

# 大数据机器学习

第一讲 概述

袁春 清华大学深圳研究生院 2017/5/26



#### • 提纲:

- 大数据机器学习的背景知识
- 机器学习与相关学科的关系
- 机器学习发展历程
- 大数据机器学习的主要特征
- •课程参考书



Question: What's the classification the images belong to?

ladles





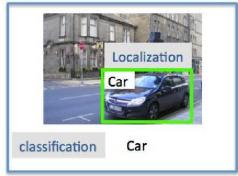
#### ■ ImageNet Challenge 2012

Task 1: Classification



- Predict a class label
- 5 predictions / image
- 1000 classes
- 1,200 images per class for training
- Bounding boxes for 50% of training.

Task 2: Detection (Classification + Localization)



- Predict a class label and a bounding box
- 5 predictions / image
- 1000 classes
- 1,200 images per class for training
- Bounding boxes for 40% of training.

Task 3: Fine-grained classification



- Predict a class label given a bounding box in test
- 1 prediction / image
- 120 dog classes (subset)
- ~200 images per class for training (subset)
- Bounding boxes for 100% of training

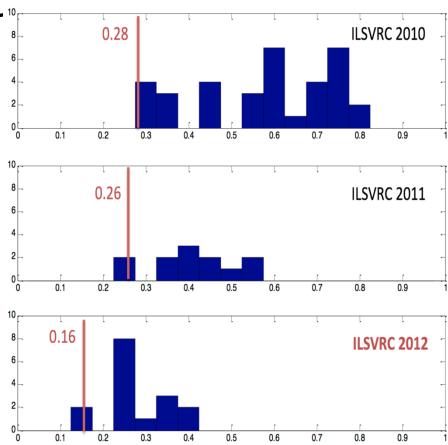


#### ■ Convolutional Neural Network

#### Promising techniques

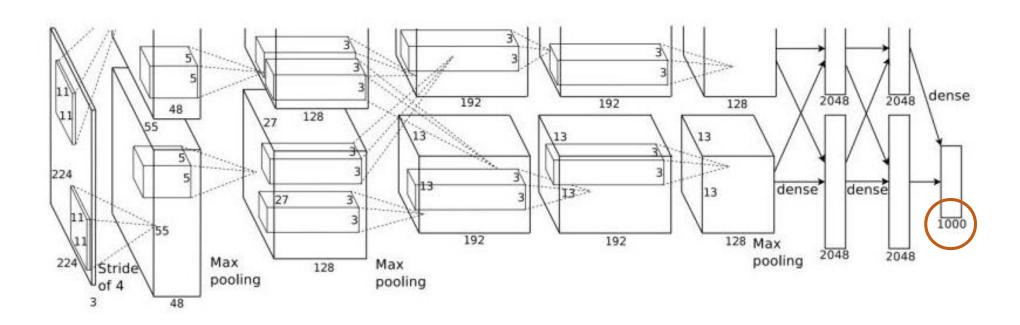
Method	Year
<b>Locality-constrained Linear Coding</b>	ILSVRC 2010
Improved Fisher Vectors	ILSVRC 2011
<b>Convolutional Neural Networks</b>	ILSVRC 2012

Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, University of Toronto 2012





#### Convolutional Neural Network



Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, University of Toronto 2012



#### ImageNet Progress:

• 2015年: 微软的神经网络系统错误率为4.94%, 低于人类测试者的5.1%。

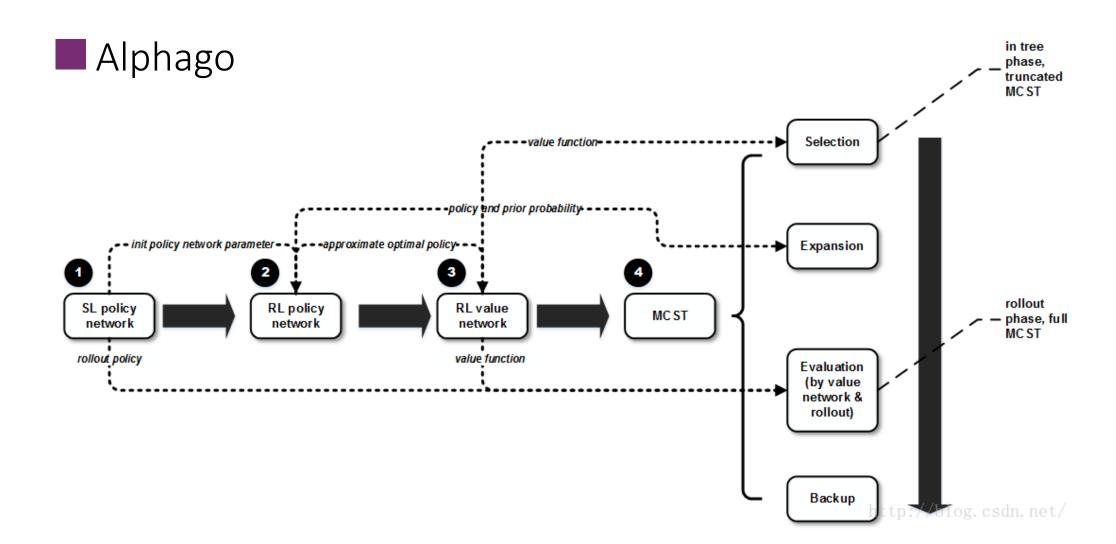
	layers	error(top 5)	
AlexNet	8	15%	2012
VGGNet	19	7.32%	2014
GooleNet	22	6.66%	2014
MSRA	152	3.57%	2016



