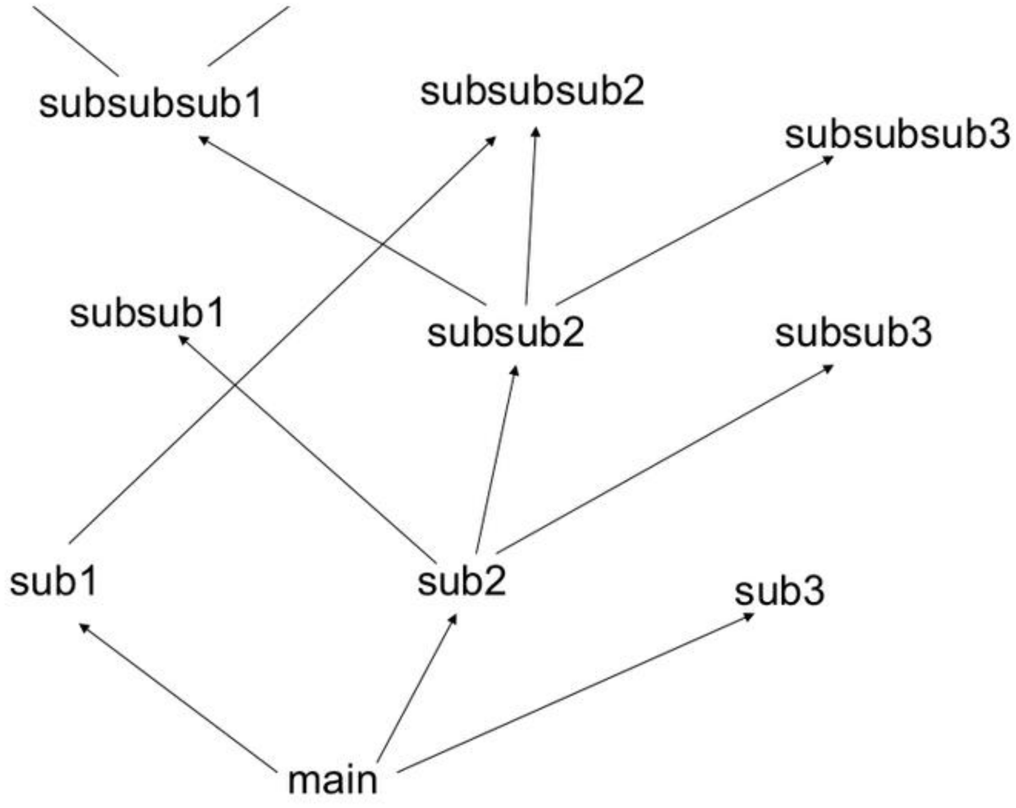
Why Deep Learning?

