Introduction to Deep Learning

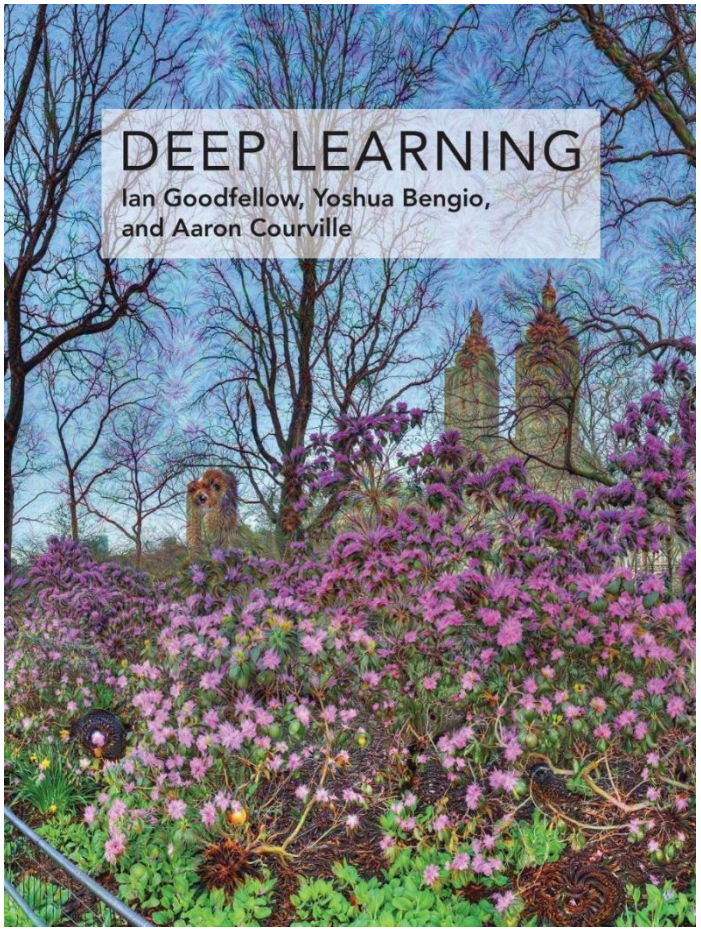
中国科学院数学与系统科学研究院

张世华

zsh@amss.ac.cn

Basic References

http://www.deeplearningbook.org/



Book

Tutorial

Deep Learning Tutorial
Hung-yi Lee (李宏毅)
http://speech.ee.ntu.edu.tw/~tlkagk/rtalk.html

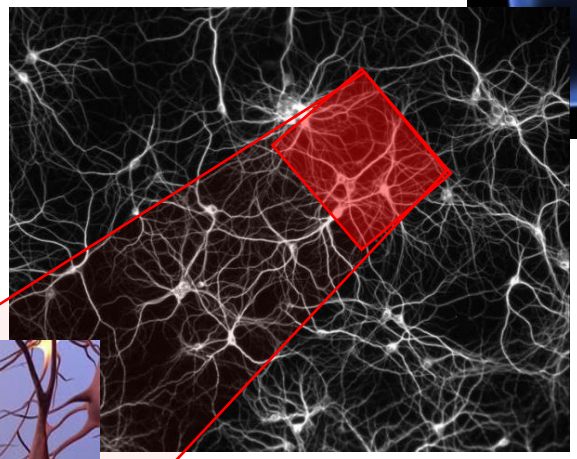
Overview of deep learning

What's Deep Learning?

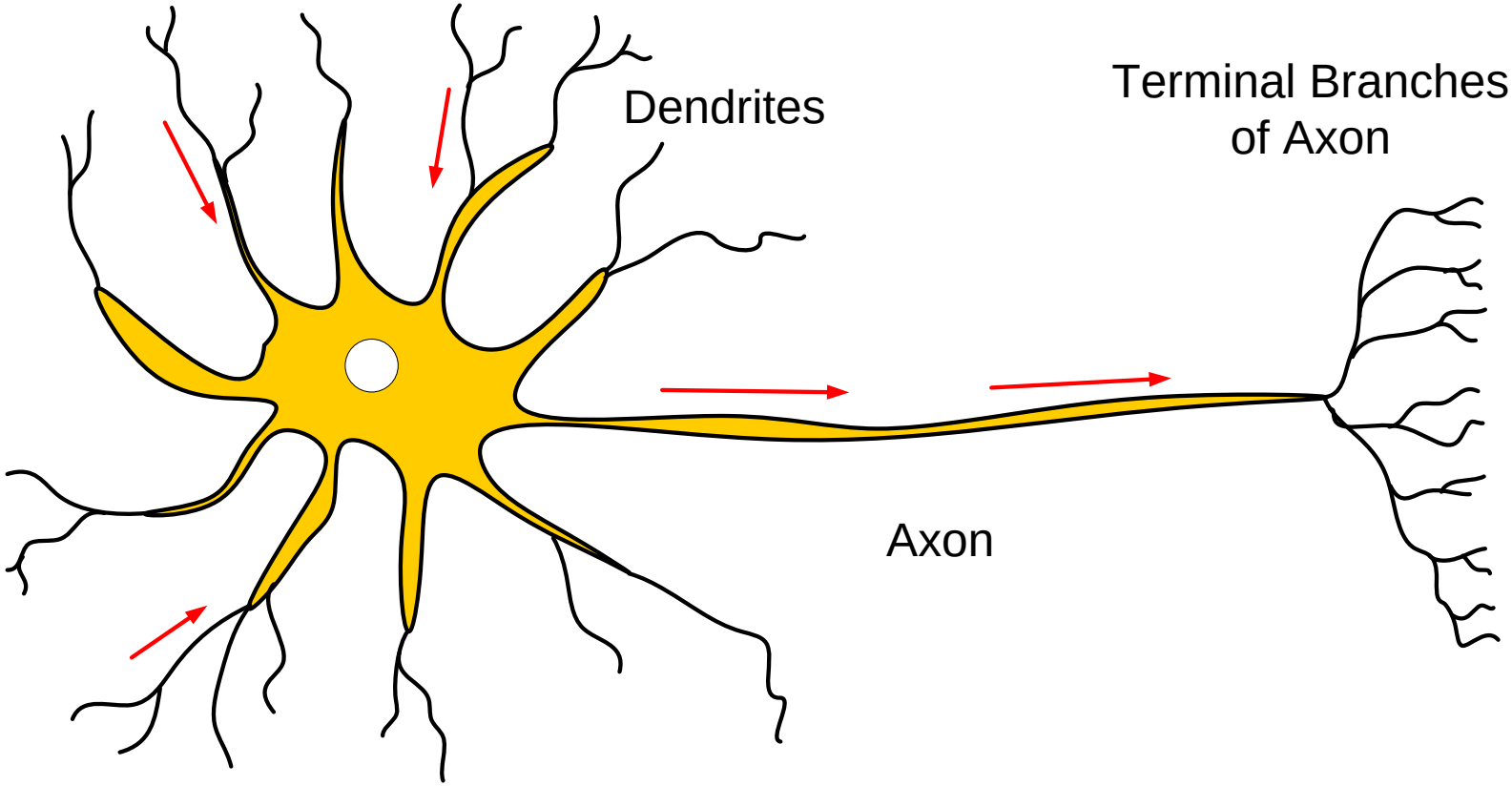
Deep learning is a subfield of machine learning, and neural networks make up the backbone of deep learning algorithms.

What's Neural Networks?

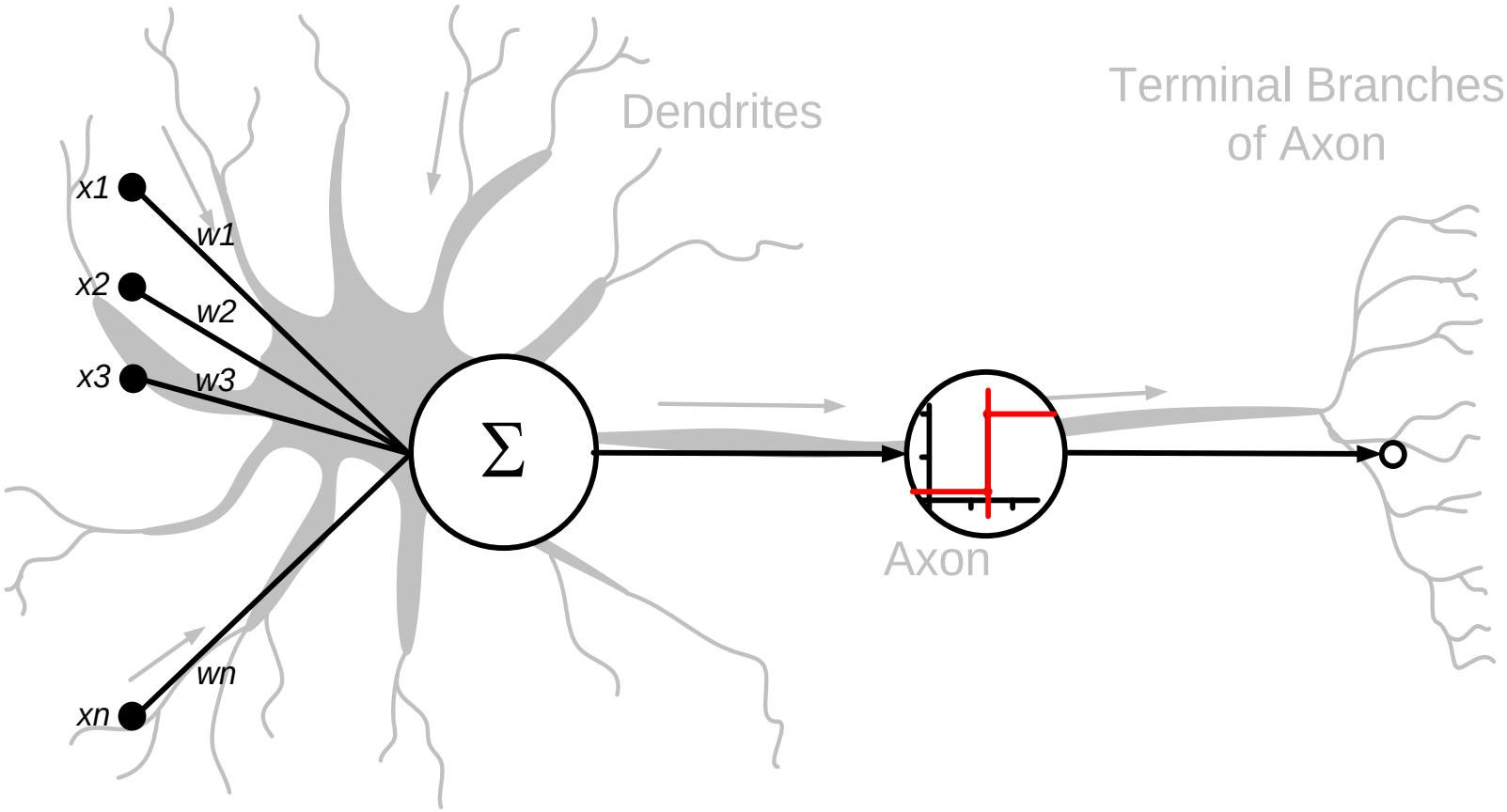
Biological
Neurons



What's Neural Networks? Biological Neurons

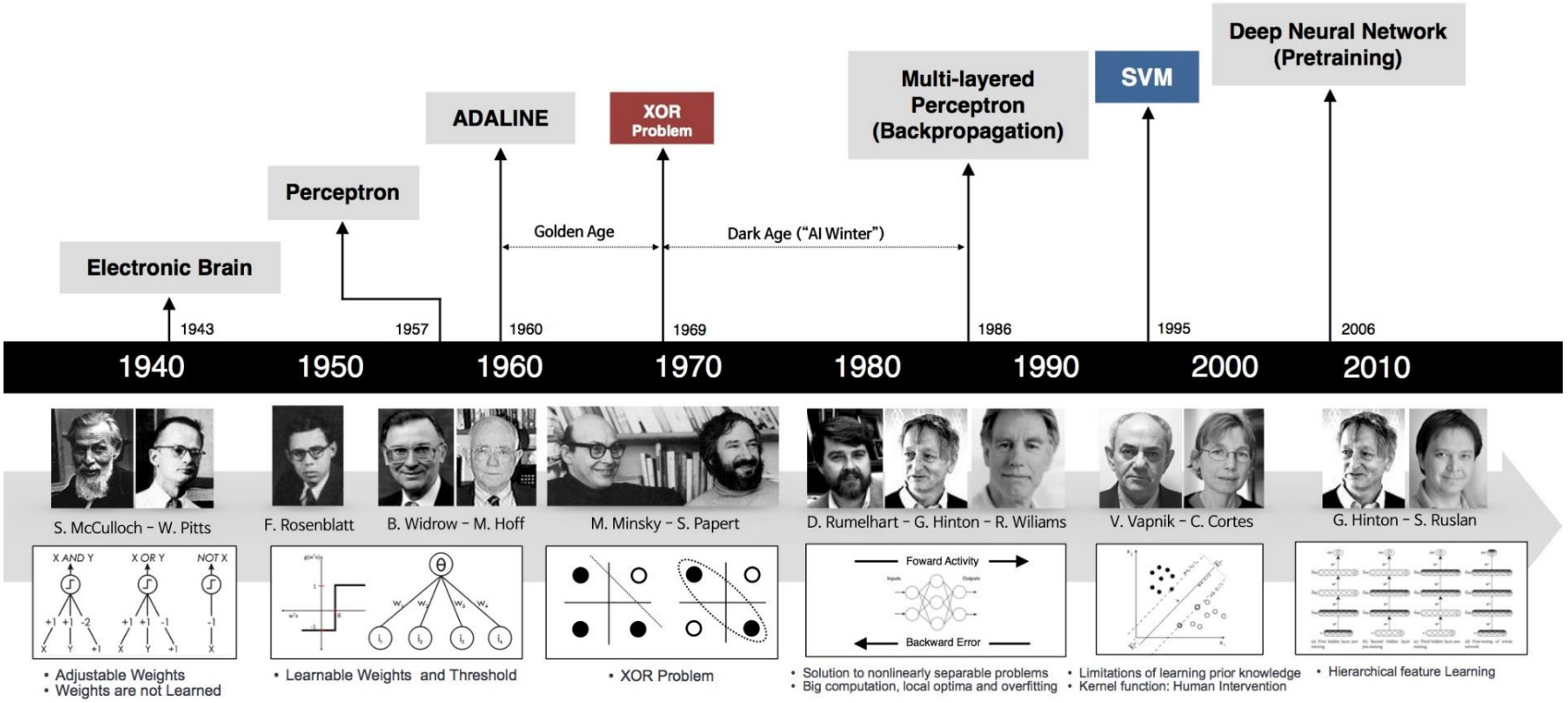


Artificial Neural Networks (ANN)



Slide credit : Andrew L. Nelson

History of Neural Networks



By Wei Di , Jianing Wei , Anurag Bhardwaj

What's Deep Learning?

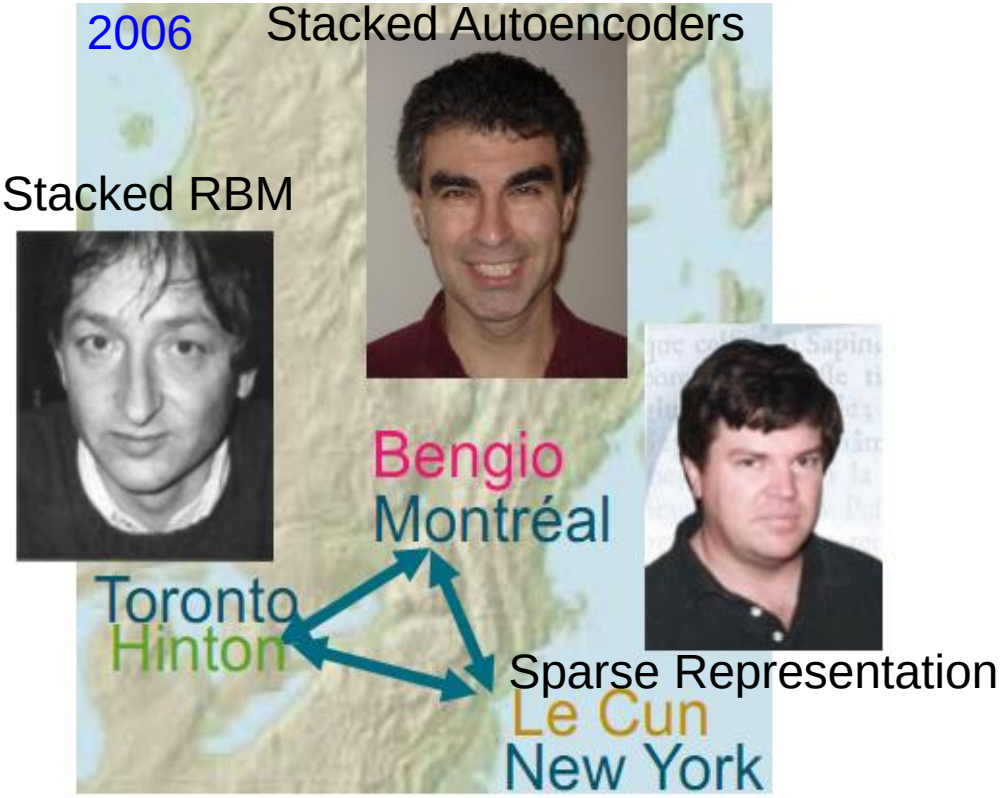
- A particular class of learning algorithms.
- Rebranded neural networks: with multiple layers.
- Inspired by the neuronal architecture of the Brain.
- Renewed interest in the area due to a few recent breakthroughs.
- Learn parameters from data.
- Nonlinear classification or clustering.

Deep architecture is an ensemble of shallow networks

Deep Learning Revolution

Before 2006

Fail to train deep architectures



Fathers of the **Deep Learning Revolution** Receive ACM A.M. Turing Award (2018) https://awards.acm.org/about/2018-turing

Old Wine in a New Bottle!

Big Data
(Digitalization)

Computation
(Moore's Law, GPUs)

Algorithmic
Progress

"2006"

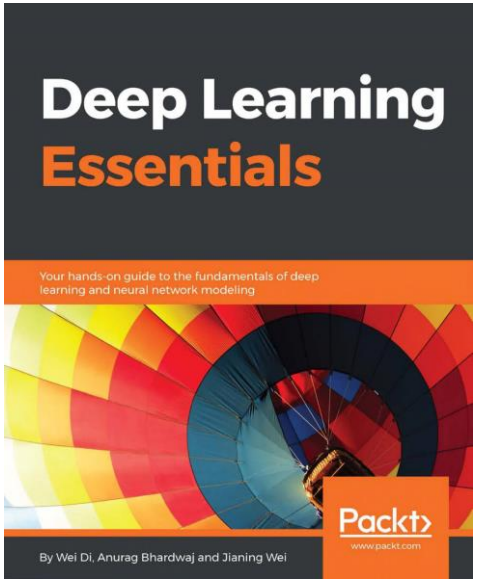
- Faster machines (GPU's!)
- More data
- New methods for unsupervised pre-training

Deep Learning Revolution

- It seems that most theoretical breakthroughs had already been made by the 1980s-1990s, so what else has changed in the past decade?
- A not-too-controversial theory is that the success of deep learning is largely a success of engineering.

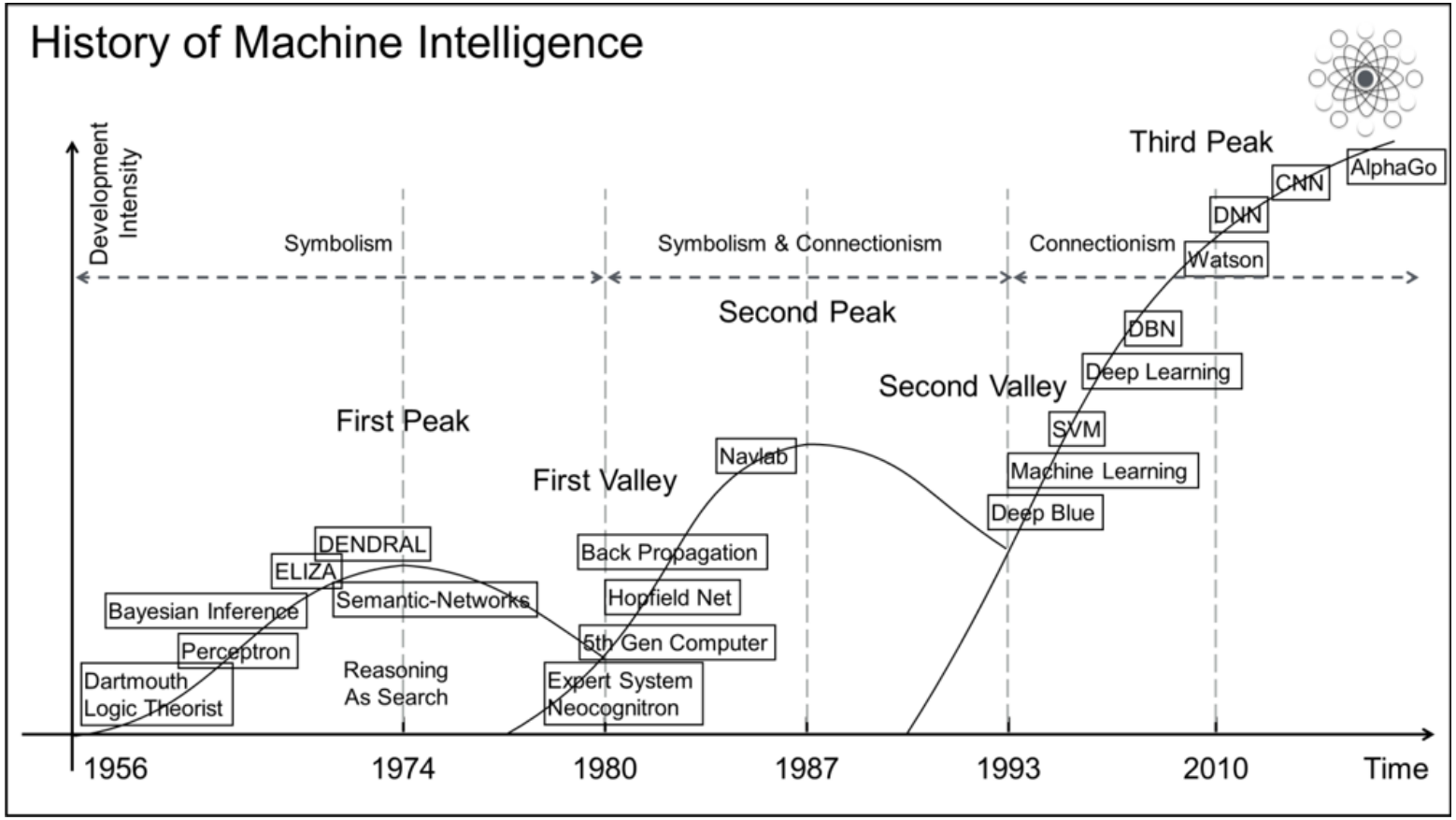
If you treat the theoretical development of deep learning as the engine, fast computer, the development of graphics processing units (GPU) and the occurrence of massive labeled datasets are the fuels.

