DEEP LEARNING

A GENTLE INTRODUCTION

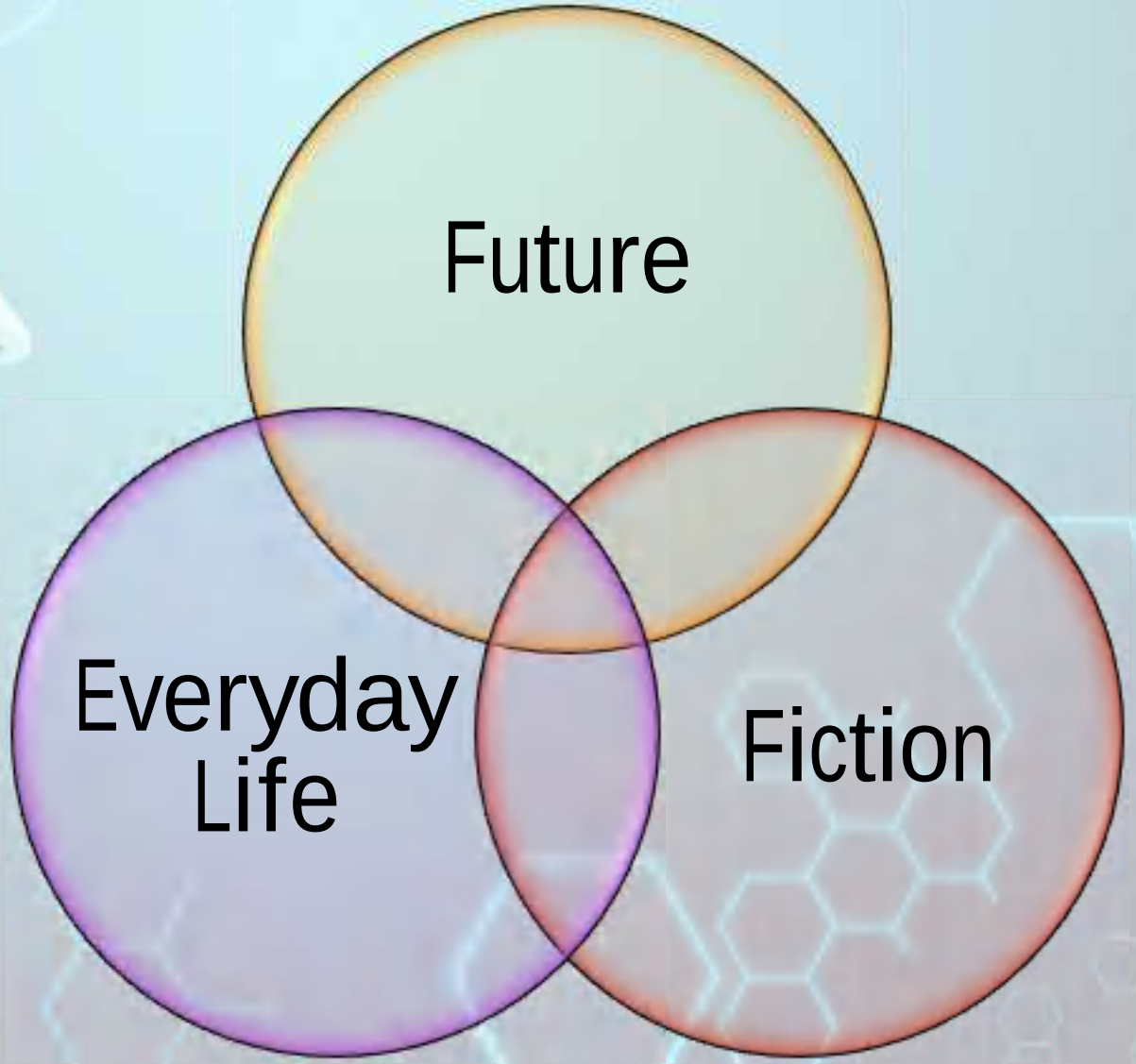
ZHANXING ZHU (朱占星)

ZHANXING.ZHU@PKU.EDU.CN

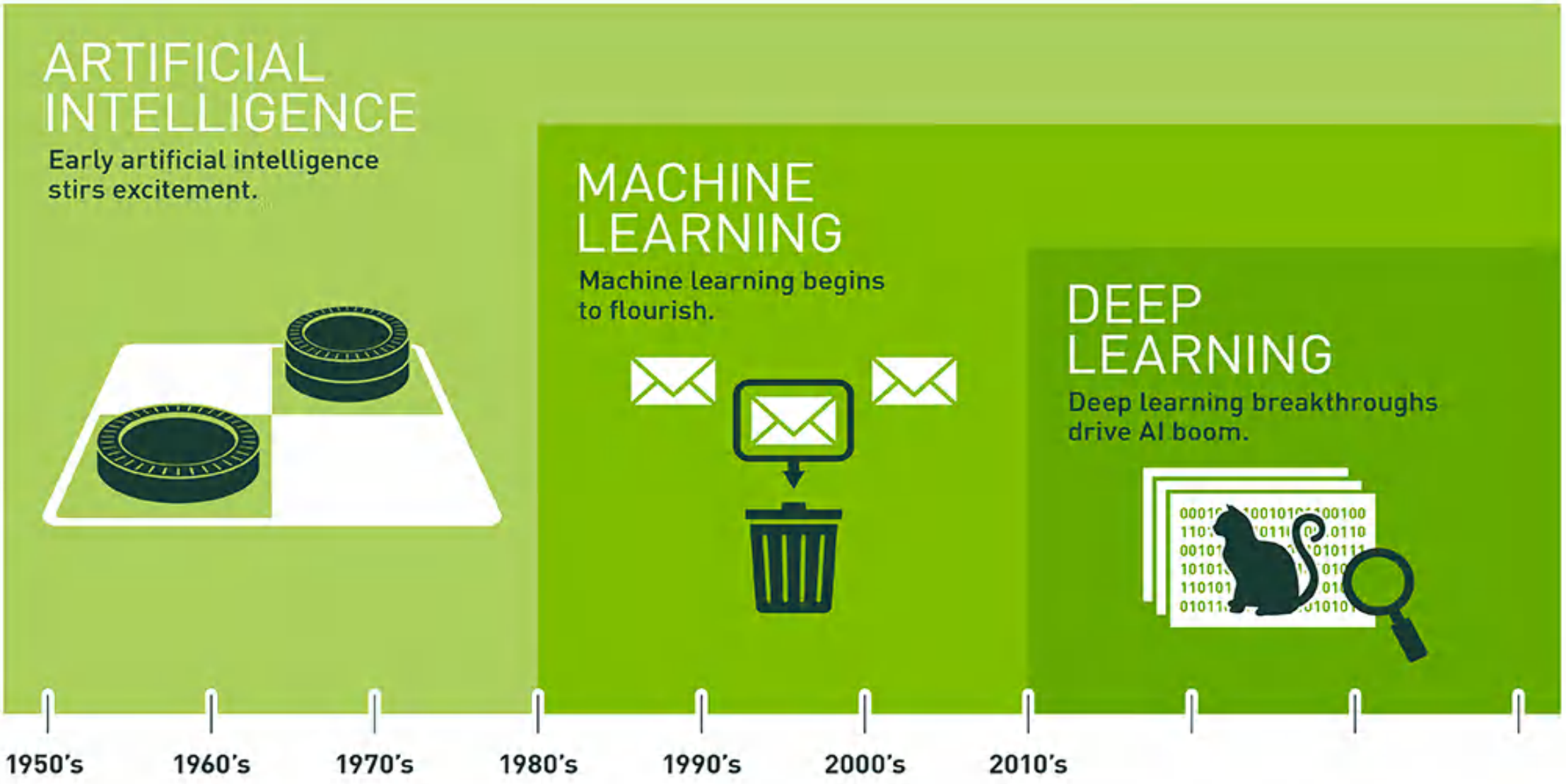
PEKING UNIVERSITY (北京大学)

BIBDR (北京大数据研究院)

ARTIFICIAL INTELLIGENCE (AI)



AI HAS EXPLODED SINCE 2015



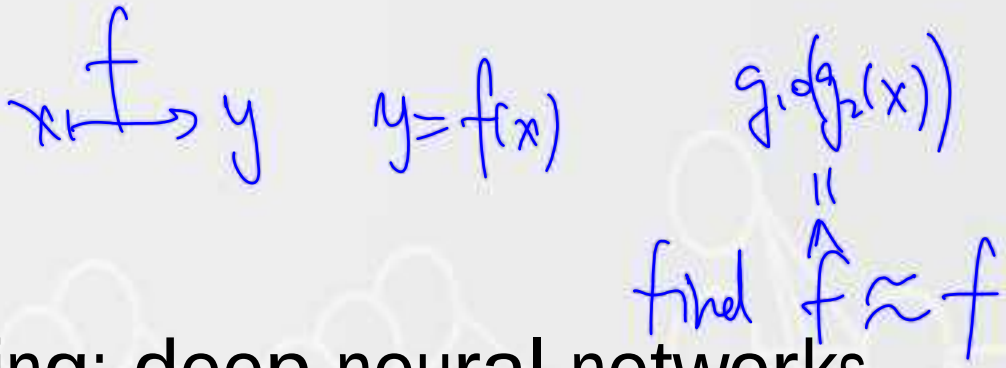
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

MACHINE LEARNING

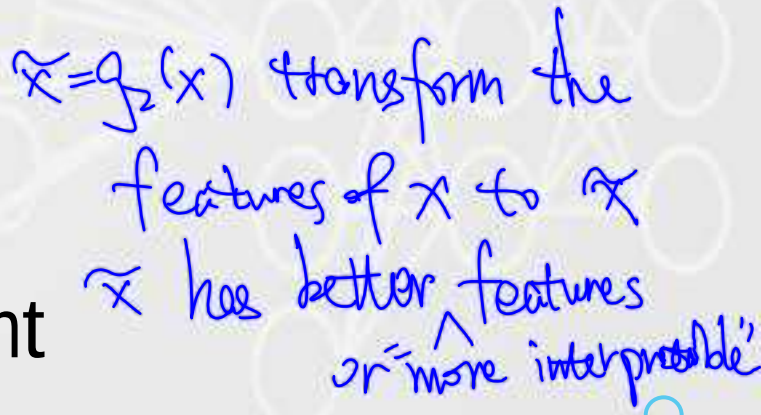
- Models learn from data
- To make predictions or learn patterns
- Models are ``trained'' by data and algorithms to obtain ability to perform the specific tasks
- Applications: computer vision, natural language processing, robot, financial markets, healthcare, etc.

Output = Model (input; parameters)

DEEP LEARNING

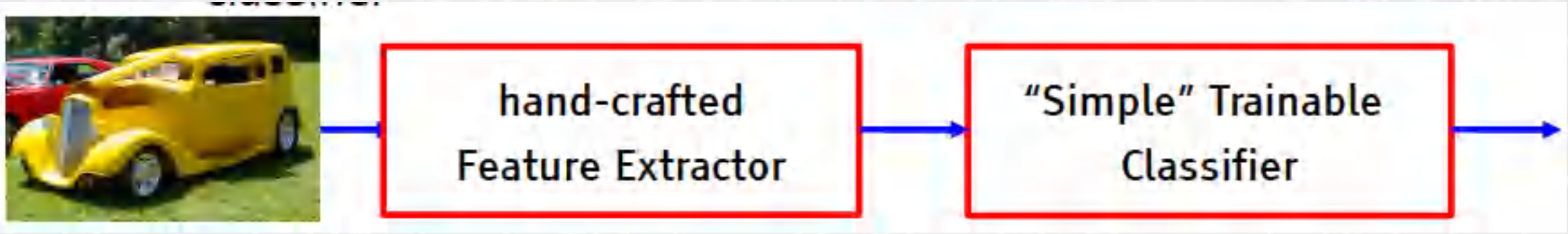


- A framework for machine learning: deep neural networks
- Feature hierarchy / Representation learning
- Function approximator
- End-to-end: raw inputs ⟶ what you want
- High-dimensional parameters, > 100,000, even billions
- Eating big data to work

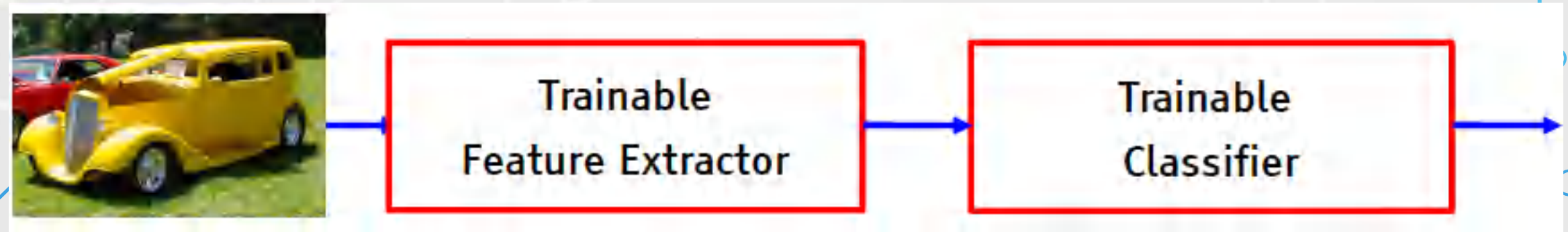


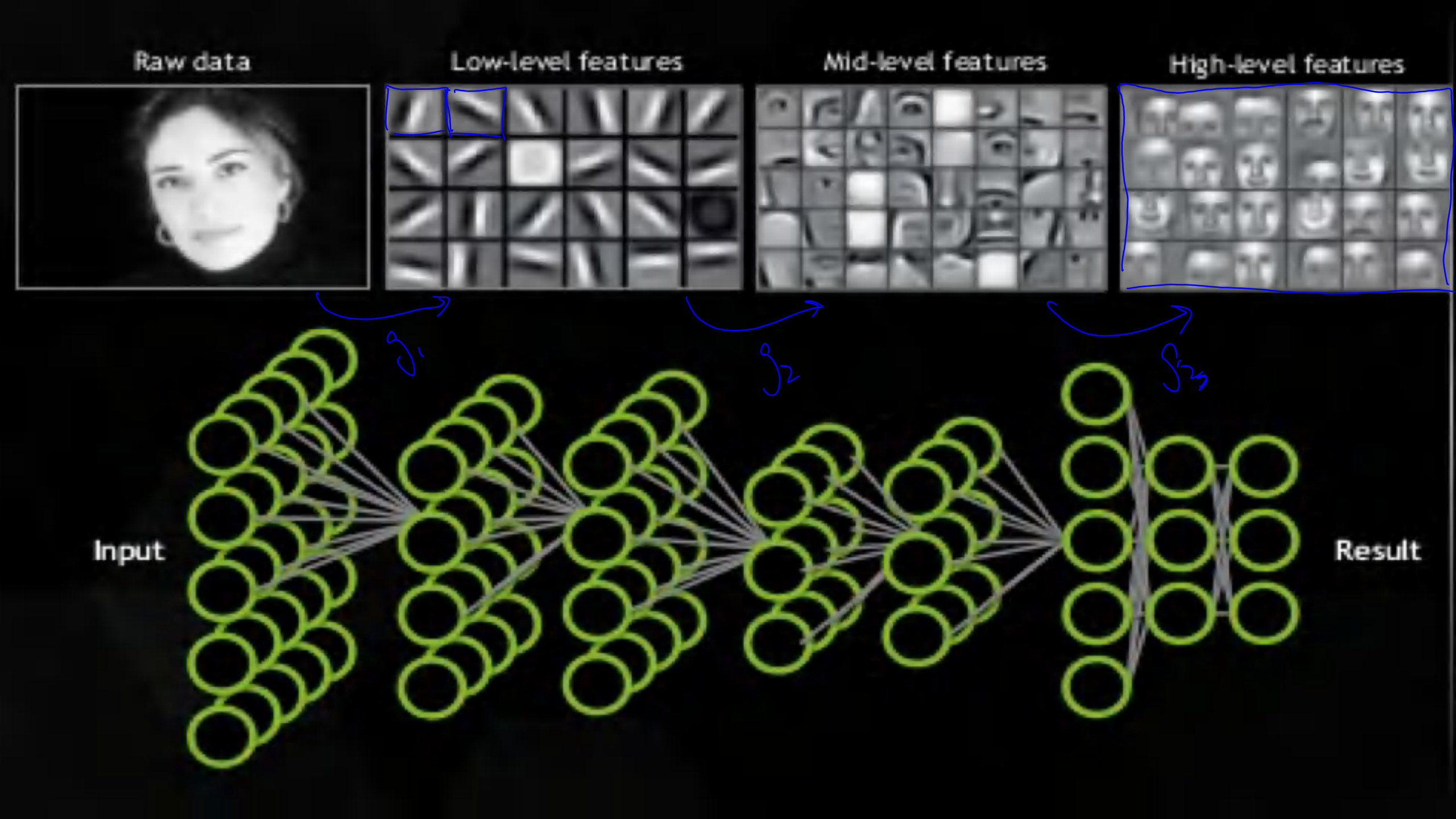
DEEP LEARNING = LEARNING REPRESENTATIONS

- Traditional pattern recognition methods



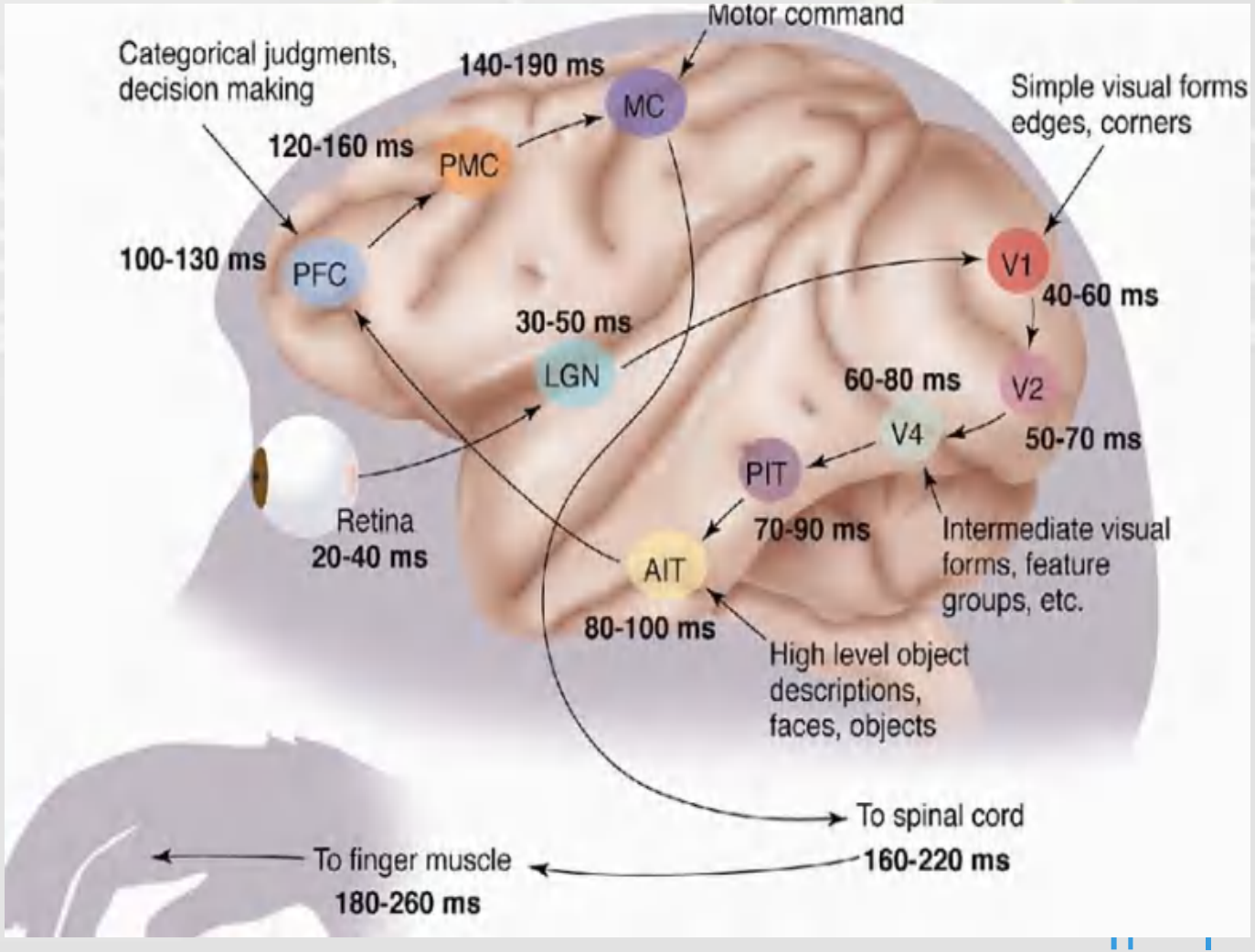
- End-to-end/Deep learning




Raw data
Low-level features
Mid-level features
High-level features
Input
Result

VISUAL CORTEX: HIERARCHY MATTERS

- Retina - LGN - V1 - V2 – V4 - PIT – AIT
- Many intermediate representations



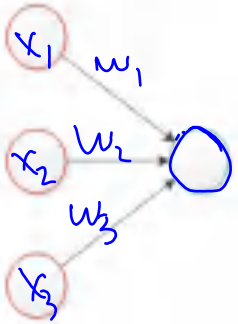
Rosenblatt

Deep neural network

Two-layer neural network

Single-layer neural network

Neuron



1943 MP 诞生

1949 Hebb

1958 Perceptron 第一次兴起

1969 "AI Winter"

1982 Hopfield

1986 BP (Hinton) 第二次兴起

1995 SVM (Vapnik)

2006 DBN (Hinton)

2012 CNN 第三次兴起

2020 Winter?

1940 1950 1960 1970 1980 1990 2000 2010 2020

COMPANIES & PEOPLE

Taken in NIPS2014, from left: Yann LeCun (Facbook, NYU), Geoffrey Hinton (Google, U of Toronto), Yoshua Bengio (U of Montreal), Andrew Ng (Baidu)

Google

facebook.