Modularization

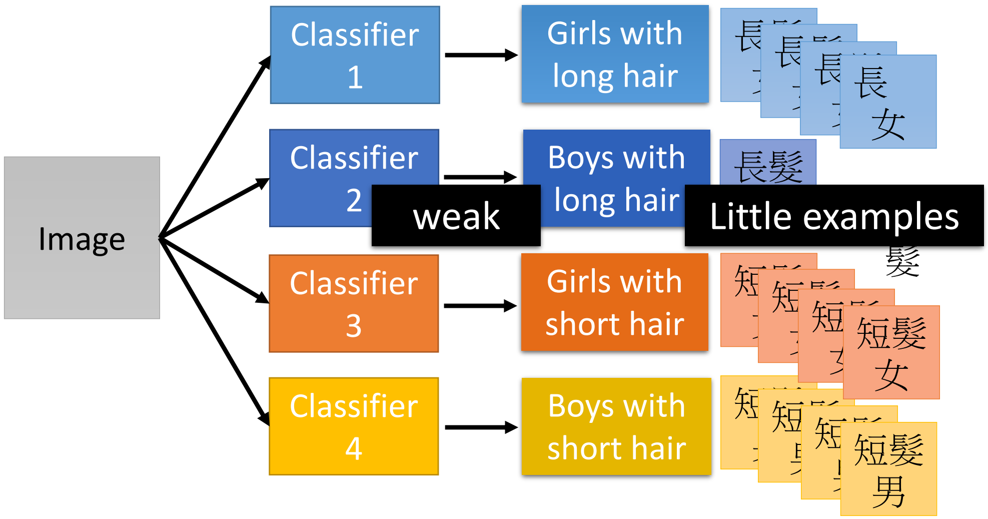
- Deep → Modularization

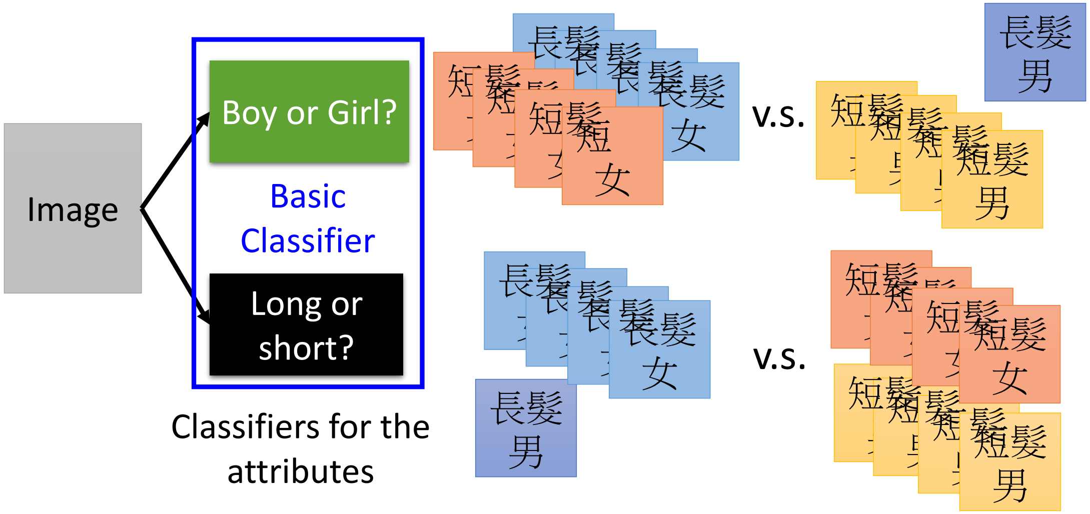
Don't put everything in your main function.



