

# Lecture 1: Introduction to Computer Vision and Deep Learning

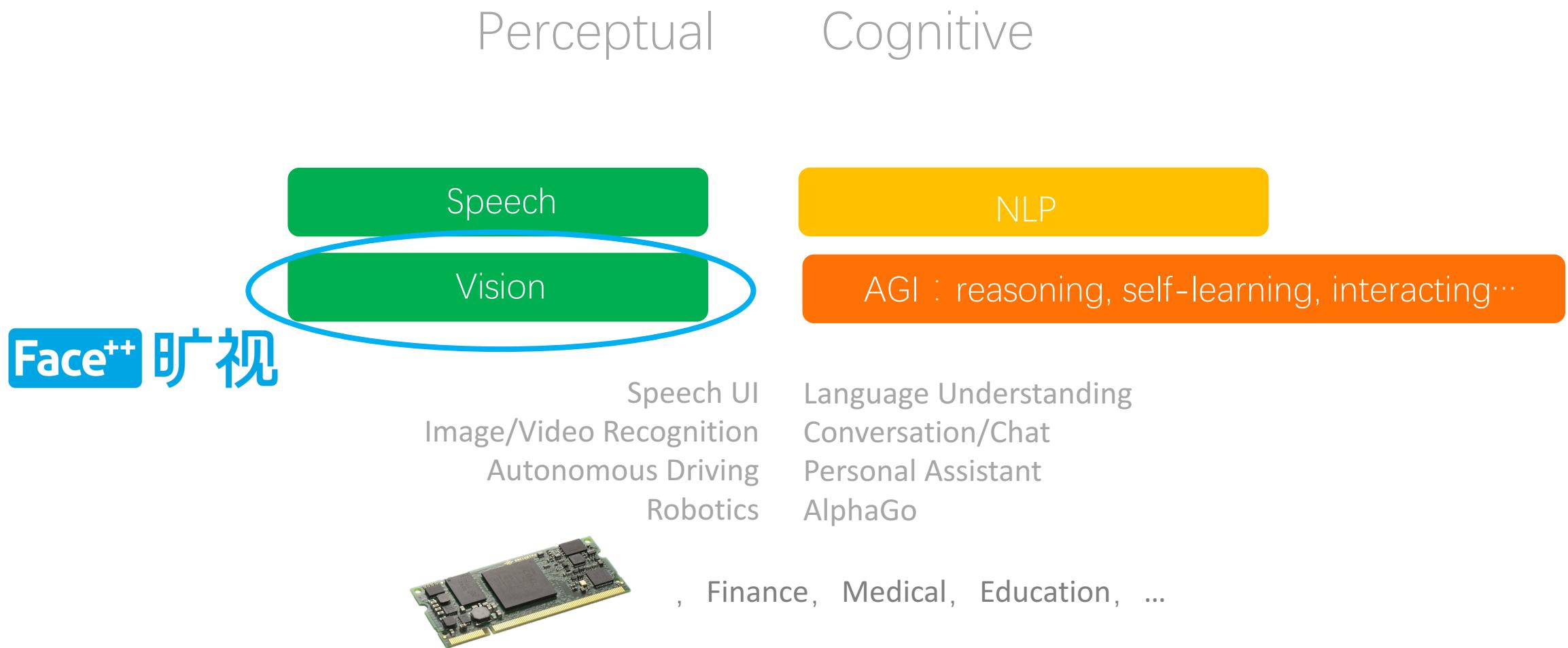
孙剑, Face++ 旷视科技首席科学家



# Agenda

- What is Computer Vision
- A few milestones in Computer Vision
- Deep learning for Computer Vision
- Applications
- Course overview

# Computer Vision in AI





## 10 Breakthrough Technologies 2017

**T**hese technologies all have staying power. They will affect the economy and our politics, improve medicine, or influence our culture. Some are unfolding now; others will take a decade or more to develop. But you should know about all of them right now.

### Paying with Your Face

Face-detecting systems in China now authorize payments, provide access to facilities, and track down criminals. Will other countries follow?



## 50 Smartest Companies 2017



- 1 Nvidia
- 2 SpaceX
- 3 Amazon
- 4 23andMe
- 5 Alphabet
- 6 iFlytek
- 7 Kite Pharma
- 8 Tencent
- 9 Regeneron
- 10 Spark Therapeutics
- 11 Face ++



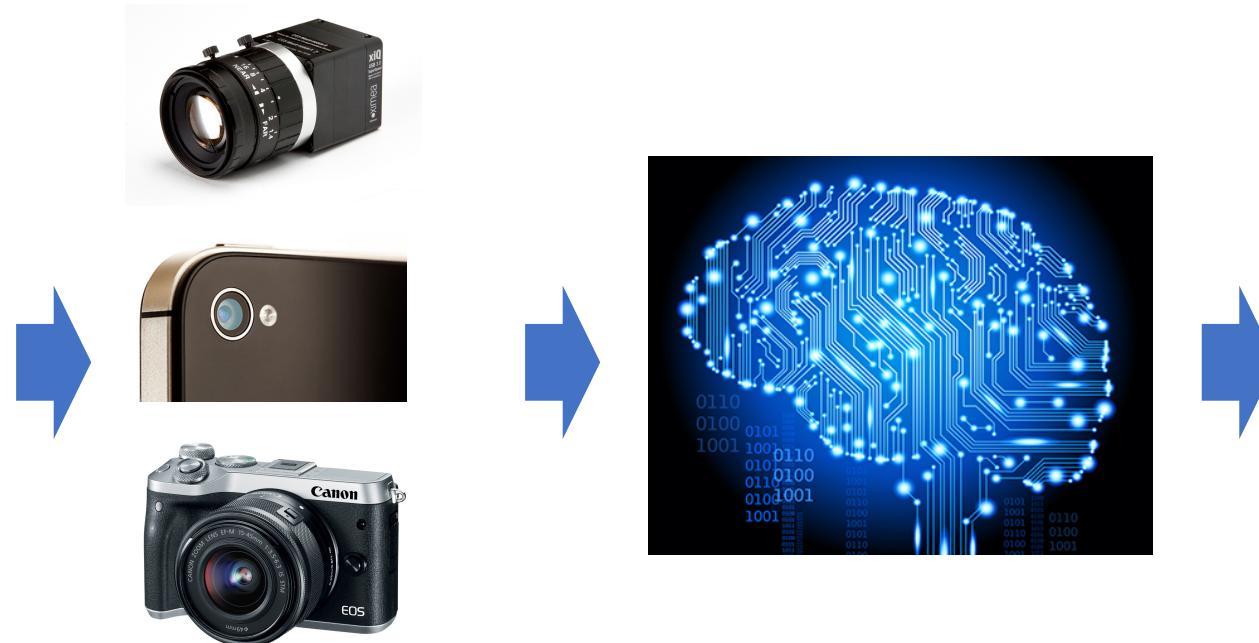
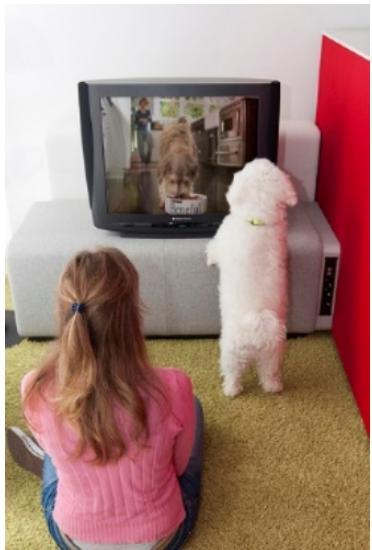
A picture is worth a thousand words



A picture is worth a thousand words

# Computer Vision

- Humans use their **eyes** and their **brains** to visually sense the world.
- Computers user their **cameras** and **computation** to visually sense the world

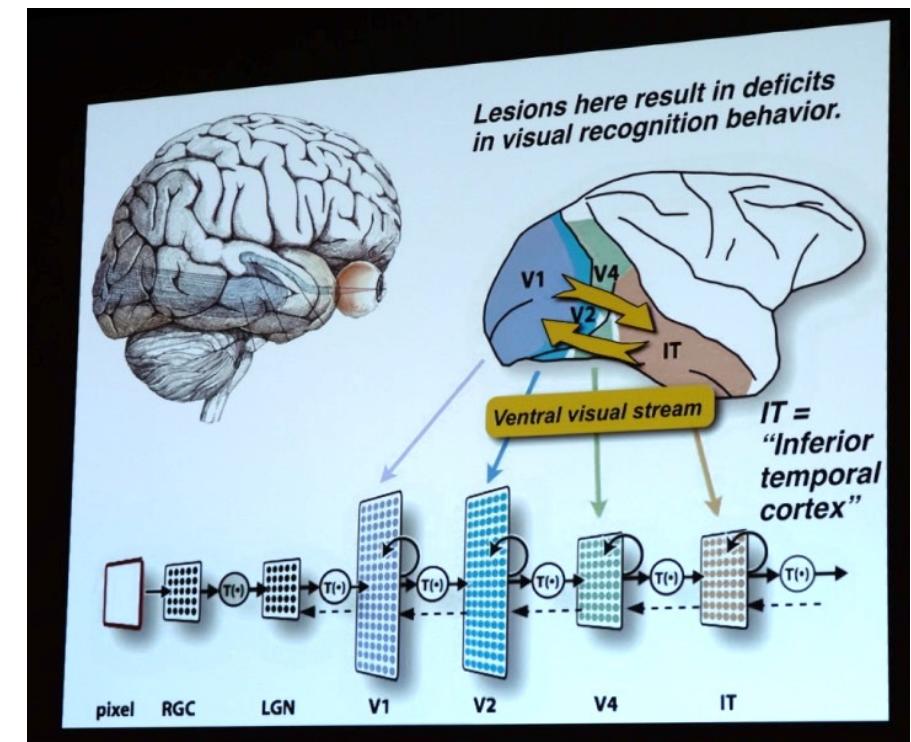
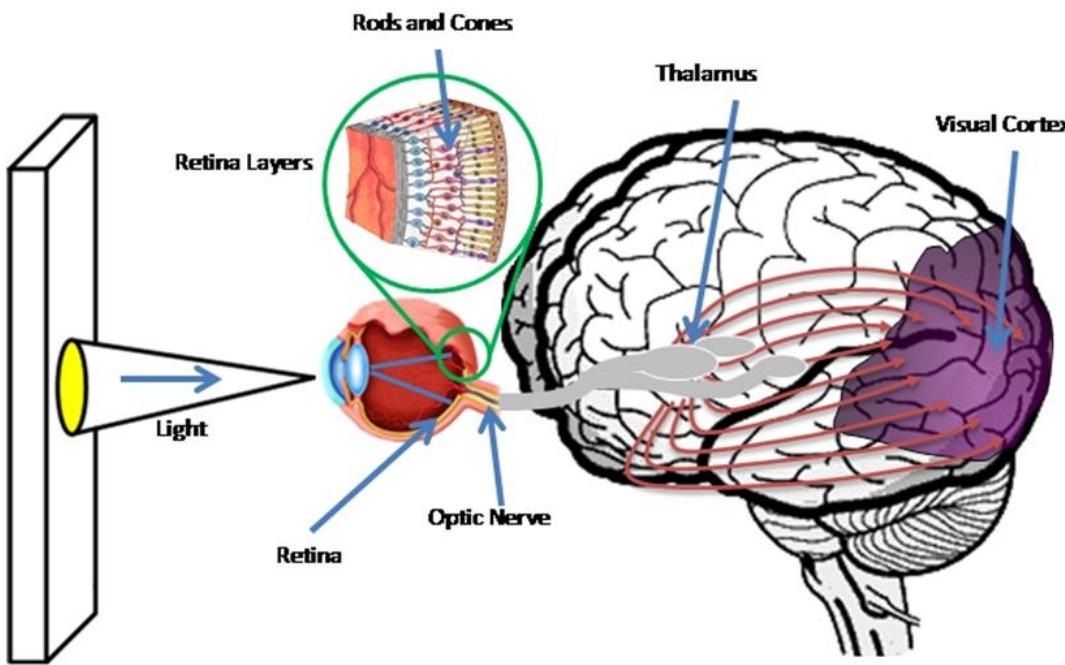


“Eyes”

“Brain”

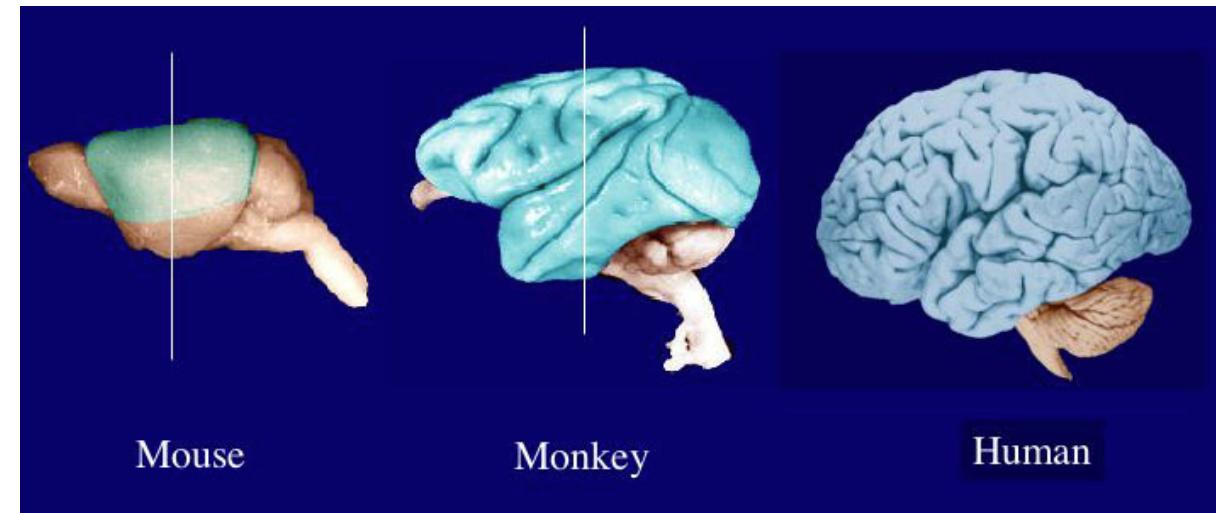
Objects  
Activities  
Scenes  
Locations  
Text  
Faces  
Gestures  
Motions  
Emotions...