Google DeepMind

Microsoft

Baidu 百度

Alibaba Group 阿里巴巴集团

Disney Research

2σ TWO SIGMA

amazon.com

NVIDIA

DiDi More than a journey

BIBDR 北京大数据研究院

CONTENTS

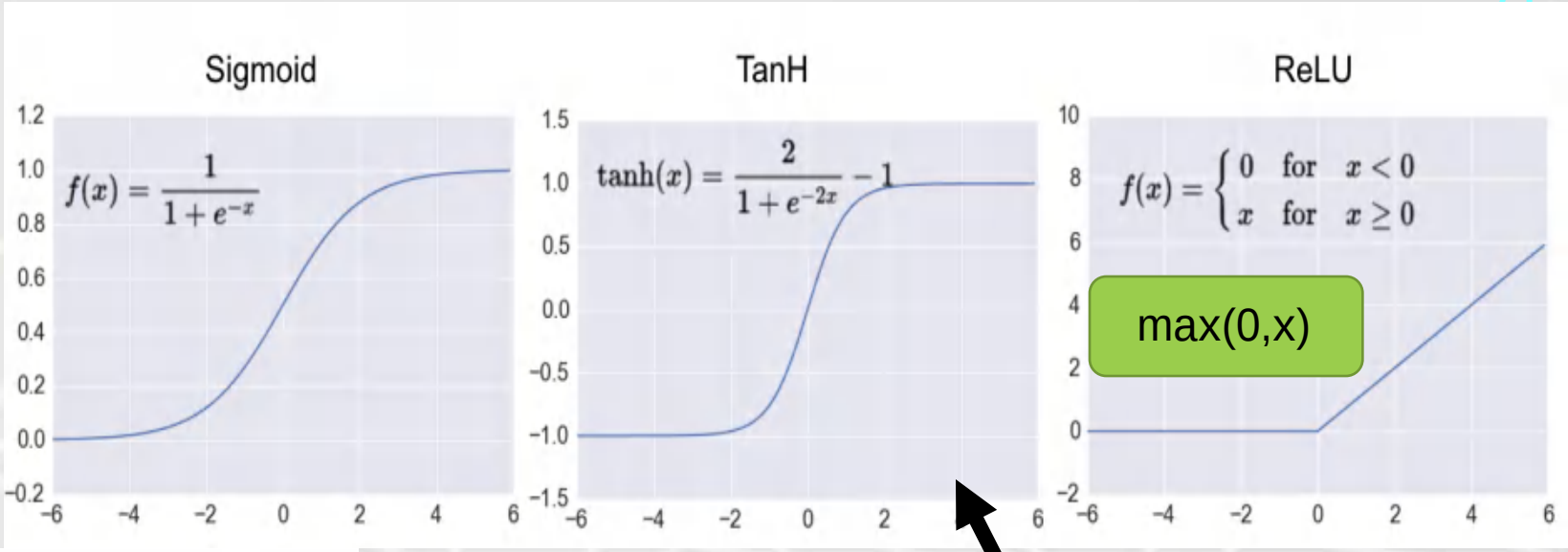
- Architecture: different types of network
- Several milestones
- Some issues
- Future

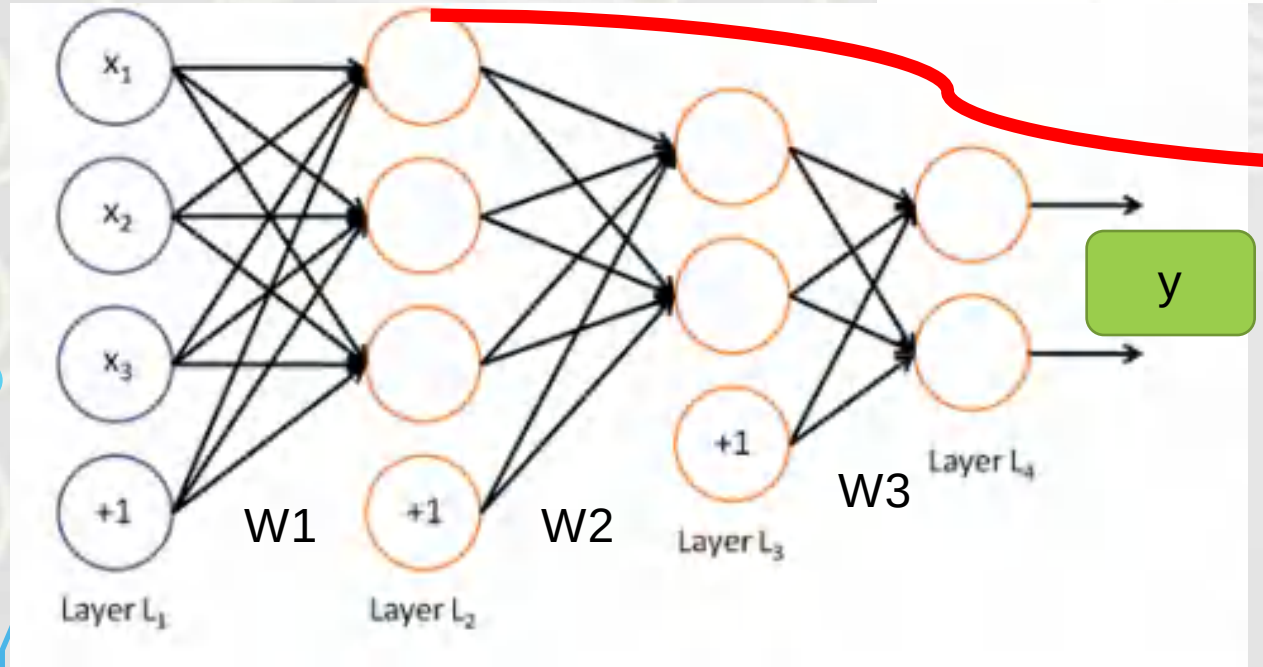
ARCHITECTURES

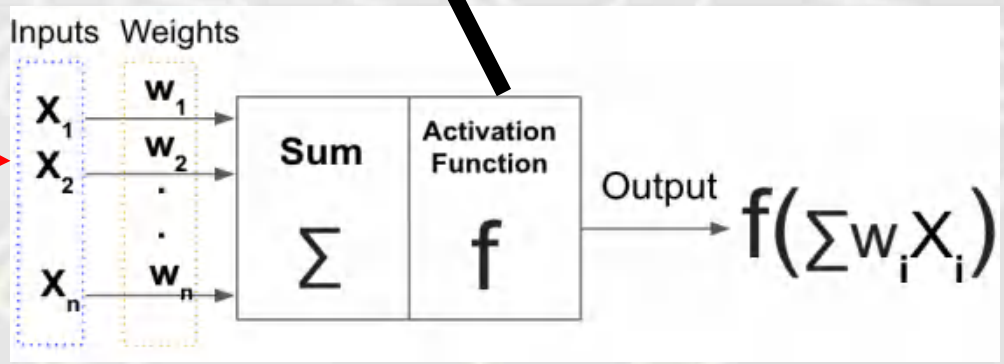
- Fully connected neural network (FCN)
  - Multi-layer perceptron
- Convolutional Neural Network (CNN)
- Recurrent Neural Network
- Autoencoder

Multi-Layer Perceptron (MLP)

Function Approximator







Error: $E(W) = \frac{1}{2}\sum_i (y_i - \hat{y}_i)^2$

y = F(x) = f3(W3, f2(W2, f1(W1, x))

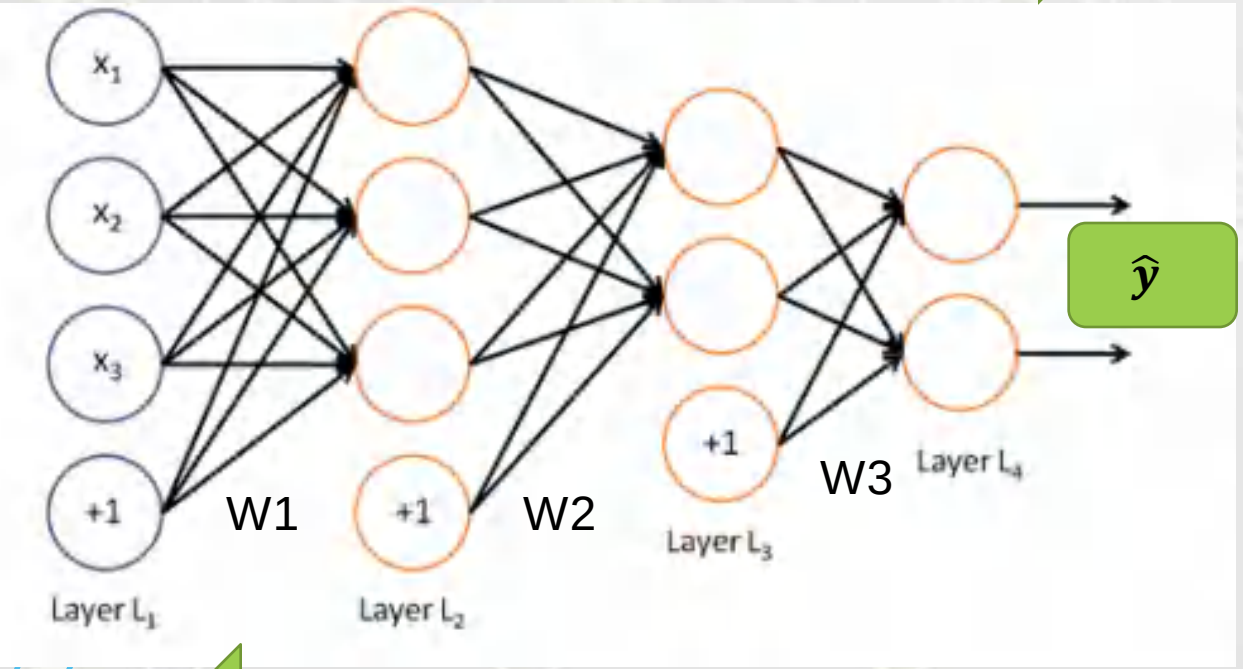
Function composition

Learning: Backpropagation

Learning internal representations by error propagation
DE Rumelhart, GE Hinton, RJ Williams - 1985 - DTIC Document
... PERSONAL AUTHOR(S) David E. Rumelhart, Geoffrey E. **Hinton**, and Ronald J. Williams 13a ...
Continue on reverse if necessary and identify by block number) FIELD GROUP SUB-GROUP
learning; networks; perceptrons; adaptive systems; learning machines; **back propagation** ...
Cited by 17926 Related articles All 36 versions Cite Save

Forward pass

$y = F(x) = f3(W3, f2(W2, f1(W1, x))$

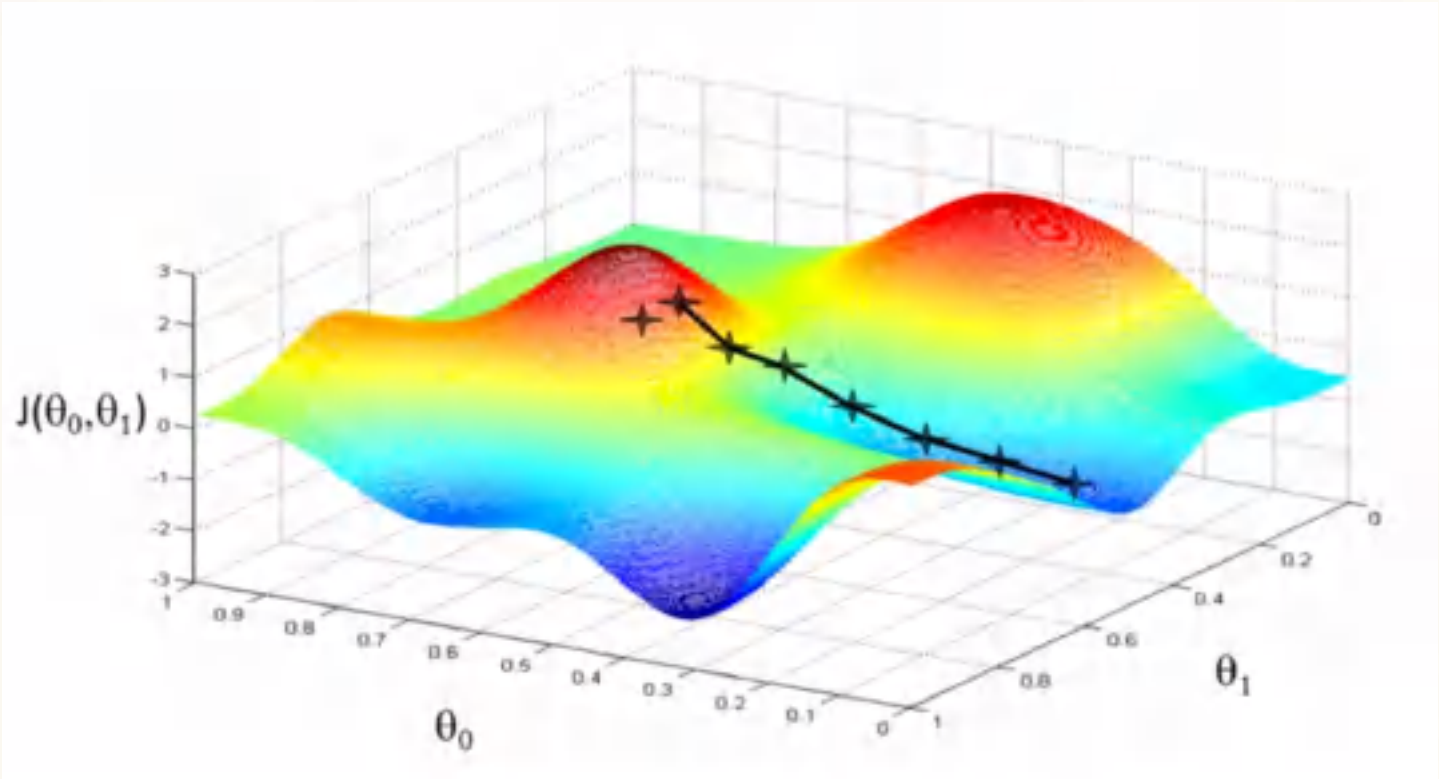


Error: $E(W) = \frac{1}{2}\sum_i (y_i - \hat{y}_i)^2$

Backward pass for computing gradient layer by layer
Chain rule
$\frac{dE(W)}{dW}$