Andrew Ng



History of Machine Intelligence

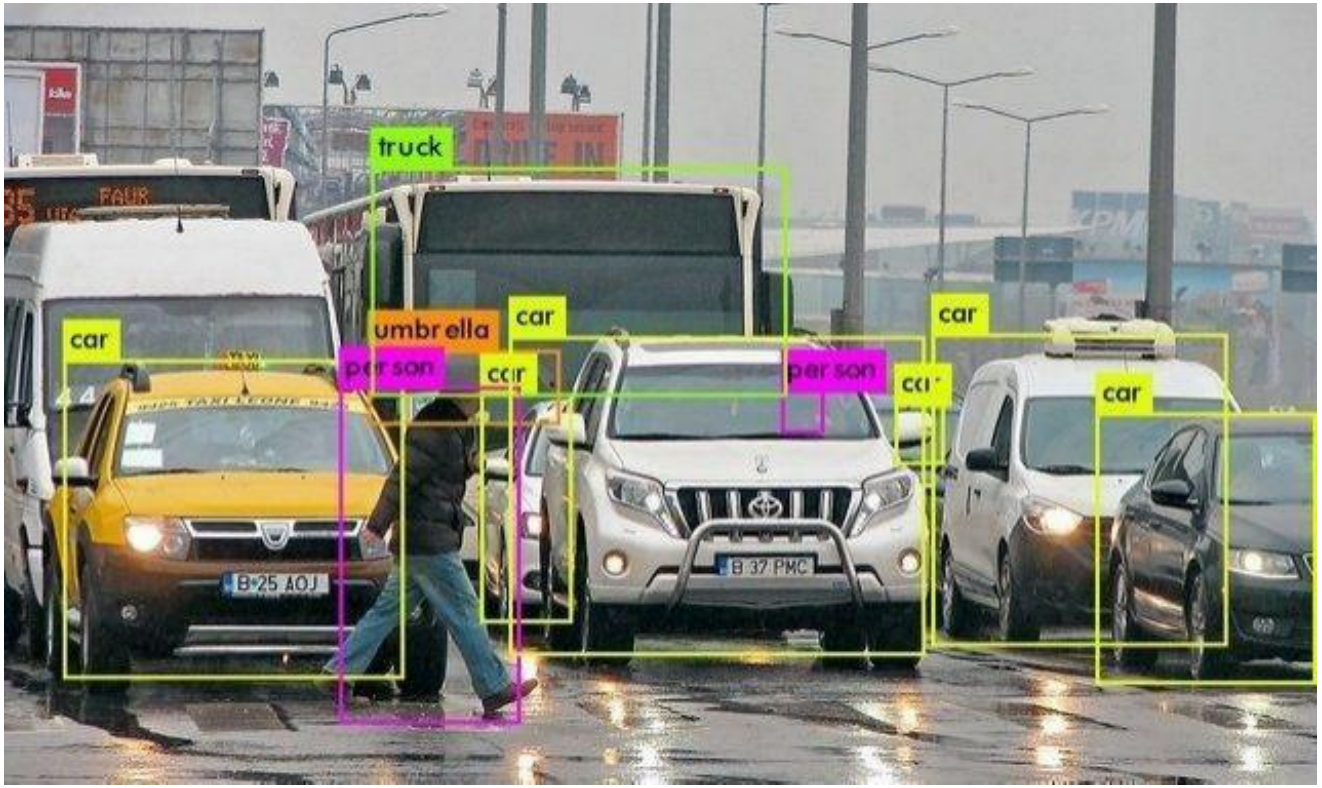
Deep learning and Artificial (Machine) Intelligence



By Chu'an at Alibaba

Deep learning is everywhere
even for scientific discovery

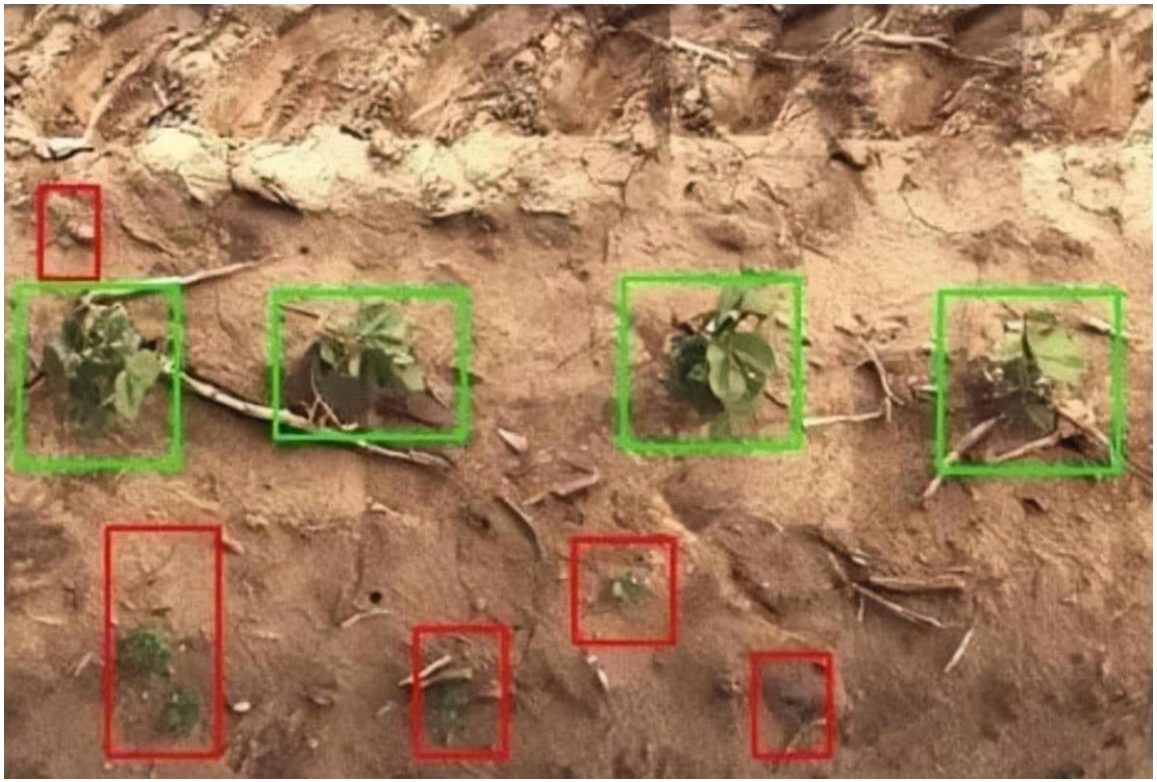
Automatic Drive



Deep learning helps cars identify objects on the road

Preprint arXiv:1811.10399, 2018

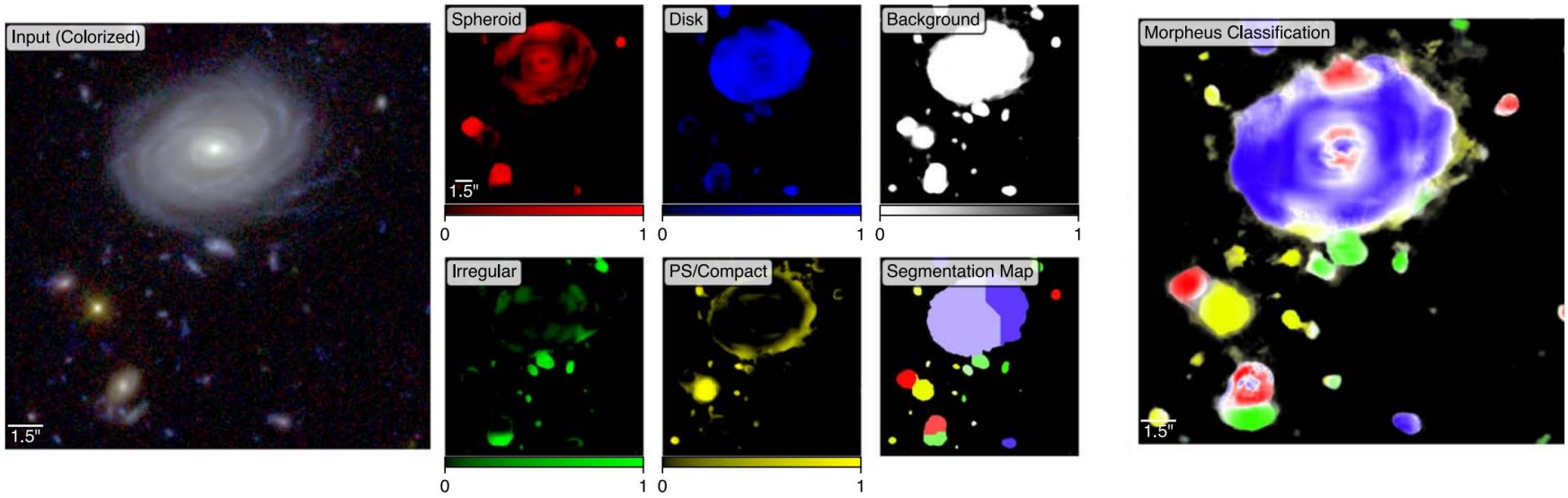
Agriculture



AI recognizes crops (green) and weeds (red), and instructs sprayer nozzles to avoid crops and spray only weeds

https://bluerivertechnology.com

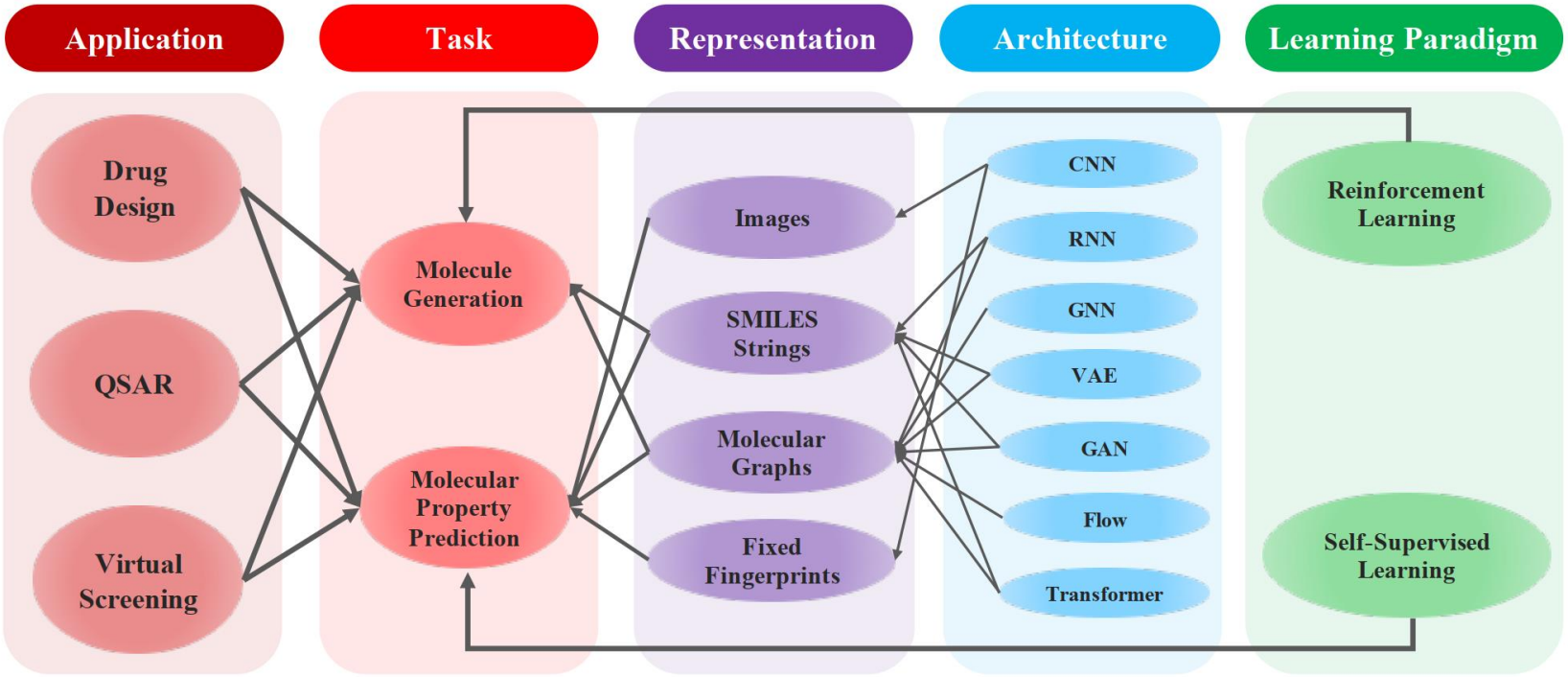
Astronomy



Morpheus helps human astronomers to complete the task of classifying space objects

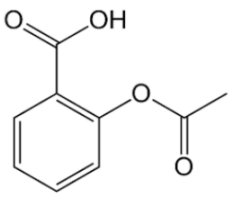
The Astrophysical Journal 248:20 (2020)

AI-driven Drug Discovery

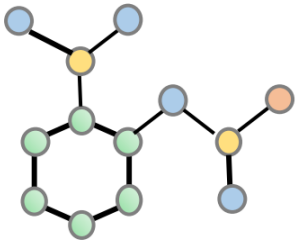


Preprint arXiv:2106.05386, 2021

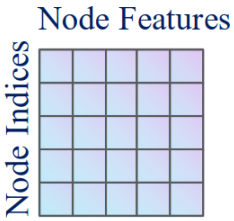
AI-driven Drug Discovery



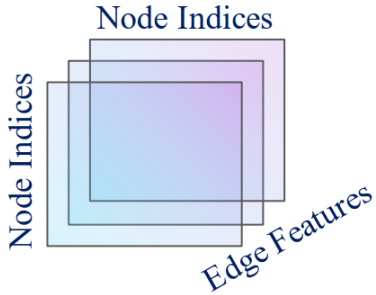
A. Kekulé Diagram



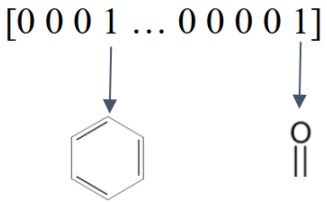
C. Molecular Graph



Node Feature Matrix



Adjacency Tensor



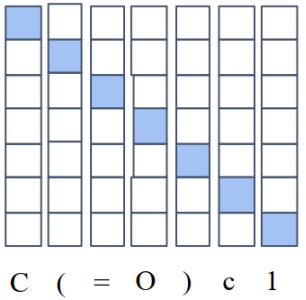
B. Fingerprints

CC(=O)Oc1ccccc1C(=O)O

D. SMILES String

Tokenization

One-Hot Encoding

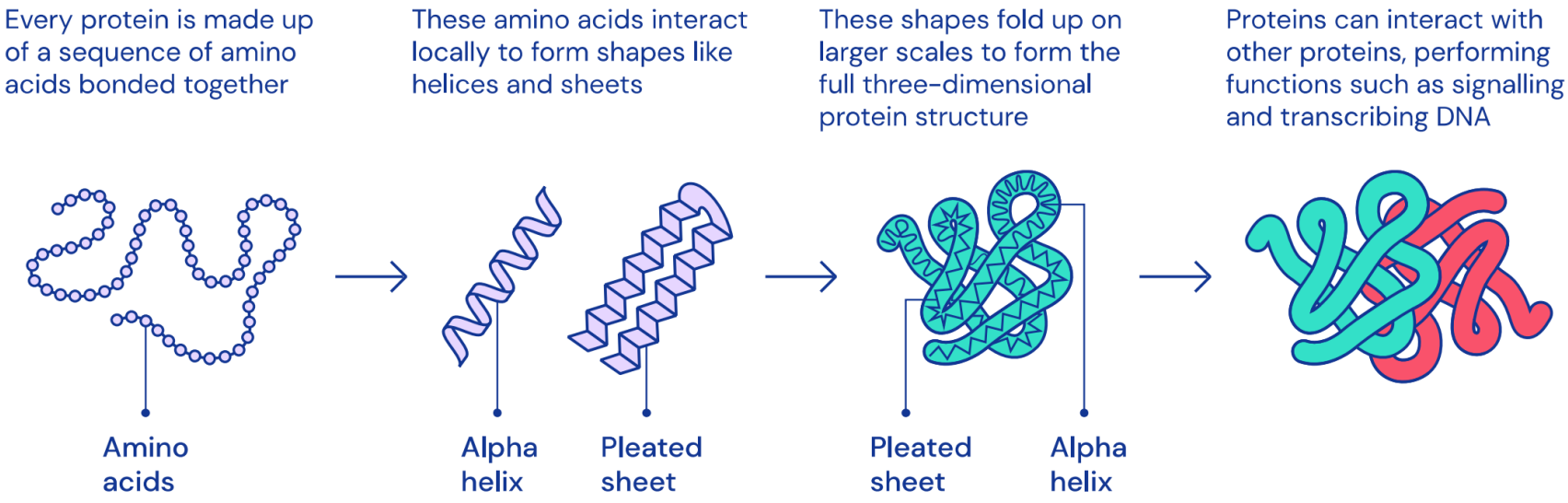


Small Molecule Representations

Preprint arXiv:2106.05386, 2021

AlphaFold: a solution to a 50-year-old grand challenge in biology

What is the protein folding problem?

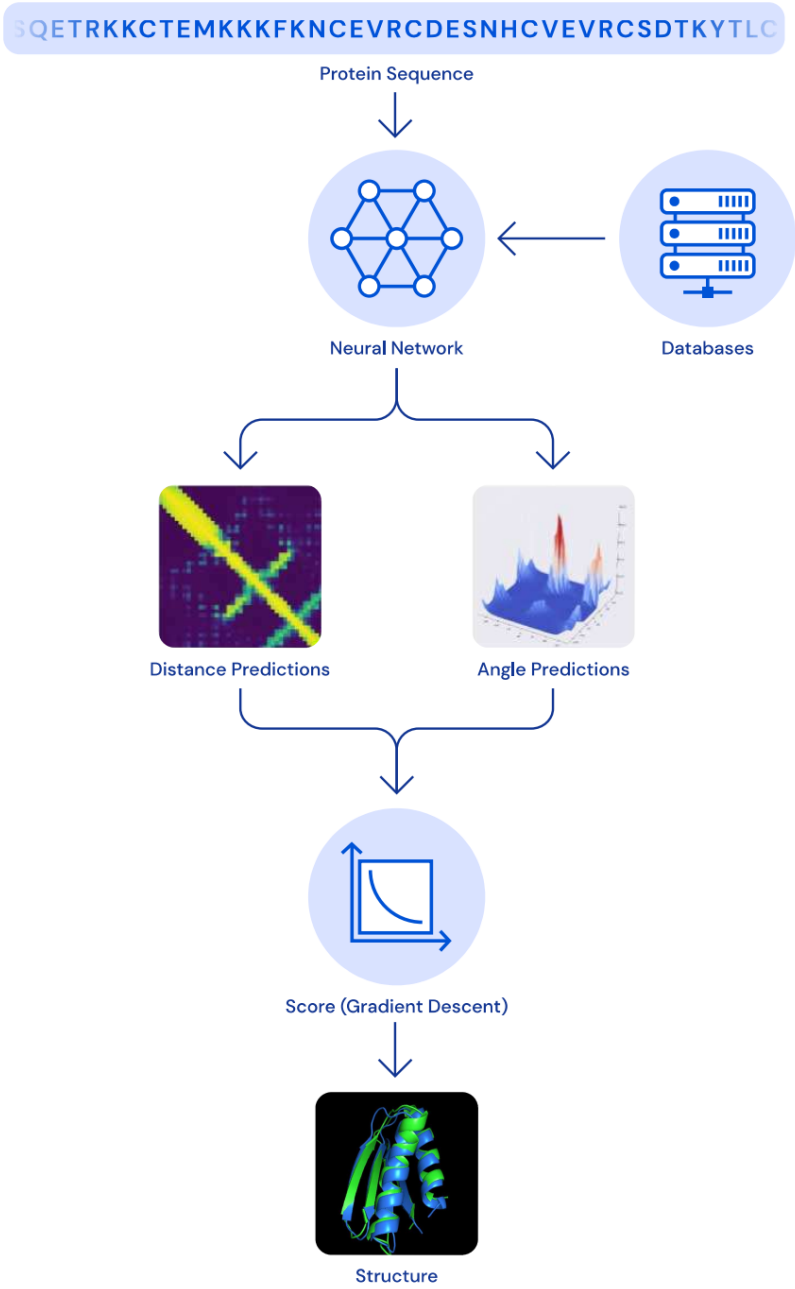


https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

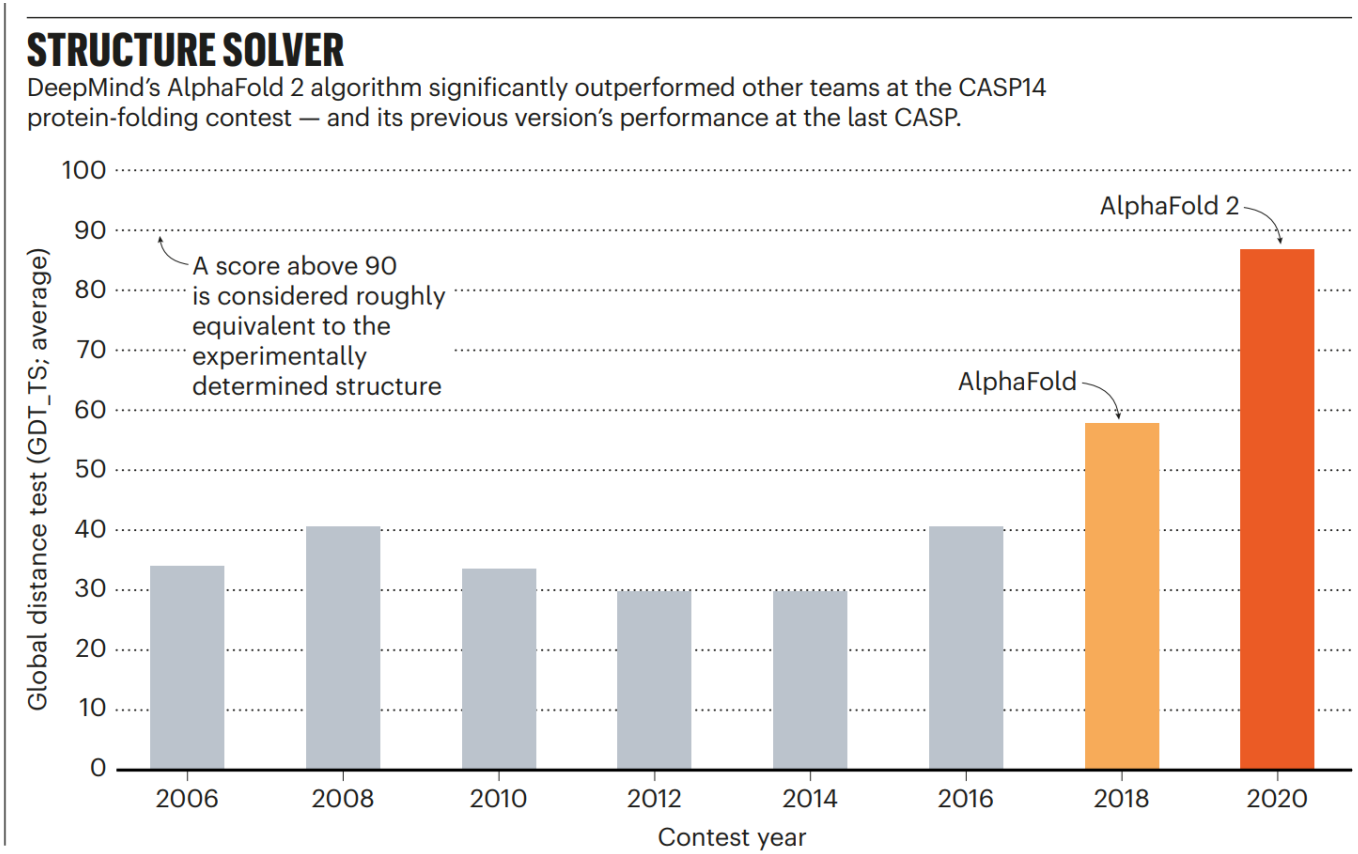
AlphaFold

The architecture of the AlphaFold system:

predicting structure from protein sequence



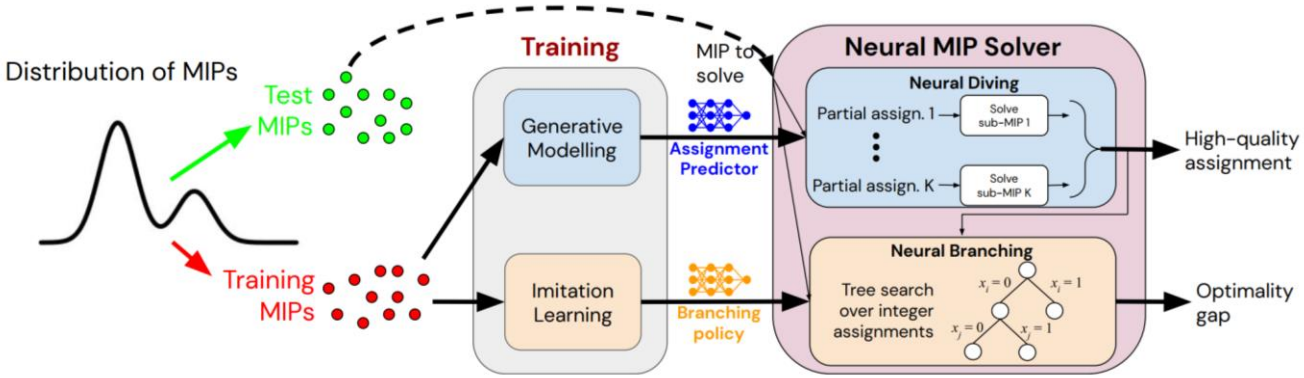
AlphaFold: a solution to a 50-year-old grand challenge in biology

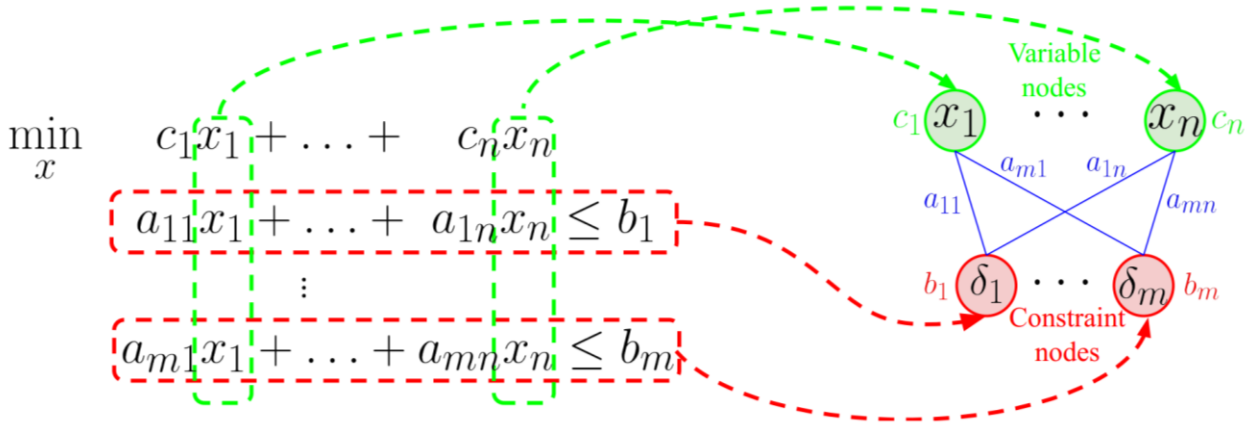


DeepMind's AlphaFold2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest.