http://rinuboney.github.io/2015/10/18/theoretical-motivations-deep-learning.html

Modularization

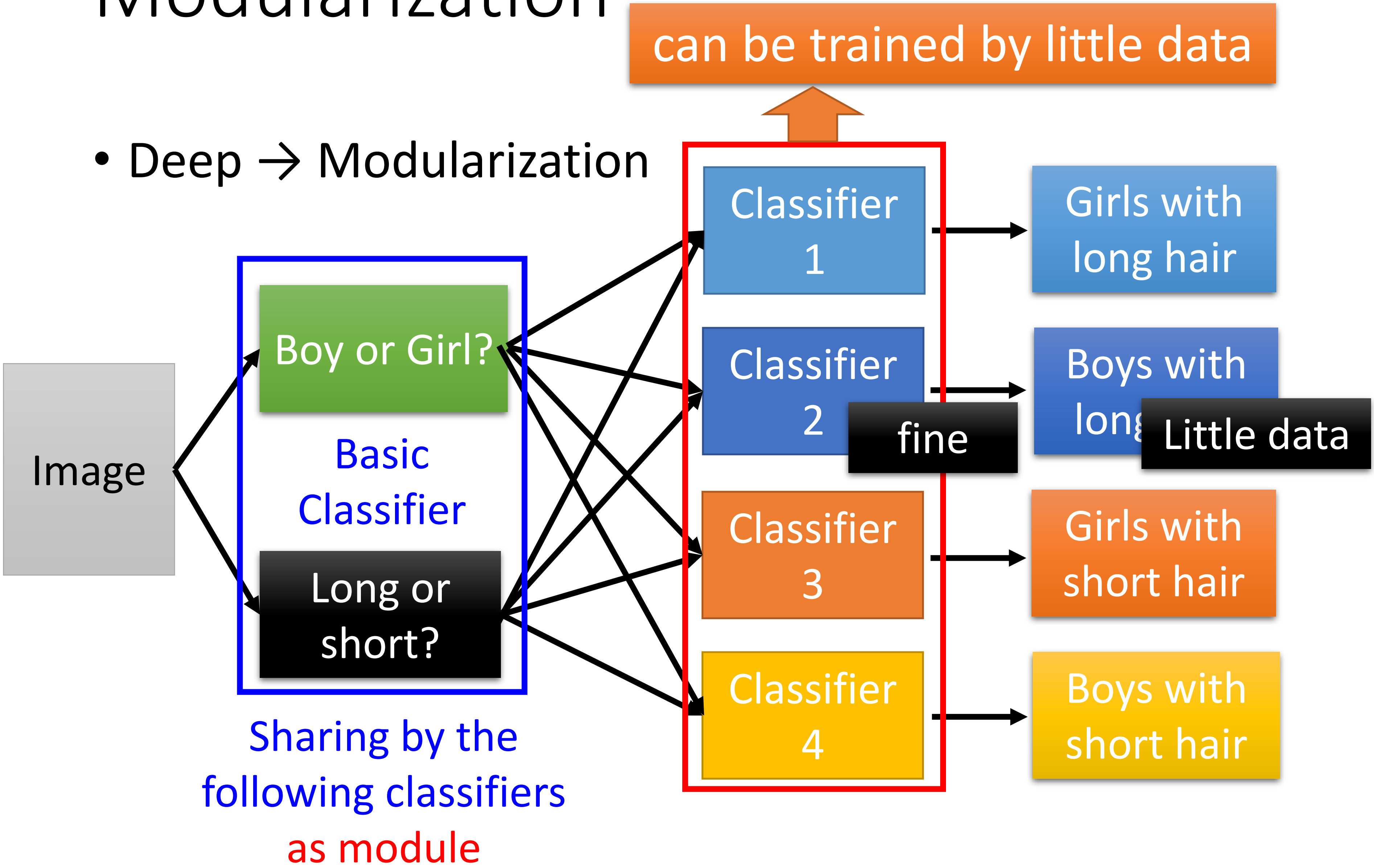




- Deep→Modularization
- Each basic classifier can have sufficient training examples.

Modularization

- Deep → Modularization



can be trained by little data

Image

Boy or Girl?

Basic Classifier

Long or short?