## Alphago









■ 机器学习进入了新的时代!



#### ■机器学习

- 维基百科:
- 机器学习是近20多年兴起的一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习理论主要是设计和分析一些让计算机可以自动"学习"的算法。机器学习算法是一类从数据中自动分析获得规律,并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论,机器学习与统计推断学联系尤为密切,也被称为统计学习理论。



#### Game



比赛结果编映了。装金归属编映了、大侠们的身份 啥时候揭晓呢?





■ Text to speech and speech recognition









# ■3D 体感游戏







#### Bioinformatics

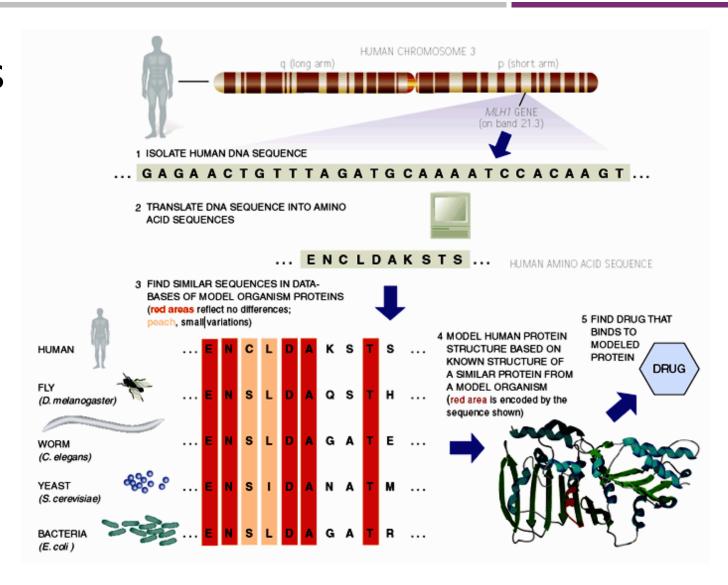
Gene

teettegageegattataeaegaeageaaaatattteagtegegettgggeaaaaggettaageaegaeteeeagteeeeettaeatttgtetteetaageeeetggageeaetateaaaettgttetaegettgeaetgaaaata aaagtaaacaatcaaaaagaccaaaaacaataacaaccagcaccgagtcgaacatcagtgaggcattgcaaaaatttcaaagtcaagtttgcgtcgtcatcgcgtctgagtccgatcaagccgggcttgtaattgaagttgttgttgatgstggattgtggggaattetggteageataettaaeageageeegetaattaageaaaataaaeatateaaatteeagaatgegaeggegeeateateetgtttgggaatteaattegegggeagategtttaatteaattaaaaggt agggagcagaagaatqcgatcgctggaatttcctaacatcacggaccccataaatttgataagcccgagctcgctgcgttgagtcagccaccccacatcccccaaatccccgccaaaagaagacagctgggttgttgactcgccagat catte taatgaagaettggagaagaettaegttatatteagaeeategtgegatagaggatgagteattteeatatggeegaaatttattatgtttaetategtttttagaggtgtttttttggaettaeeaaaagaggeatttgttttacttg aggaattg gegateg tag egattteece ege egta aa ag tteet gateet eg ttg tttttg tacatea ta aa g tee ggattet geteg tege ega ag at gggaa ega ag et gee aa ag et gegat get get gggateg tege gang at gegat ge ${\it scaget}$  take the contract of the contrac stecttegetggggatteeeetgegggtaageegettgettggaetegttteeaaateeeatageeaegeeageagaggagtaaeaggeteyhereisthegenetgattaaaaatateetttaagaaageeeatgggtataae pogtoctatgogaggaatggtotttaggttotttatggcaaagttotogoctogottgcccagcoggtacgttottggtgatottttaggaa/aatcotggactactgtogcotggottatggccacaagacccaccaagag actgttatgattctcatgctgatgcgactgaagcttcacctgactcctgctccacaattggtggcctttatatagcgagatccacccgcatgttgcgtggaatagaaatgcgggtgactccaggaattagcattatcgatcggaaag toaaatagotgogatoaotogggagoagggtaaagtogootogaaacaggaagotgaagotottotataaatacaotoaaagogatofttoogaggogagtotggttagaaatttacatggactgoaaaaagggtatagoocoacaa ategetgegttteggeagetaattgeettttagaaaattatttteeeatttegagaaaetegtgtgggatgeeggatgeggettteaa $\chi$ eaettetggeeegggateggg $\chi$ etgggteaeattgtetgegggetetattgtetegatee ycagttcgcgtgcttagcggtcagaaaggcaggattcggttcggattgatgcgctggcagcagggcacaaagatctaatgactggyaaatcgctacaaataaattaaaytccggcggctaattaatgagcggactgaagccacttt