Course number: 80240743

Deep Learning

Xiaolin Hu (胡晓林) & Jun Zhu (朱军)
Dept. of Computer Science and Technology
Tsinghua University

Lecture 1: Introduction

Xiaolin Hu
Department of Computer Science and Technology
Tsinghua University

Outline

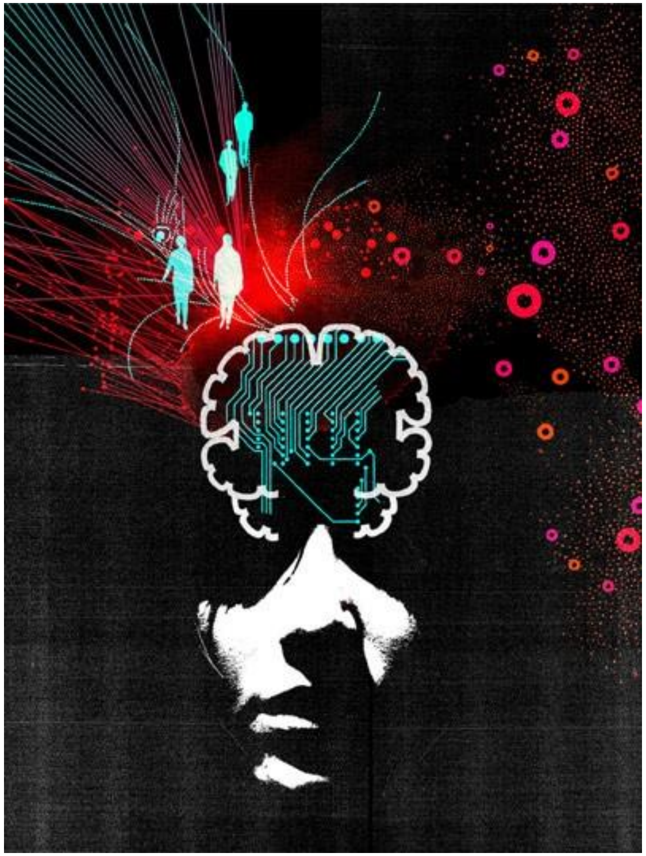
1. General concepts
2. History
3. Applications
4. Risks
5. Summary

MIT Technology Review

10 Breakthrough Technologies 2013

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.



Deep learning in industry

Autonomous car

Face identification

Speech recognition

Web search

…

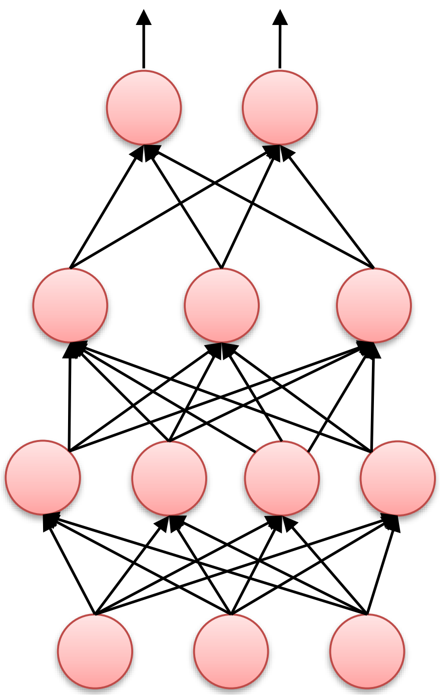
Google
Microsoft
阿里巴巴 Alibaba.com
Baidu 百度
HUAWEI
Tencent 腾讯
NVIDIA
AMD

…

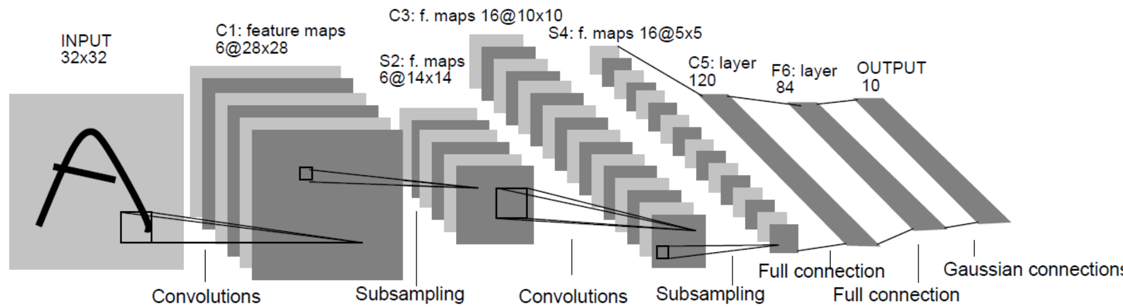
What is deep learning

- Narrow sense: artificial neural networks

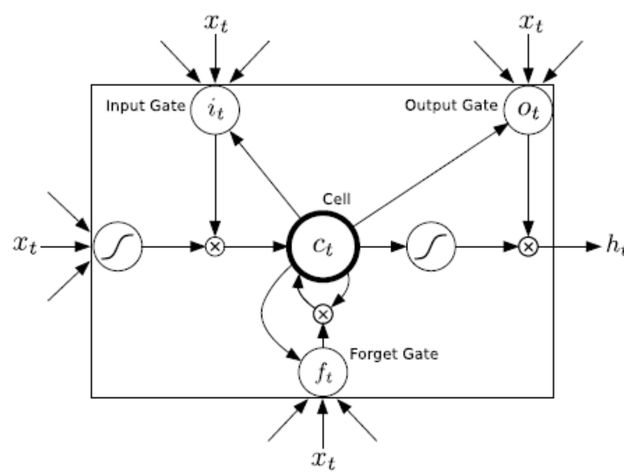
Multilayer Perceptron



Convolutional neural network



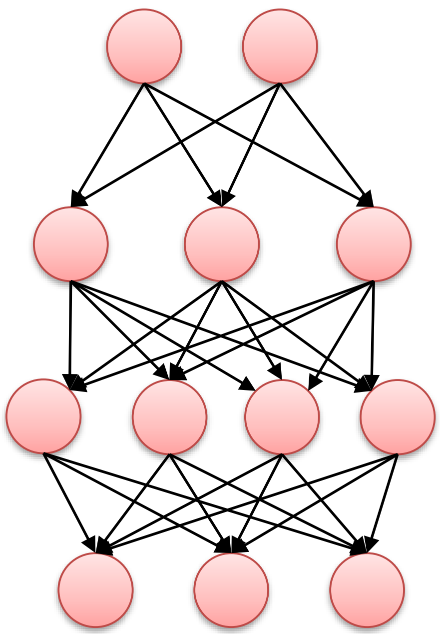
Recurrent neural network



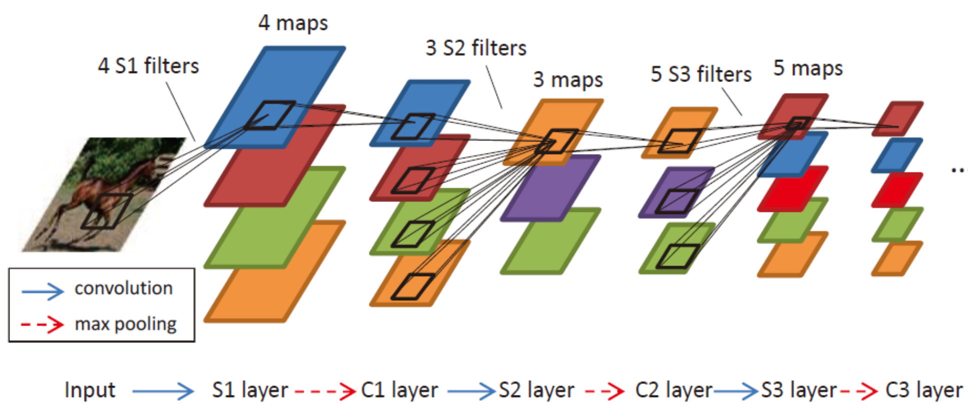
What is deep learning

- Broad sense: hierarchical machine learning models

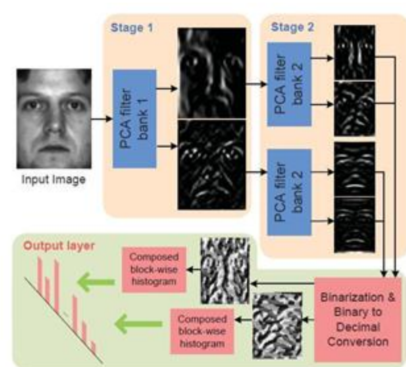
Deep belief network



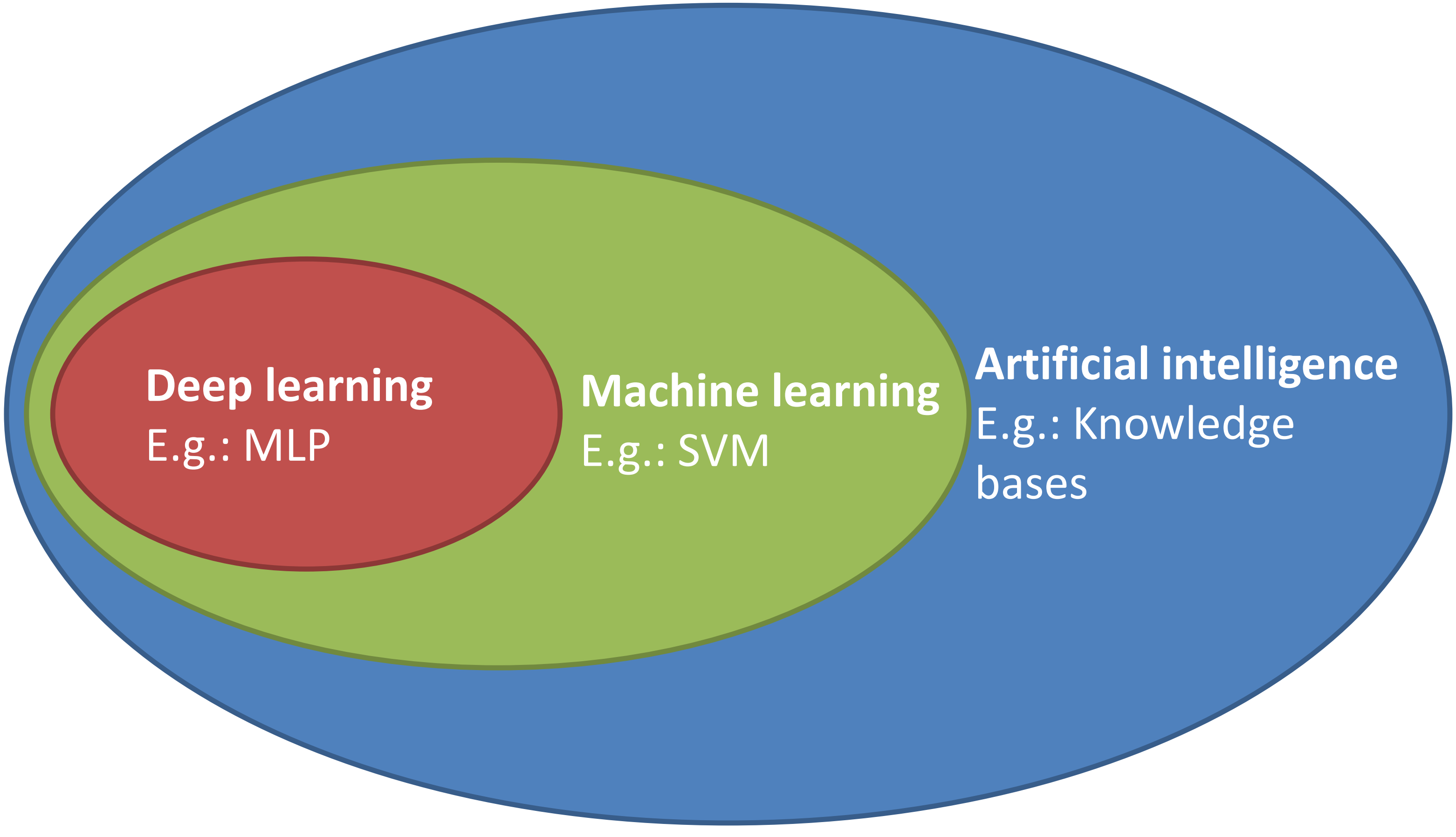
Sparse HMAX (Hu et al., 2014)



