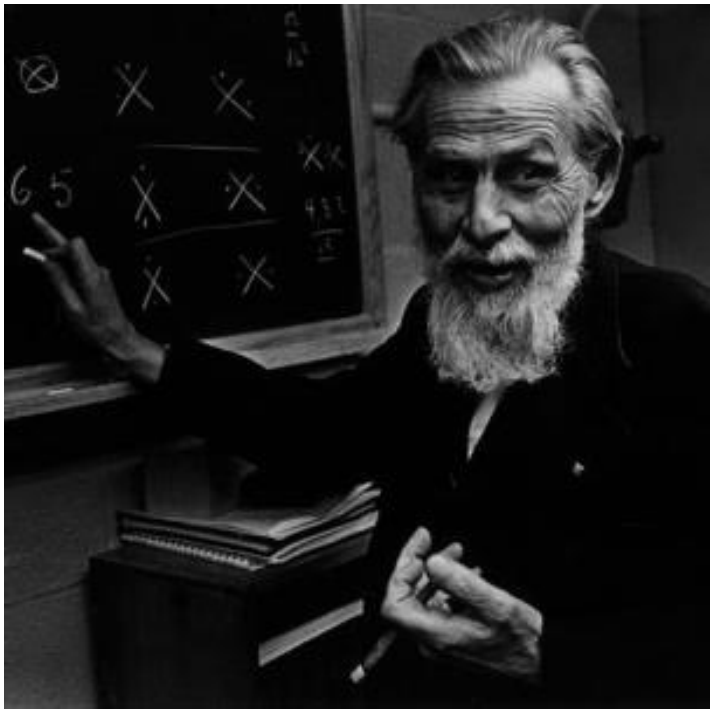
Neural Network and Deep Learning

Lecture 2

Yanwei Fu
School of Data Science, Fudan University

神经元进入计算领域的视野

Walter Pitts
(1923-1969)



Warren McCulloch
(1898-1969)

BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

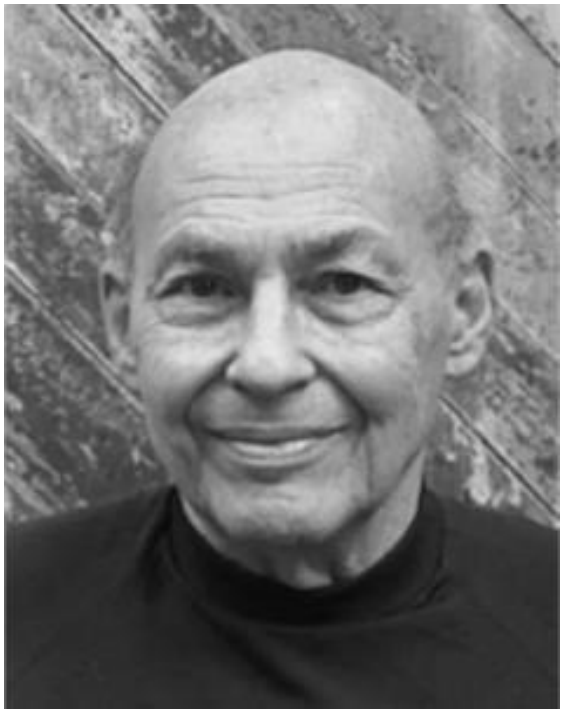
A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

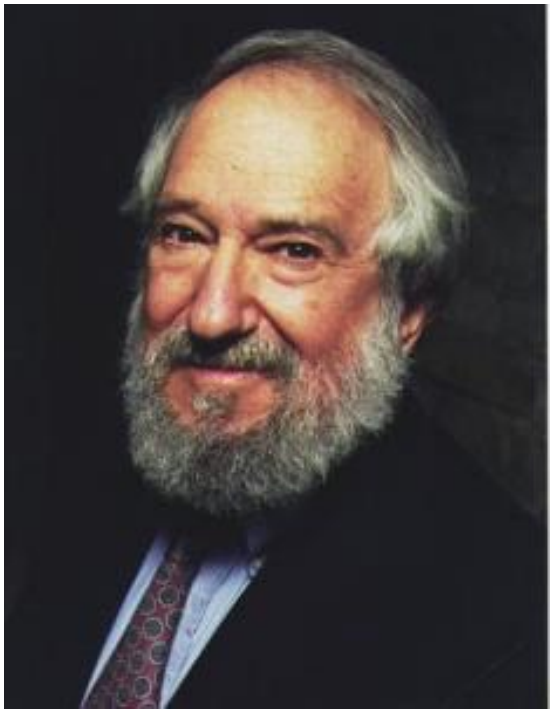
FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

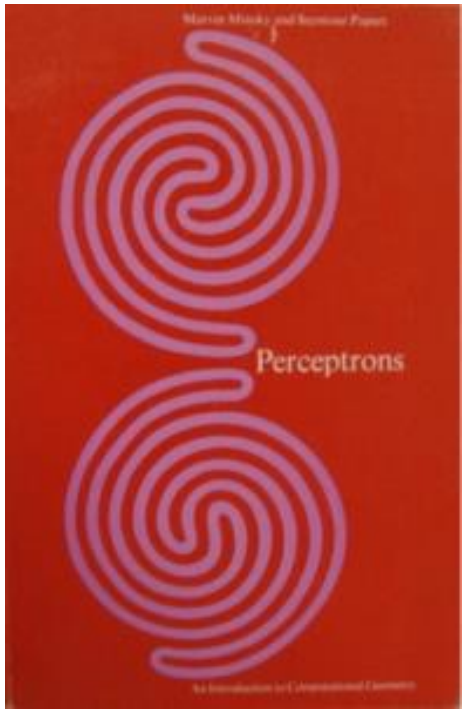
神经网络的第一次寒冬

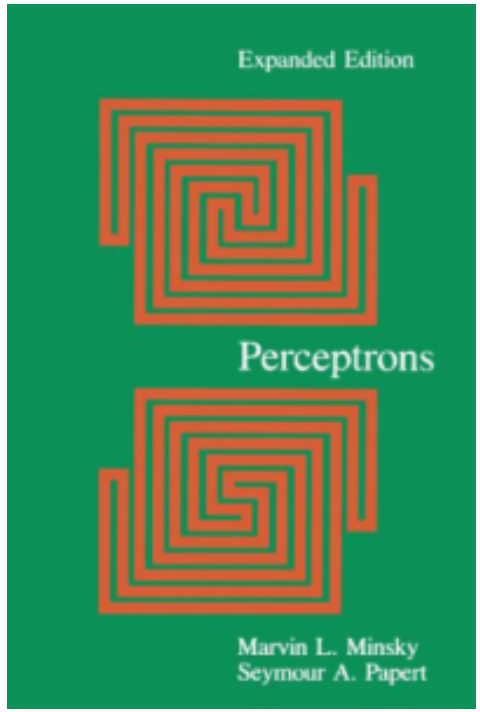


Marvin Minsky (1927-2016)



Seymour Papert (1928-)





1969年出版《Perceptrons》一书，认为仅靠局部连接的神经网络无法有效开展训练

以及很多被后来的读者们以讹传讹的观点

Perception: From Rosenblatt, 1962..

Optional subtitle

- "Perception, then, emerges as that relatively primitive, partly autonomous, institutionalized, ratiomorphic subsystem of cognition which achieves prompt and richly detailed orientation habitually concerning the vitally relevant, mostly distal aspects of the environment on the basis of mutually vicarious, relatively restricted and stereotyped, insufficient evidence in uncertainty-geared interaction and compromise, seemingly following the highest probability for smallness of error at the expense of the highest frequency of precision. "
  - – From "Perception and the Representative Design of Psychological Experiments, " by Egon Brunswik, 1956 (posthumous).

- "That's a simplification. Perception is standing on the sidewalk, watching all the girls go by."
  - – From "The New Yorker", December 19, 1959

反向传播算法（BP）的提出

Geoffrey Hinton
(1947-)

David Rumelhart
(1942-2011)

Learning representations by back-propagating errors

David E. Rumelhart*, Geoffrey E. Hinton† & Ronald J. Williams*

* Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California 92093, USA
† Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Philadelphia 15213, USA

1986年发表在Nature上的文章将BP算法用于神经网络模型，极大降低了计算量（$O(n^2)$->$O(n)$）

当时的计算机普及率、计算能力也远胜60年代

“首个”重要应用：手写数字识别

Yann LeCun
(1960-)

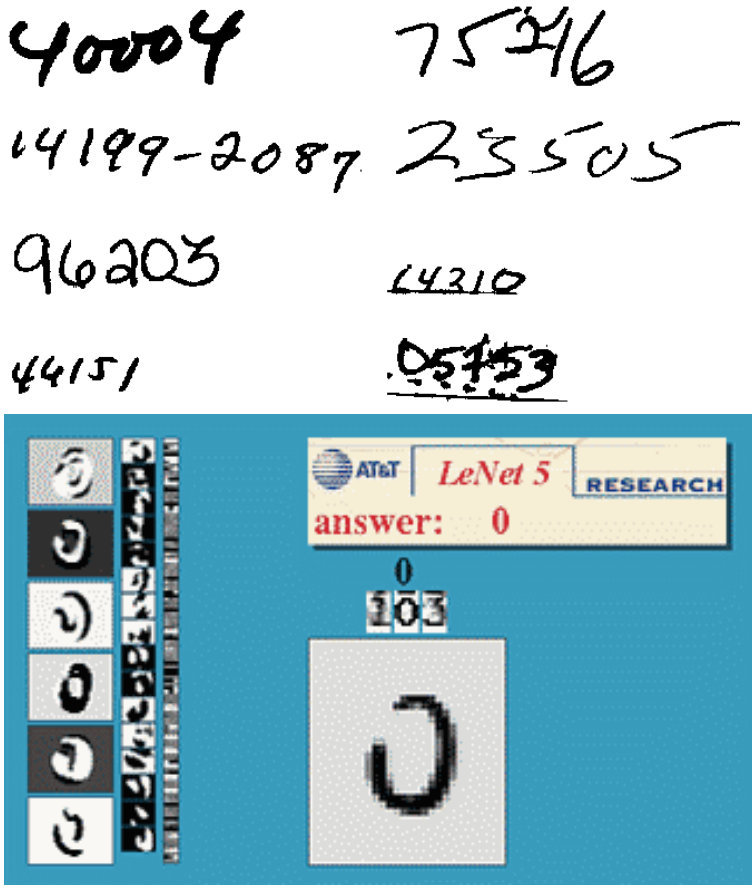
Handwritten Digit Recognition with a Back-Propagation Network

Y. Le Cun, B. Boser, J. S. Denker, D. Henderson,
R. E. Howard, W. Hubbard, and L. D. Jackel
AT&T Bell Laboratories, Holmdel, N. J. 07733

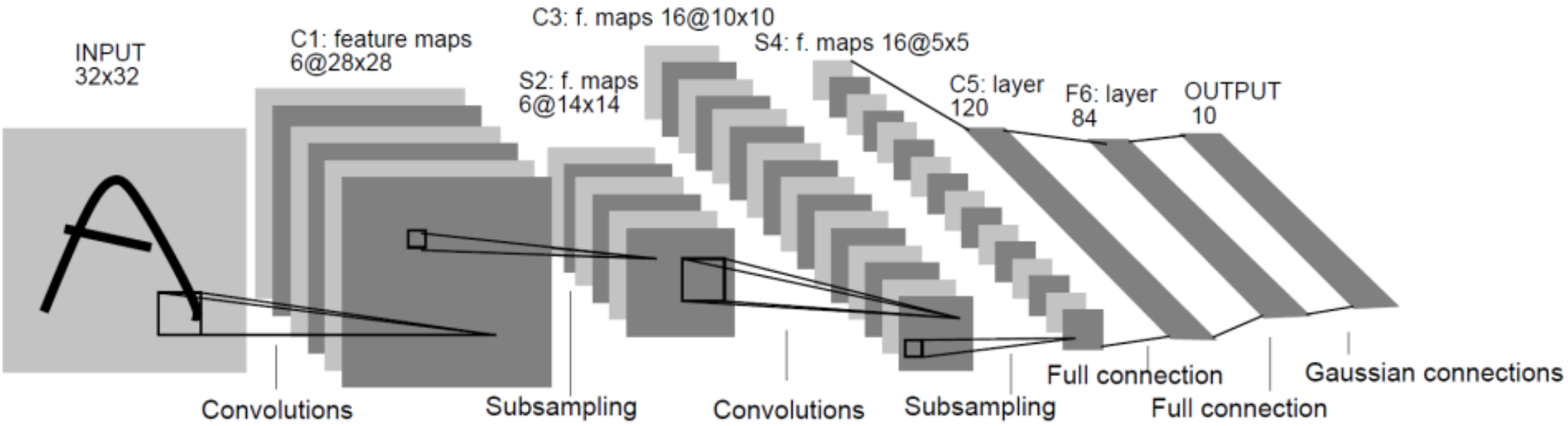
ABSTRACT

We present an application of back-propagation networks to handwritten digit recognition. Minimal preprocessing of the data was required, but architecture of the network was highly constrained and specifically designed for the task. The input of the network consists of normalized images of isolated digits. The method has 1% error rate and about a 9% reject rate on zipcode digits provided by the U.S. Postal Service.