aaccaaaaaacagcagataaacaaaaacggcaaagaaaattgccacagagttgtcacgctttgttgcacaaacatttgtgcagaaaagtgaaaagcttttagccattat<mark>t</mark>aagtttttcctcagctcgctggcagcacttgcgaatg cogoaggogtotatttatgocaaggacogttottottottoagottoggotogagtatttgttgtgocatgttggttacgatggoaatogoggtacagttatgcaaatgaapaaaggaatacogotoactgacaatgaacggogtottgt tcatgotgacattcatattcattcotttggttttttgtcttcgacggactgaaaagtgcggagagaaacccaaaaacaga (gcgcgcaaagcgcgttaatatgcgaactcattgaatttatcacaacaccatatcca atocatatoaatatoaatatogotattattaaogatoatgototgotgatoaagtattoagogotgogotagattogao (gattgaatogagotoaatagaotoaacag cotocaotogaoagatgogoaatgocaaggacaattgo  $attact t tegeograp teg teatggtgtegttgetgettt tatgttgegtacteeggactacaeggagagtteaggg{} attegtgeteegtgateegtgateegtgtteegtgtteegtgtteegtgtteegtgtteegtgtteegtgtteegtgtteegtgtgtgtaceettegt$ tttttttttagggeeeaataaaagegettttgtggeggettgatagattateaettggttteggtggetageeaagtggetttettetgteegaegeaettaattgaatta oggeagegggtgggggategggageeeeeegaaaaaaaaeagaggegeaegtegatgeeateggggaattggaaeeteaatgtgtgggaatgtttaaatattetgtgtgttaggtagtgtagttteatagaetatagatteteatacaga to etto gago cogatta ta cacega cacaa a a tattic agtog gott gago cacaa aggott aggo accepta cacattit gto the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacago coct gago cacata to a acctif give the cacata gago cacataaaagtaaacaatcaaaaagaccaaaaaacaatcaaccagcaccgagtcgaacatcagtgaggcattgcaaaaatttcaaagttcaagtttgcgtcgtcatcgcgtctggagtccgatcaagccgggcttgtaattgaagttgttgatg stggattgtggggaattetggteageataettaaeageageeegetaattaageaaaataaaeatateaaatteeagatgegaeggegeeateateetgtttgggaatteaattegegggeagategtttaatteaattaaaaggt agggaggagaagaatqggatcqctggaatttcctaacatcacggaccccataaatttgataaqcccgagctcqctgcgttgagtcagccaccccacatccccaaatccccqccaaaagaagacagctgggttgttgactcqccagatycagtggagtggacctggtcaaagaagcaccgttaatgtgctgattccattcgattccatccgggaatgcgataaagaaggctctgatccaagcaactgcaatccggattttctcttttccattttgttttttttttatttacgt  ${\it catte}$  tata  ${\it catte}$  and  ${\it catte}$  aactgaaaagatatttaaatttttettettggaecatttteaaggtteeggatatattttgaaacacactagetageagtgttggtaagttaeatgtatttetataatgteatatteetttgteegtatteaaategaataeteeaeate tacttgaggaattggcgatcgtagcgatttcccccgccgtaaagttcctgatcctcgttgtttttgtacatcataaagtccggattctgctcgtcgccgaagatgggaacgaagctgccaaagctgccaaagctgctgcttgaggtgctg  ${\tt scagetggataacettgctgtacagateggcatetgcctggagggcacgategaaatccttccagtggacgaacttcacctgctcgctgggaatagcgttgttgtcaagcagctcaaggagcgtattcgaggttgacggactcacca$ 