Deep Learning and Combinatorial Optimization

Solving Mixed Integer Programs Using Neural Networks

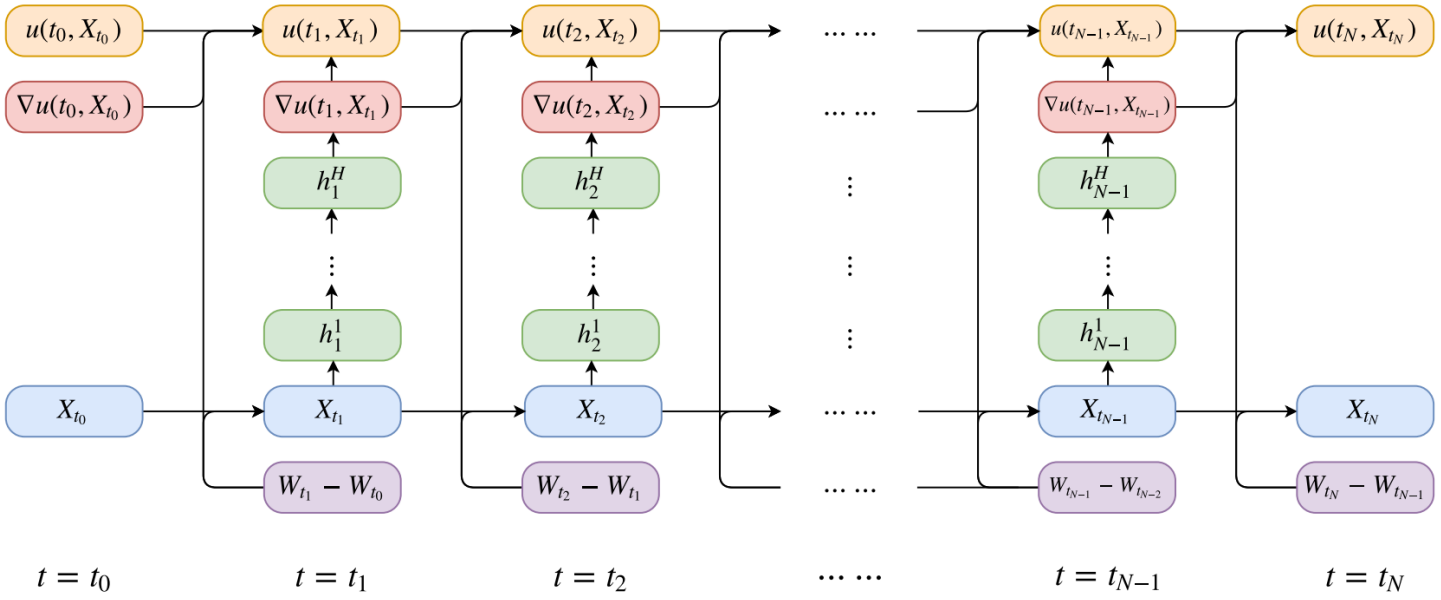




http://www.ipam.ucla.edu/programs/workshops/deep-learning-and-combinatorial-optimization/

Preprint arXiv:2012.13349

Deep Learning and Numerical Methods



Rough sketch of the architecture of the deep BSDE solver

Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations

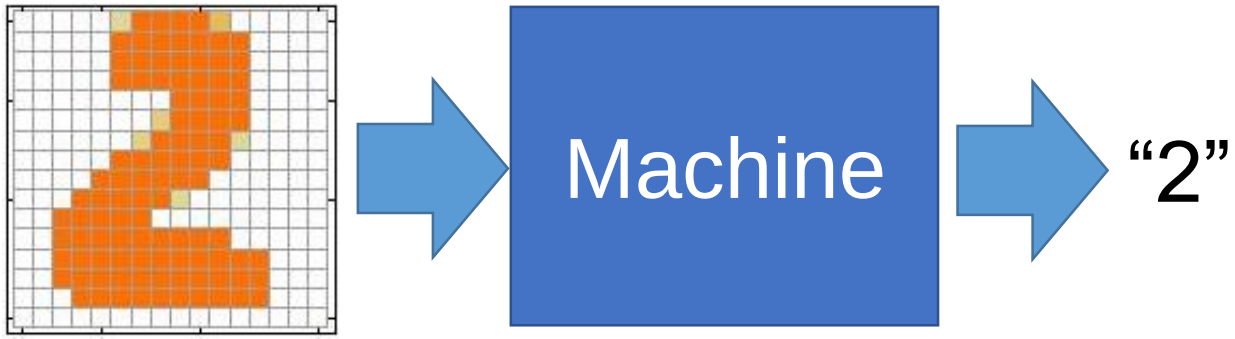
Commun. Math. Stat. 5, 349-380 (2017)

Essentials of Deep Learning

What people already knew in 1980s

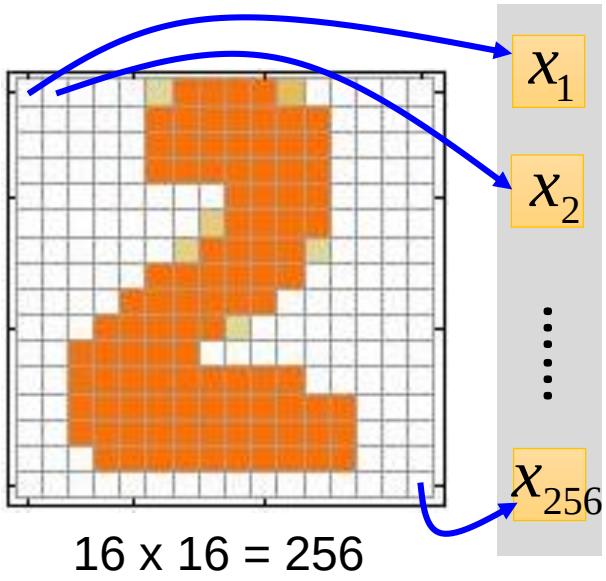
Example Application

- Handwriting Digit Recognition



Handwriting Digit Recognition

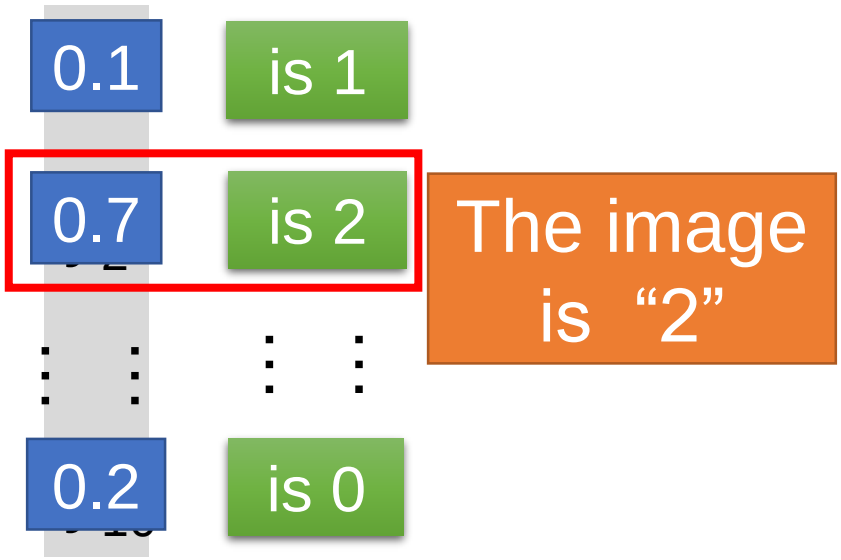
Input



Ink → 1

No ink → 0

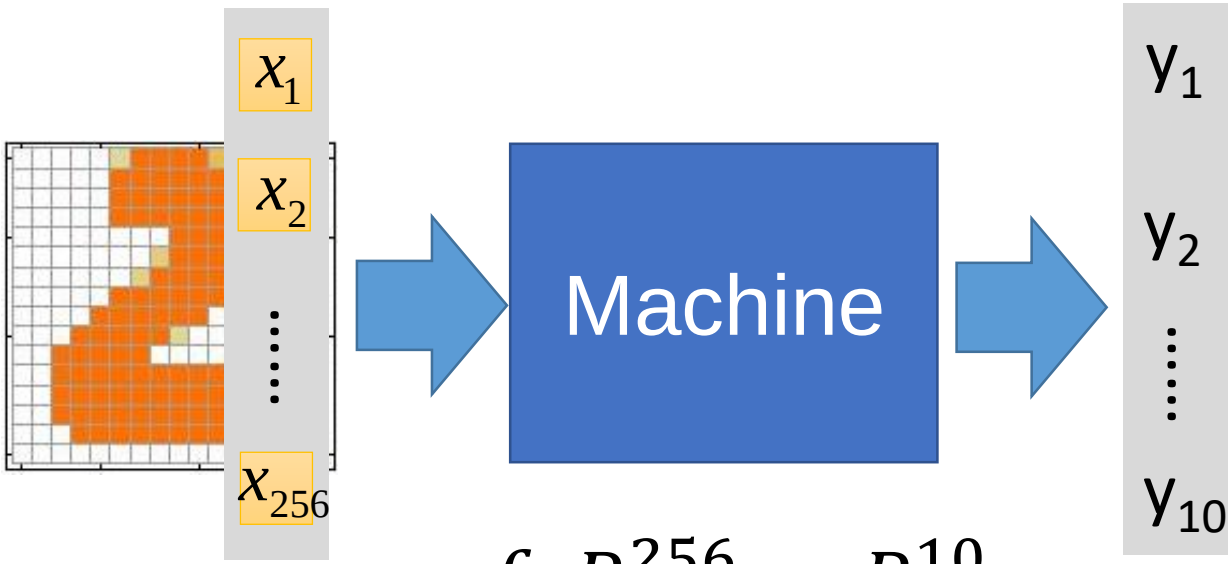
Output



Each dimension represents the confidence of a digit.

Example Application

- Handwriting Digit Recognition

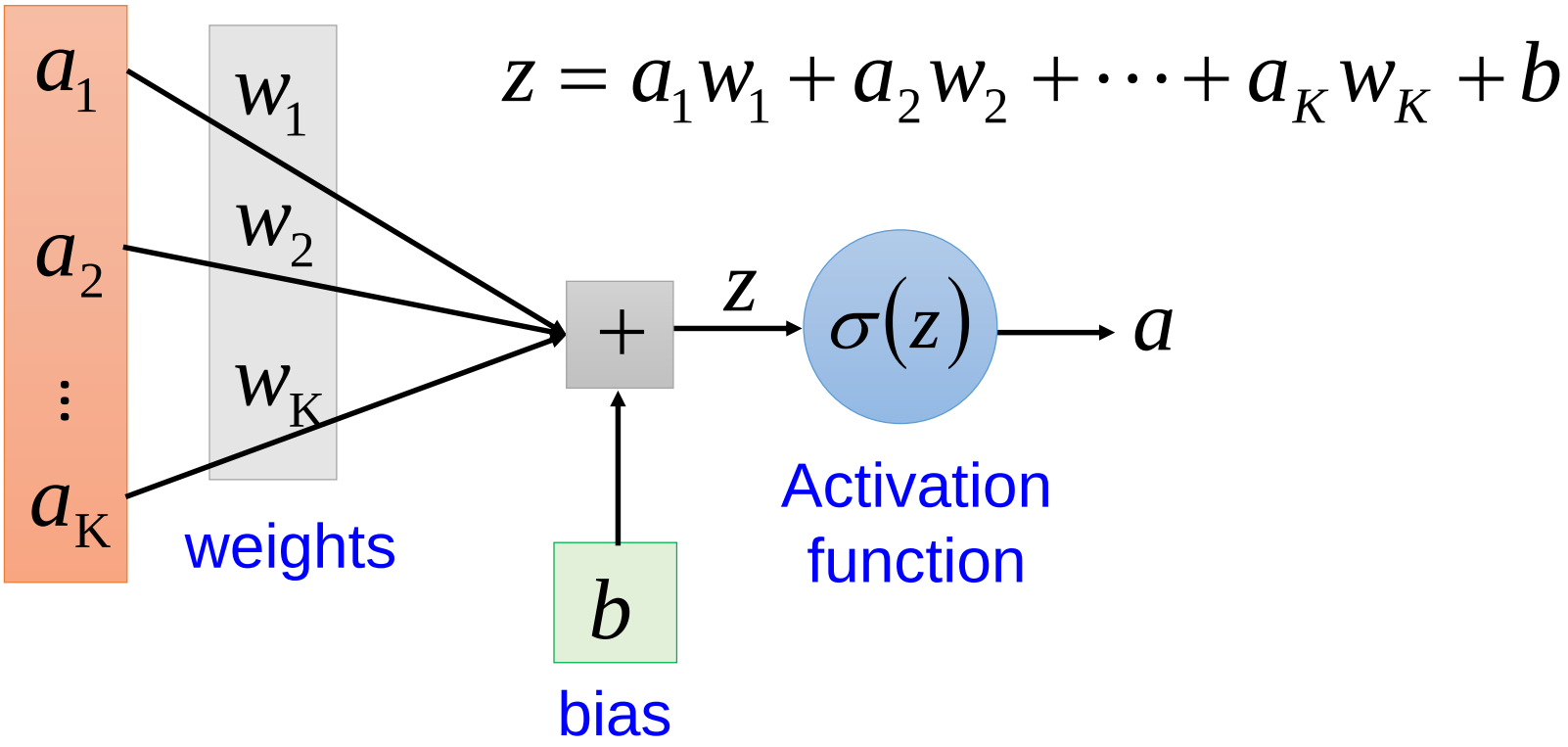


The task of Machine learning is to look for a function f.