Backward pass

Gradient Descent Learning
$W^{t+1} = W^t - \eta_t \frac{dE(W)}{dW}$


J(θ0,θ1)
θ0
θ1

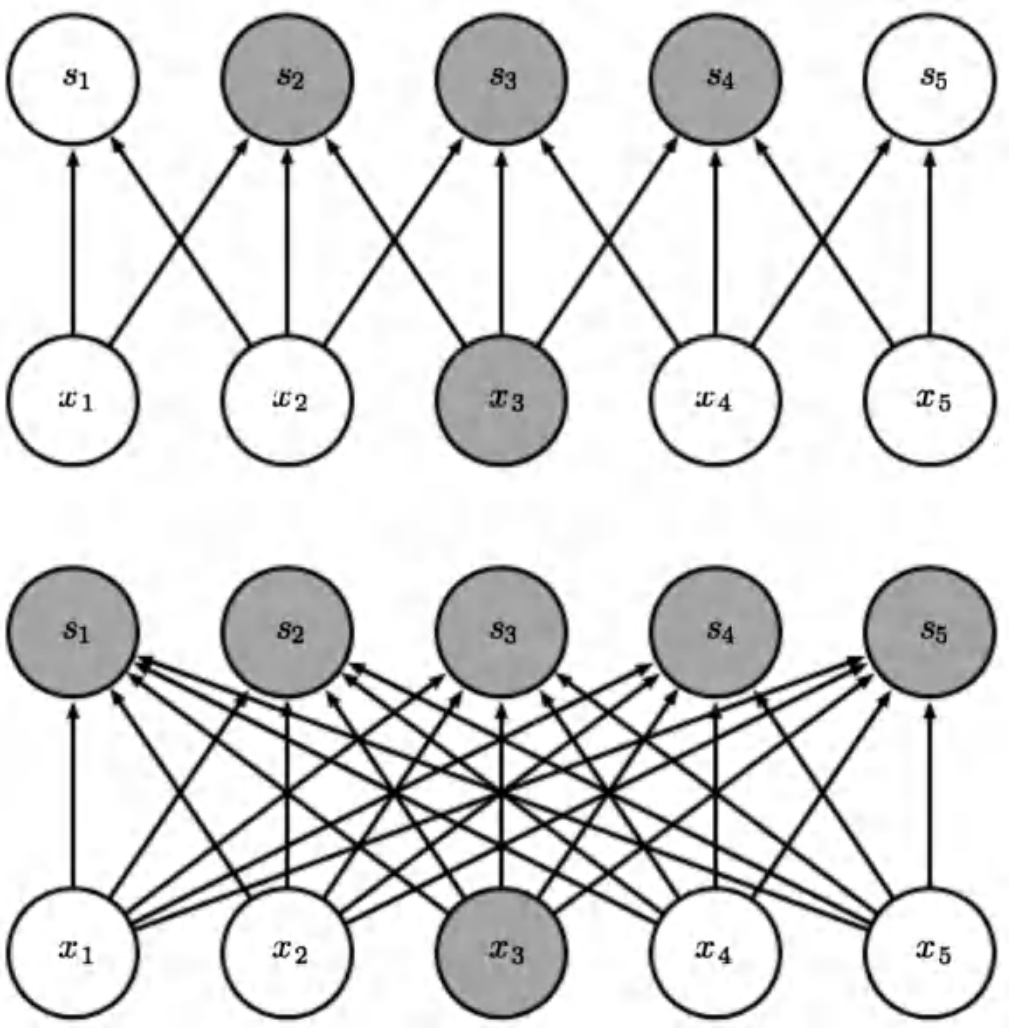
GD IS NOT EFFICIENT ENOUGH

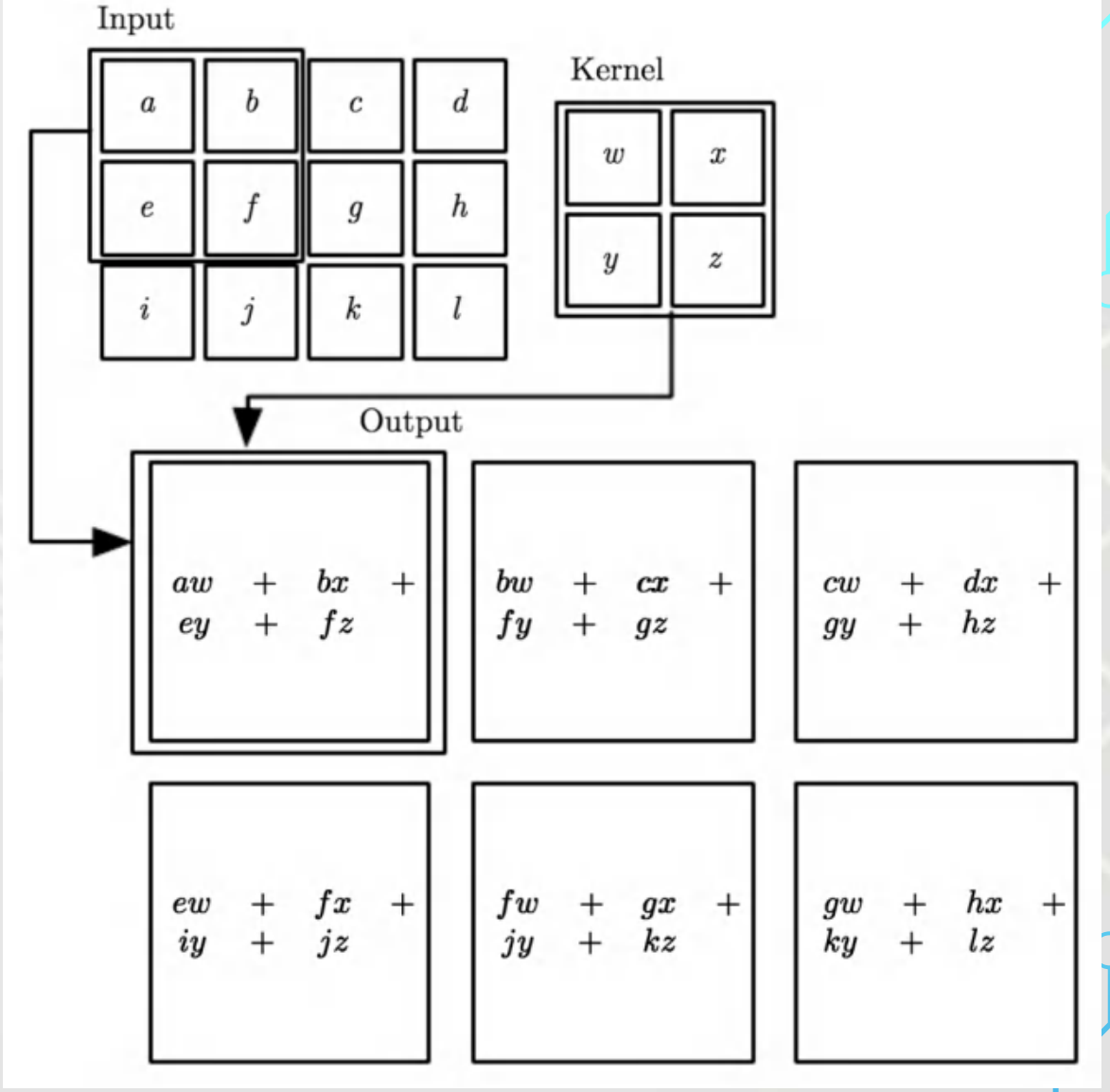
- Stochastic Gradient Descent (SGD)
  - $\frac{dE(W)}{dW}$ involves many additive terms (each with a sample)
  - SGD only select a randomsubset from N (size of dataset)
    - m << N
  - Unbiased estimator of true gradient
  - Computationally feasible

CONVOLUTION NEURAL NETWORK (CNN) (LECUN, 1989)

- Suitable for data with grid topology
  - Especially important in computer vision
    Object recognition, image classification, 2D grid
  - Time series, 1D grid
- Fully-connected not practical for images
  - Huge number of pixels: parameter explosion
  - Image size: 1000*1000, O(1,000,000) first layer
- CNN: convolution
  - Sparse connections; Parameter sharing: Equivalent representations

CNN: CONVOLUTION

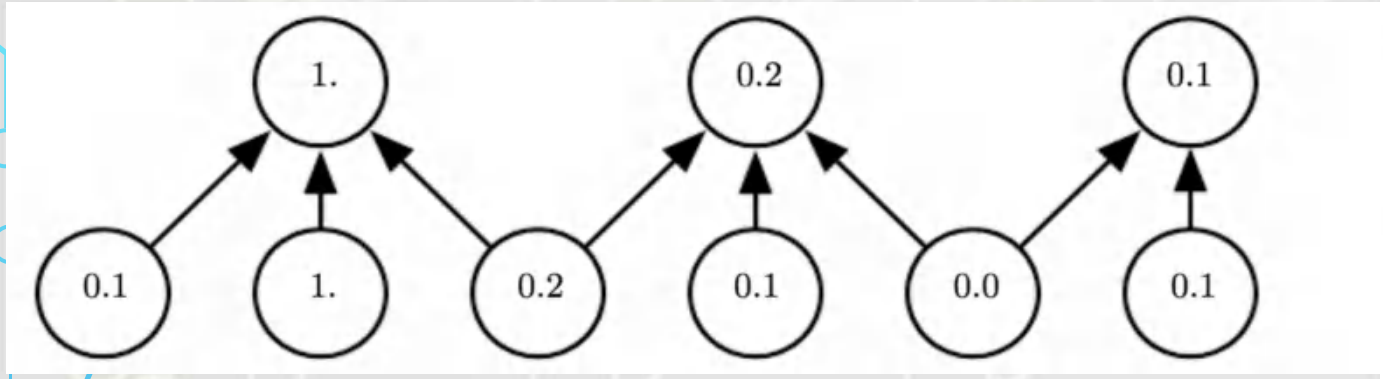


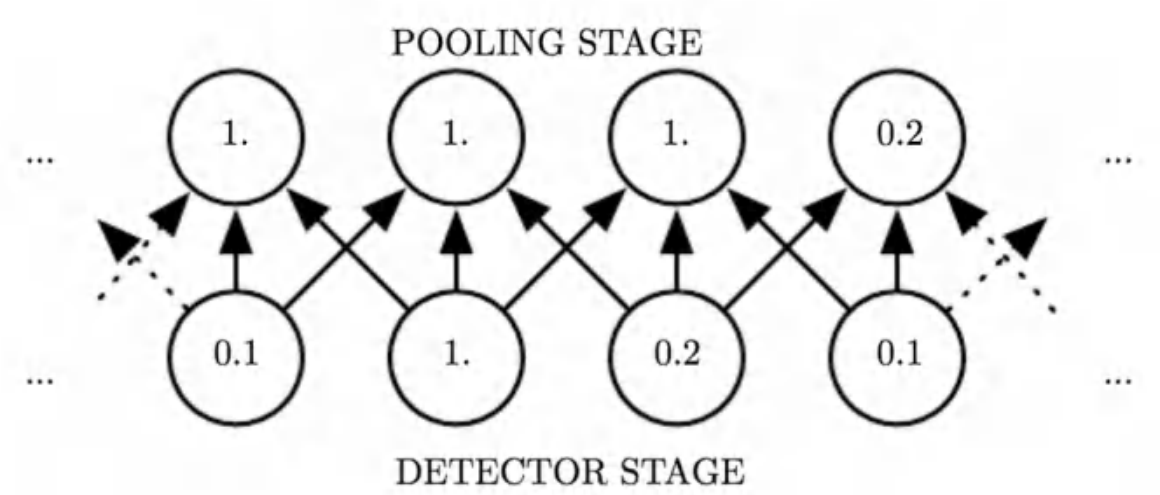


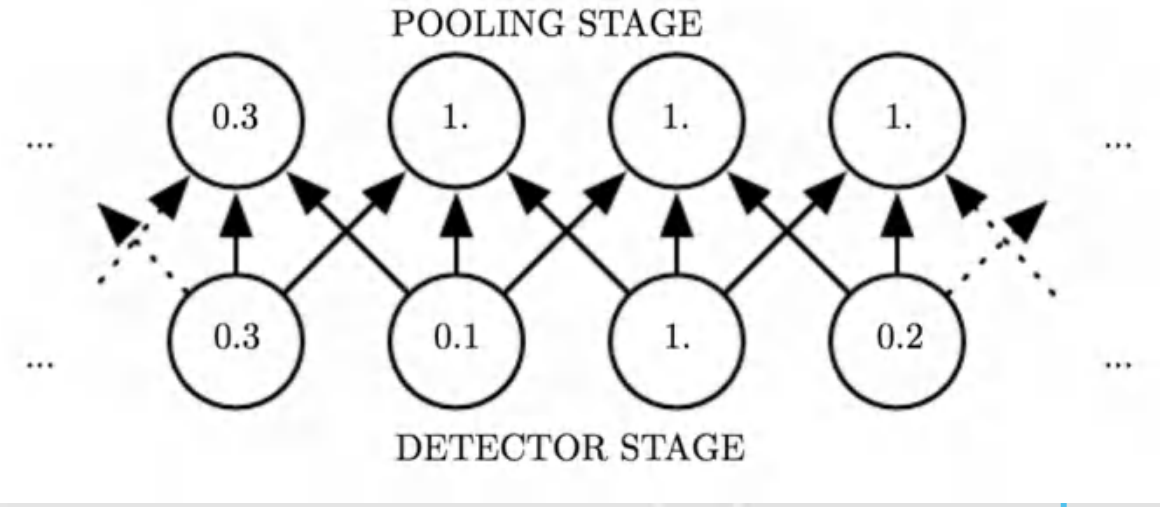
$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n).$$

CNN: POOLING

- Invariant to small translation
- Operating over neighborhood
- Pooling with downsampling