PCA net (Chan et al., 2014)



Deep learning, machine learning and AI



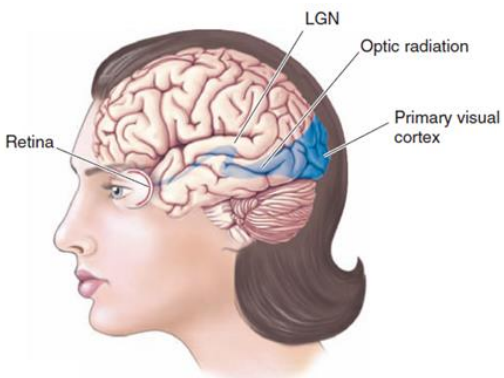
Why go deep?

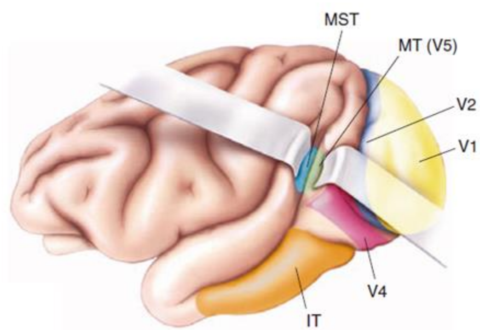
Why go deep?

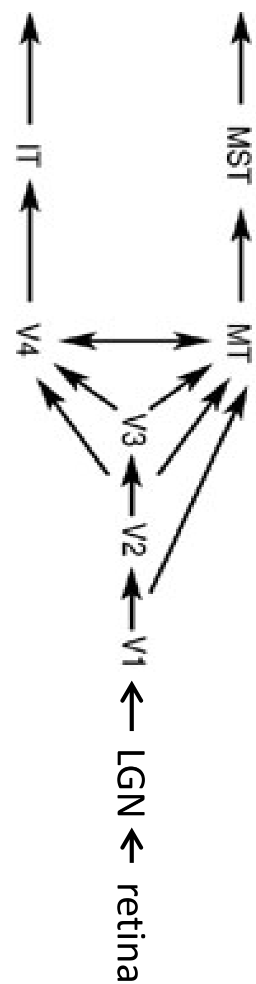
- Data are often high-dimensional.
- There is a huge amount of structure in the data, but the structure is too complicated to be represented by a simple model.
- Insufficient depth can require more computational elements than architectures whose depth matches the task.
- Deep nets provide simpler but more descriptive model of many problems.

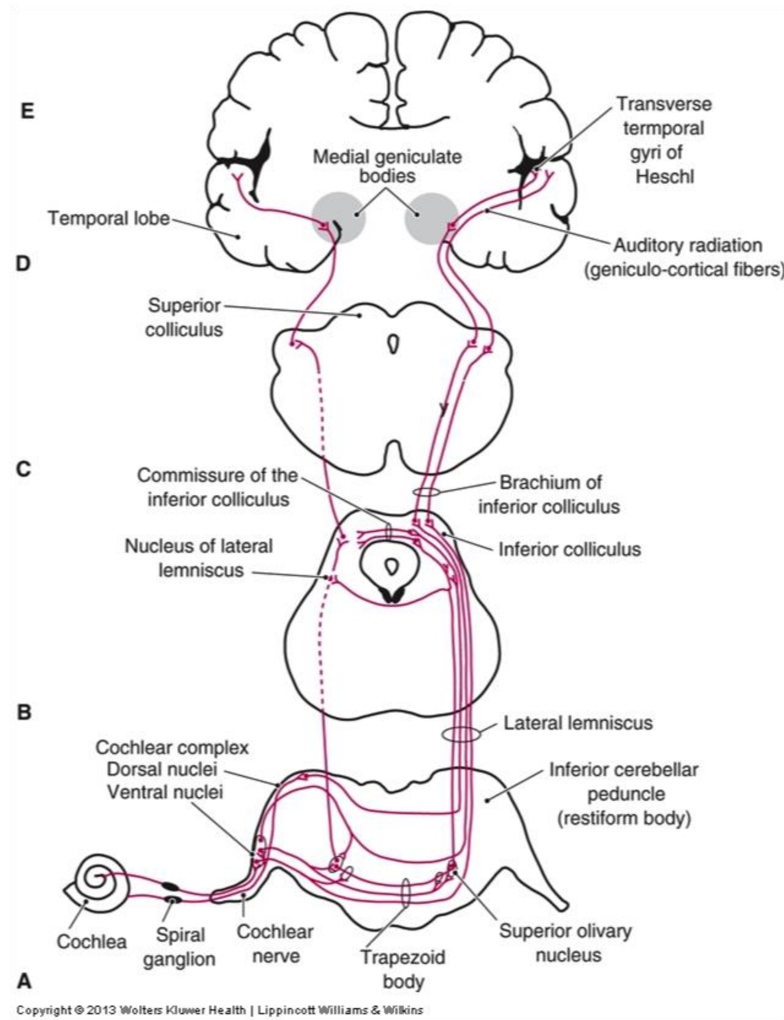
-By Geoffery Hinton

Hierarchical structures in the brain

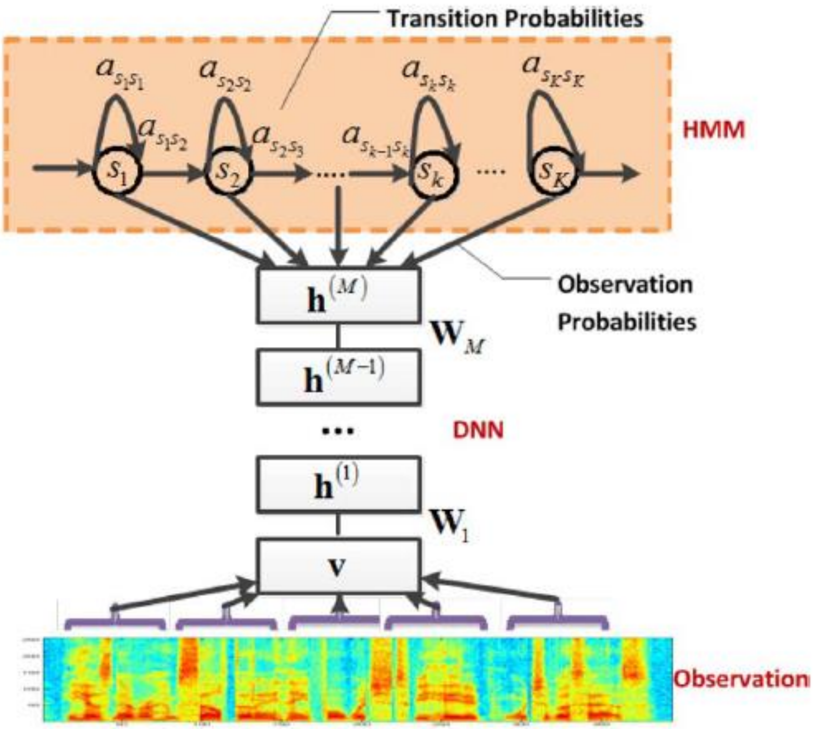




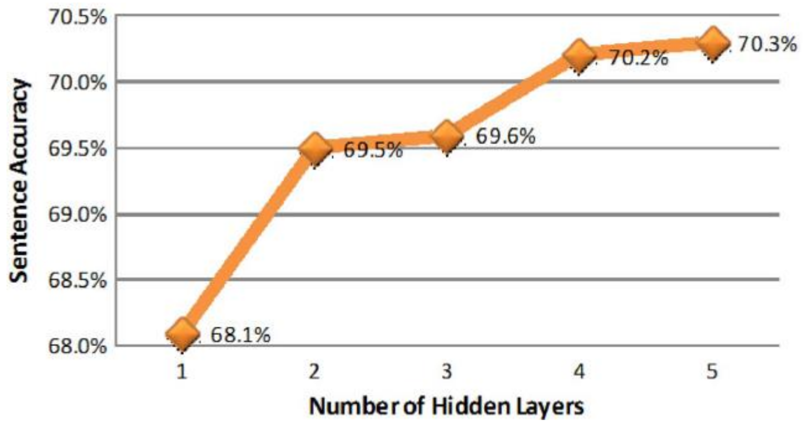




When did deep learning become popular (1)

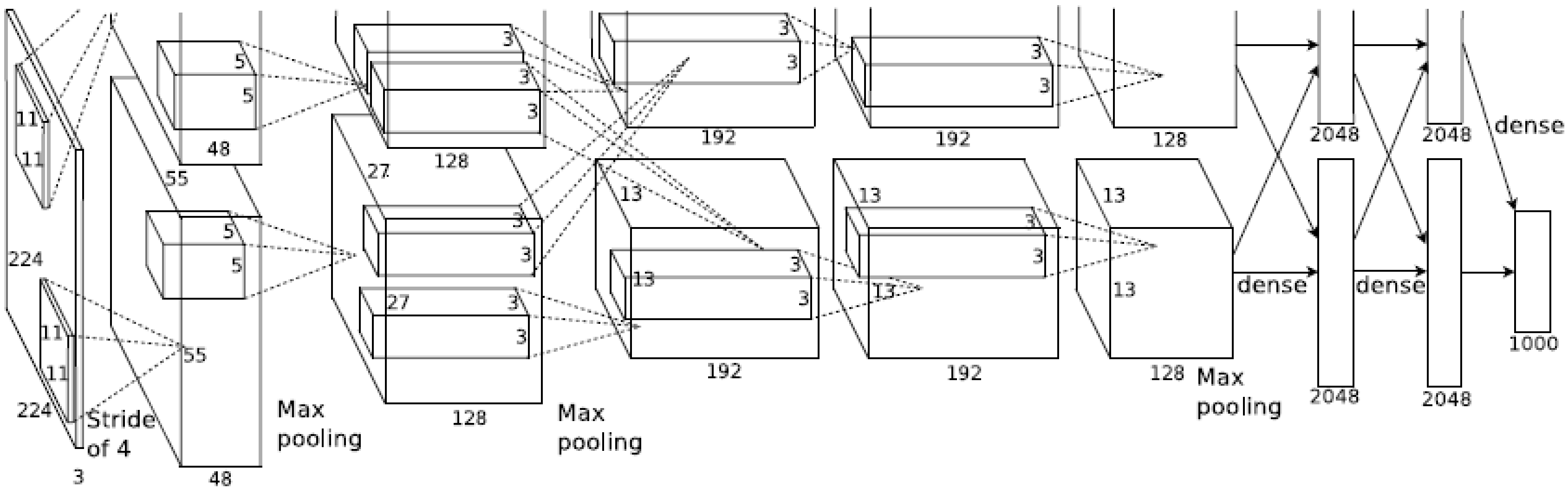


Dahl, Yu, Deng, Acero, IEEE TASLP, 2012



http://v.youku.com/v_show/id_XNDc0MDY4ODI0.html

When did deep learning become popular (2)