# Human Visual System

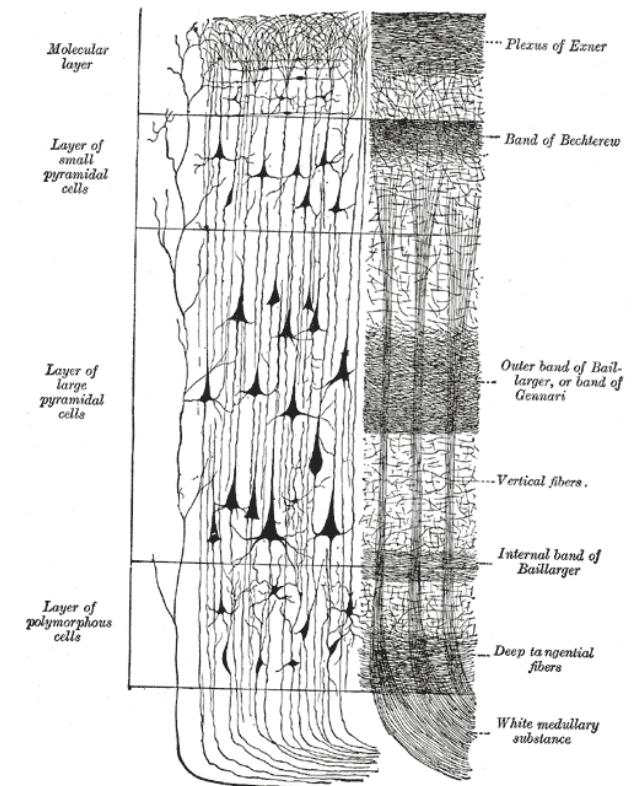


# 大脑皮层

- 6层，每层2毫米厚



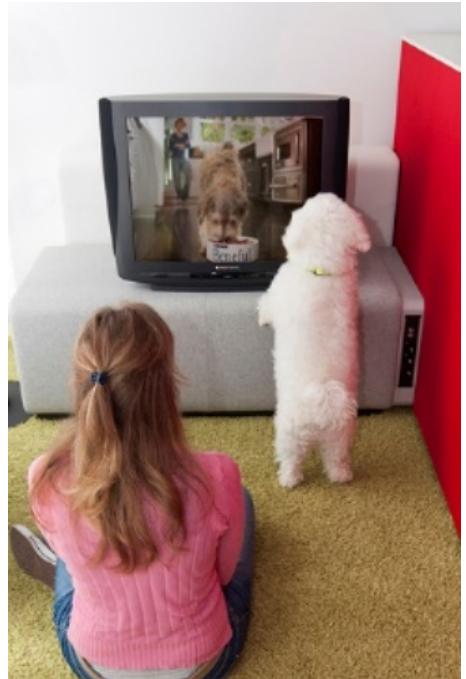
- 300亿个神经元(neurons), 每个有几千个其他相连接。他们包含了我们所有的记忆, 知识, 技能, 生活经验。物理上是毫无二至的表面, 没有明显的边界和分区。
- 层级结构：视觉有V1, V2, V4, IT区, MT负责运动检测, A1的听觉区域, S1的初级体感。还有联合区域接受多个感官的输入。M1负责向脊髓传送指令驱动肌肉



# Computer Vision and AI

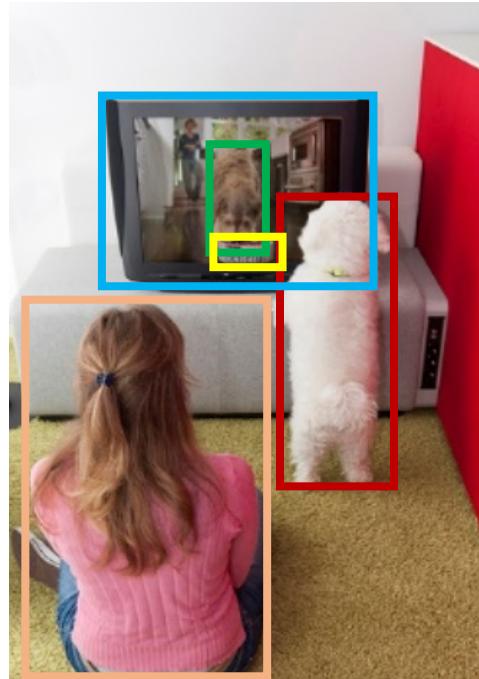
- A quest to AI
  - “If we want our machines to think, we need to teach them to see.” – Feifei Li
- A driving force of AI
  - Convolutional Neural Network
  - Batch normalization, attention, residual learning
- Key application of AI

# A Few Core Problems



Classification

Image



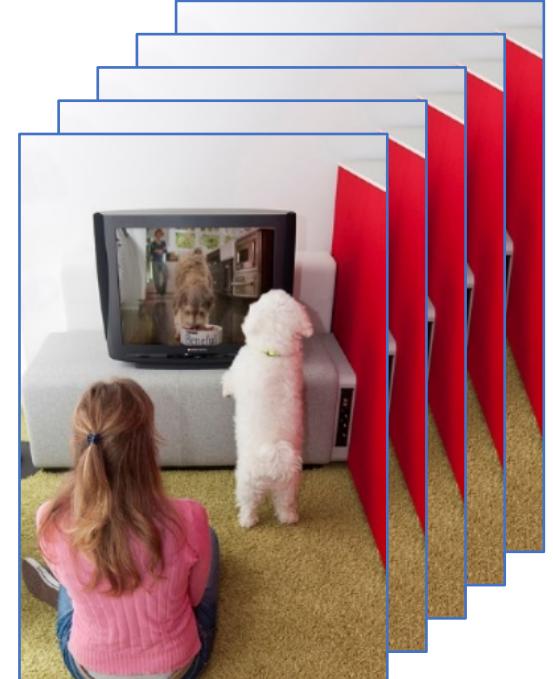
Detection

Region



Segmentation

Pixel

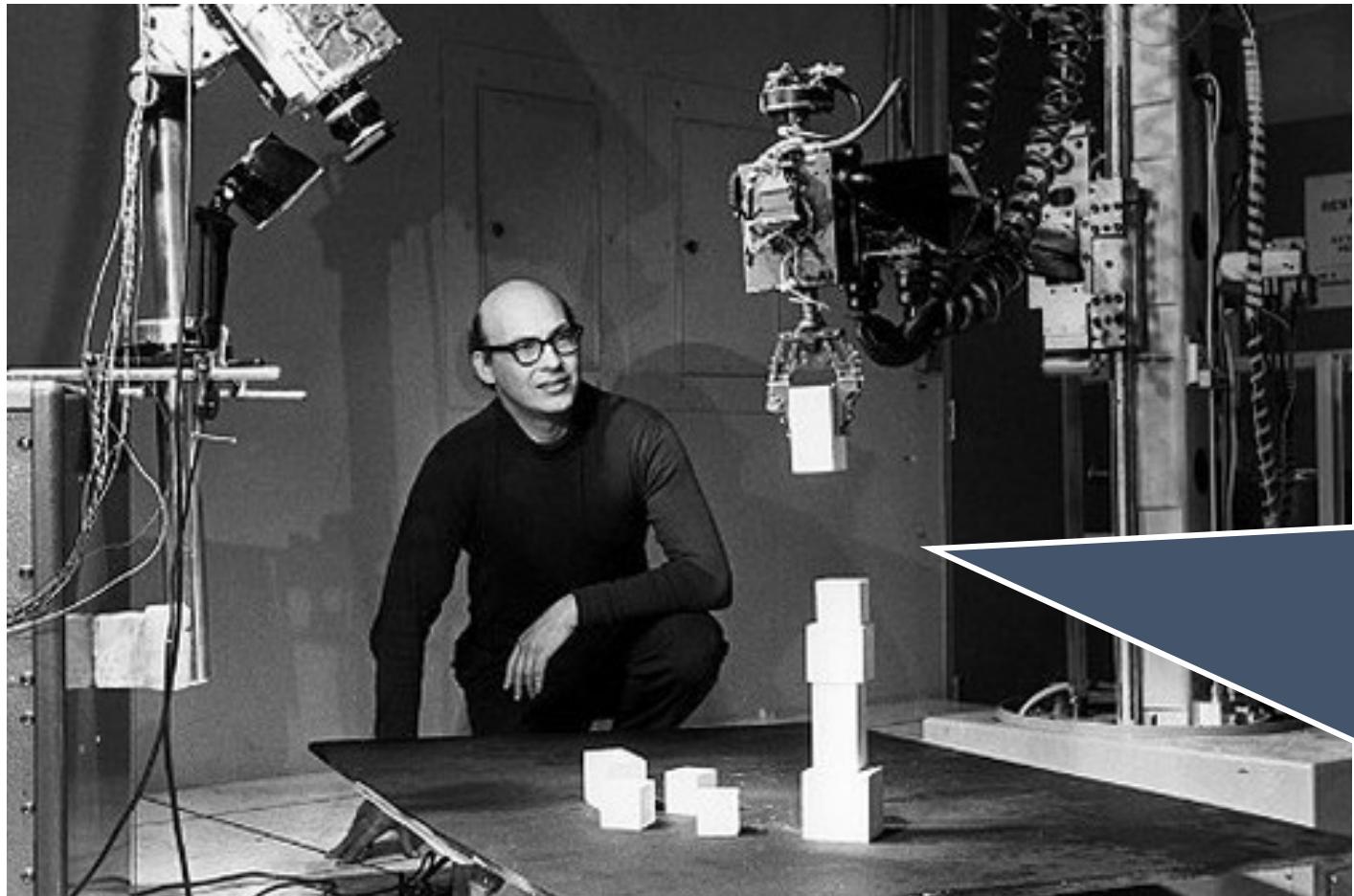


Sequence

Video

# Computer Vision

## Since the beginning of Artificial Intelligence



“Connect a television camera to a computer and get the machine to describe what it sees.”

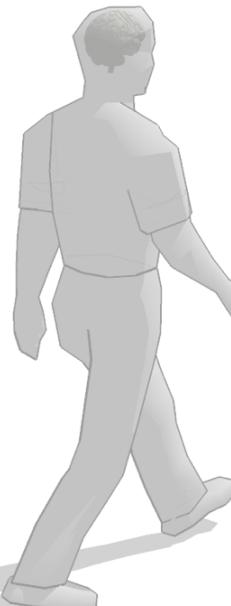
—Marvin Minsky (1966)

# David Marr

## Three levels of description (*David Marr, 1982*)

### Computational

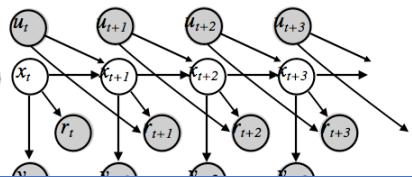
Why do things work the way they do?  
What is the goal of the computation?  
What are the unifying principles?



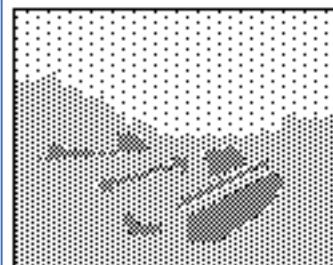
### Algorithmic

What representations can implement such computations?  
How does the choice of representations determine the algorithm?

maximize:  
 $R_t = r_{t+1} + r_{t+2} + \dots + r_T$



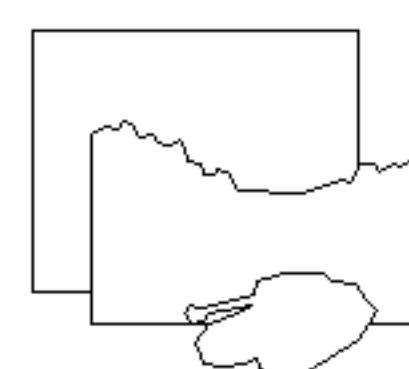
input image



edge image



2 $\frac{1}{2}$ -D sketch



3-D model



### Implementational

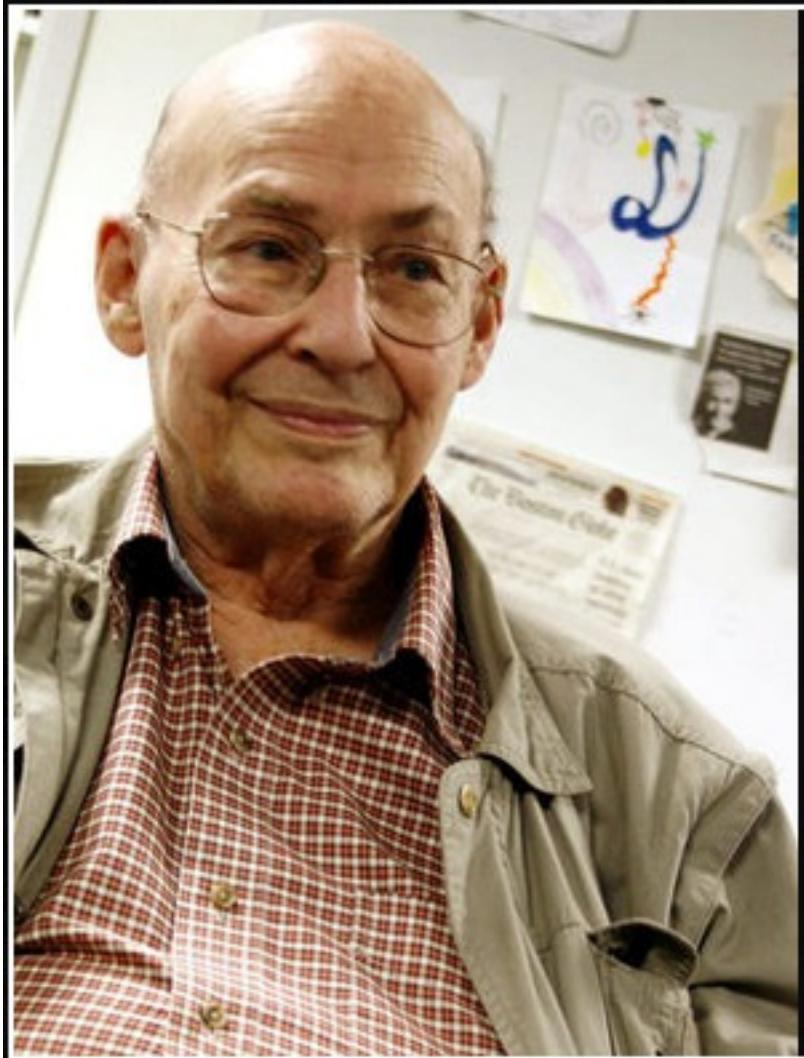
How can such a system be built in hardware?  
How can neurons carry out the computations?

# VISION



David Marr

FOREWORD BY  
Shimon Ullman  
AFTERWORD BY  
Tomaso Poggio

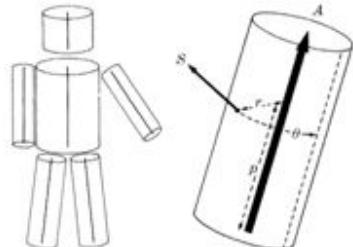


When David Marr at MIT moved into computer vision, he generated a lot of excitement, but he hit up against the problem of knowledge representation; he had no good representations for knowledge in his vision systems.