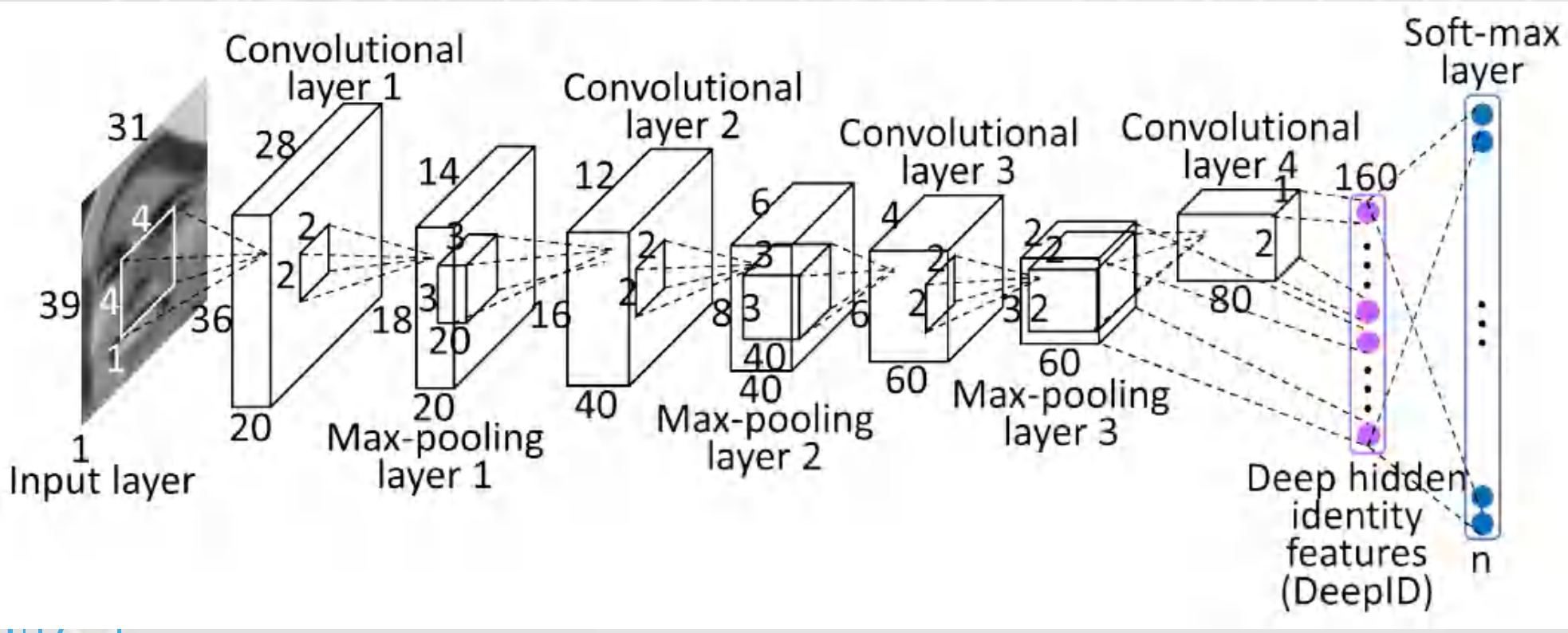


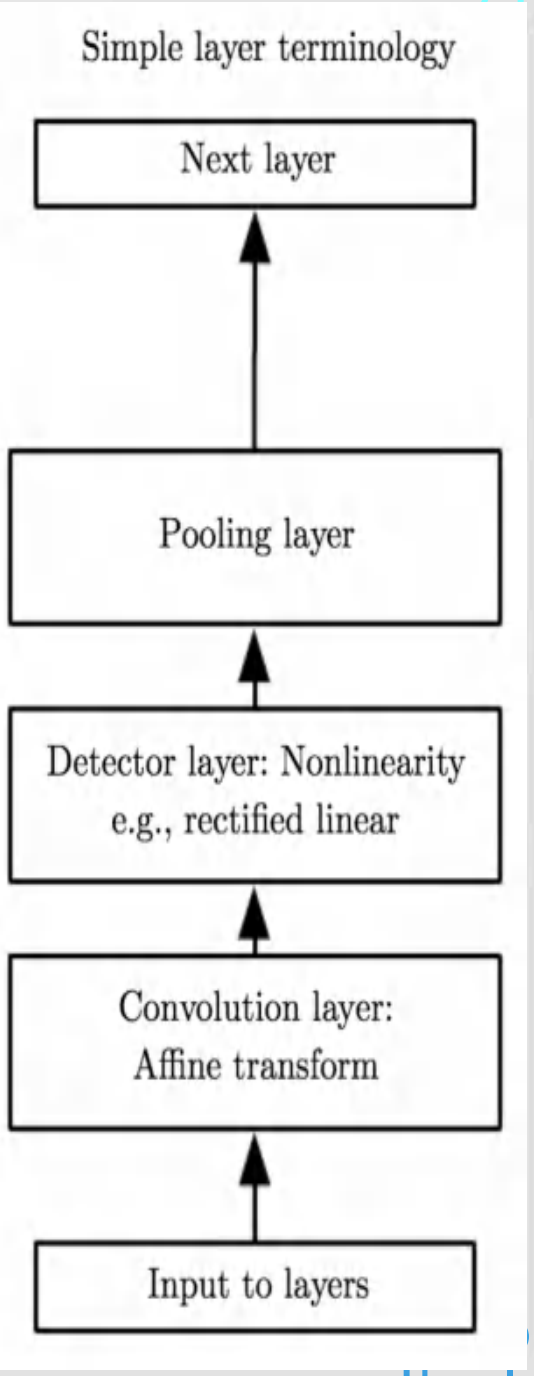


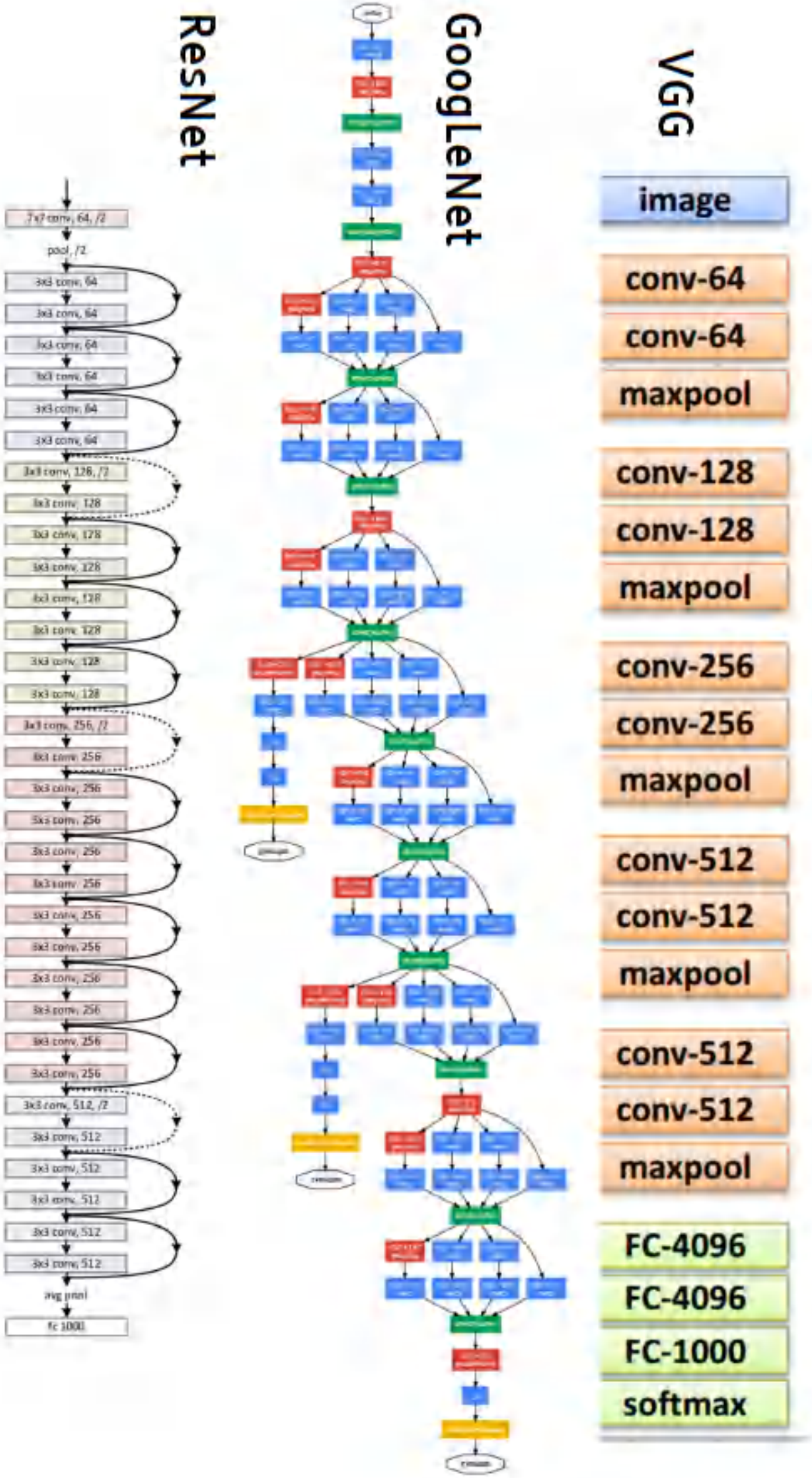
Max pooling

CNN AS A WHOLE

- Stack convolution layers + fully-connected layers





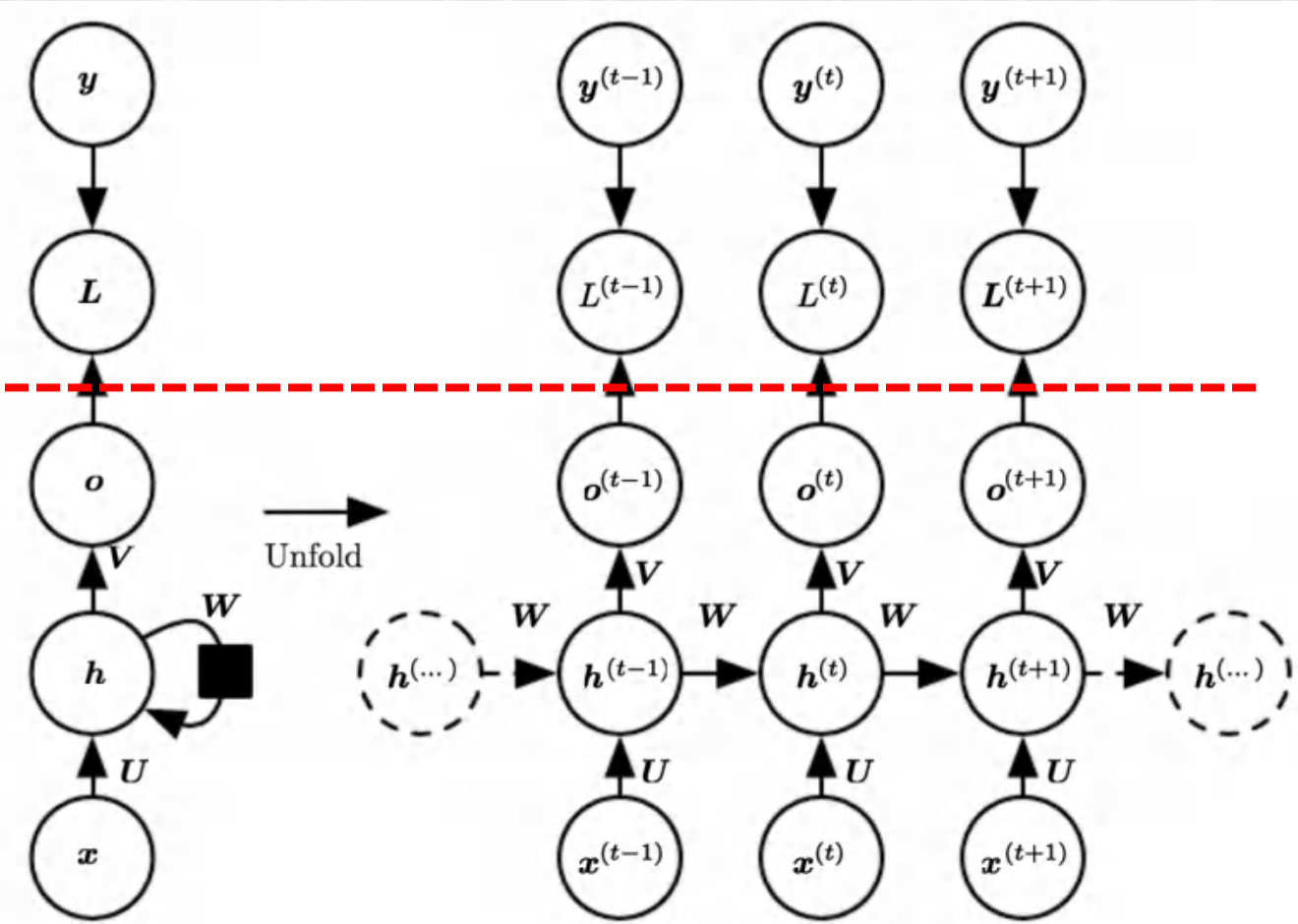

ResNet
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
avg pool
fc 1000
GoogLeNet
VGG
image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax

RECURRENT NEURAL NETWORK (RNN) (RUMELHART, 1986)

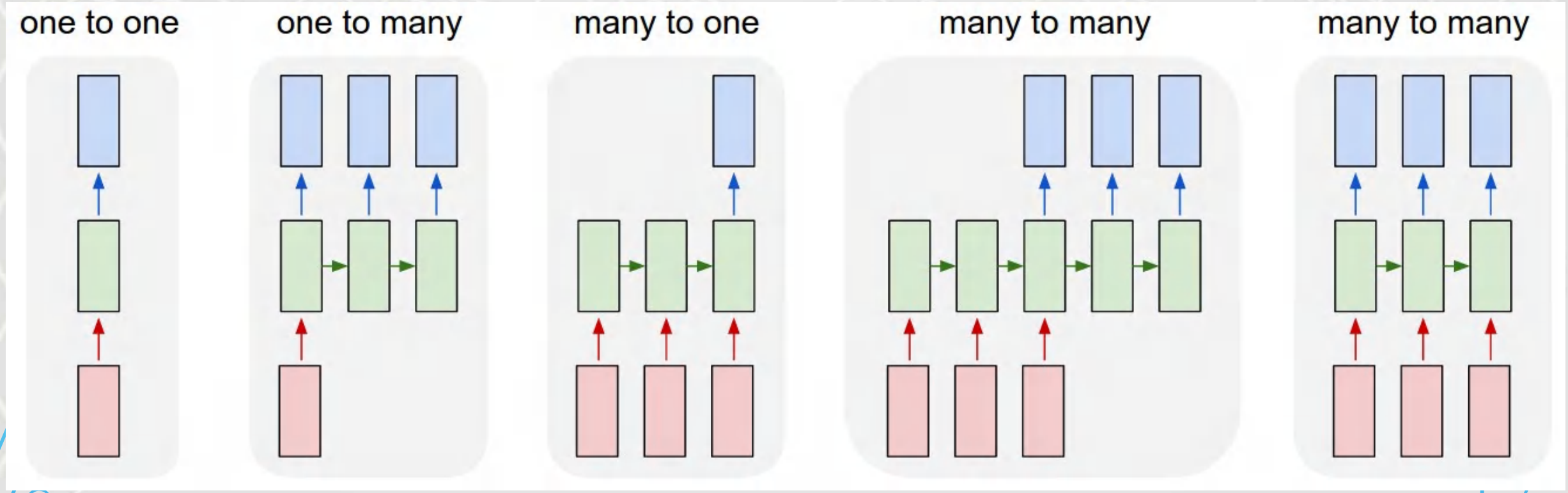
- Suitable for sequential data
  - Time series
  - Natural language processing
  - Speech
  - Video
- Training is not easy, especially for long windows

$$\begin{aligned} \boldsymbol{h}^{(t)} &= g^{(t)}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(t-2)}, \ldots, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) \\ &= f(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}; \boldsymbol{\theta}) \end{aligned}$$

Compute loss



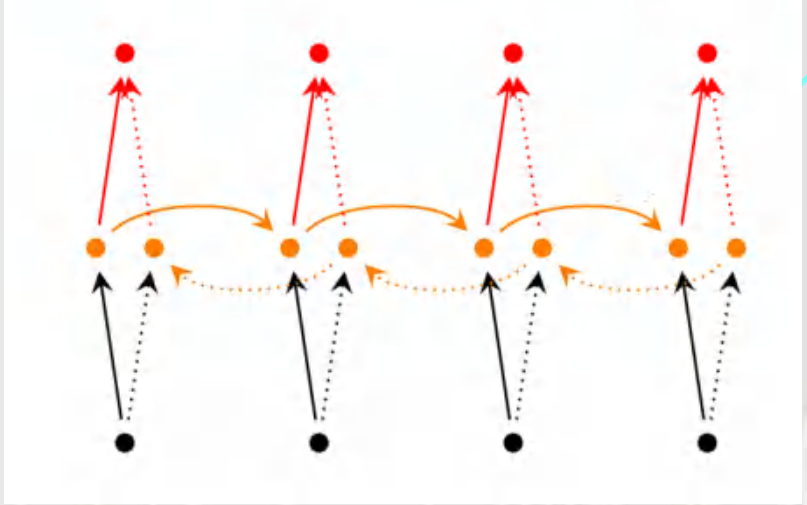
DIFFERENT INPUT-OUTPUT TYPES

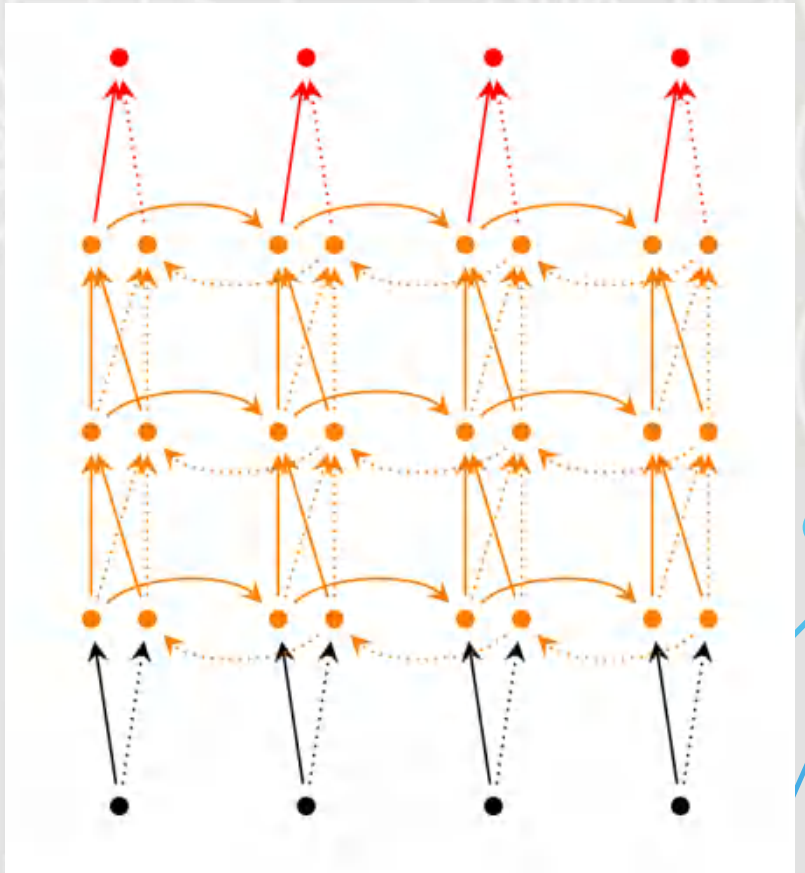


VARIANTS OF RNN

- Bidirectional RNN
  - The output of time *t might also depend on future elements*

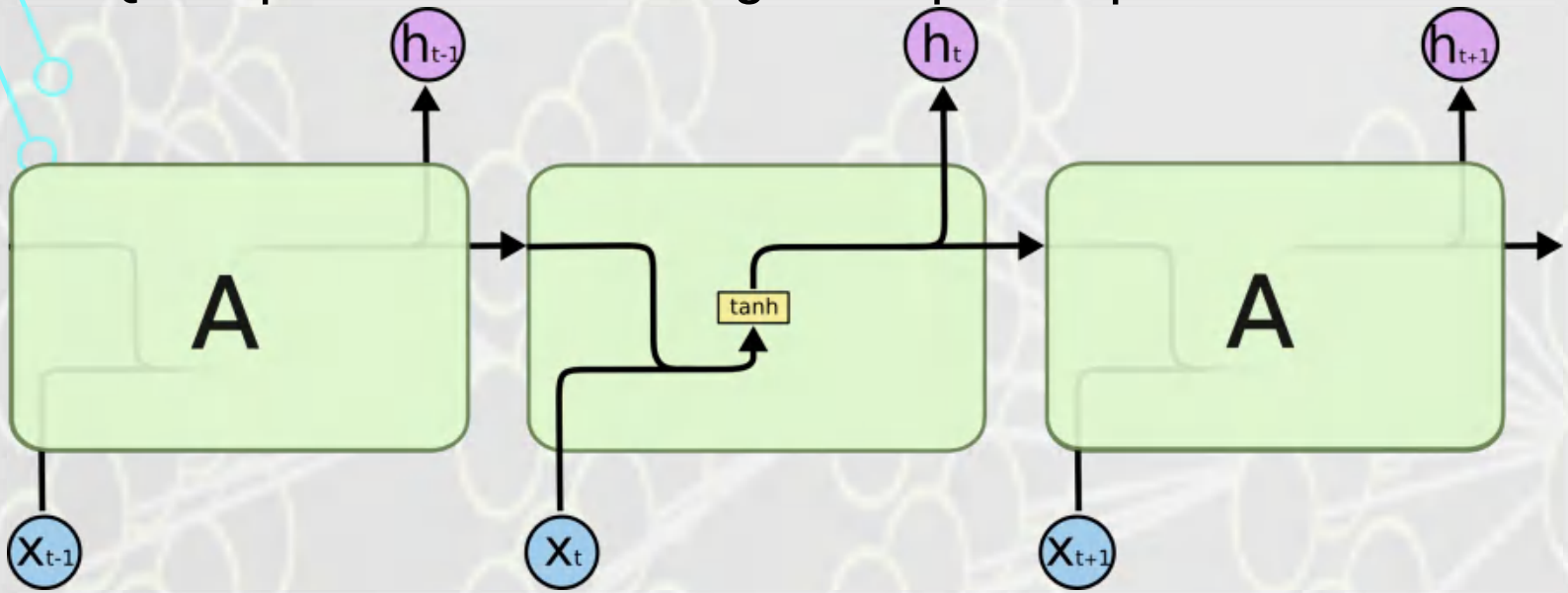
Make it “deep”

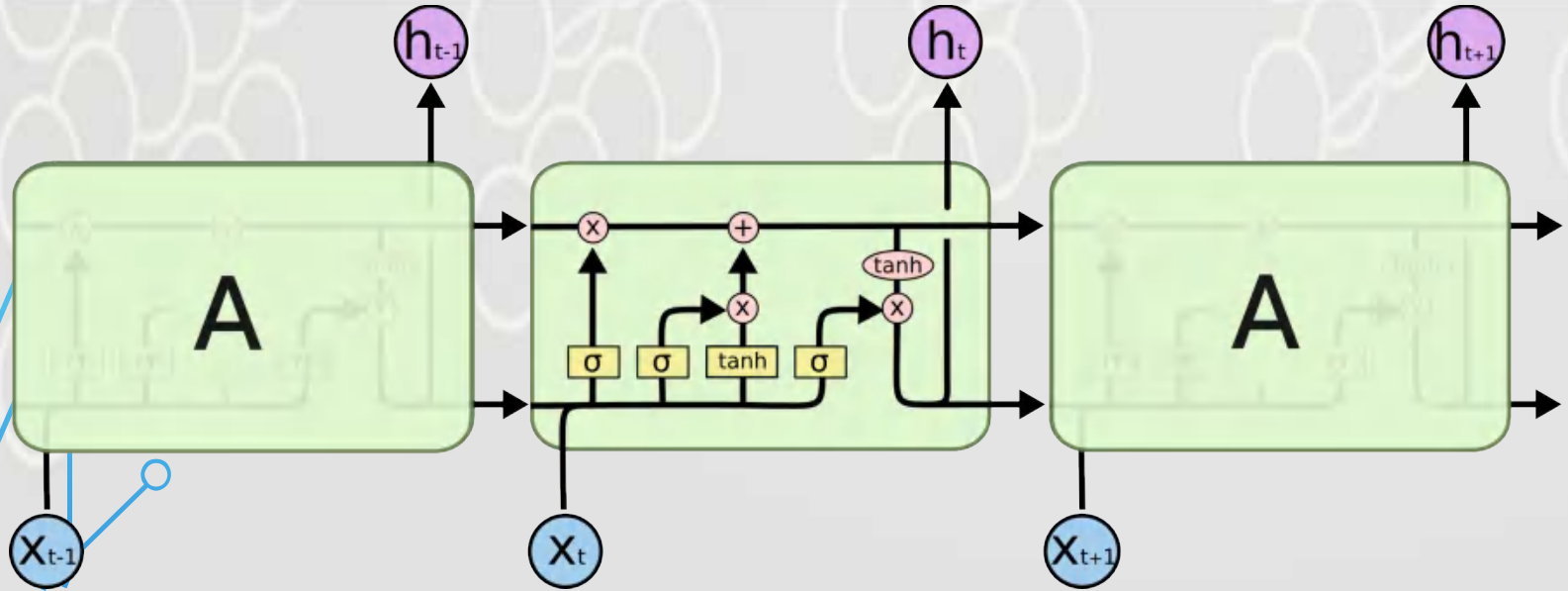




LONG SHORT TERM MEMORY NETWORKS (LSTM)

- Quite powerful according to empirical performances

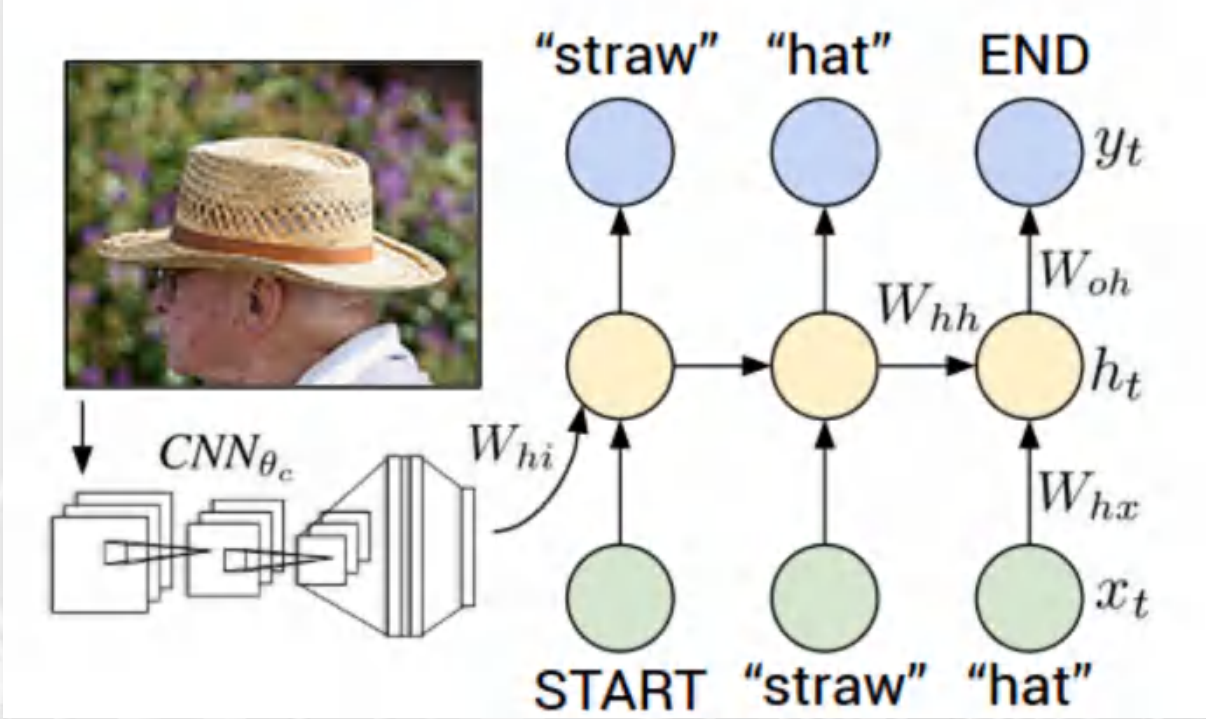




- What information to throw away?
- What to store?
- What to output?