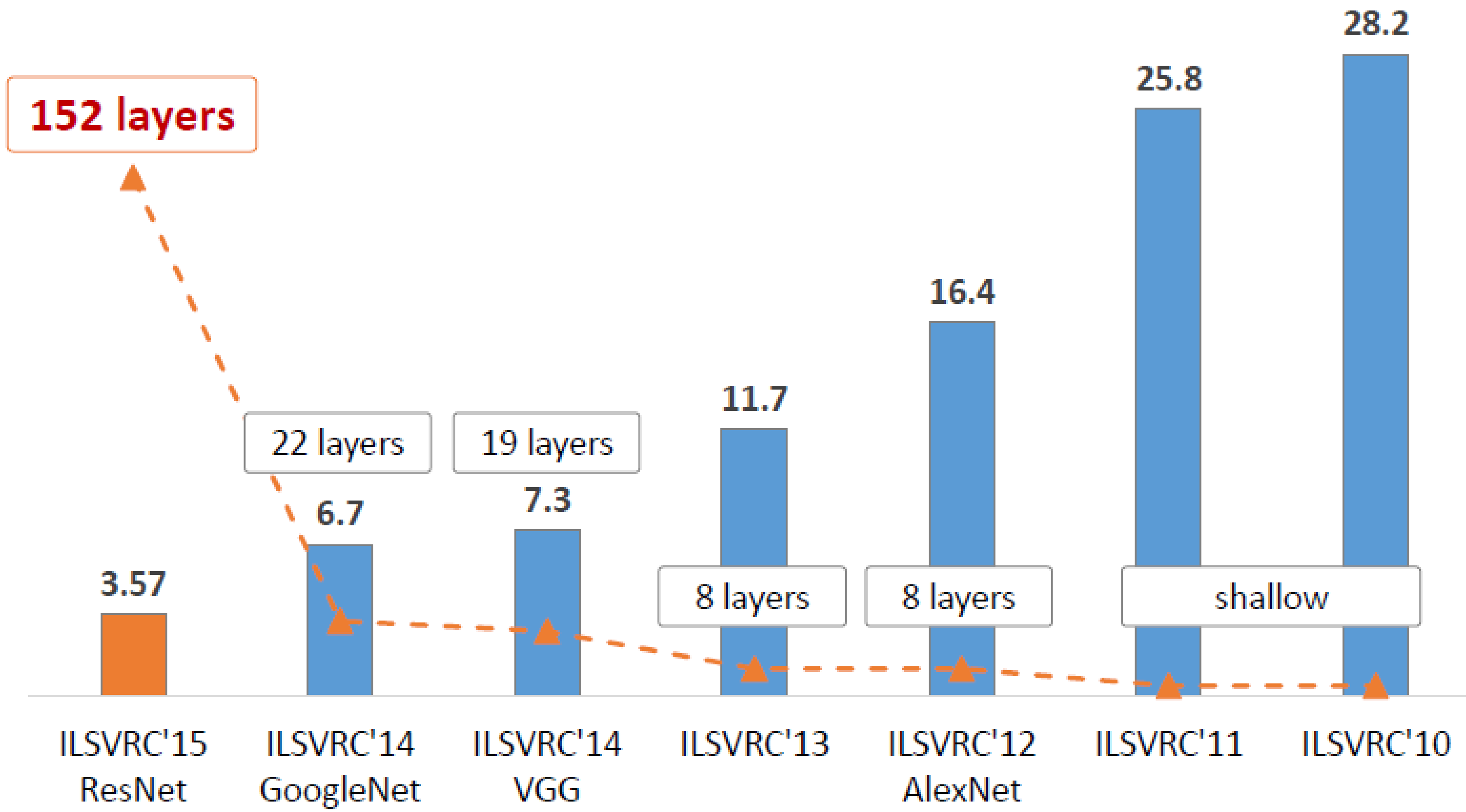


- Classify images in the ImageNet LSVRC-2010 contest into 1000 different classes
- Trained on 1.2M images
- 60 million parameters

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Krizhevsky, Sutskever and Hinton, NIPS, 2012

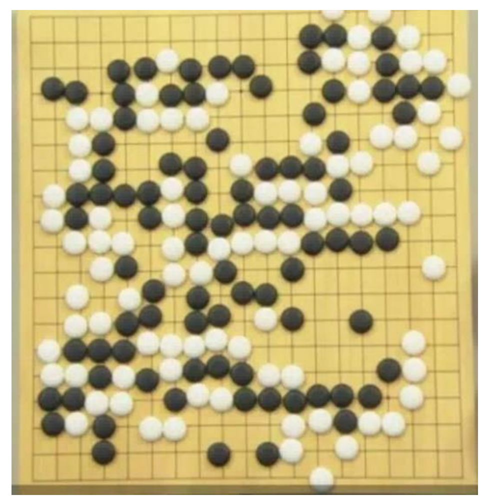
Revolution of depth



Slide credit: Kaiming He

The world is astonished by AlphaGo

28 January 2016

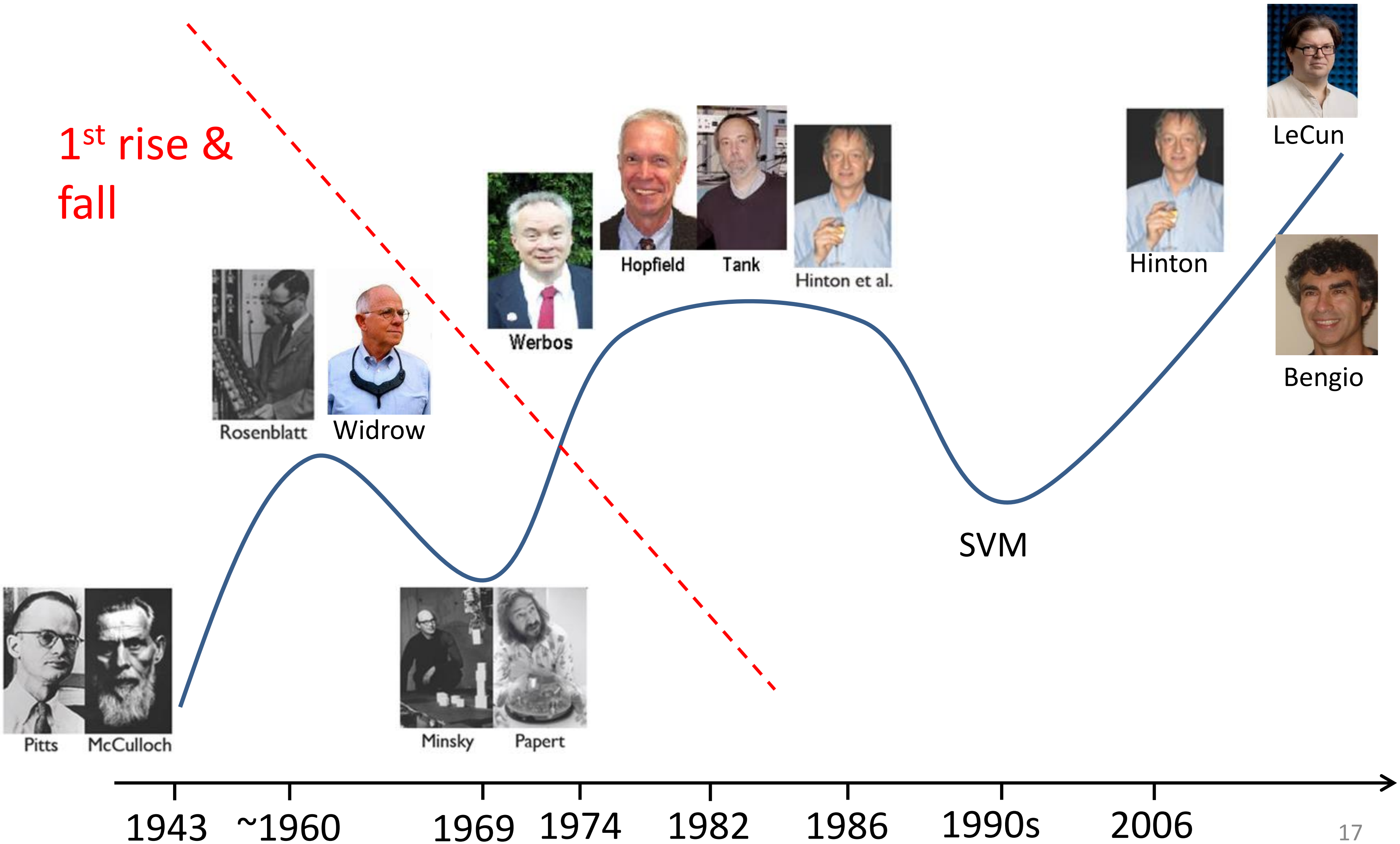


AlphaGo: black

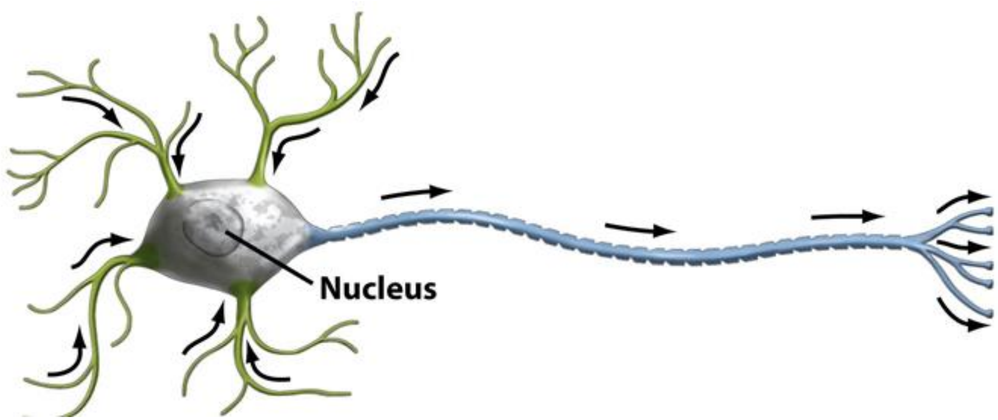
Outline

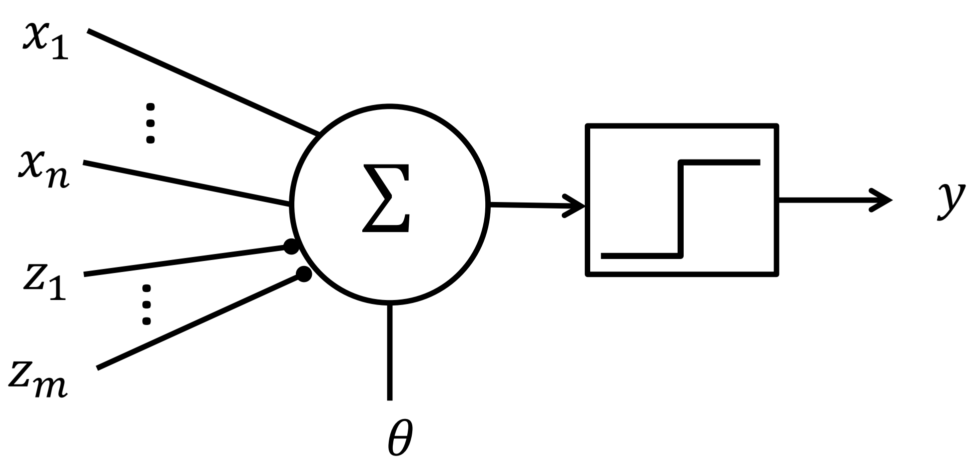
1. General concepts
2. History
3. Applications
4. Risks
5. Summary