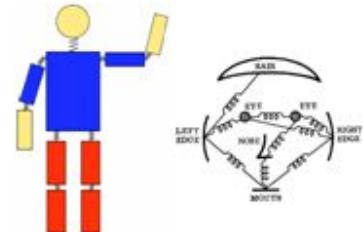
— Marvin Minsky —

AZ QUOTES

# Part Representation

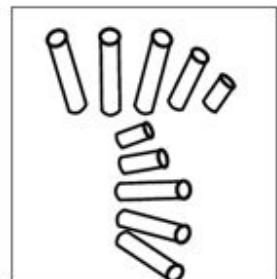


Marr & Nishihara 1978



Fischler & Elschlager 1973

Felzenszwalb & Huttenlocher 2005



## Part Representation

- Head, Torso, Arm, Leg
- Location, Rotation, Scale

## Pictorial Structure

- Unary Templates
- Pairwise Springs

Lan & Huttenlocher 2005

Sigal & Black 2006

Ramanan 2007

Epshteyn & Ullman 2007

Wang & Mori 2008

Ferrari etc. 2008

Andriluka etc. 2009

Eichner etc. 2009

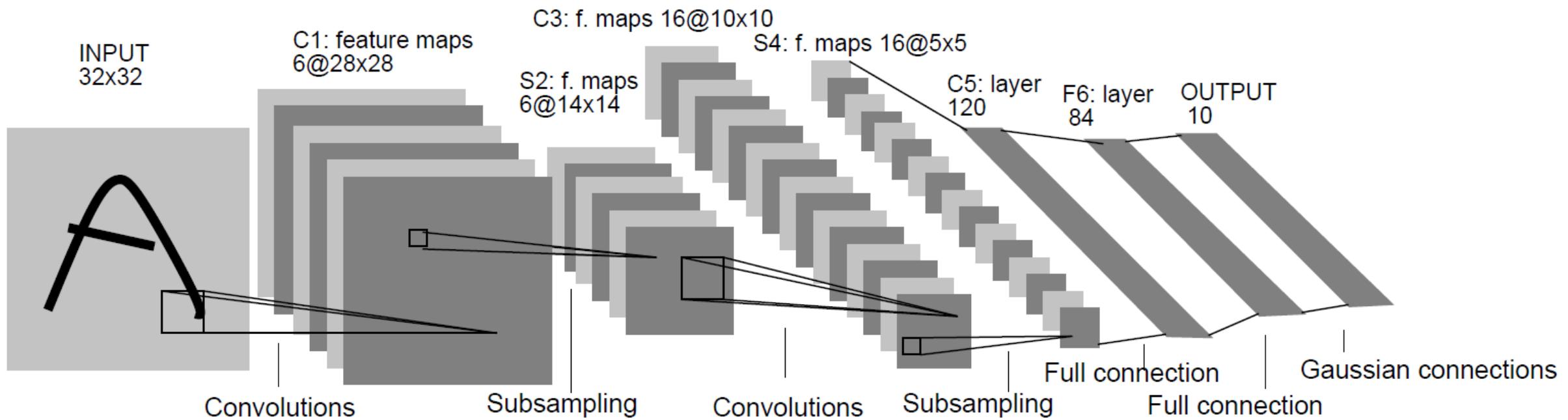
Singh etc. 2010

Johnson & Everingham 2010

Sapp etc. 2010

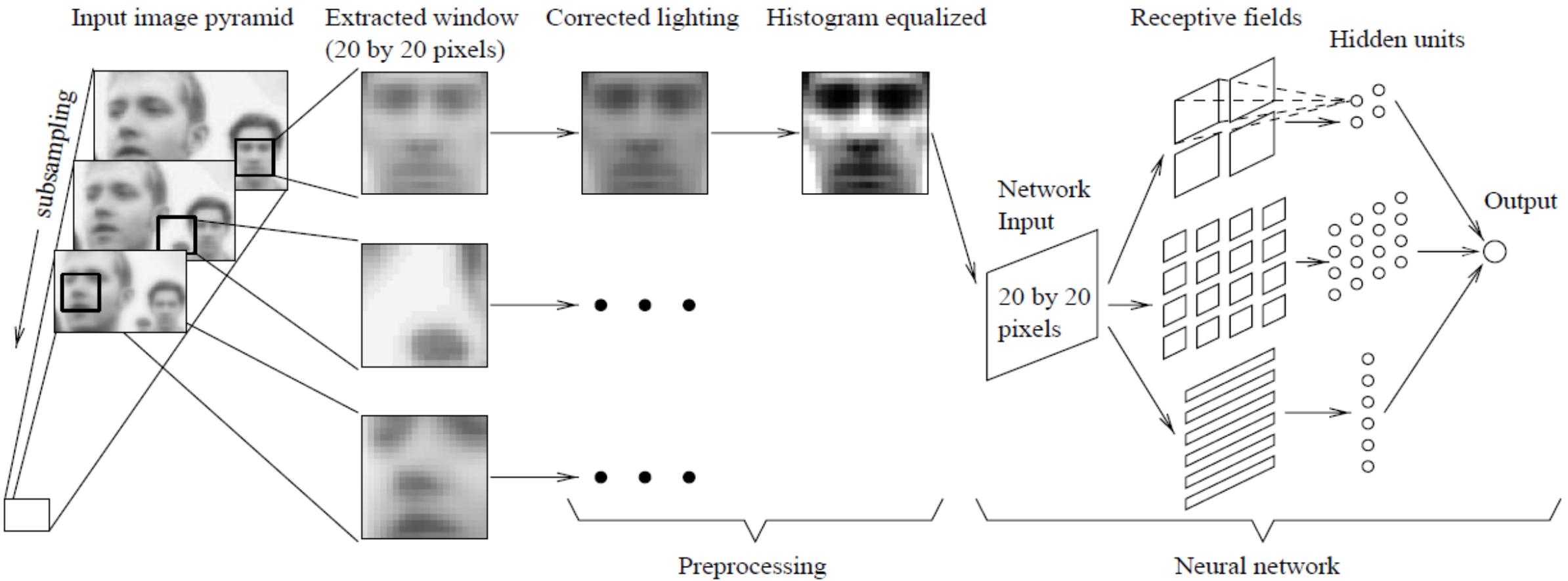
Tran & Forsyth 2010

# 1989 Digits



Source: Backpropagation Applied to Handwritten Zip Code Recognition, Lecun et al., Neural Computation 1989

# 1998 Faces



# The fall of neural networks

Became unpopular in the mid 1990s

## The rise of Support Vector Machines (SVMs)

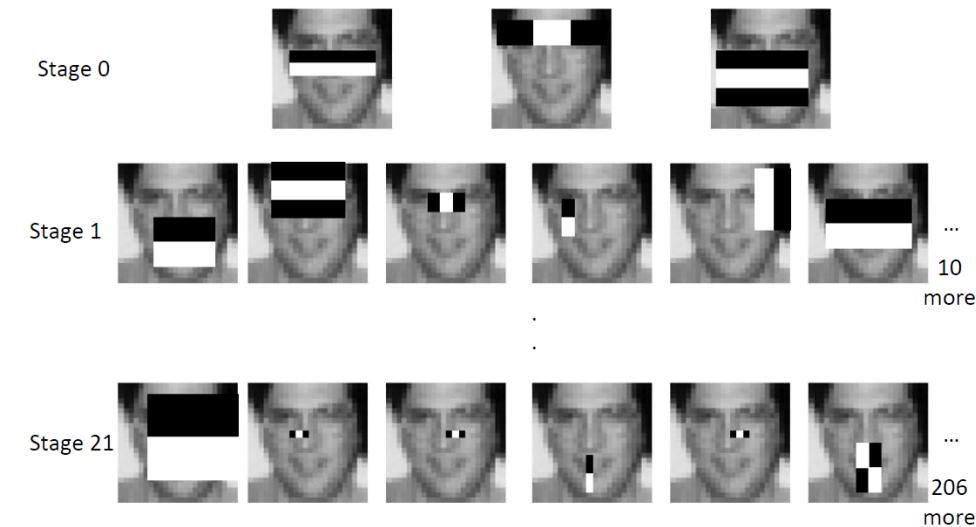
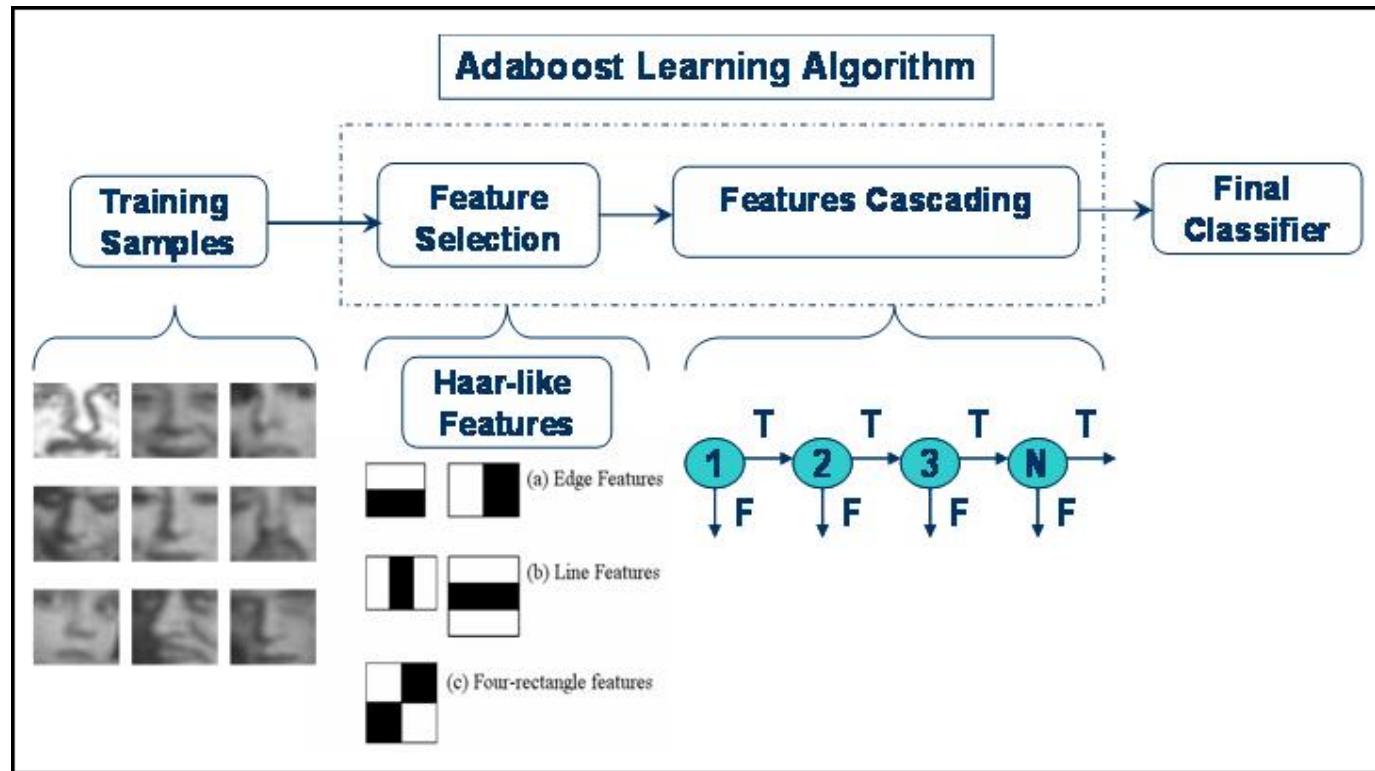
Mathematical advantages (theory, convex optimization)

## Inensitive datasets

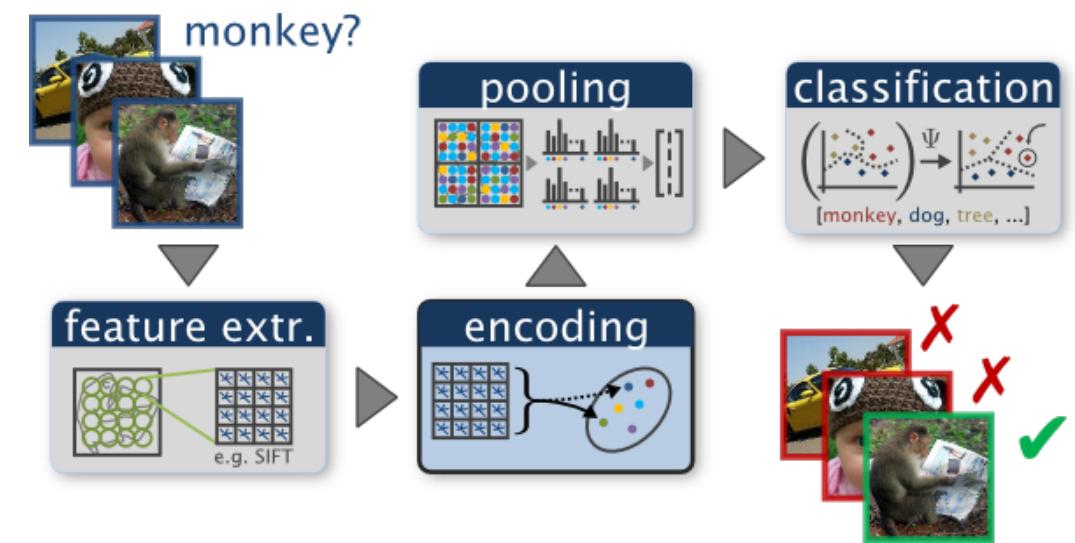
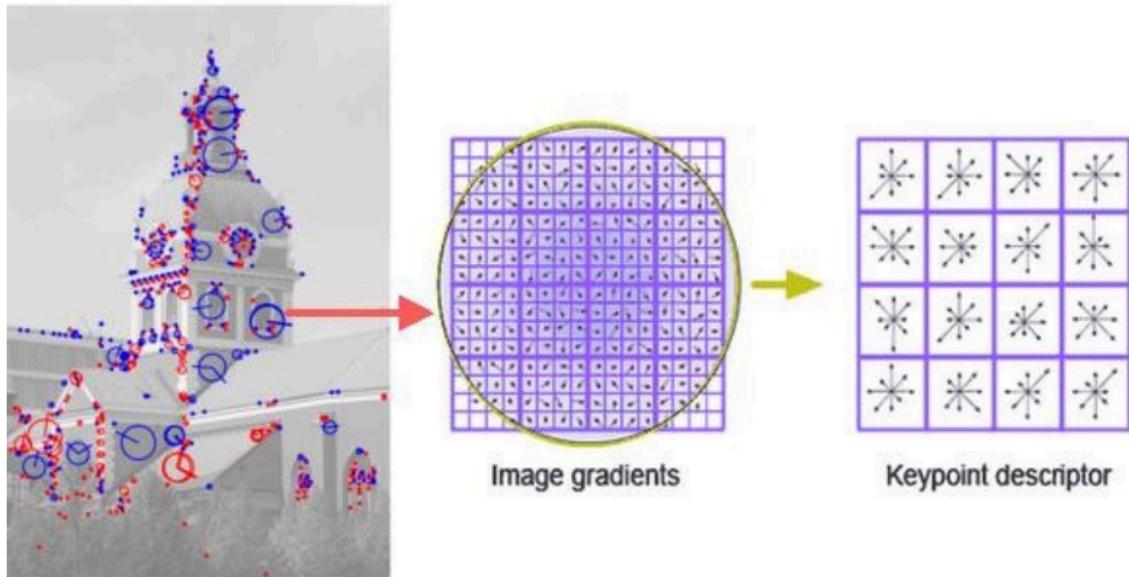
Competitive performance on MNIST digit classification

All methods look competitive

# Learning-based Representation



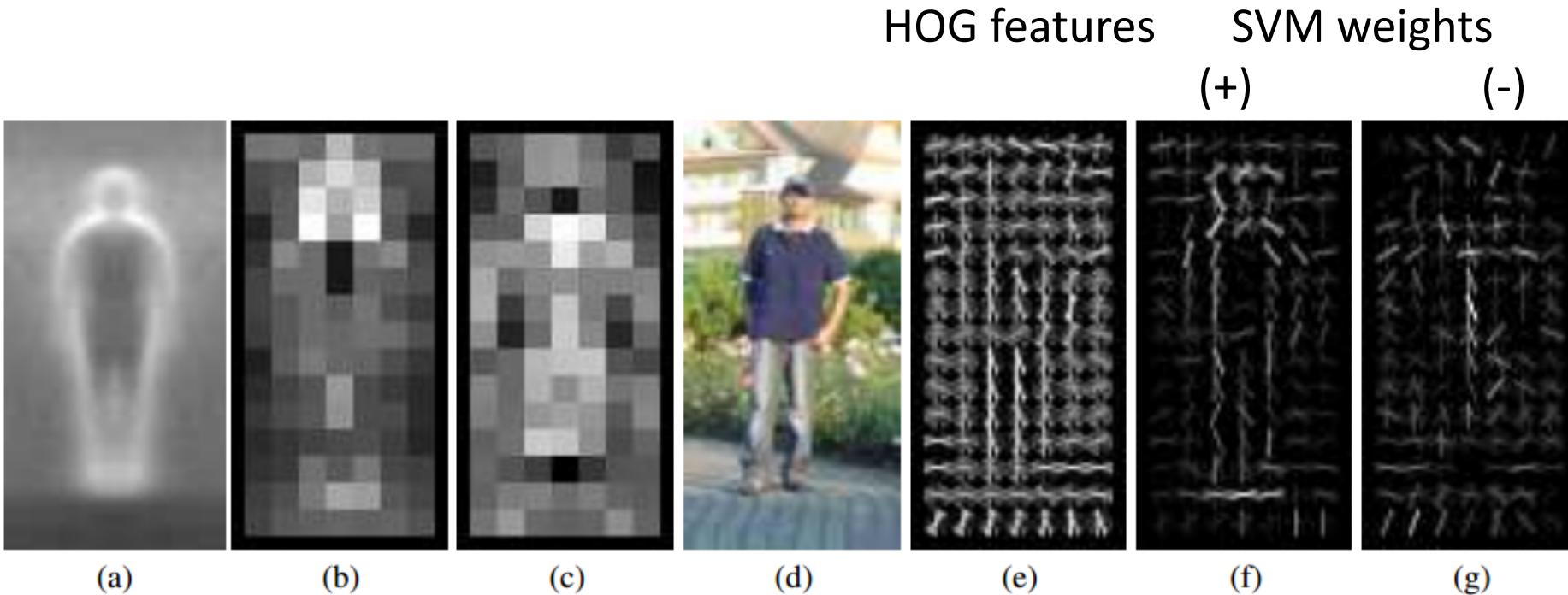
# Feature-based Representation: SIFT/HOG



# The key to SVM/Random Forest

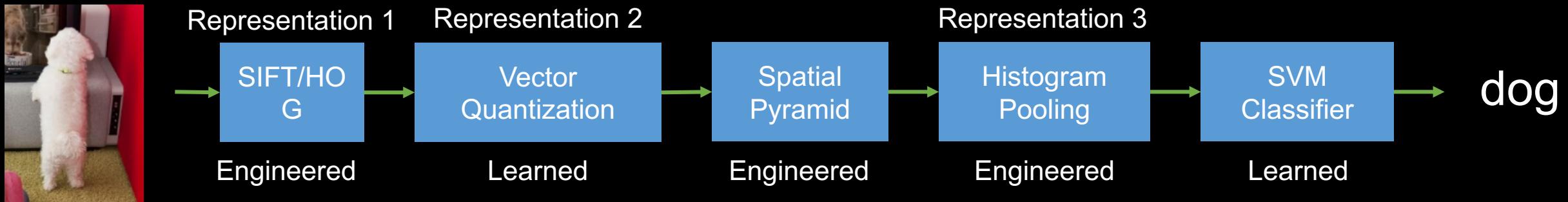
*It's all about having good features*

---



Source: Histograms of Oriented Gradients for Human Detection, Dalal and Triggs, CVPR 2005 (**9.6k** citations)  
Distinctive Image features from scale-invariant keypoints, Lowe, ICCV 1999, IJCV 2004 (**35k** citations)

