

Why Deep Learning?

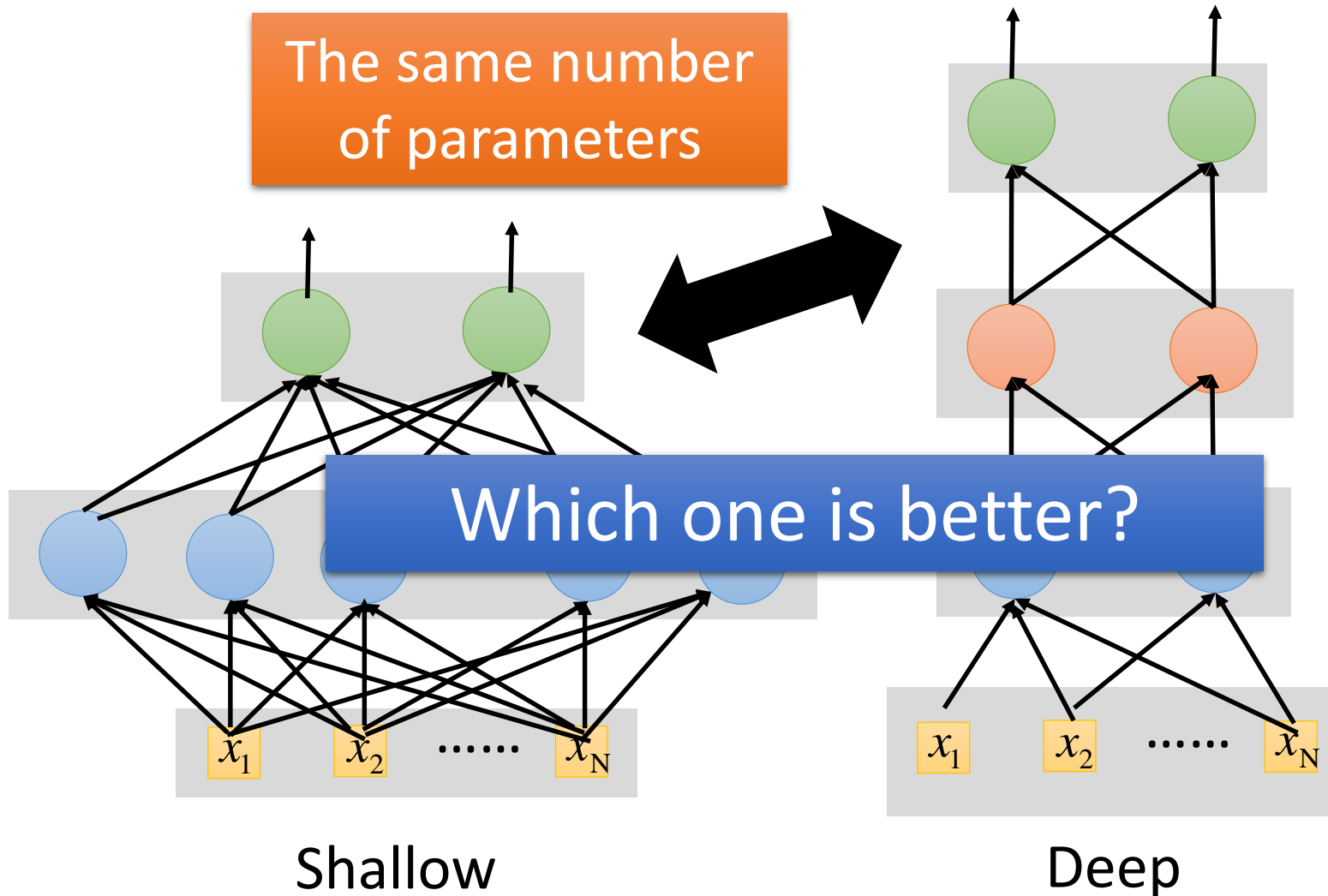
Deeper is Better?

Layer X Size	Word Error Rate (%)
1 X 2k	24.2
2 X 2k	20.4
3 X 2k	18.4
4 X 2k	17.8
5 X 2k	17.2
7 X 2k	17.1

Not surprised, more parameters, better performance

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

Fat + Short v.s. Thin + Tall



Fat + Short v.s. Thin + Tall

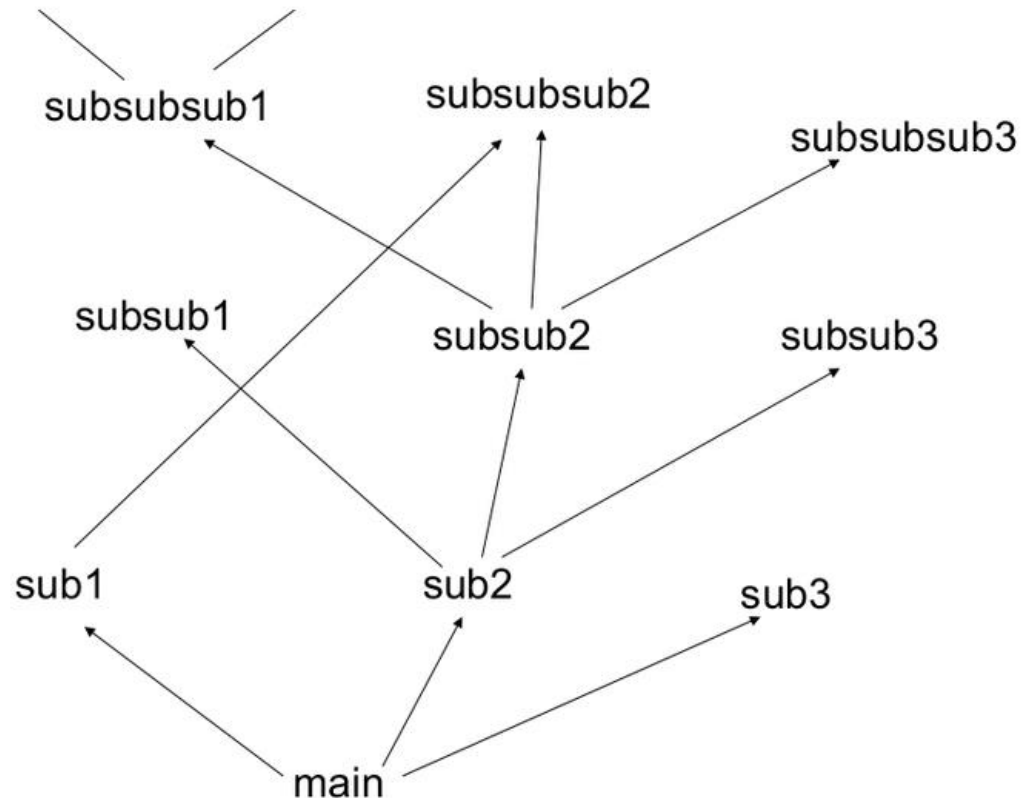
Layer X Size	Word Error Rate (%)	Layer X Size	Word Error Rate (%)
1 X 2k	24.2		
2 X 2k	20.4		
3 X 2k	18.4		
4 X 2k	17.8		
5 X 2k	17.2	1 X 3772	22.5
7 X 2k	17.1	1 X 4634	22.6
		1 X 16k	22.1

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

Modularization

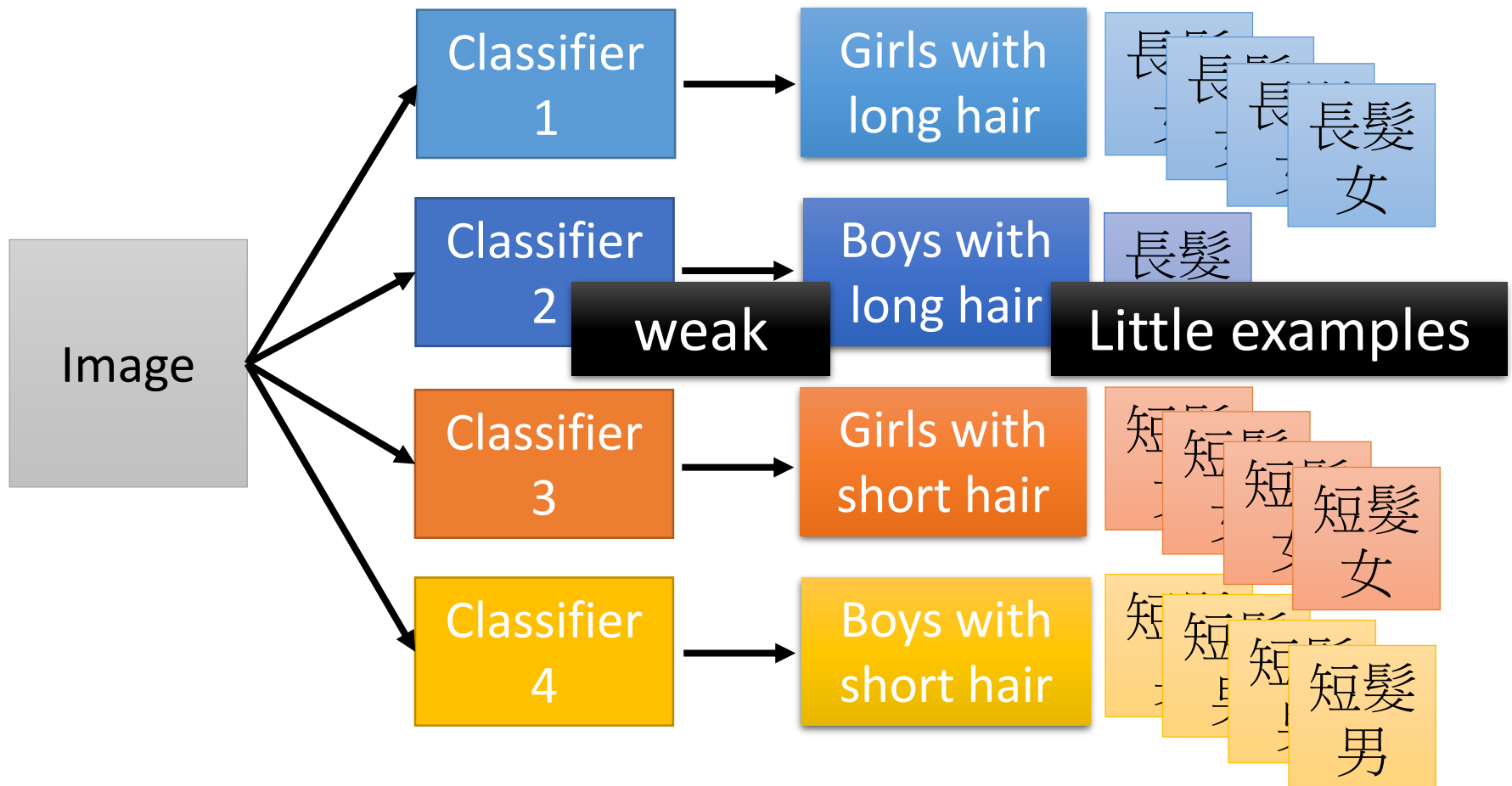
- Deep → Modularization

Don't put
everything in your
main function.