In deep learning, the function f is represented by neural network

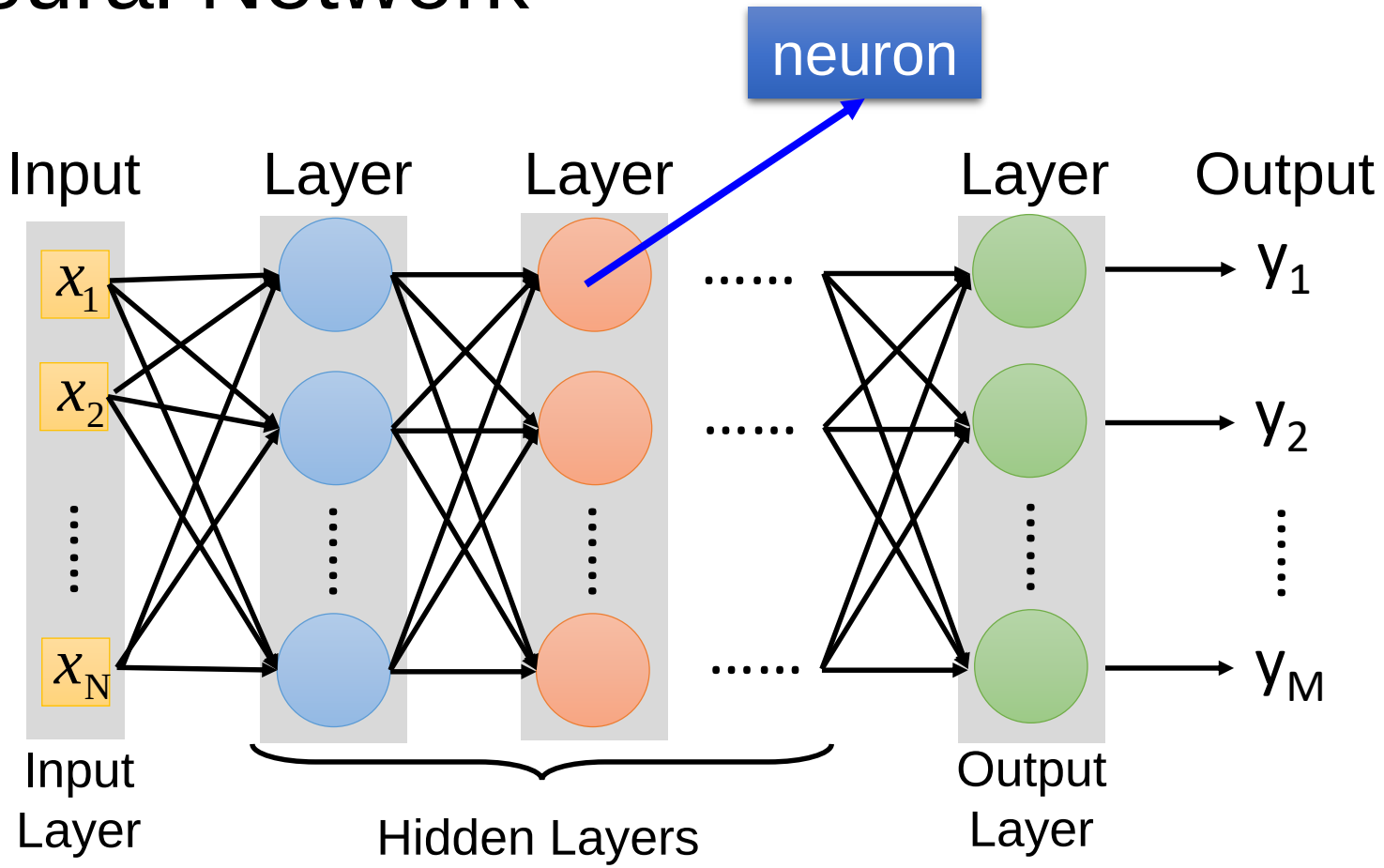
Element of Neural Network

Neuron $f: R^K \rightarrow R$



$$z = a_1 w_1 + a_2 w_2 + \cdots + a_K w_K + b$$

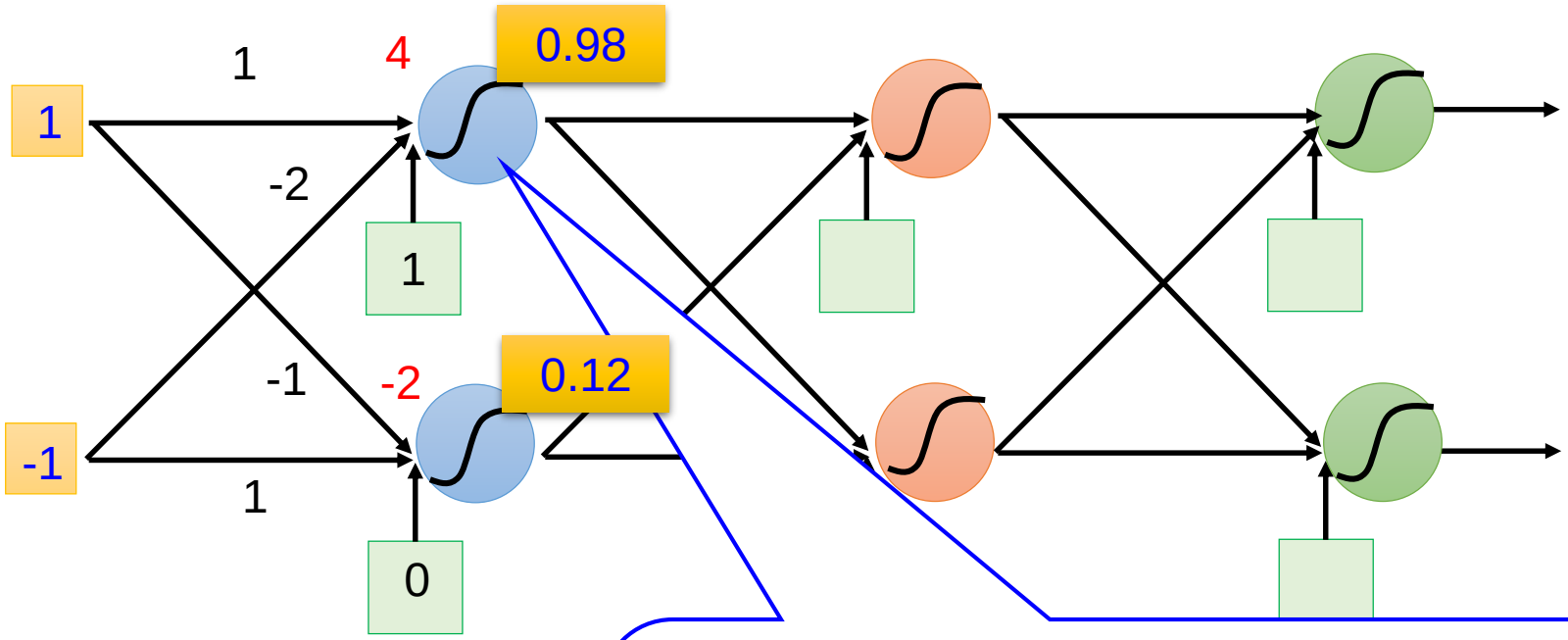
Neural Network



Deep means many hidden layers

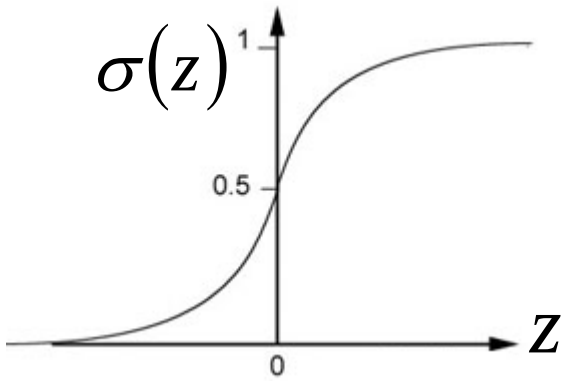
Fully Connected Feedforward Network

Example of Neural Network

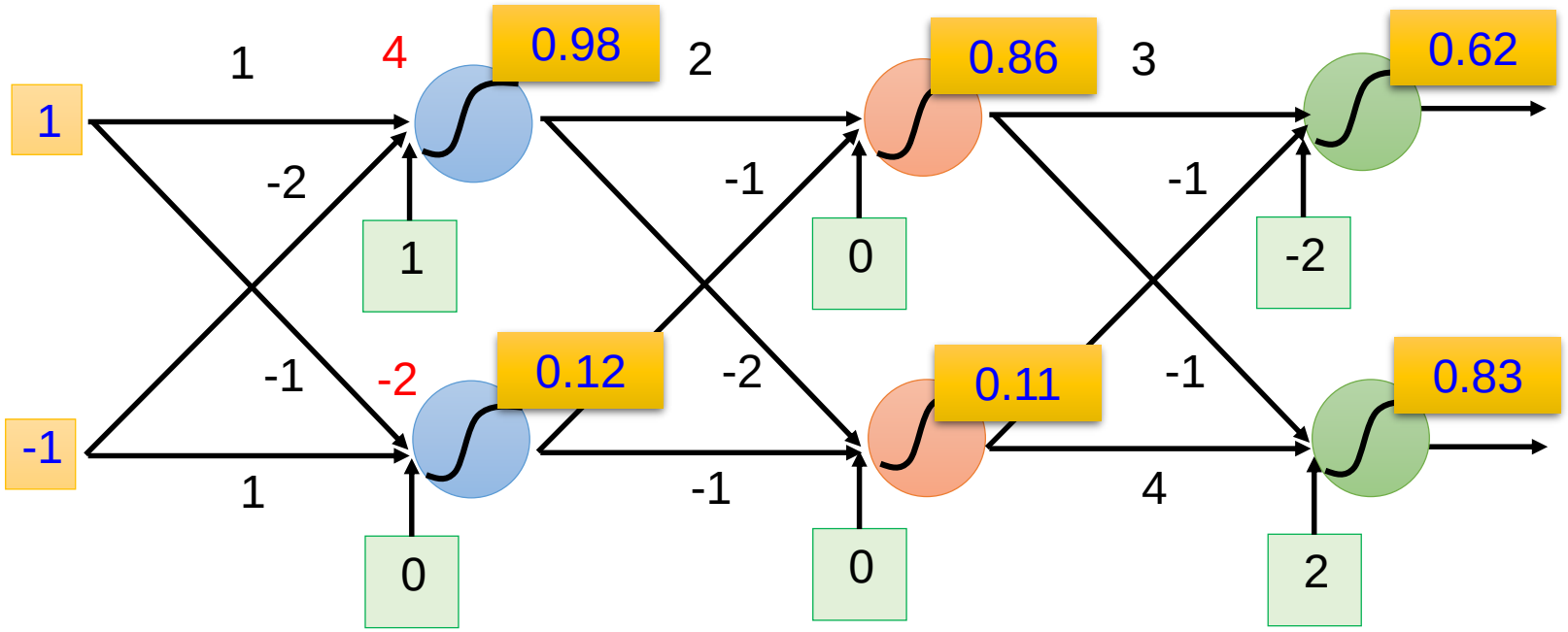


Sigmoid Function

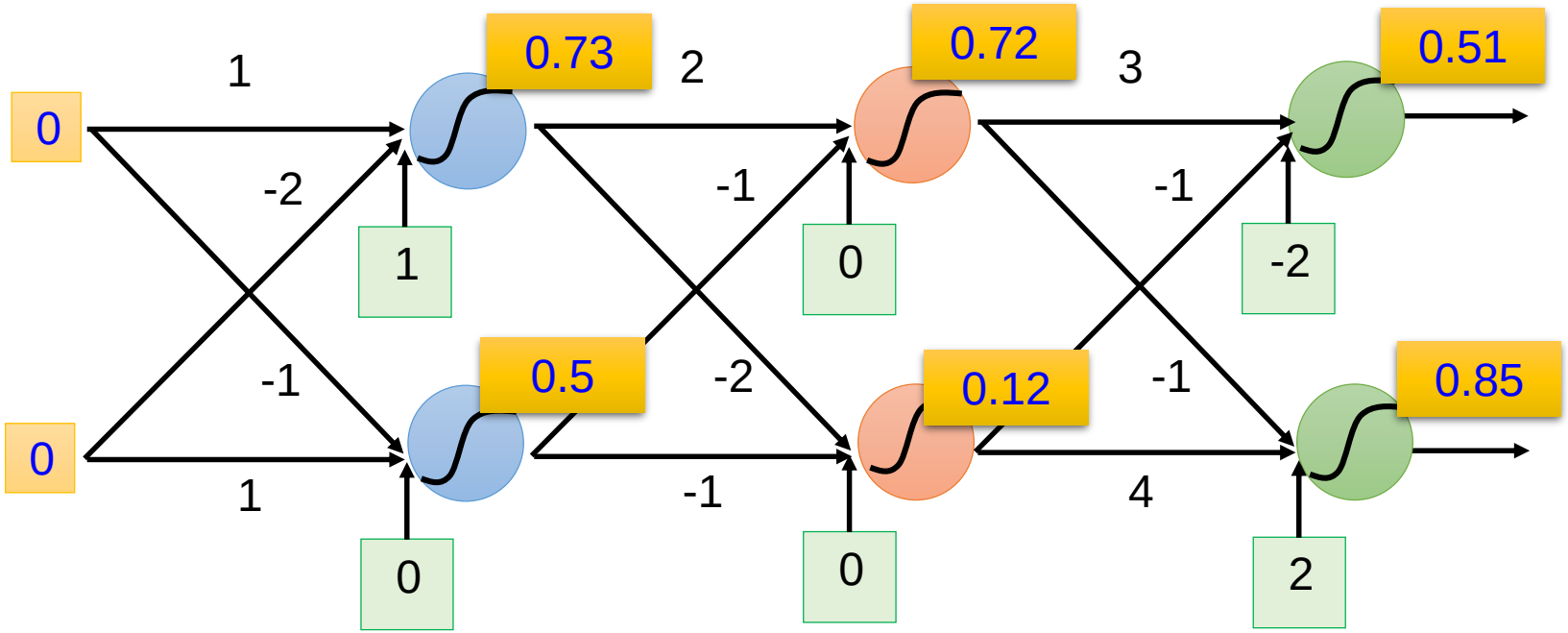
$$\sigma(z) = \frac{1}{1+e^{-z}}$$



Example of Neural Network



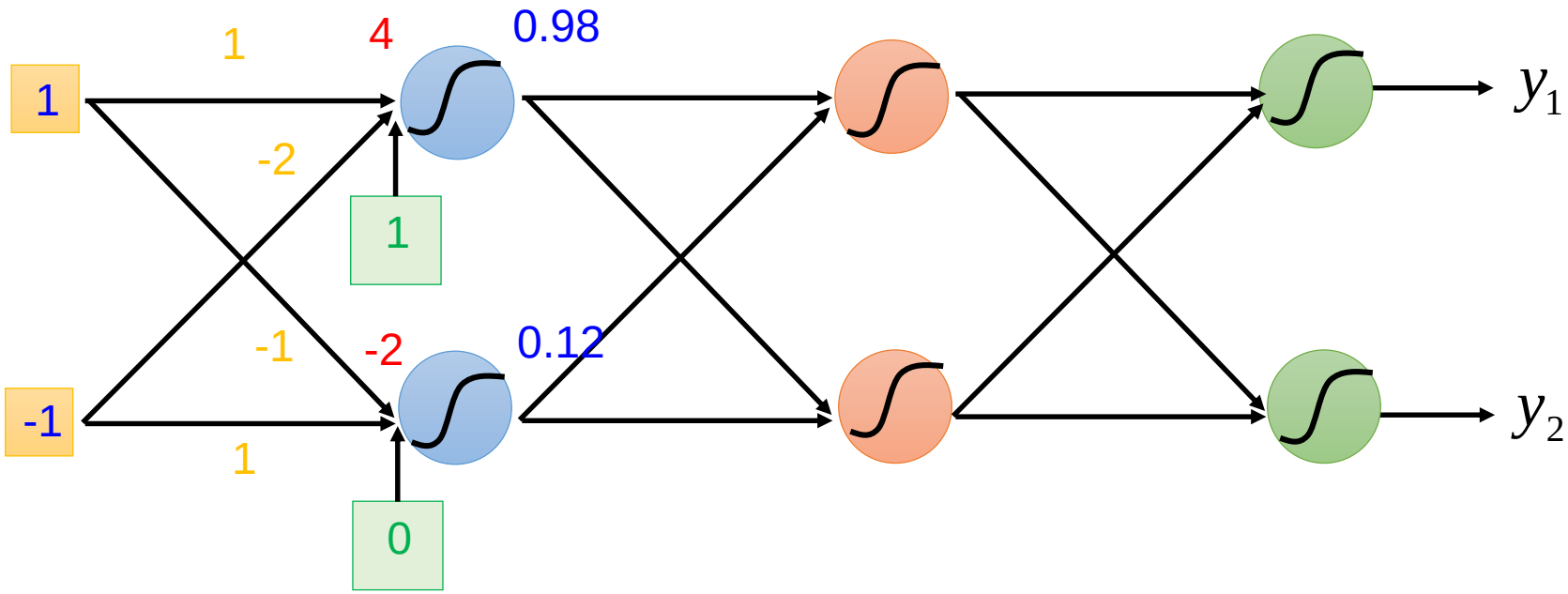
Example of Neural Network



$f: R^2 \rightarrow R^2$ $\quad f\left(\begin{bmatrix}1\\-1\end{bmatrix}\right) = \begin{bmatrix}0.62\\0.83\end{bmatrix}$ $\quad f\left(\begin{bmatrix}0\\0\end{bmatrix}\right) = \begin{bmatrix}0.51\\0.85\end{bmatrix}$

Different parameters define different function

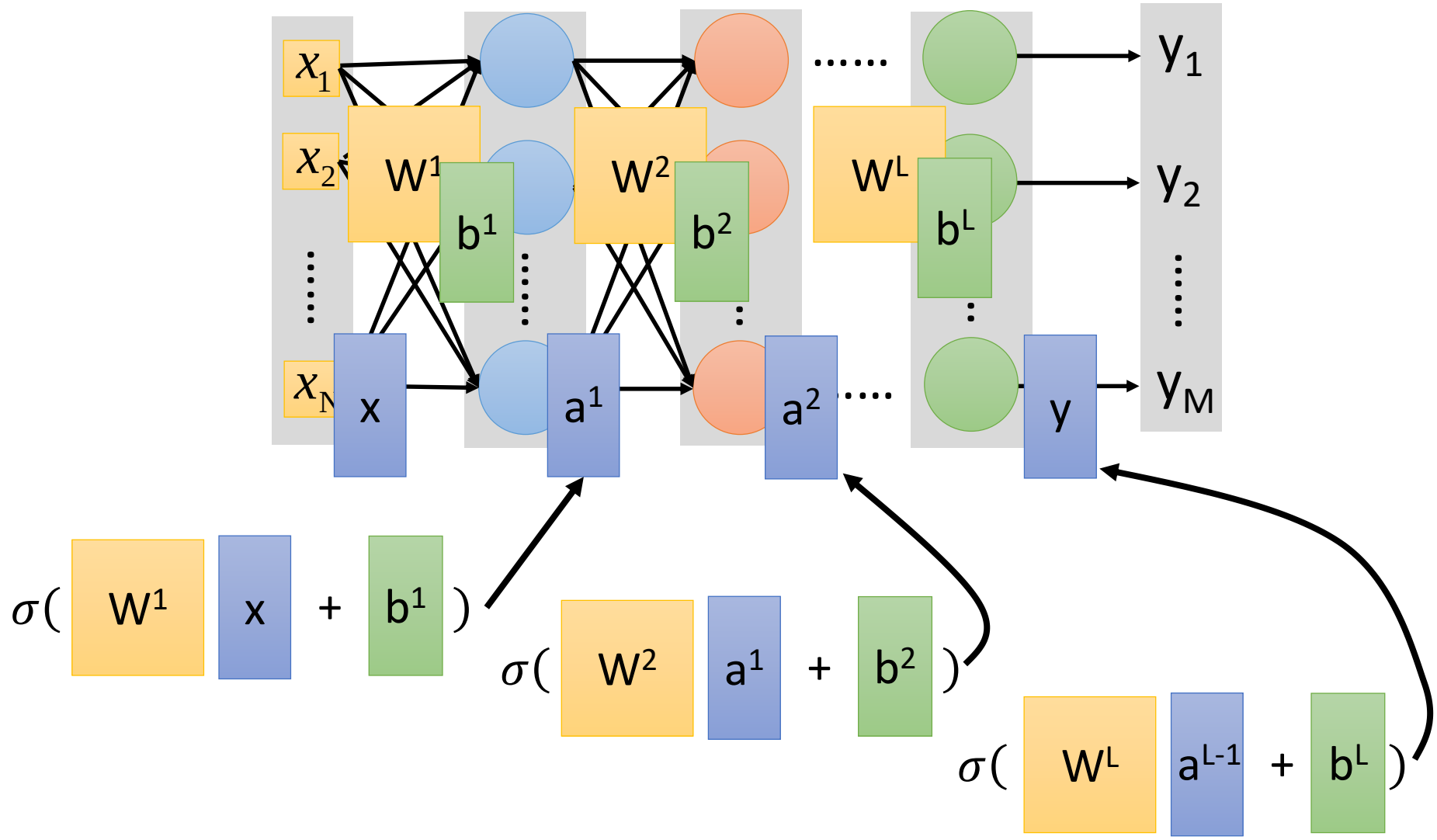
Matrix Operation



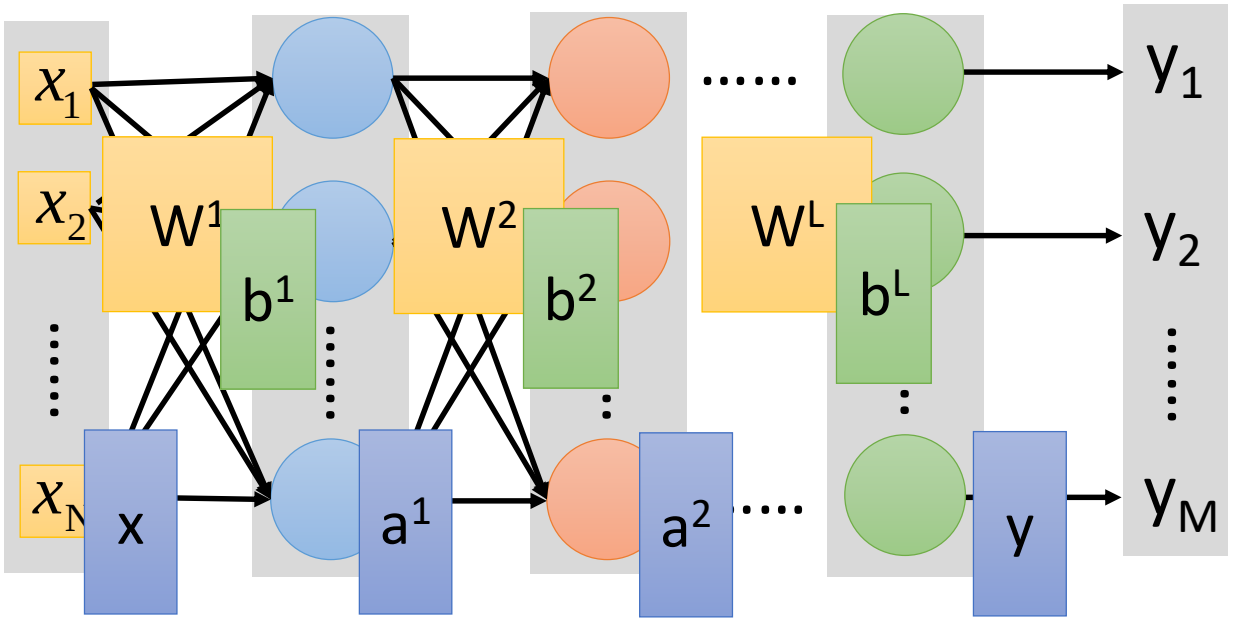
$$\sigma\left(\begin{bmatrix} 1 & -2 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 0.98 \\ 0.12 \end{bmatrix}$$

$$\begin{bmatrix} 4 \\ -2 \end{bmatrix}$$

Neural Network



Neural Network



$$y = f(x)$$

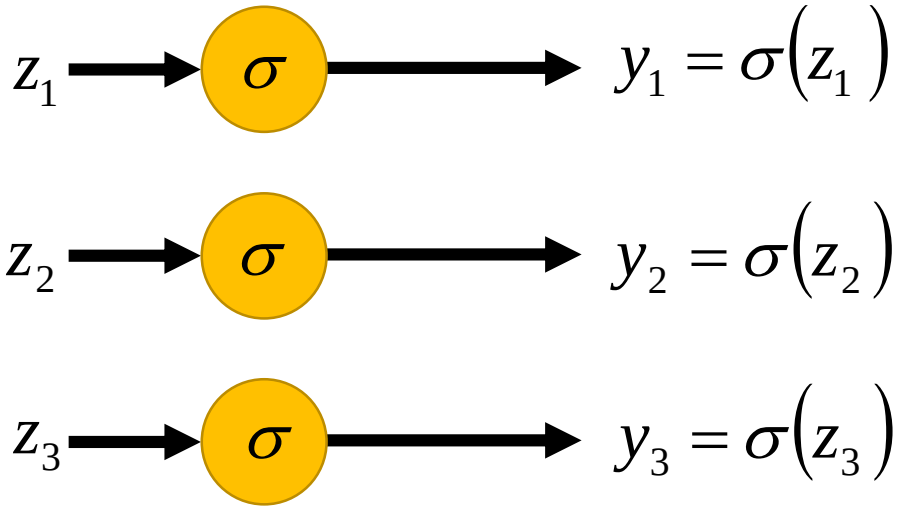
Using parallel computing techniques to speed up matrix operation

$$= \sigma(W^L \cdots \sigma(W^2 \sigma(W^1 x + b^1) + b^2) \cdots + b^L)$$

Softmax

- Softmax layer as the output layer

Ordinary Layer



In general, the output of network can be any value.

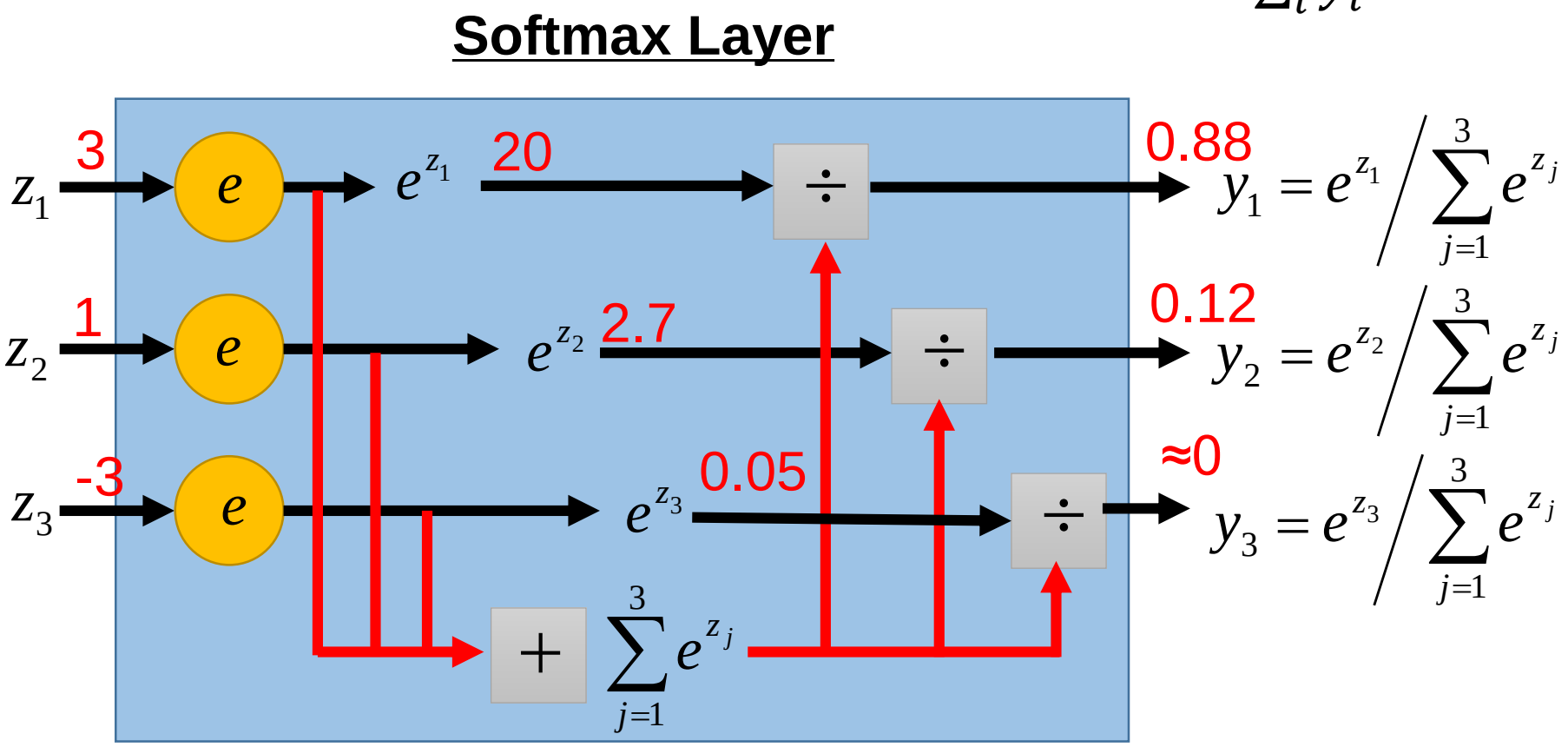
May not be easy to interpret !

Softmax

- Softmax layer as the output layer

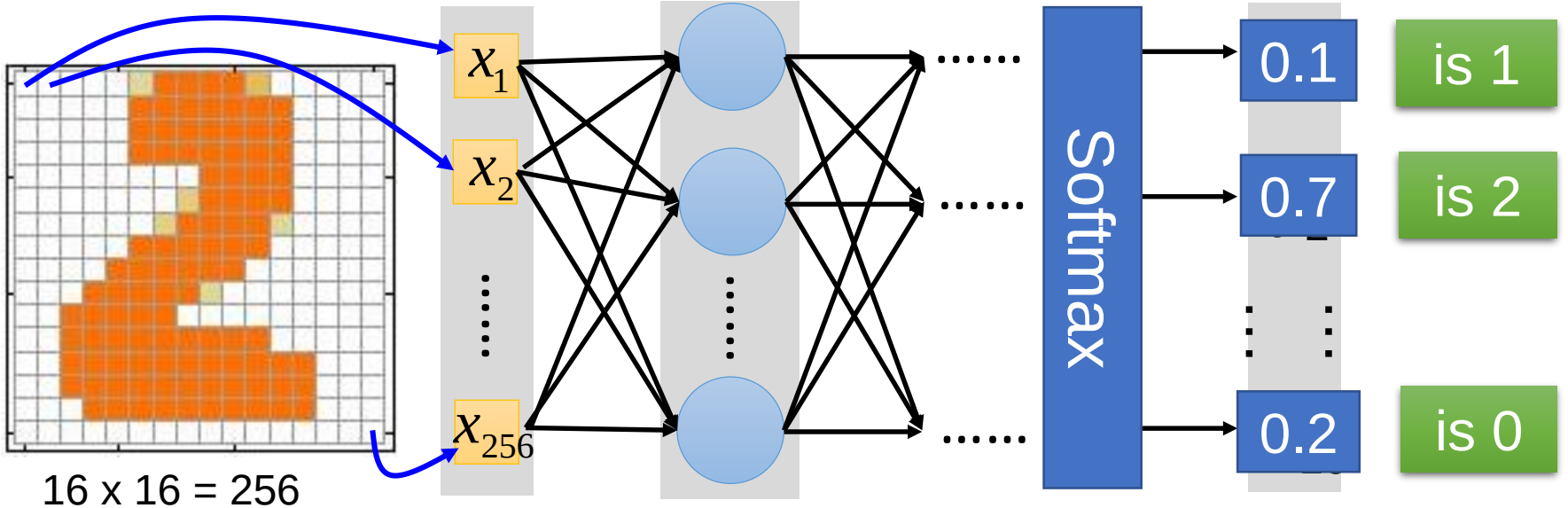
Probability:

- $1 > y_i > 0$
- $\sum_i y_i = 1$



How to Set Network Parameters

$$\theta = \{W^1, b^1, W^2, b^2, \cdots W^L, b^L\}$$



16 x 16 = 256

Ink → 1

No ink → 0

Set the network parameters θ such that ……

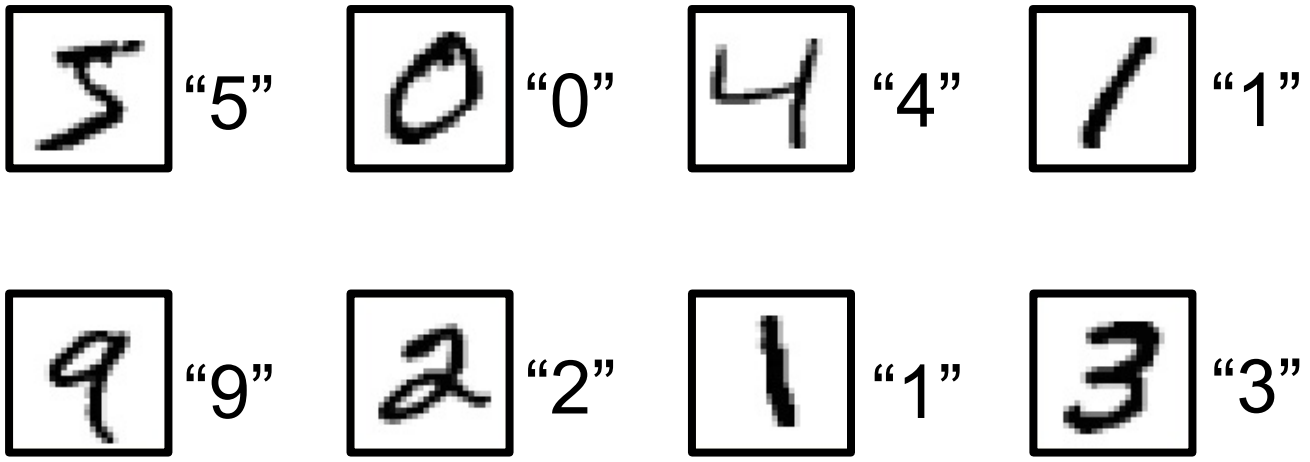
Input:  → y_1 has the maximum value

Input:  → y_2 has the maximum value

How to let the neural network achieve this

Training Data

- Preparing training data: images and their labels



Using the training data to find the network parameters.