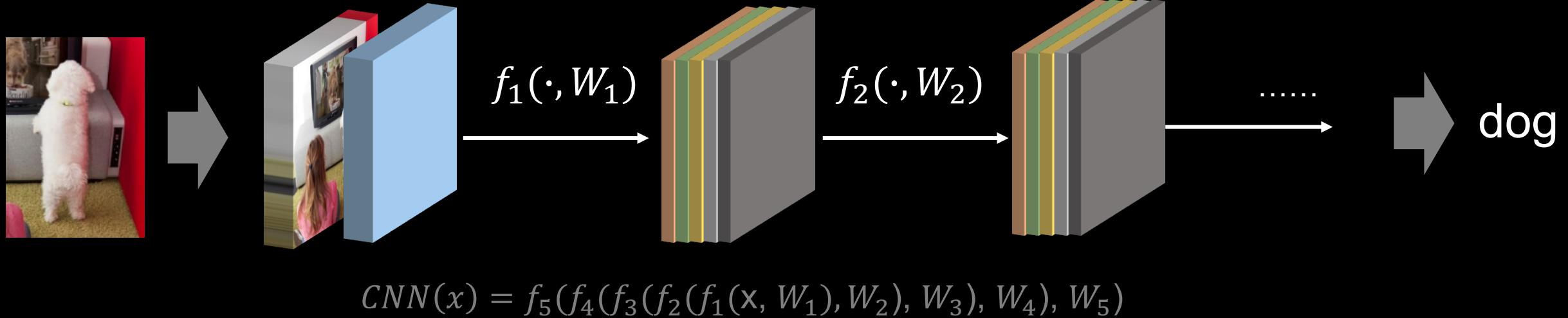
# Features Engineering + SVM/Random Forest



A short sequence of learned *non-linear* transformations

Too much “hand engineering” of features/parameters

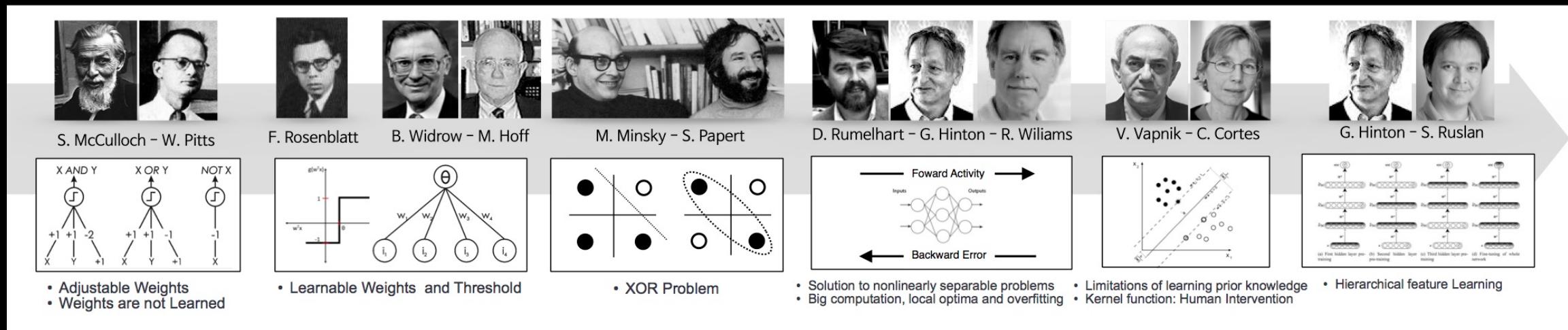
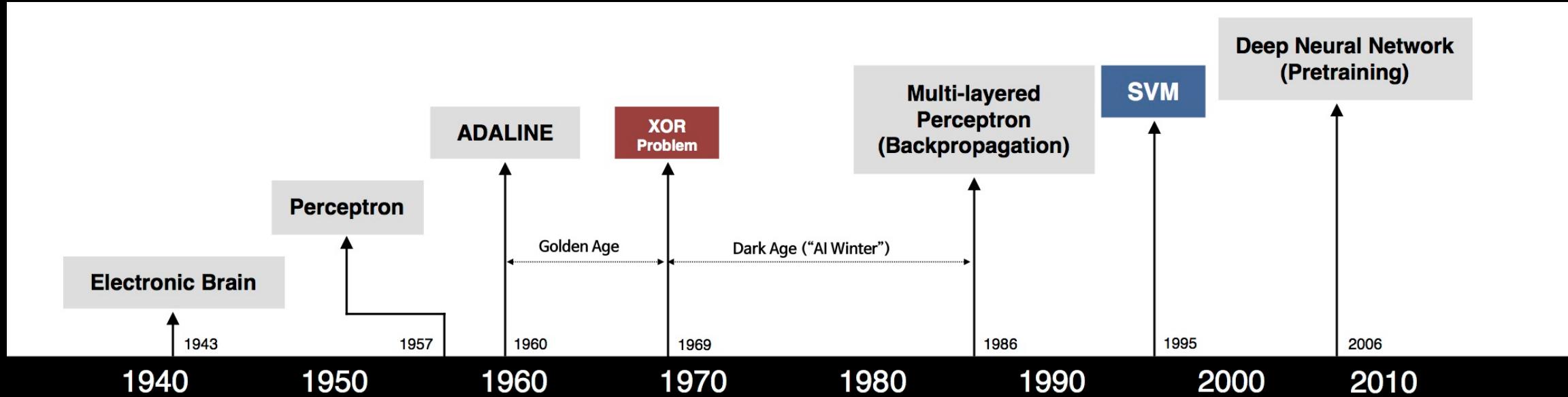
# Deep Convolutional Neural Networks



A **long** sequence of learned *non-linear* transformations

*End-to-end learning* - all of the free parameters are jointly optimized

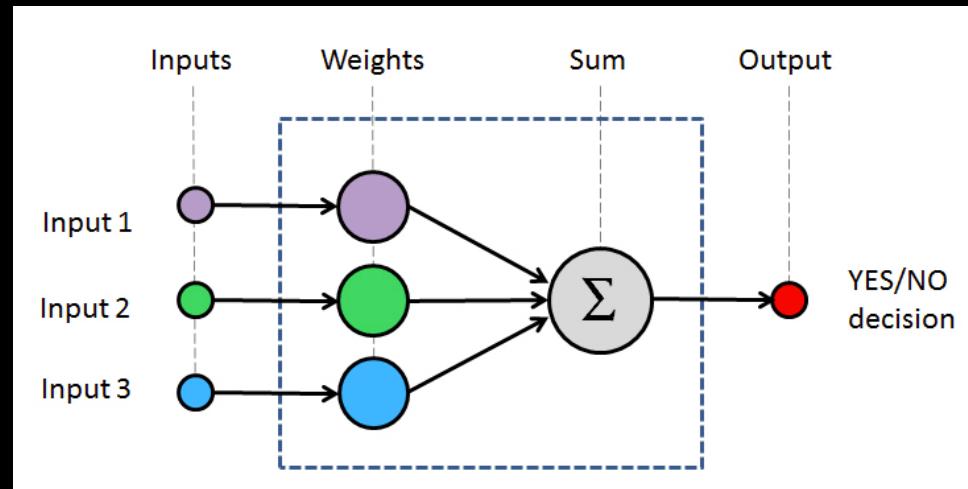
# A Bit of History



# The Perceptron (1957)

- “*The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence*”

– The New Yorker Times

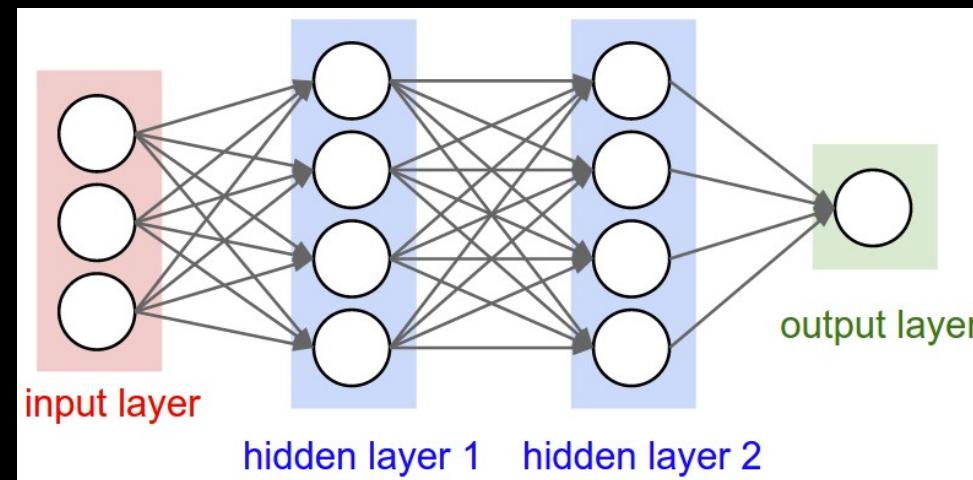


“Neuron”

A	B	
0	0	0
0	1	1
1	0	1
1	1	0

# MultiLayer Perceptron (MLP)

- “Any function can be approximated arbitrarily closely by a MLP.”



- A visual proof: <http://neuralnetworksanddeeplearning.com/chap4.html>

# Backpropagation (BP)

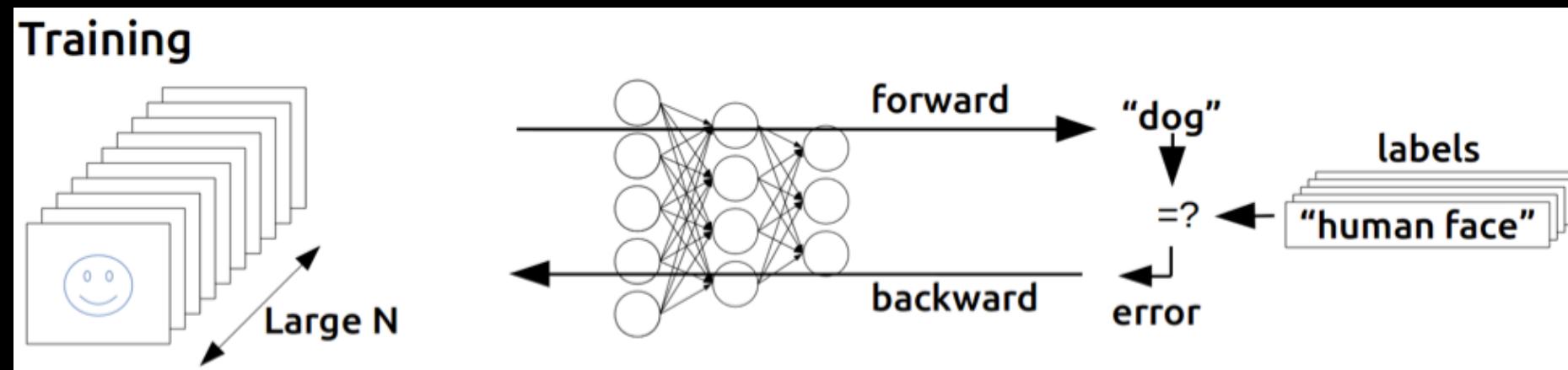
- Discovered multiple times (1968, 1982, 1986)

Learning representations  
by back-propagating errors

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

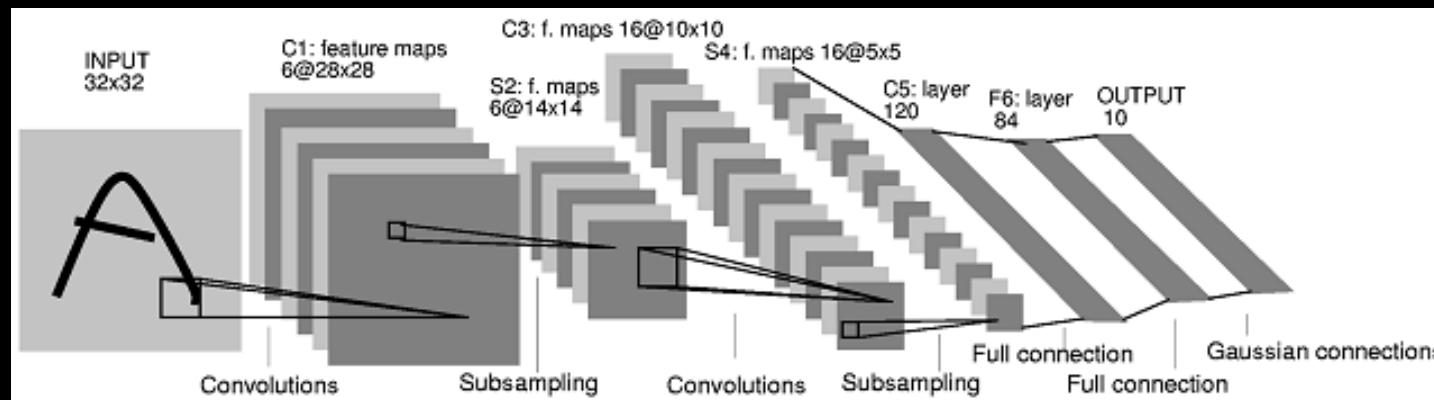
\* Institute for Cognitive Science, C-015, University of California,  
San Diego, La Jolla, California 92093, USA

† Department of Computer Science, Carnegie-Mellon University,  
Pittsburgh, Philadelphia 15213, USA



# Neural Networks in 90'

- Convolutional Neural Networks [LeCun et.al. 1989]



- Autoencoder
- Boltzmann Machine
- Belief Nets
- Recurrent Neural Networks (RNNs)

# Winter of Neural Networks (mid 90' – 2006)

- The rises of SVM, Random forest
- No theory to play
- Lack of training data
- Benchmark is insensitive
- Difficulties in optimization
- Hard to reproduce results

魔咒：

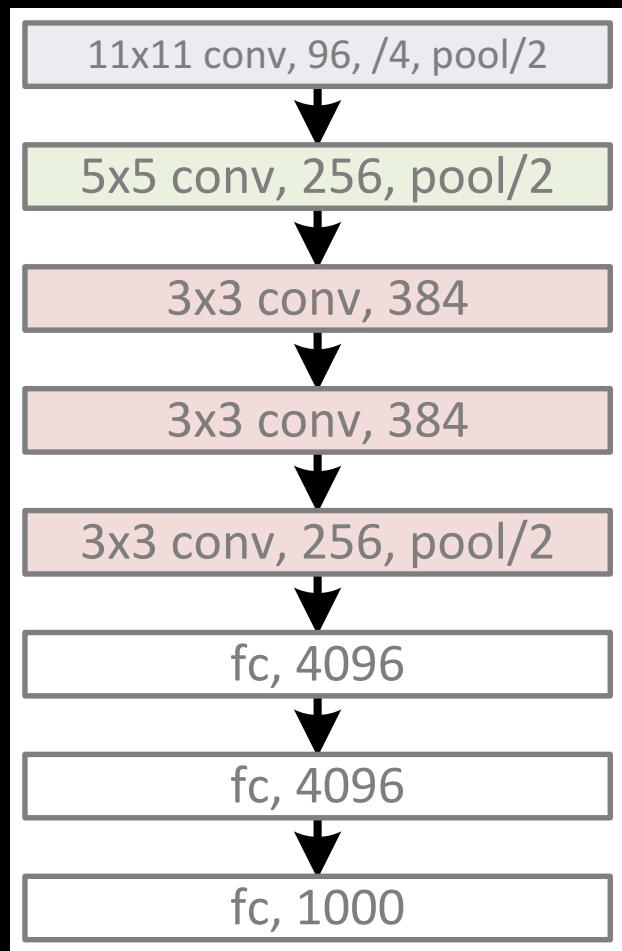
“Deep neural networks are no good and could never be trained.”