Modularization

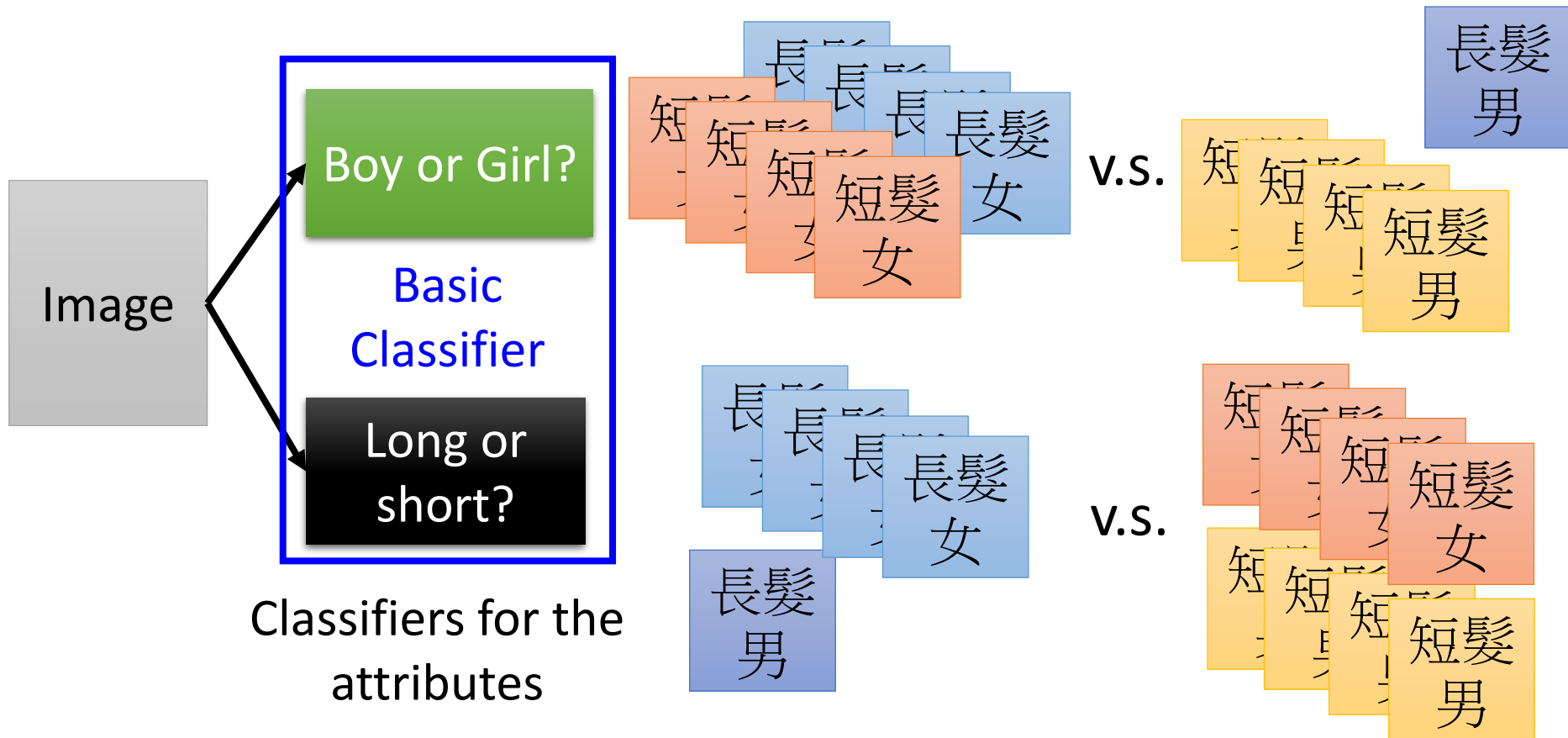
- Deep \rightarrow Modularization



Modularization

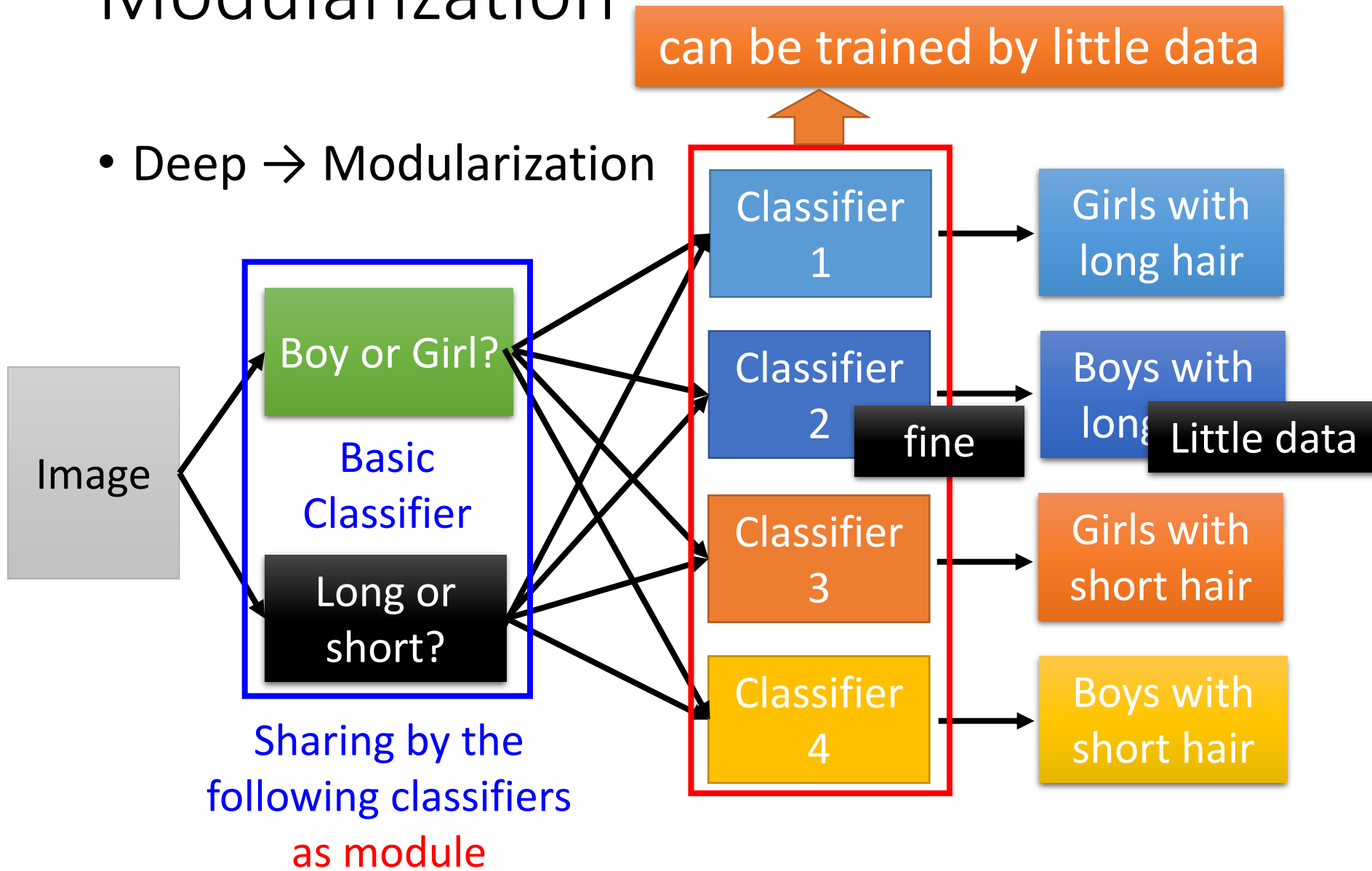
Each basic classifier can have sufficient training examples.

- Deep → Modularization



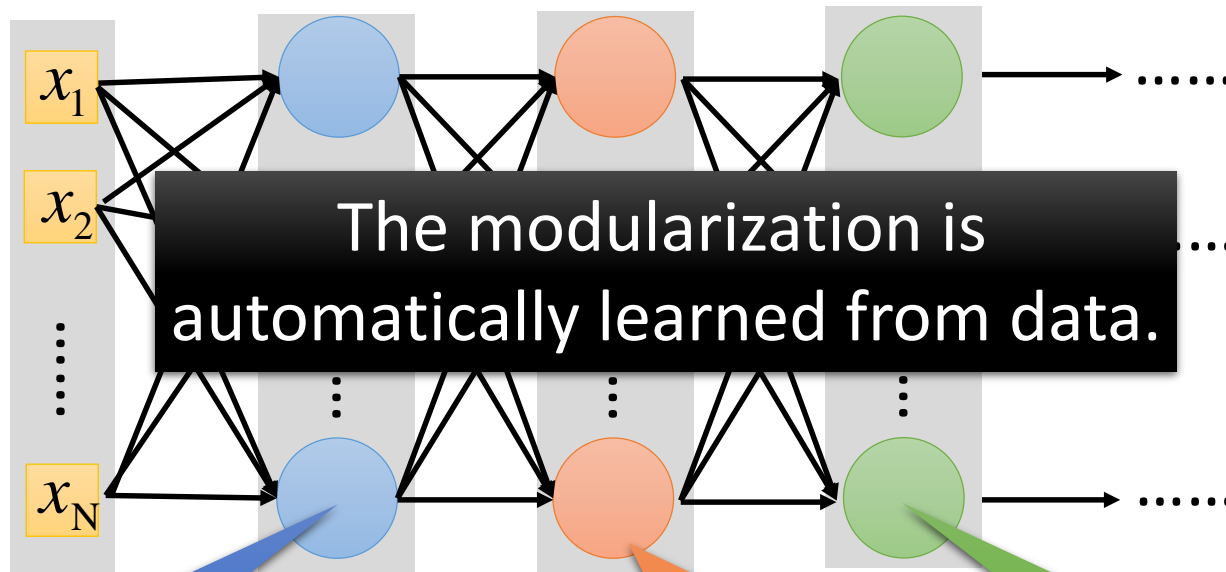
Modularization

- Deep → Modularization



Modularization

- Deep → Modularization → Less training data?