到上世纪九十年代末，超过10%的美国支票识别采用了相关技术



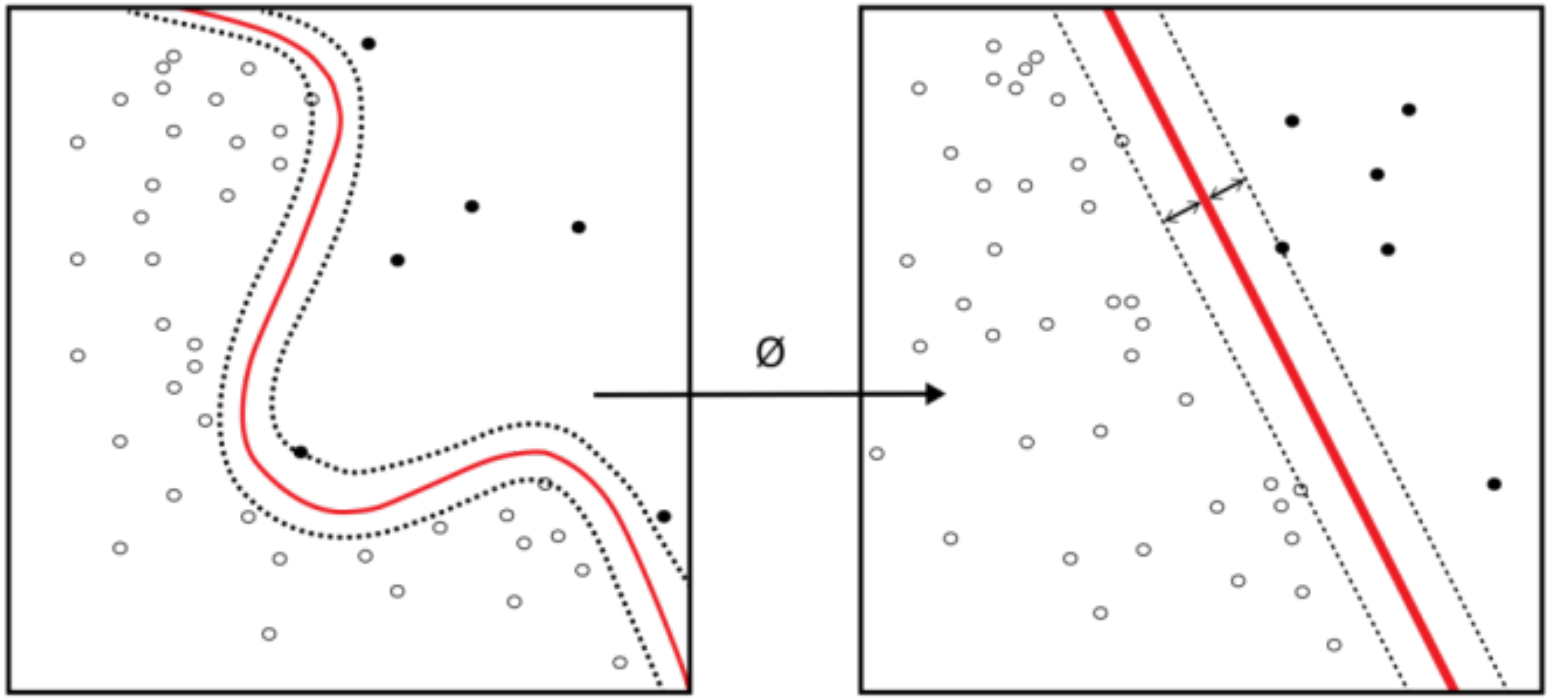
LeNet-5



1998年发表的Gradient-based learning applied to Document Recognition一文中提出了LeNet-5
已是Convolutional Neural Network的基本框架

神经网络的第二次寒冬

Vladmir Vapnik
(1936-)

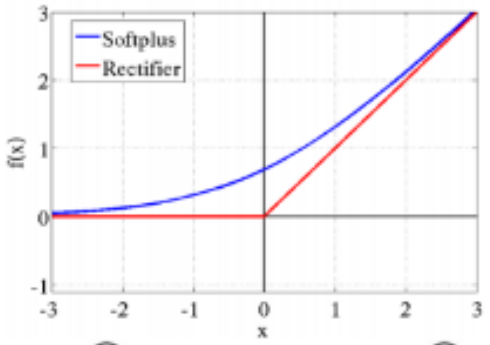


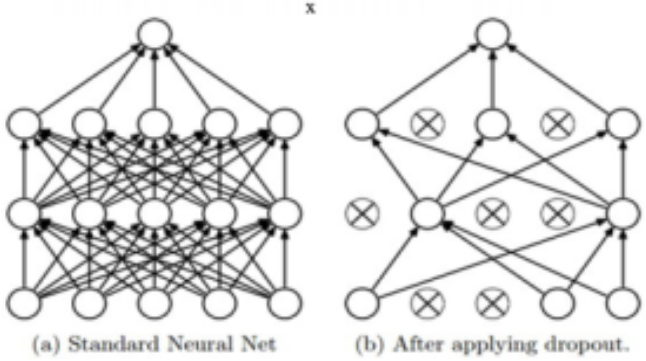
在1963年发表的论文中即提出Support Vector Machine的概念
在2002年左右，不断改进的SVM方法将手写数字识别的错误率降到了0.56%
不需要大量样本（事实上也难以支撑大量样本），速度相对可以接受

卷土重来的（深度）神经网络

优化策略

梯度弥散问题通过ReLU、Dropout、Deep Residual Learning等方法得到缓解

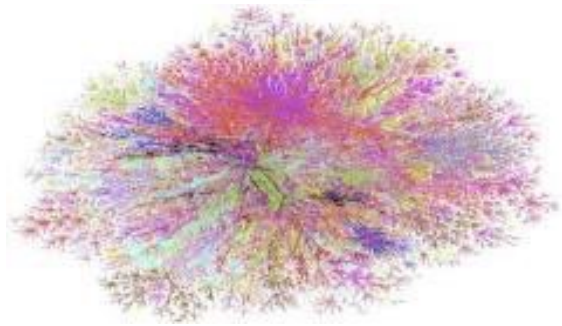




数据规模

真正意义上的大数据出现

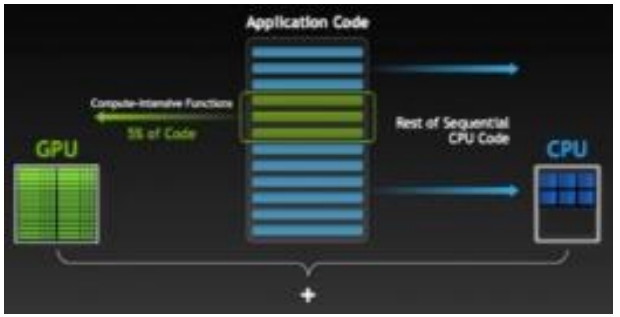
社会标注机制



计算能力

CPU、GPU的长足进步

共享权值



AlexNet

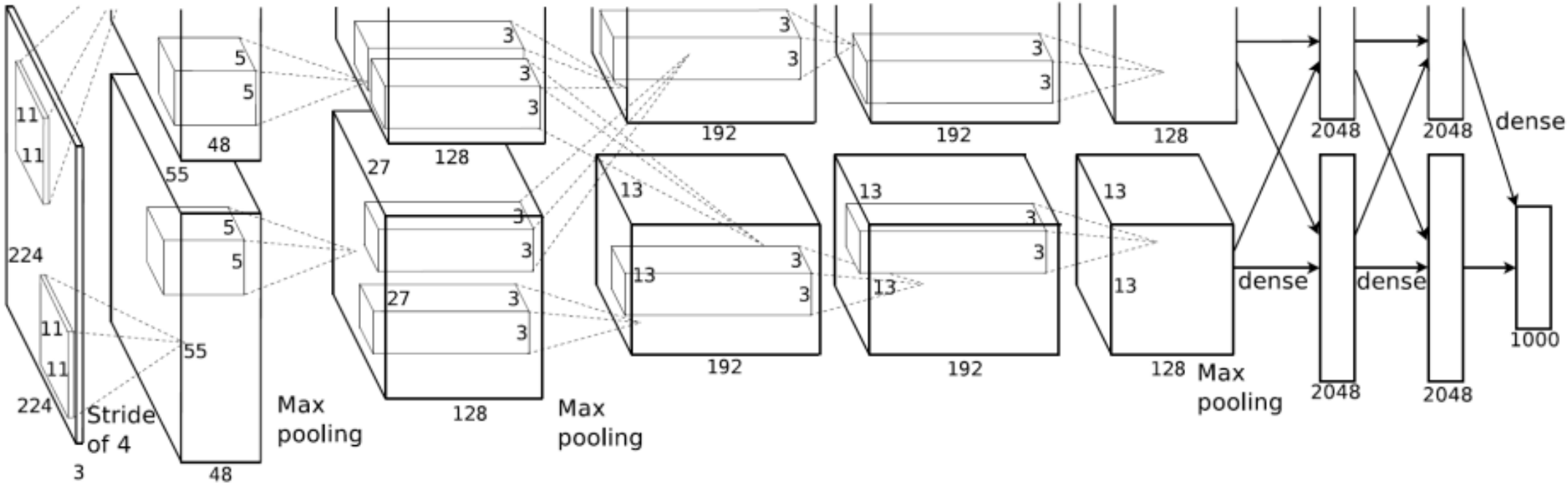
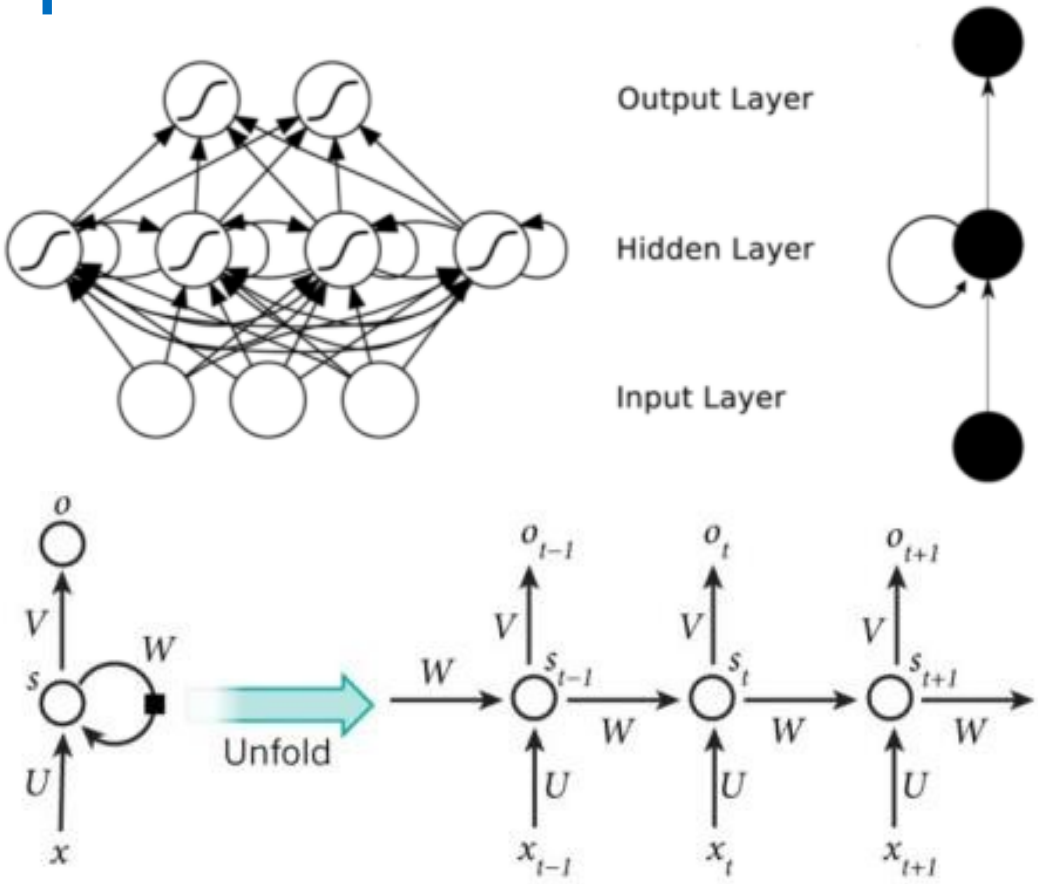


