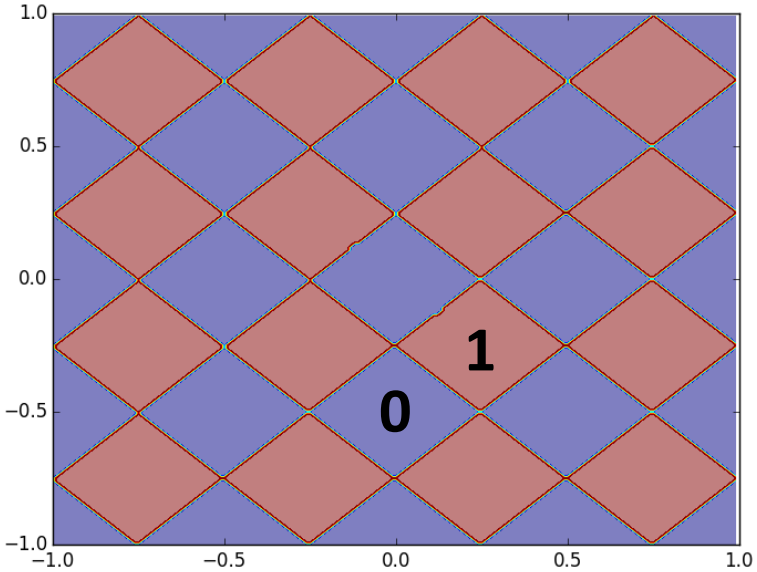


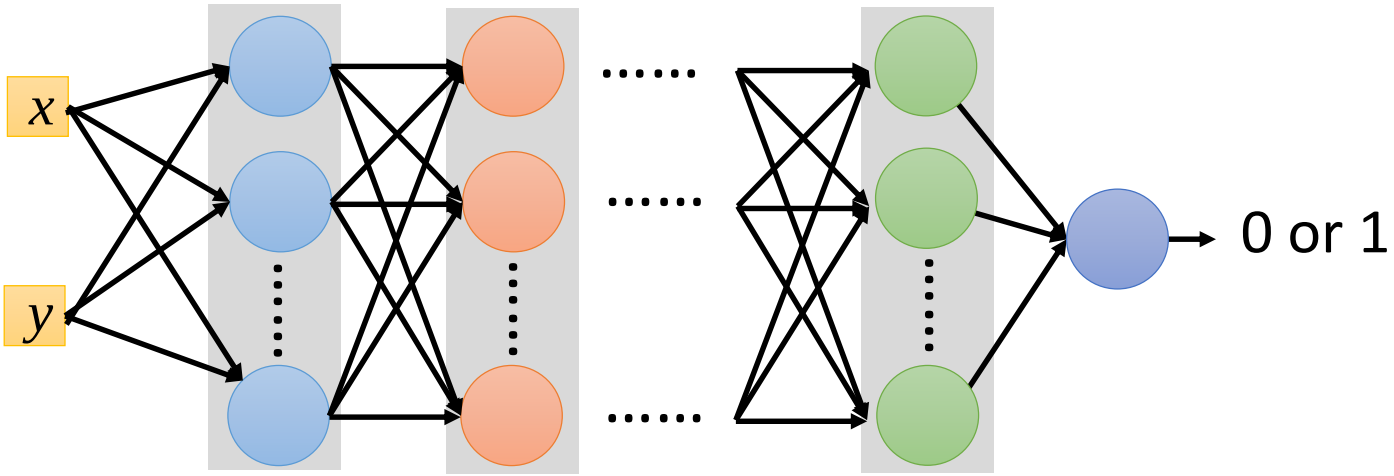

Deep

# Fat + Short v.s. Thin + Tall Toy Example

$f : R^2 \rightarrow \{0,1\}$

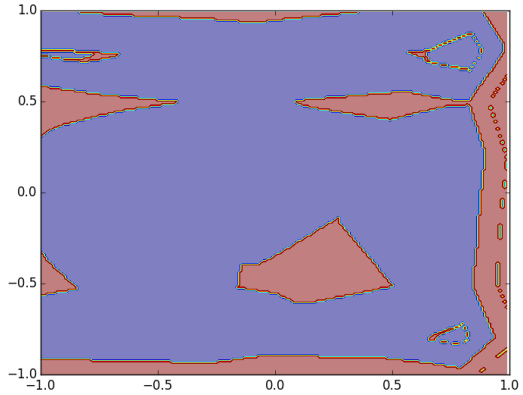
Sample 10,0000 points as training data



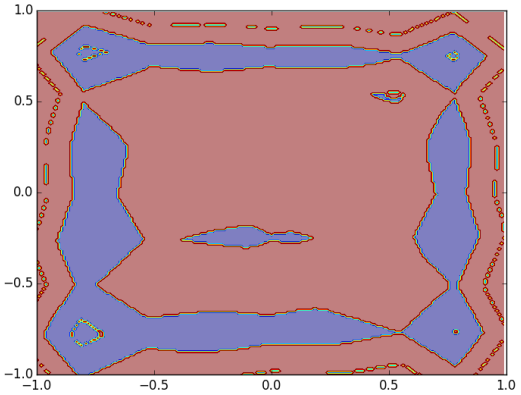



# Fat + Short v.s. Thin + Tall
# Toy Example

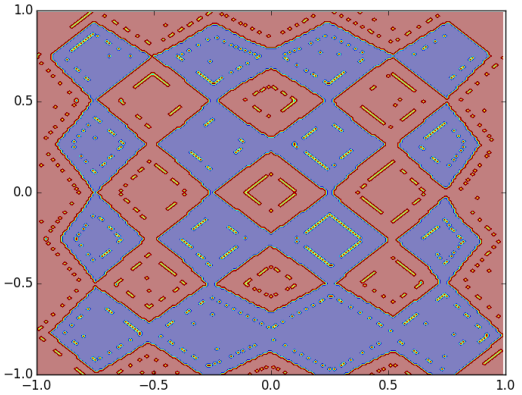
**1 hidden layer:**




(A) 125 neurons

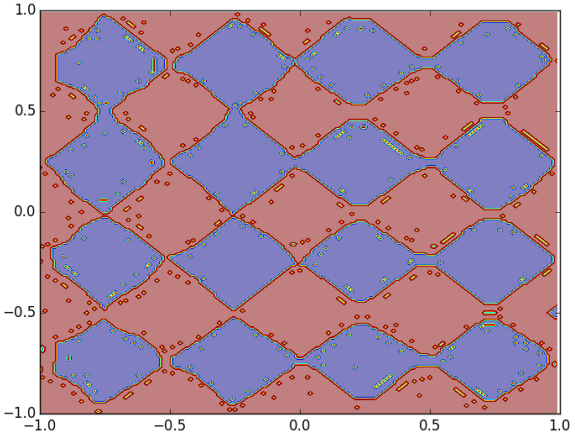



(B) 500 neurons




(C) 2500 neurons

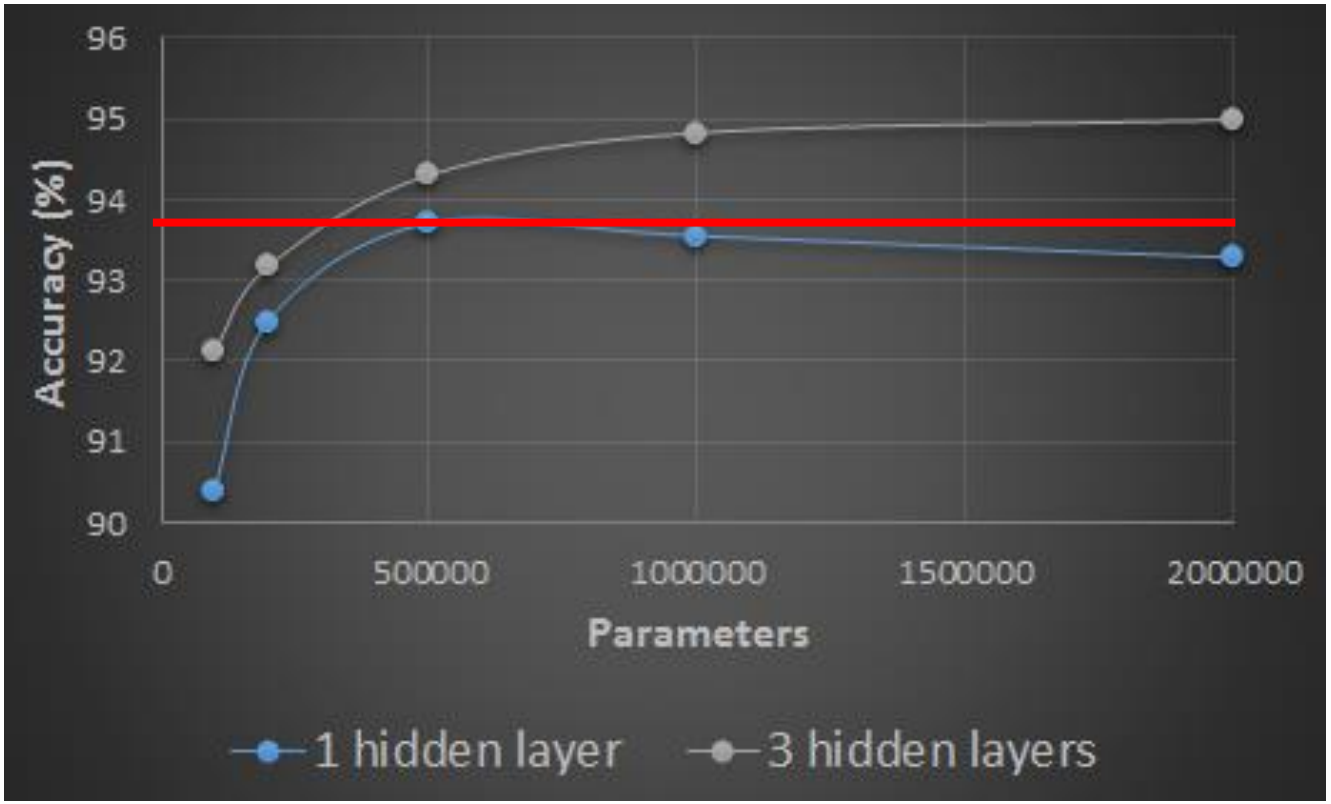