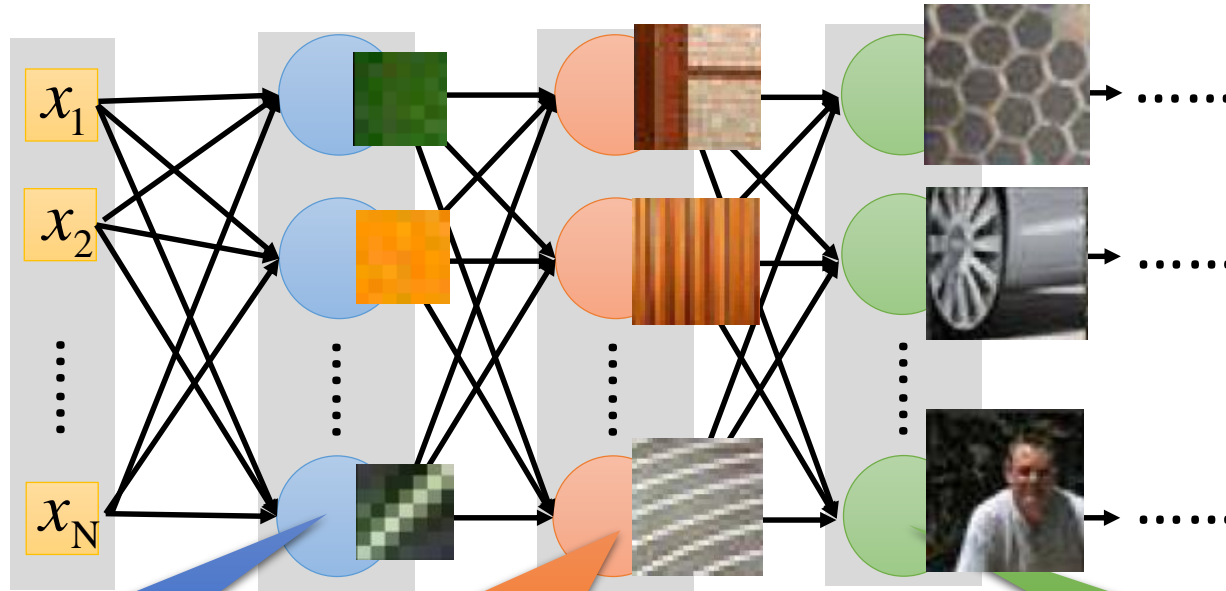
The most basic classifiers

Use 1st layer as module to build classifiers

Use 2nd layer as module

Modularization - Image

- Deep \rightarrow Modularization



The most basic
classifiers

Use 1st layer as module
to build classifiers

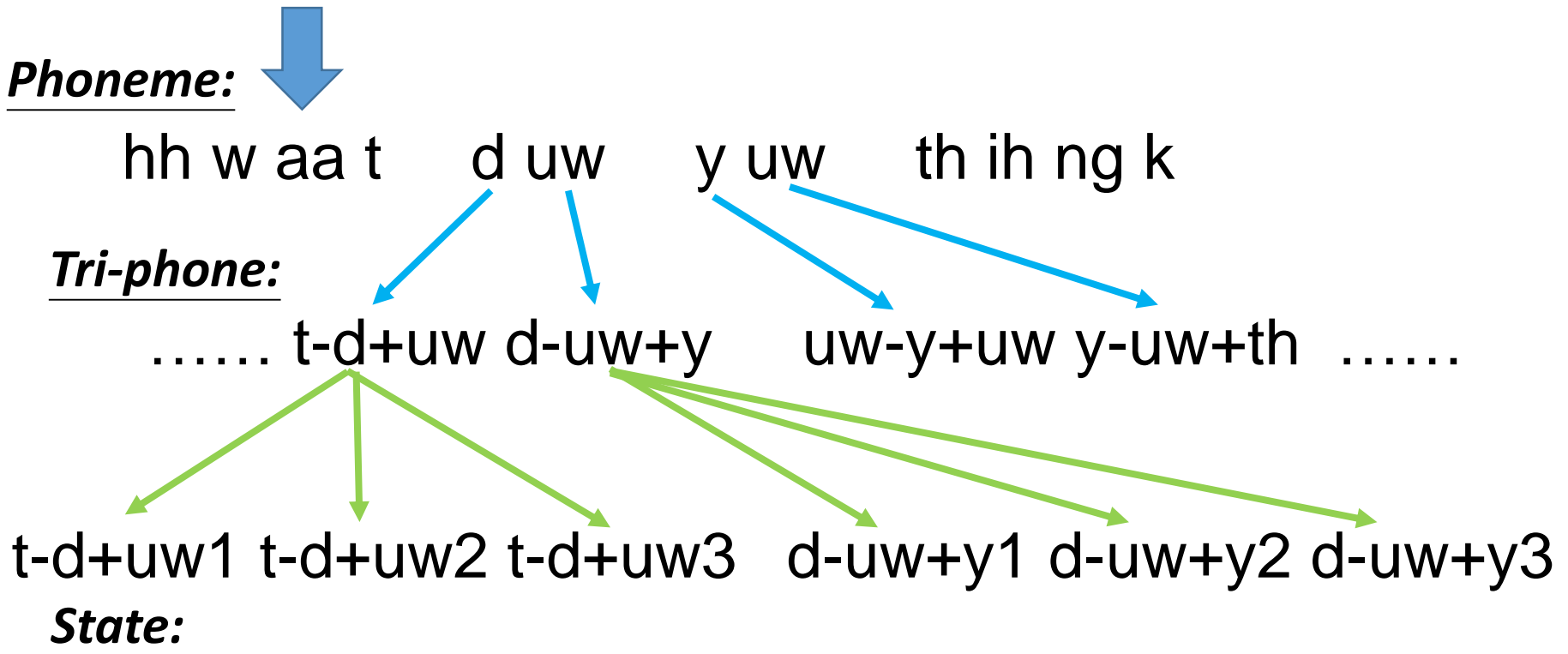
Use 2nd layer as
module

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

Modularization - Speech

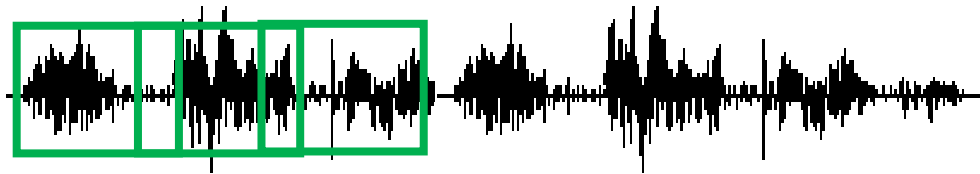
- The hierarchical structure of human languages

what do you think

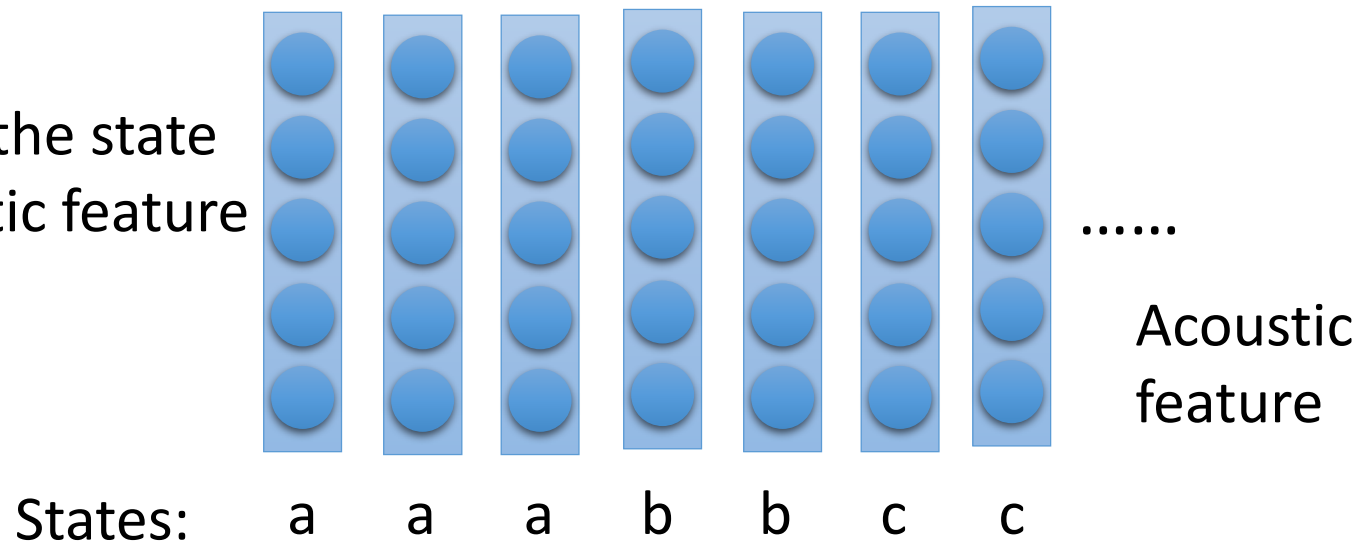


Modularization - Speech

- The first stage of speech recognition
 - Classification: input \rightarrow acoustic feature, output \rightarrow state



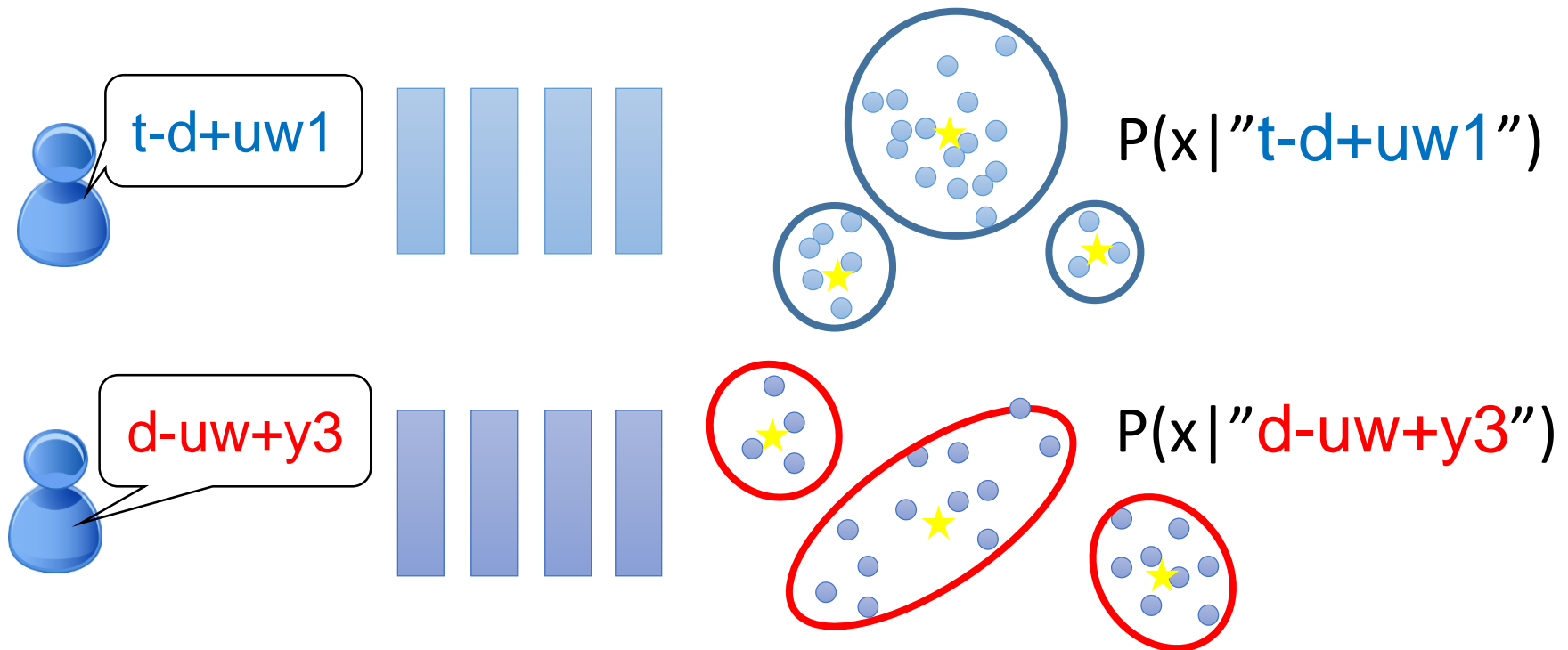
Determine the state
each acoustic feature
belongs to