# Renaissance of Deep Learning (2006- )

- A fast learning algorithm for deep belief nets. [Hinton et.al 1996]
  - Layer-by-layer, unsupervised pre-training
- Data + Computing + Industry Competition
  - NVidia's GPU, Google Brain (16,000 CPUs)
- A few breakthroughs
  - Speech: Microsoft [2010], Google [2011], IBM
  - Image: AlexNet, 8 layers [Krizhevsky et.al 2012] (26.2% -> 15.3%)
  - NTM:

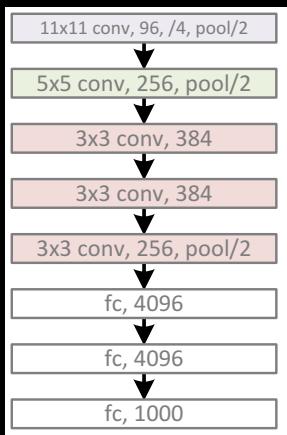
# Revolution of depth

AlexNet, 8 layers  
(ILSVRC 2012)

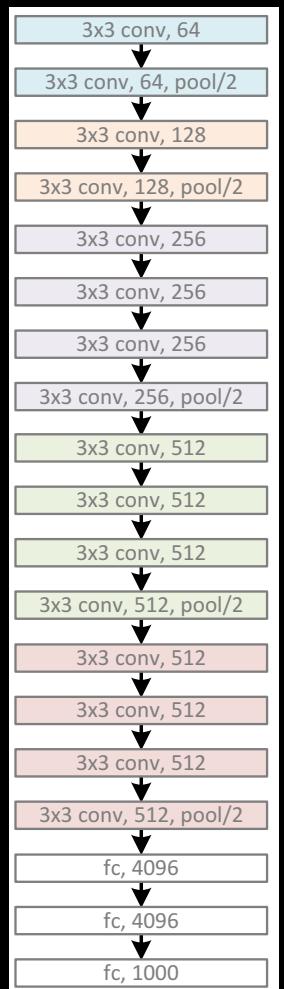


# Revolution of depth

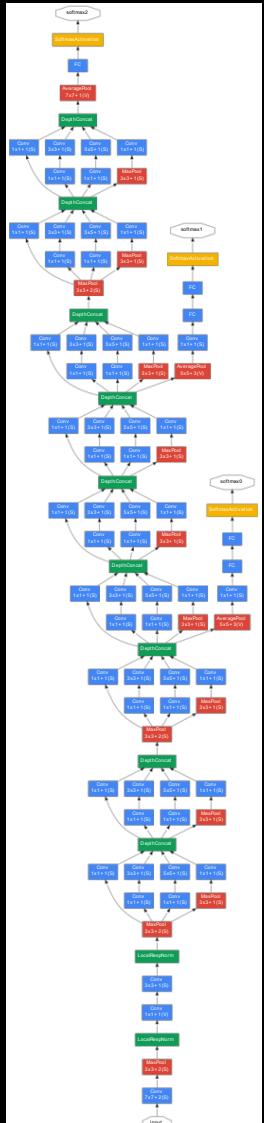
AlexNet, 8 layers  
(ILSVRC 2012)



VGG, 19 layers  
(ILSVRC 2014)



GoogleNet, 22 layers  
(ILSVRC 2014)



# Revolution of depth

AlexNet, 8 layers  
(ILSVRC 2012)



VGG, 19 layers  
(ILSVRC 2014)

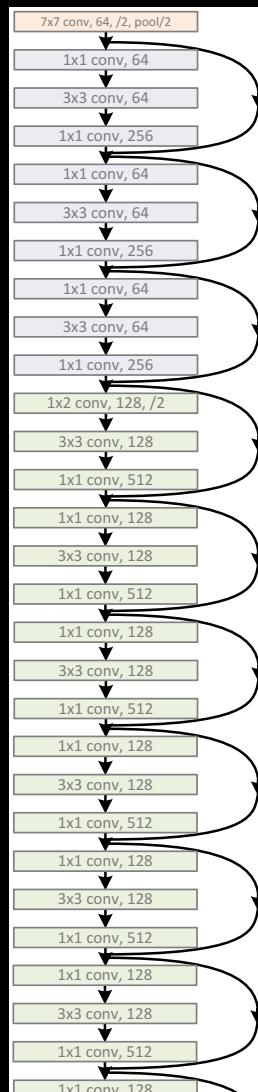


ResNet, 152 layers  
(ILSVRC 2015)

*Microsoft*

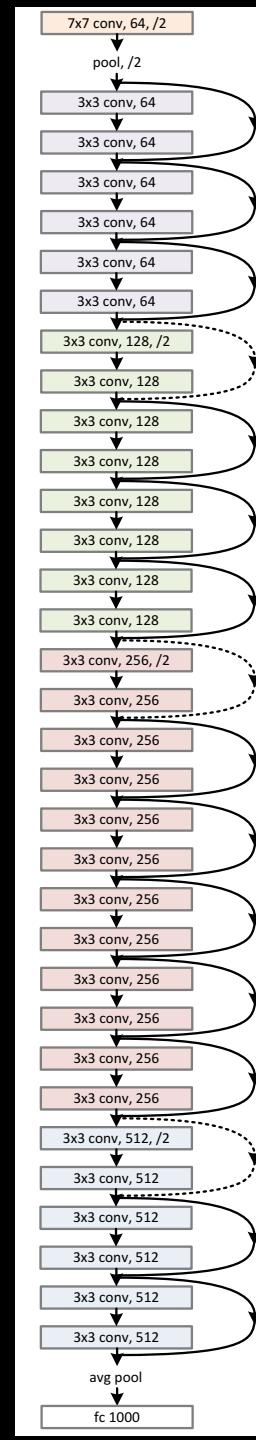
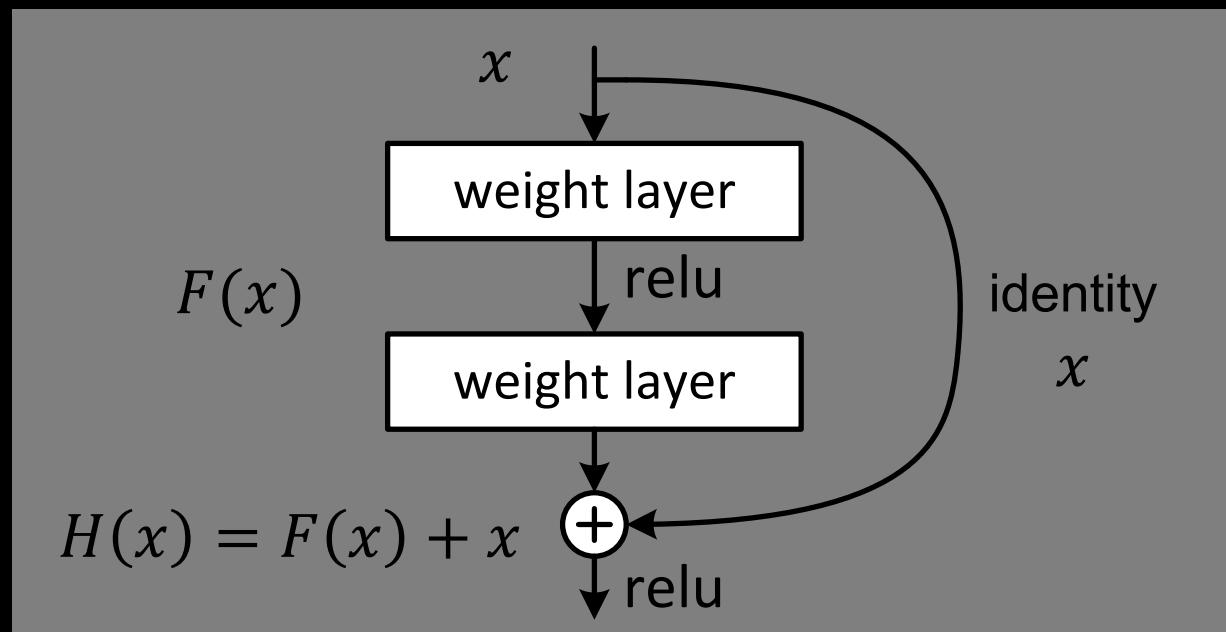


# ResNet (152 layers)



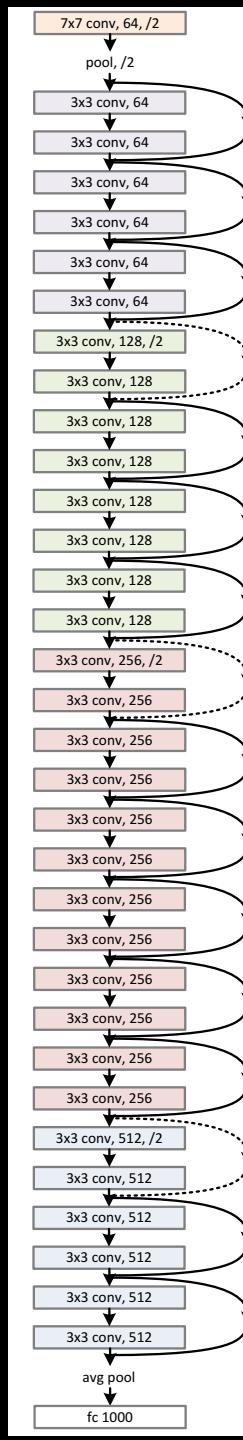
# Design of ResNet

- Key ideas:
  - skip connection = “residual function”
  - the shortest path contains only a few layers



# Design of ResNet

- Key ideas:
    - skip connection = “residual function”
    - the shortest path contains only a few layers
  - Forward view: “shallow-to-deep” dynamics in training
    - Early stage – train shallow networks
    - Later stage – train deep networks

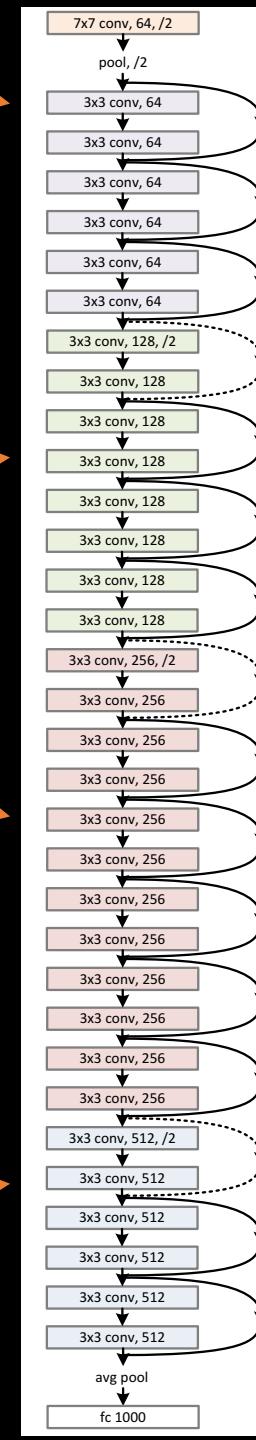


# Design of ResNet

- Key ideas:
  - skip connection = “residual function”
  - the shortest path contains only a few layers
- Backward view: identity gradient path

$$\frac{\partial \varepsilon}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L} \frac{\partial x_L}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L} \left( 1 + \frac{\partial}{\partial x_l} \sum_{i=1}^{L-1} F(x_i, W_i) \right)$$

$$\frac{\partial \varepsilon}{\partial x_L}$$



# ImageNet Large Scale Visual Recognition Challenge



# Dataset – ImageNet



14,197,122 images, 21841 synsets indexed

[Explore](#) [Download](#) [Challenges](#) [Publications](#) [CoolStuff](#) [About](#)

Not logged in. [Login](#) | [Signup](#)

ImageNet is an image database organized according to the [WordNet](#) hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.

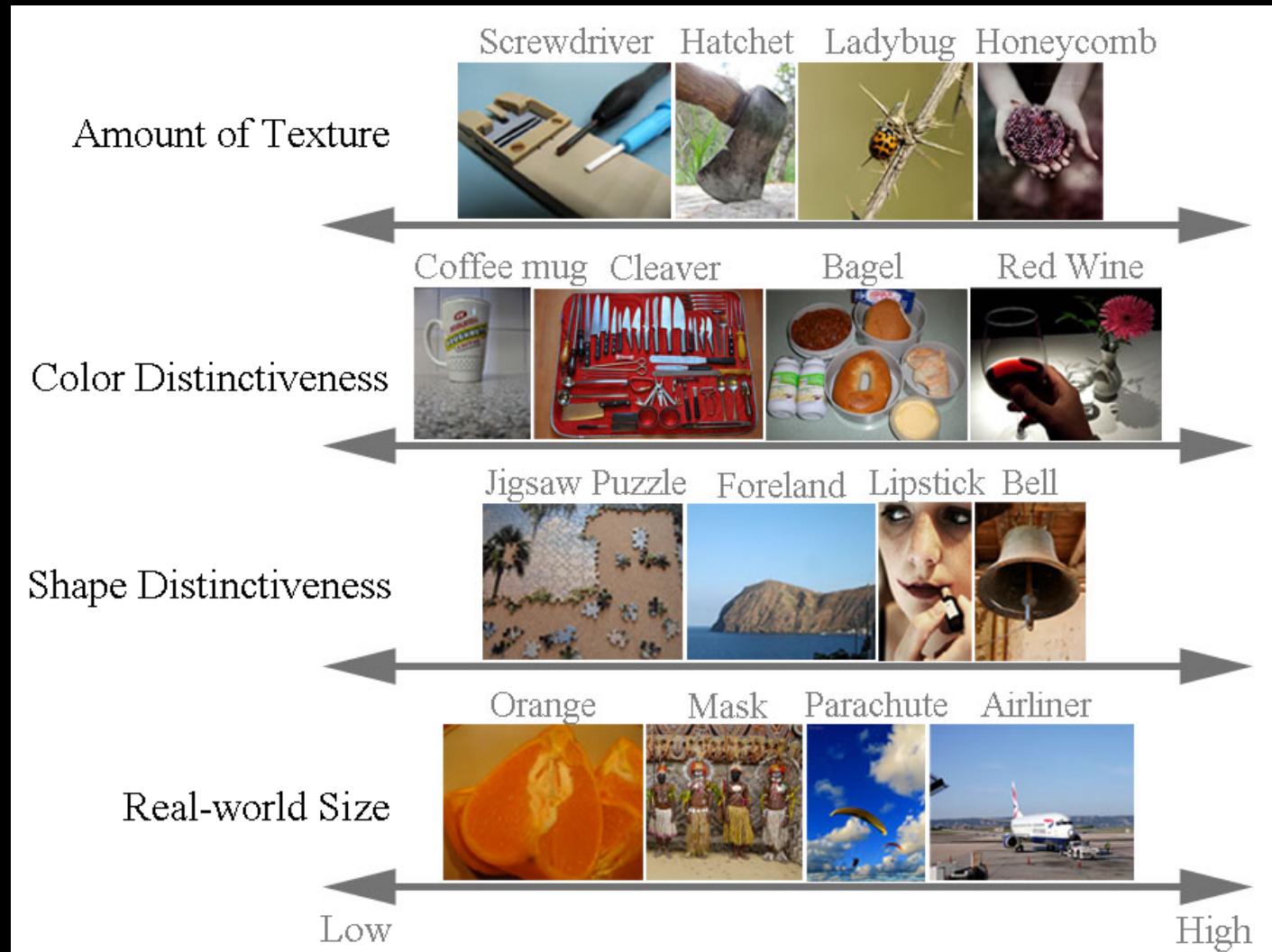
[Click here](#) to learn more about ImageNet, [Click here](#) to join the ImageNet mailing list.



What do these images have in common? *Find out!*

- An image dataset organized according to the WordNet hierarchy.
- 14 M images, 22K synsets
  - Synset - multiple words for the same visual concept
  - majority of them are nouns
- Images of each concept are quality-controlled and human-annotated

# Variety of Object Classes in ILSVRC



# ILSVRC Classification Task (top-5 error)

Steel drum



**Output:**  
Scale  
T-shirt  
Steel drum  
Drumstick  
Mud turtle



**Output:**  
Scale  
T-shirt  
Giant panda  
Drumstick  
Mud turtle



Top-5 error is to allow an algorithm to identify multiple objects in an image and not be penalized if one of the objects identified was in fact present, but not included in the ground truth.