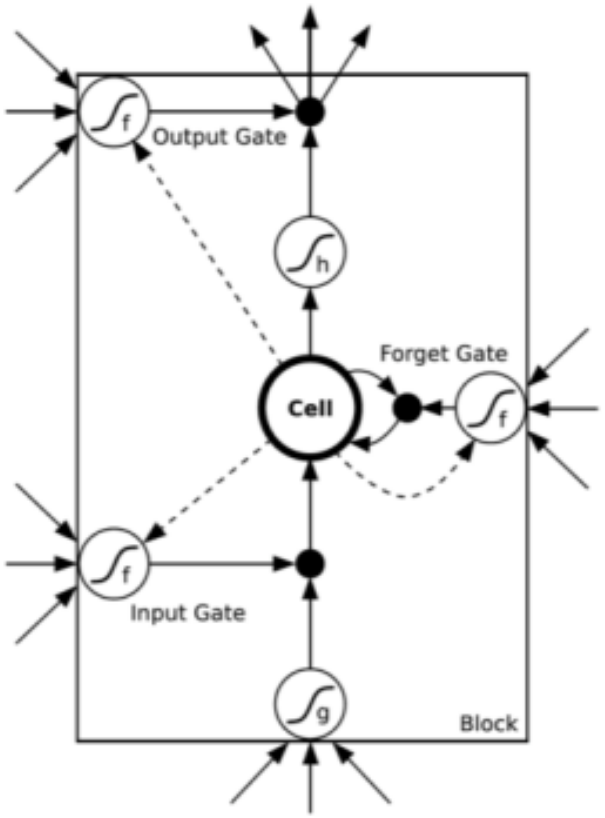
Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

可处理时序数据的深度学习机制

Recursive Neural Networks



Long Short Term Memory



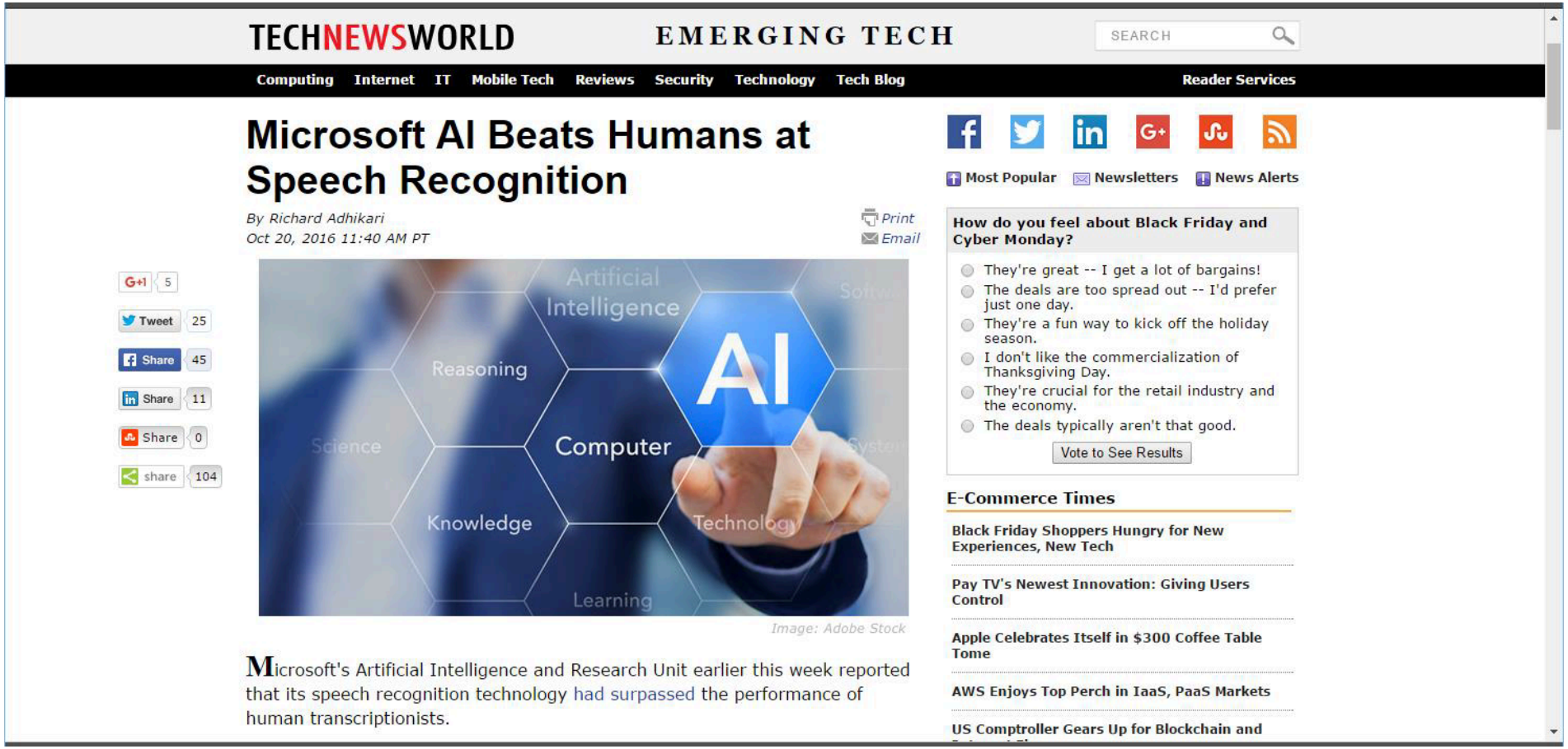
大型企业不同于以往的密切关注

开源框架

名称	基本语言	是否支持多GPU	速度	应用领域
TensorFlow	Python、C++	是	★★☆	通用
Caffe	C++	是	★★★☆	图像分类
Torch	Lua	是	★★☆	通用
Theano	Python	默认为否	★★★☆	通用
IBM DL Platform	SystemML等	是	?	通用

Breakthroughs with neural networks

Optional subtitle



TECHNEWSWORLD EMERGING TECH SEARCH

Computing Internet IT Mobile Tech Reviews Security Technology Tech Blog Reader Services

Microsoft AI Beats Humans at Speech Recognition

By Richard Adhikari
Oct 20, 2016 11:40 AM PT

Print Email

G+1 5 Tweet 25 Share 45 Share 11 Share 0 share 104

Image: Adobe Stock

Microsoft's Artificial Intelligence and Research Unit earlier this week reported that its speech recognition technology had surpassed the performance of human transcriptionists.

Most Popular Newsletters News Alerts

How do you feel about Black Friday and Cyber Monday?

- They're great -- I get a lot of bargains!
- The deals are too spread out -- I'd prefer just one day.
- They're a fun way to kick off the holiday season.
- I don't like the commercialization of Thanksgiving Day.
- They're crucial for the retail industry and the economy.
- The deals typically aren't that good.

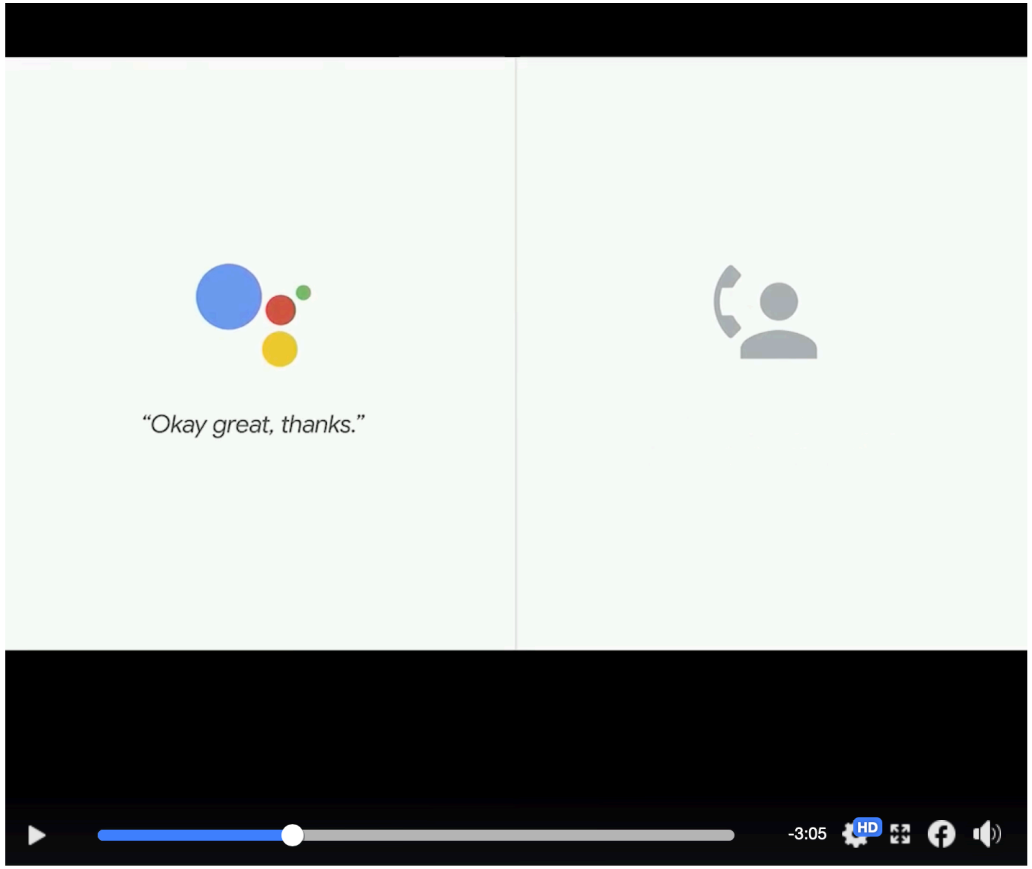
Vote to See Results

E-Commerce Times

- Black Friday Shoppers Hungry for New Experiences, New Tech
- Pay TV's Newest Innovation: Giving Users Control
- Apple Celebrates Itself in $300 Coffee Table Tome
- AWS Enjoys Top Perch in IaaS, PaaS Markets
- US Comptroller Gears Up for Blockchain and

Google just gave a stunning demo of Assistant making an actual phone call

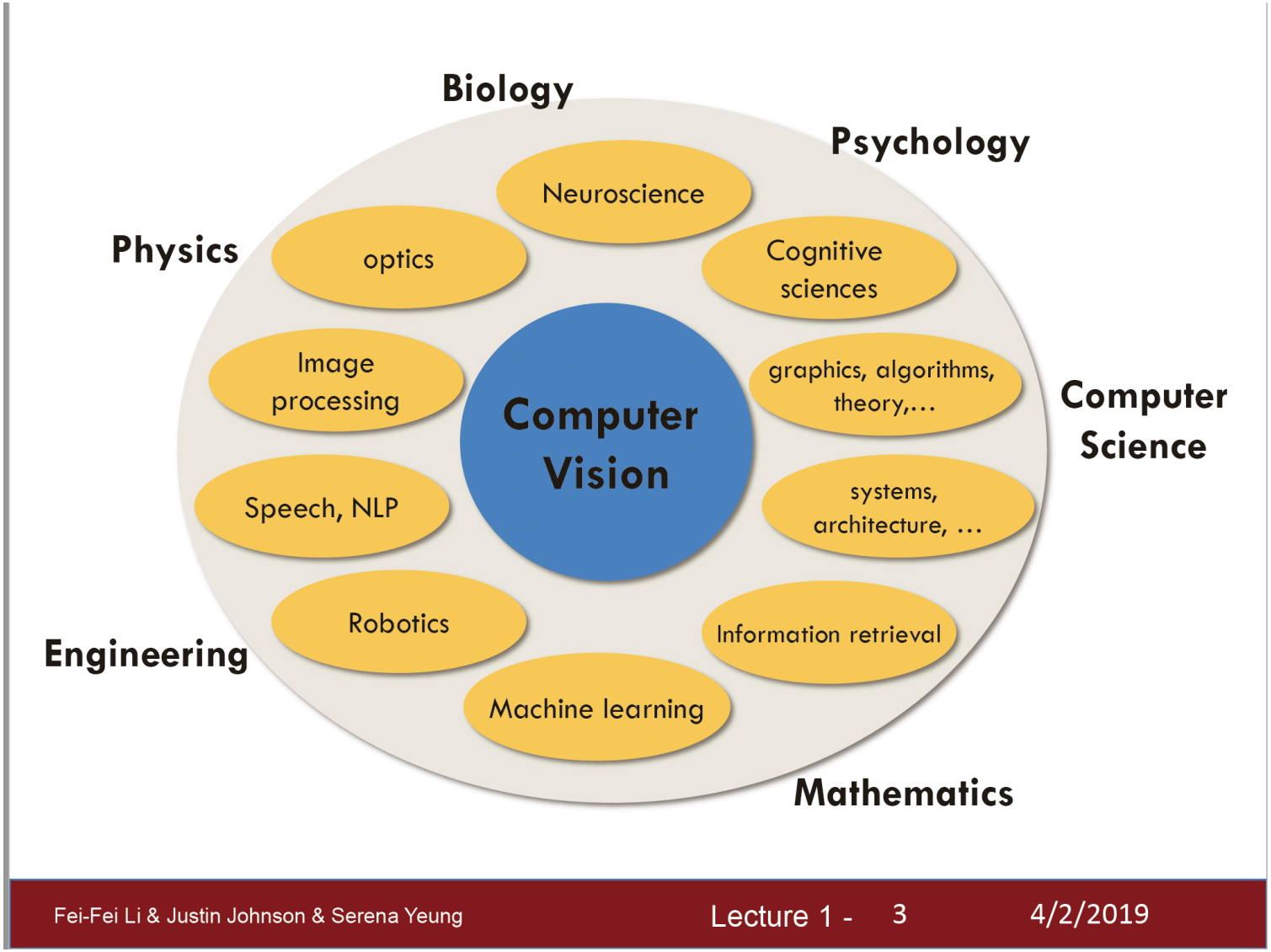
https://www.theverge.com/2018/5/8/17332070/google-assistant-makes-phone-call-demo-duplex-io-2018