Modularization - Speech

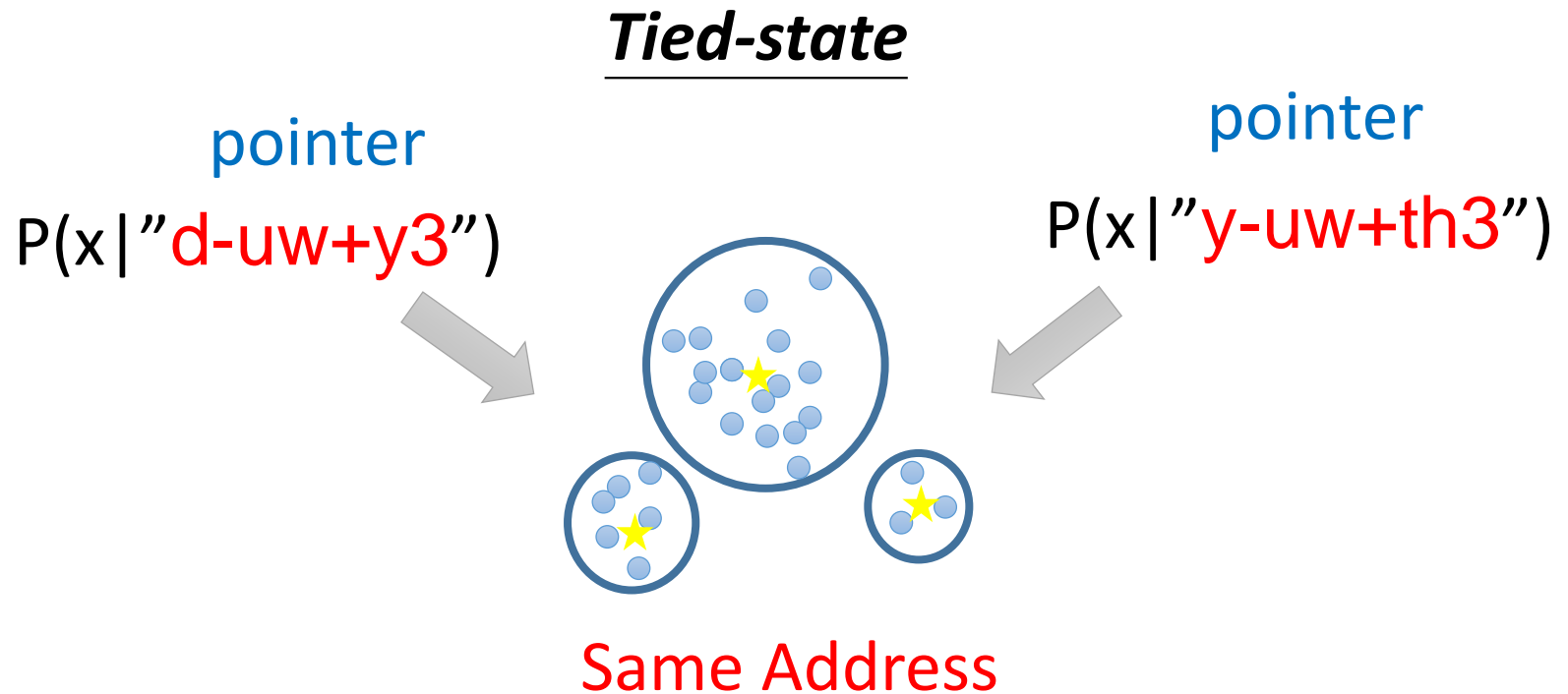
- Each state has a stationary distribution for acoustic features

Gaussian Mixture Model (GMM)



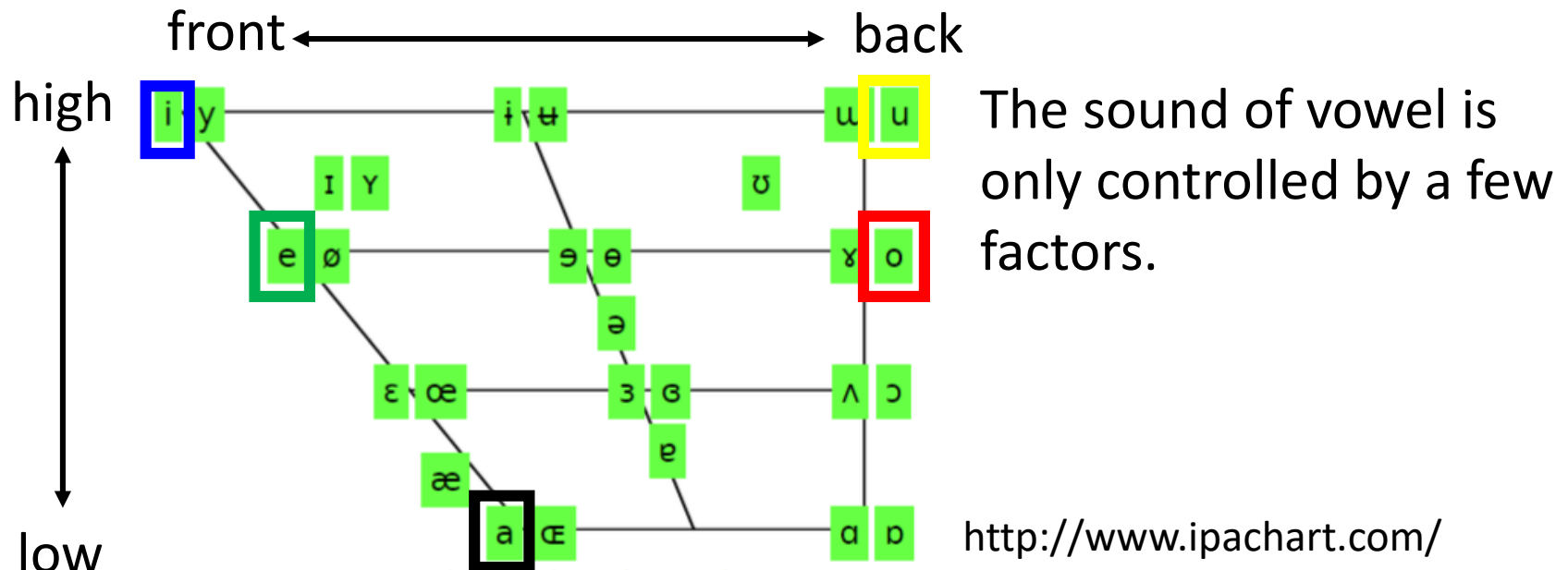
Modularization - Speech

- Each state has a stationary distribution for acoustic features



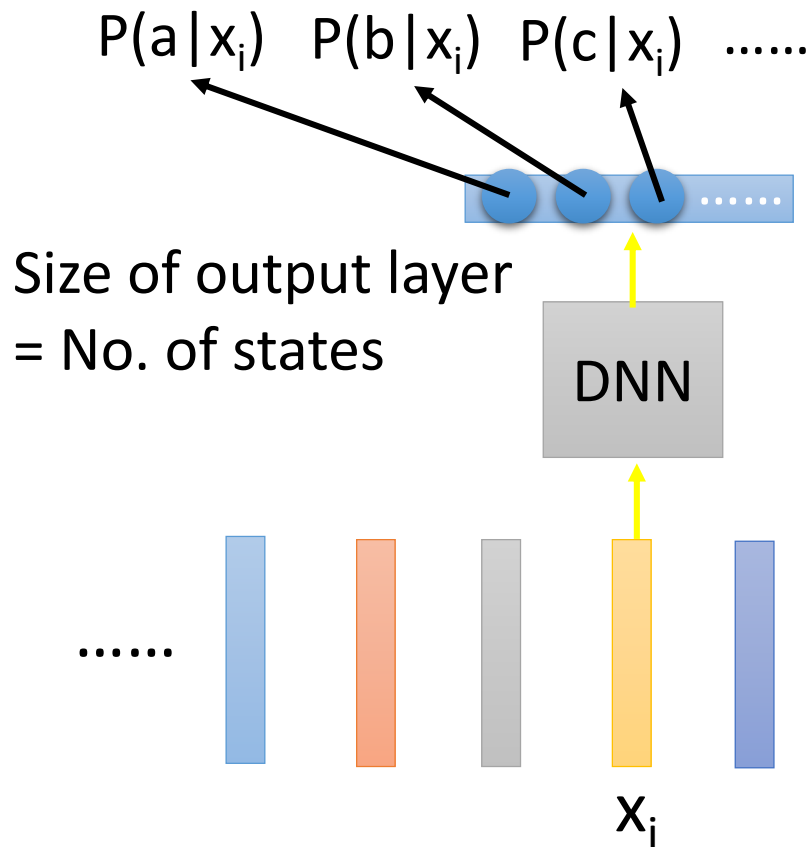
Modularization - Speech

- In HMM-GMM, all the phonemes are modeled independently
 - Not an effective way to model human voice



Modularization - Speech

- DNN input:
One acoustic feature
- DNN output:
Probability of each state

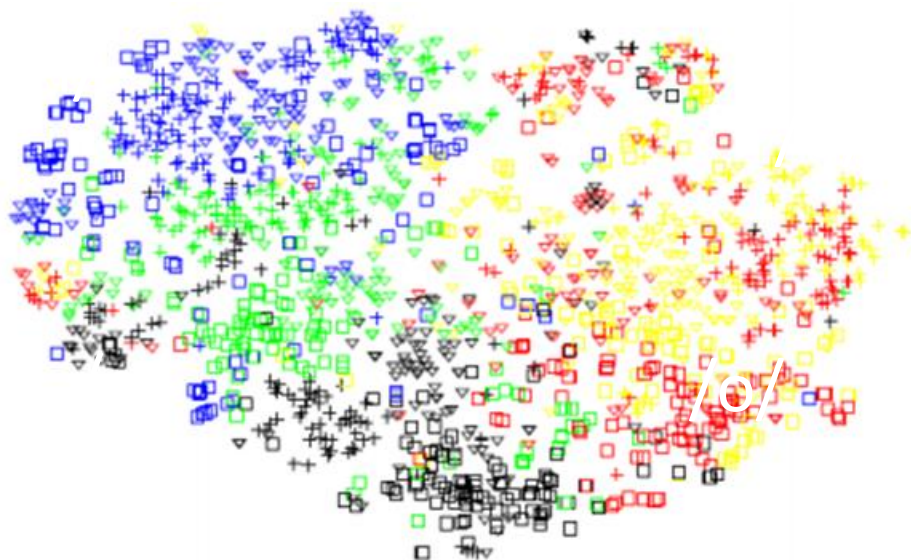
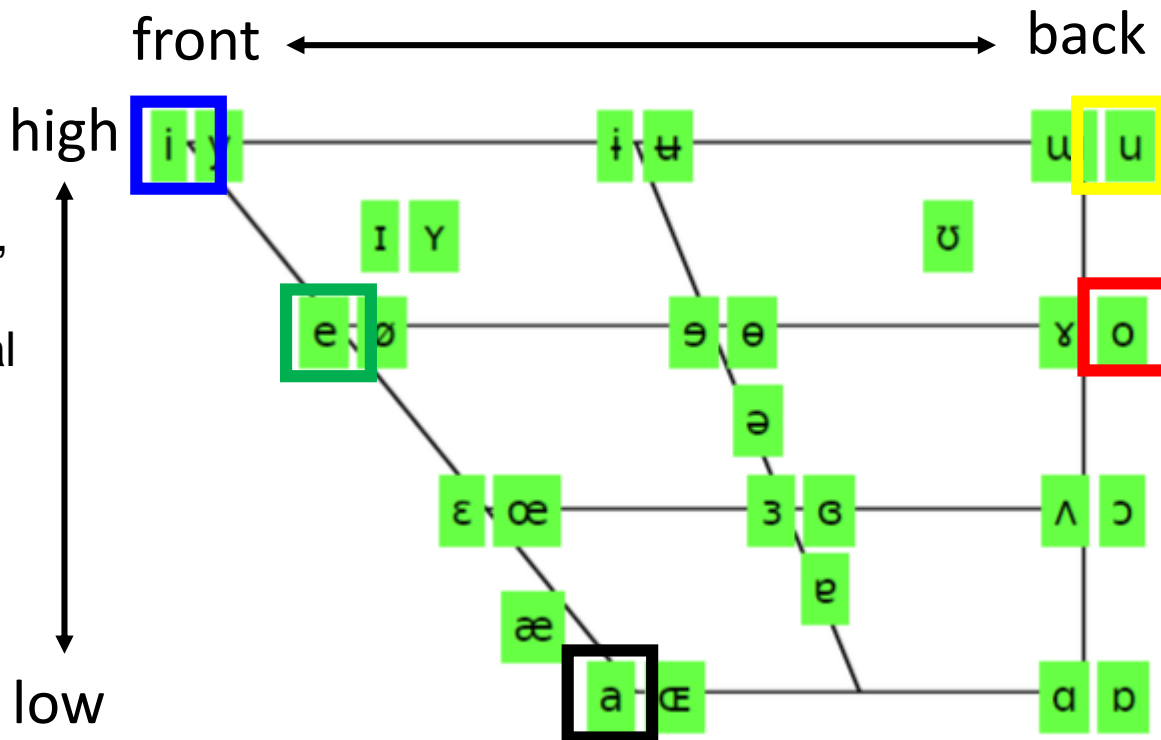