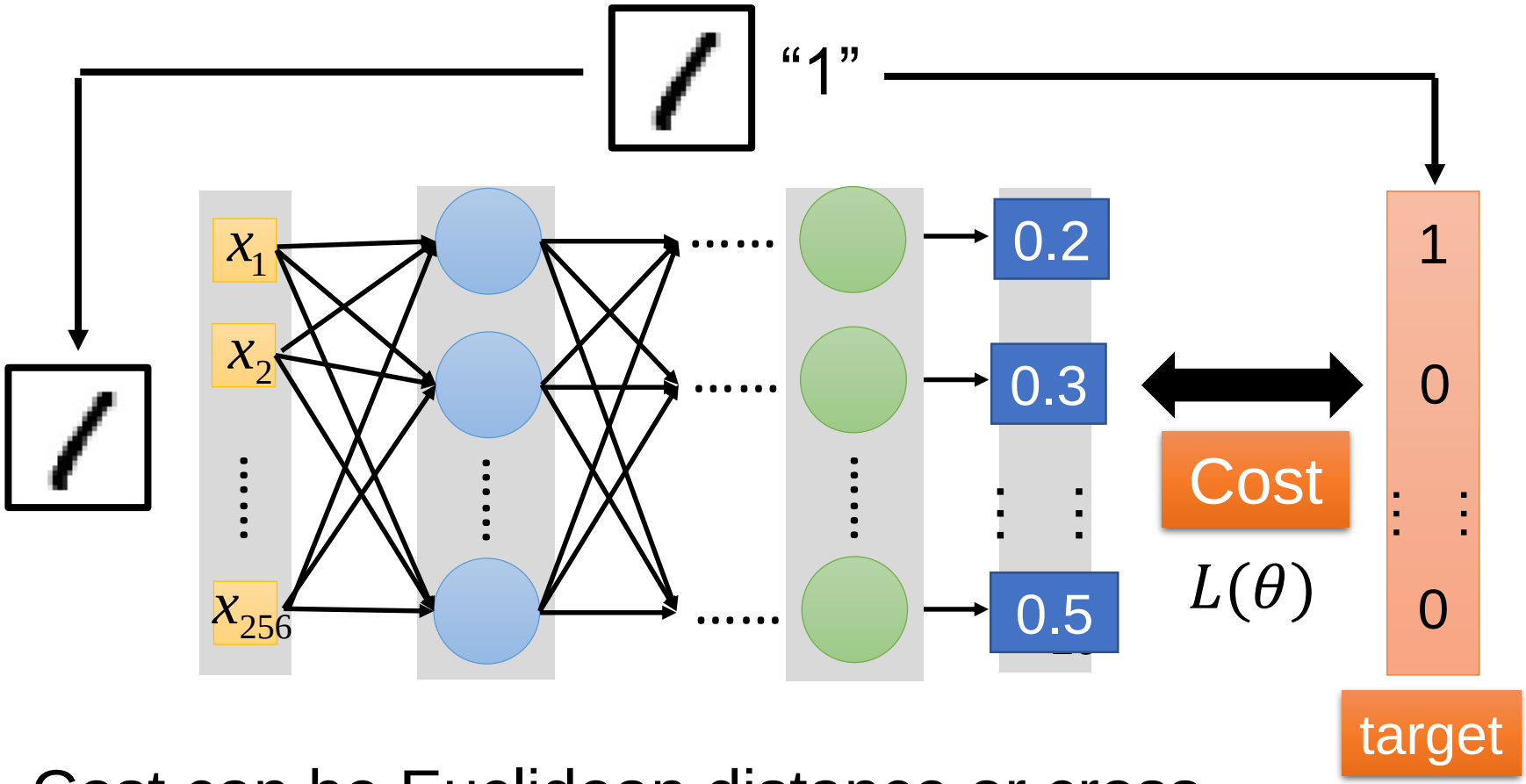
The MNIST Database
of handwritten digits

- Training set of 60,000 examples vs Testing set of 10,000 examples.
- It is a subset of a larger set available from NIST.
- The digits have been size-normalized and centered in a fixed-size image.
- It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data.

http://yann.lecun.com/exdb/mnist/

Cost

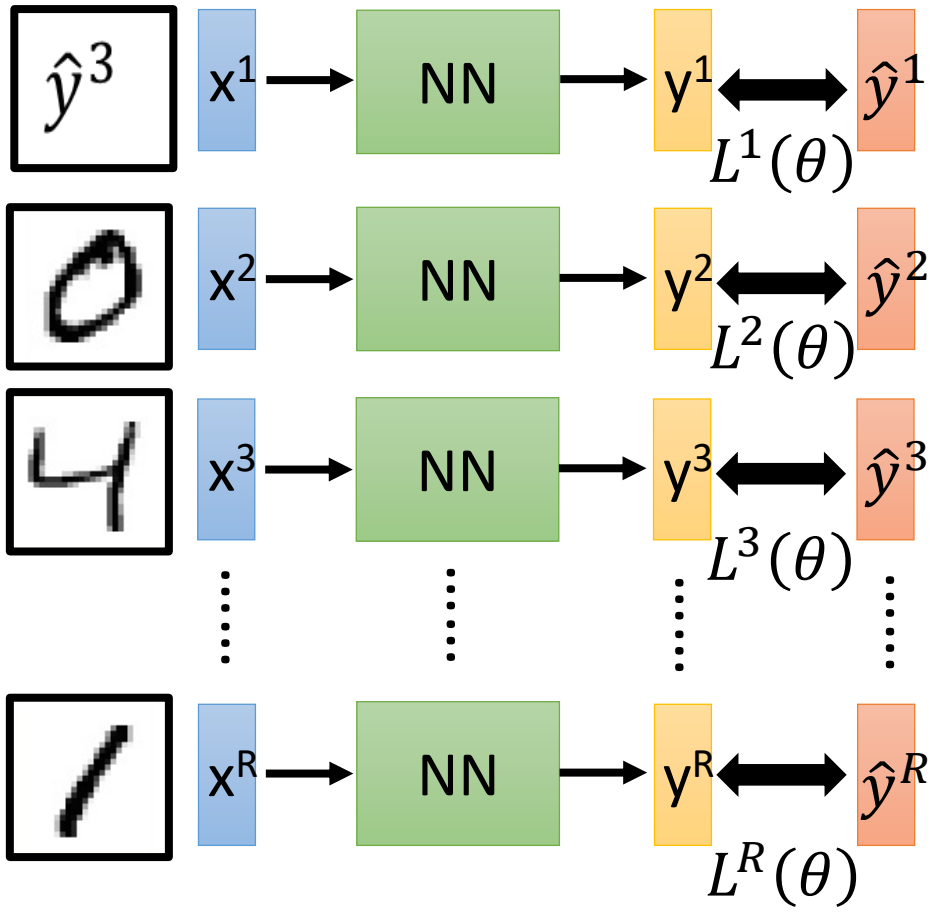
Given a set of network parameters θ, each example has a cost value.



Cost can be Euclidean distance or cross entropy of the network output and target

Total Cost

For all training data …



Total Cost:

$$C(\theta) = \sum_{r=1}^{R} L^r(\theta)$$

How bad the network parameters θ is on this task?

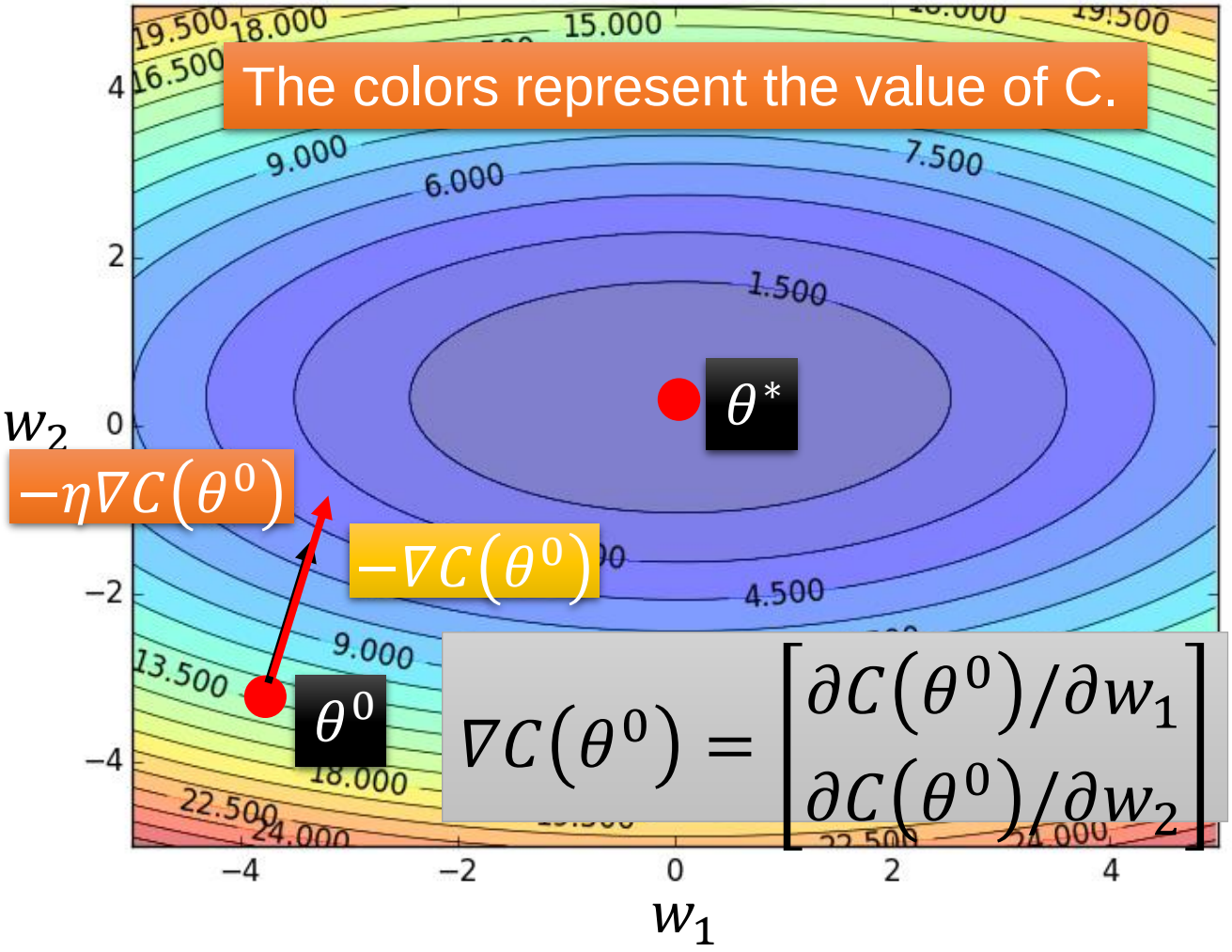
Find the network parameters θ^* that minimize this value

Gradient Descent

Assume there are only two parameters w_1 and w_2 in a network.

$$\theta = \{w_1, w_2\}$$

Error Surface



Randomly pick a starting point θ^0

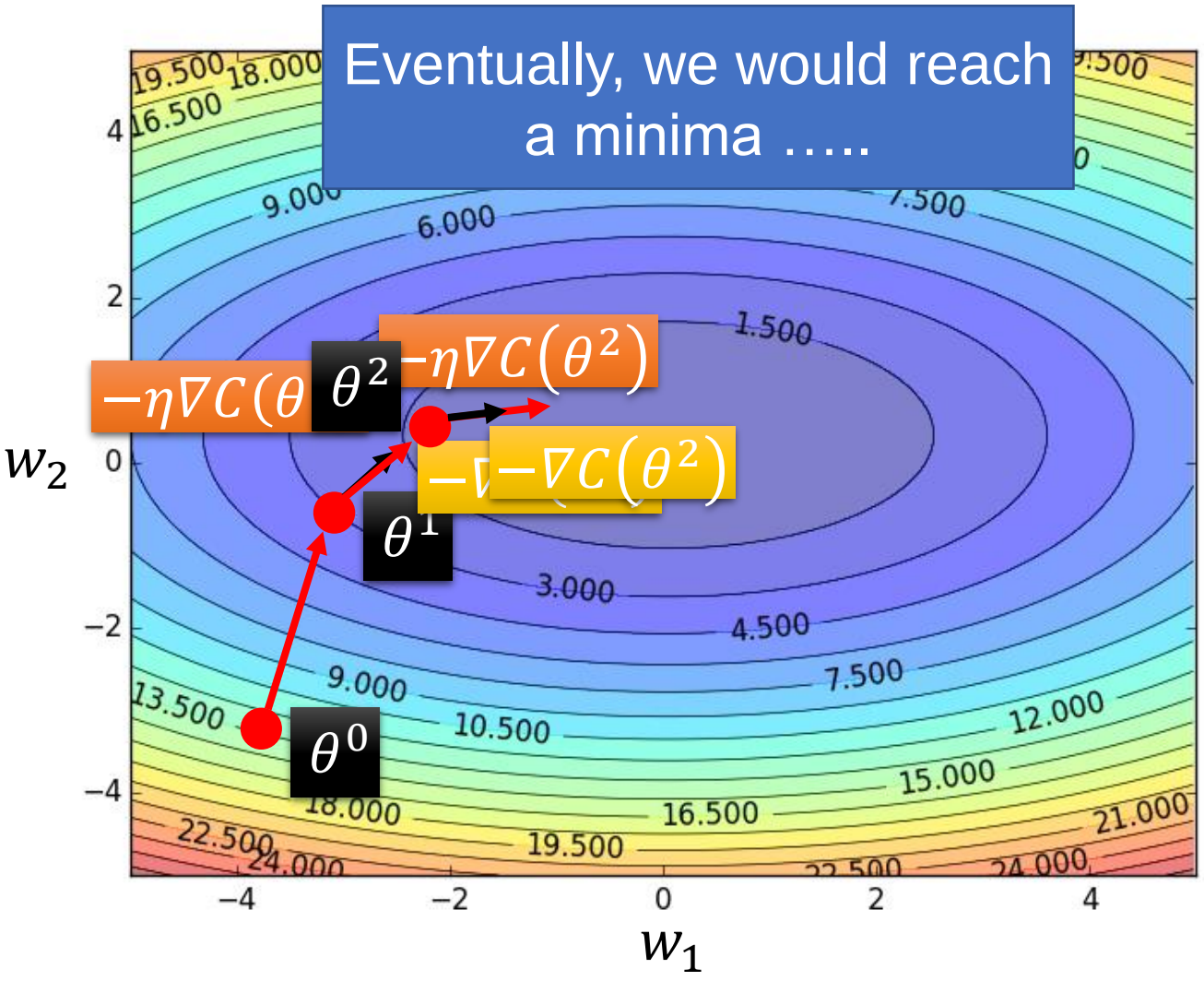
Compute the negative gradient at θ^0

$\Rightarrow -\nabla C(\theta^0)$

Times the learning rate η

$\Rightarrow -\eta\nabla C(\theta^0)$

Gradient Descent



Randomly pick a starting point θ^0

Compute the negative gradient at θ^0

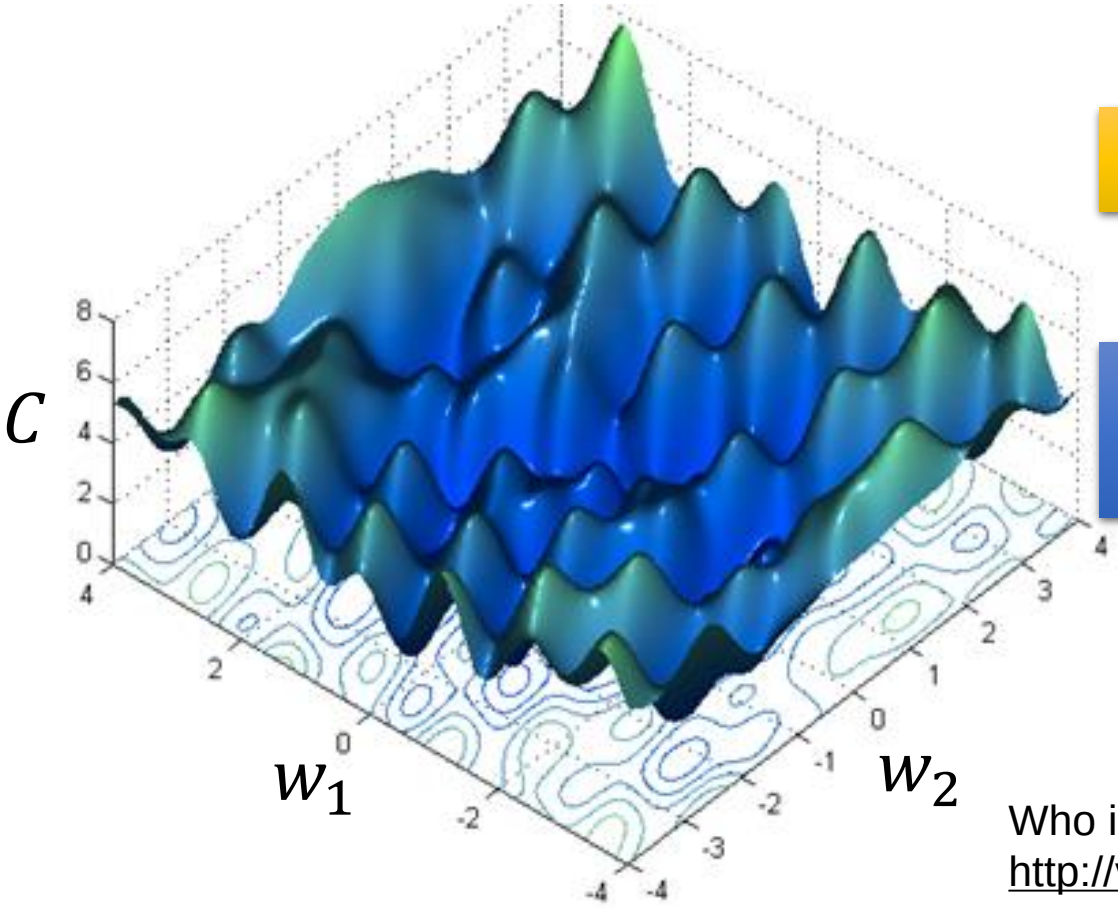
$-\nabla C(\theta^0)$

Times the learning rate η

$-\eta\nabla C(\theta^0)$

Local Minima

- Gradient descent never guarantee global minima

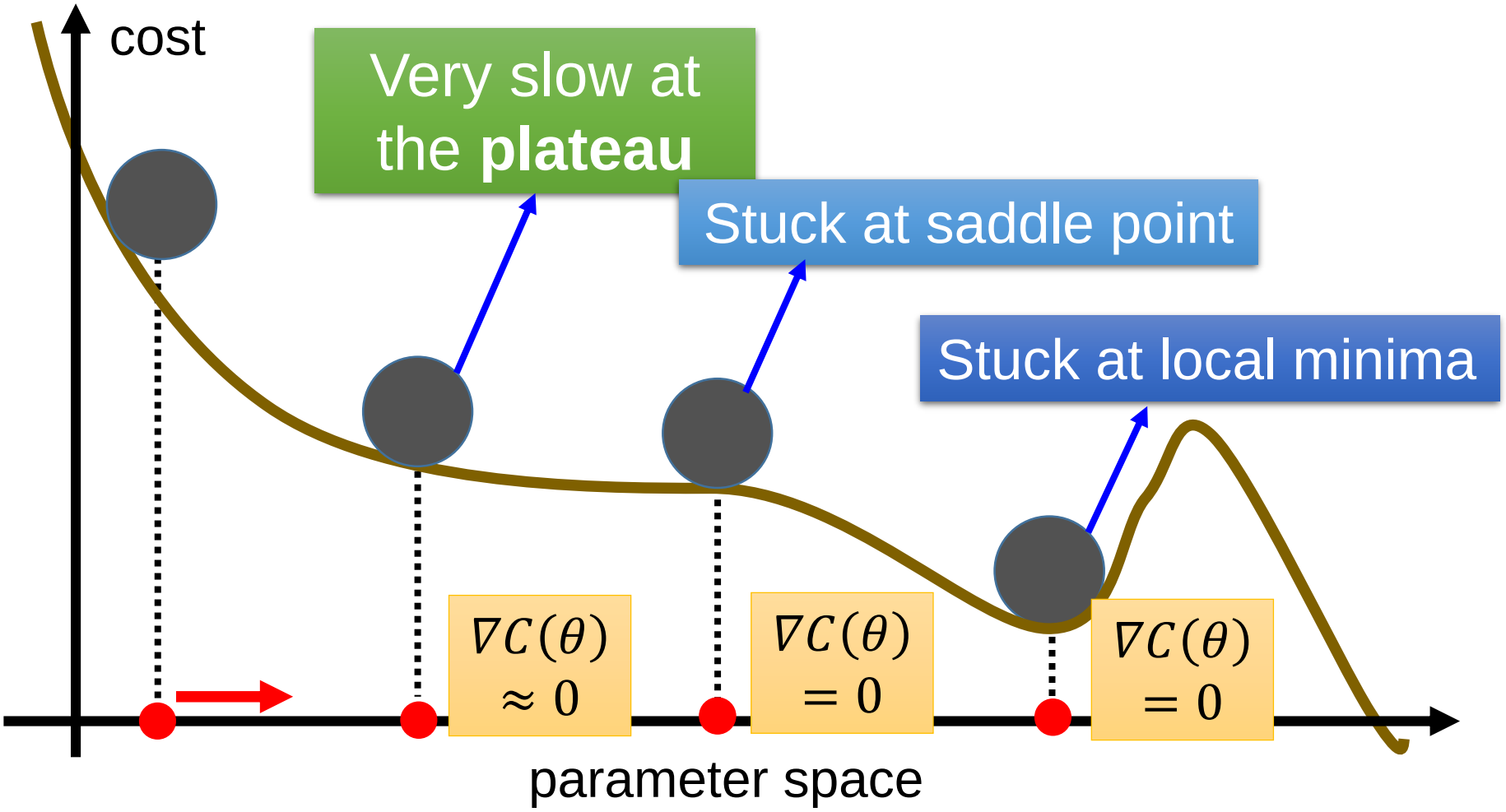


Different initial point θ^0

Reach different minima, so different results

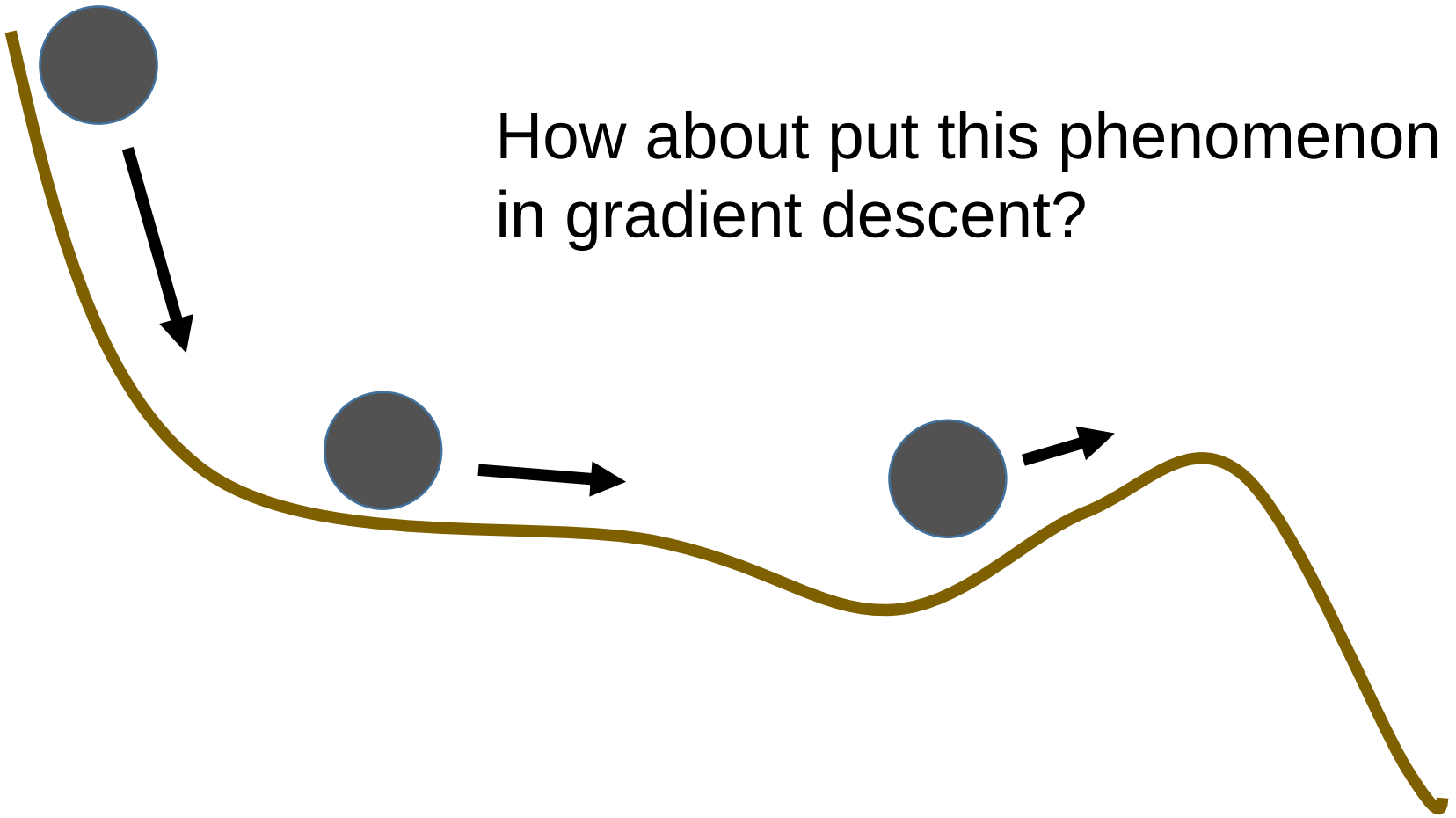
Who is Afraid of Non-Convex Loss Functions?
http://videolectures.net/eml07_lecun_wia/