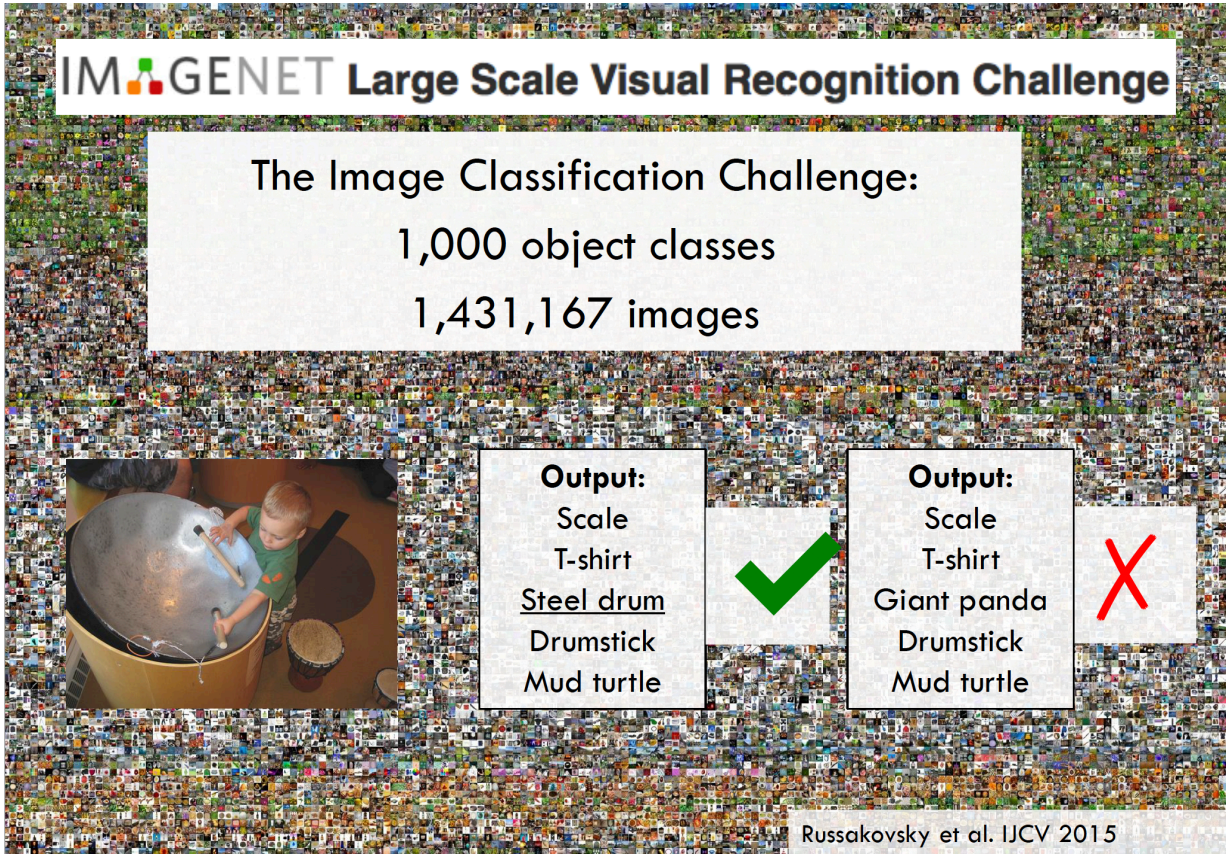


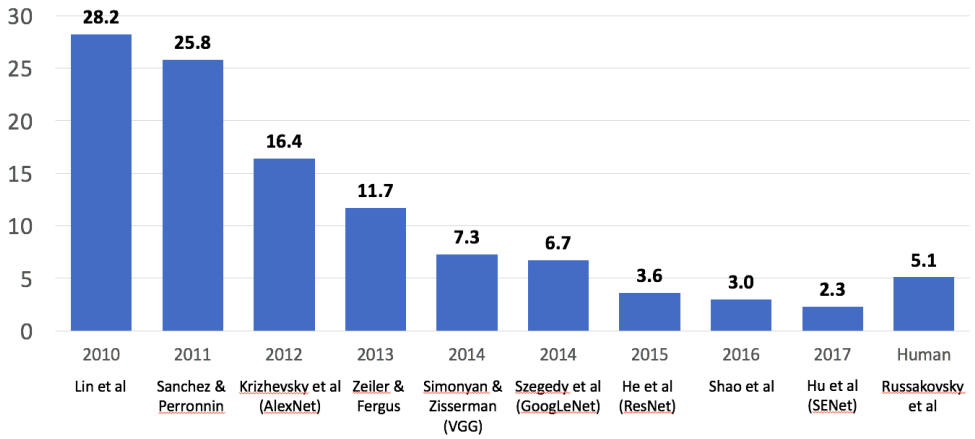
Onstage at I/O 2018, Google showed off a jaw-dropping new capability of Google Assistant: in the not too distant future, it's going to make phone calls on your behalf. CEO Sundar Pichai played back a phone call recording that he said was placed by the Assistant to a

Breakthroughs with neural networks

Optional subtitle




IMAGENET Large Scale Visual Recognition Challenge
The Image Classification Challenge:
1,000 object classes
1,431,167 images
Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle
Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle
Russakovsky et al. IJCV 2015


30
25
20
15
10
5
0
28.2
25.8
16.4
11.7
7.3
6.7
3.6
3.0
2.3
5.1
2010 Lin et al
2011 Sanchez & Perronnin
2012 Krizhevsky et al (AlexNet)
2013 Zeiler & Fergus
2014 Simonyan & Zisserman (VGG)
2014 Szegedy et al (GoogLeNet)
2015 He et al (ResNet)
2016 Shao et al
2017 Hu et al (SENet)
Human Russakovsky et al

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

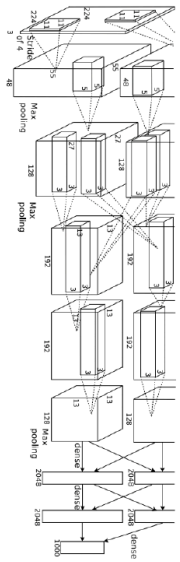
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

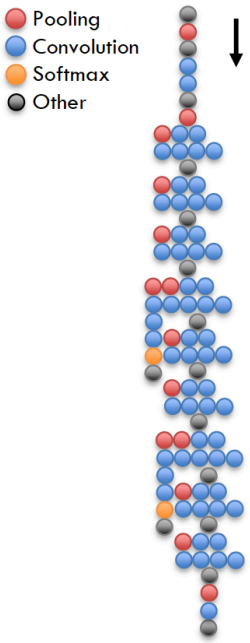
SuperVision



[Krizhevsky NIPS 2012]

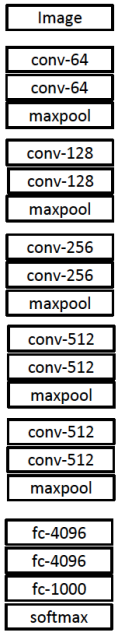
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

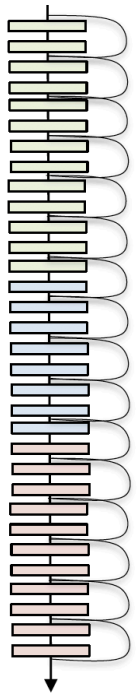
VGG



[Simonyan arxiv 2014]