All the states use
the same DNN

Modularization

Vu, Ngoc Thang, Jochen Weiner, and Tanja Schultz. "Investigating the Learning Effect of Multilingual Bottle-Neck Features for ASR." *Interspeech*. 2014.

Output of hidden layer reduce to two dimensions



- The lower layers detect the manner of articulation
- All the phonemes share the results from the same set of detectors.
- Use parameters effectively

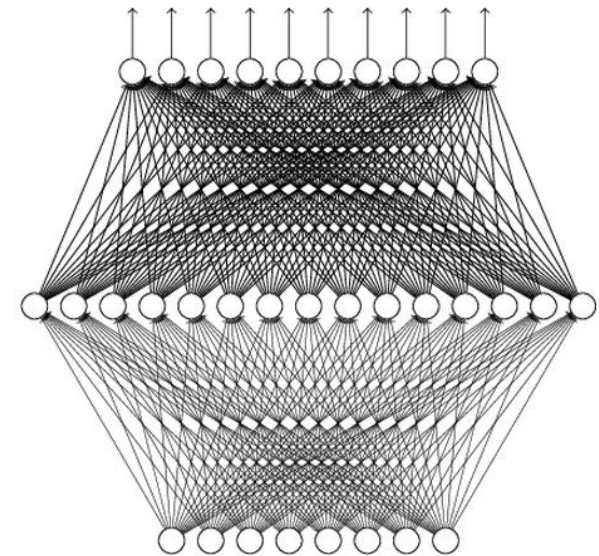
Universality Theorem

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.

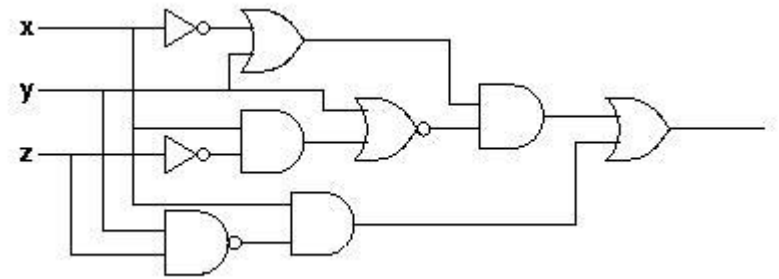
Analogy

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



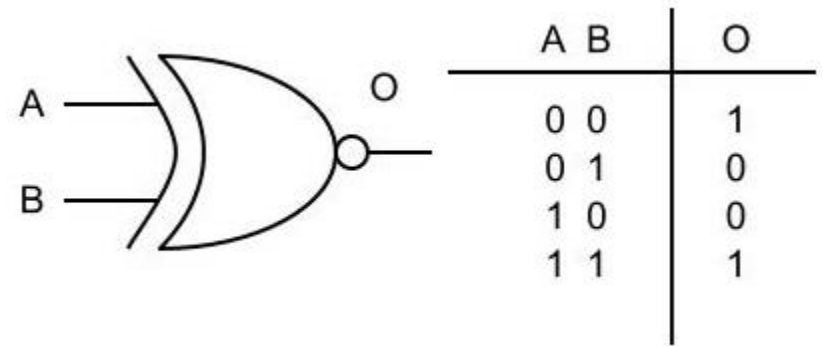
less parameters



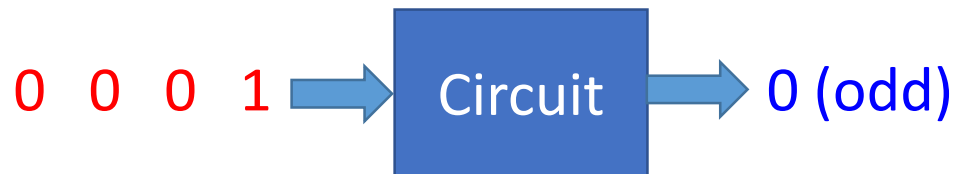
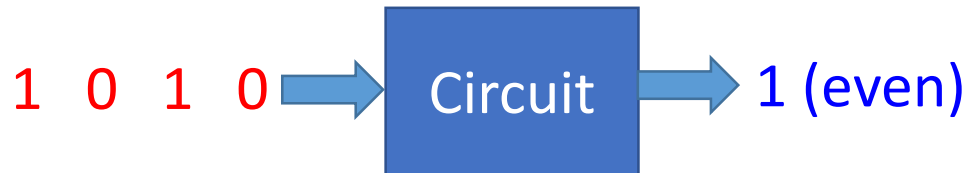
less data?

This page is for EE background.

Analogy

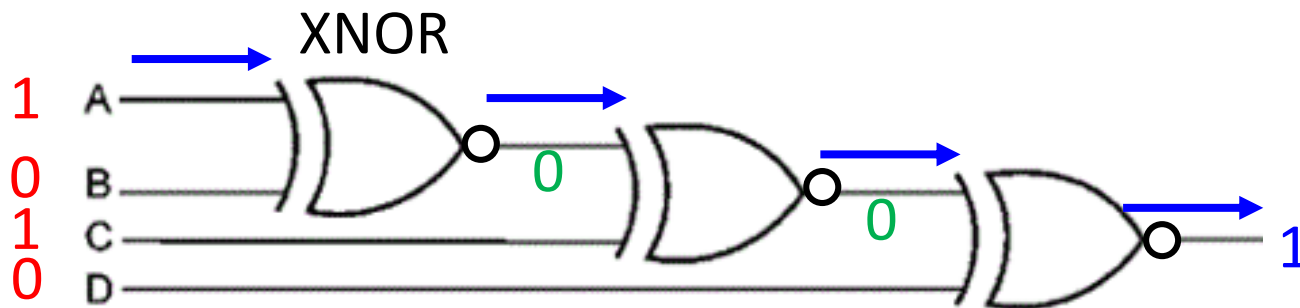


- E.g. parity check



For input sequence with d bits,

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

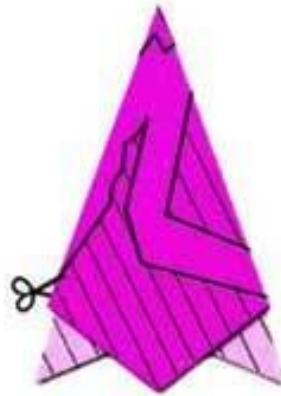


With multiple layers, we need only $O(d)$ gates.

More Analogy



① 画



② 剪



③ 展开, 完成



① 画



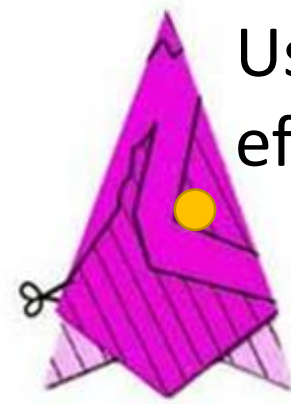
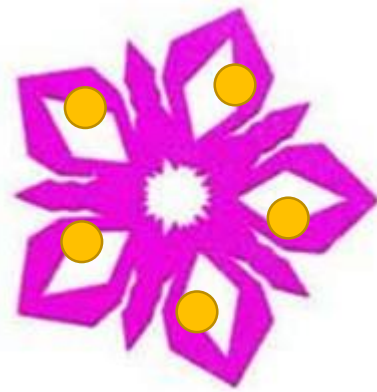
② 剪



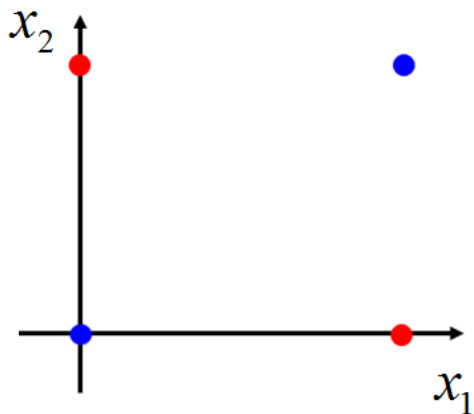
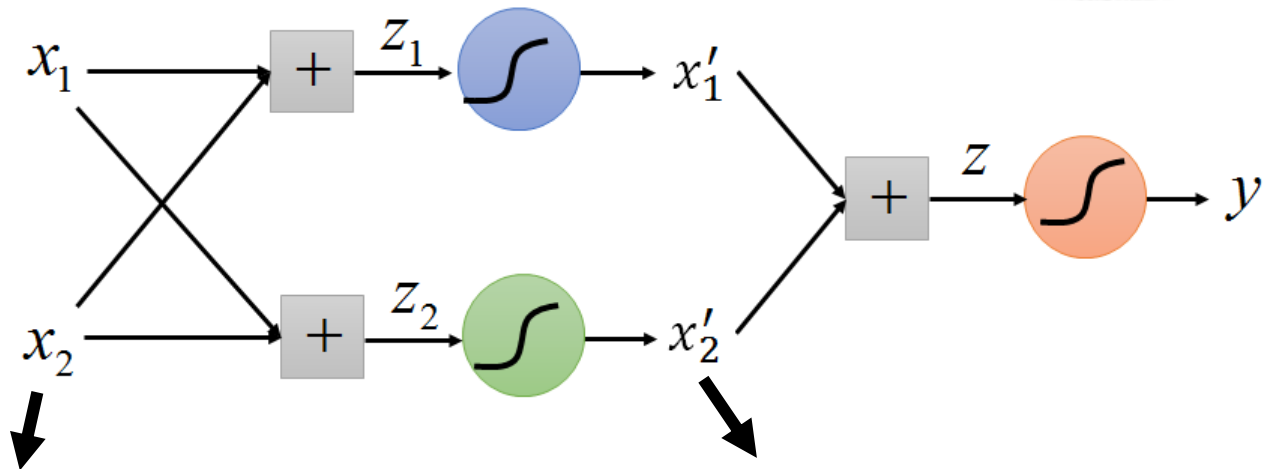
③ 展开, 完成