**3 hidden layers:**




Q: the number of parameters close to (A), (B) or (C)?

# Fat + Short v.s. Thin + Tall Hand-writing digit classification

- Same parameters




Deeper: Using less parameters to achieve the same performance

# Fat + Short v.s. Thin + Tall Speech Recognition

- Word error rate (WER)

Multiple layers

| LxN | DBN-PT (%) |
|---|---|
| 1×2k | 24.2 |
| 2×2k | 20.4 |
| 3×2k | 18.4 |
| 4×2k | 17.8 |
| 5×2k | 17.2 |
| 7×2k | 17.1 |
| | |

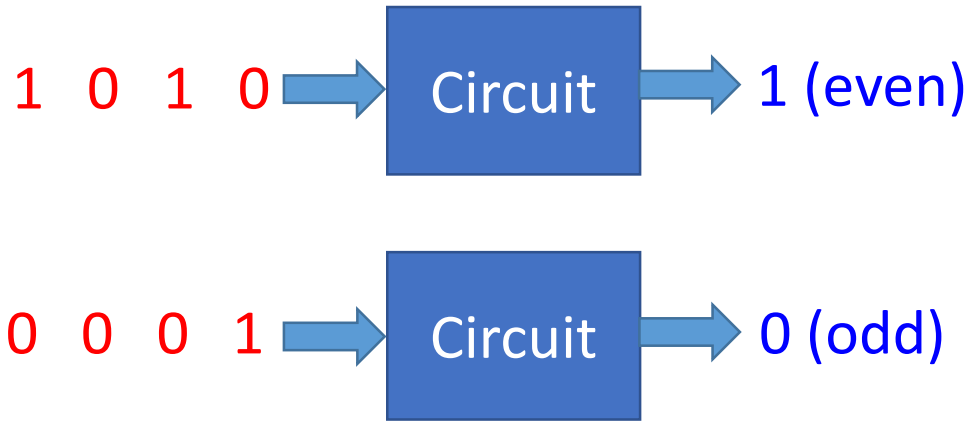
1 hidden layer

| 1xN | DBN-PT (%) |
|---|---|