Sharing by the following classifiers as module

Classifier 1

Classifier 2

fine

Classifier 3

Classifier 4

Girls with long hair

Boys with long

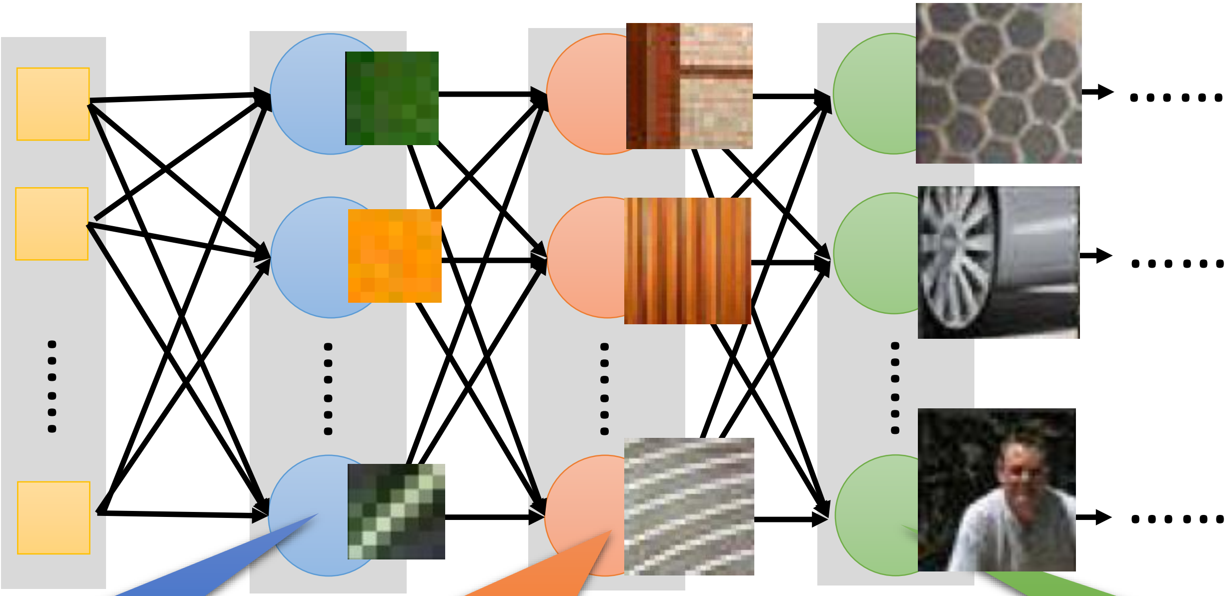
Little data

Girls with short hair

Boys with short hair

Modularization - Image

- Deep → Modularization



The most basic classifiers

Use 1st layer as module to build classifiers

Use 2nd layer as module ……

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

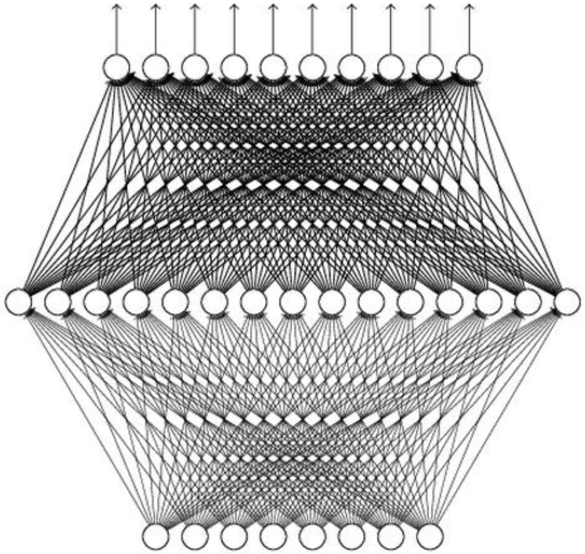
Universality Theorem

Any continuous function f

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

Can be realized by a network with one hidden layer

(given **enough** hidden neurons)

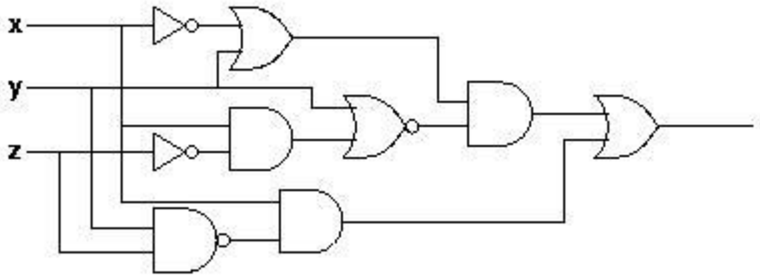


Reference for the reason: http://neuralnetworksanddeeplearning.com/chap4.html

Yes, shallow network can represent any function.

However, using deep structure is more effective.

Analogy



Logic circuits

- Logic circuits consists of **gates**
- **A two layers of logic gates** can represent **any Boolean function.**
- Using multiple layers of logic gates to build some functions are much simpler

→ less gates needed

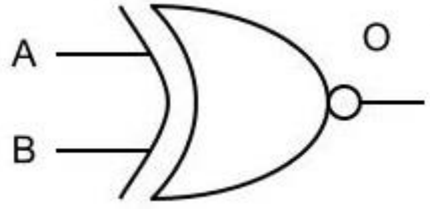
Neural network

- Neural network consists of **neurons**
- **A hidden layer network** can represent **any continuous function.**
- Using multiple layers of neurons to represent some functions are much simpler

→ less parameters → less data?

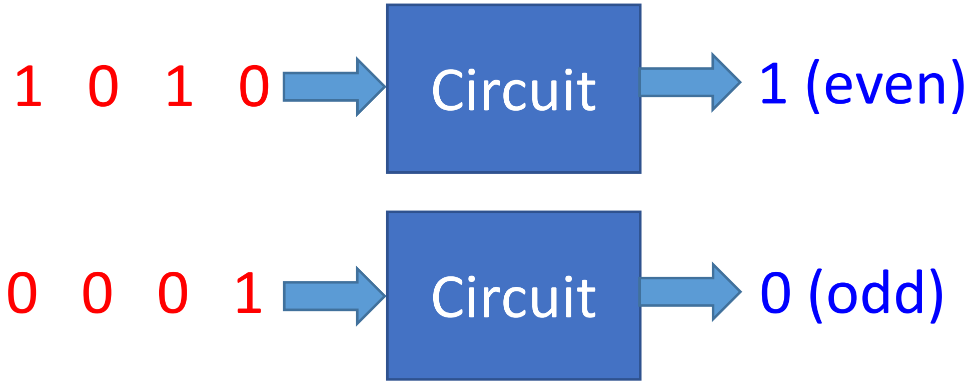
This page is for EE background.