History of deep learning

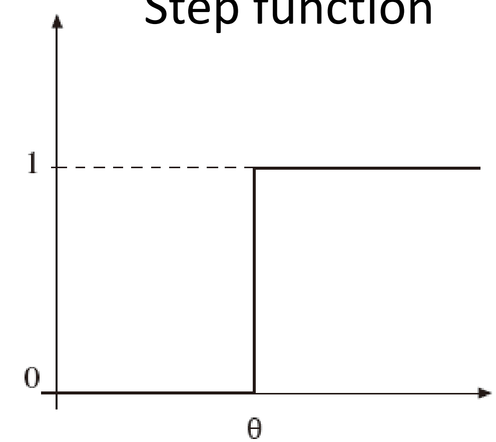


Threshold Logic Unit (TLU)



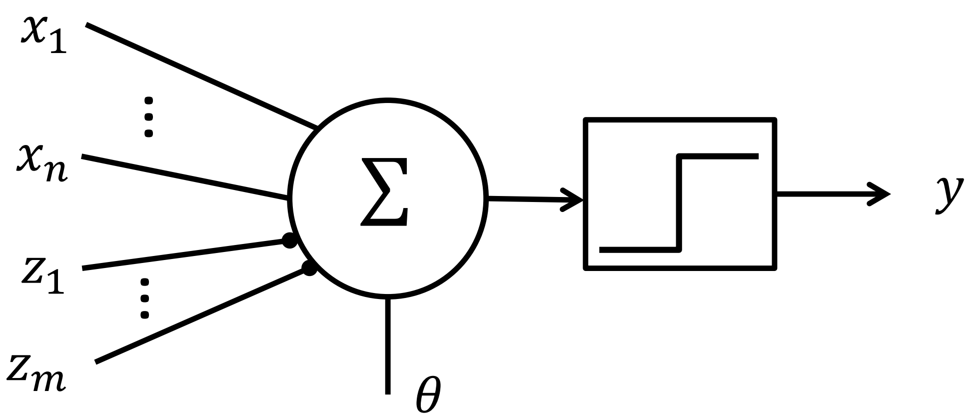


Step function



- Excitatory input x_i
- Inhibitory input z_i
- Binary output y_i
- Threshold θ

McCulloch–Pitts unit (M-P unit)



- If at least one of $z_1, z_2, \ldots, z_m$ is 1, the unit is inhibited and $y = 0$
- Otherwise the total excitation $T = \sum_{i=1}^{n} x_i$ is computed and compared with the threshold θ of the unit (if $n = 0$ then $x = 0$)
  – If $T \geq \theta$ the unit fires a 1
  – If $T < \theta$ the result is 0.
- The MP unit can be inactivated by a single inhibitory signal, as is the case with some real neurons

Rojas: Neural Networks, Springer-Verlag, Berlin, 1996

Boolean function

- A Boolean function $f: \{0,1\}^n \rightarrow \{0,1\}$
- It can be represented by a table

Input	Output
1	0
0	1

NOT

Input	Output
(0, 1)	0
(1, 0)	0
(1, 1)	1
(0, 0)	0

AND

Input	Output
(0, 1)	1
(1, 0)	1
(1, 1)	1
(0, 0)	0

OR

Boolean function

- A Boolean function $f: \{0,1\}^n \to \{0,1\}$
- It can be represented by a table

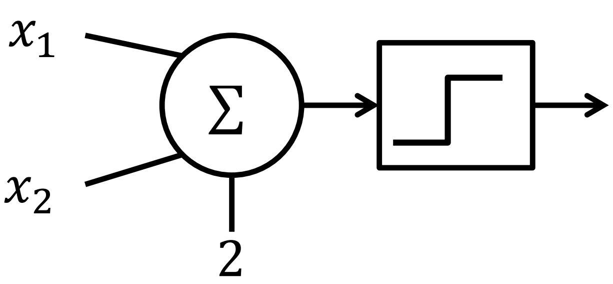
Input	Output
(0, 1, 1, 1)	1
(0, 0, 1, 1)	1
(1, 0, 0, 1)	1
All others	0

Synthesis of Boolean functions

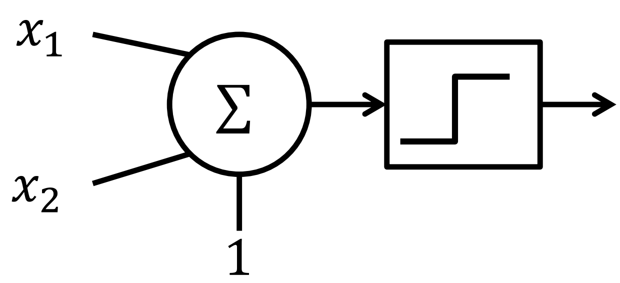
Boolean function: $\{0,1\}^n \rightarrow \{0,1\}$

- Conjunction, disjunction, negation

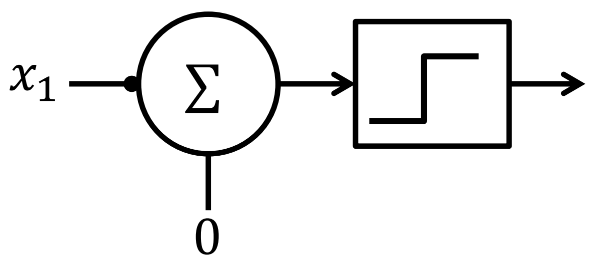
AND



OR



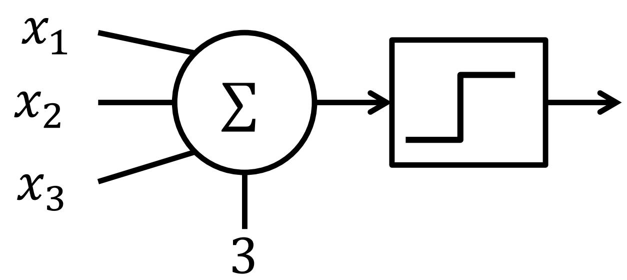
NOT



Multiple Choice(single) Points: 1

Setting

What function does this unit implement?

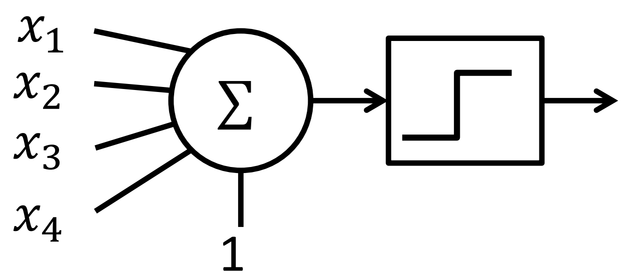


A AND (conjunction)

B OR (disjunction)

Submit

What function does this unit implement?



A AND (conjunction)

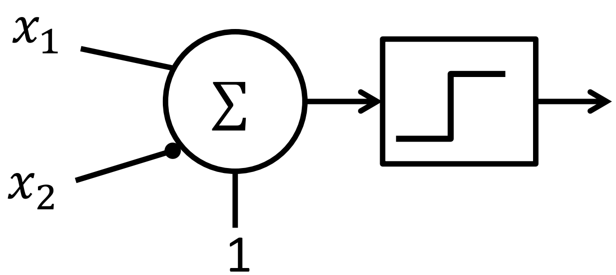
B OR (disjunction)

Submit

Multiple Choice(single) Points: 1

Setting

What function does this unit implement?



A x_1 OR $\neg x_2$

B x_1 AND $\neg x_2$

C x_1 AND x_2

("$\neg$" means NOT)

Submit

Can any logical function be implemented by MP units?

- Every logical function of n variables can be written in tabular form

Consider an example with $n = 3$

input vectors	F
(0,0,1)	1
(0,1,0)	1
all others	0

Method

Suppose there are K rows that have results 1

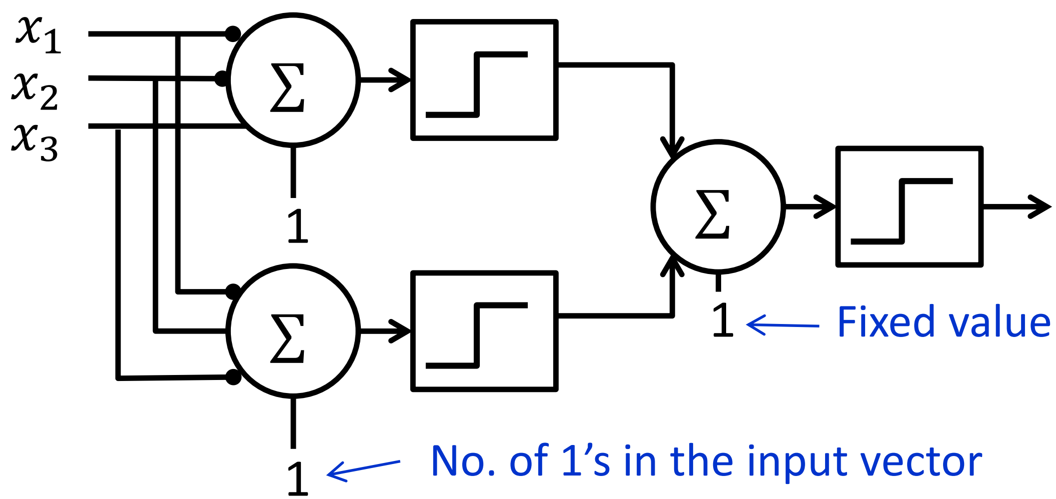
① Use a M-P unit to represent n values which lead to the result of 1

② Use a disjunction unit to connect the K M-P unit

Constructive synthesis

- Consider the previous example

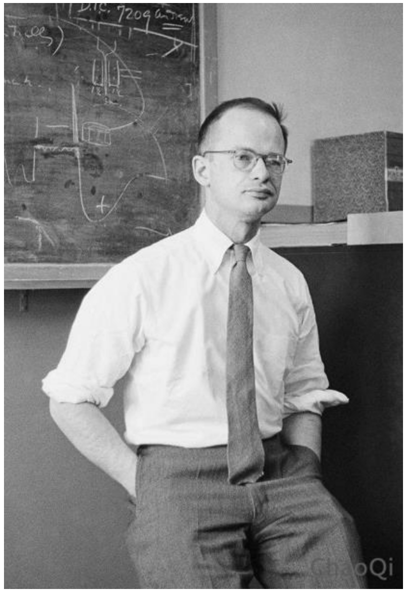
input vectors	F
(0,0,1)	1
(0,1,0)	1
all others	0



Proposition. Any logical function $F : \{0, 1\}^n \to \{0, 1\}$ can be computed with a M-P network of two layers.