APPLICATIONS:
IMAGE DESCRIPTION

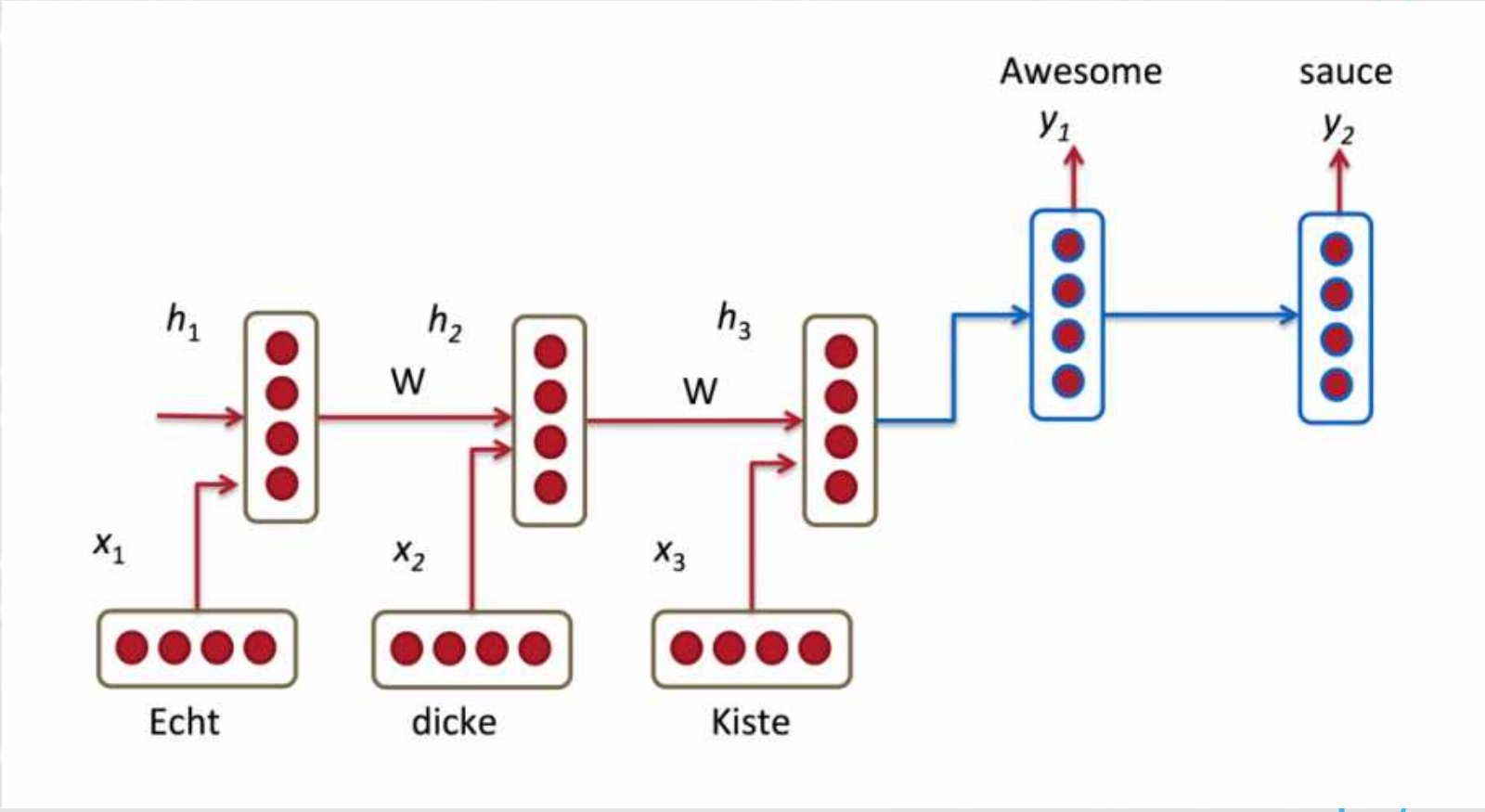
- Joint RNN and CNN



Andrej Karpathy, Li Fei-Fei
Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR2015

APPLICATIONS:
MACHINE TRANSLATION

- RNN encoder-decoder approach
- “Neural machine translation”

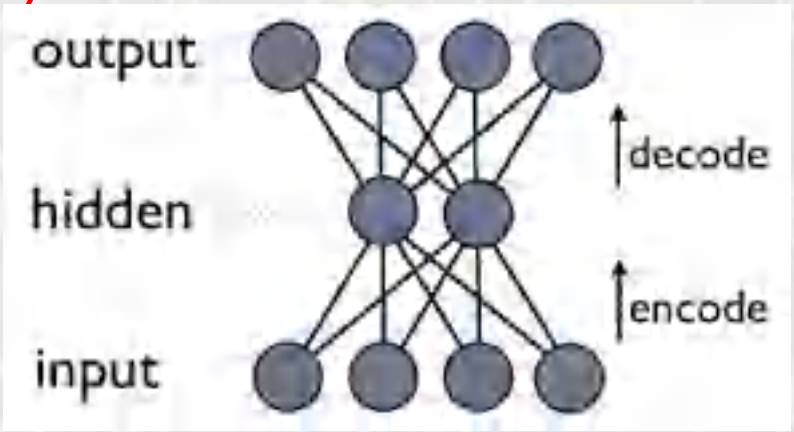


K. Cho et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP 2014.

AUTOENCODER

- Unsupervised learning
  - Output tries to copy input. (undercomplete)
  - Feature extraction
  - Nonlinear: more powerful than PCA

$$L(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$$



$$\mathbf{h} = f(\mathbf{x}) := s_f(W\mathbf{x} + \mathbf{p});$$
$$\mathbf{y} = g(\mathbf{h}) := s_g(\widetilde{W}\mathbf{h} + \mathbf{q}),$$

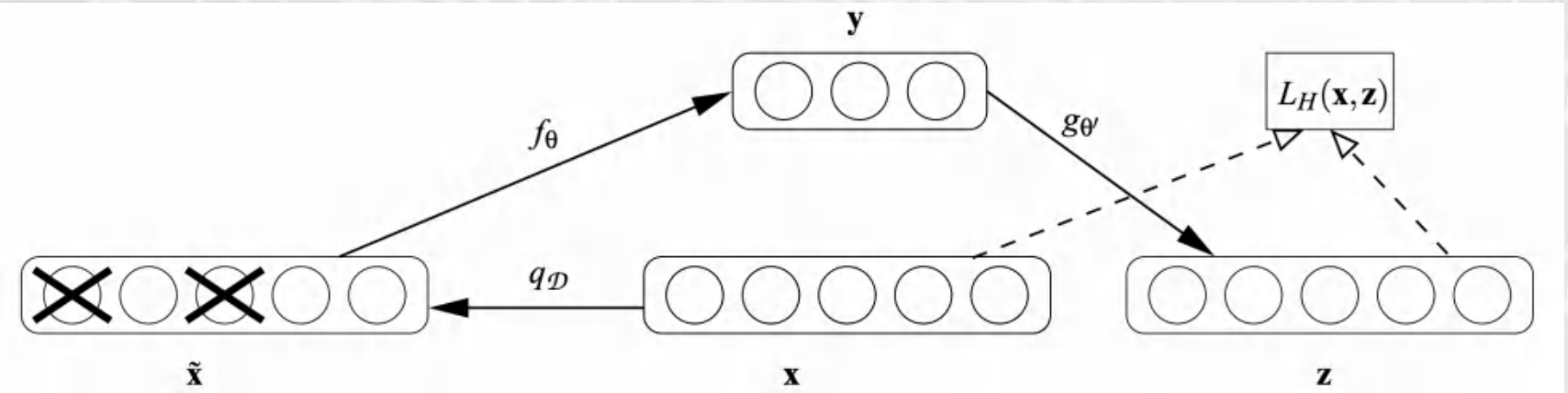
$$s_f(z) = \frac{1}{1+e^{-z}};$$

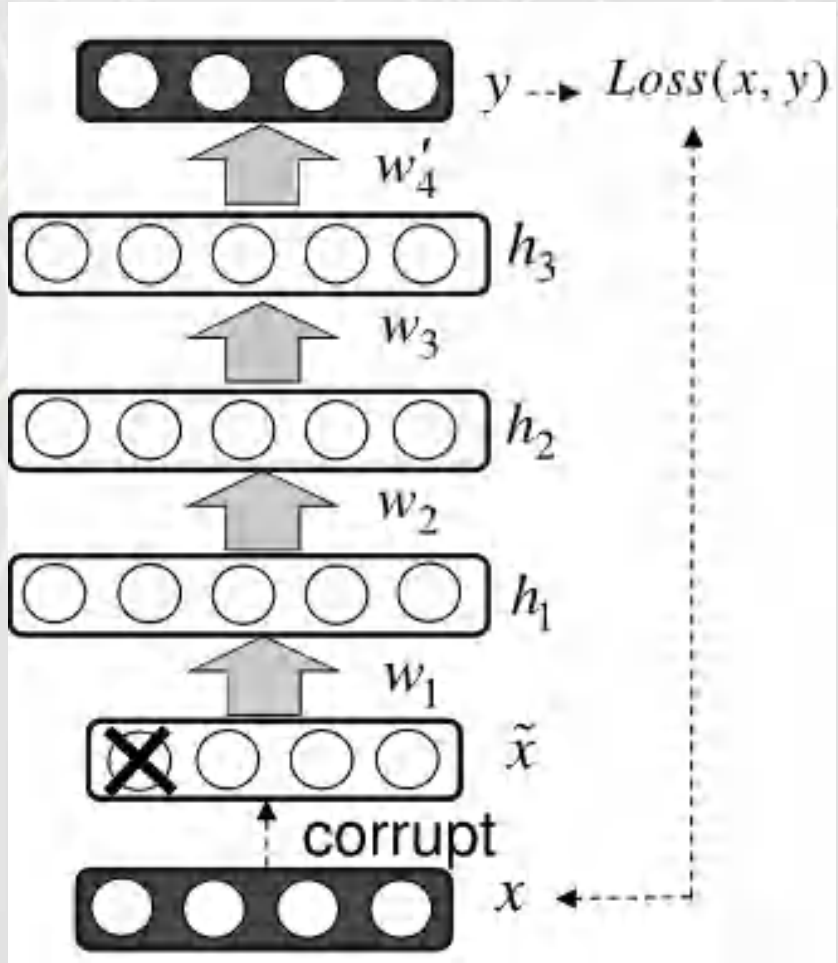
$$s_g(z) = \frac{1}{1+e^{-z}} \text{ 或 } s_g(z) = z.$$

Sometimes : $\widetilde{W} = W^T$

DENOISING AUTOENCODER

- Adding noise to corrupt the input x
- Allow overcomplete hidden representation



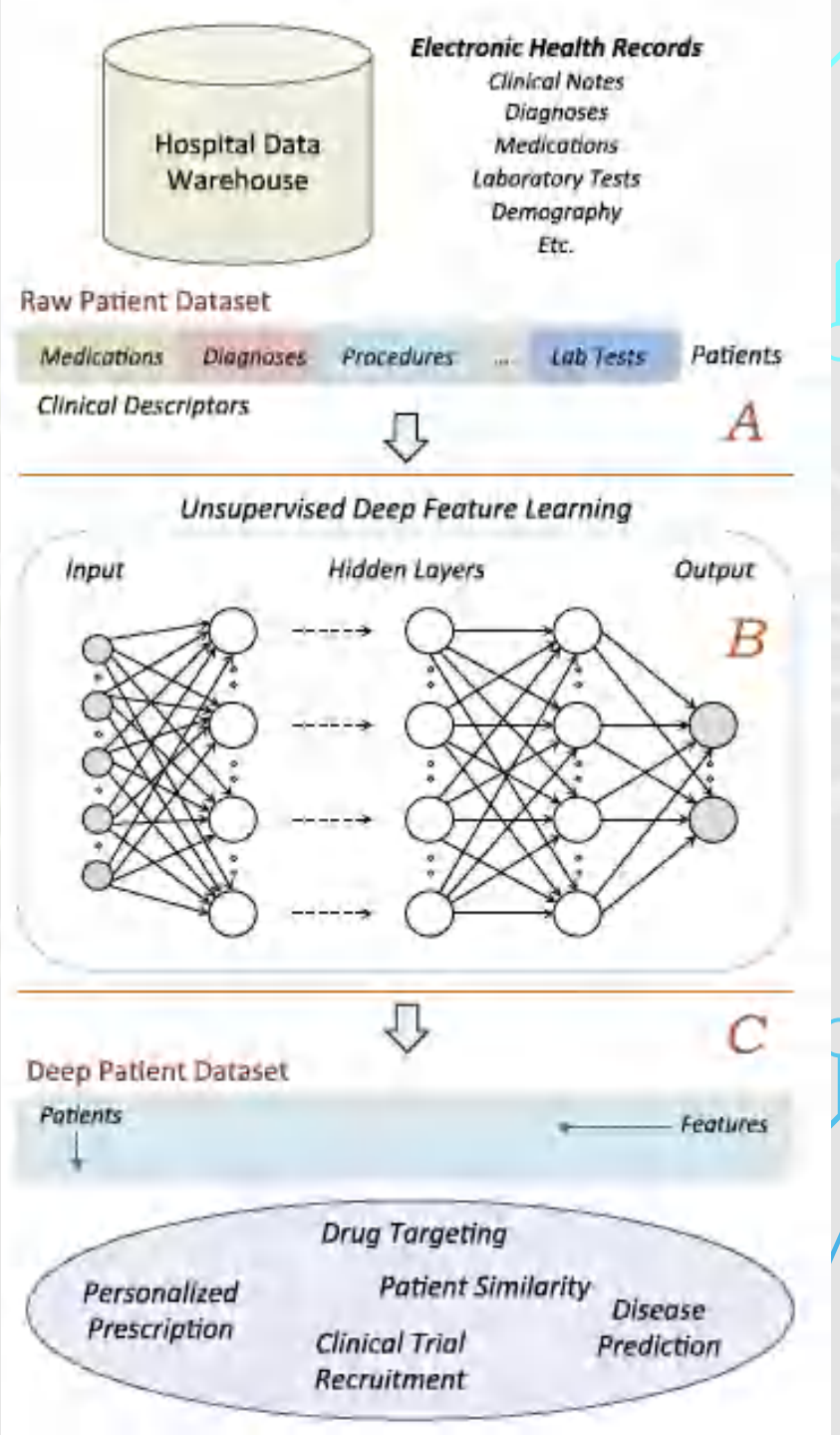


- Stacked denoising autoencoder (popular)

HEALTHCARE: DEEP PATIENT

- Electronic health records (EHRs)
- Stacked denoising autoencoders
- Training: 700,000 patients
- Test: 76,214 patients, 78 diseases
- Accuracy: 92.9%

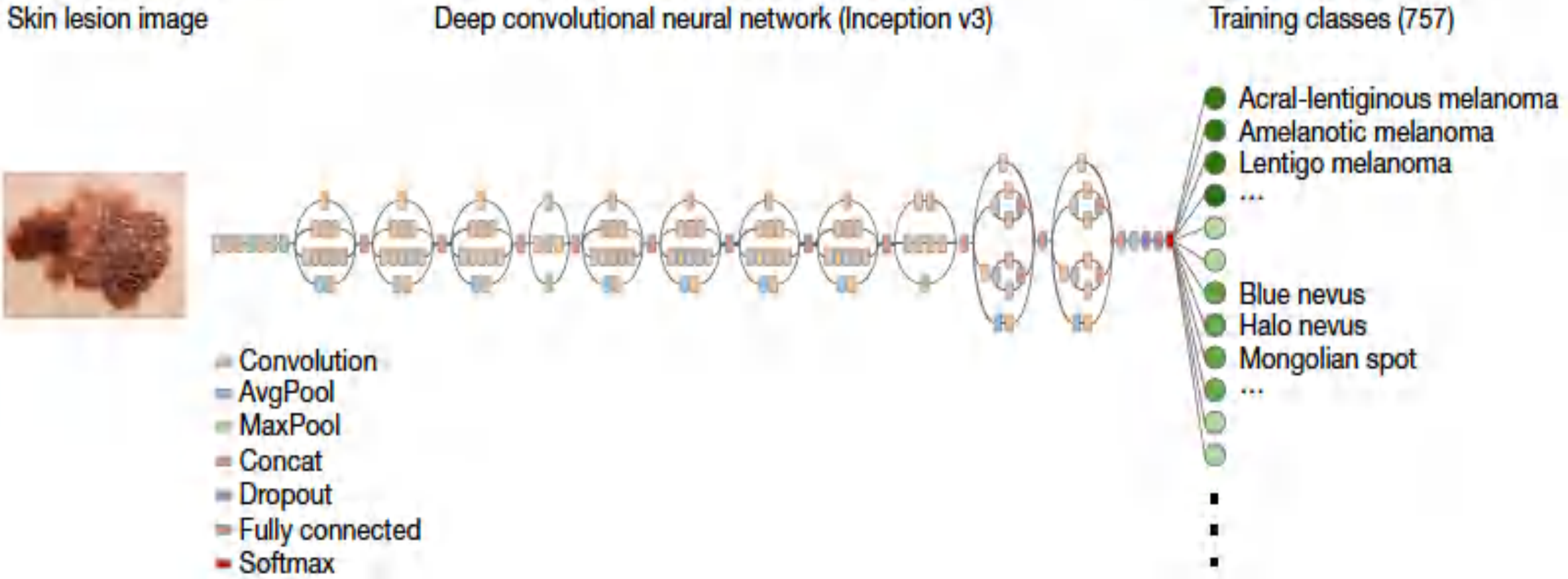
Miotto, Riccardo, et al. “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records.” Nature *Scientific reports* 6 (2016)



HEALTHCARE: SKIN CANCER CLASSIFICATION

- Skin images
- CNN with pretrained parameters from ImageNet
- Training: 129,450 clinical images, 2,032 diseases
- Overall Accuracy: 72.1%

Andre Esteva, et al. “Dermatologist-level classification of skin cancer with deep neural networks.” *Nature (2017)*


Skin lesion image
Deep convolutional neural network (Inception v3)
Training classes (757)
Acral-lentiginous melanoma
Amelanotic melanoma
Lentigo melanoma
...
Blue nevus
Halo nevus
Mongolian spot
...
Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmax

MILESTONES OF DEEP LEARNING

Deng Li@Microsoft &
Hinton DNN
speech recognition
30%
2011

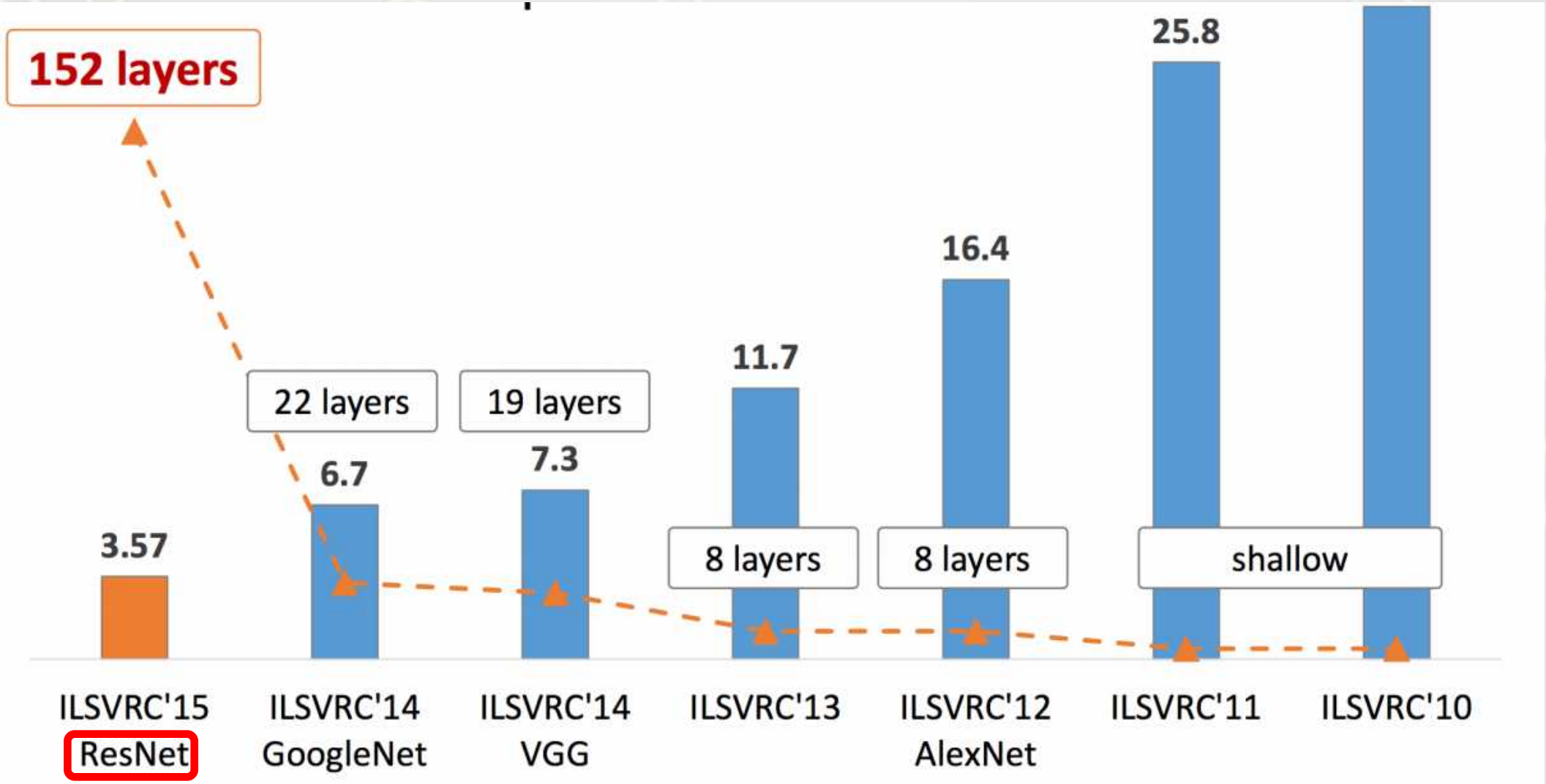
AlexNet won
ImageNet Challenge
2012

ResNet won
ImageNet Challenge
2015

DeepMind
AlphaGo : Sedol Lee
4 : 1

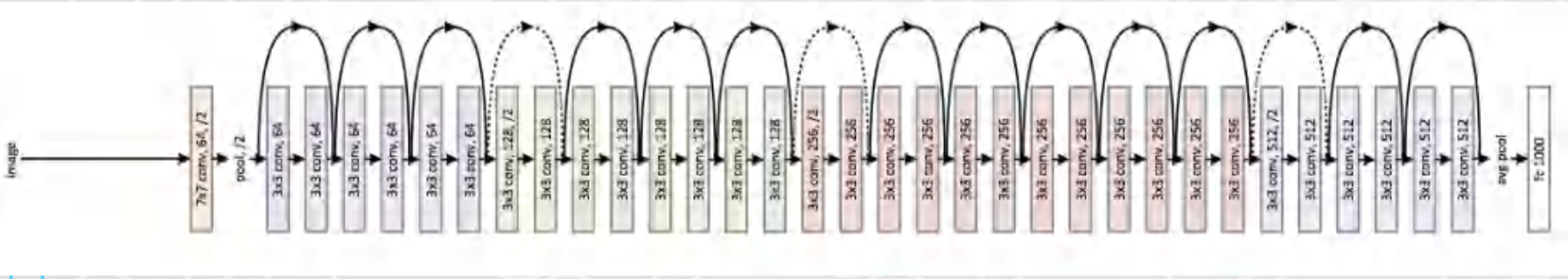
DeepMind
Differentiable Neural
Computers