Analogy



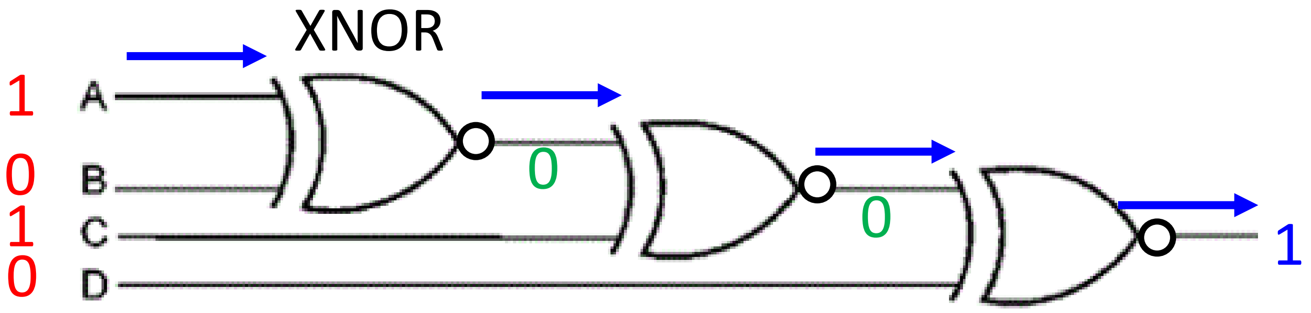
A	B	O
0	0	1
0	1	0
1	0	0
1	1	1

- E.g. ***parity check***



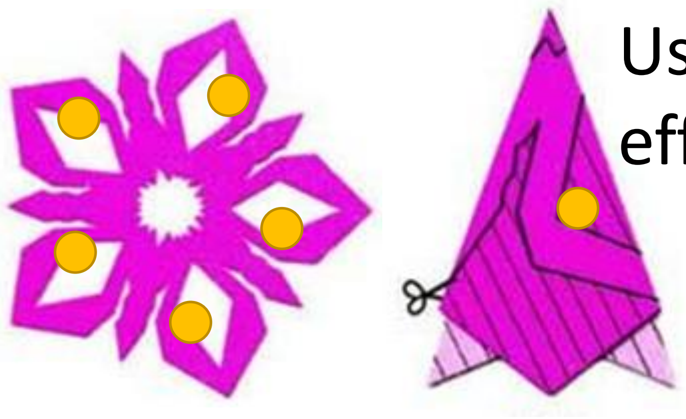
For input sequence with d bits,

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

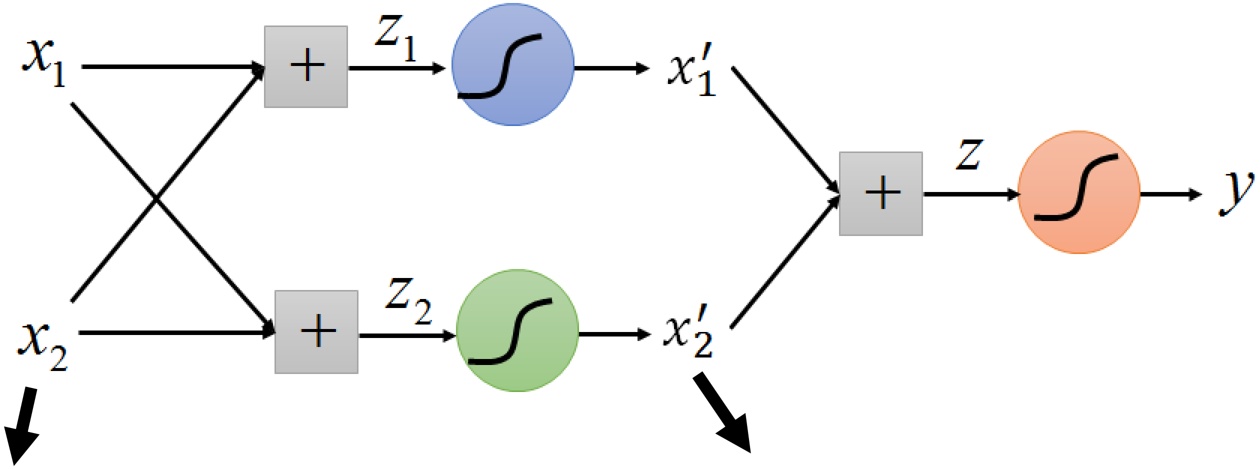


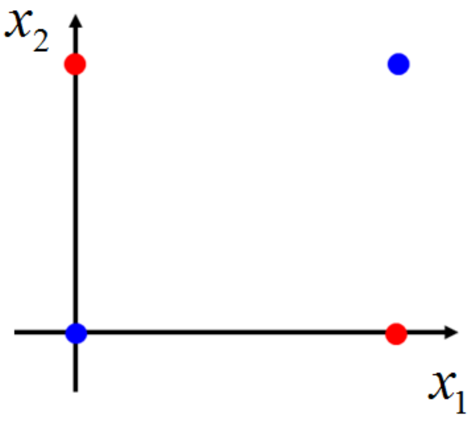
With multiple layers, we need only O(d) gates.