# ILSVRC Classification Task (top-5 error)

Steel drum



Output:  
Scale  
T-shirt  
Steel drum  
Drumstick  
Mud turtle

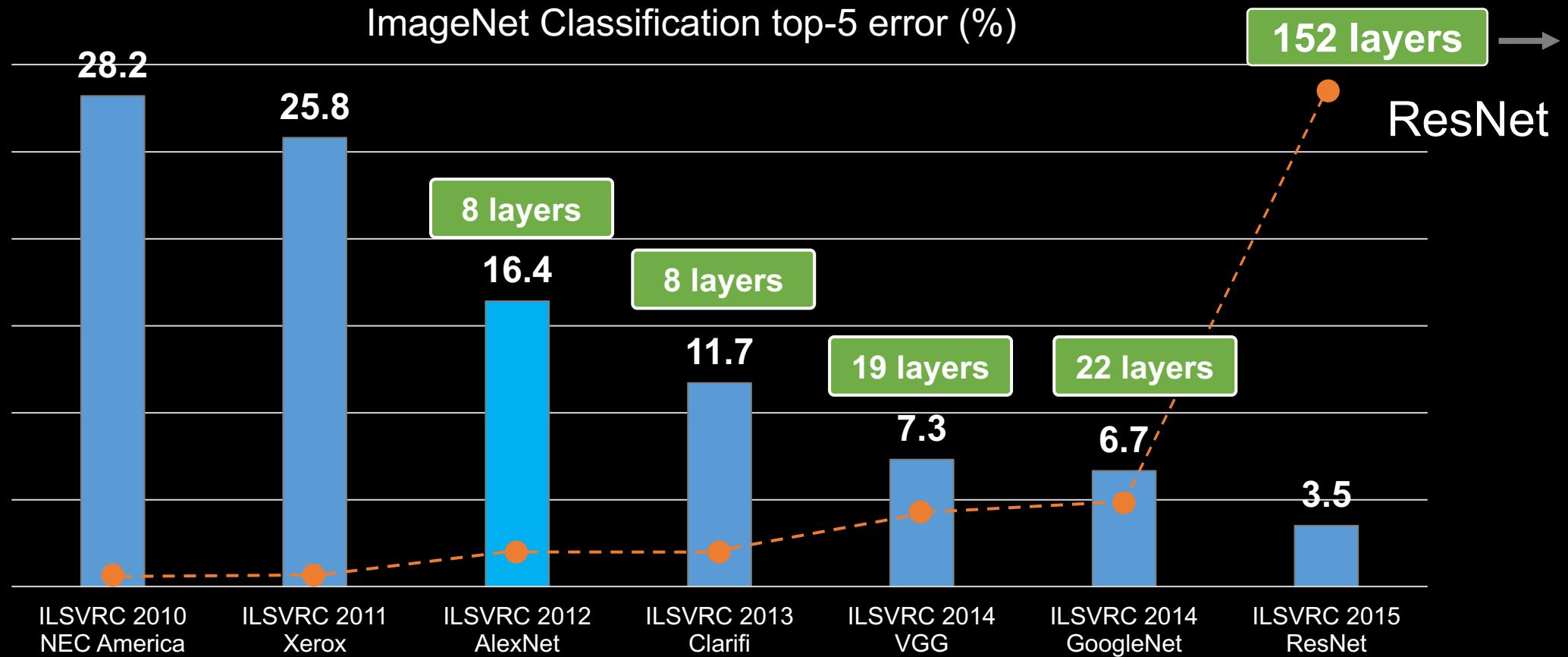


Output:  
Scale  
T-shirt  
Giant panda  
Drumstick  
Mud turtle



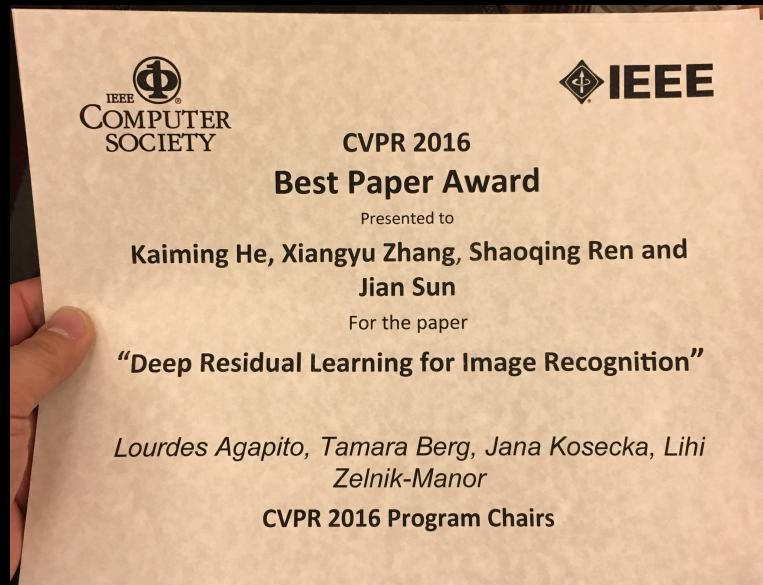
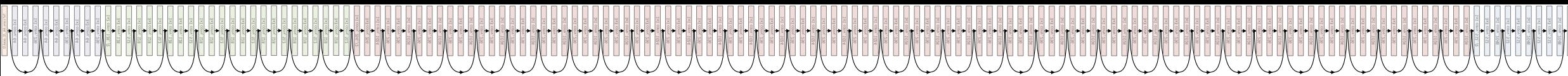
$$\text{Accuracy} = \frac{1}{100,000} \sum_{\text{100,000 images}} 1[\text{correct on image } i]$$

# ImageNet Classification: Surpassed Human



# 152层网络意味着什么？解决了什么问题？

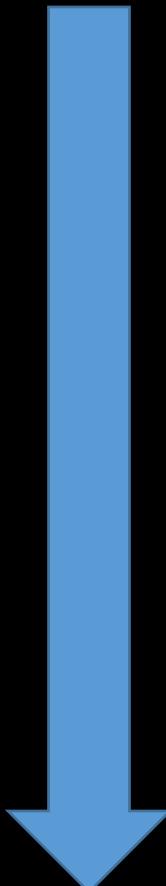
ResNet, 152 layers



# 深度学习：破除魔咒之旅

- Geoffrey Hinton's points:
  - Our labeled datasets were thousands of times too small.
  - Our computers were millions of times too slow.
  - We initialized the weights in a stupid way.
  - We used the wrong type of non-linearity.
- ResNets imply:
  - Our network architectures were optimization-unfriendly.

# Architecture Design

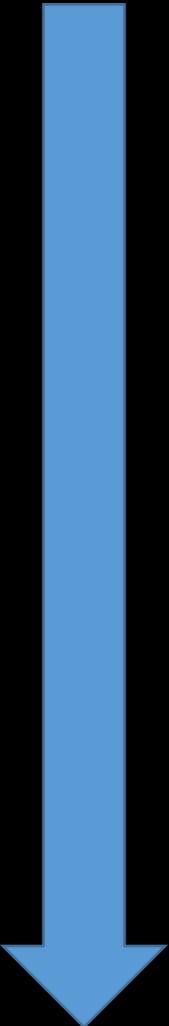
- 
- 2012
    - AlexNet [Krizhevsky et.al. 2012]
  - 2013
    - Maxout Networks [Goodfellow. et.al. 2013]
    - Network in Network [Lin et.al. 2014]
    - Deeply-supervised Nets [Lee et al. 2014]
  - 2014
    - VGG [Simonyan et al. 2014]
    - GoolgeNet [Szegedy et al. 2014]
    - Residual Networks [He et al. 2015]
  - 2015
    - Highway networks [Srivastava et al., 2015]
    - FitNets [Romero et al., 2015]
  - 2016
    - Inception-ResNet [Szegedy et al. 2016]
    - Residual Networks with pre-activation [He et al. 2016] 1001 layers

2016

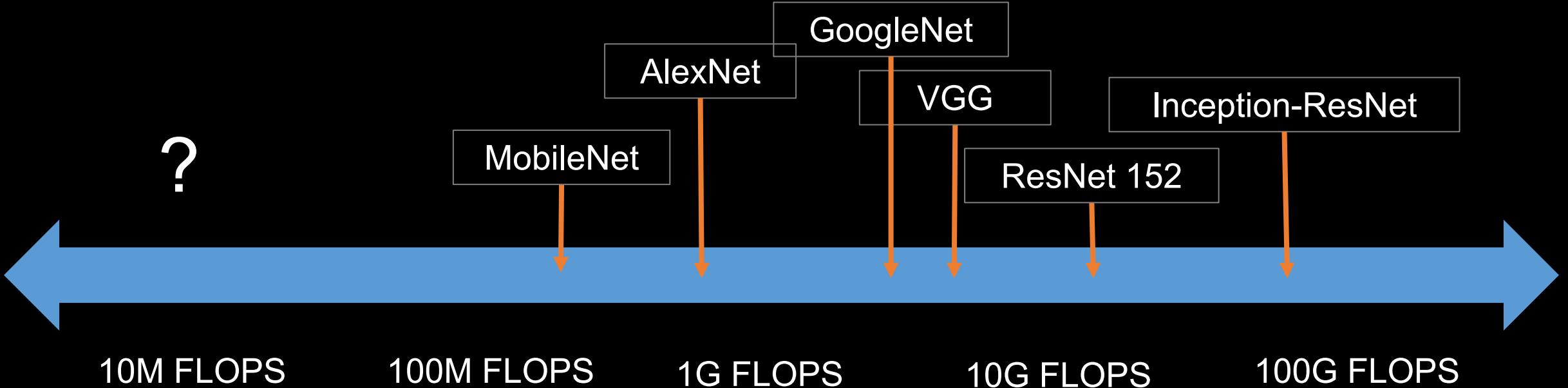
- SqueezeNet[Forrest N. Iandola et.al. 2016]
- FractalNet[Larsson et al. 2016]
- DenseNet [Huang et al. 2016]
- ResNet in ResNet [Targ el al. 2016]
- Wide Residual Networks [Zagoruyko el al. 2016]

2017

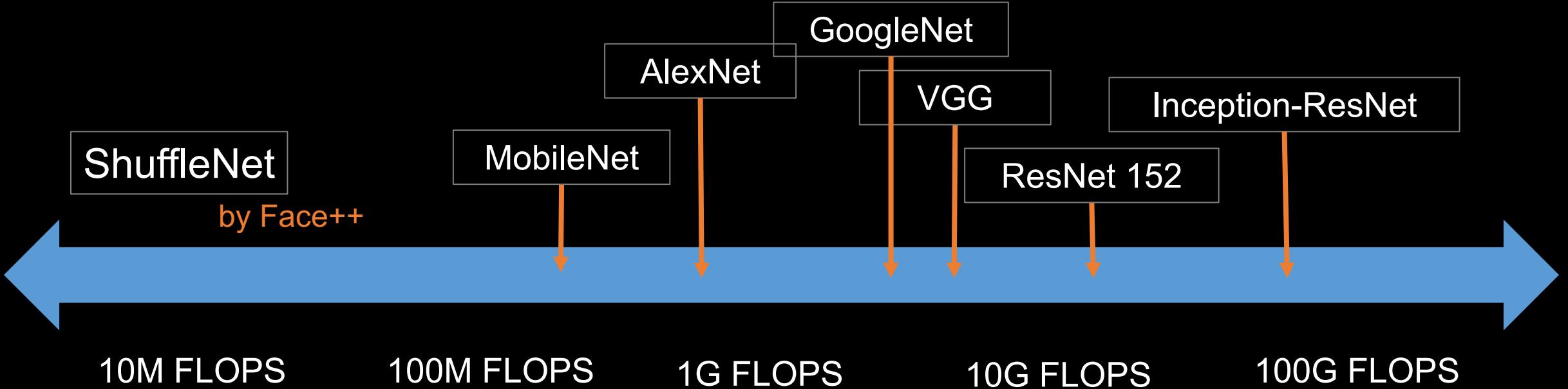
- Stochastic Depth [Huang et al. 2016]
- PolyNet [Zhang et al. 2016]
- ResNeXt [He et al. 2016]
- Xception [Chollet et al. 2016]
- BinaryNet [Matthieu et al. 2016]
- XNOR-Net [Rastegari et al. 2016]
- DeReFa-Net [Zhou et al. 2016]
- MobileNets [Howard et al. 2017]