Walter Pitts 1923-1969

- Born in a tough family in Prohibition-era Detroit, where his father, a boiler-maker, had no trouble raising his fists to get his way
- In 1935, he read *Principia Mathematica*, a three-volume tome written by Bertrand Russell and Alfred Whitehead, which attempted to reduce all of mathematics to pure logic

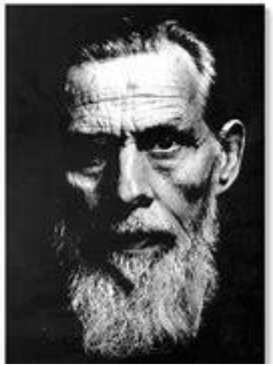


1923-1969

- He found several mistakes and wrote to Russel
- In 1938, when he heard that Russell would be visiting the University of Chicago, he ran away from home and headed for Illinois. He never saw his family again

Work with Warren McCulloch

- In 1923, the year that Walter Pitts was born, a 25-year-old Warren McCulloch was also digesting the Principia
- McCulloch was born into a well-to-do East Coast family of lawyers, doctors, theologians, and engineers



1898-1969

- Working together, they would create the first mechanistic theory of the mind, the first computational approach to neuroscience, the logical design of modern computers, and the pillars of artificial intelligence

A Logical Calculus of Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics, 1943

Start of ANN

Work with Norbert Wienner

- In 1943, Pitts became a PhD student of Wienner at MIT
- Wienner realized that it ought to be possible for Pitts' neural networks to be implemented in man-made machines, ushering in his dream of a cybernetic revolution

Norbert Wienner

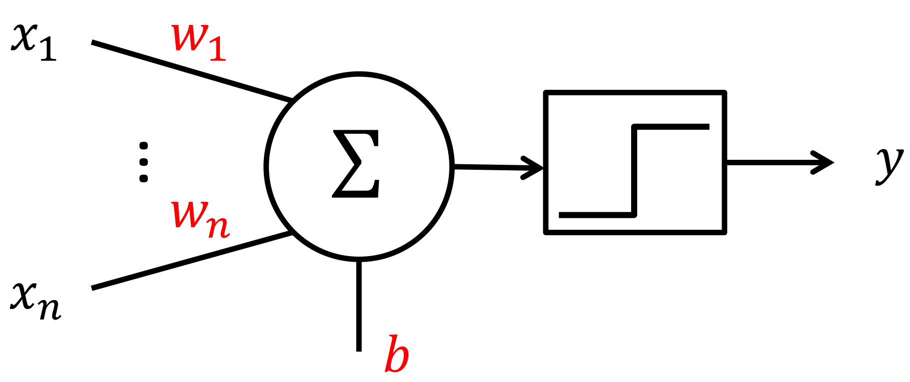
von Neumann

- The beginnings of the group who would become known as the *cyberneticians* was formed, with Wiener, Pitts, McCulloch, Lettvin, and von Neumann its core.
- von Neumann suggested modeling the computer after Pitts and McCulloch's neural networks

Collapse of logical brain idea

- Wiener’s wife invented a story. She sat Wiener down and informed him that when their daughter, Barbara, had stayed at McCulloch’s house in Chicago, several of “his boys” had seduced her.
- Wiener never spoke to Pitts again. And he never told him why
- Experiments with frog’s eyes. “The eye speaks to the brain in a language already highly organized and interpreted,” they reported in the now-seminal paper “What the Frog’s Eye Tells the Frog’s Brain,” published in 1959
- The results shook Pitts’ worldview to its core
- In 1969 Pitts died alone in a boarding house in Cambridge. Four months later, McCulloch passed away

Perceptron



$$y = \begin{cases} 1 & \boldsymbol{w}^\top \boldsymbol{x} + b > 0 \\ 0 & \text{else} \end{cases}$$

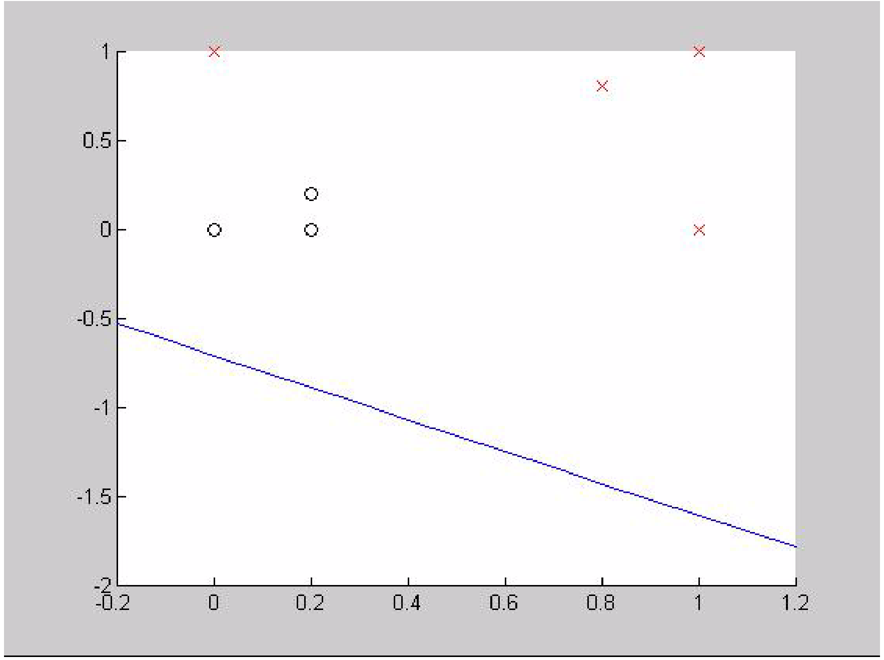
Could also be -1

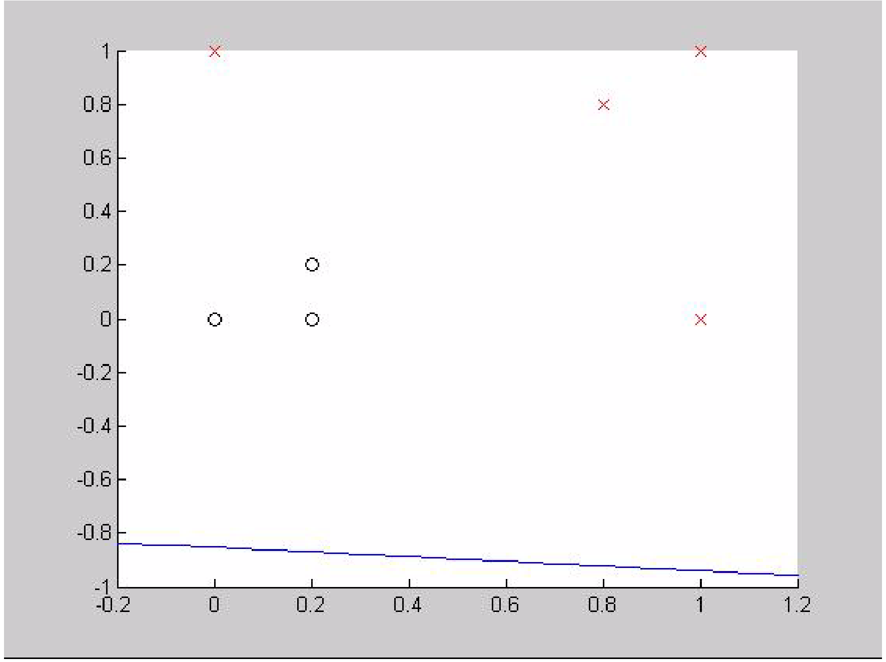
- Add weights to the input connections of the M-P unit
- Propose a supervised learning algorithm: For each data points $\boldsymbol{x}^{(j)} \in R^m$ and the corresponding labels $t^{(j)}$
  - Calculate the actual output $y^{(j)}$
  - Update the weights: $\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{old} + \eta\big(t^{(j)} - y^{(j)}\big)\boldsymbol{x}^{(j)}$;

$$b^{\text{new}} = b^{old} + \eta\big(t^{(j)} - y^{(j)}\big)$$

where $\eta > 0$ is the learning rate

Example



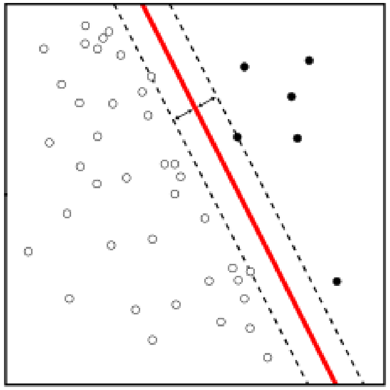


From two different sets of initial weights

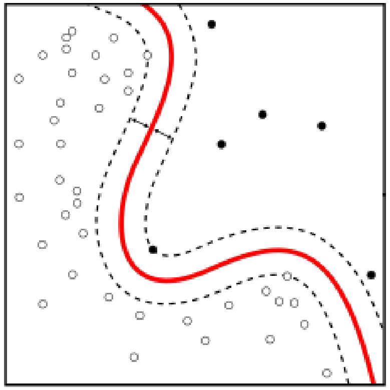
Convergence

Proposition 4: *If the training set is linearly separable, then the perceptron is guaranteed to converge. Furthermore, there is an upper bound on the number of times the perceptron will adjust its weights during the training.*

Proof. See (Novikoff, 1962)



linearly separable

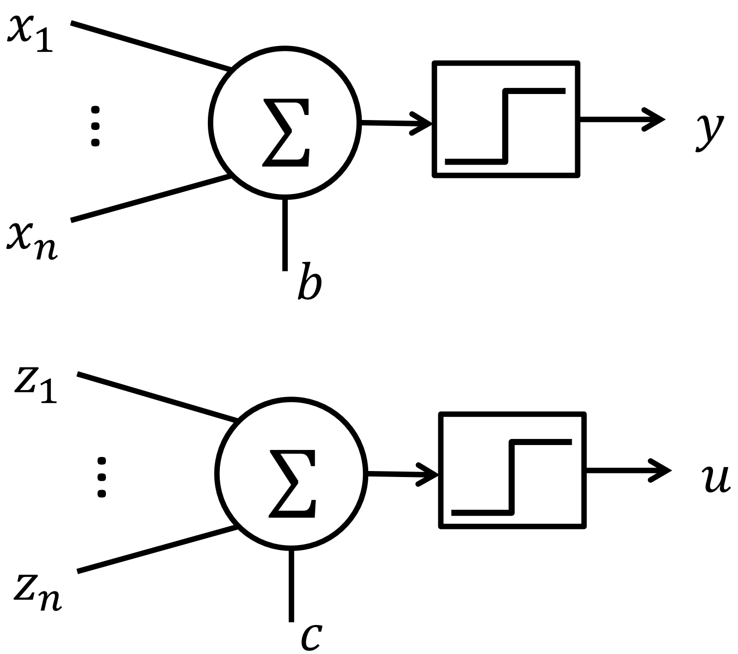


linearly non-separable

Percepton cannot solve linearly non-separable problems (Minsky & Papert, 1969)