#### Bioinformatics

Gene



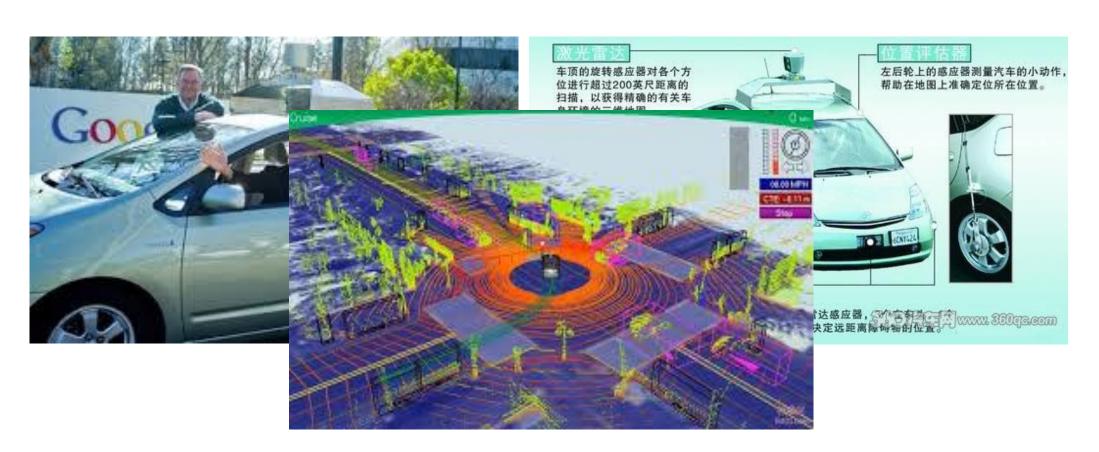


### Quantitative trading





#### Robotic Control





# ■再现古代陶瓷工艺







### ■机器学习相关概念

人工智能

机器学习

深度学习

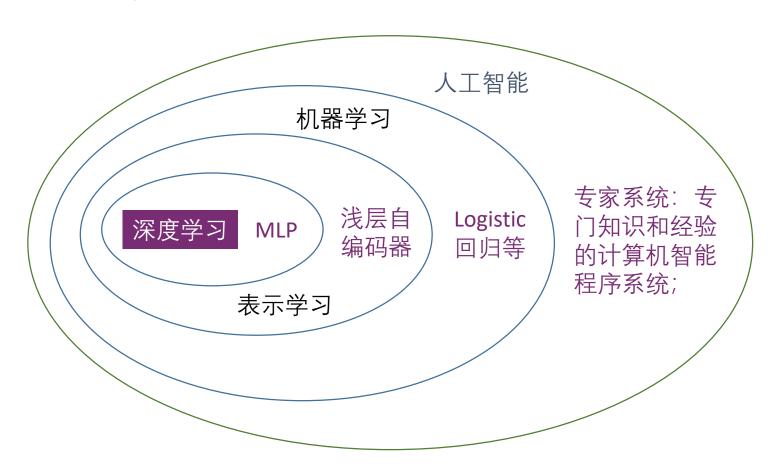
数据挖掘

计算机视觉

统计学习



### ■人工智能/机器学习/深度学习



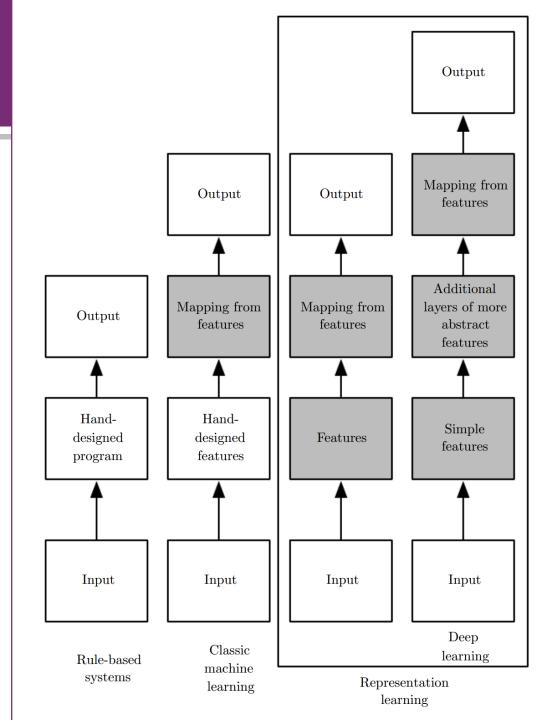
人工智能:是科学,为机器赋予视觉/听觉/触觉/推理等智能。

机器学习:人工智能的计算方法。



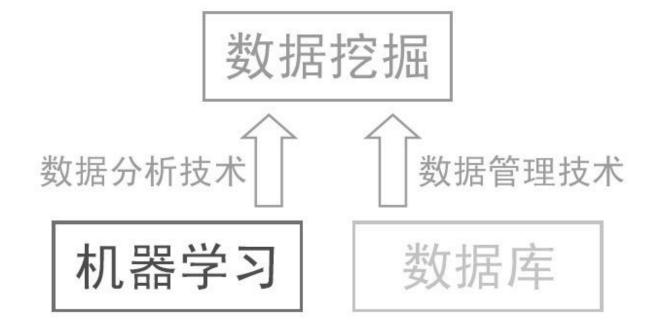
■深度学习和人工智能其它方法

阴影: 可学习的部分





■机器学习和数据挖掘





■机器学习的一个形象描述





■机器学习和计算机视觉

播放视频

播放视频

计算机视觉是机器学习最重要的应用

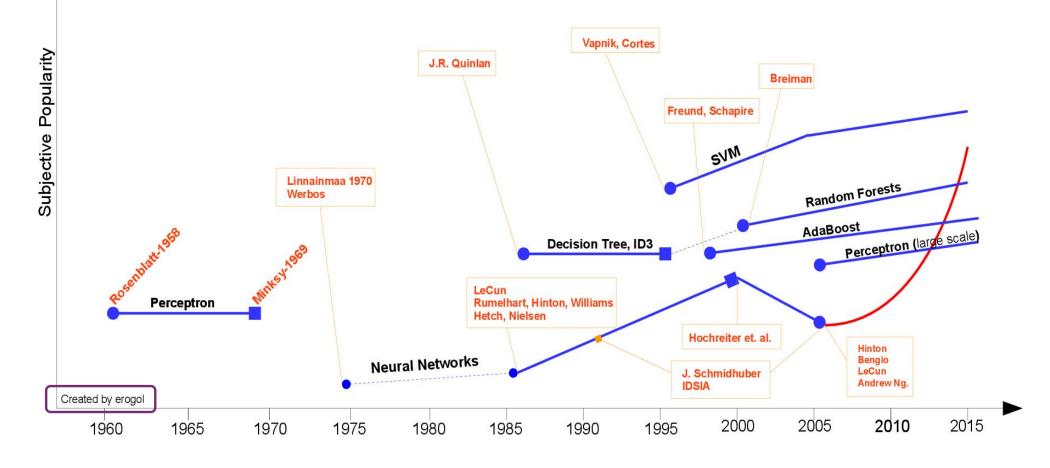


### ■机器学习和统计学习

- ---Simon Blomberg:
  - From R's fortunes package: To paraphrase provocatively, 'machine learning is statistics minus any checking of models and assumptions'.
- ---Andrew Gelman:
  - In that case, maybe we should get rid of checking of models and assumptions more often. Then maybe we' d be able to solve some of the problems that the machine learning people can solve but we can' t!

机器学习=统计—模型和假设的检验







Al Magazine Volume 27 Number 4 (2006) (F AAAS)

#### A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

John McCarthy, Marvin L. Minsky, Nathaniel Bochester, and Claude E. Stamoon

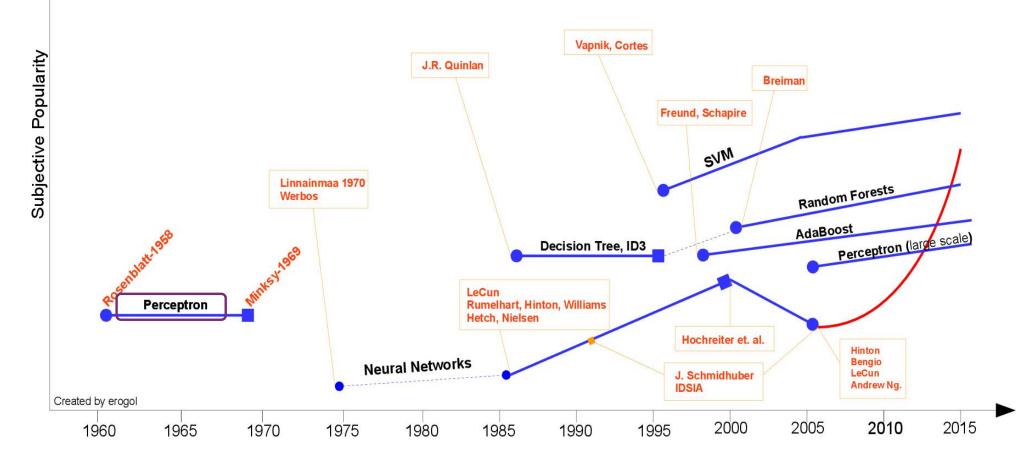




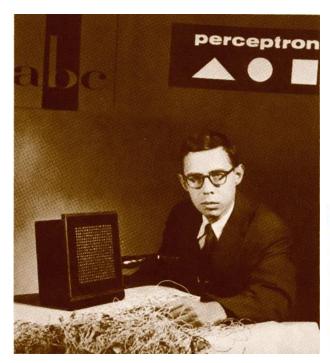
在人工智能50年大会上,5位1956年Dartmouth人 工智能夏季研究会的与会者再相最