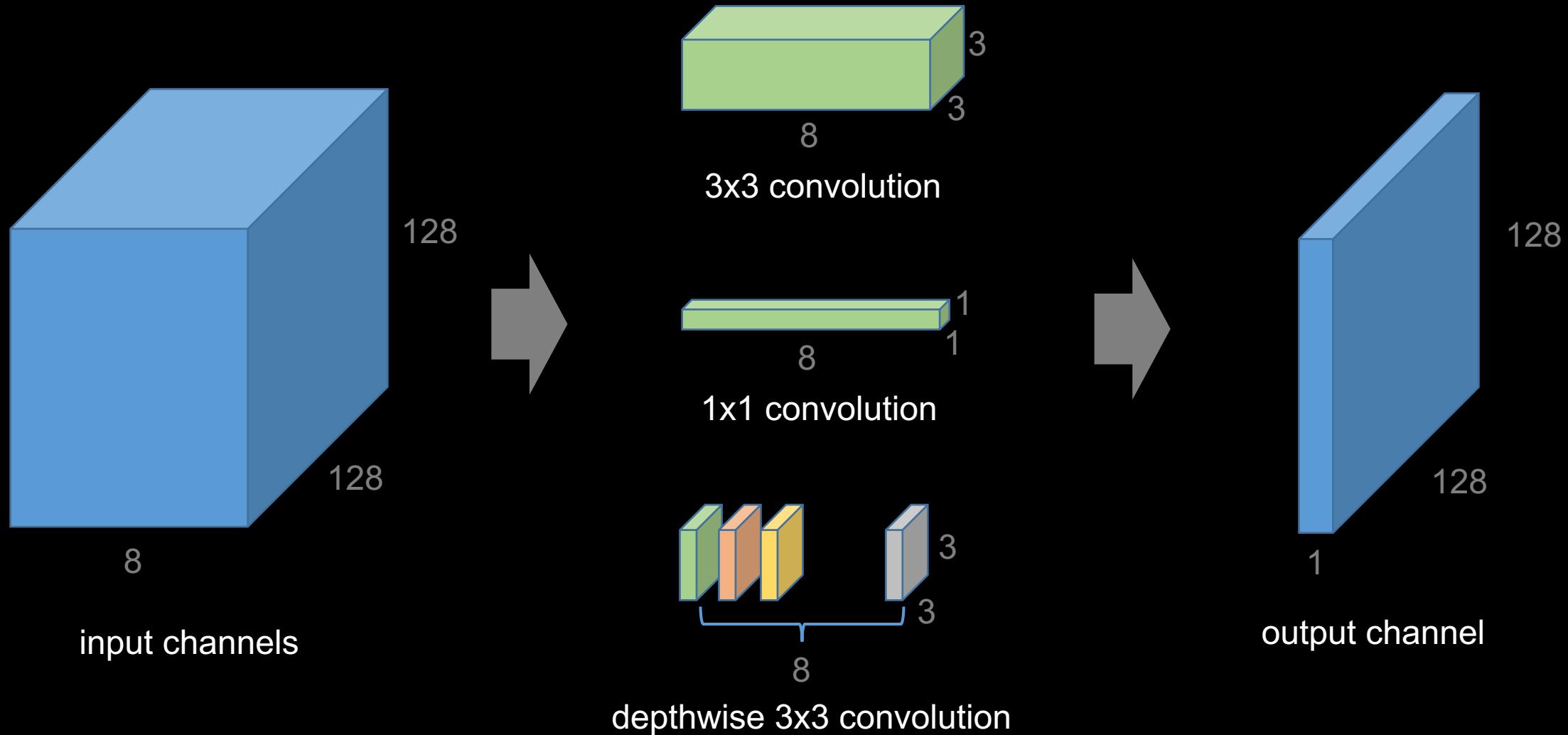
# Computing Spectrum

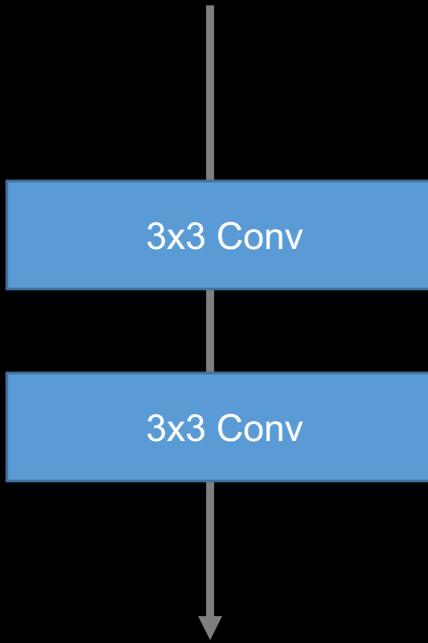


# Computing Spectrum

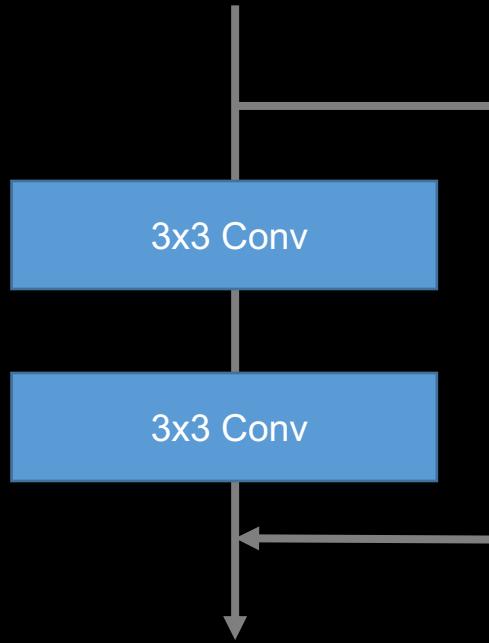


# Convolution: 3x3, 1x1, depthwise 3x3



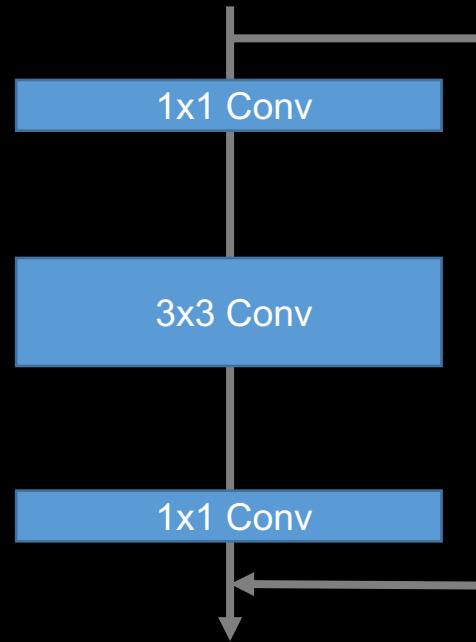


**Plain Net**



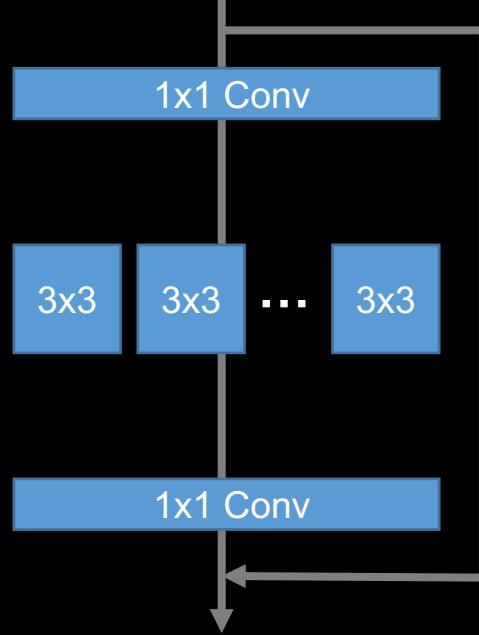
**ResNet**

[He et al. 2015]



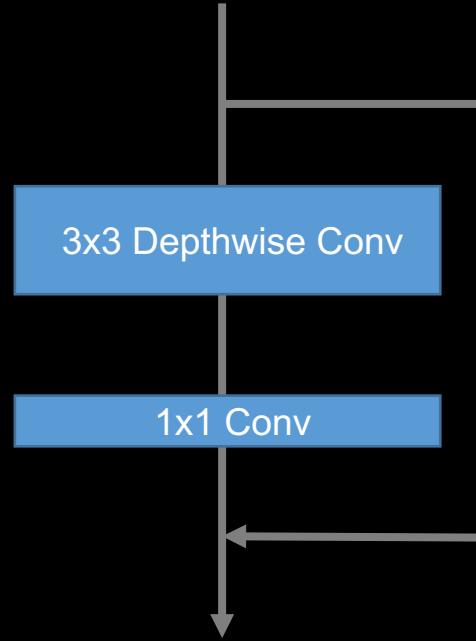
**Bottleneck**

[He et al. 2015]



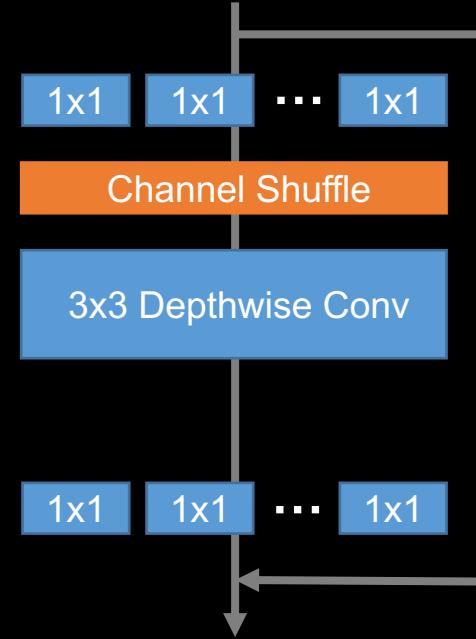
**ResNeXt**

[Xie et al. 2017]



**Xception/MobileNets**

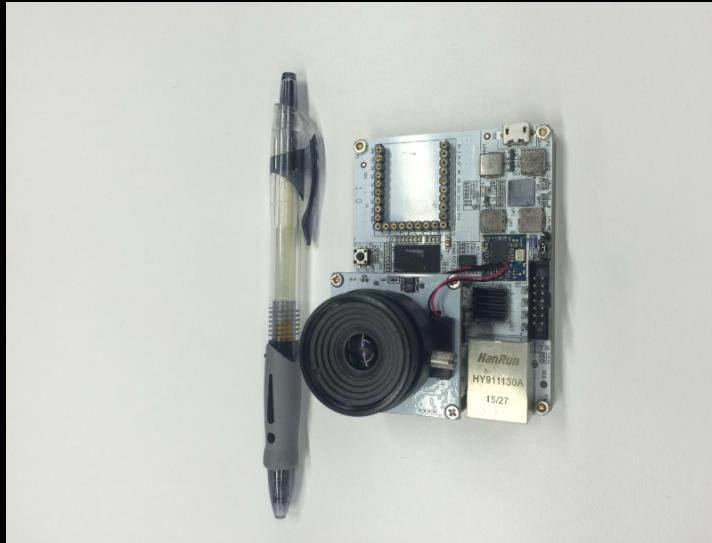
[Francois Chollet. 2017]  
[Howard et al. 2017]



**ShuffleNet**

[Zhang et al. 2017]

# ShuffleNet on FPGA



MegEye-C3S  
智能人像抓拍机

# Face Detection on FGPA, 1080p, 30fps, No Tracking

# ShuffleNet on 手机



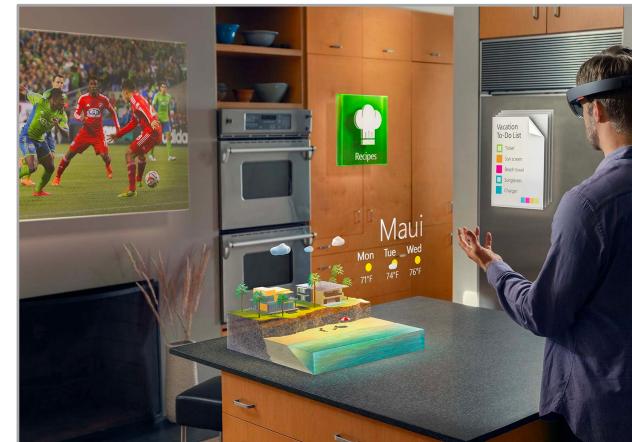
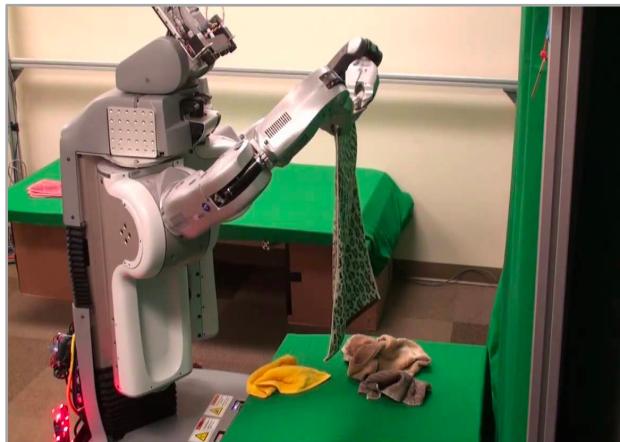
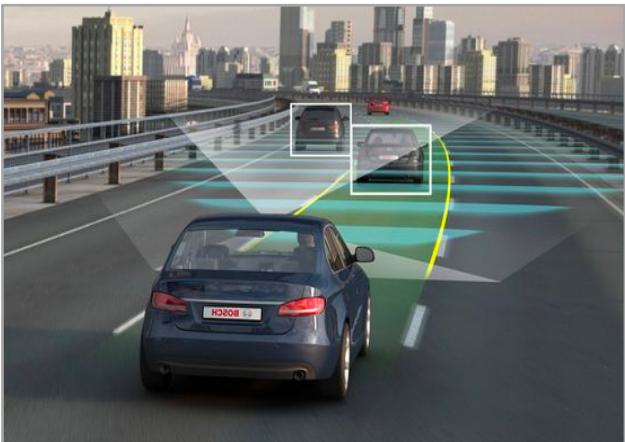
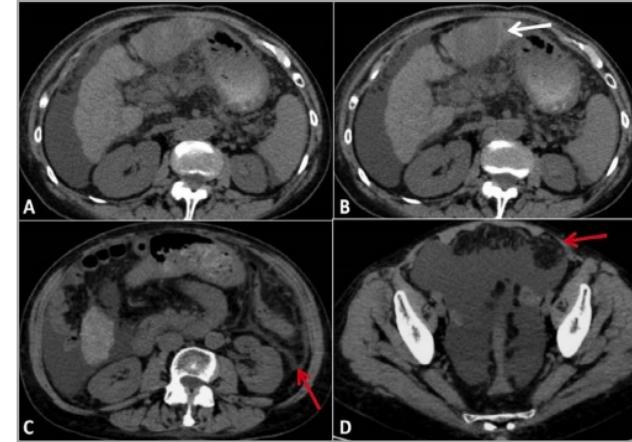
Vivo V7+ 第一款上市搭载完整人脸识别技术的国产手机