| | |
| | |
| | |
| | |
| 1×3,772 | 22.5 |
| 1×4,634 | 22.6 |
| 1×16K | 22.1 |

Seide, Frank, Gang Li, and Dong Yu. "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks." *Interspeech*. 2011.

# Think about Logic Circuits ……

- A two-layer circuit of logic gates can represent any Boolean function.
- Using multiple layers of logic gates to build some functions are much simpler (less gates needed).
- E.g. ***parity check***

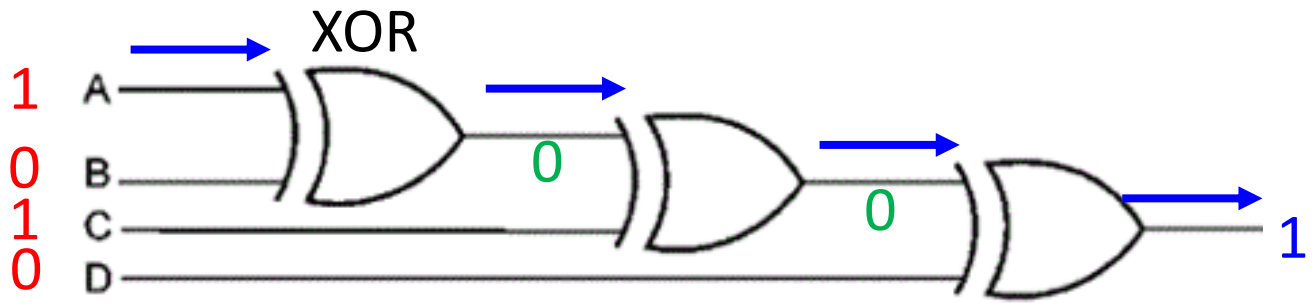



For input sequence with d bits,

Two-layer circuit need $O(2^d)$ gates.