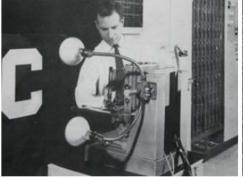
照片以充支右: Trenchard More, John McCarthy, Marvin Minsky, Oliver Selfridge, 以 3ay Solomonoff (Photo by Joseph Mehling 39)



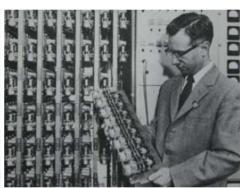






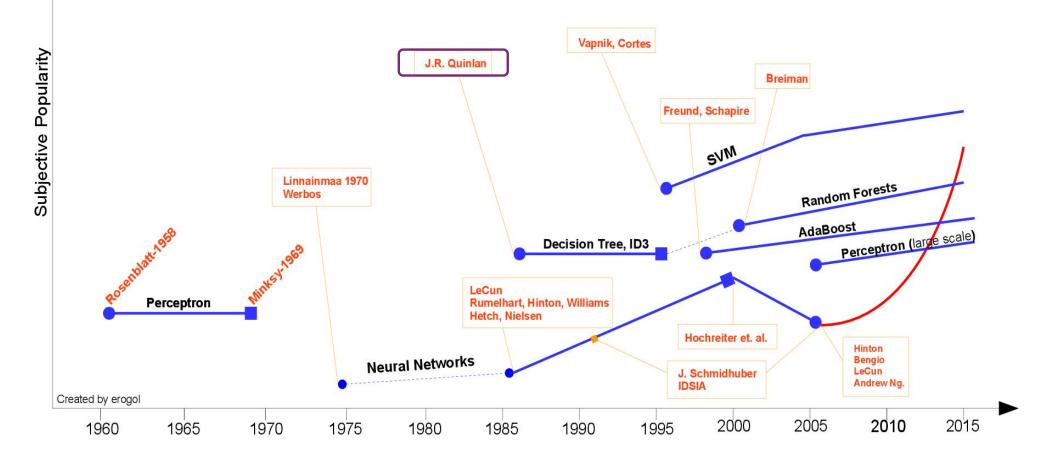






**Figure 4.8** Illustration of the Mark 1 perceptron hardware. The photograph on the left shows how the inputs were obtained using a simple camera system in which an input scene, in this case a printed character, was illuminated by powerful lights, and an image focussed onto a  $20 \times 20$  array of cadmium sulphide photocells, giving a primitive 400 pixel image. The perceptron also had a patch board, shown in the middle photograph, which allowed different configurations of input features to be tried. Often these were wired up at random to demonstrate the ability of the perceptron to learn without the need for precise wiring, in contrast to a modern digital computer. The photograph on the right shows one of the racks of adaptive weights. Each weight was implemented using a rotary variable resistor, also called a potentiometer, driven by an electric motor thereby allowing the value of the weight to be adjusted automatically by the learning algorithm.