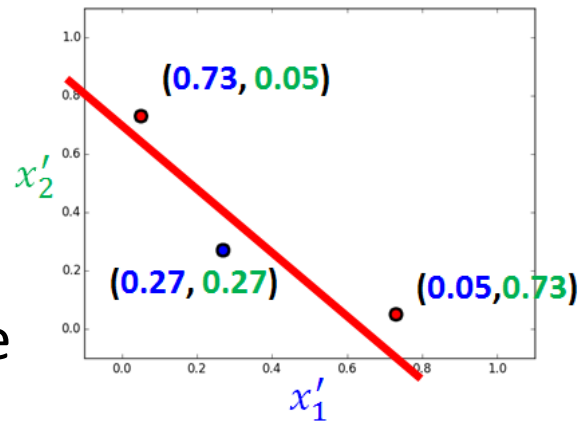
More Analogy



Use data
effectively



Folding
the space



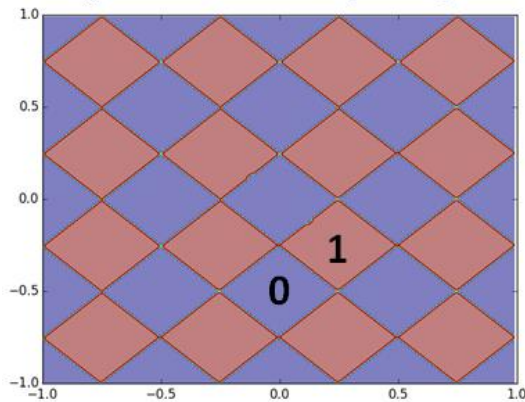
More Analogy - Experiment

Different numbers of training examples

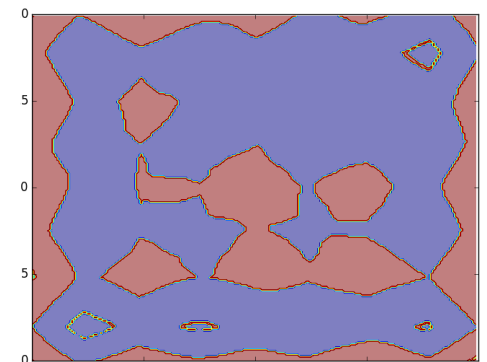
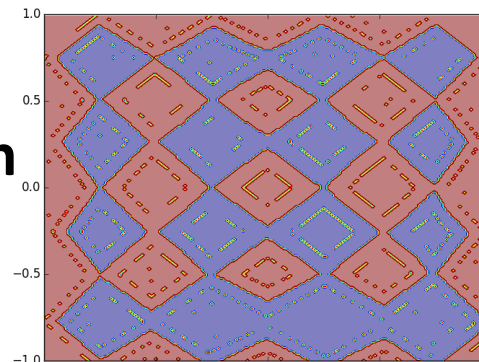
10,000

2,000

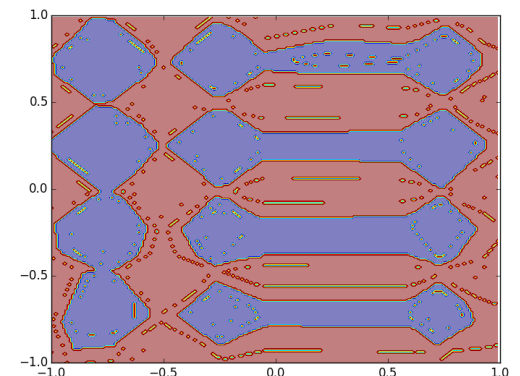
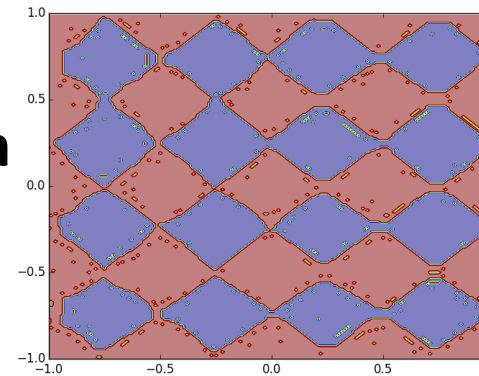
$$f: \mathbb{R}^2 \rightarrow \{0,1\}$$



**1 hidden
layer**

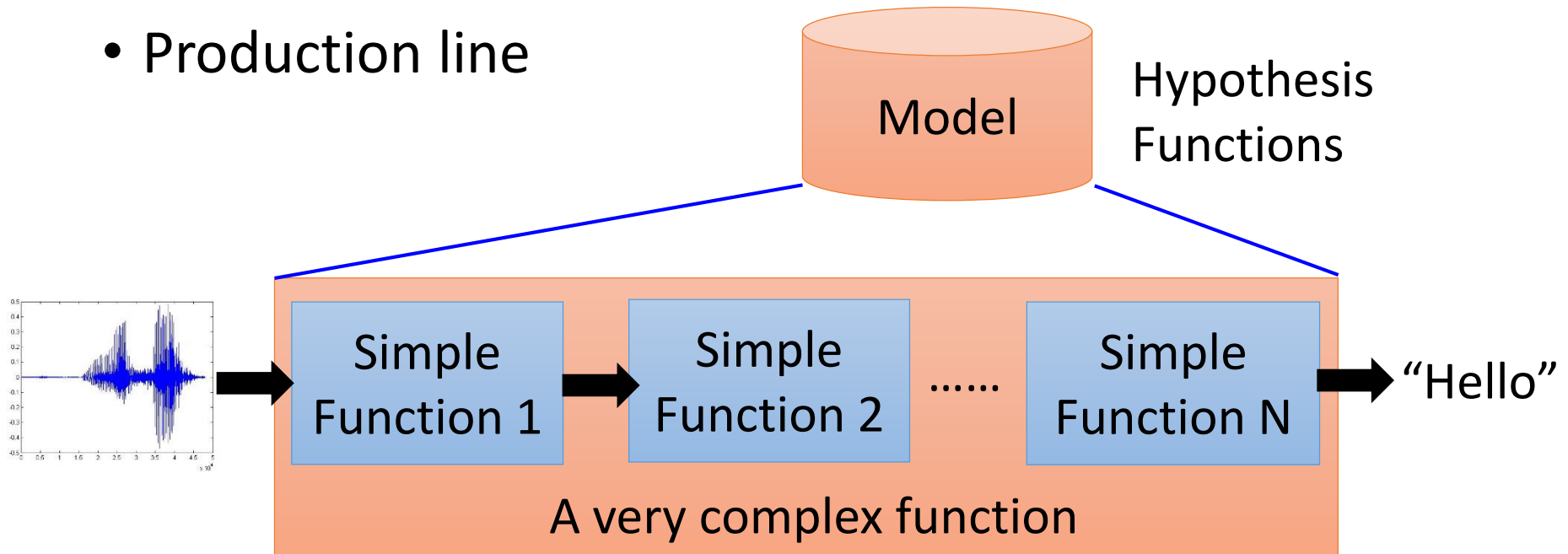


**3 hidden
layers**



End-to-end Learning

- Production line



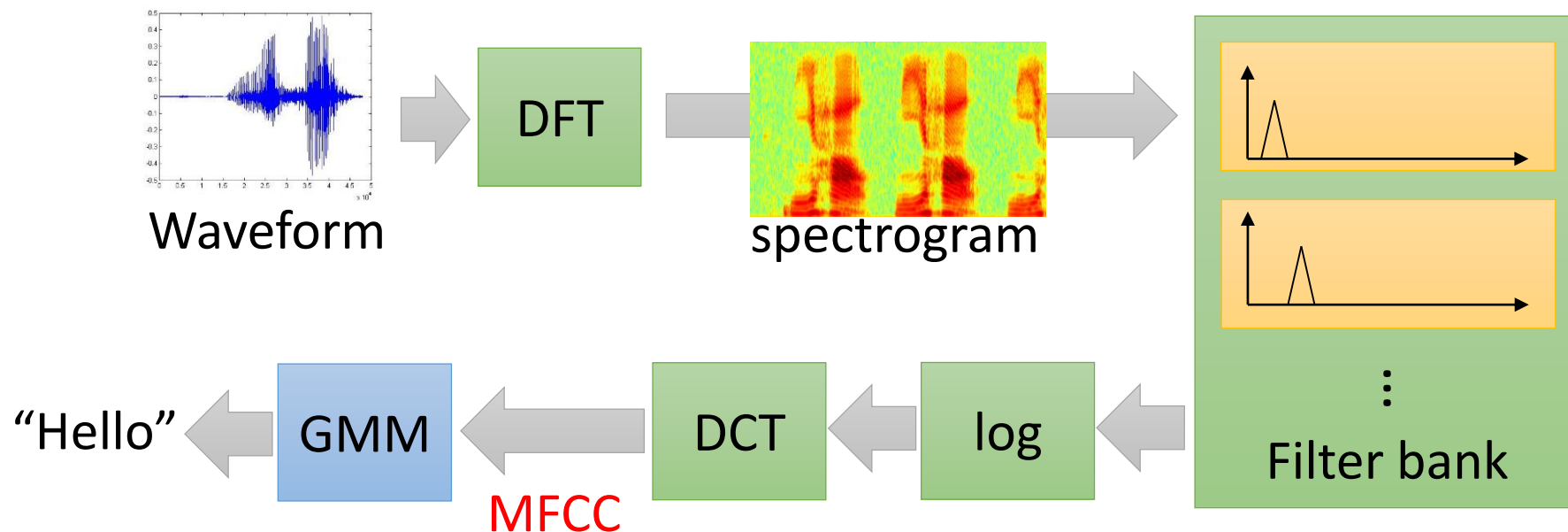
End-to-end training:

What each function should do is learned automatically

End-to-end Learning

- Speech Recognition

- Shallow Approach



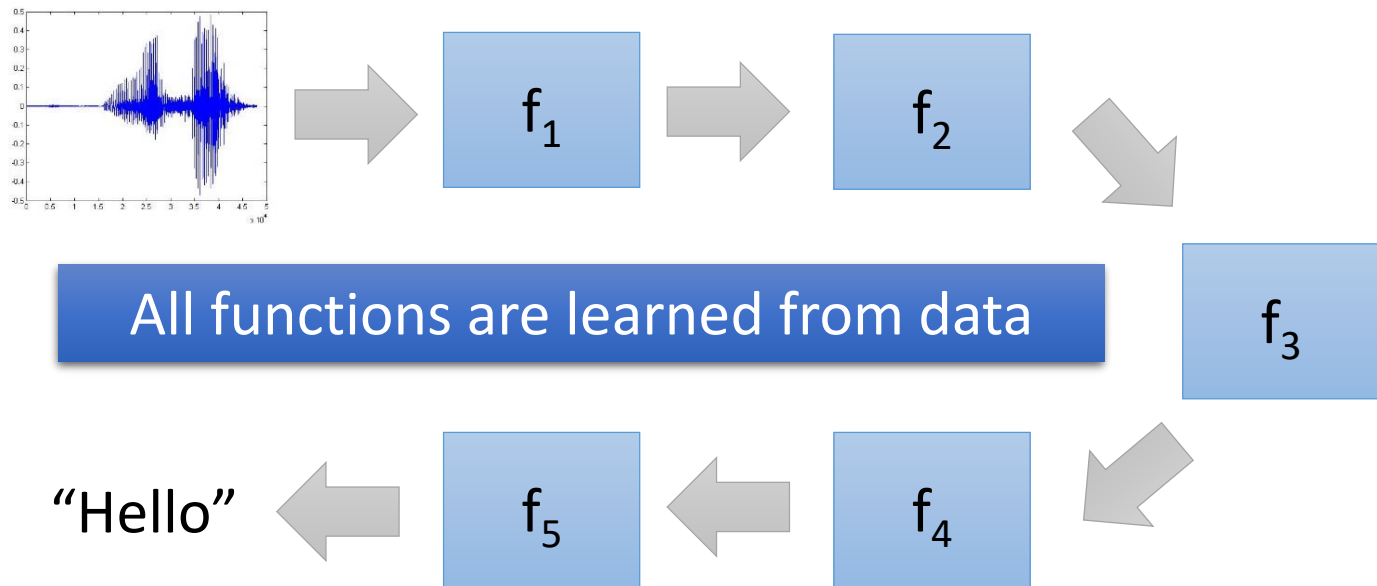
Each box is a simple function in the production line:

 :hand-crafted  :learned from data

End-to-end Learning

- Speech Recognition

- Deep Learning



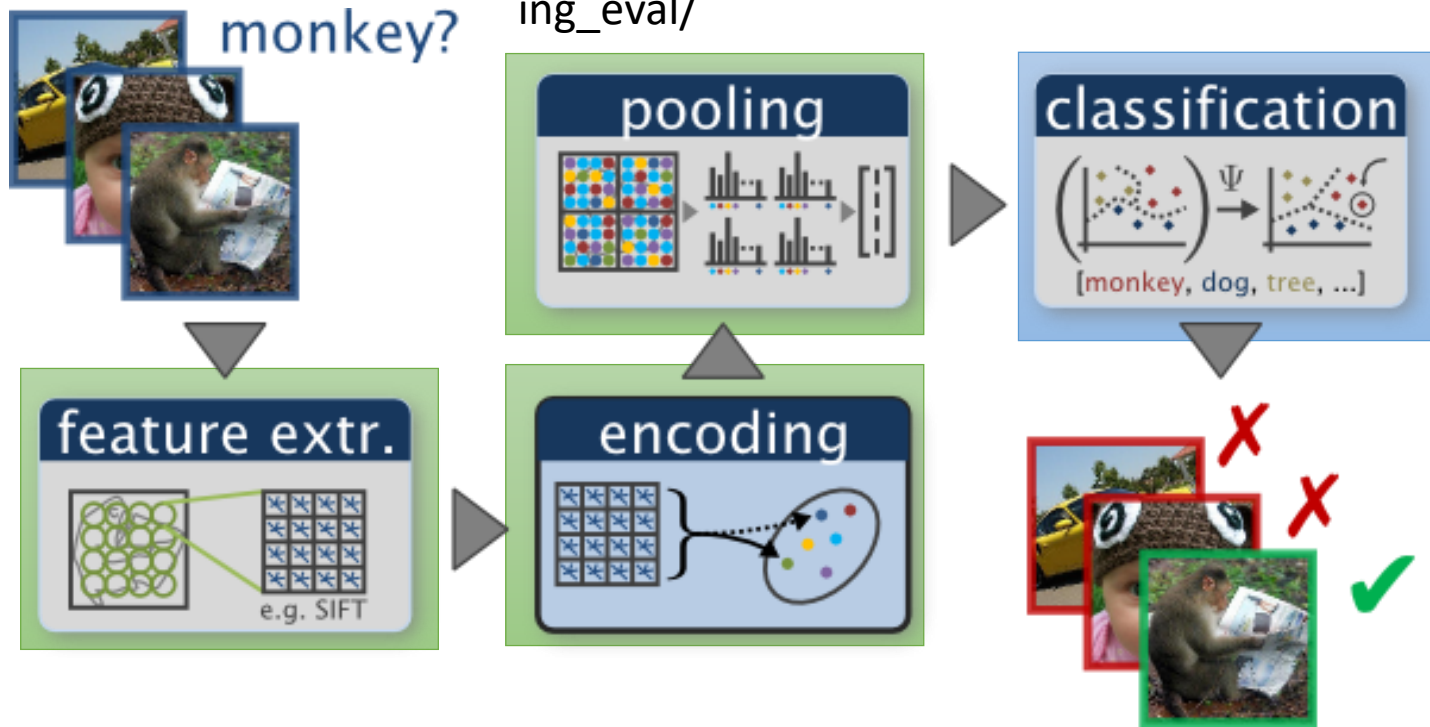
Less engineering labor, but machine learns more

End-to-end Learning

- Image Recognition

- Shallow Approach

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



:hand-crafted

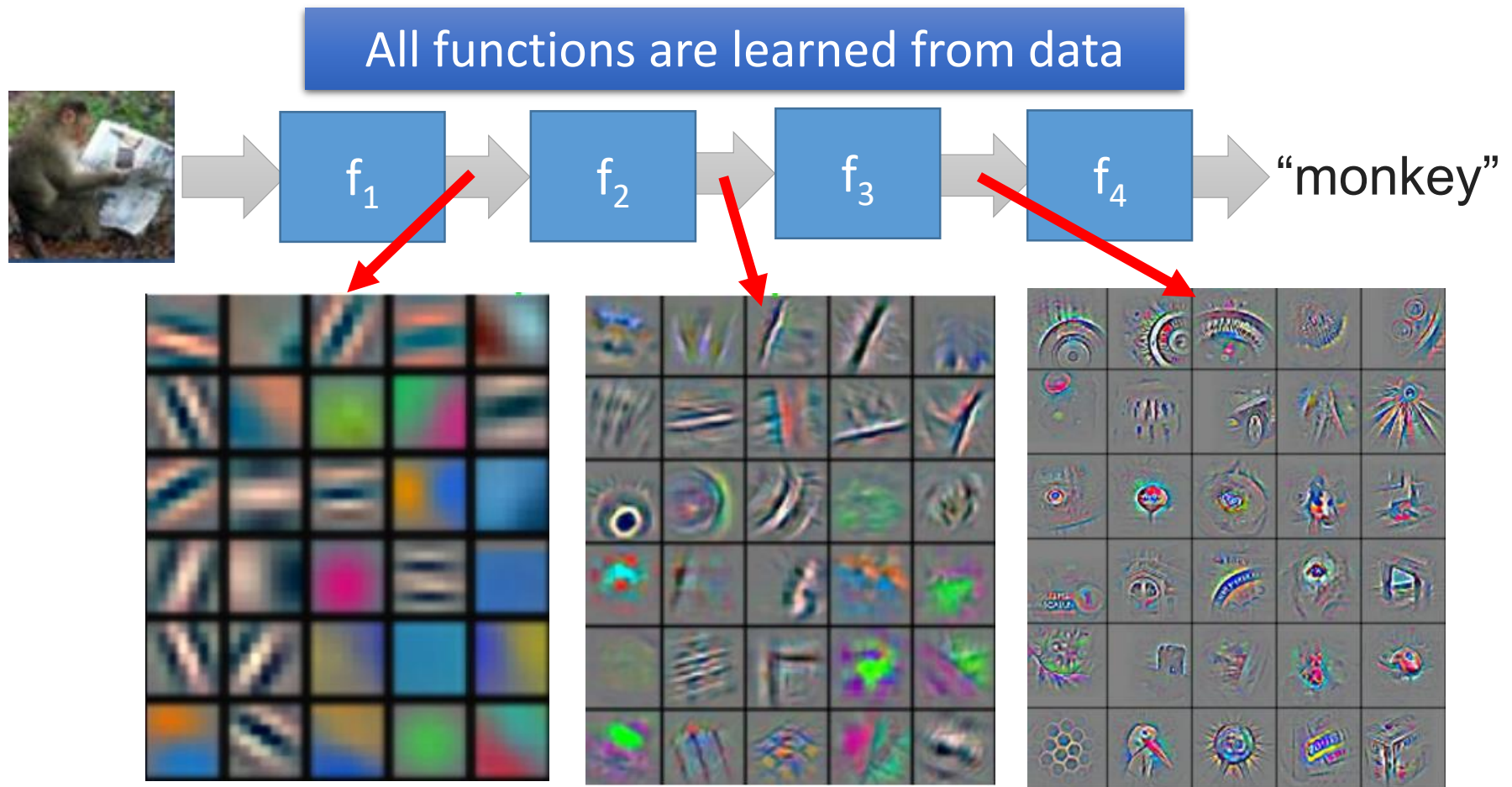


:learned from data

End-to-end Learning - Image Recognition

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

- Deep Learning



Complex Task ...

- Very similar input, different output



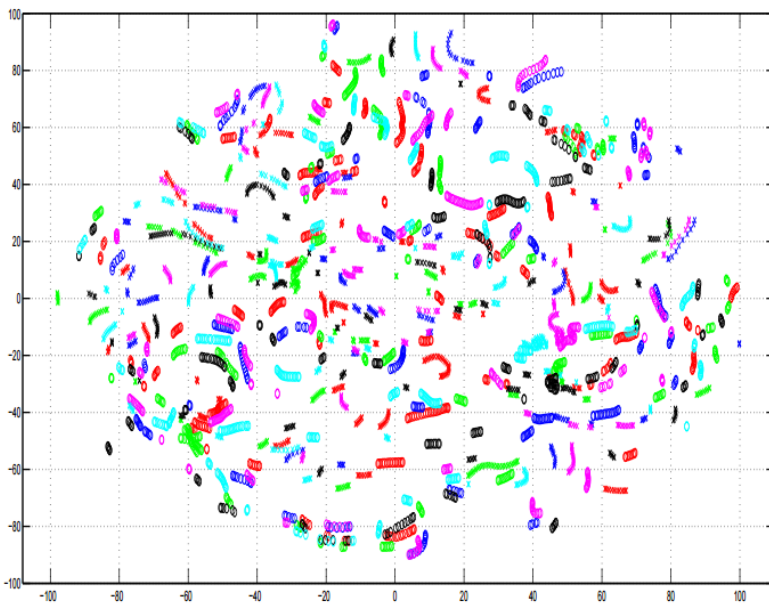
- Very different input, similar output



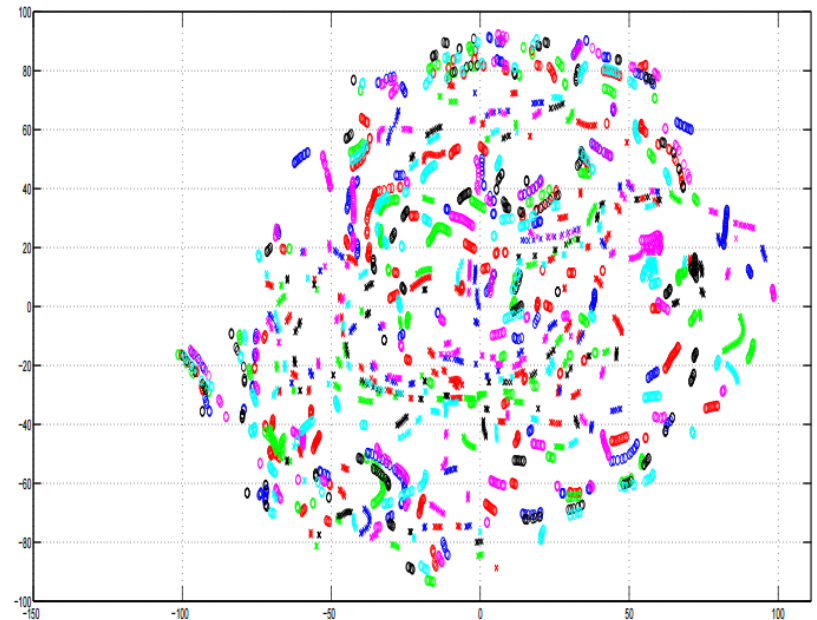
Complex Task ...