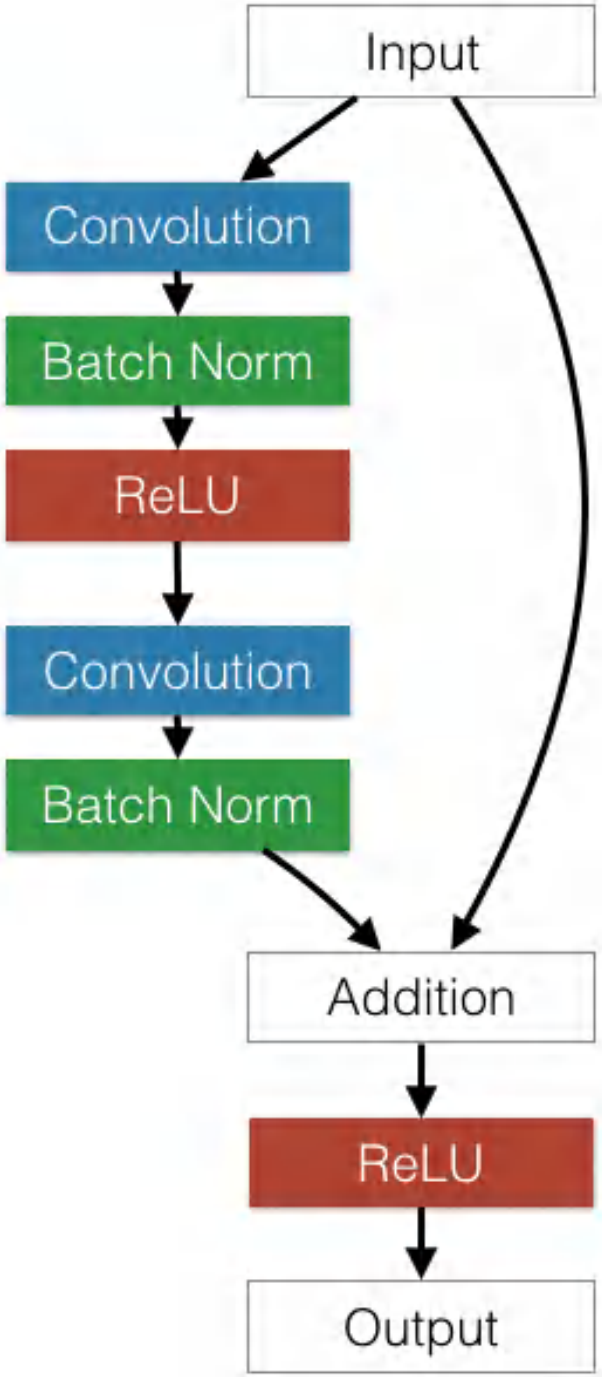
MUCH DEEPER NETWORKS



RESNET

- Residual mapping shortcut connections





Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”, CVPR 2016 Best Paper

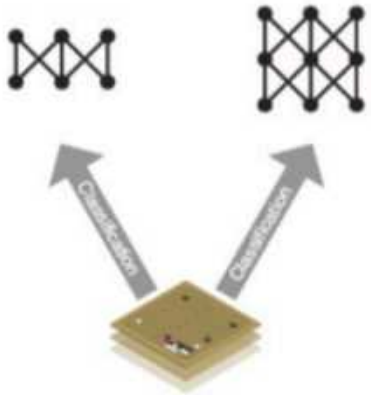
MORE COMPLICATED TASK: GO PLAY

- AlphaGo vs Sedol Lee, 4:1
- Go play is hard for computers!
  - Possible games: $\mathbf{10^{761}}$
  - Search tree: $\mathbf{250^{150}}$
  - Atoms of universe: “only” $\mathbf{10^{80}}$
  - Brute force intractable!
  - Need reduce search space

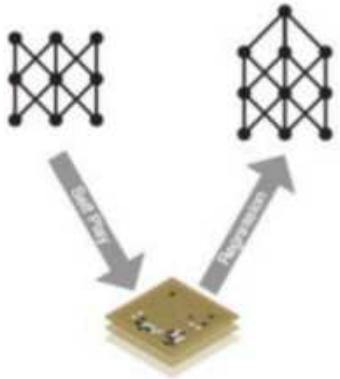
DEEPMIND: ALPHAGO

Silver, D., et. al. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016

Deep Supervised Learning
(Imitating humans)

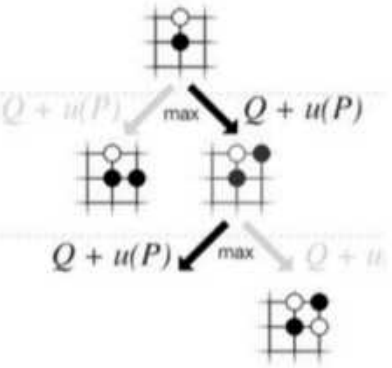


Deep Reinforcement Learning
(Self learning)



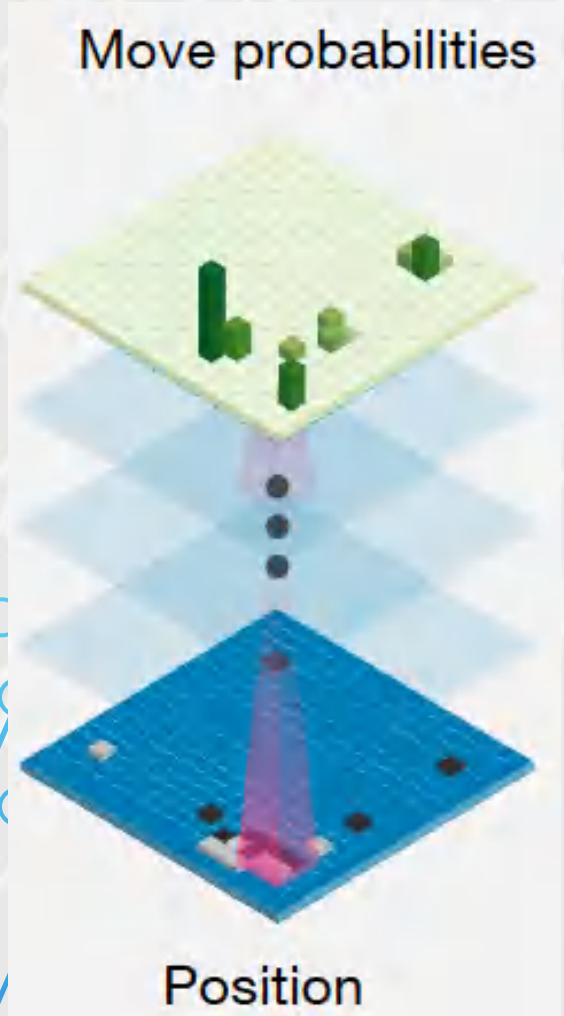
AlphaGo

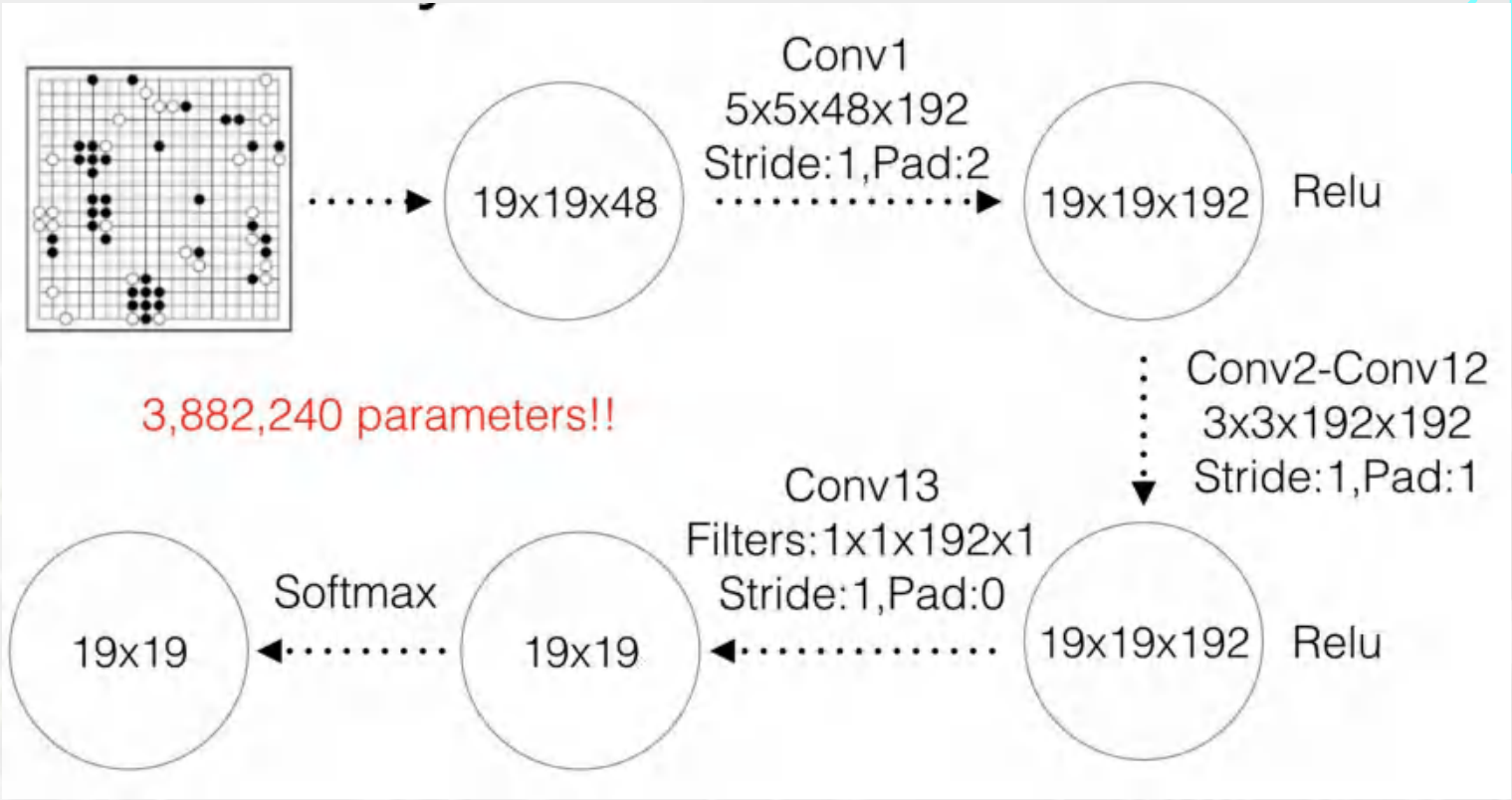
Monte Carlo Tree Search
(Increase stability)



ALPHAGO: PART 1

- Deep Supervised Learning





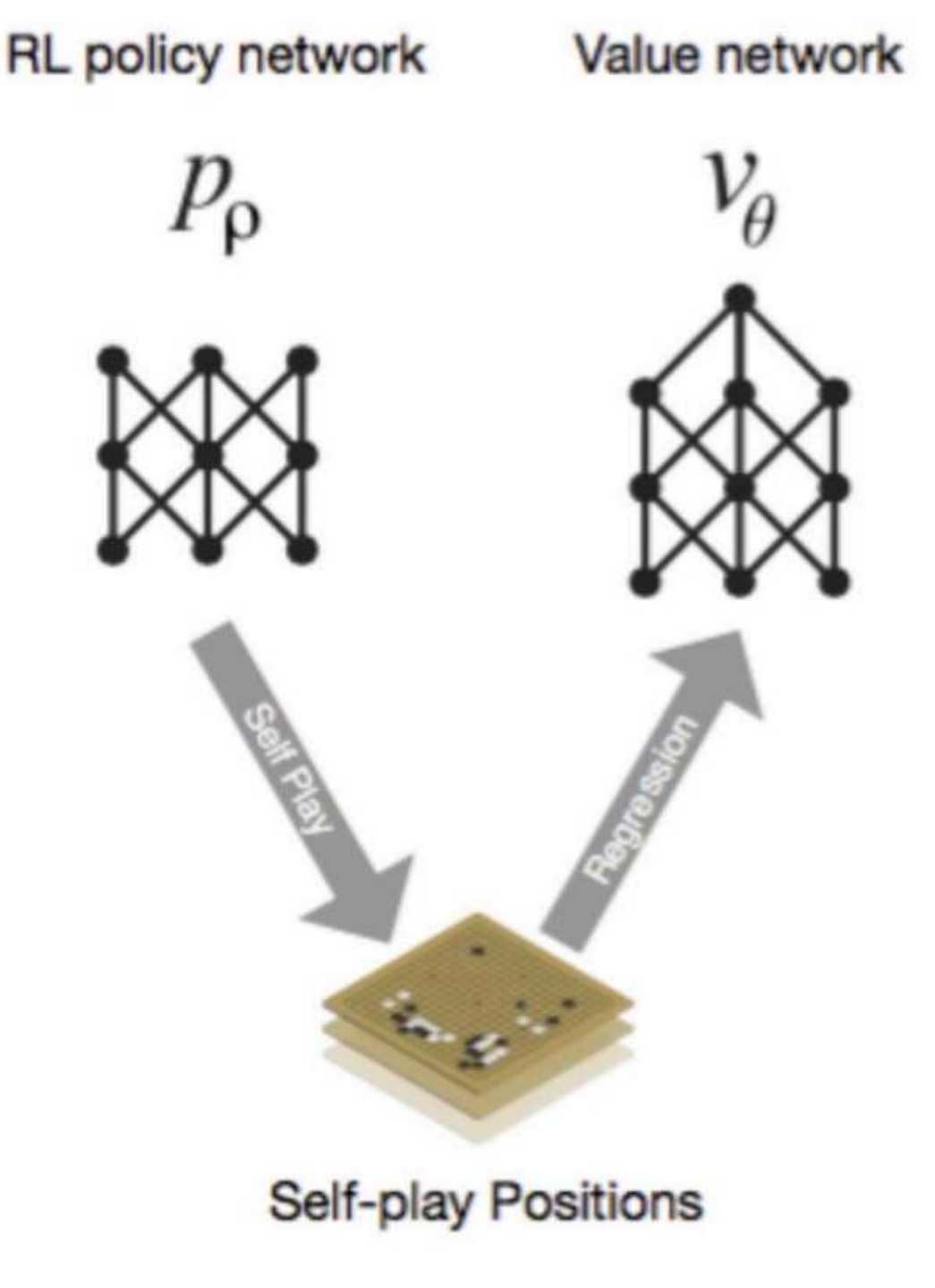
Training data: 160,000 expert games, KGS sever, 30 million positions

Hardware and time: 50 GPU, 3 weeks

Test Accuracy: 57%

ALPHAGO: PART 2

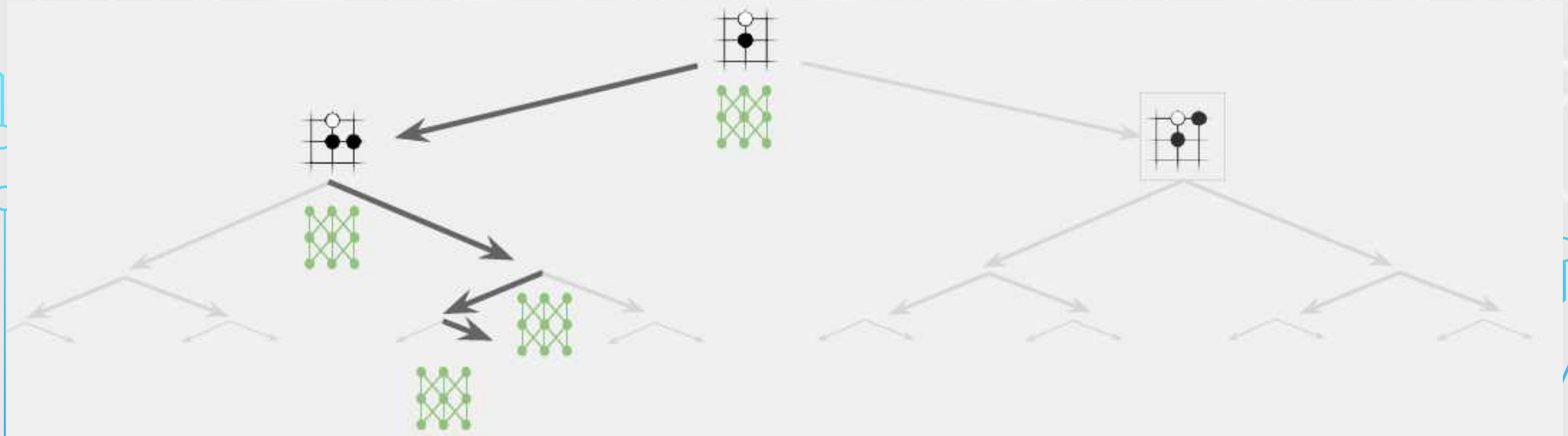
- Deep Reinforcement Learning (Self learning)
- RL policy network: initialize by CNN
- Value network: evaluate the winning probability
- Accuracy: 80%



ALPHAGO: PART 3

99.8% winning over other Go programs

- Putting together: Monte Carlo Tree Search
- Simulating games and choose the best by value network and policy network