# ShuffleNet on 手机

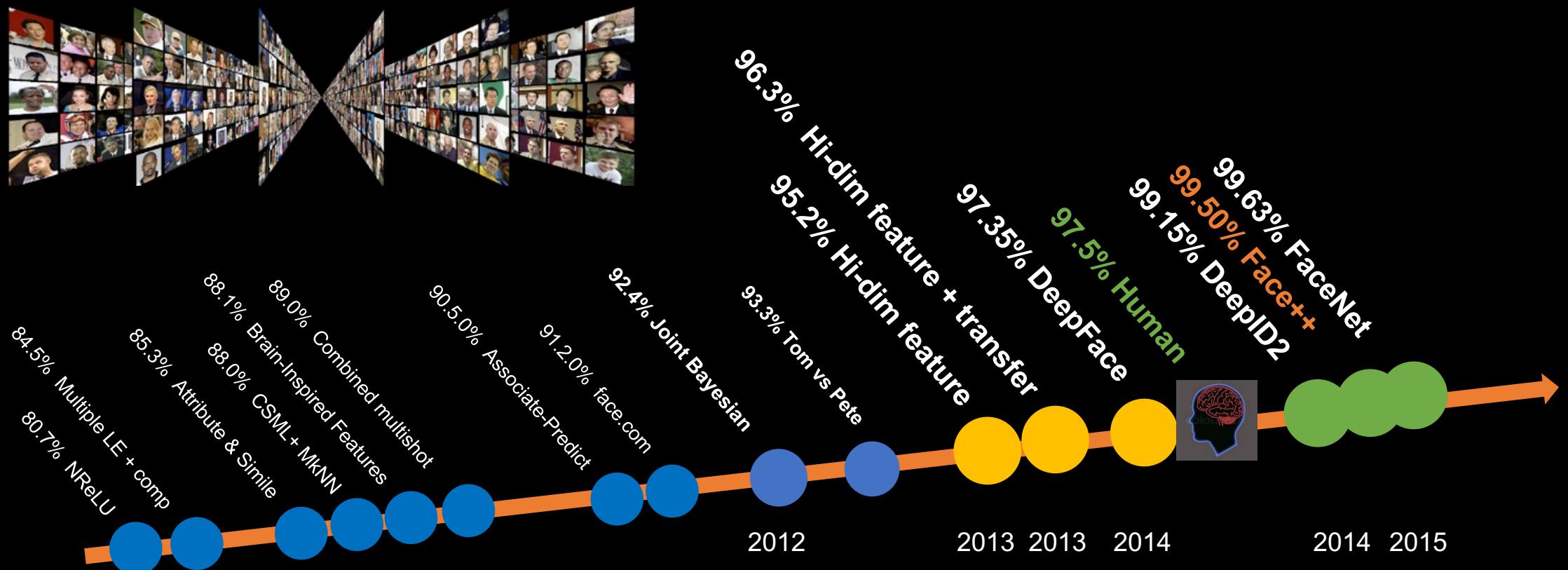




# Applications



# Face Recognition : Beyond Human



Face Verification in LFW Benchmark

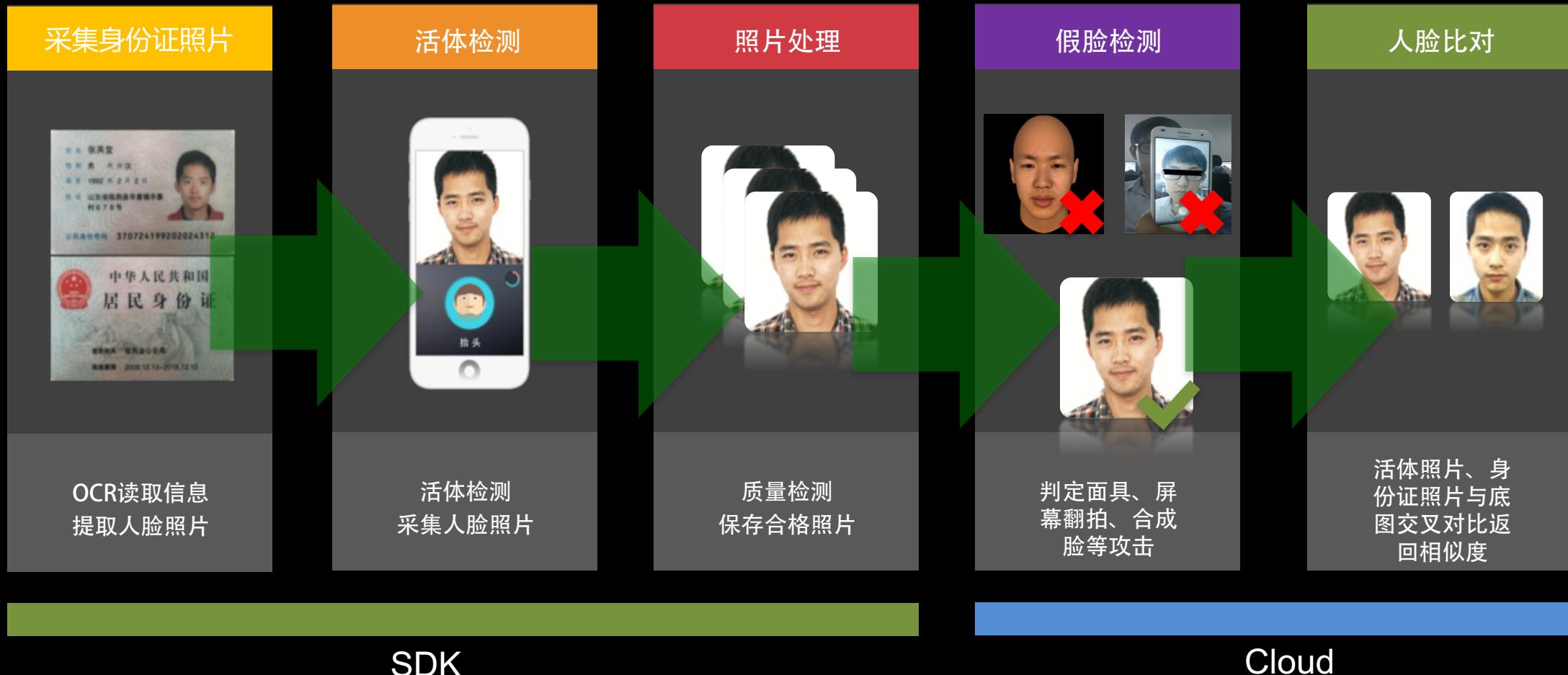
# FaceID 在线身份认证服务



典型客户



# FaceID 在线身份认证服务





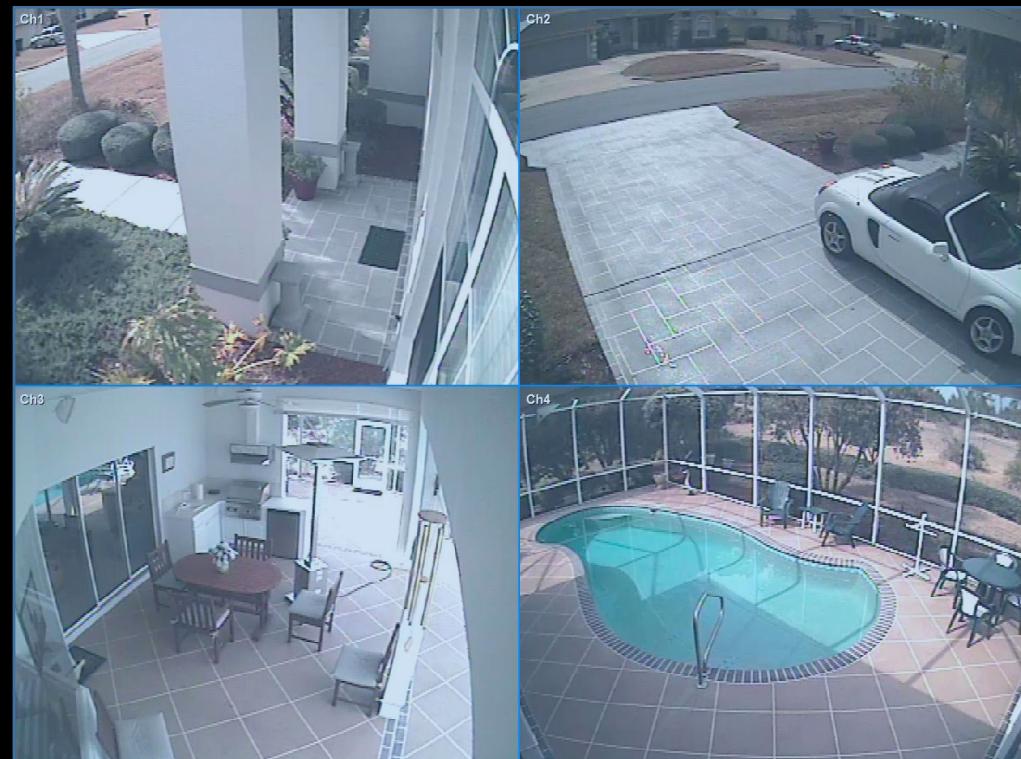
世界最大的身份验证平台  
2016年实现2.1亿人身份验证

# 智慧城市



Top1 accuracy >90% , in 100M database

# 智慧城市



旷视人像卡口大数据应用系统

卡口实况 智能查询 人员布控 数据研判 库管理 系统管理

GO 1 4 9 16 V P

海淀区 科学院南路 苏州街 海淀南路

石景山区 王泉路 阜石路 莲石路

01-26-2016 星期二 14:55

实时告警 查看历史

75

抓拍时间: 2015-03-16-09-09-09  
抓拍位置: A120-101-110  
姓名:  
身份证:

95

抓拍时间: 2015-03-16-09-09-12  
抓拍位置: A120-101-110  
姓名:  
身份证:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

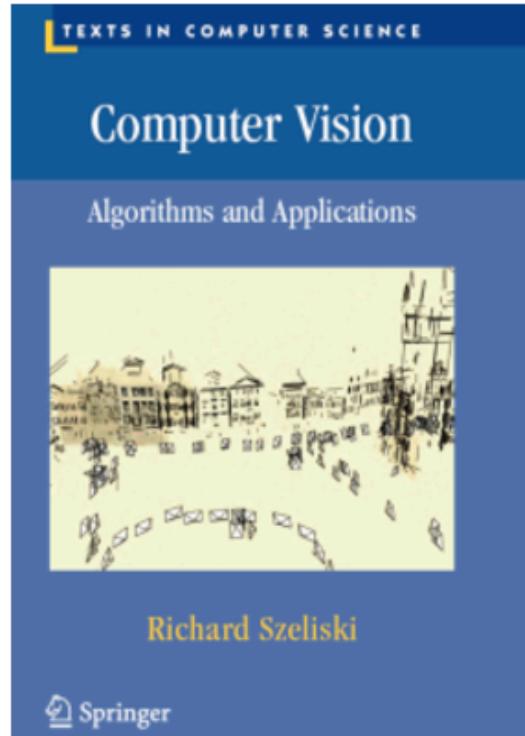
This screenshot shows the user interface of the 'Megvii Human Face Recognition Data Application System'. It features a main monitoring area with four live feeds (Ch1-Ch4) and a larger 4x4 grid of surveillance frames. On the left, there are dropdown menus for 'District' (Beihai District, Shijingshan District) and 'Road' (Kexueyuan South Road, Suzhou Street, Haidian South Road; Wangquan Road, Fushili Road, Lianshi Road). The top navigation bar includes tabs for 'Real-time Monitoring', 'Smart Search', 'Personnel Deployment', 'Data Analysis', 'Database Management', and 'System Management'. Below the feeds, a timeline shows '01-26-2016 星期二 14:55'. Two specific events are highlighted with callouts: one for a male subject (ID 75) and another for a female subject (ID 95), each with details about the capture time, location, name, and ID number. At the bottom, a row of 16 small thumbnail images represents a scrollable list of detected faces.

# Video Analytics (Vehicle/Person Detection & Tracking)

Week 1	Intro to Deep Learning and Computer Vision
Week 2	Deep Learning Toolkits, Math Revisit, Dataset
Week 3	Neural Network Basis & Architecture Design
Week 4	Deep Learning Platform: from Theano to Tensorflow/Megbrain
Week 5	Neural Network Approximation: Pruning, Factorization, and Low-bit Representation
Week 6	Modern Object Detection: SSD, Faster-RCNN, R-FCN
Week 7	Semantic Image Segmentation
Week 8	Invited Talk
Week 9	Text Detection and Recognition
Week 10	Recurrent Neural Network (RNN) and LSTM
Week 11	Generative Models and GANs
Week 12	3D Reconstruction: Conventional and Modern (Deep Learning based) Approaches
Week 13	Visual Object Tracking
Week 14	Human Understanding: ReID and Pose and Attributes
Week 15	Engineering in Deep Learning Research/System
Week 16	Student Spotlight Talk

# Computer Vision: Algorithms and Applications

© 2010 [Richard Szeliski](#)





**Stanford University**  
**CS 131 Computer Vision: Foundations and Applications**  
Fall 2016-2017

[Course home](#)

[Syllabus, lectures and assignments](#)

[Discussion](#)

[FAQ](#)

---

**Announcements:**

- Welcome to CS131!
  - Schedule information may change during the quarter; please visit the Syllabus page regularly to stay up to date.
  - **Lecture location has changed to 370-370 due to the high volume of student enrollment.**
- 

**Course Instructor:**

**Prof. Fei-Fei Li**

Office: Room 246, Gates Building

Office hours: Tuesday, 3-4pm

---

## CS231n: Convolutional Neural Networks for Visual Recognition

Spring 2017

### Course Description

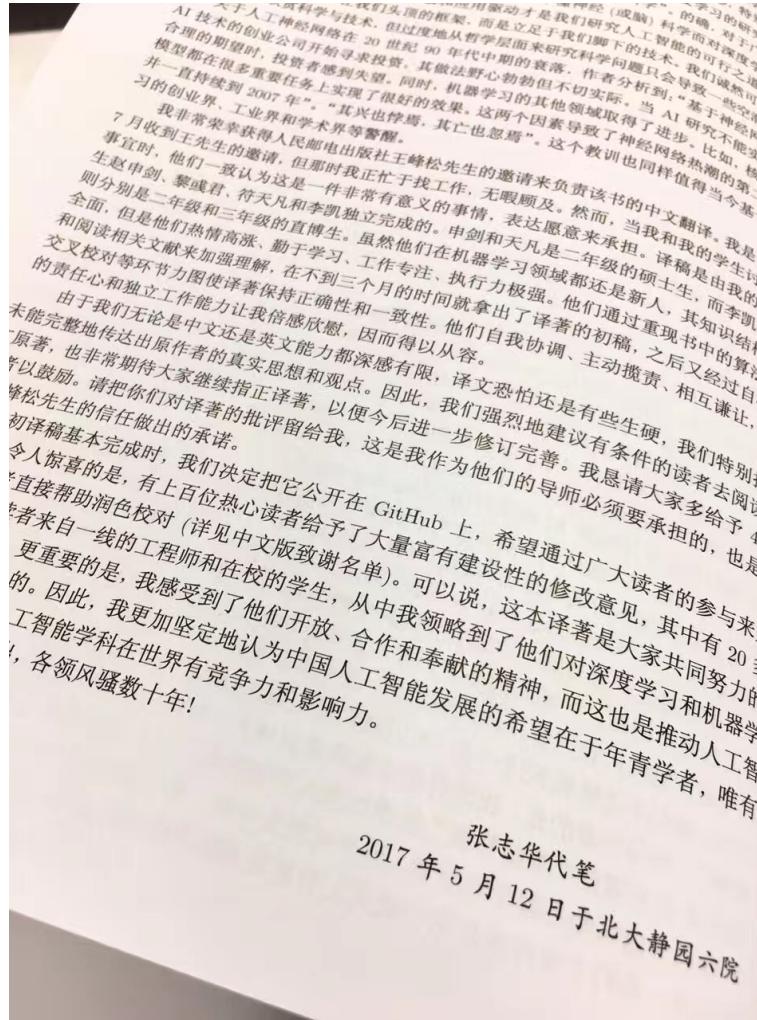
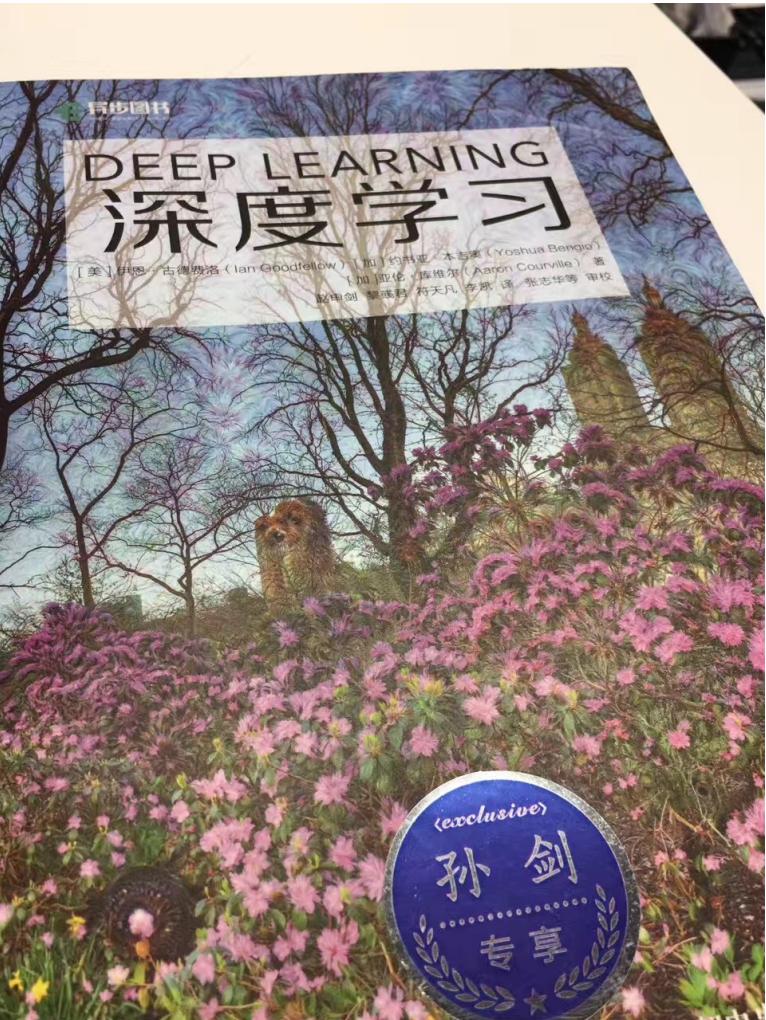
Computer Vision has become ubiquitous in our society, with applications in search, image understanding, apps, mapping, medicine, drones, and self-driving cars. Core to many of these applications are visual recognition tasks such as image classification, localization and detection. Recent developments in neural network (aka "deep learning") approaches have greatly advanced the performance of these state-of-the-art visual recognition systems. This course is a deep dive into details of the deep learning architectures with a focus on learning end-to-end models for these tasks, particularly image classification. During the 10-week course, students will learn to implement, train and debug their own neural networks and gain a detailed understanding of cutting-edge research in computer vision. The final assignment will involve training a multi-million parameter convolutional neural network and applying it on the largest image classification dataset (ImageNet). We will focus on teaching how to set up the problem of image recognition, the learning algorithms (e.g. backpropagation), practical engineering tricks for training and fine-tuning the networks and guide the students through hands-on assignments and a final course project. Much of the background and materials of this course will be drawn from the [ImageNet Challenge](#).

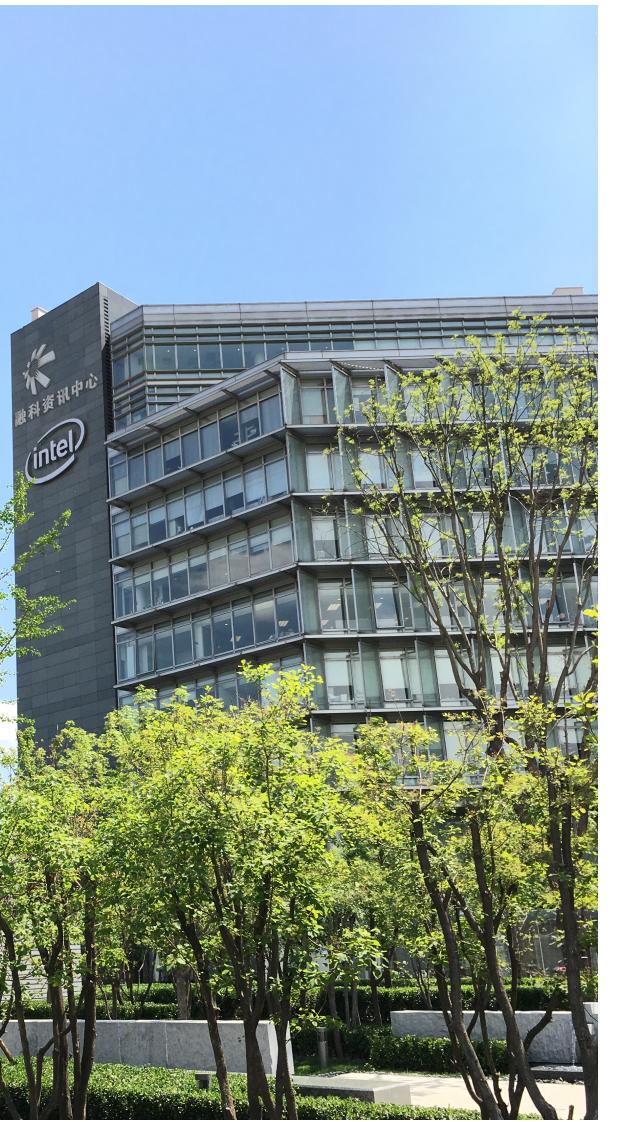
### Instructors



### Teaching Assistants







[career@megvii.com](mailto:career@megvii.com) (工作、实习)  
<https://megvii.com/campus/>

# Power Human with AI

追求 极致 简单 可靠

孙剑

[www.jiansun.org](http://www.jiansun.org)



构建人工智能云

赋能智能感知网络