# Think about Logic Circuits ……

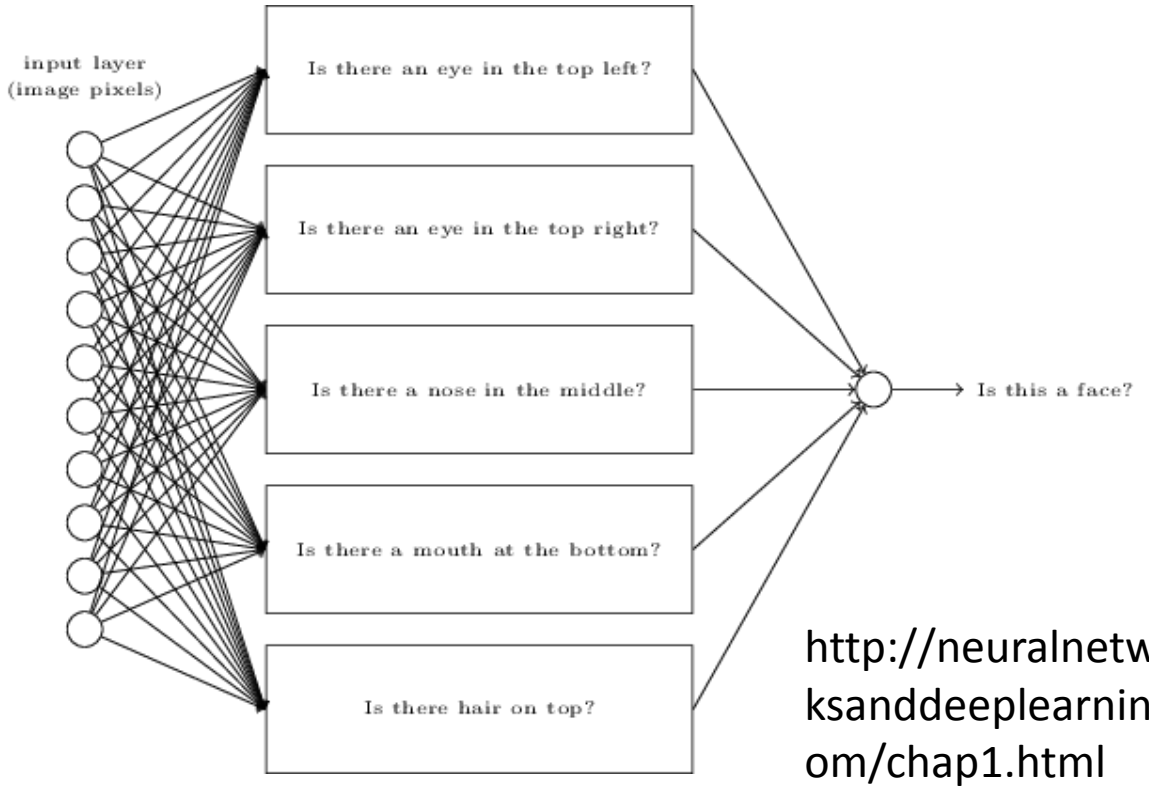
- A two-layer circuit of logic gates can represent any Boolean function.
- Using multiple layers of logic gates to build some functions are much simpler (less gates needed).
- E.g. ***parity check***




With multiple layers, we need only O(d) gates.

# Back to Deep Learning

- *Some functions* can be easily represented by deep structure
  - Perhaps the functions that can be naturally decomposed into several steps
  - E.g. image