Besides Local Minima ……



In Physical World ……

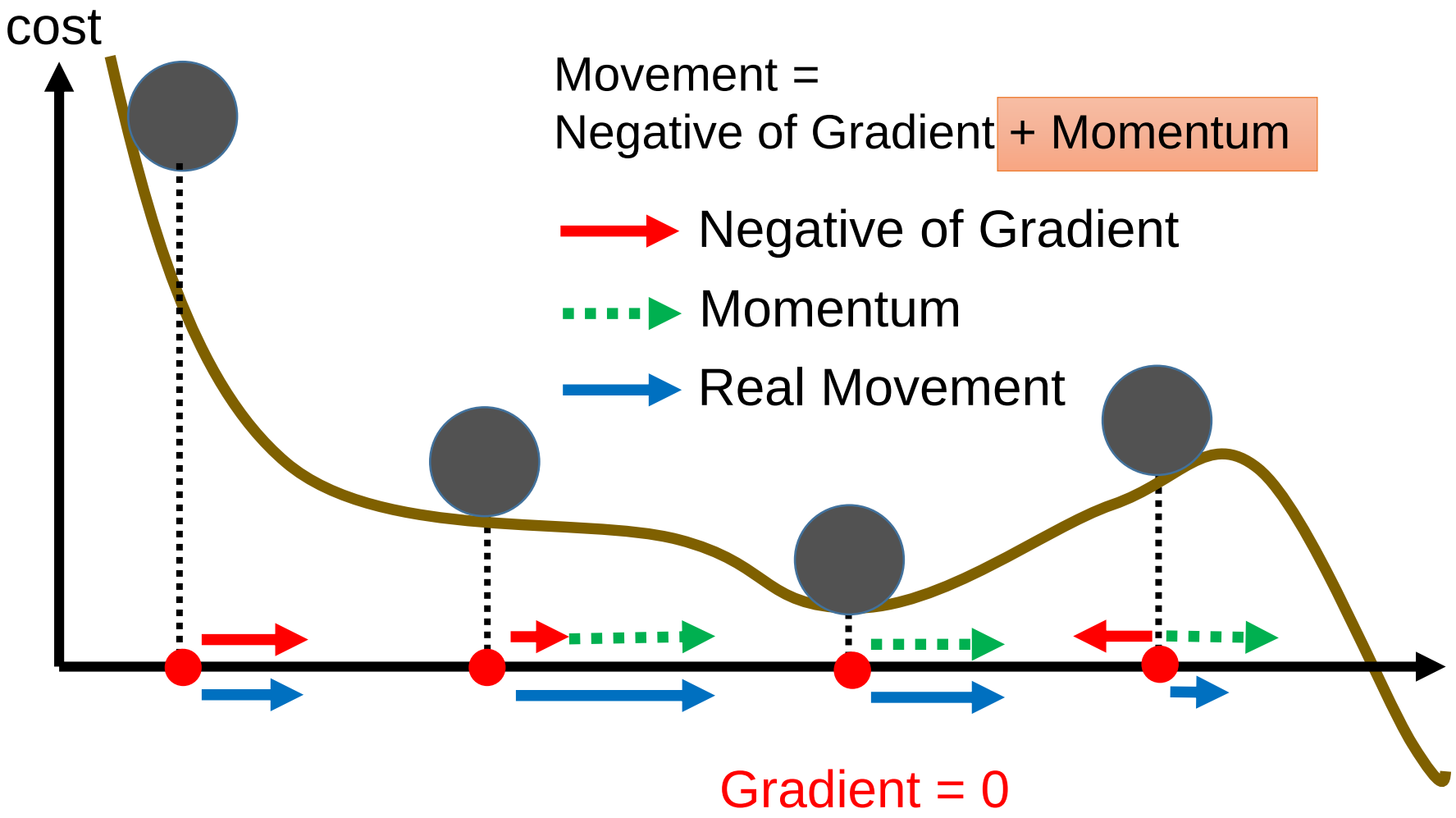
- Momentum



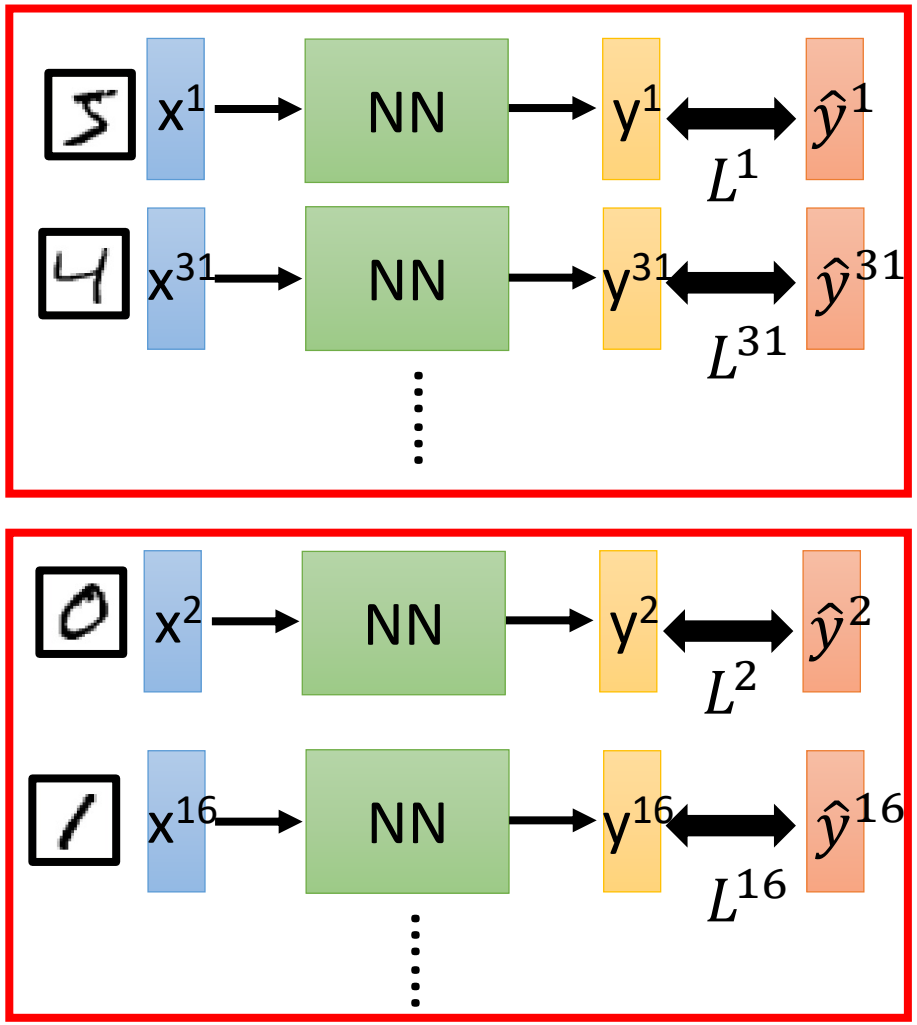
How about put this phenomenon in gradient descent?

Momentum

Still not guarantee reaching global minima, but give some hope ……



Mini-batch



- Randomly initialize θ^0
- Pick the 1st batch

$$C = L^1 + L^{31} + \cdots$$

$$\theta^1 \leftarrow \theta^0 - \eta \nabla C(\theta^0)$$

- Pick the 2nd batch

$$C = L^2 + L^{16} + \cdots$$

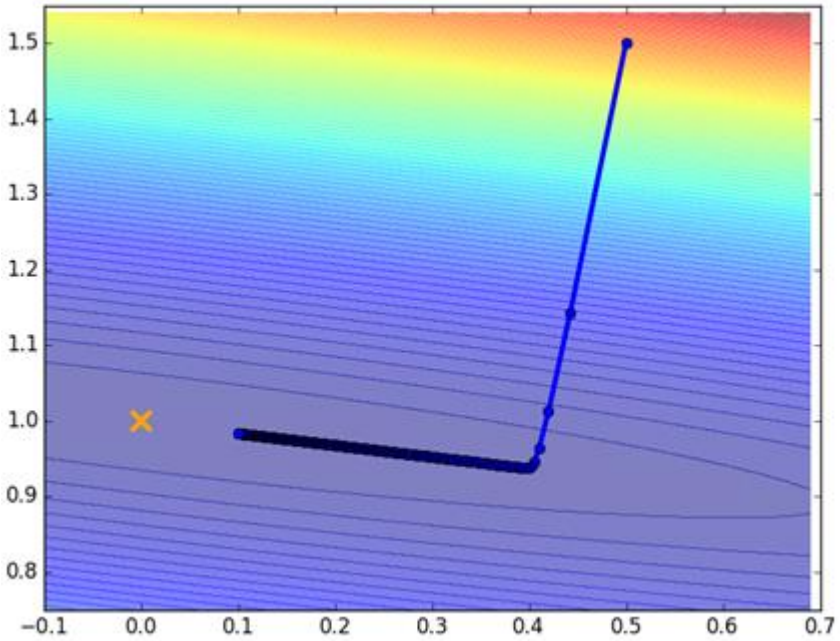
$$\theta^2 \leftarrow \theta^1 - \eta \nabla C(\theta^1)$$

⋮

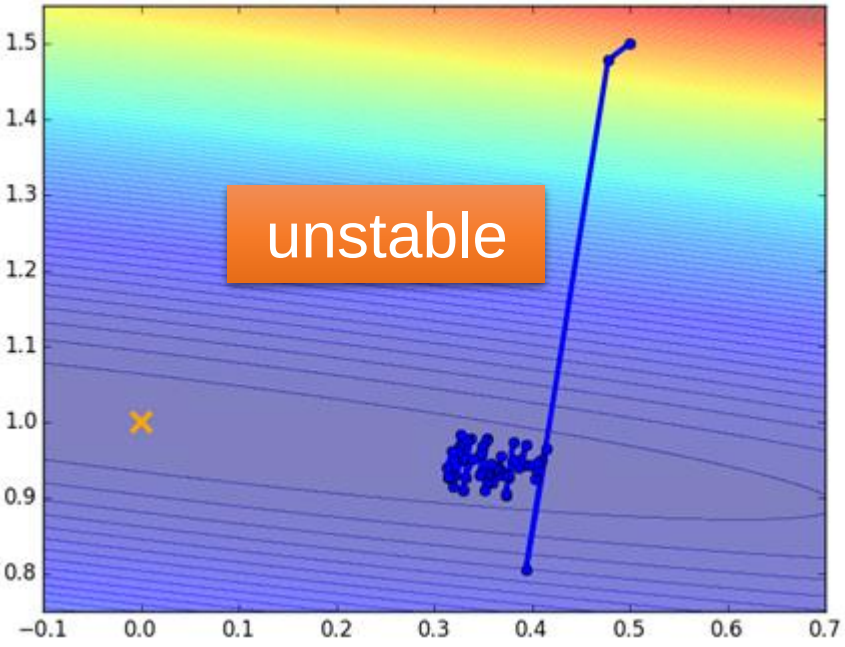
C is different each time when we update parameters!

Mini-batch

Original Gradient Descent



With Mini-batch

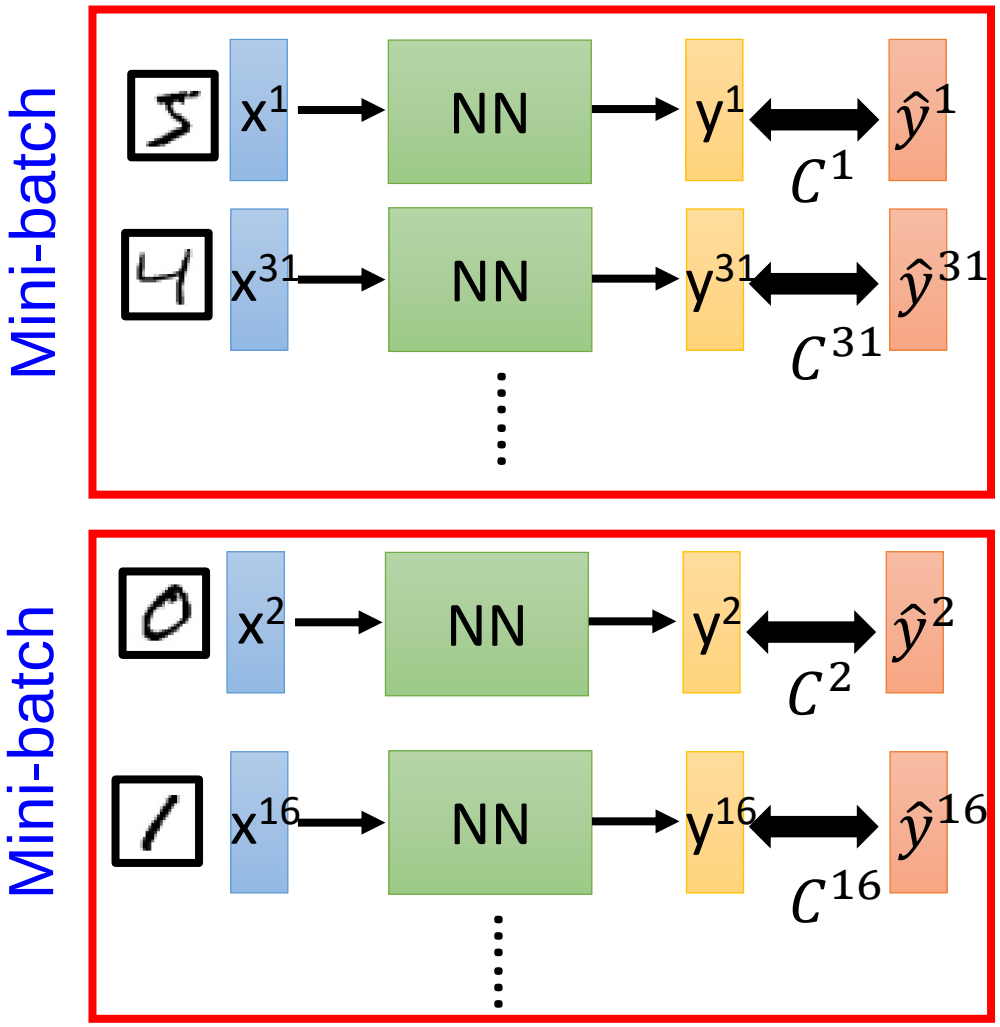


The colors represent the total C on all training data.

Mini-batch

Faster

Better!



- Randomly initialize θ^0
- Pick the 1st batch

$$C = C^1 + C^{31} + \cdots$$

$$\theta^1 \leftarrow \theta^0 - \eta \nabla C(\theta^0)$$

- Pick the 2nd batch

$$C = C^2 + C^{16} + \cdots$$

$$\theta^2 \leftarrow \theta^1 - \eta \nabla C(\theta^1)$$

$\vdots$

- Until all mini-batches have been picked

one epoch

Repeat the above process

Backpropagation

- A network can have millions of parameters.
  - Backpropagation is the way to compute the gradients efficiently (not today)
  - http://speech.ee.ntu.edu.tw/~tlkagk/courses/MLDS_2015_2/Lecture/DNN%20backprop.ecm.mp4/index.html
- Many toolkits can compute the gradients automatically

theano

TensorFlow

http://speech.ee.ntu.edu.tw/~tlkagk/courses/MLDS_2015_2/Lecture/Theano%20DNN.ecm.mp4/index.html

Conclusion

Three Steps for Deep Learning

- Define a set of function (architecture)
- Goodness of function (loss)
- Pick the best function (optimization)

When can we use deep learning to solve a problem?

- First, you want find a function
- Second, you have lots of function input/output as training data

Diverse Architectures

- Convolutional Neural Network (CNN)
- Residual Neural Network (ResNet)
- Recurrent Neural Network (RNN)
- Transformer
- Graph Neural Network (GNN)

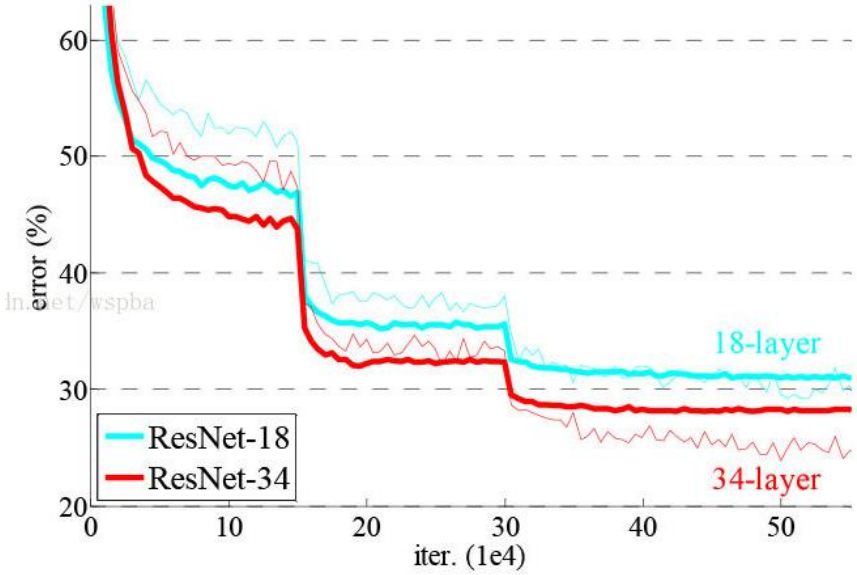
……

- Auto-encoder
- Generative Adversarial Network (GAN)
- Contrastive learning

……

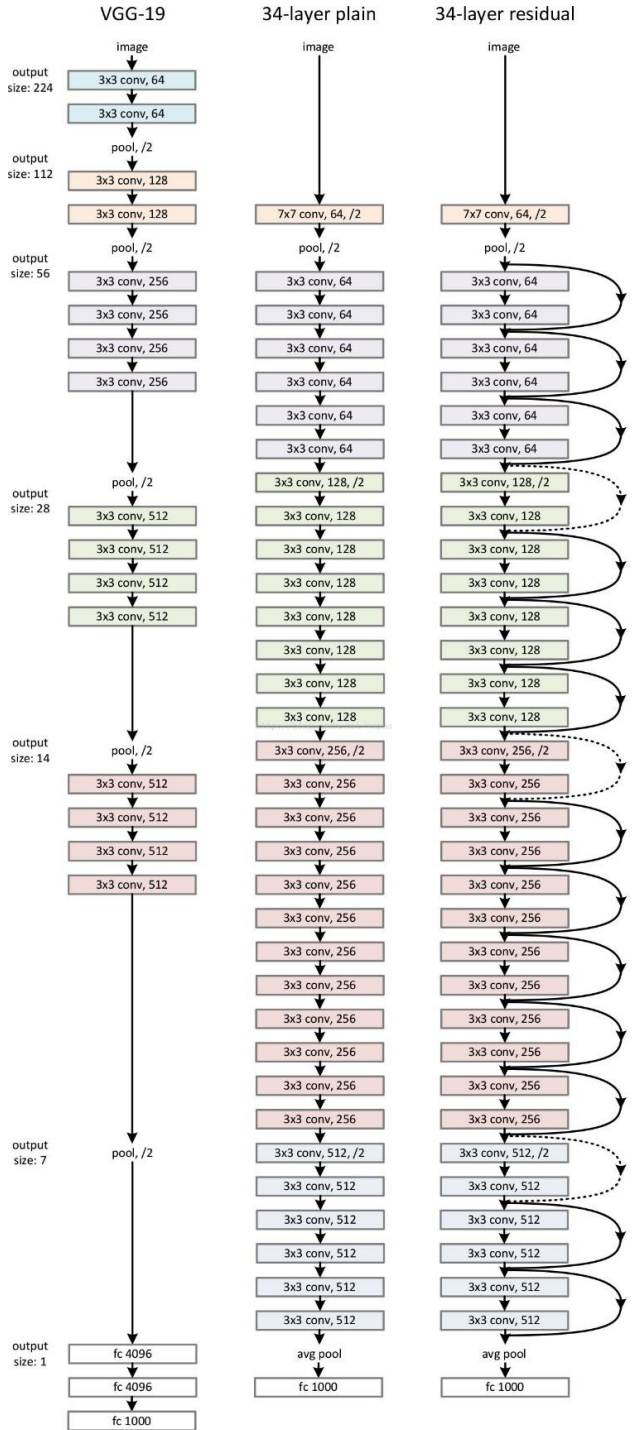
Large Model Change Space

Deep Neural Networks



Deeper is Better?

Not surprised, more parameters, better performance



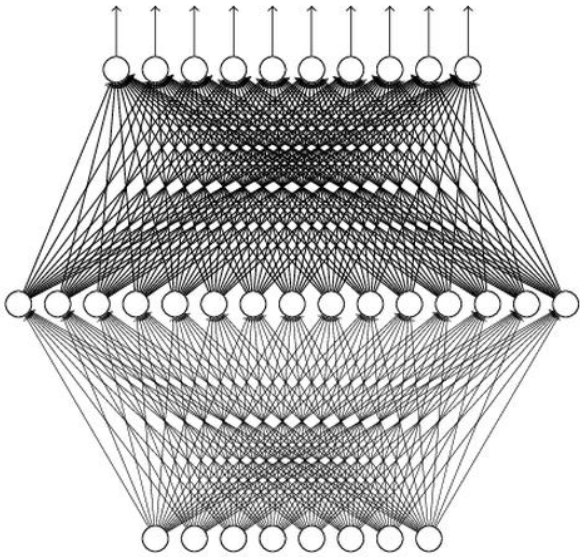
Universality Approximation Theorem

Any continuous function f

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

Can be realized by a network with one hidden layer (given **enough** hidden neurons)

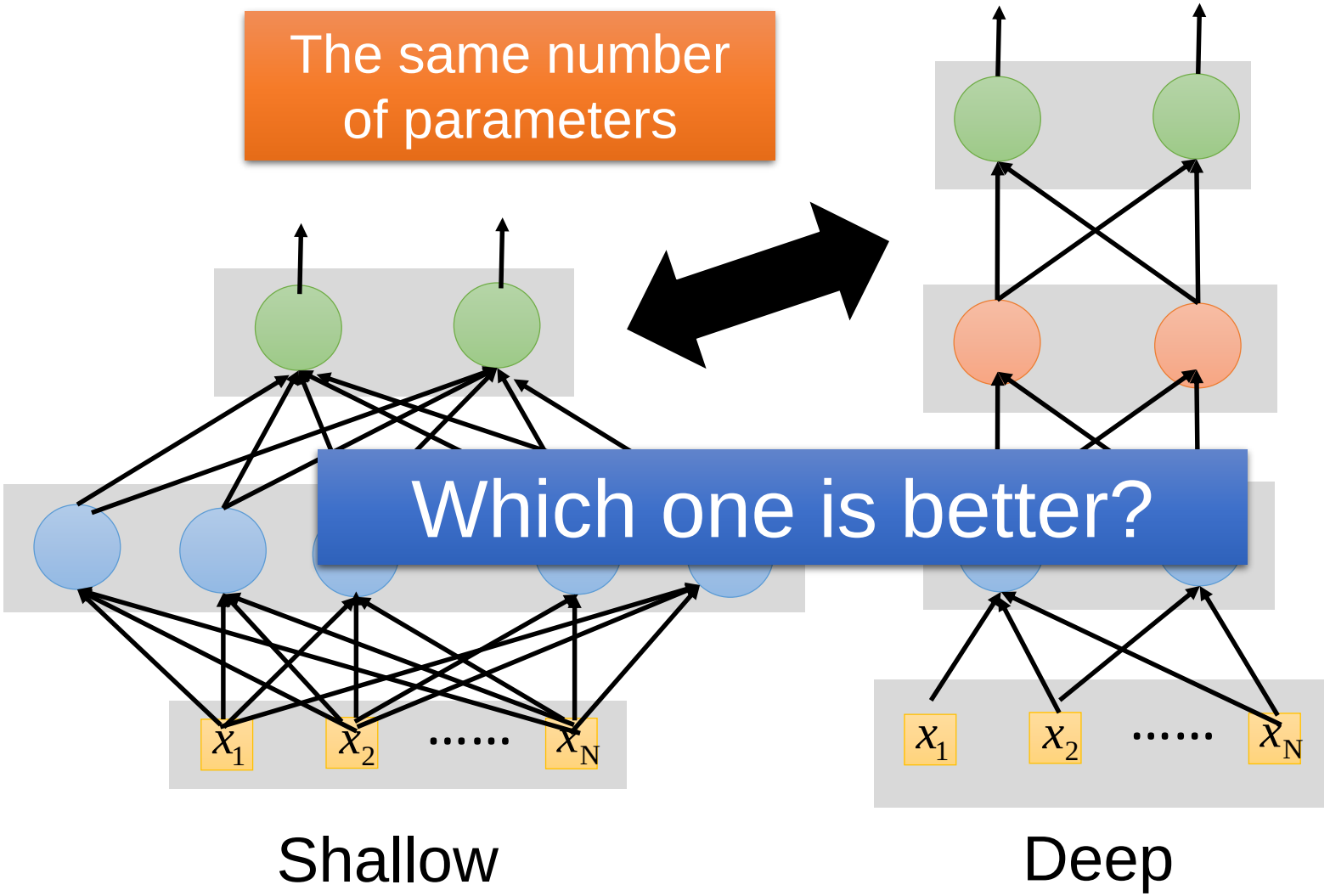
Neural Networks, 1989, 2(5):359-366.