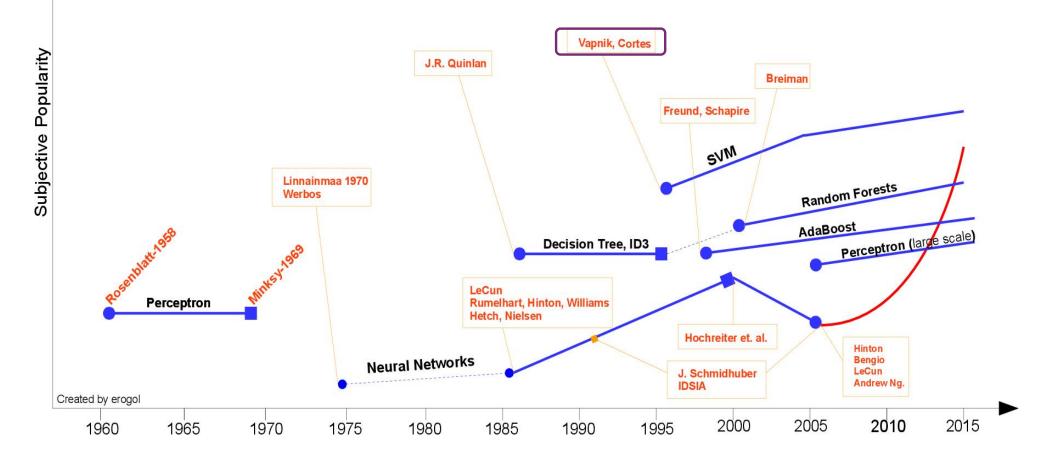








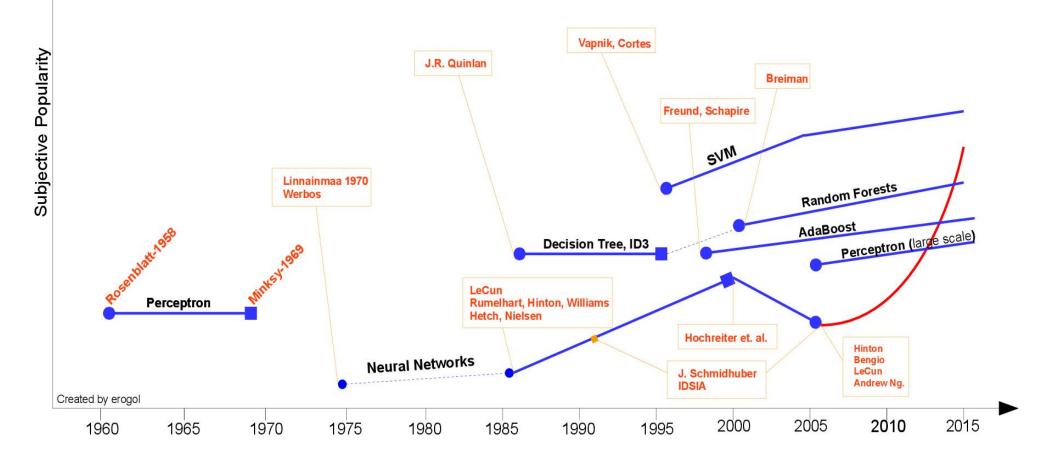




















■大数据机器学习的主要特征



#### ■与日俱增的数据量: Facebook

- facebook月活跃用户接近8.5亿
- 每天上传的照片总量为2.5亿张
- 4.25亿移动用户
- 1000亿个connections
- Zygna游戏为facebook总收入贡献了12%
- 57%的用户为女性用户



#### ■与日俱增的数据量: Tencent

• QQ: 月活跃用户超 8 亿, 最高同时在线1.9亿;在线人际关系链超1000亿;

• 微信: 月活跃用户超3.5亿;日均消息量超50亿;

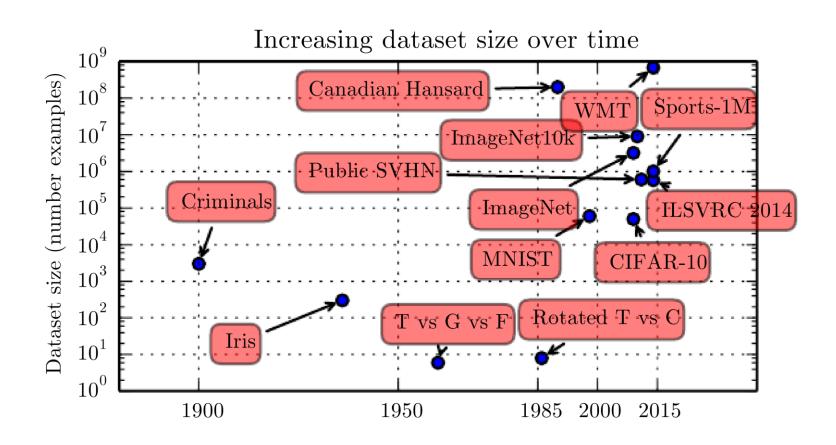
• 空间: 月活跃用户超6亿;日均相册上传超过4亿;

• 游戏: 腾讯游戏月活跃用户4.5亿; 手机游戏月活跃用户近2亿;