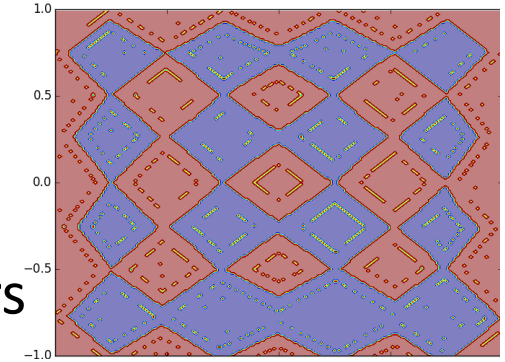
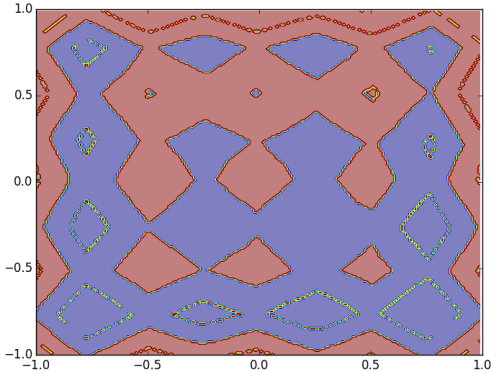
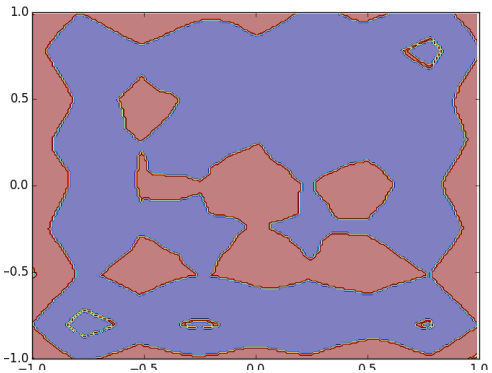
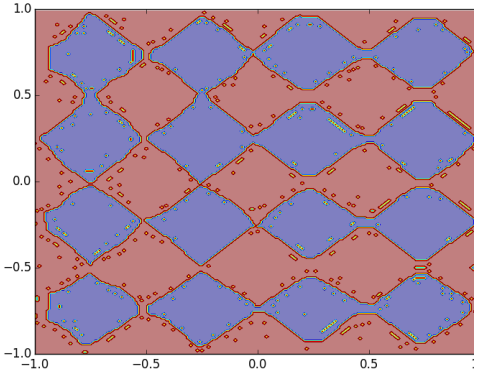
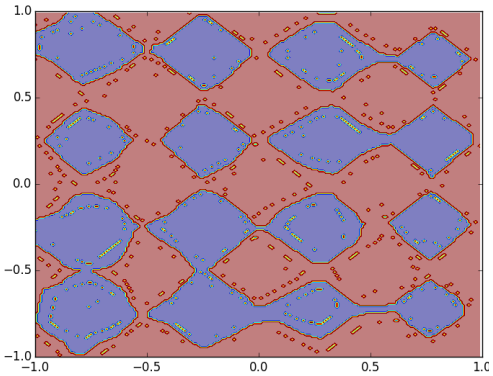
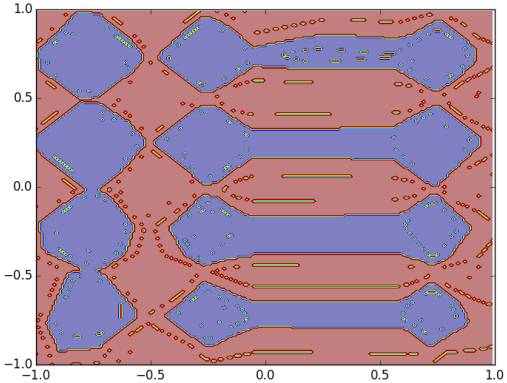

http://neuralnetworksanddeeplearning.com/chap1.html

# Back to Deep Learning

- *Some functions* can be easily represented by deep structure
  - Perhaps the functions that can be naturally decomposed into several steps
  - E.g. image
- To represent the functions with shallow structure needs much more parameters
  - More parameters imply more training data needed
- To achieve the same performance, deep learning needs less training data
- With the same amount of data, deep learning can achieve better performance.