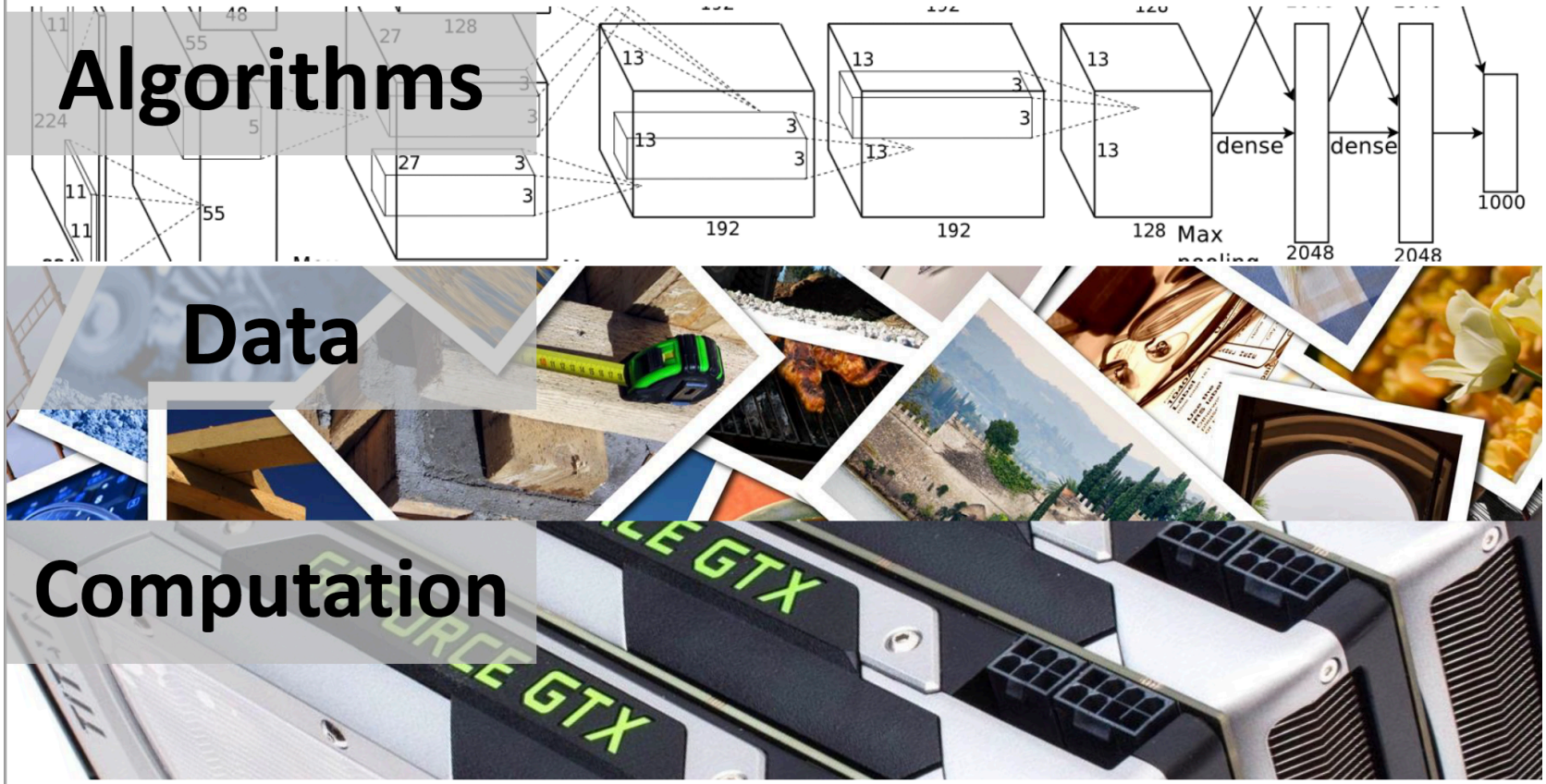
Year 2015

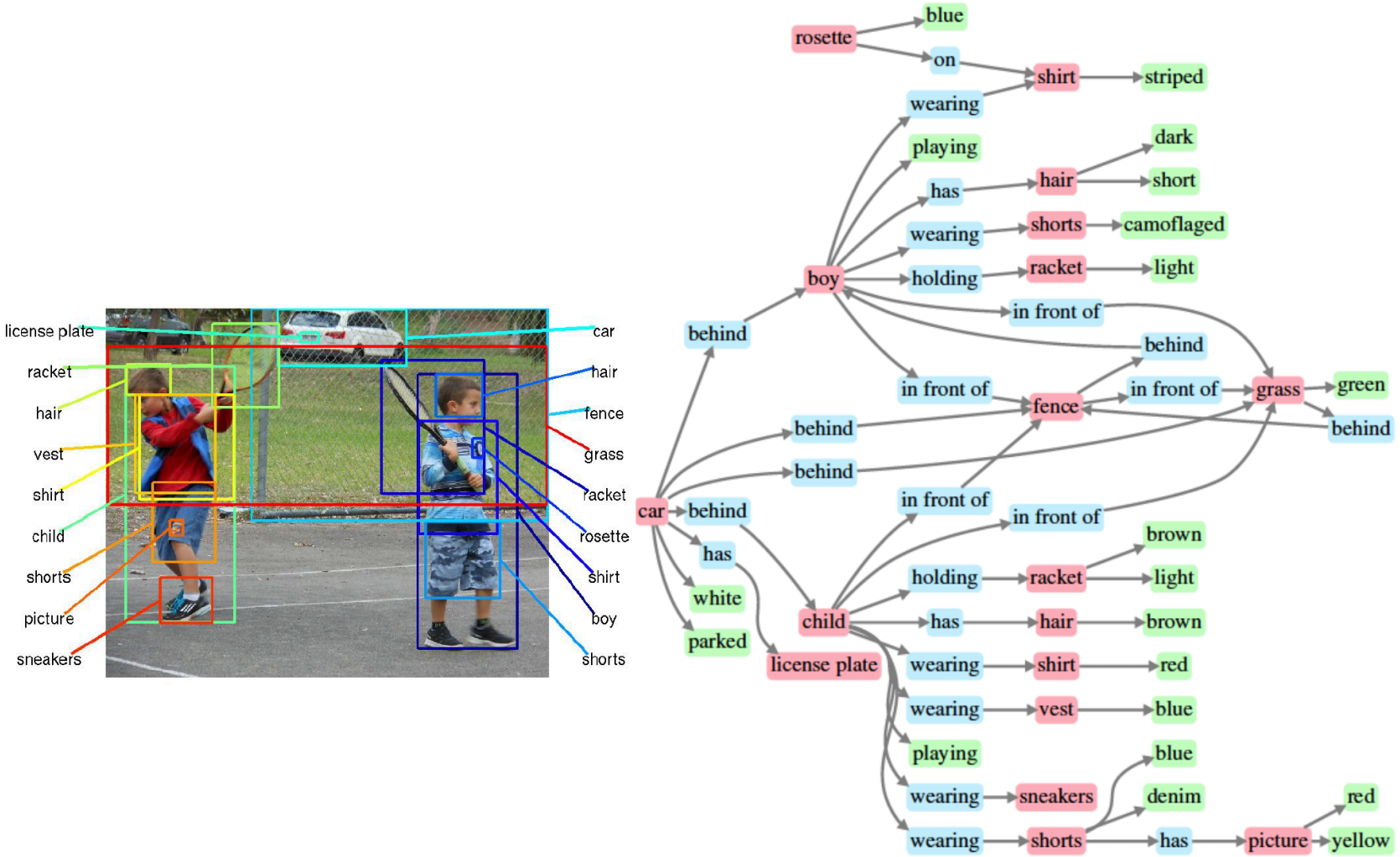
MSRA



[He ICCV 2015]

Ingredients for Deep Learning





Johnson *et al.*, "Image Retrieval using Scene Graphs", CVPR 2015

Image Recognition

Optional subtitle

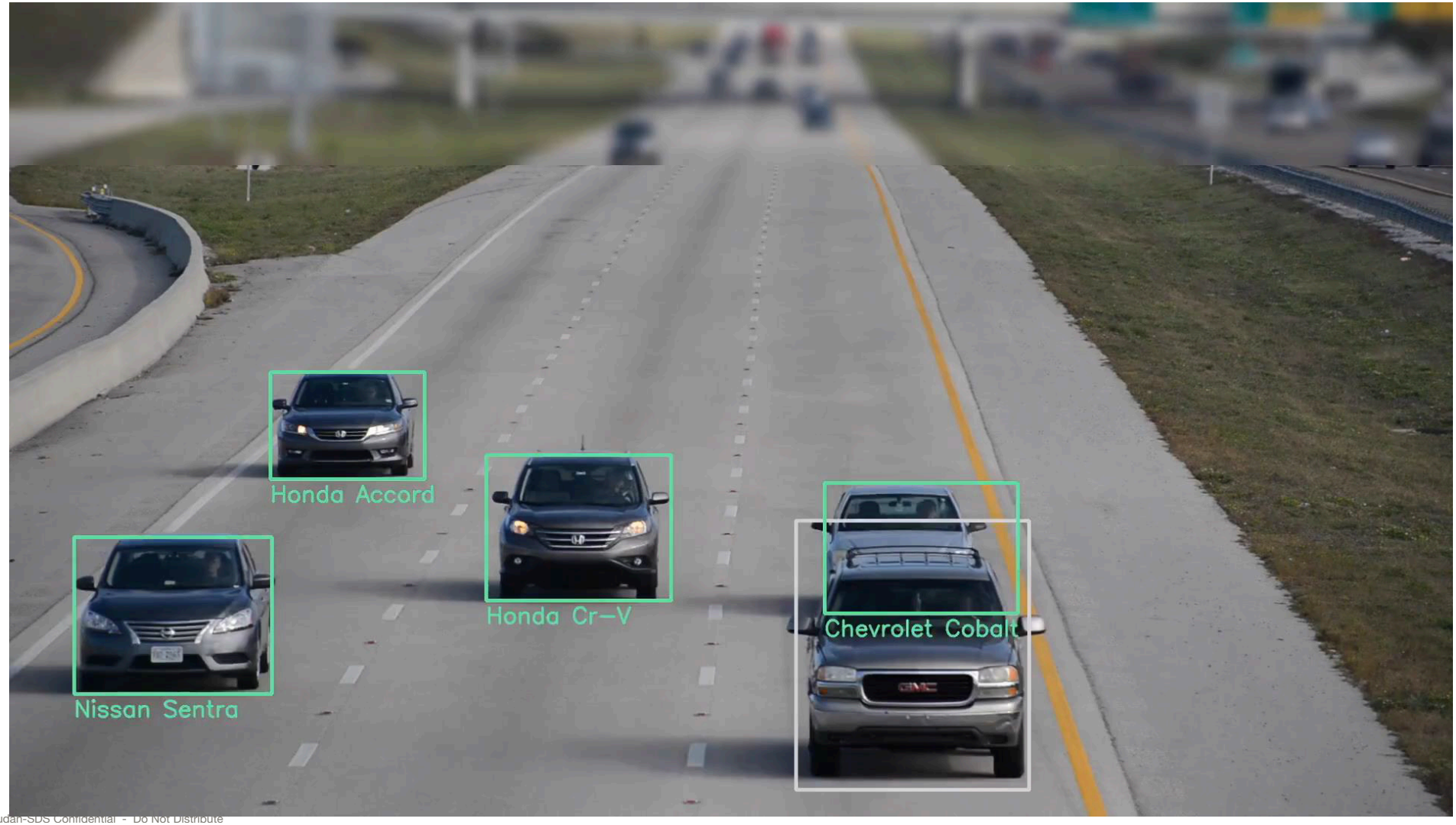
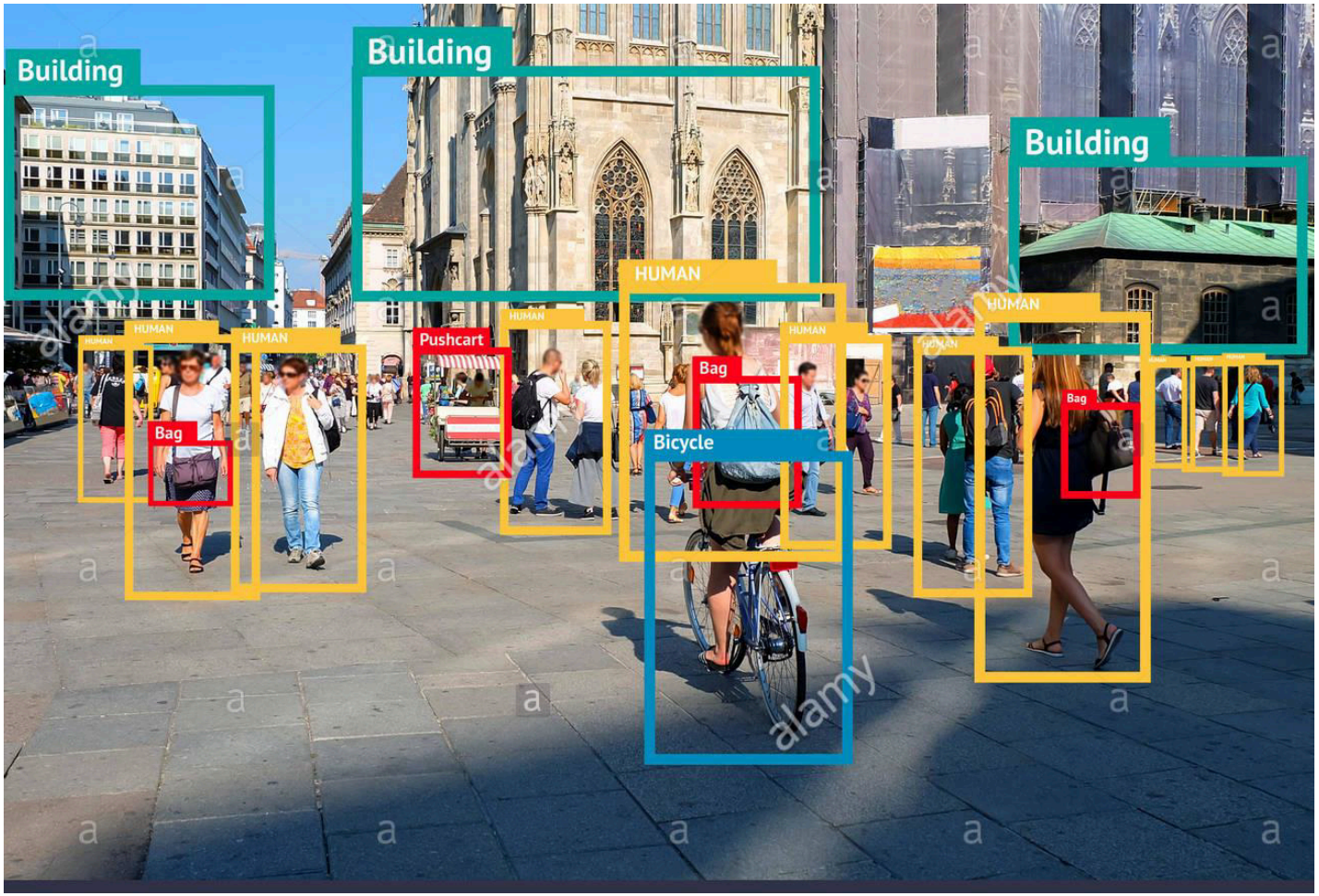


Image segmentation & recognition

Optional subtitle



Breakthroughs with neural networks

Optional subtitle

AlphaGo

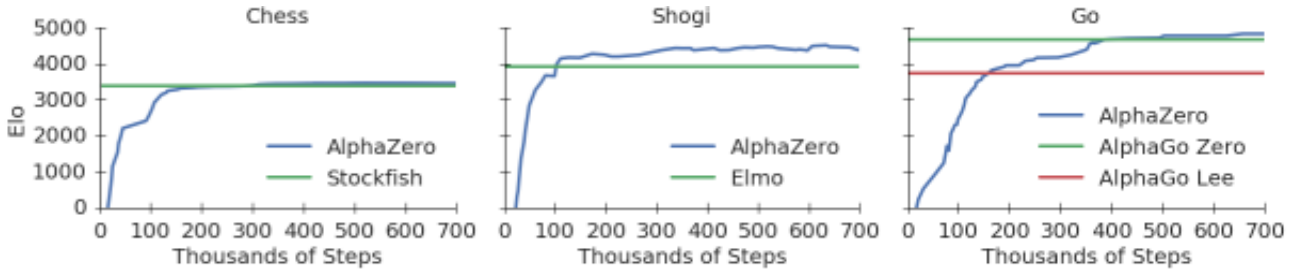
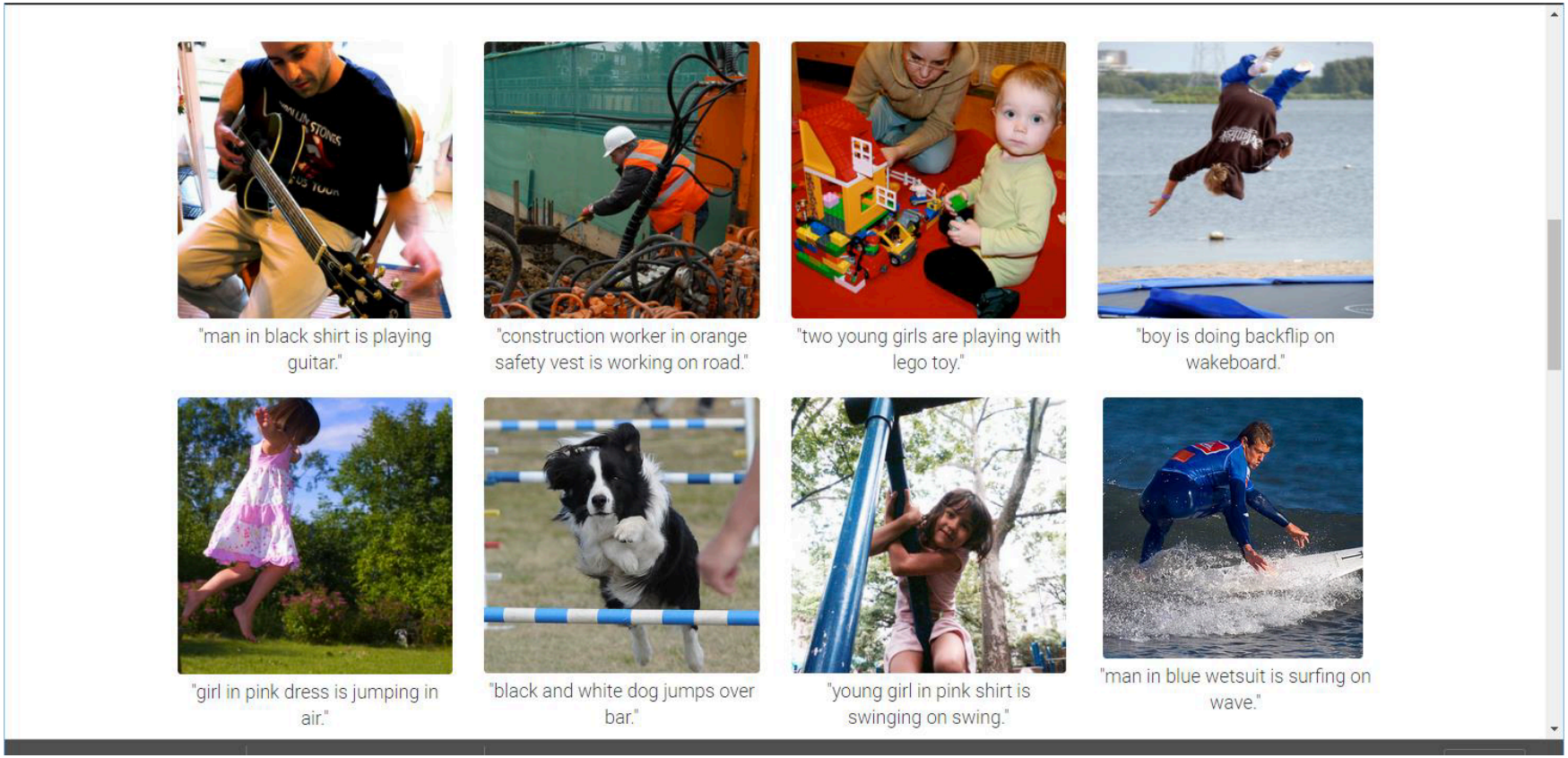