More Analogy

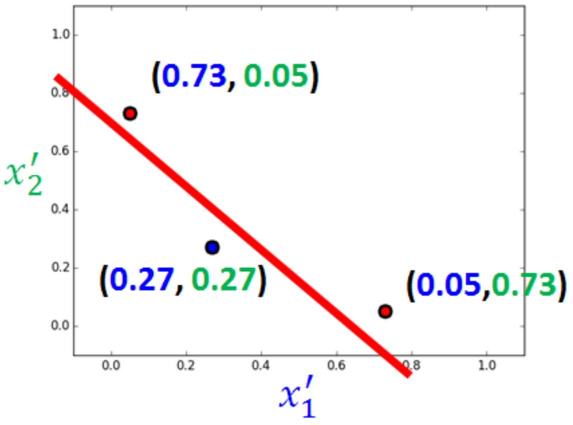


Use data effectively



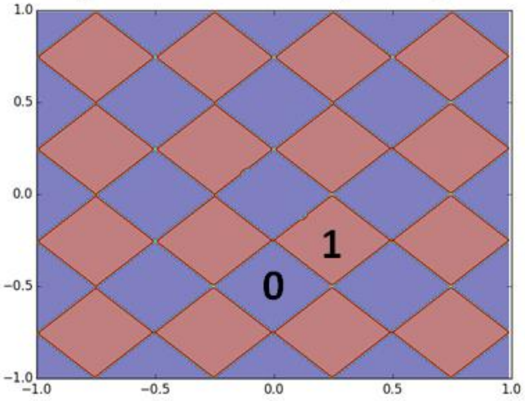


Folding the space



More Analogy - Experiment

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



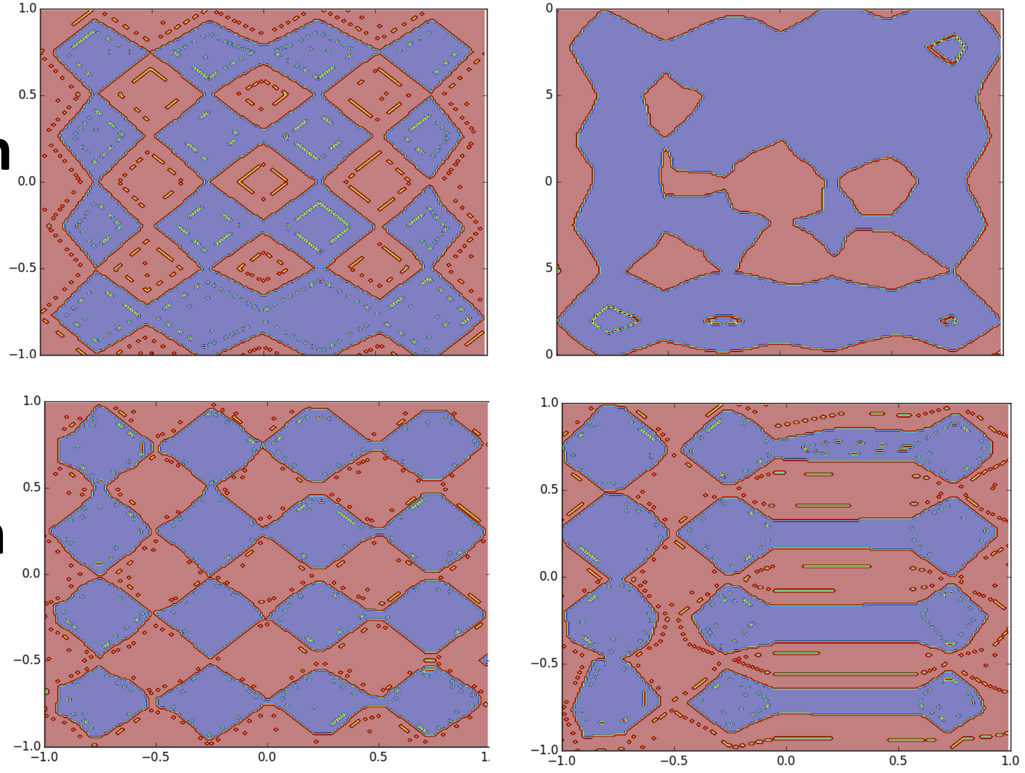
Different numbers of training examples

10,0000

2,0000

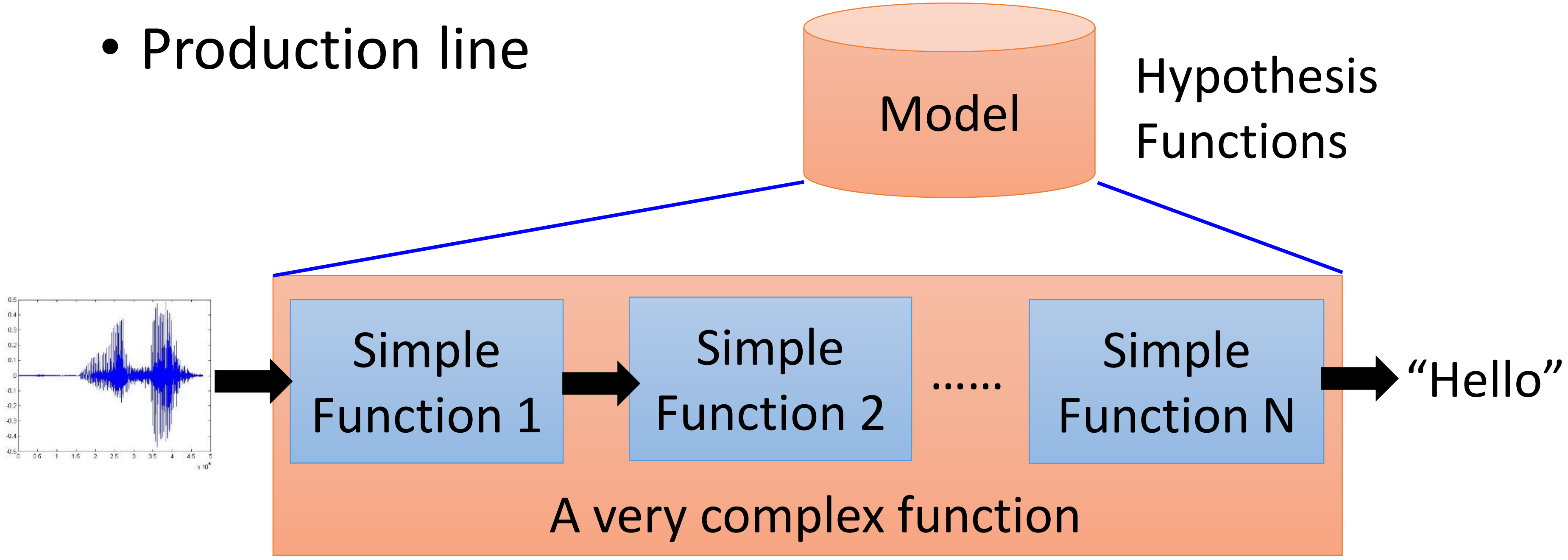
1 hidden layer

3 hidden layers



End-to-end Learning