Multiple Perceptrons in one layer

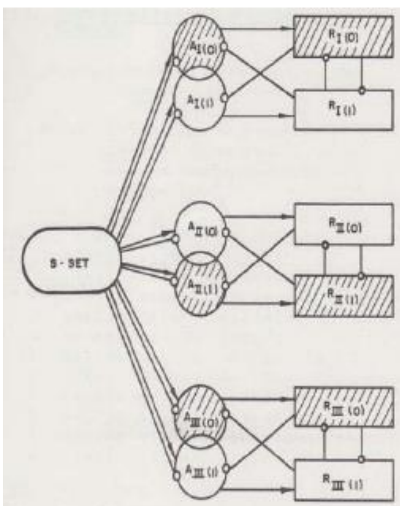
- When multiple Perceptrons are combined, each output neuron operates independently of all the others; thus, learning each output can be considered in isolation



No good training method for multi-layer Perceptrons

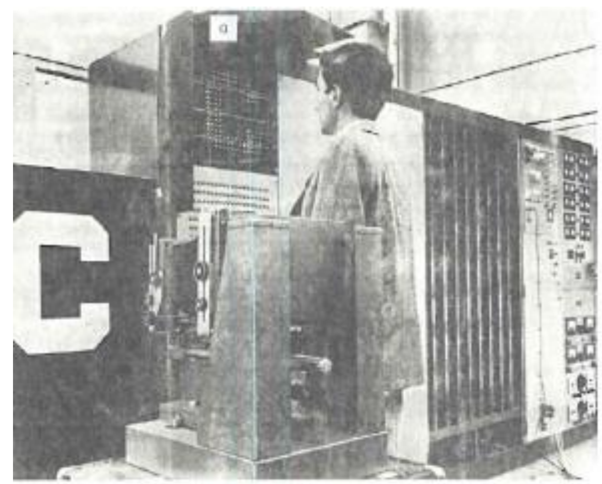
Perceptron

- Definition(p. 83, Neurodynamics): A perceptron is a network of S, A, and R units with a variable interaction matrix V which depends on the sequence of past activity states of the network.



NOTE:
THE SHADING SHOWS THE ASSOCIATION SETS AND R-UNITS WHICH WOULD BE INHIBITED WHEN THE RESPONSE (0) IS ACTIVE.

FIGURE 3
ORGANIZATION OF A PERCEPTRON WITH THREE BINARY RESPONSE SETS



Mark I: 400 S-units, 512 A-units, 8 R-units

By George Nagy in 2011, Rosenblatt‘s PhD student

Frank Rosenblatt

1928 -1971

- Bronx High School of Science
- Cornell student (1946 –1956)
- Cornell Aeronautical Laboratories
- Cognitive Systems Research Program
- Neurobiology

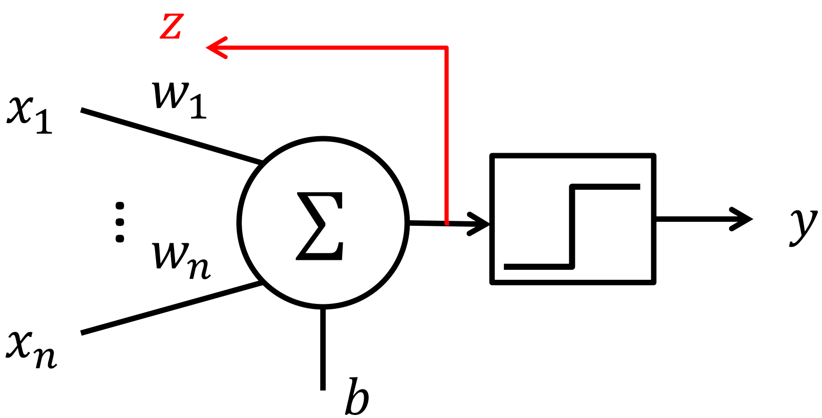
- *political campaigns in NY, NH, VT, CA*
- *music (piano, composition)*
- *astronomy and cosmology*
- *mountain climbing and sailing*

1950 Social Psychology

The gravestone of Frank Rosenblatt, Brooktondale, NY.

By George Nagy in 2011, Rosenblatt's PhD student

ADALINE



$$y = \begin{cases} 1 & \boldsymbol{w}^\top \boldsymbol{x} + b > 0.5 \\ 0 & \text{else} \end{cases}$$

Or

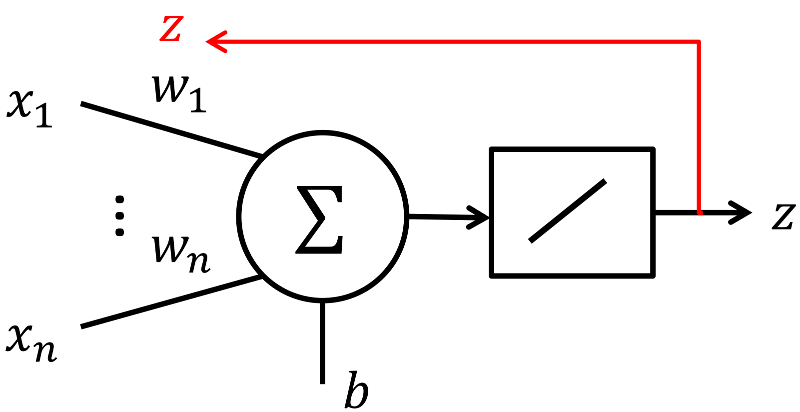
$$y = \begin{cases} 1 & \boldsymbol{w}^\top \boldsymbol{x} + b > 0 \\ -1 & \text{else} \end{cases}$$

- Same architecture as Perceptron; different training algorithm
  - $z = \boldsymbol{w}^\top \boldsymbol{x} + b$ instead of y is used to adjust the weights and bias
- Minimize MSE $E = \frac{1}{N}\sum_j \left(t^{(j)} - z^{(j)}\right)^2$. The learning algorithm:

$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{old} + \eta\left(t^{(j)} - z^{(j)}\right)\boldsymbol{x}^{(j)}$$
$$b^{\text{new}} = b^{old} + \eta\left(t^{(j)} - z^{(j)}\right)$$

  where $\eta > 0$ is the learning rate
- Different names: LMS rule, Delta rule, Widrow-Hoff rule, actually SGD

Another view



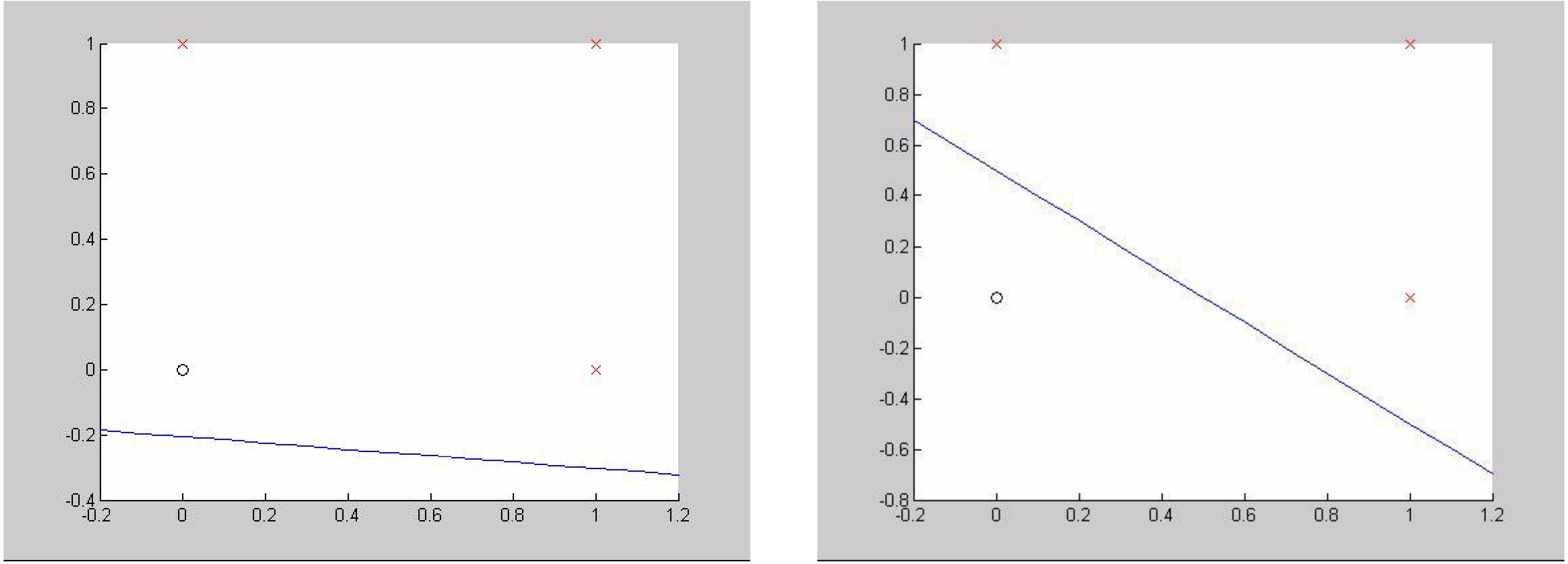
$$z = \boldsymbol{w}^\top \boldsymbol{x} + b$$

$$y = \begin{cases} 1 & z > 0 \\ -1 & \text{else} \end{cases}$$

- There is a linear activation function for the variable z
  - This is where the name *Adaptive Linear Neuron* comes
- The step function is only used for output y and the output is not involved in the learning process

Example

Training data: x1=(0, 0), t1=-1; x2=(0, 1), t2=1;
x3=(1, 0), t3=1; x4=(1, 1), t4=1;