Figure 1: Training *AlphaZero* for 700,000 steps. Elo ratings were computed from evaluation games between different players when given one second per move. **a** Performance of *AlphaZero* in chess, compared to 2016 TCEC world-champion program *Stockfish*. **b** Performance of *AlphaZero* in shogi, compared to 2017 CSA world-champion program *Elmo*. **c** Performance of *AlphaZero* in Go, compared to *AlphaGo Lee* and *AlphaGo Zero* (20 block / 3 day) (*29*).

Breakthroughs with neural networks

Optional subtitle



Neural Network In General

Background

A Recipe for Machine Learning

1. Given training data:

$$\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$$

Face *Face* *Not a face*

2. Choose each of these:
   - Decision function

$$\hat{\boldsymbol{y}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

   - Loss function

$$\ell(\hat{\boldsymbol{y}}, \boldsymbol{y}_i) \in \mathbb{R}$$

Examples: Linear regression, Logistic regression, Neural Network

Examples: Mean-squared error, Cross Entropy

Background

A Recipe for Machine Learning

1. Given training data:

$$\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$$

2. Choose each of these:
   - Decision function

$$\hat{\boldsymbol{y}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

   - Loss function

$$\ell(\hat{\boldsymbol{y}}, \boldsymbol{y}_i) \in \mathbb{R}$$

3. Define goal:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

4. Train with SGD:

(take small steps opposite the gradient)

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_t \nabla \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i)$$

Background

A Recipe for

1. Given training d

$\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$

2. Choose each of these:

– Decision function

$\hat{\boldsymbol{y}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$

– Loss function

$\ell(\hat{\boldsymbol{y}}, \boldsymbol{y}_i) \in \mathbb{R}$

$\boldsymbol{y}_i)$

opposite the gradient)

$\boldsymbol{\theta}^{(t}$ $^{(t)} - \eta_t \nabla \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i)$

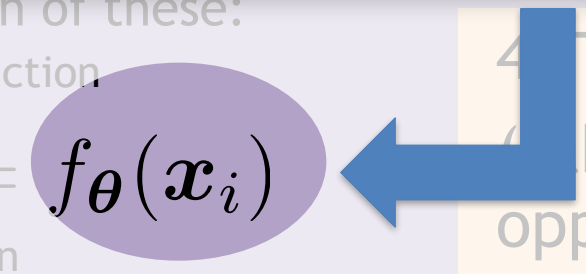
Gradients

Backpropagation can compute this gradient!

And it's a **special case of a more general algorithm** called reverse-mode automatic differentiation that can compute the gradient of any differentiable function efficiently!

Goals for This Course

1. Explore a **new class of decision functions** (Neural Networks)
2. Consider **variants of this recipe** for training



Representation Learning

Causal Inference **in Deep Learning?**

Optional subtitle

“无人问津”的贝叶斯网络之父Judea Pearl在NIPS 2017上到底报告了啥

Judea Pearl, 2011年，

因通过概率和因果推理的算法研发

在人工智能取得的杰出贡献而获得图灵奖

Representations Matter

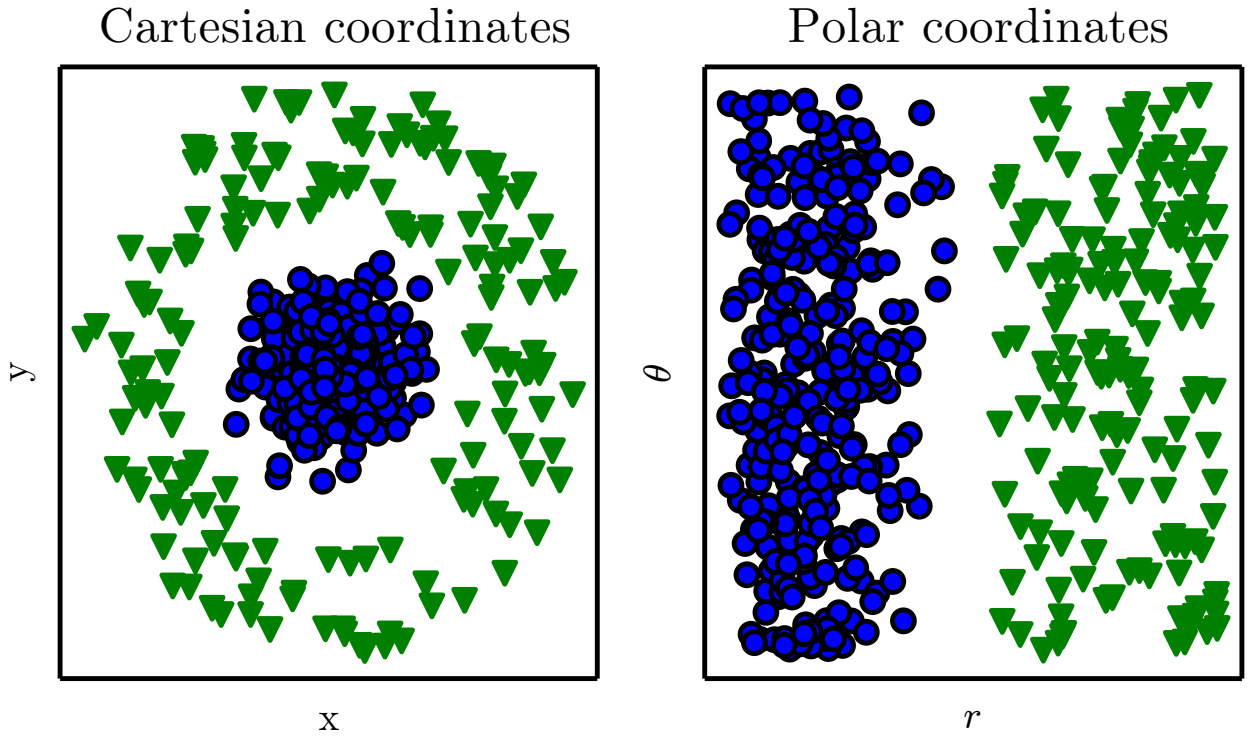


Figure 1.1

(Goodfellow 2016)