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)

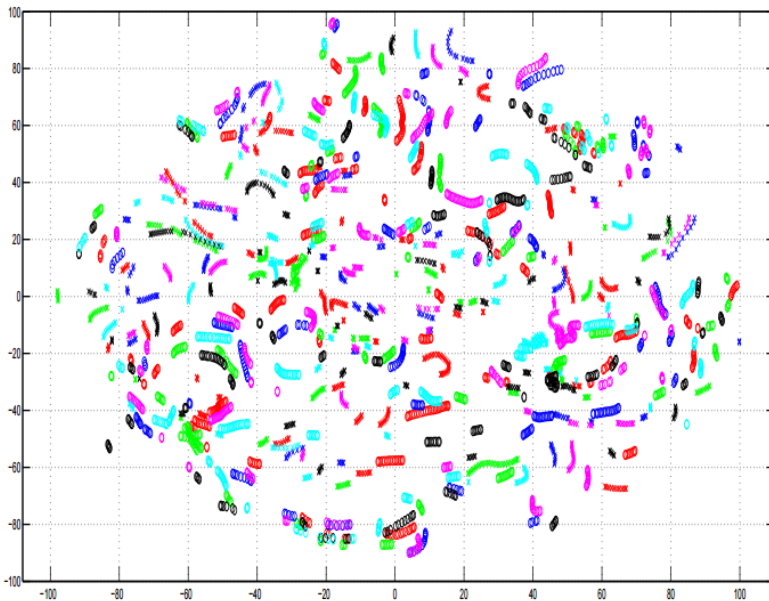


1-st Hidden Layer

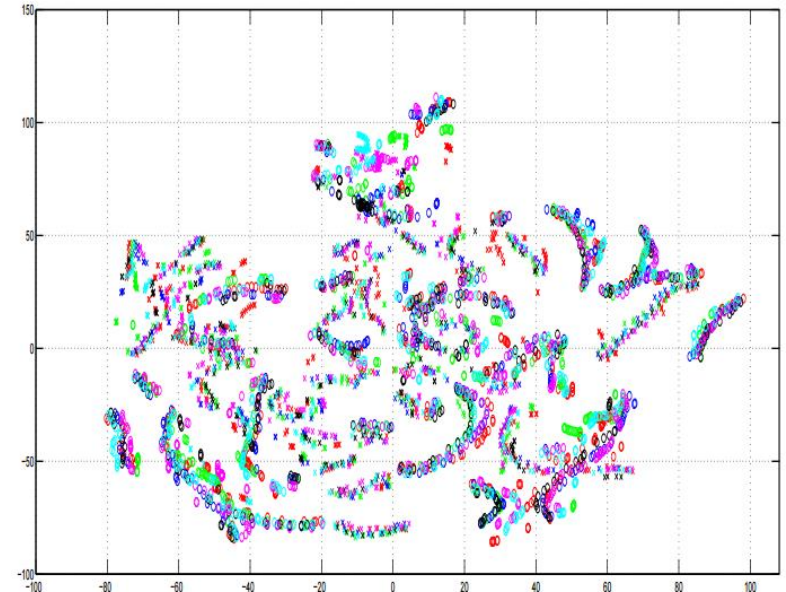
Complex Task ...

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

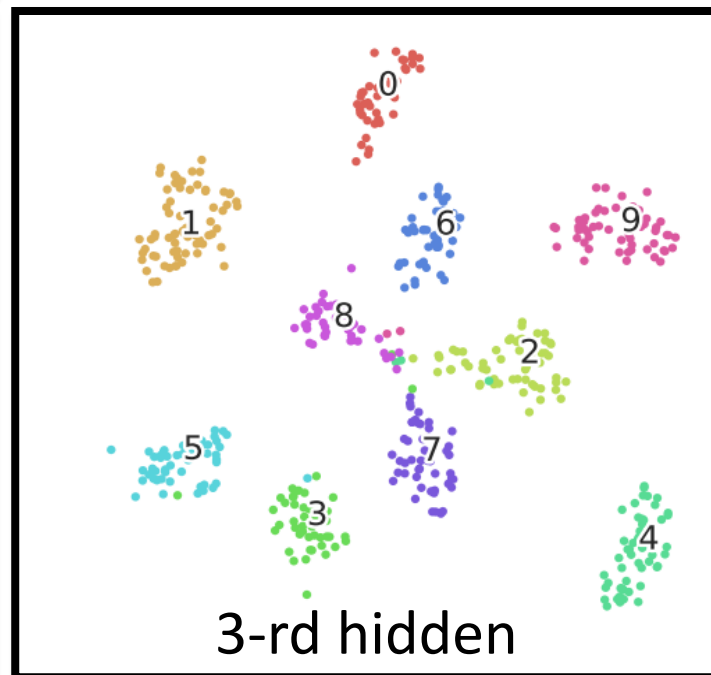
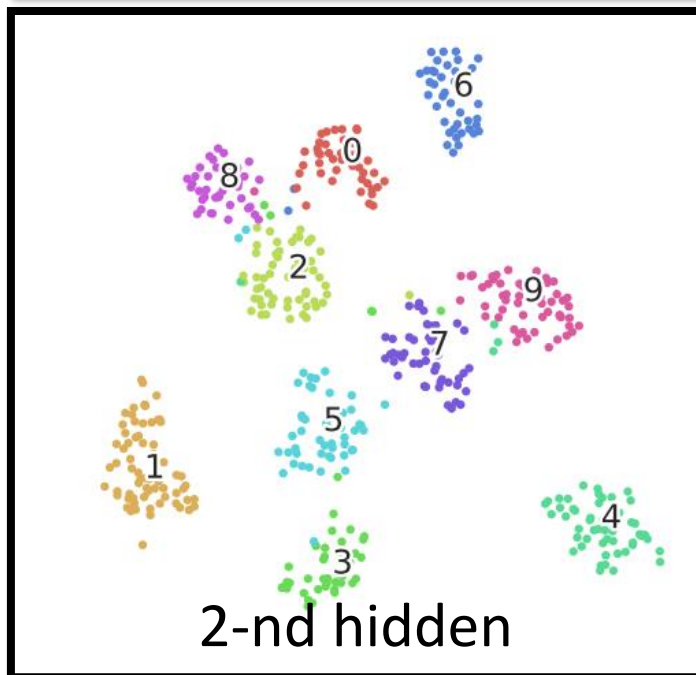
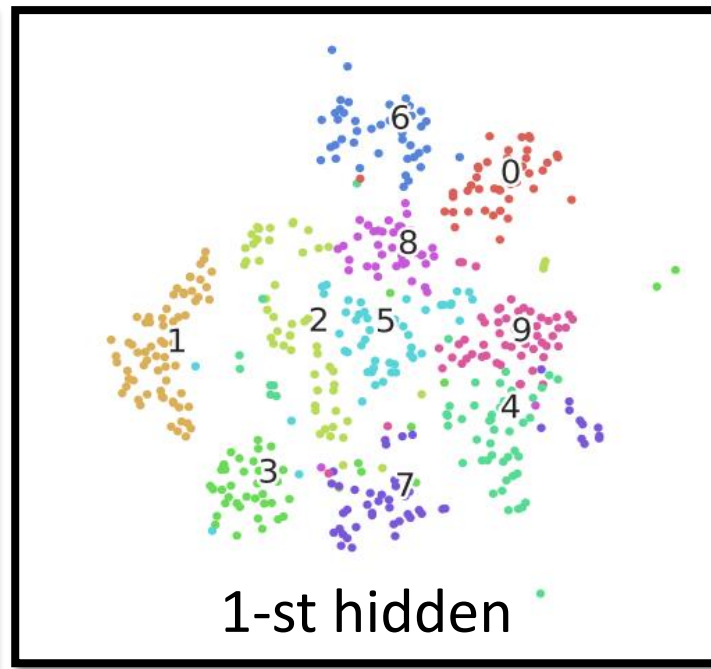
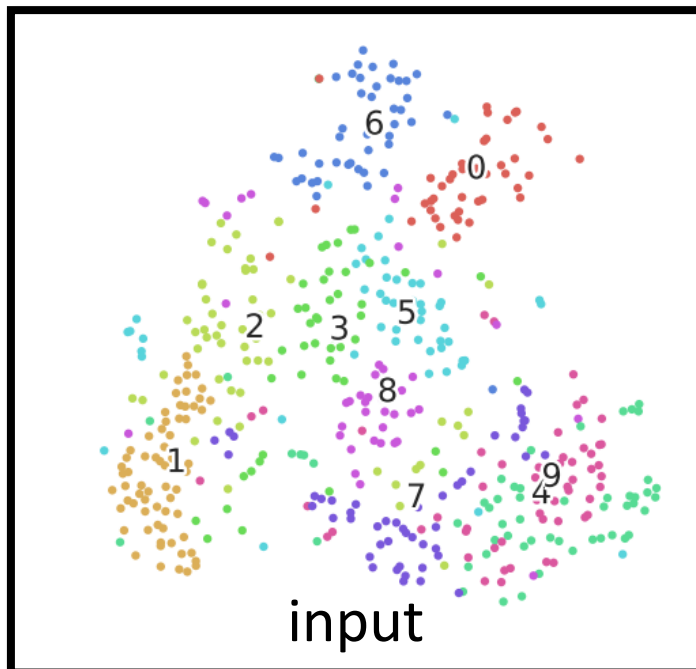


Input Acoustic Feature (MFCC)



8-th Hidden Layer

MNIST



To learn more ...

- Do Deep Nets Really Need To Be Deep? (by Rich Caruana)
- <http://research.microsoft.com/apps/video/default.aspx?id=232373&r=1>

Do deep nets really
need to be deep?

Rich Caruana
Microsoft Research

Lei Jimmy Ba
MSR Intern, University of Toronto

*Thanks also to: Gregor Urban, Krzysztof Geras, Samira Kahou, Abdelrahman Mohamed,
Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong*

Yes!

Thank You

Any Questions?

To learn more ...

- Deep Learning: Theoretical Motivations (*Yoshua Bengio*)
 - http://videolectures.net/deeplearning2015_bengio_theoretical_motivations/
- Connections between physics and deep learning
 - <https://www.youtube.com/watch?v=5MdSE-N0bxs>
- Why Deep Learning Works: Perspectives from Theoretical Chemistry
 - <https://www.youtube.com/watch?v=kIbKHIPbxiU>