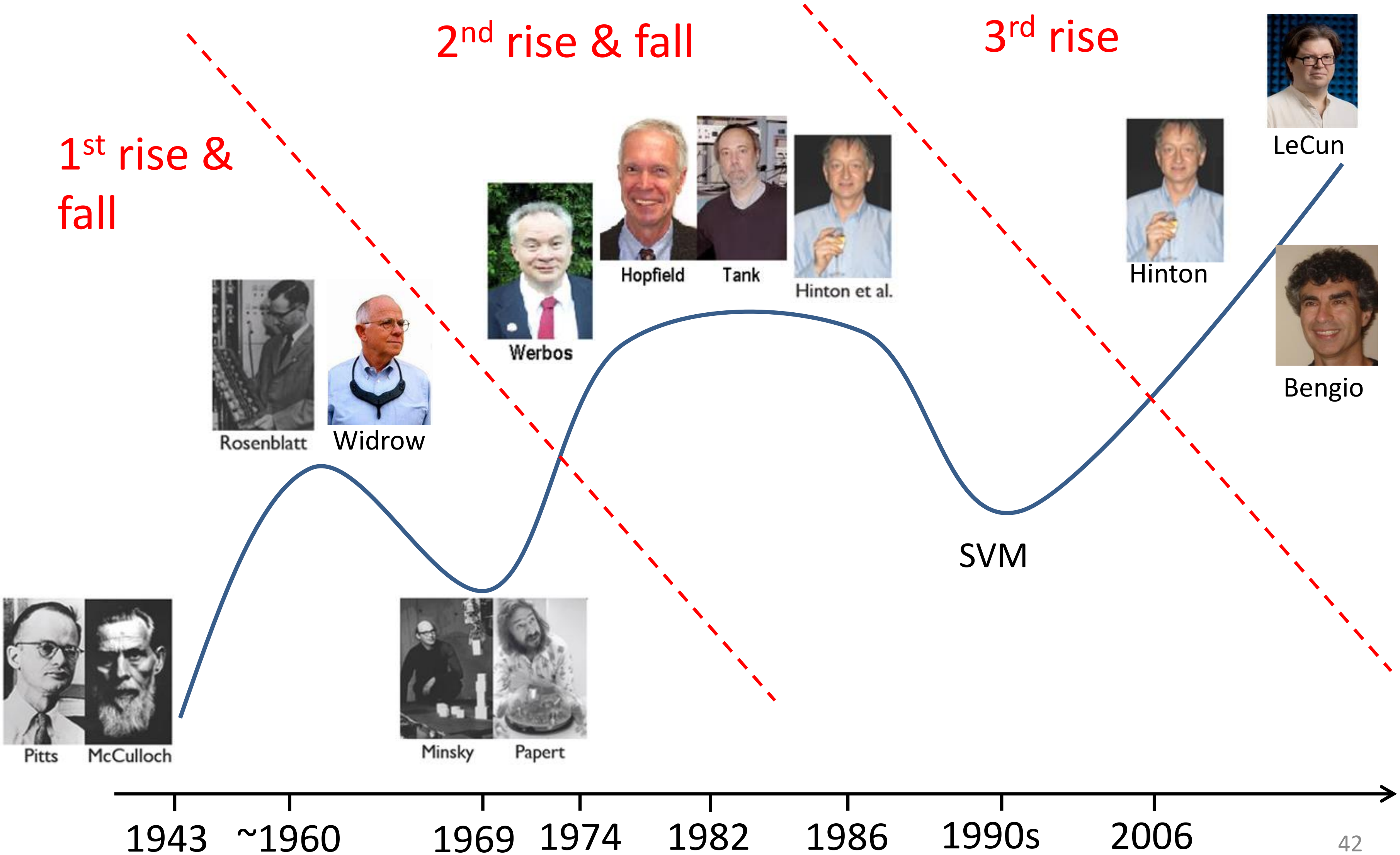
From two different sets of initial weights

Bernard Widrow

Born in 1929

- Widrow and his doctoral student Ted Hoff invented the least mean squares filter (LMS) adaptive algorithm in 1960
- The LMS algorithm led to the ADALINE and MADALINE and to the backpropagation technique
- LMS algorithm minimizes the mean squared error (MSE), and is a stochastic gradient descent (SGD) method
  - It was proposed for signal processing and achieved great success in that field, but not so successful in training muli-layer neural networks
- In early 1960 Widrow turned to study signal processing, and returned to neural networks in 1980s

The 2nd rise and fall



The 3rd rise

Turing award 2018

Geoffrey Hinton、Yann LeCun、Yoshua Bengio

Jürgen Schmidhuber

Established in 1982, CIFAR is a Canadian-based, international research institute with nearly 400 fellows, scholars and advisors from 18 countries.

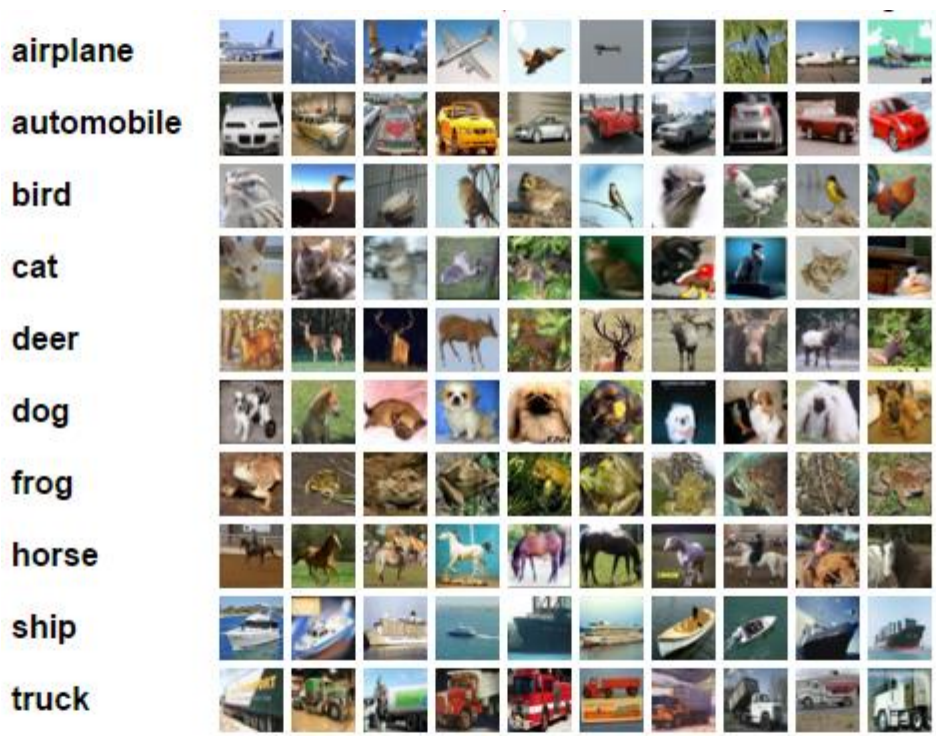
Hinton’s interview by Ng

Outline

1. General concepts
2. History
3. Applications
4. Risks
5. Summary

General object classification

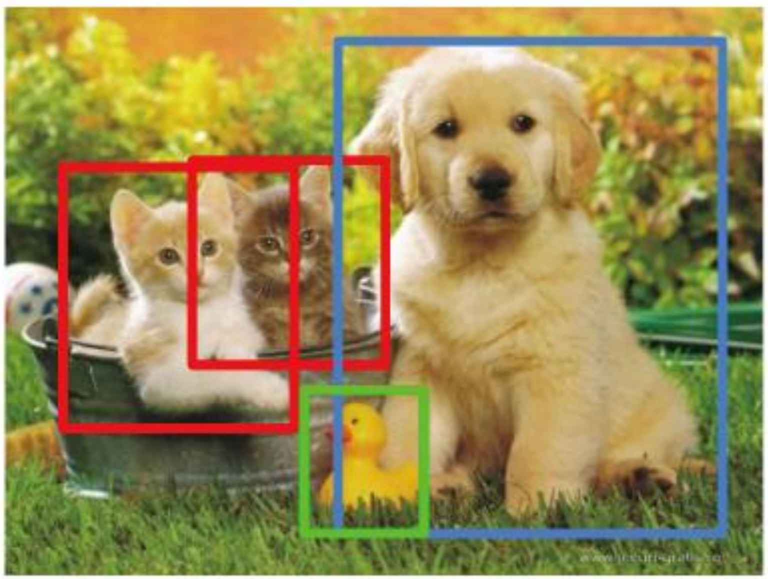
CIFAR-10 & CIFAR100 datasets
- 50,000 training, 10,000 test
- 32x32 RGB imgs



ILSVRC2012 dataset
- ~128M training
- 50,000 validation
- 100,000 test

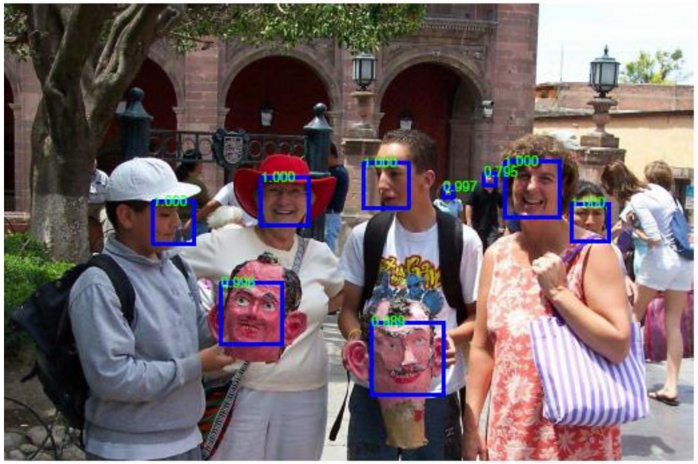
Specific object classification

General object detection

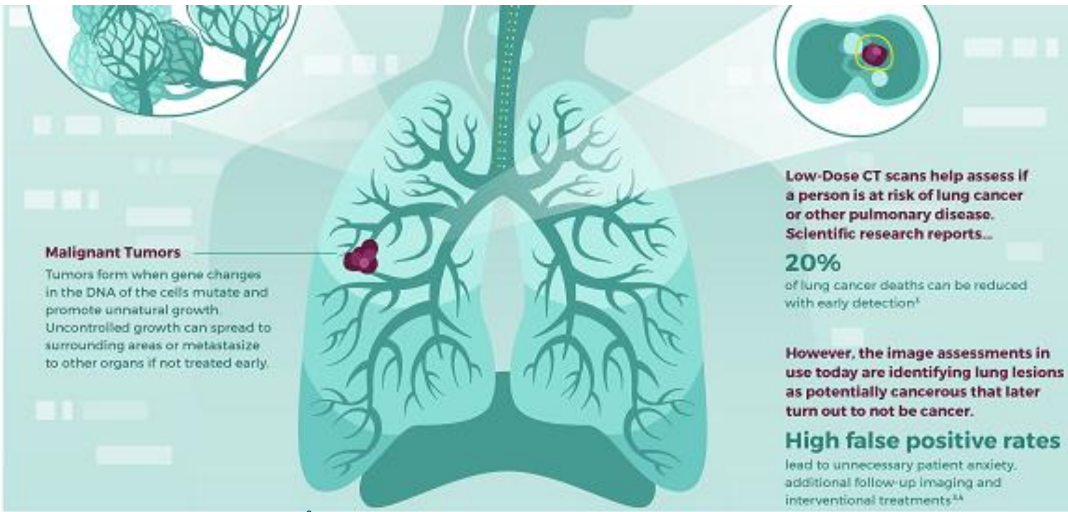


DOG, (x, y, w, h)
CAT, (x, y, w, h)
CAT, (x, y, w, h)
DUCK (x, y, w, h)

Specific object detection



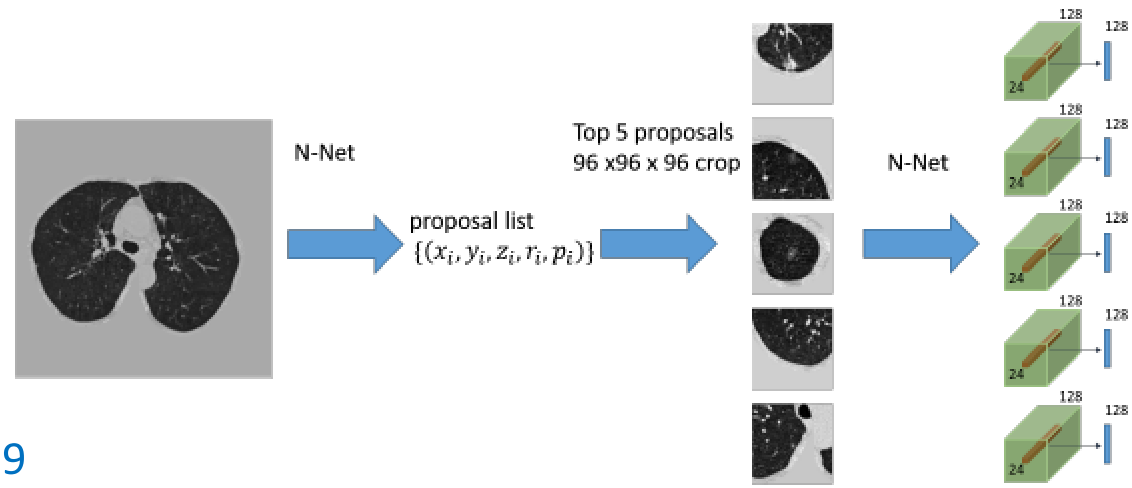
Medical image analysis



DATA SCIENCE BOWL

Passion. Curiosity. Purpose.