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

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

Fat + Short v.s. Thin + Tall



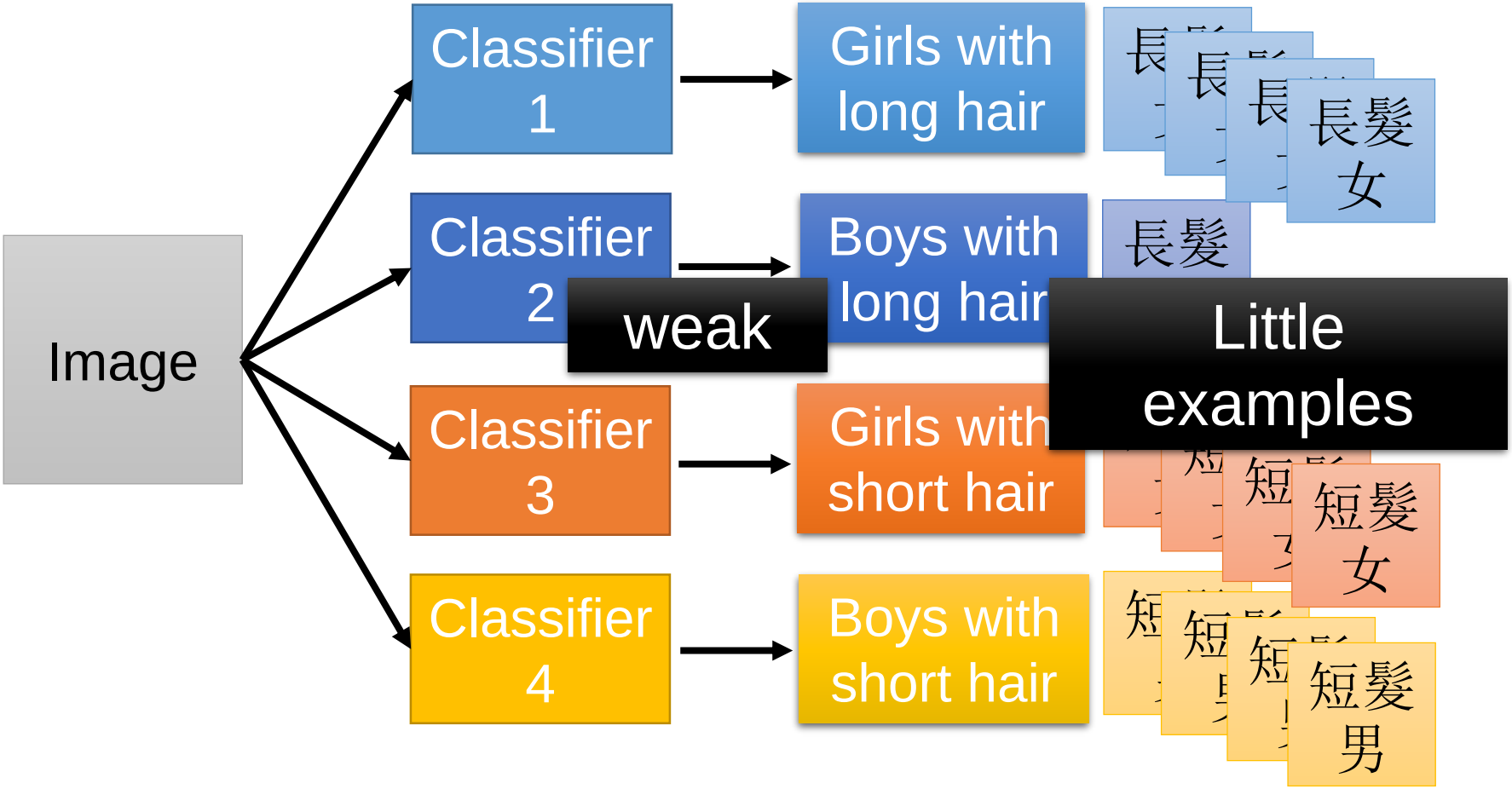
Fat + Short v.s. Thin + Tall

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.

Why Deep?

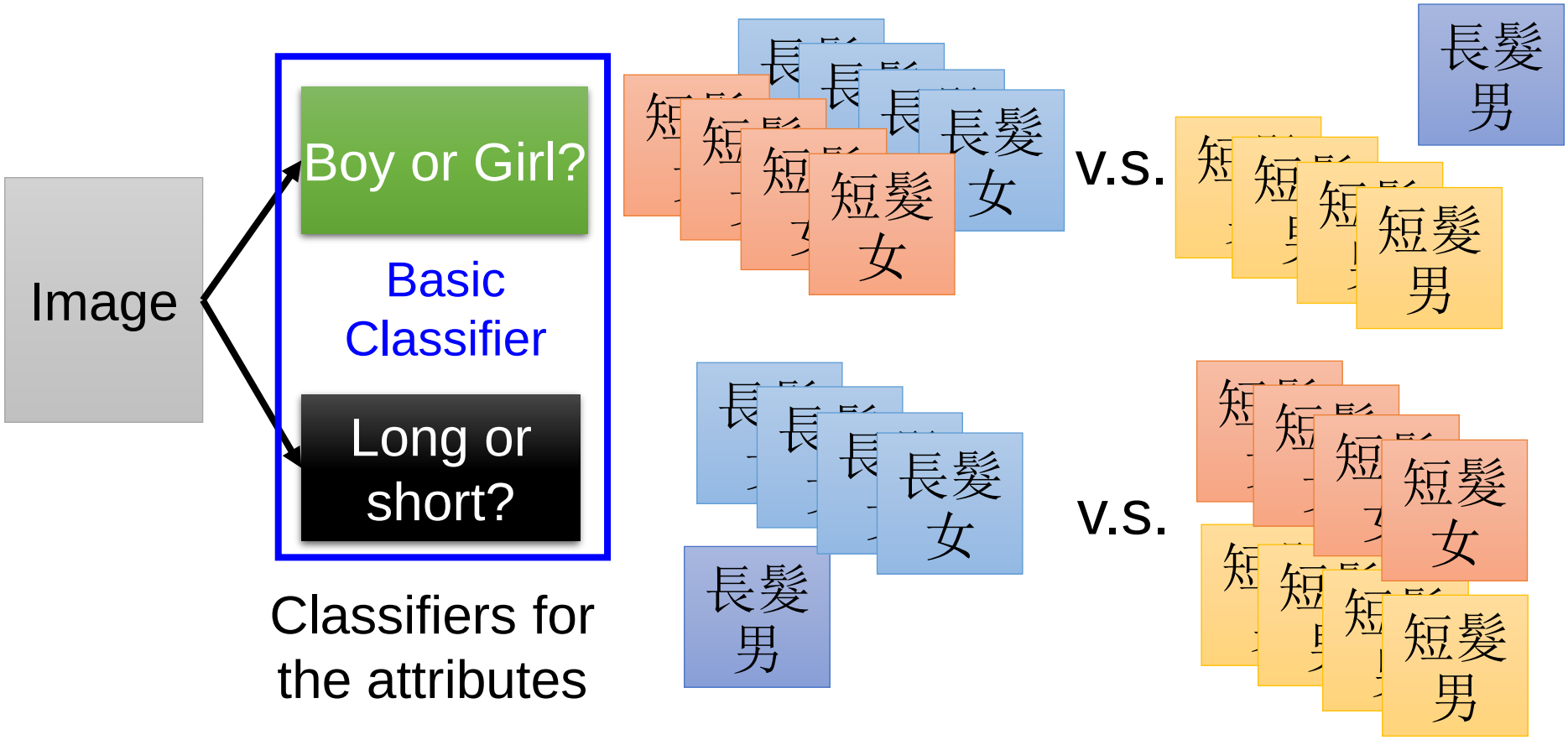
- Deep → Modularization



Why Deep?

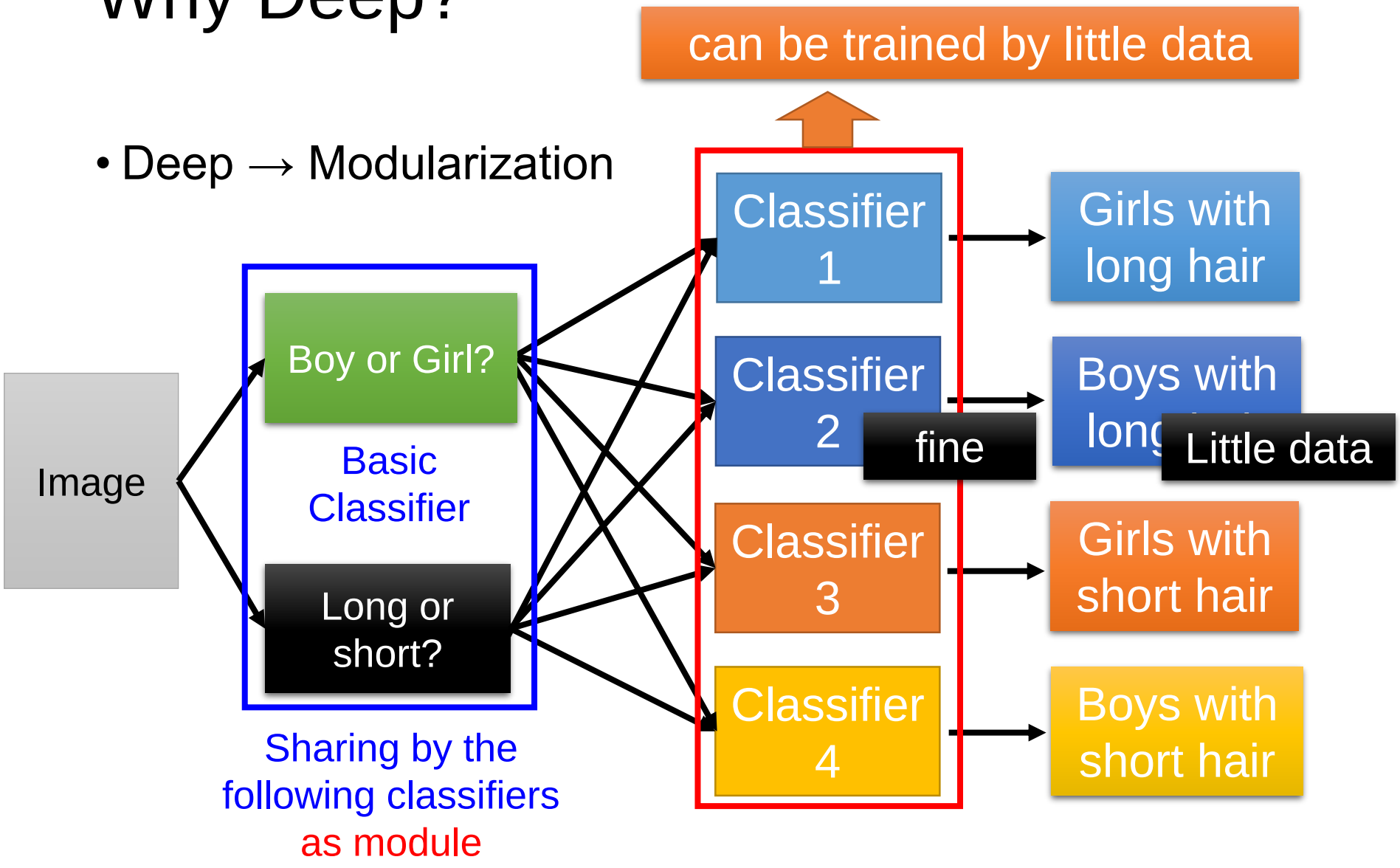
Each basic classifier can have sufficient training examples.

- Deep → Modularization



Why Deep?

- Deep → Modularization

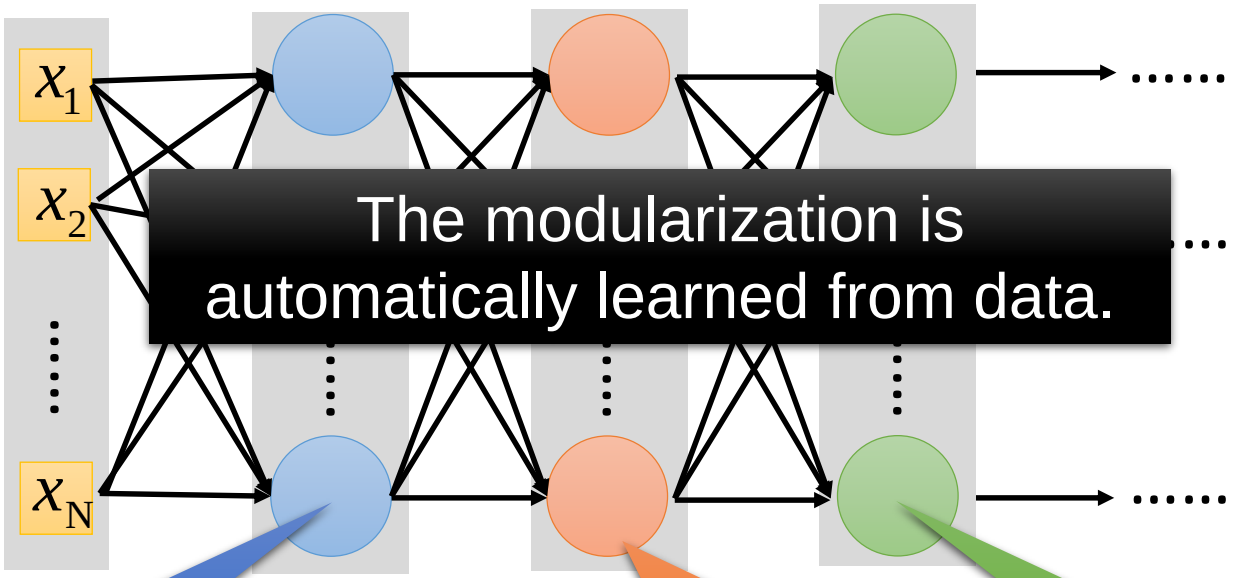


Why Deep?

Deep Learning also works on small data set like TIMIT.

- Deep → Modularization

→ Less training data?



The most basic classifiers

Use 1st layer as module to build classifiers

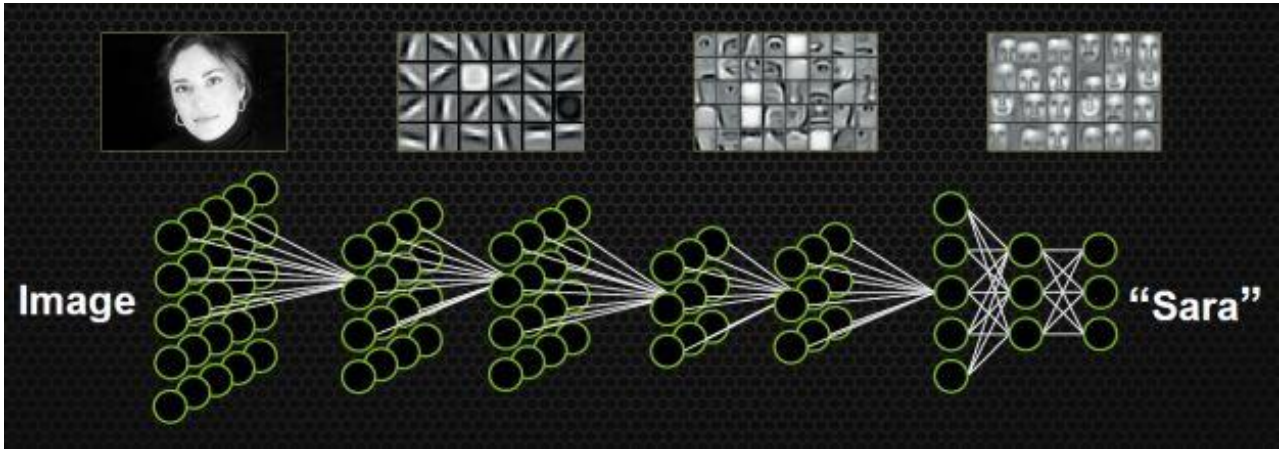
Use 2nd layer as module ……

Why Deep?

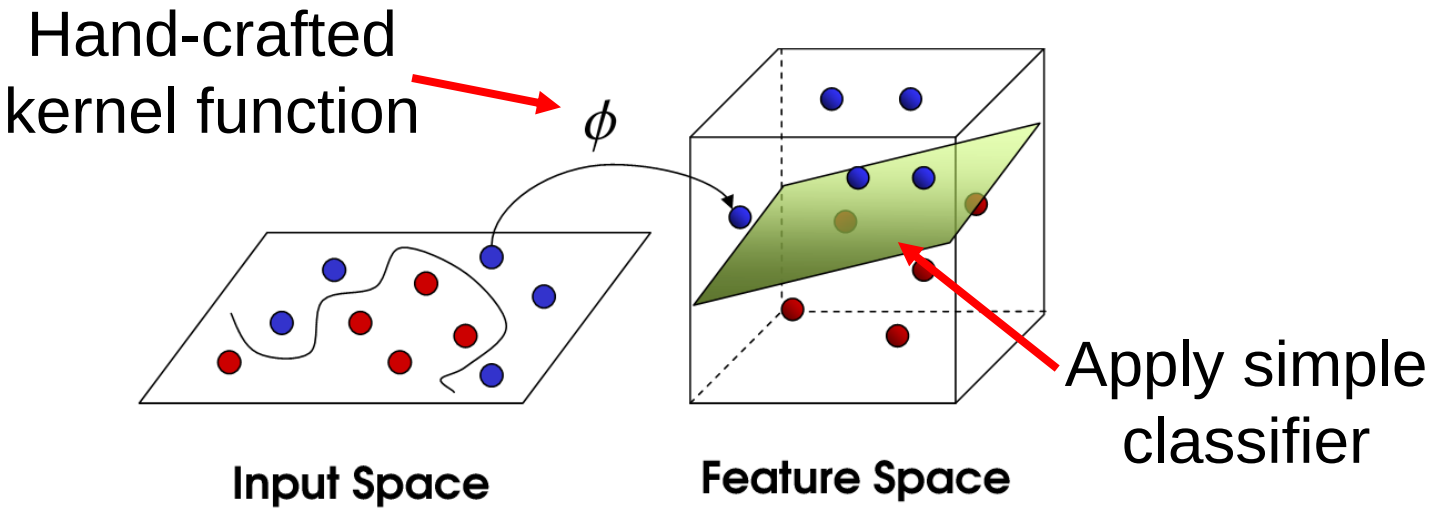
The deep neural network need less parameters to describe a function.

- That means that we need less data to determine the function.
- It can effectively reduce the risk of overfitting.

In the deep neural network, the deeper layers can pick up more complex features to help us complete the task.

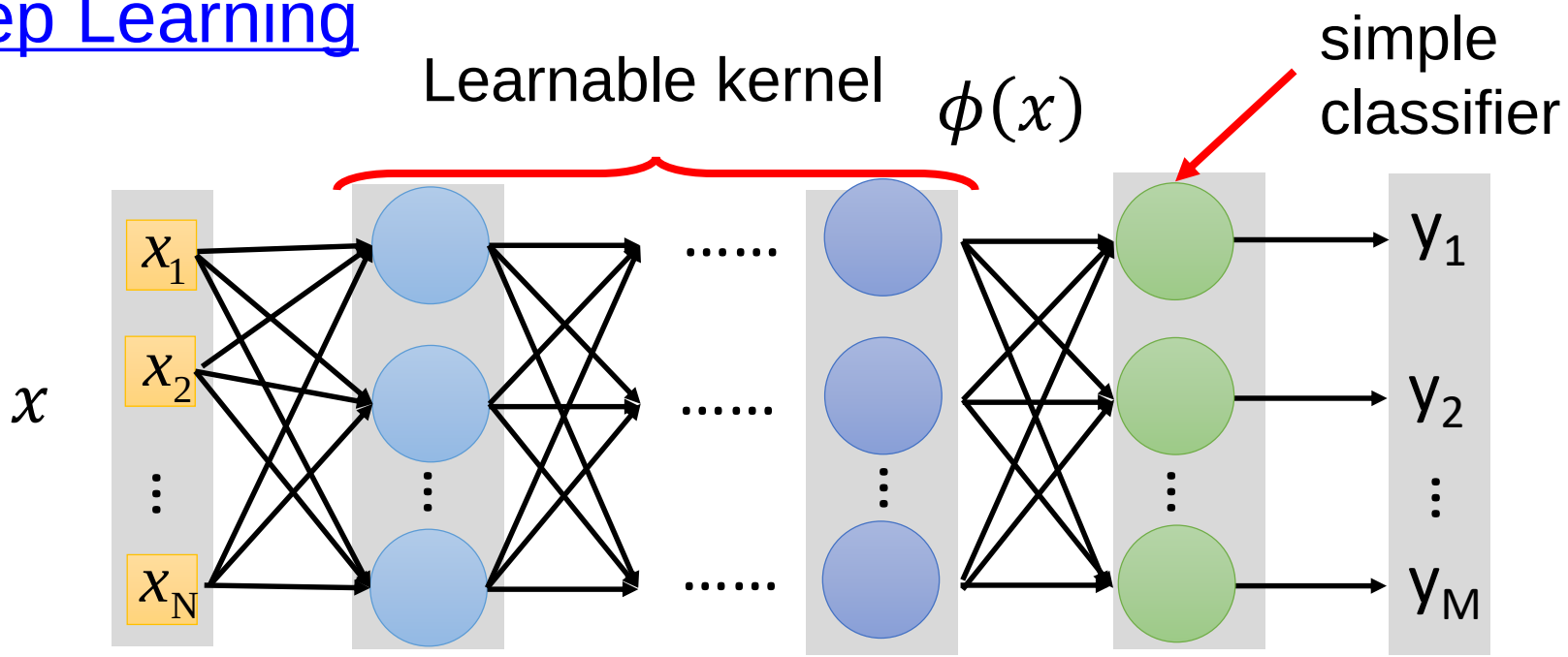


SVM

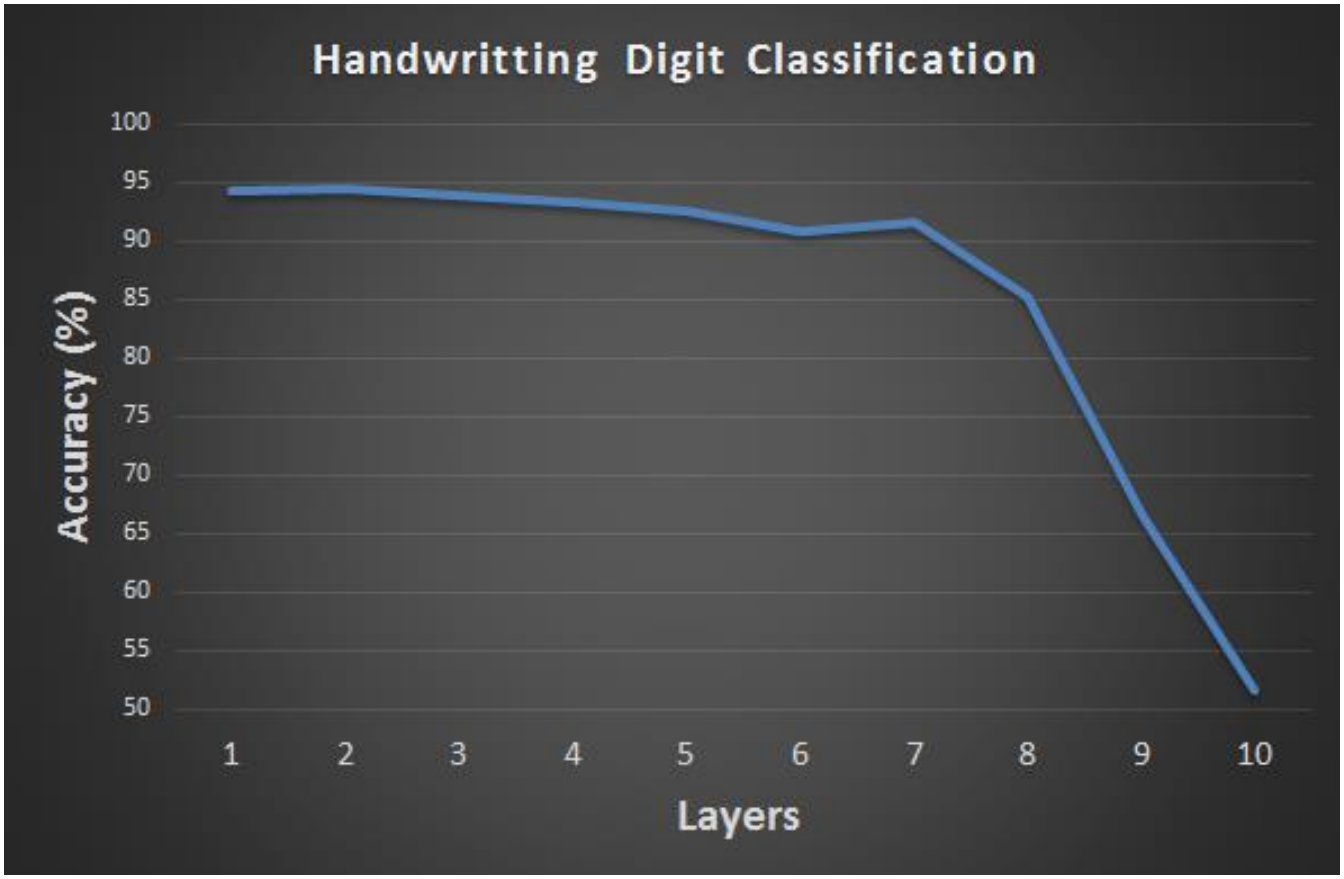


Source of image: http://www.gipsa-lab.grenoble-inp.fr/transfert/seminaire/455_Kadri2013Gipsa-lab.pdf