Depth: Repeated Composition

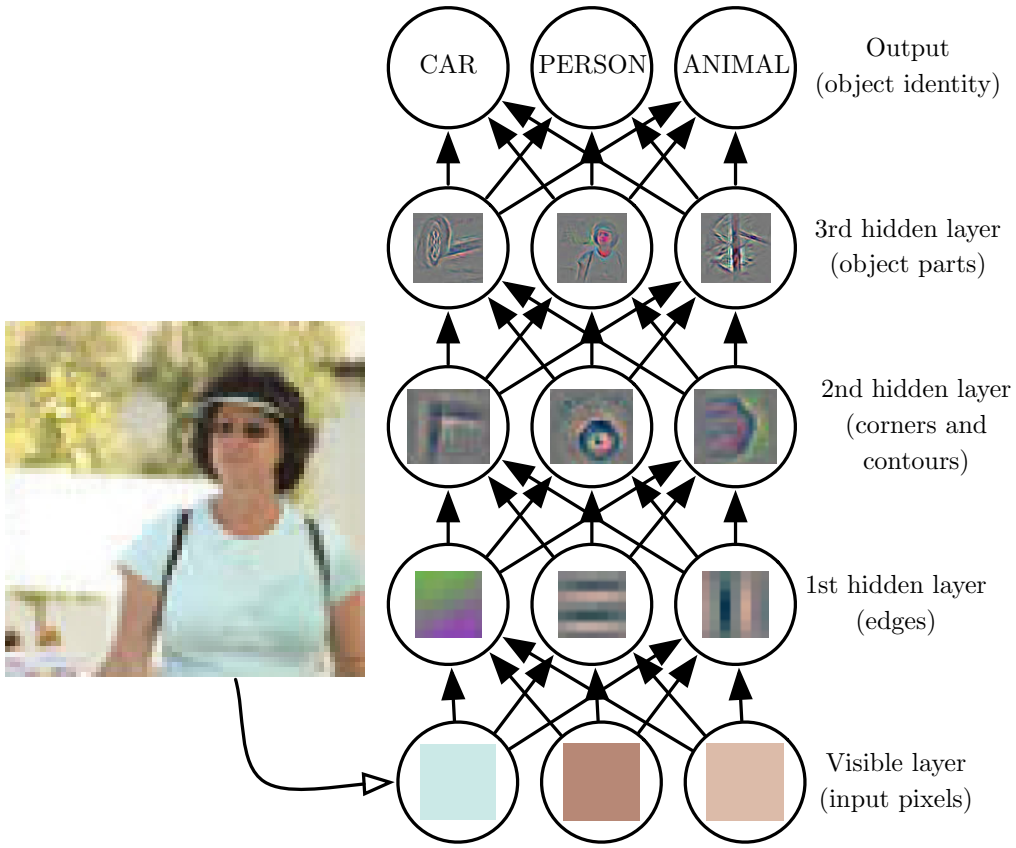


Figure 1.2

Computational Graphs

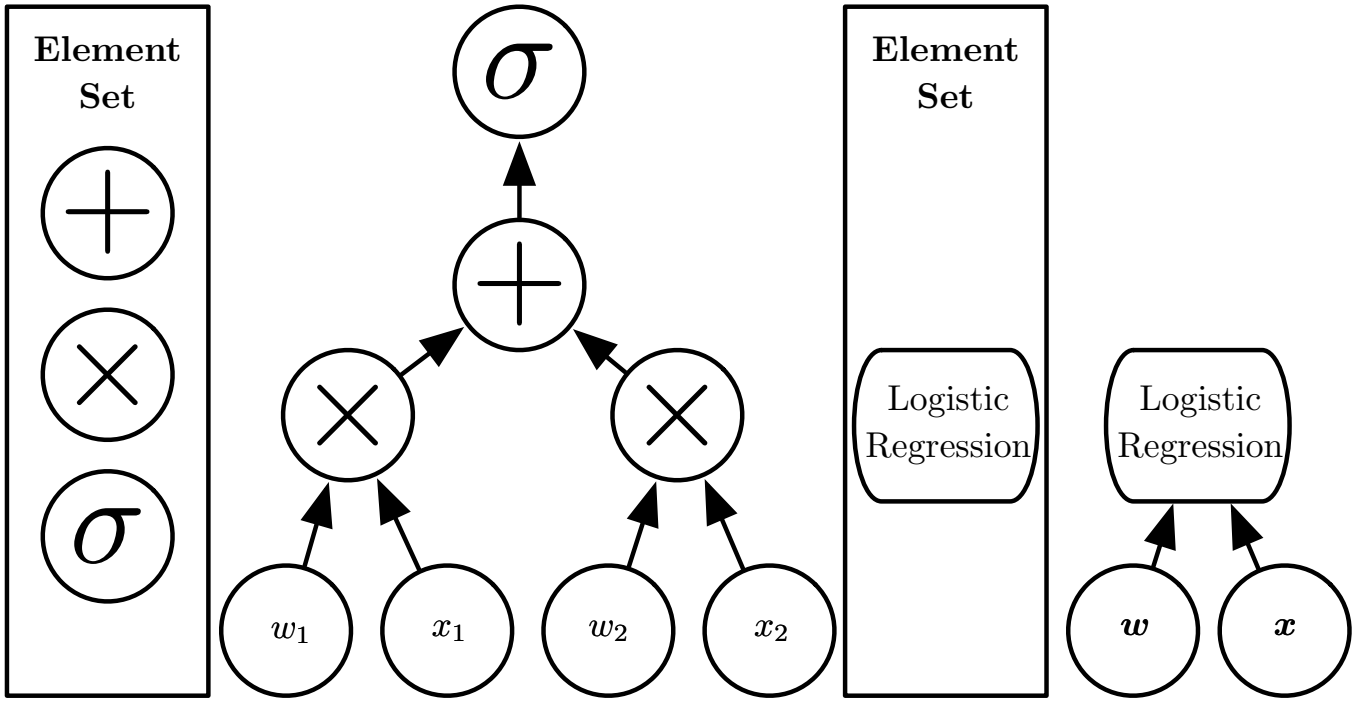


Figure 1.3

Machine Learning and AI

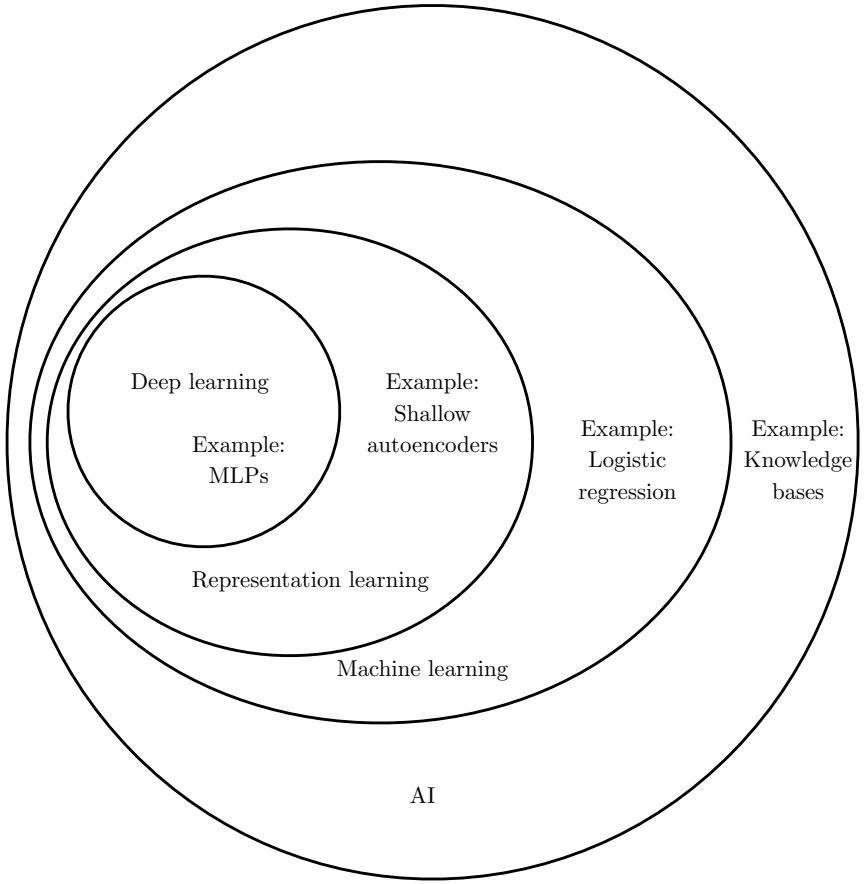
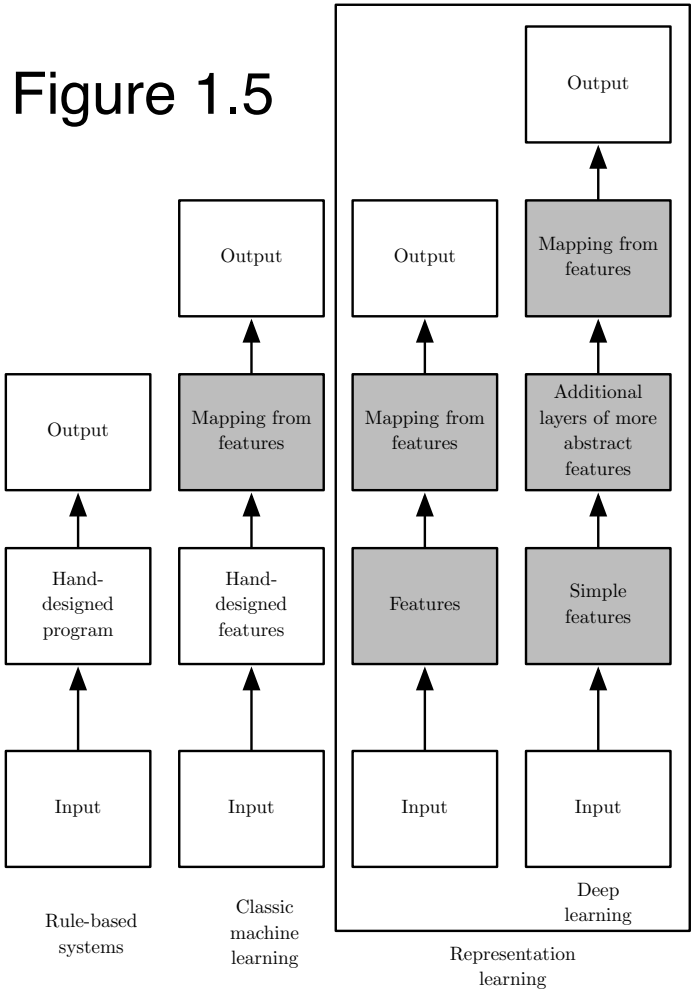


Figure 1.4

(Goodfellow 2016)

Learning Multiple Components

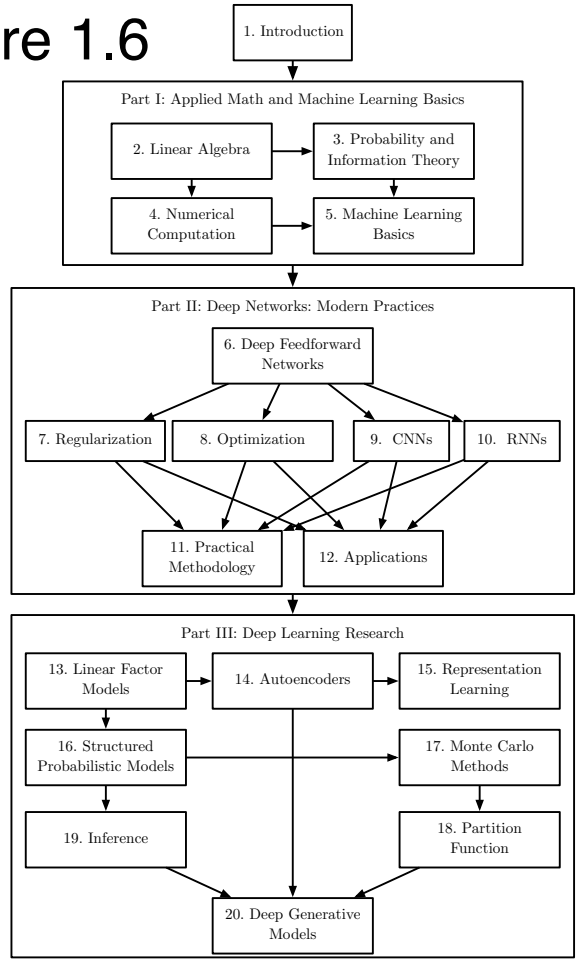
Figure 1.5



(Goodfellow 2016)

Organization of the Book

Figure 1.6



(Goodfellow 2016)

Historical Waves

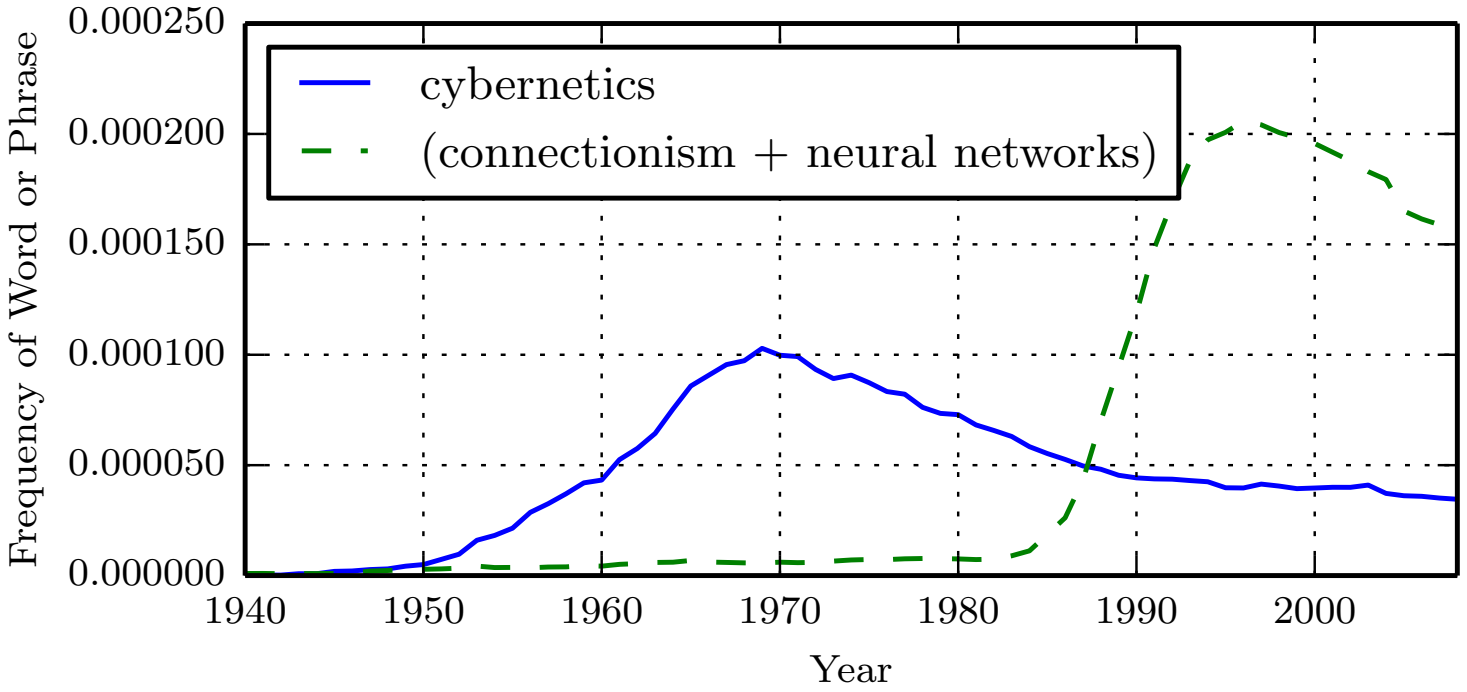


Figure 1.7

Historical Trends: Growing Datasets

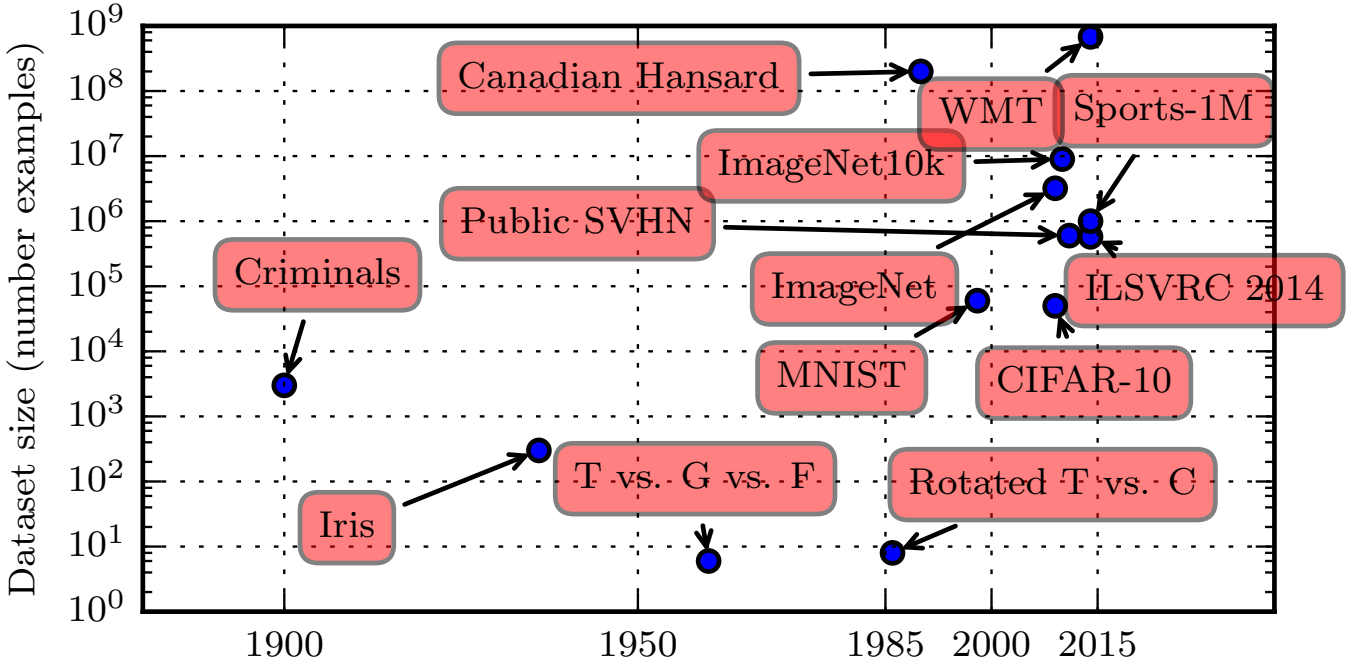


Figure 1.8

(Goodfellow 2016)