- Production line



Model

Hypothesis Functions

Simple Function 1

Simple Function 2

……

Simple Function N

A very complex function

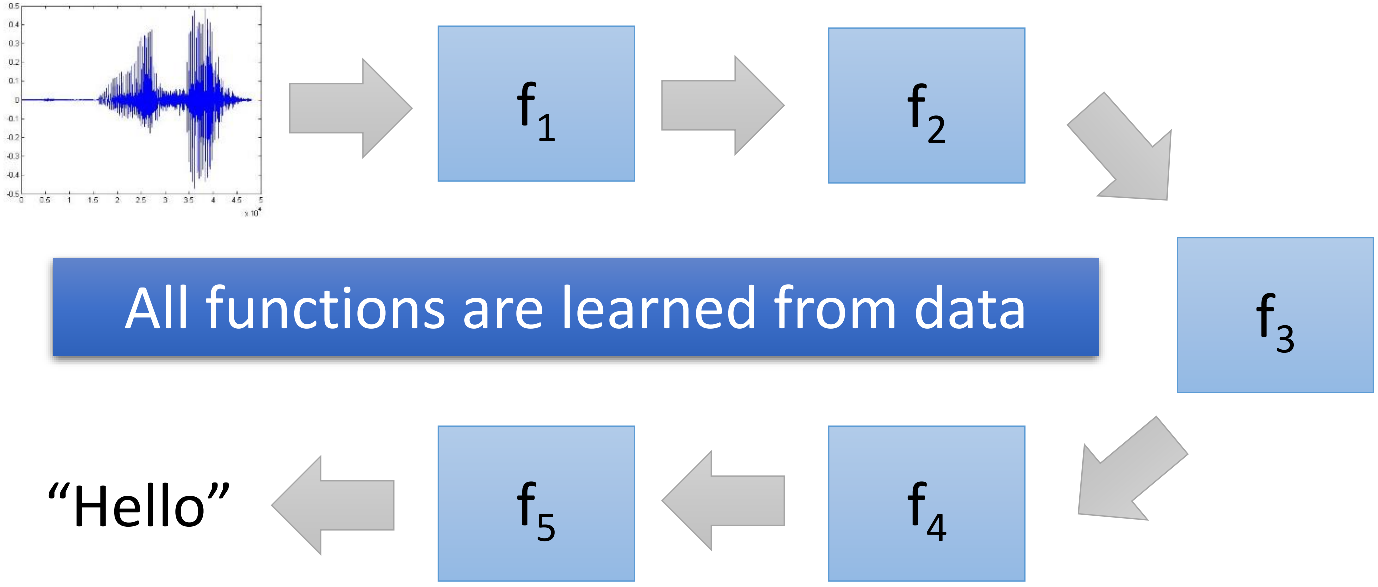
“Hello”

End-to-end training:

What each function should do is learned automatically

End-to-end Learning
- Speech Recognition

- Deep Learning



f_1

f_2

f_3

f_4

f_5

“Hello”

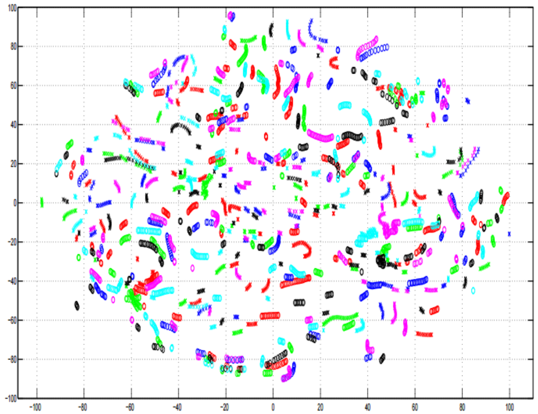
All functions are learned from data