Deep Learning



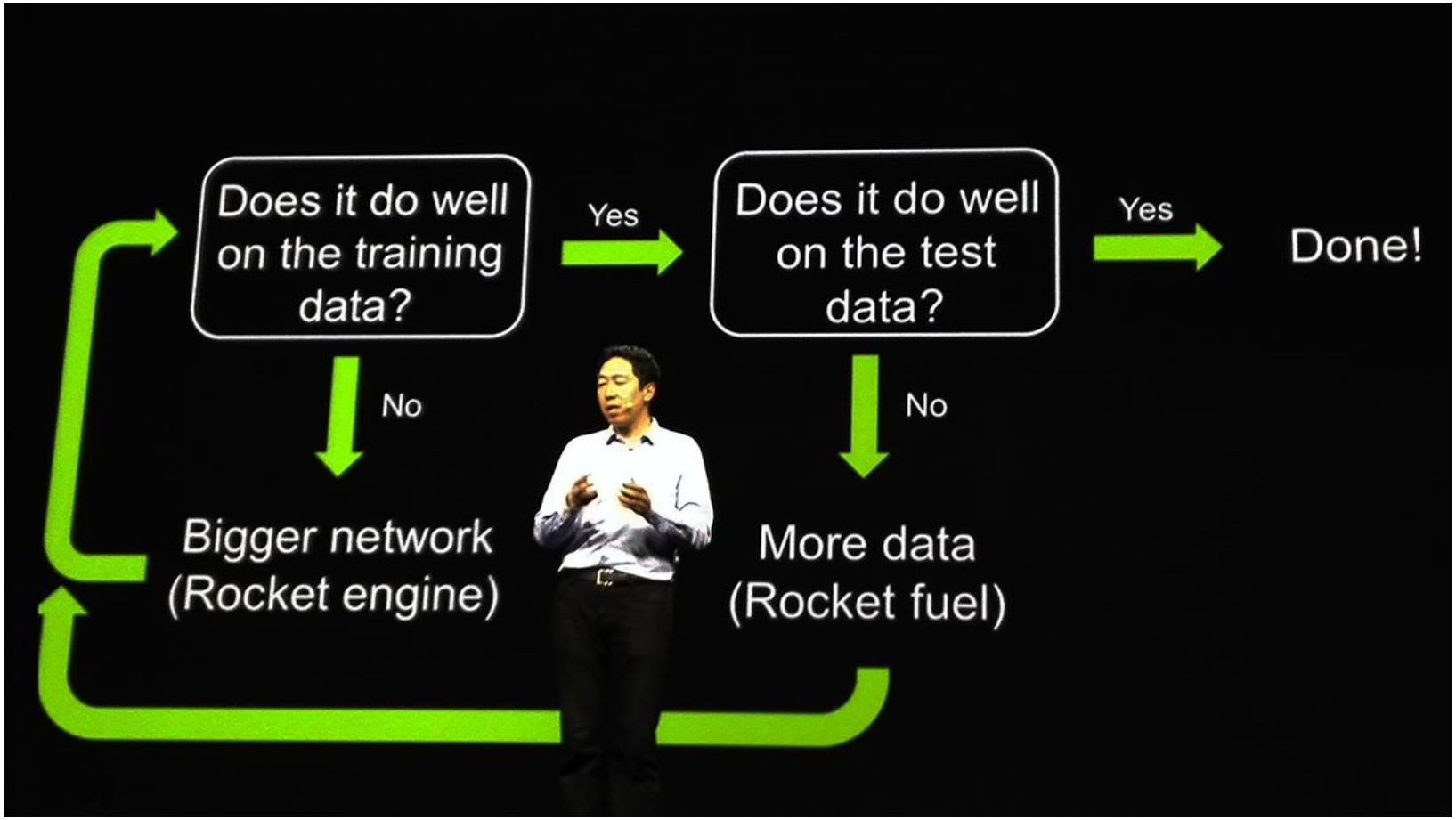
Hard to get the power of deep …



Before 2006, deeper usually does not imply better.

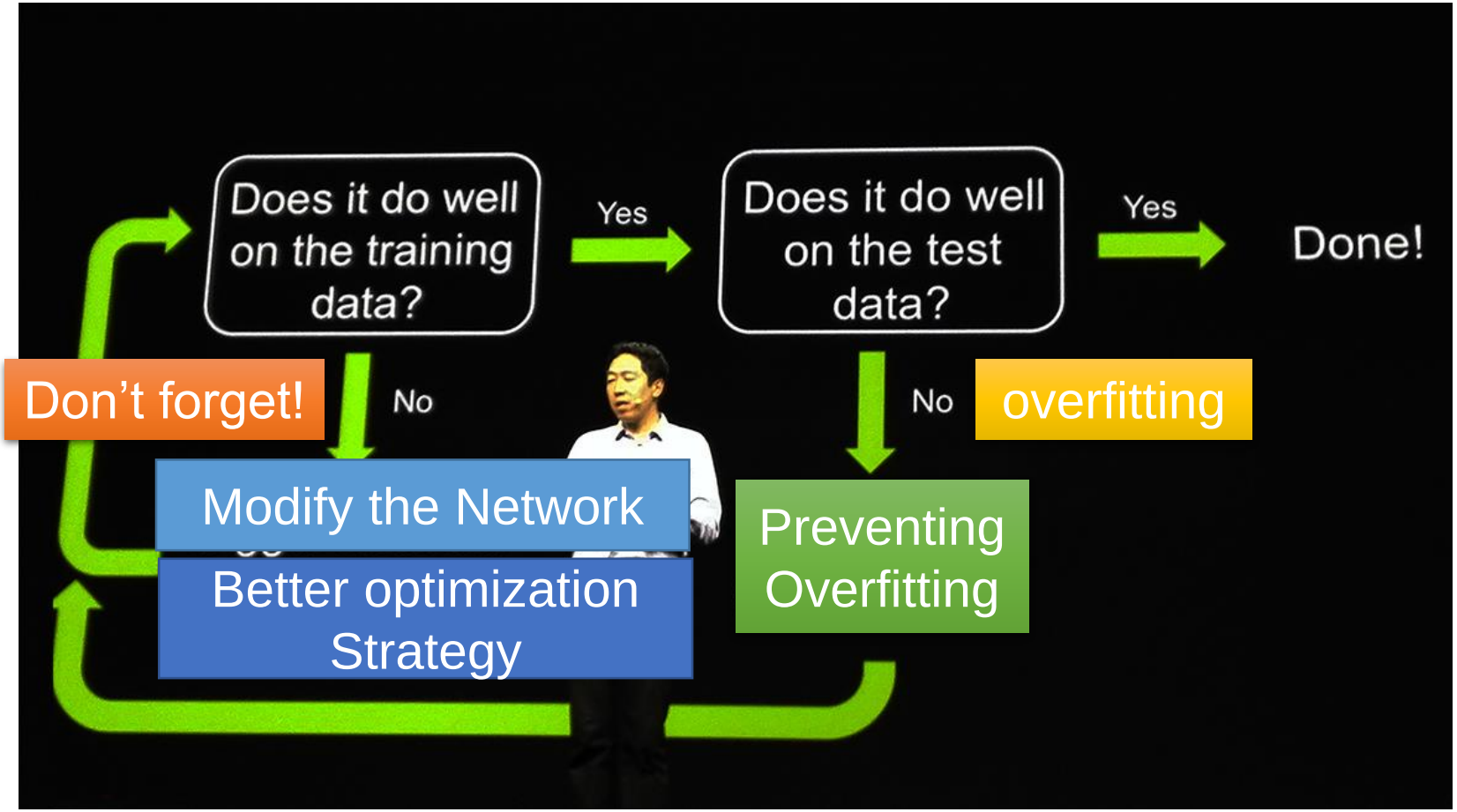
Recipe for Learning

Tips for Training DNN



http://www.gizmodo.com.au/2015/04/the-basic-recipe-for-machine-learning-explained-in-a-single-powerpoint-slide/

Recipe for Learning



http://www.gizmodo.com.au/2015/04/the-basic-recipe-for-machine-learning-explained-in-a-single-powerpoint-slide/

Recipe for Learning

Modify the Network

- New activation functions, for example, ReLU or Maxout

Better Optimization Strategy

- Loss function/Mini-batch/Optimizer (Adaptive learning rates)

Prevent Overfitting

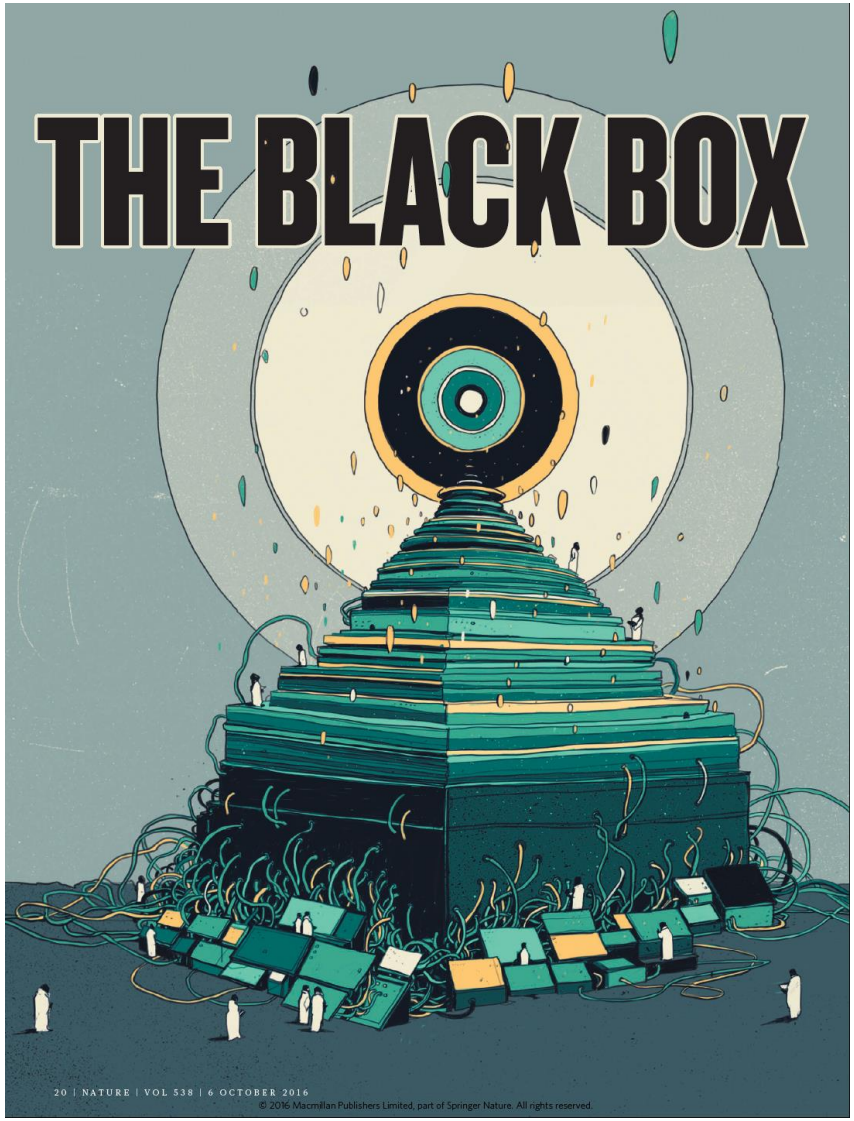
- Increasing training data/Early stopping Regularization/Dropout

We get a “perfect” deep learning model, which works well.

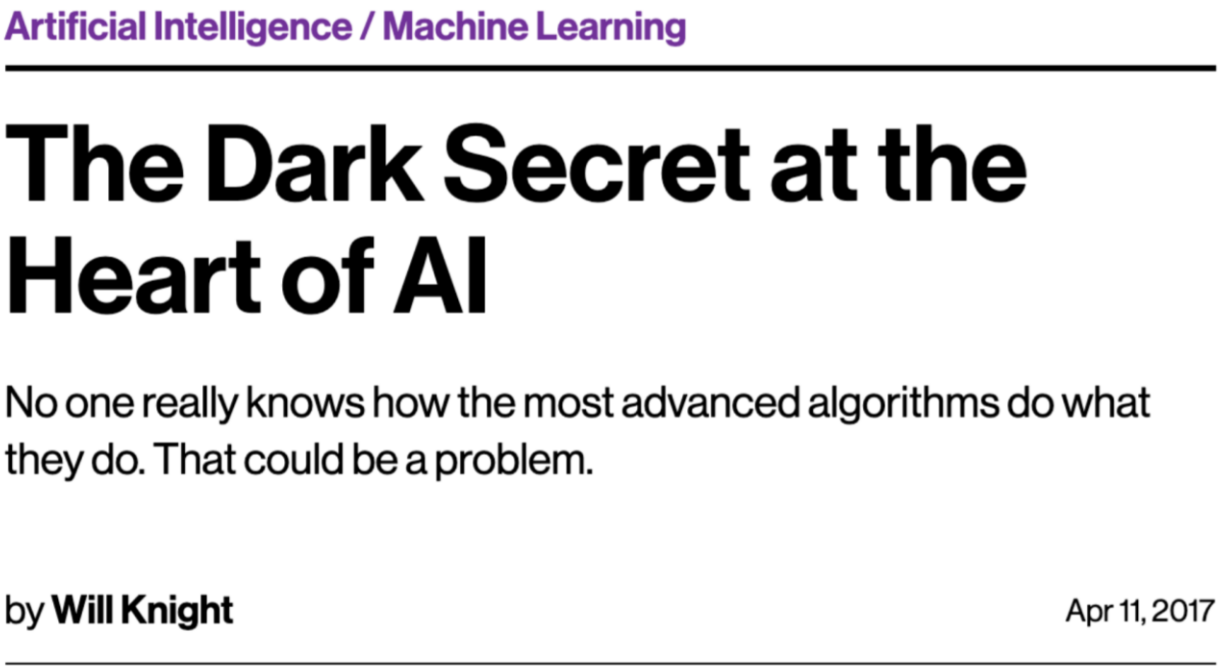
> *“The problem is that the knowledge gets baked into the network, rather than into us.”*

Nature 538, 20–23 (2016)

The Black Box of Deep Learning



Demystifying deep learning



Artificial Intelligence / Machine Learning

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

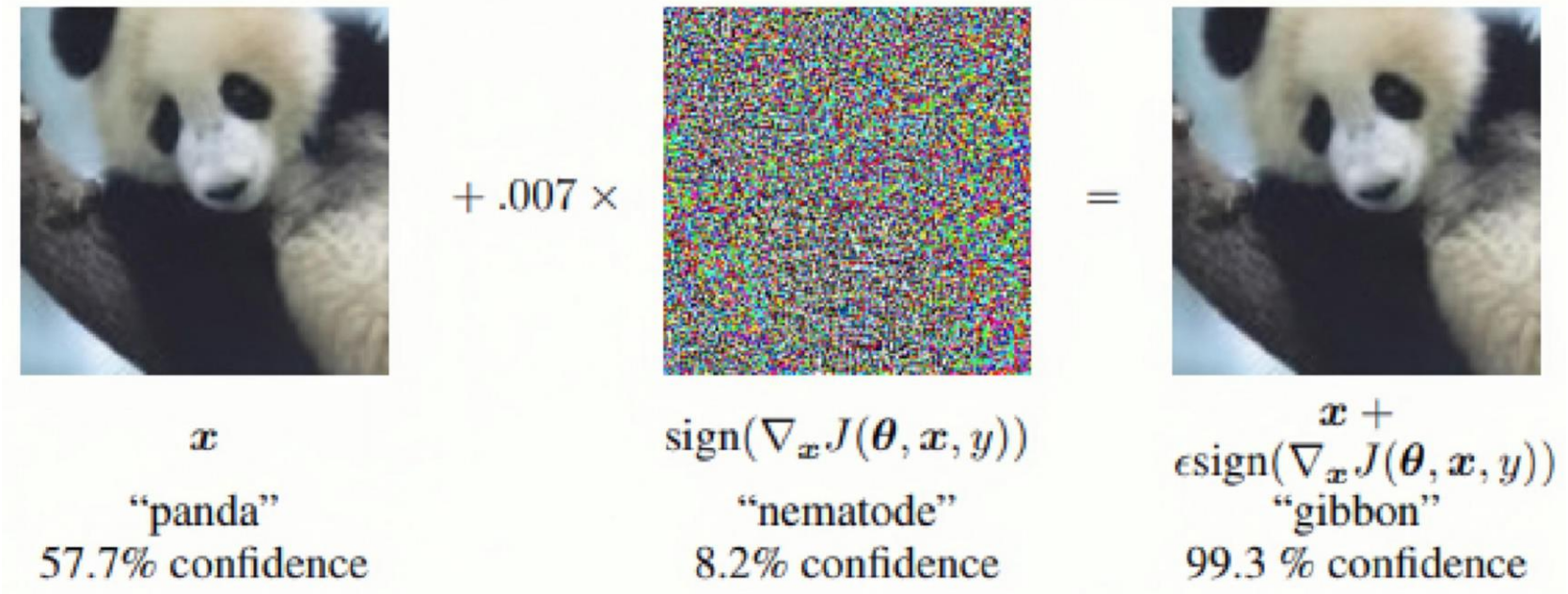
by **Will Knight** Apr 11, 2017

We can build these models, but we don't know how they work.

How well can we get along with machines that are unpredictable and inscrutable?

https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/

Easy to Be Attacked



Preprint arXiv:1312.6199, 2013

Larger and Larger DNN Models

- The number of parameters of DNN is bigger and bigger.
- For example, the number of parameters of Switch Transformer has reached the trillion level. There are some problems when training so huge network.
- Requirement of a lot of annotation data
- Difficulty of model training

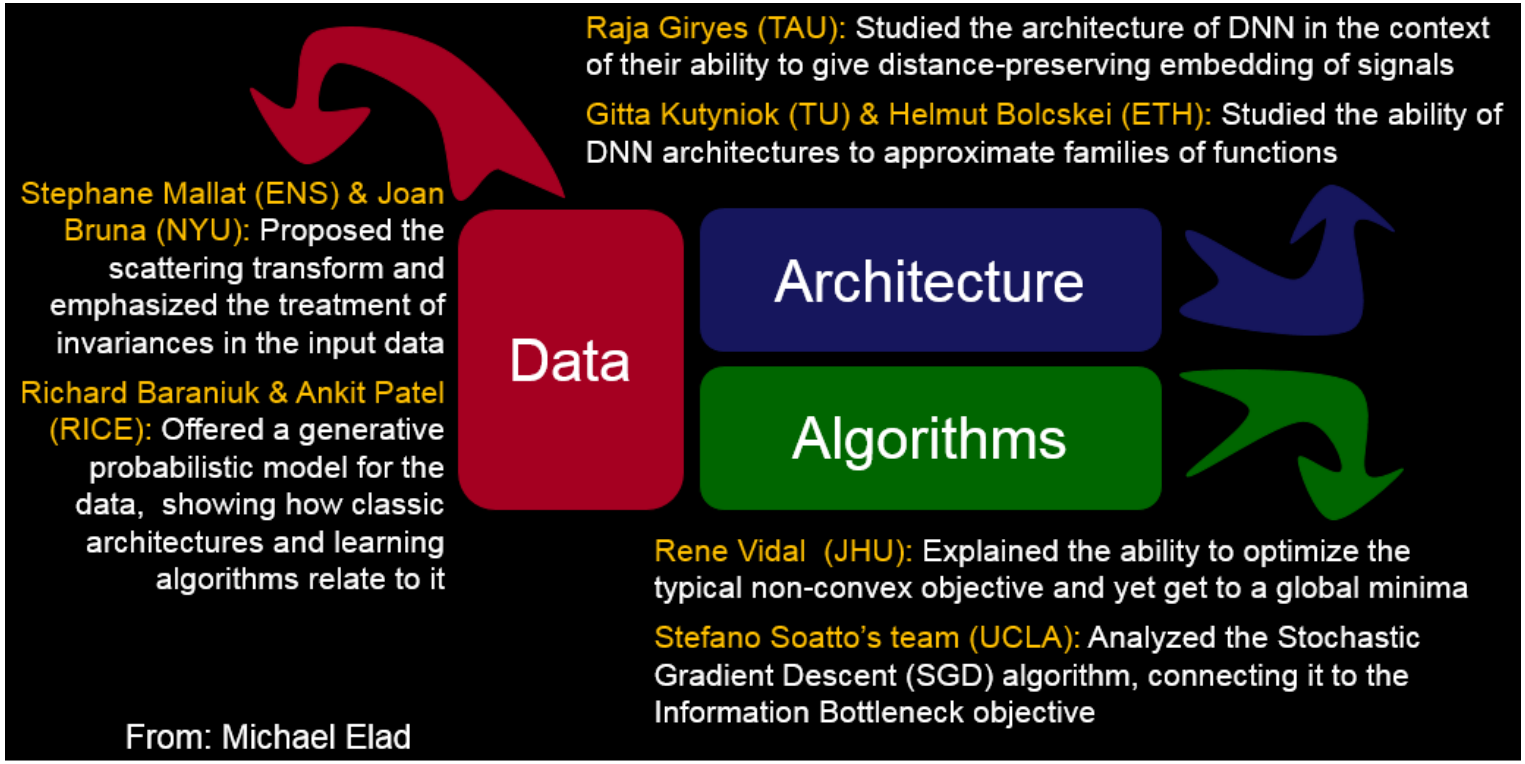
Preprint arXiv:2101.03961 (2021)

Deep Learning Theory

- **Representation:** Why are deep neural networks better than shallow ones?
- **Optimization:** Why is SGD (Stochastic Gradient Descent) good at finding minima and what are good minima?
- **Generalization:** Why is it that we don’t have to worry about overfitting despite overparameterization?

Tomaso Poggio on Deep Learning Representation, Optimization, and Generalization

Deep Learning Theory



深度学习理论2021

Thanks

zsh@amss.ac.cn

该二维码7天内(10月20日前)有效，重新进入将更新