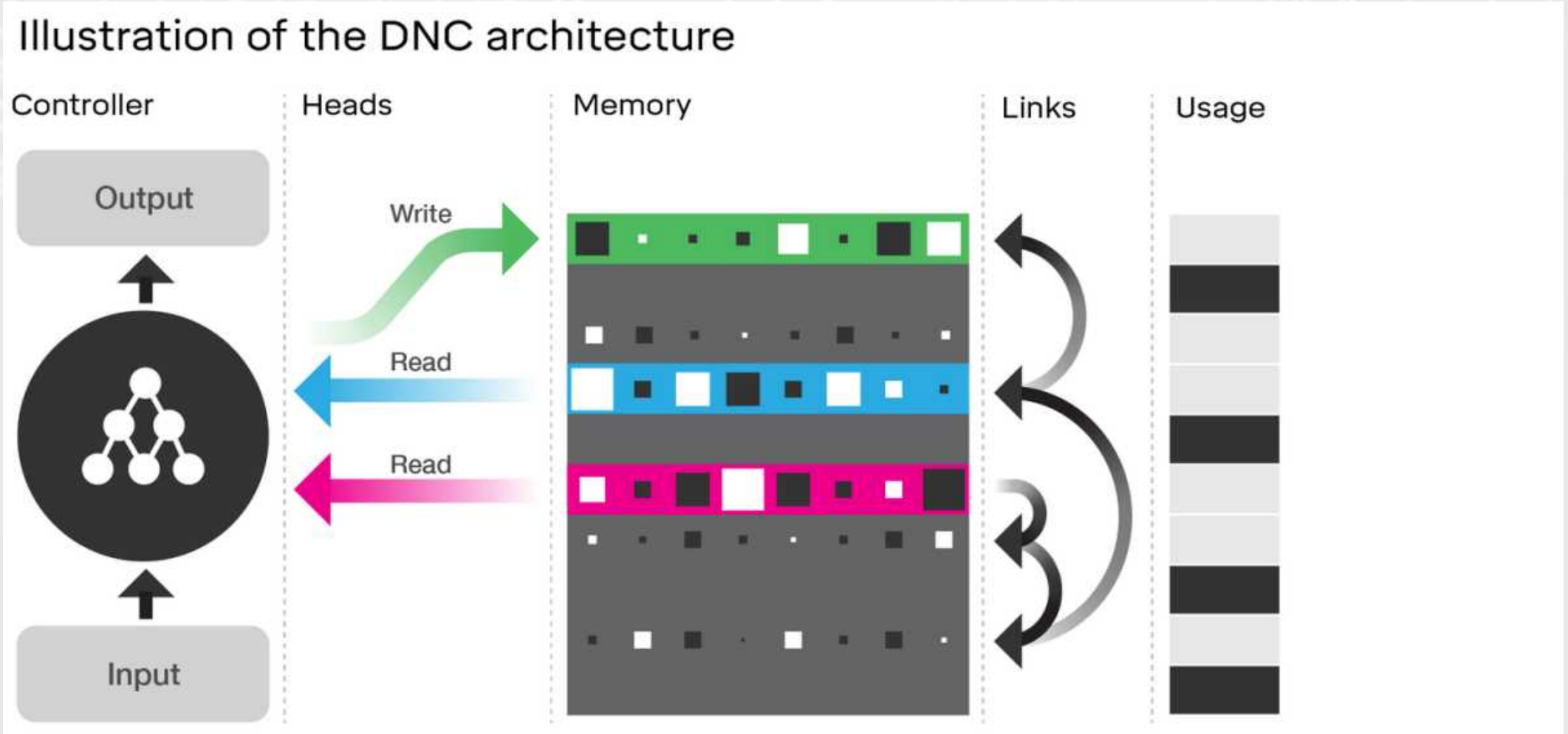
DIFFERENTIABLE NEURAL COMPUTERS
FROM DEEPMIND

- Augment memory (RAM) unit to neural network
- Addressing for reading and writing from memory
- Answer questions about complex, structured data

Graves, A. et.al. Hybrid computing using a neural network with dynamic external memory. *Nature* 2016

DIFFERENTIABLE NEURAL COMPUTERS (DNC)

- Deep layers of LSTMs

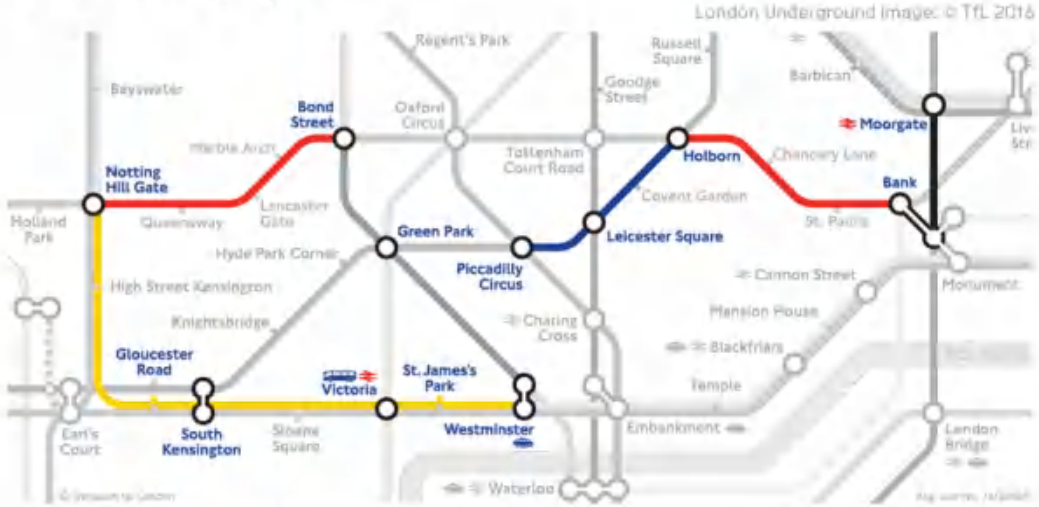


TASKS FOR DNC

- Path searching

Random Training Graph

London Underground



Underground Input:

(OxfordCircus, TottenhamCtRd, Central)
(TottenhamCtRd, OxfordCircus, Central)
(BakerSt, Marylebone, Circle)
(BakerSt, Marylebone, Bakerloo)
(BakerSt, OxfordCircus, Bakerloo)
...
(LeicesterSq, CharingCross, Northern)
(TottenhamCtRd, LeicesterSq, Northern)
(OxfordCircus, PiccadillyCircus, Bakerloo)
(OxfordCircus, NottingHillGate, Central)
(OxfordCircus, Euston, Victoria)

- 84 edges in total

Traversal Question:

(BondSt, _, Central),
(_, _, Circle), (_, _, Circle),
(_, _, Circle), (_, _, Circle),
(_, _, Jubilee), (_, _, Jubilee),

Answer:

(BondSt, NottingHillGate, Central)
(NottingHillGate, GloucesterRd, Circle)
...
(Westminster, GreenPark, Jubilee)
(GreenPark, BondSt, Jubilee)

Shortest Path Question:

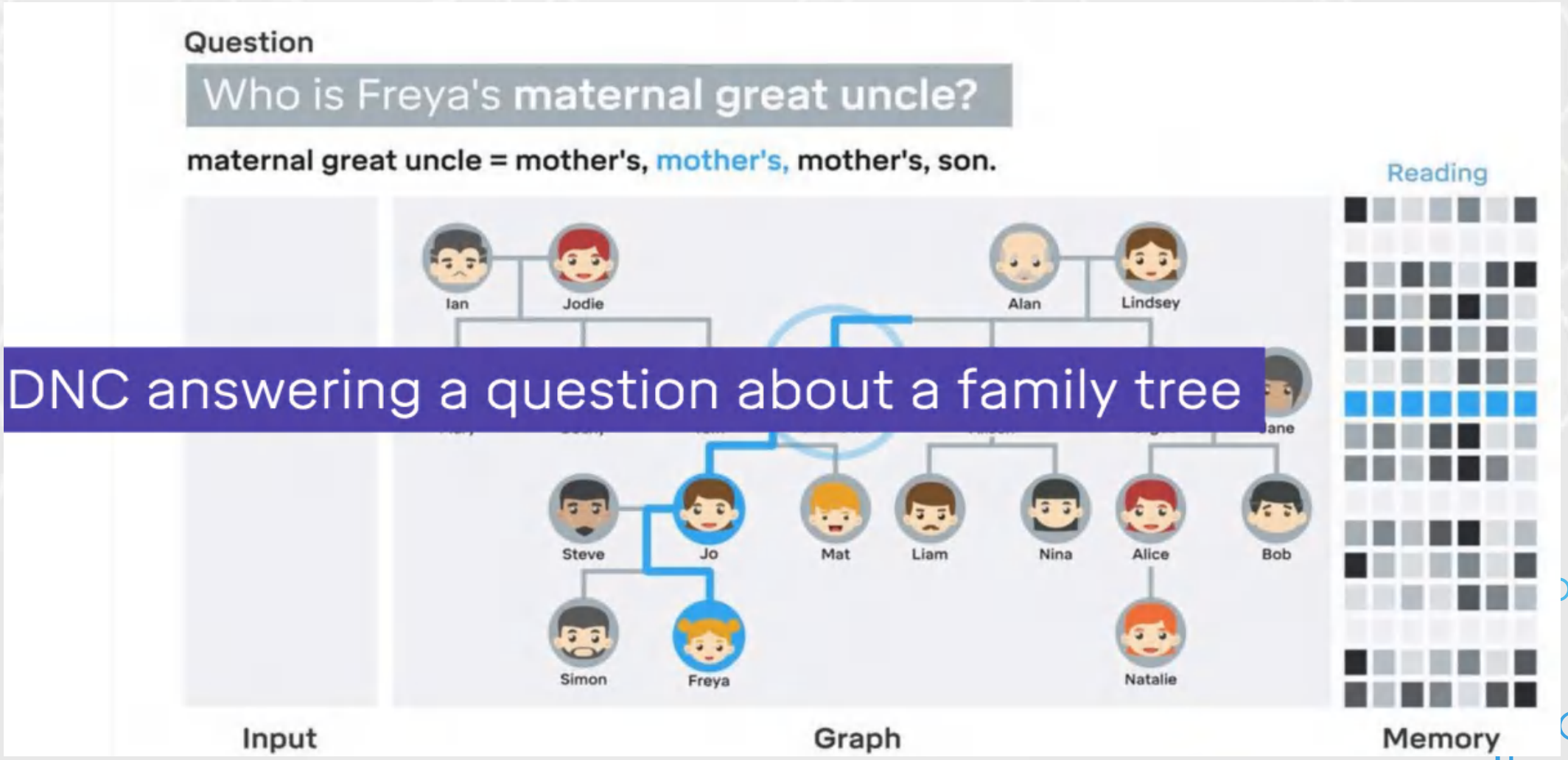
(Moorgate, PiccadillyCircus, _)

Answer:

(Moorgate, Bank, Northern)
(Bank, Holborn, Central)
(Holborn, LeicesterSq, Piccadilly)
(LeicesterSq, PiccadillyCircus, Piccadilly)

TASKS FOR DNC

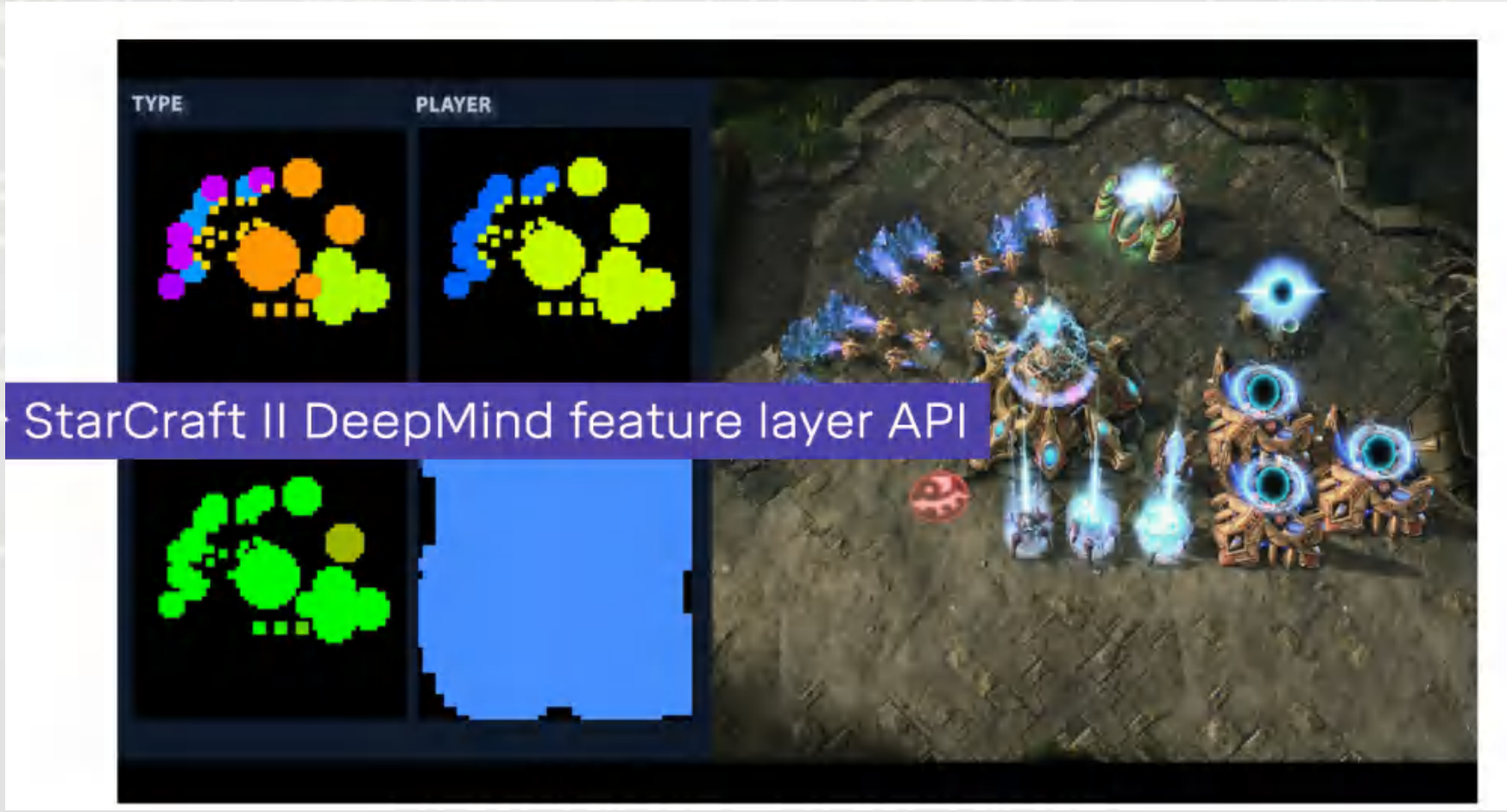
- Family tree



DNC answering a question about a family tree

NEXT MOVE OF DEEPMIND

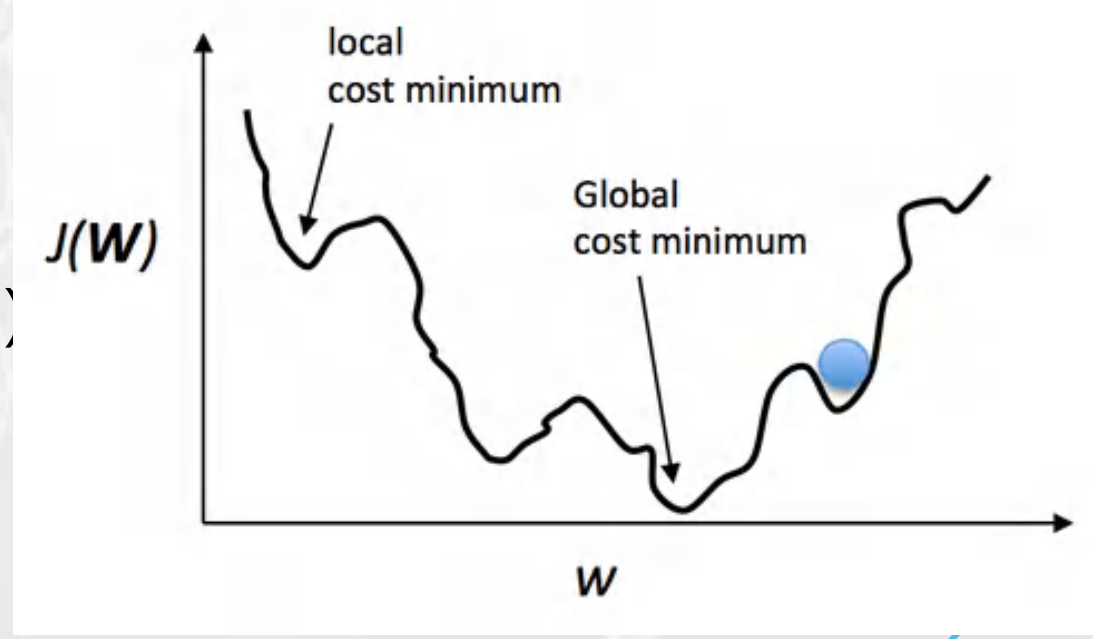
- DeepMind and Blizzard to release StarCraft II as an AI research environment



SOME ISSUES OF DEEP LEARNING

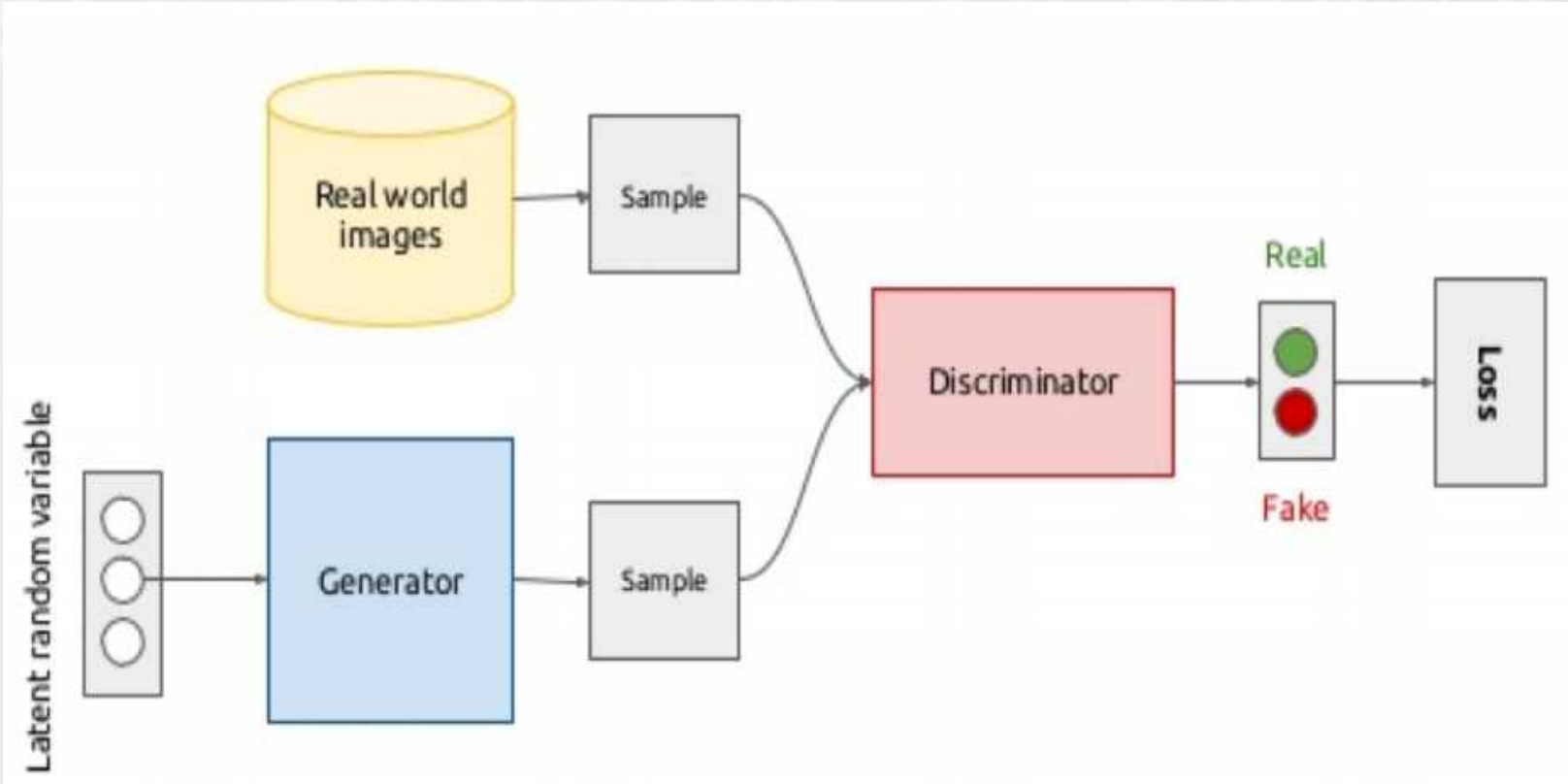
$$W \propto \exp(-J(W)/T(W))$$

- Learning
  - Better optimization methods for high-dimensional non-convex problems? (AdaGrad, Adam, Sampling-based…)
  - Better network structure for learning? (ResNet, Batch Normalization, path-normalized)
  - Dropout