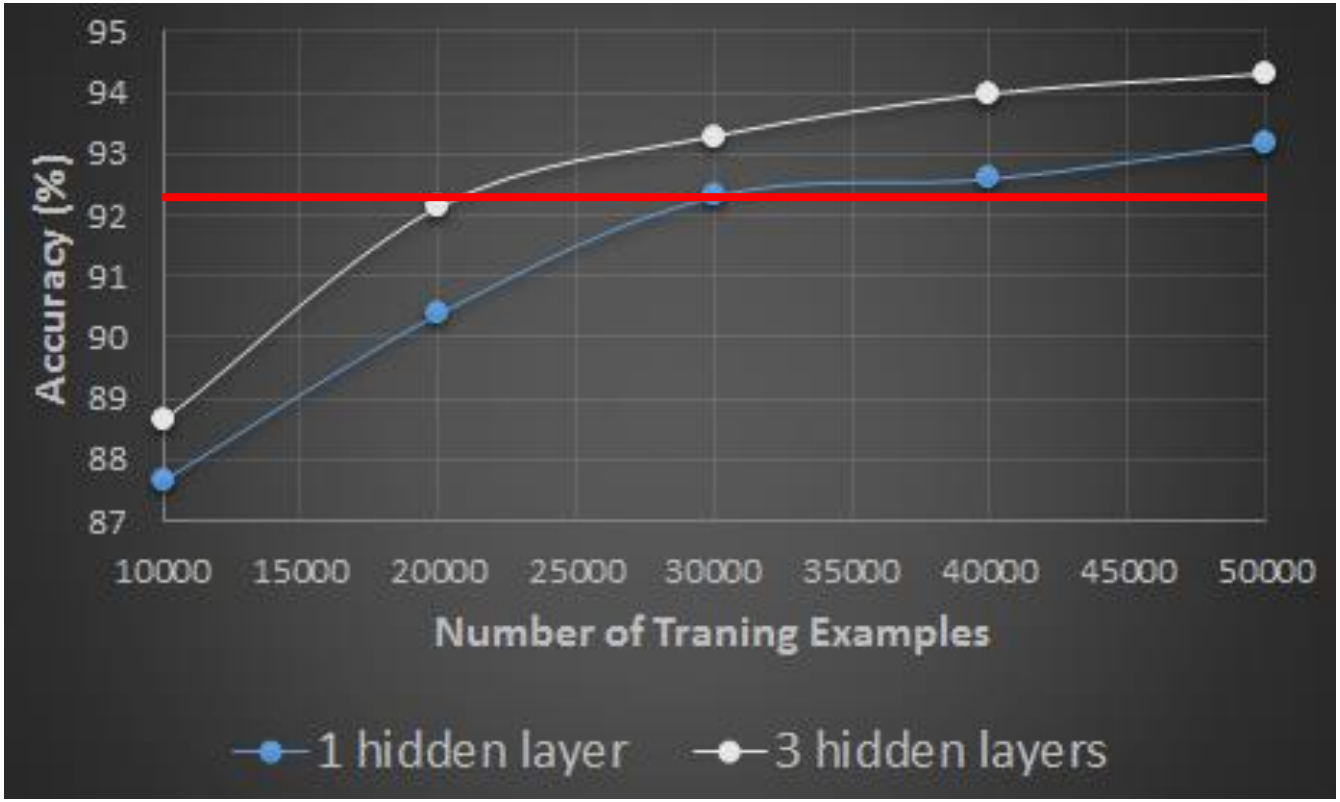
# Size of Training Data

- Different numbers of training examples

| | 10,0000 | 5,0000 | 2,0000 |
|---|---|---|---|
| **1 hidden layer** More parameters |  |  |  |
| **3 hidden layers** Less parameters |  |  |  |

# Size of Training Data

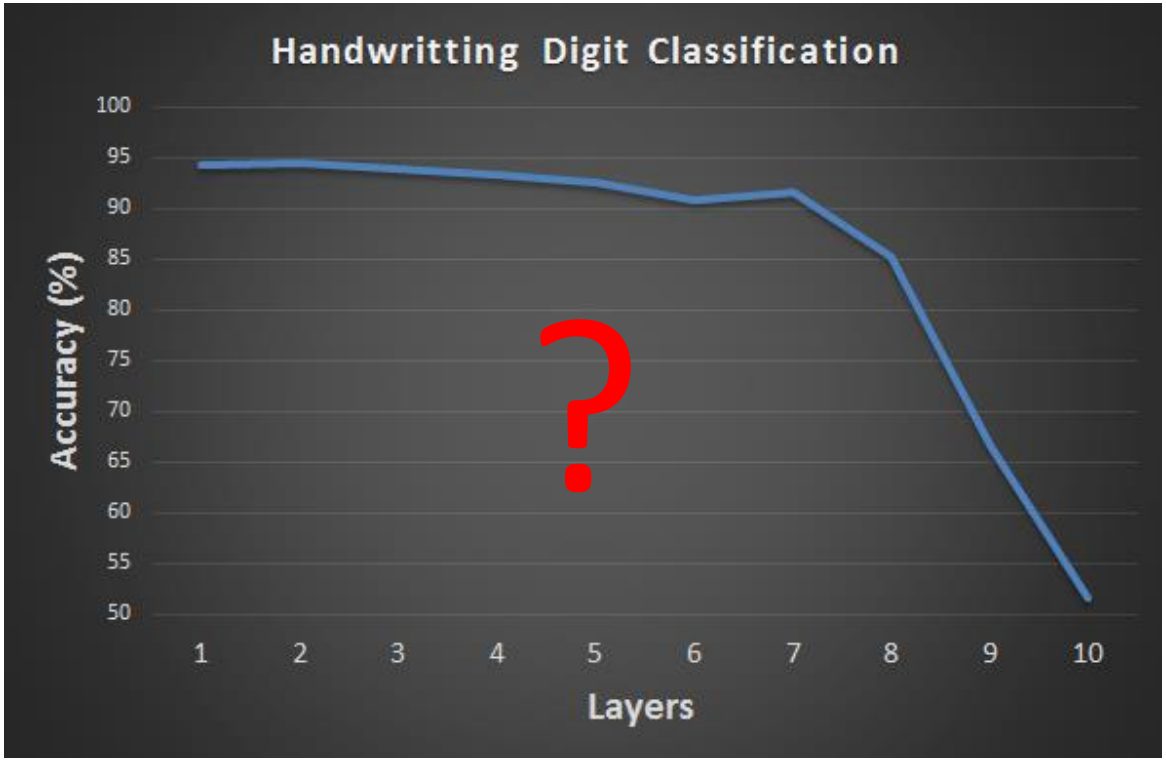
- Hand-writing digit classification




Deeper: Using less training data to achieve the same performance

# Why deep is not popular before?

- In the past, usually deep does not work ……




We will go back to this issue in the following lectures.