The MNIST Dataset

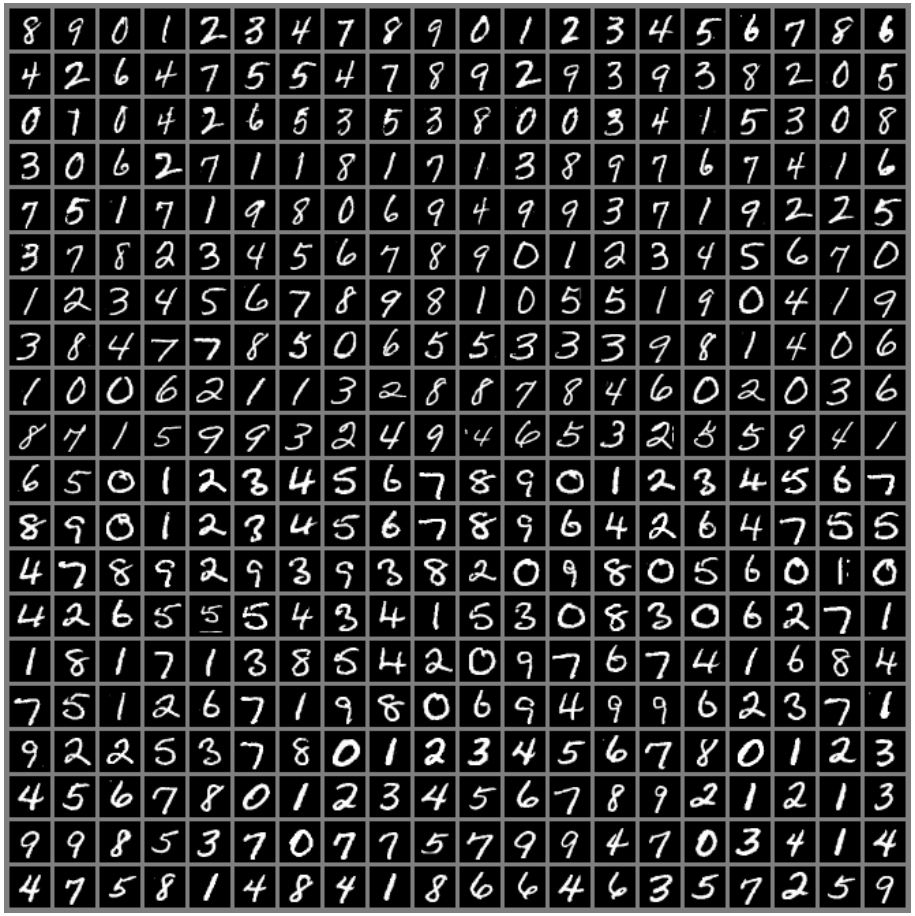


Figure 1.9

Connections per Neuron

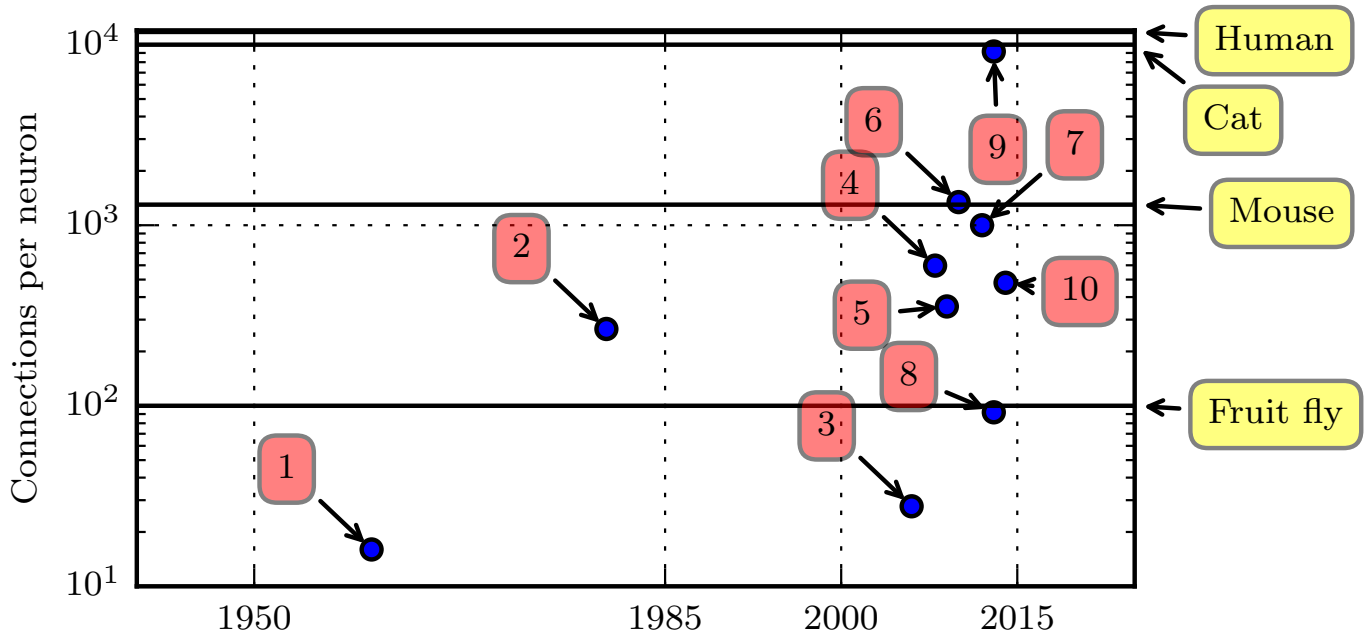


Figure 1.10

Number of Neurons

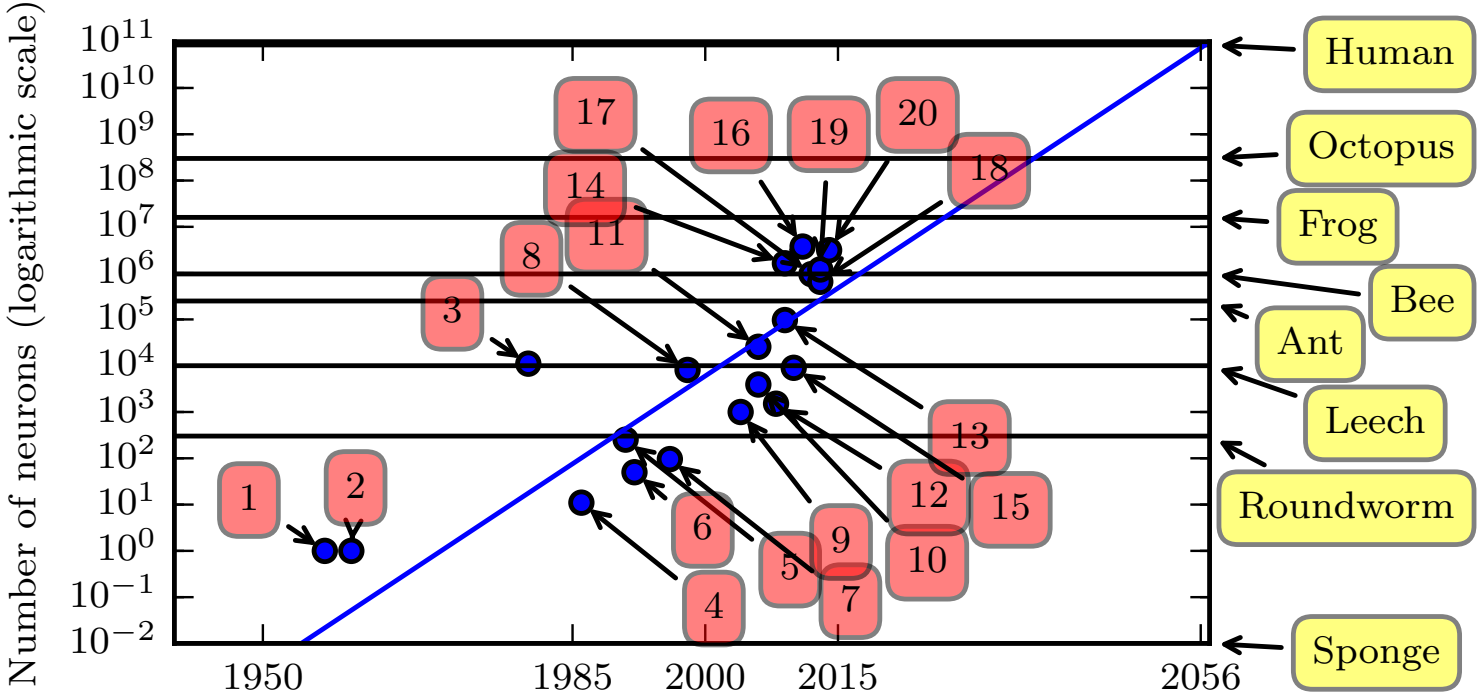


Figure 1.11

(Goodfellow 2016)

Solving Object Recognition

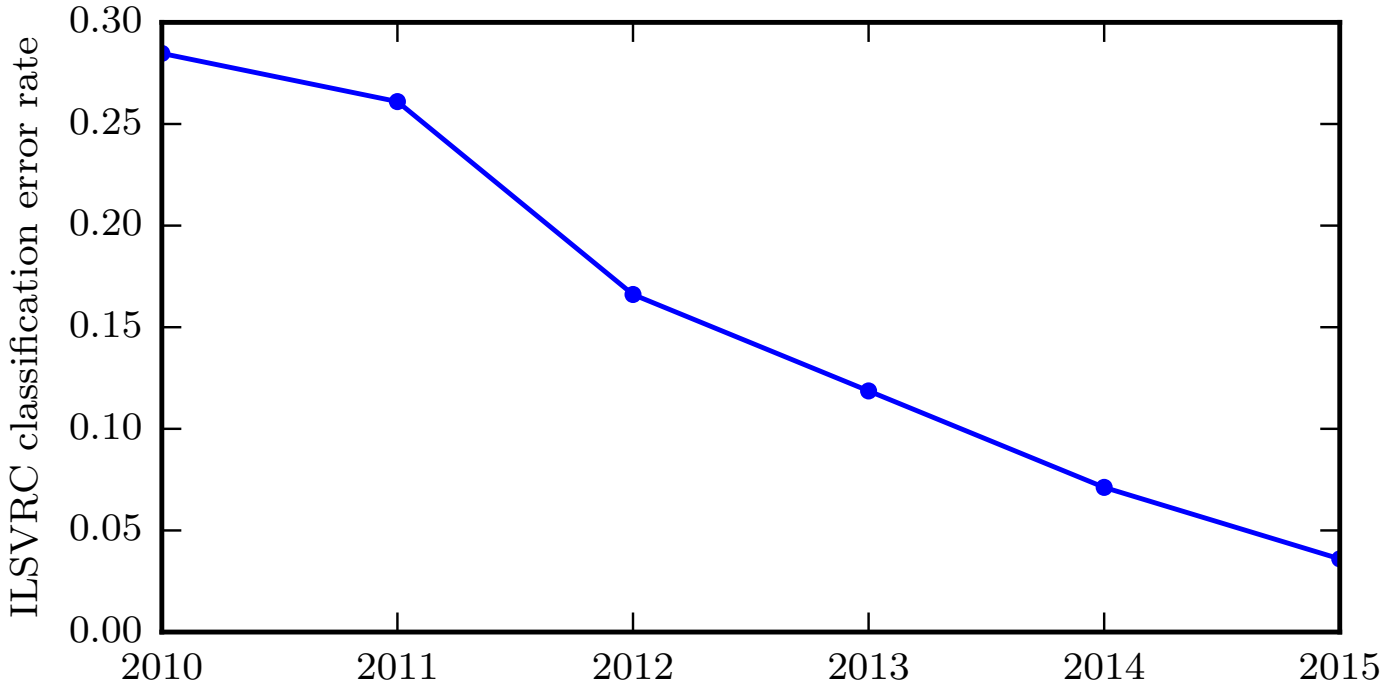


Figure 1.12

Thanks

This is the end of this lecture!

Neurons