Less engineering labor, but machine learns more

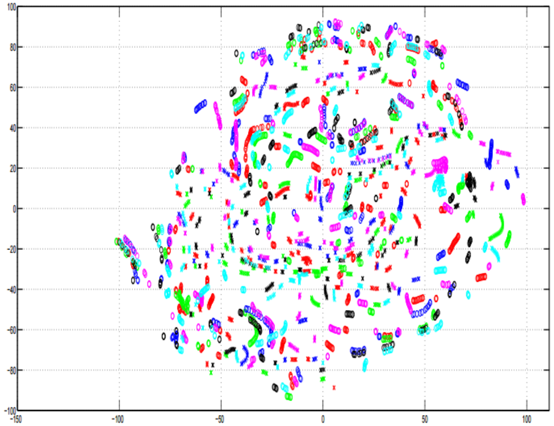
Complex Task …

A. Mohamed, G. Hinton, and G. Penn, “Understanding how Deep Belief Networks Perform Acoustic Modelling,” in ICASSP, 2012.

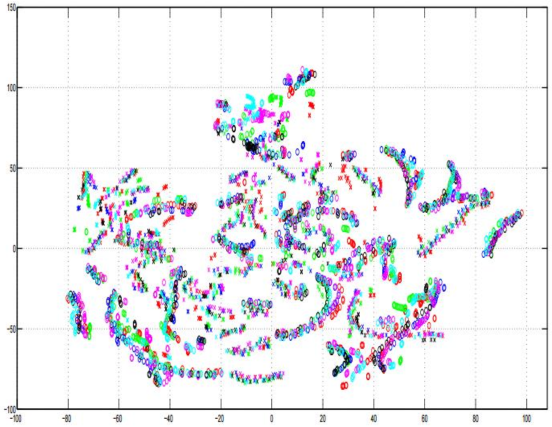
- Speech recognition: Speaker normalization is automatically done in DNN



Input Acoustic Feature (MFCC)



1-st Hidden Layer



8-th Hidden Layer