网站:日均浏览量PC侧超17亿;手机侧近13亿;最高日接入消息条数8000亿;日接入数据量
200TB; 并发分拣业务接口10000个。

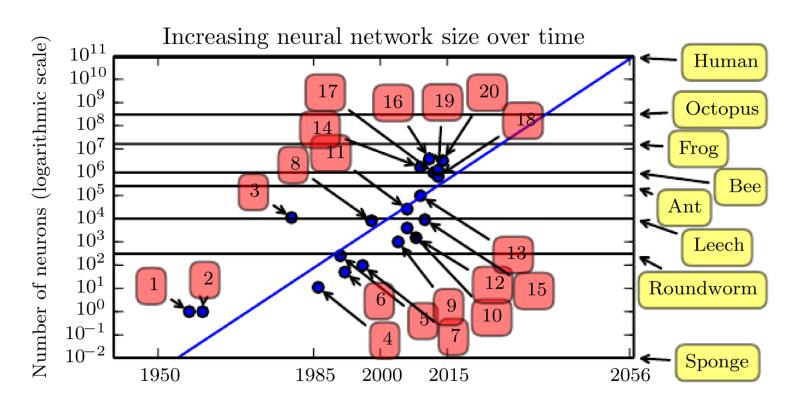


## ■实验数据量的增加



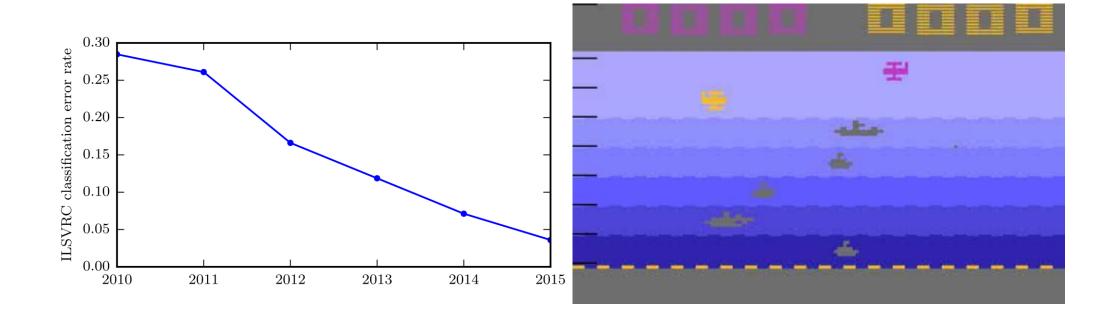


# ■与日俱增的神经网络模型规模



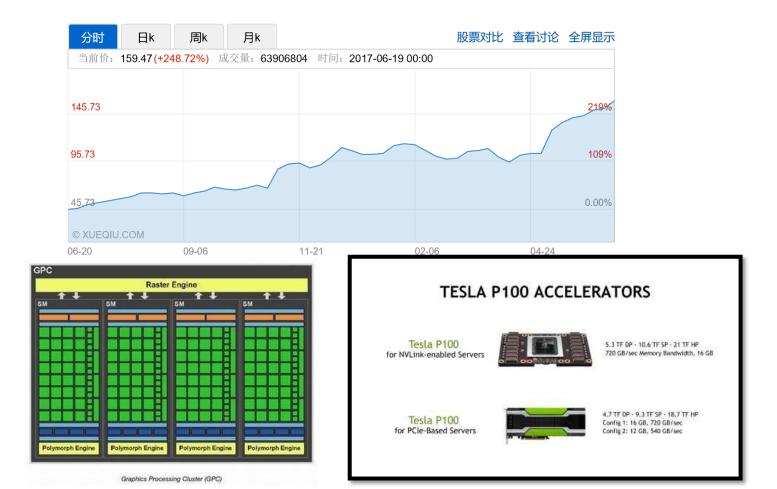


# ■与日俱增的精度、复杂度和对现实世界的冲击



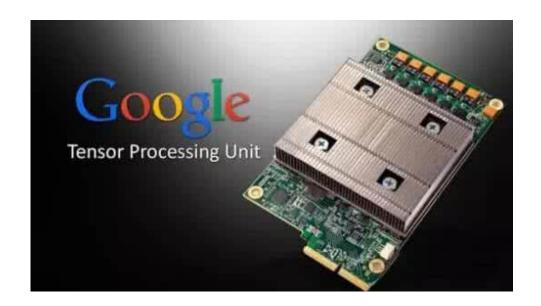


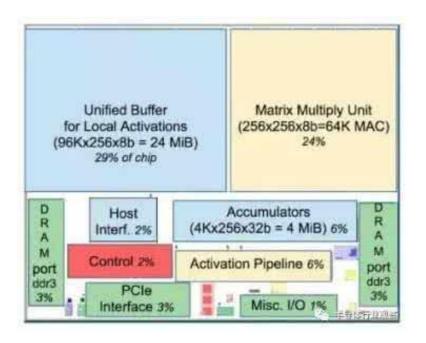
### ■ GPU (Graphic Processing Unit)





#### ■ TPU Tensor Processing Unit







## ■深度学习框架

**TensorFlow** 

Pytorch

Caffe

**CNTK** 

Keras

**MXNet** 

Theano

Scikit-learning

Spark MLlib



## ■课程特点

- 以机器学习方法为主干
- 以深度学习模型为重点
- 实现大数据机器学习的应用为目标

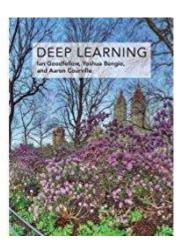


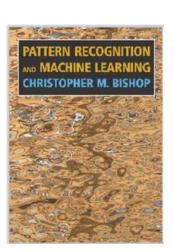
# ■课程内容

• 参考教材











# Q&A?