# What is Structured Learning?

# In the real world ……

$$f : X \rightarrow Y$$

***X* (Input domain)**:
Sequence, graph structure, tree structure ……

***Y* (Output domain)**:
Sequence, graph structure, tree structure ……

# Retrieval

$$f : X \rightarrow Y$$

$X$ :

"Machine learning"

(keyword)

$Y$ :


A list of web pages (Search Result)

# Translation

$$f : X \rightarrow Y$$

$X$ :

“Machine learning and having it deep and structured”

(One kind of sequence)

$Y$ :

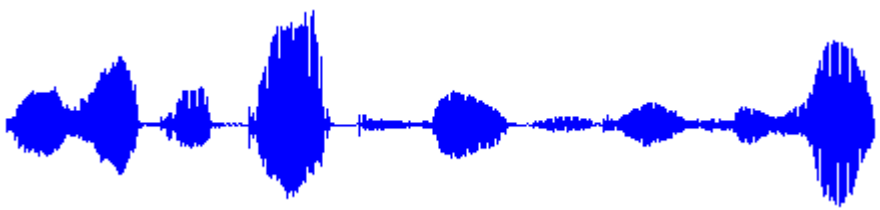
“機器學習及其深層與結構化”

(Another kind of sequence)

# Speech Recognition

$$f : X \rightarrow Y$$

$X$ :



(One kind of sequence)

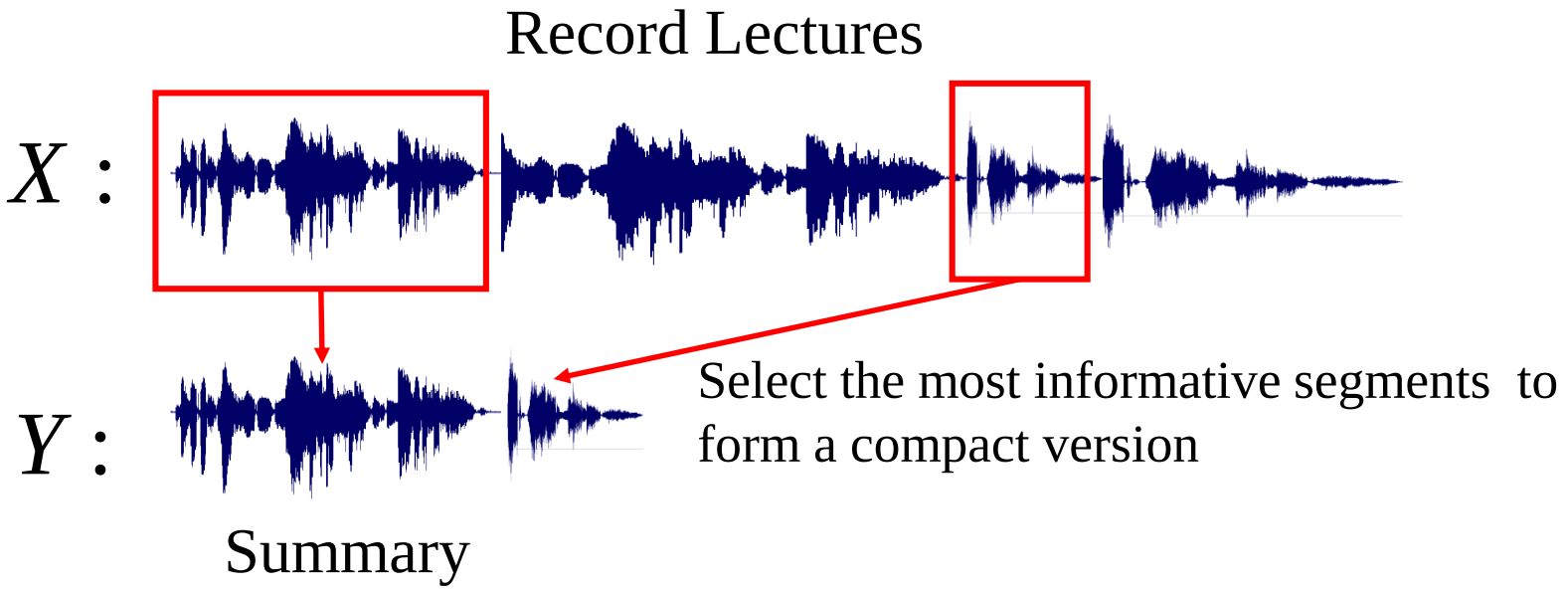
$Y$ :