ONE FRONTIER OF MODERN DEEP LEARNING

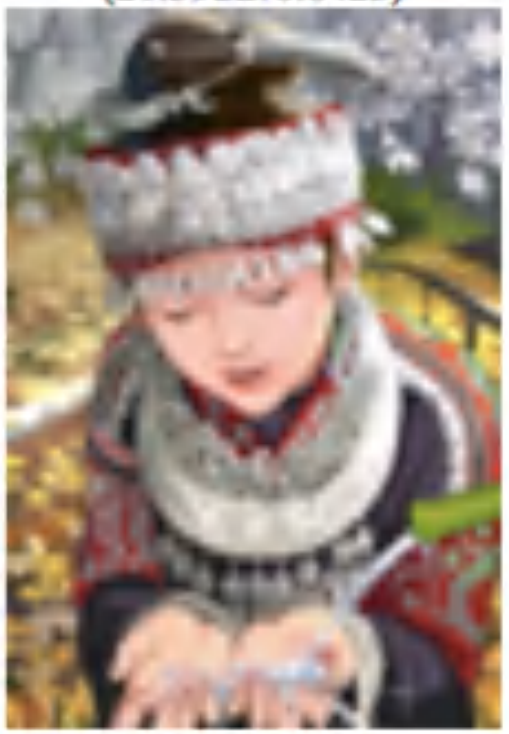
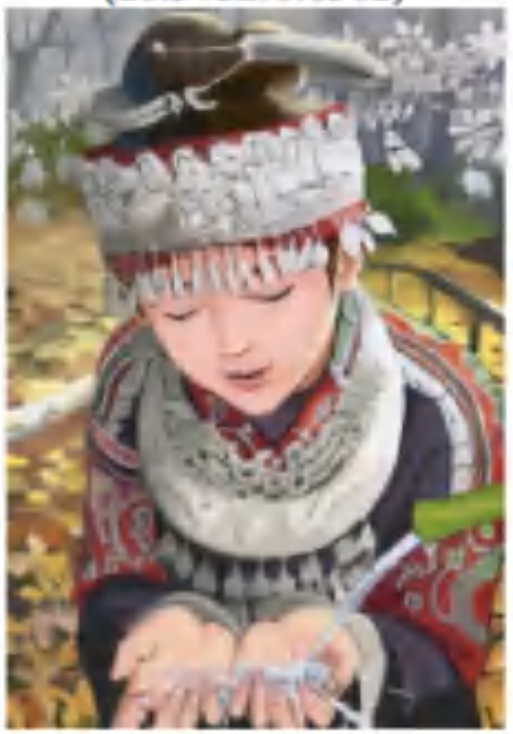
- Unsupervised & semi-supervised learning:
  Deep generative models (DRBM, NADE, Variational autoencoder, Generative Adversarial Networks (GAN)…)
  - (implicitly) learning a distribution over data: p(x)
  - Not realistic enough
  - Hard to manipulate
  - Much space to explore



Objective function of GAN

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

Single Image Super-Resolution

original	bicubic (21.59dB/0.6423)	SRResNet (23.44dB/0.7777)	SRGAN (20.34dB/0.6562)
			

(Ledig et al 2016)

This small blue bird has a short pointy beak and brown on its wings

This bird is completely red with black wings and pointy beak

A small sized bird that has a cream belly and a short pointed bill

A small bird with a black head and wings and features grey wings

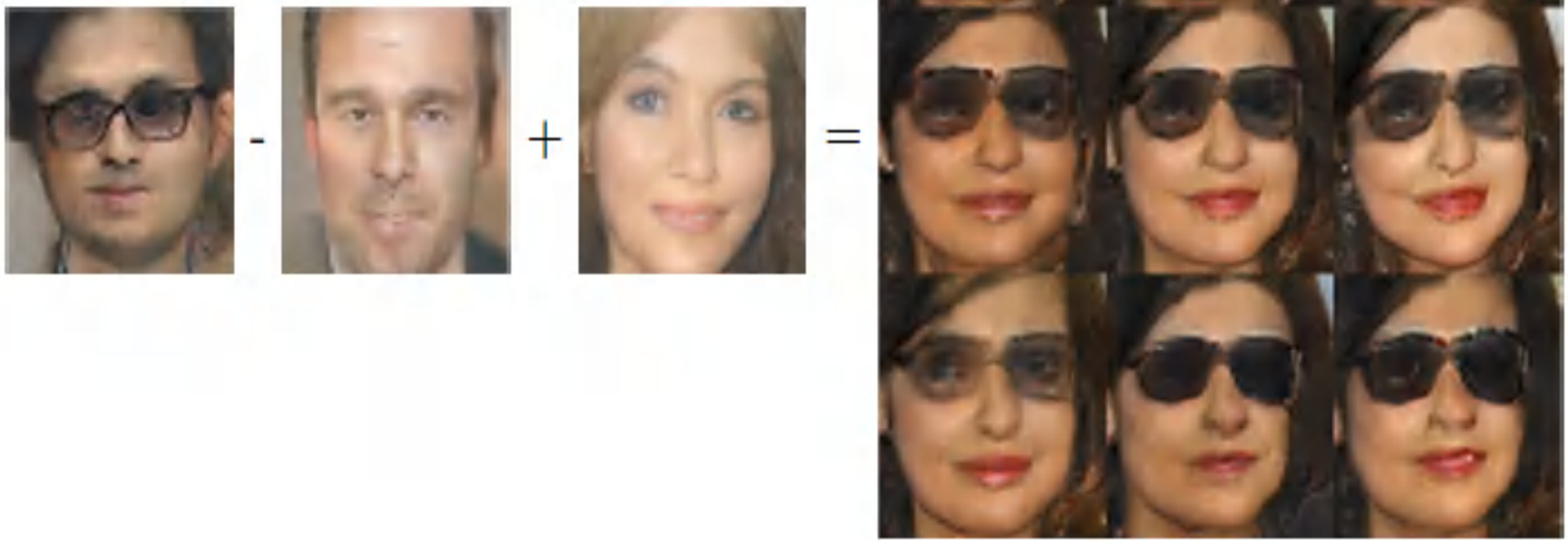
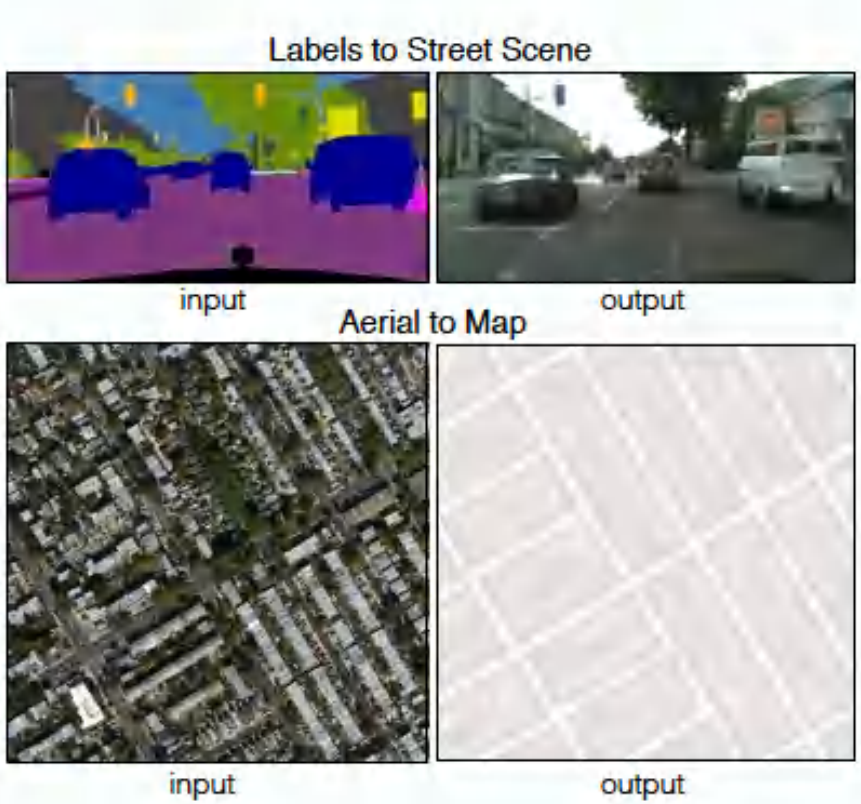
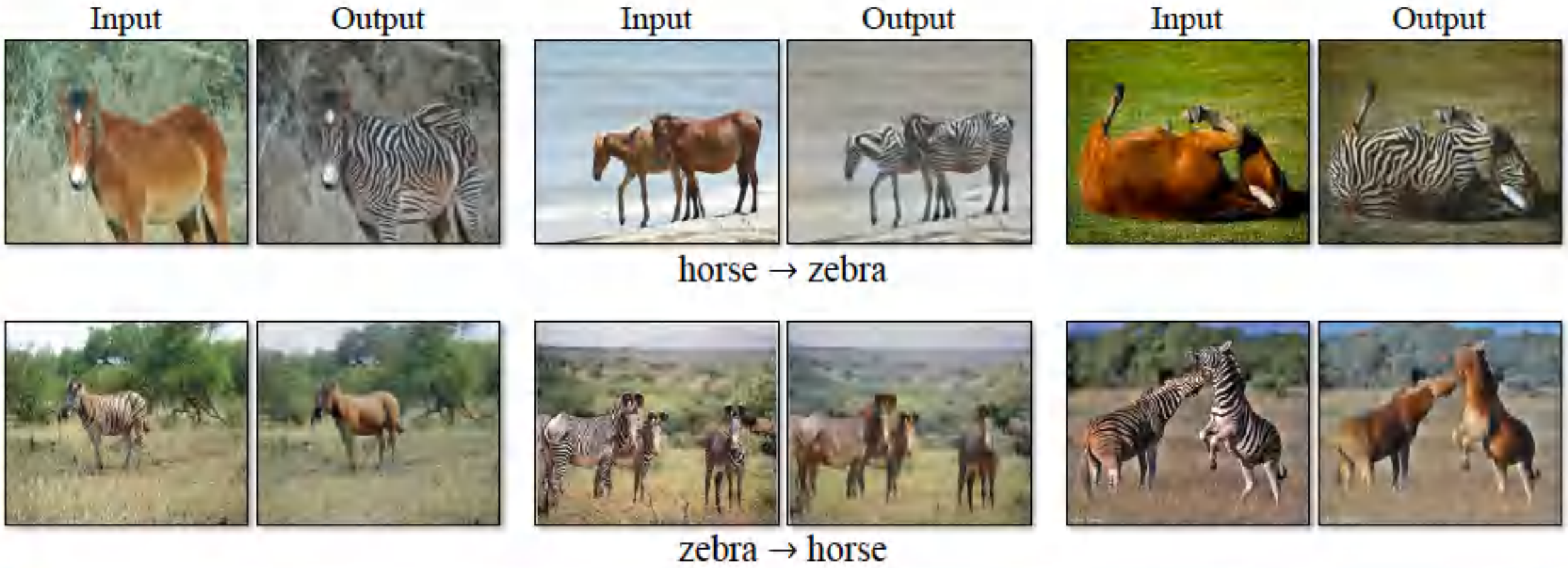


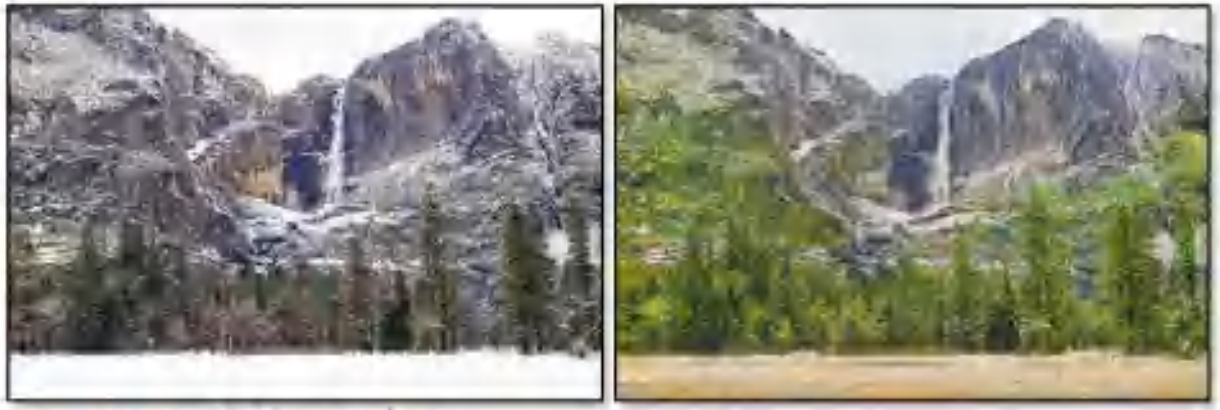
Image to Image Translation



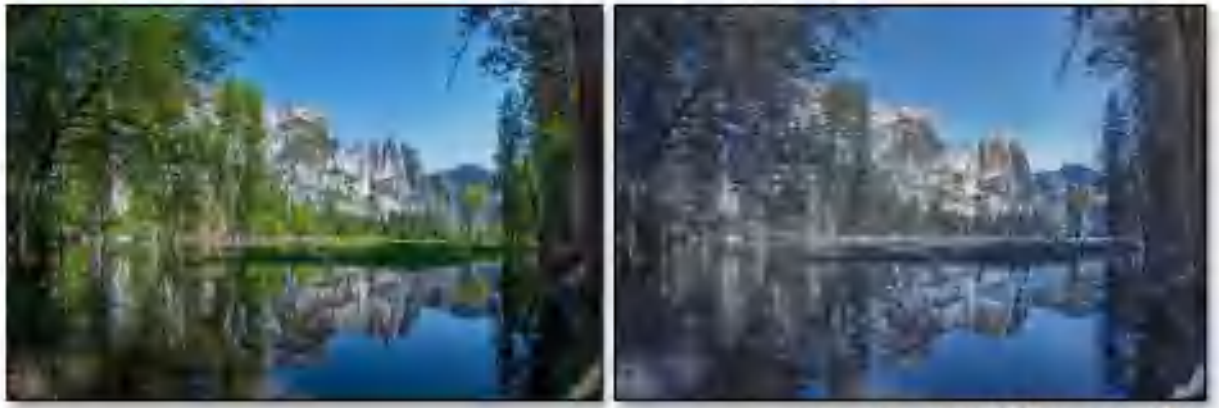
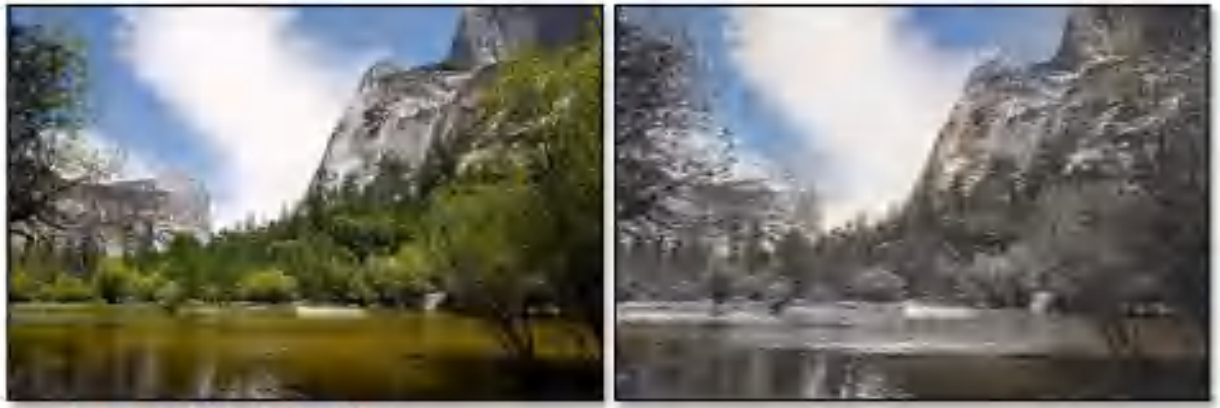
(Isola et al 2016)

(Goodfello


Input
Output
Input
Output
Input
Output
horse → zebra
zebra → horse

winter Yosemite → summer Yosemite

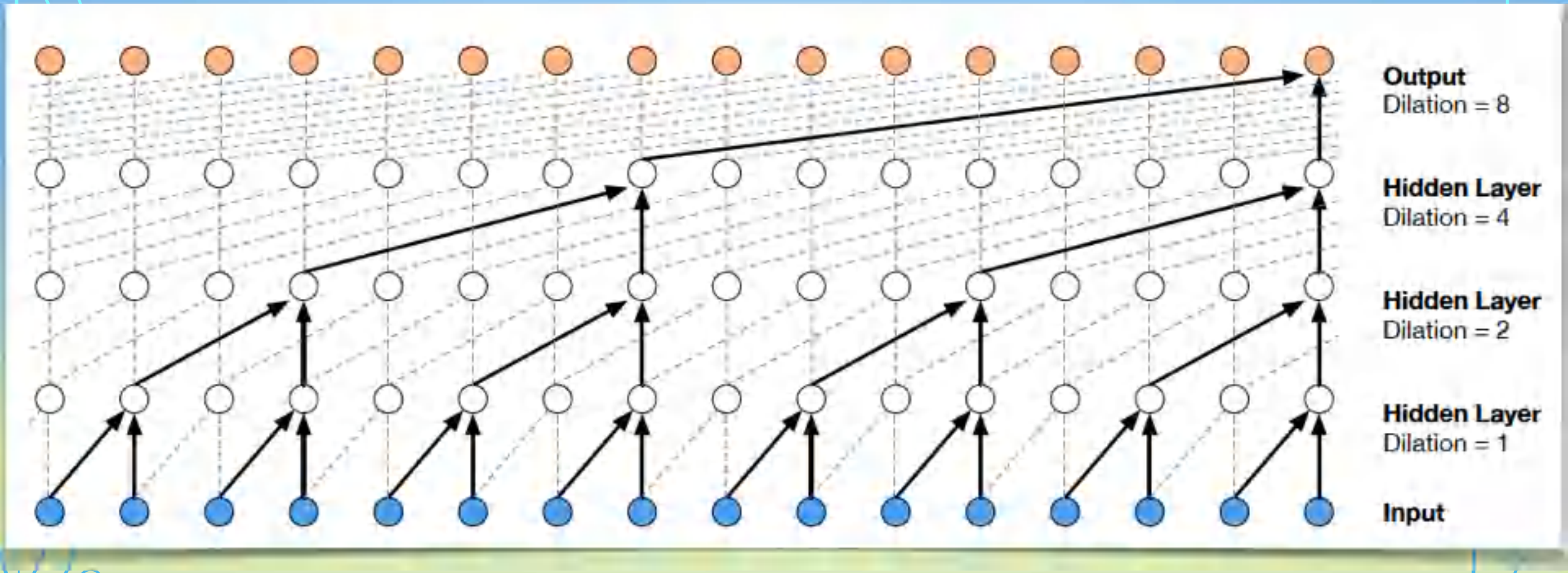



summer Yosemite → winter Yosemite

apple → orange

orange → apple

WaveNet (den Oord, 2016)



FUTURE: LONG WAY TO GO

- Theory: why deep learning works? How? Interpretability
- Capacity: how deep is enough?/Network structure selection
- Manipulate network
- Geometry of loss function
- Deep learning with less data?
- Interdisciplinary fields: statistics, physics, geometry, game theory…
- More exciting applications!
  - Healthcare, industrial process, finance, AI game play, animation…

TOOLS

- Tensorflow by Google
- Torch, PyTorch by Facebook
- Caffe
- Theano
- Keras
- Lasagne
- …

Auto-differentiation

Thanks!

Q&A

zhanxing.zhu@pku.edu.cn