A 500,000 USD solution!



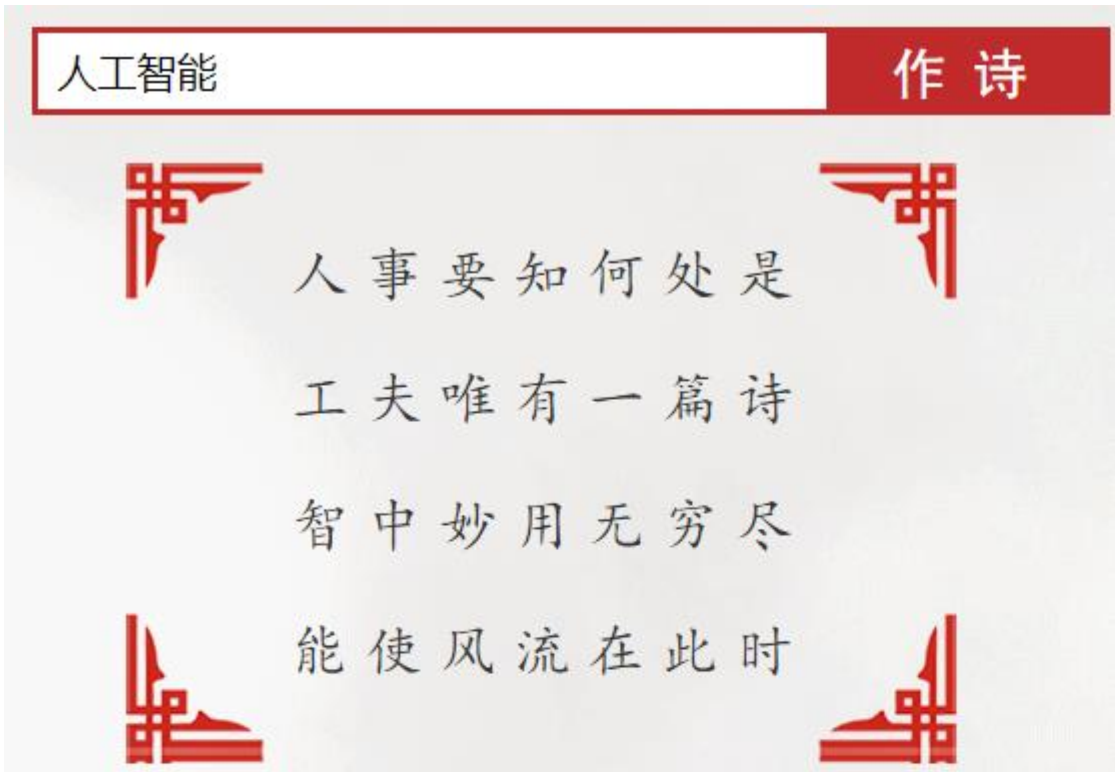
Liao et al., IEEE TNNLS 2019

Image generation

64*64

Generated bedrooms after five epochs of training of a GAN

Chinese poem generation



http://jiuge.thunlp.org/

Music generation

Discussion

- What interesting applications do you know?

  [AI Spots Mysterious Signals Coming from Deep in Space]

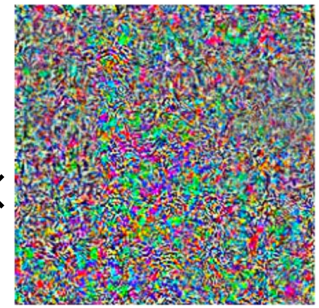
- What problem do you mostly want to solve with deep learning?

Outline

1. General concepts
2. History
3. Applications
4. Risks
5. Summary

Adversarial examples

+ .008×



=

ResNet： Kangaroo: 99.31% Airplane: 99.99%

Deepfake

Facebook is launching a project with $10M
https://ai.facebook.com/blog/deepfake-detection-challenge/

AI weapon


StratoEnergetics LIVE STREAM
http://www.stratoenergetics.com
Buenos Aires Event
TV Truck 02
YOUKU

Discussion

- How to prevent abuse of AI including deep learning?

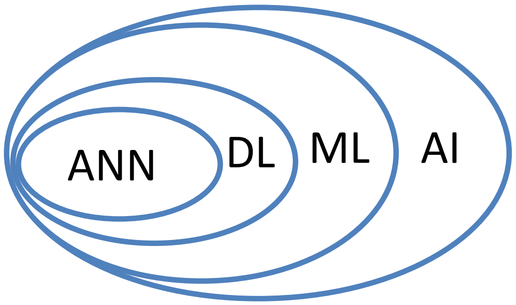
Outline

1. General concepts
2. History
3. Applications
4. Risks
5. Summary

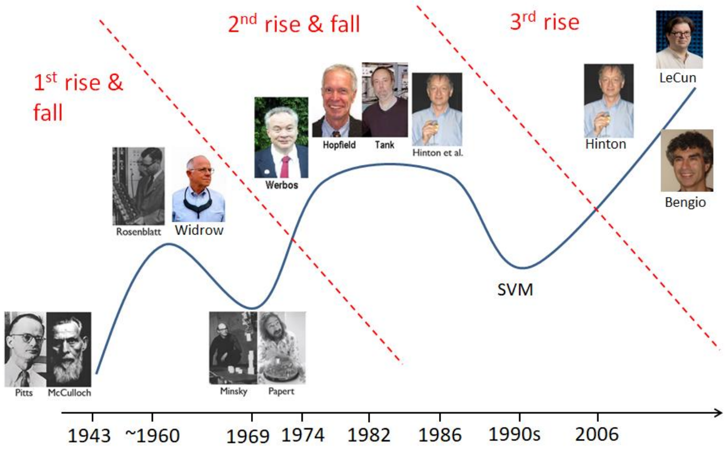
Summary of this lecture

Knowledge

1. General concepts



2. History



3. Applications

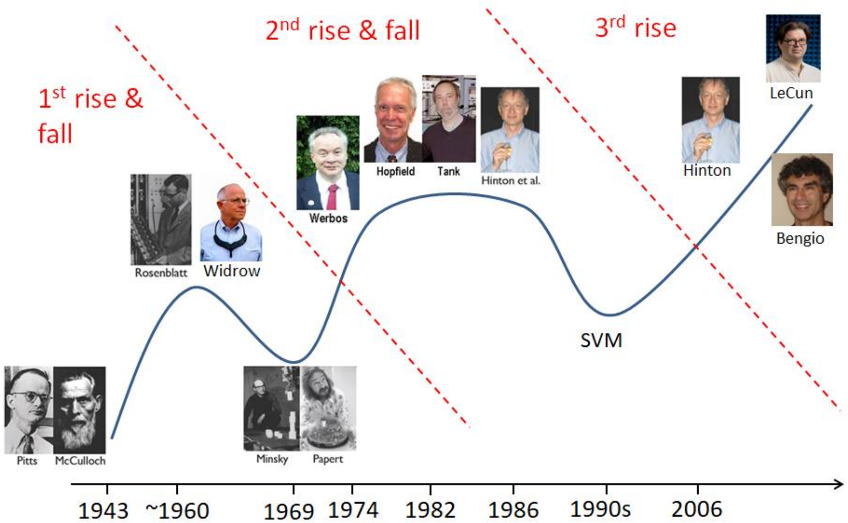
   Computer vision, speech recognition, natural language processing, etc.

4. Risks

Summary of this lecture

Capability and value

Scientific research has rises & falls



"Don't be evil"

1. Cherish the heritage

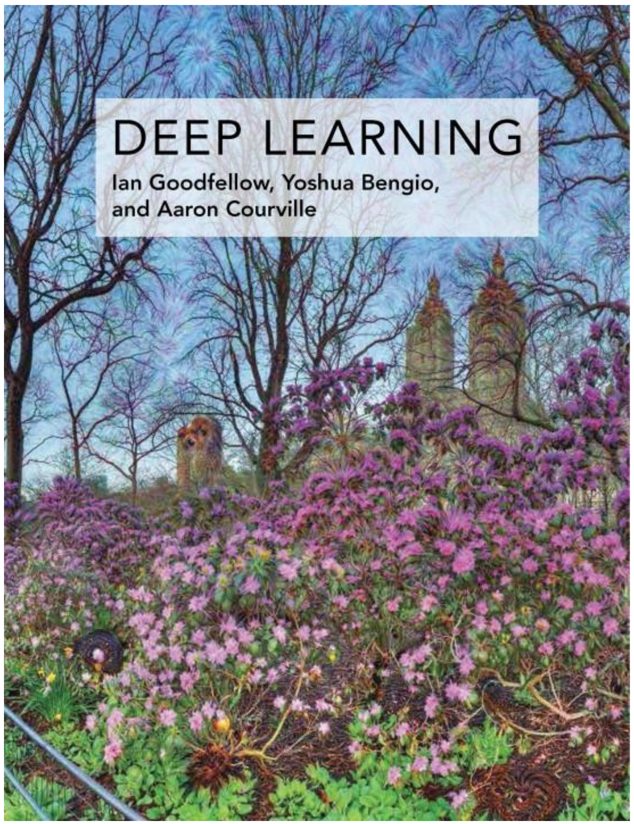
2. Perseverance is important

Recommended reading

- MP unit material on Web Learning
- Walter Pitts: The Man Who Tried to Redeem the World with Logic

  http://nautil.us/issue/21/information/the-man-who-tried-to-redeem-the-world-with-logic

Prepare for the next lecture



[Deep Learning](https://github.com/janishar/mit-deep-learning-book-pdf)
Ian Goodfellow, Yoshua Bengio and Aaron Courville
The MIT Press, 2018

https://github.com/janishar/mit-deep-learning-book-pdf

1. Chapters 2-5
2. Handout about Math Basics