“大家好，歡迎大家來修機器學習及其深層與結構化”

(Another kind of sequence)

# Speech Summarization

$$f : X \rightarrow Y$$

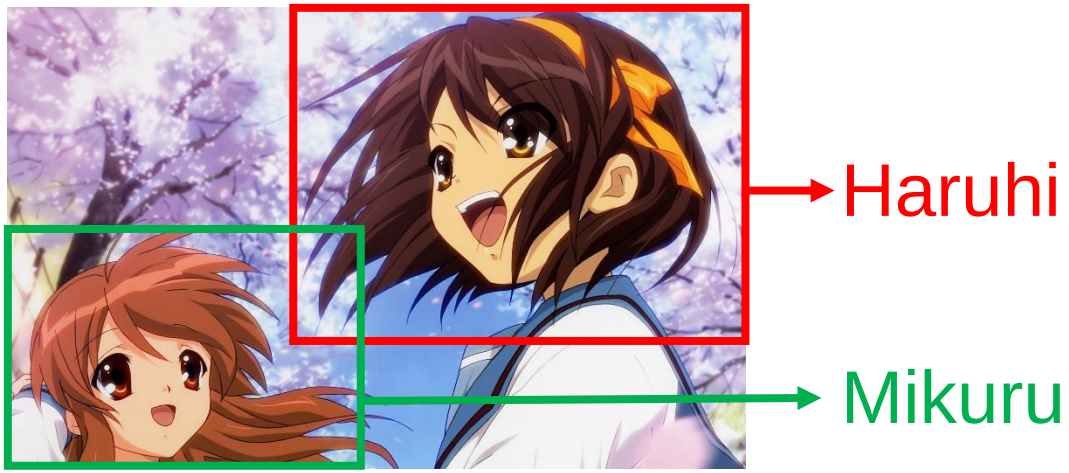


# Object Detection

$$f : X \rightarrow Y$$

$X$ : Image

$Y$ : Object Positions



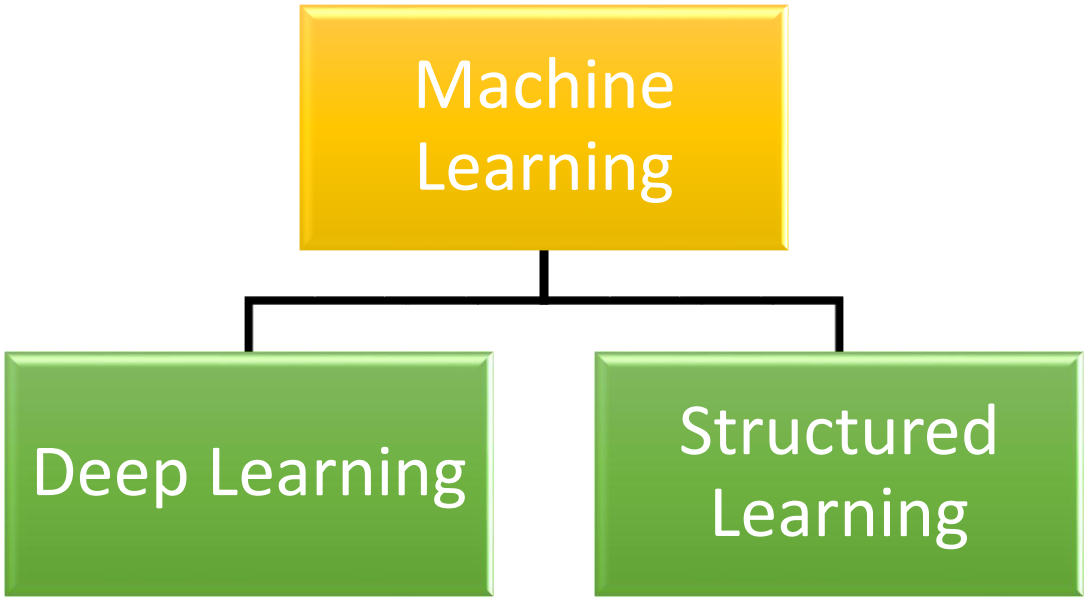
# Pose Estimation

$$f : X \rightarrow Y$$

$X$ : Image $Y$ : Pose

Source of images: http://groups.inf.ed.ac.uk/calvin/Publications/eichner-techreport10.pdf

# Concluding Remarks



# Reference

- No Textbook
- Deep Learning
  - “Neural Networks and Deep Learning”
    - written by Michael Nielsen
    - http://neuralnetworksanddeeplearning.com/
  - “Deep Learning” (not finished yet)
    - Written by Yoshua Bengio, Ian J. Goodfellow and Aaron Courville
    - http://www.iro.umontreal.ca/~bengioy/dlbook/
- Structured Learning
  - No suggested reference

Thank you!

# Human Brains are Deep