机器学习简介

什么是机器学习

- 赫伯特·西蒙（1959）：如果一个系统，能够通过执行某个过程，就此改进了它的性能，那么这个过程就是学习。
- Langley（1996）：机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能”
- Tom Mitchell(1997):机器学习是对能通过经验自动改进的计算机算法的研究。
- Tom Mitchell《机器学习》： A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E（强调学习效果）

什么是机器学习

- Tom Mitchell《机器学习》：如果一个程序在使用既有的经验E(Experience)来执行某类任务T(Task)的过程中被认为是具备学习能力的，那么它一定要展现出：利用现有的经验E，不断改善其完成既定任务T的性能(Performance)的特质。
- Alpaydin（2004）：机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。”
- 特雷弗·哈斯蒂《统计学习基础》：机器学习就是抽取重要的模式和趋势，理解数据的内涵表达，即从数据中学习（突出学习任务的分类）
- 弗拉基米尔·万普尼克《统计学习理论的本质》:机器学习就是一个基于经验数据的函数估计问题（侧重可操作性）

机器学习

机器学习是近40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。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

机器学习是人工智能(Artificial Intelligence, AI）的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，它主要使用归纳、综合而不是演绎。

机器学习：探究和开发一系列算法来如何使计算机不需要通过外部明显的指示，而可以自己通过数据来学习，建模，并且利用建好的模型和新的输入来进行预测的学科。

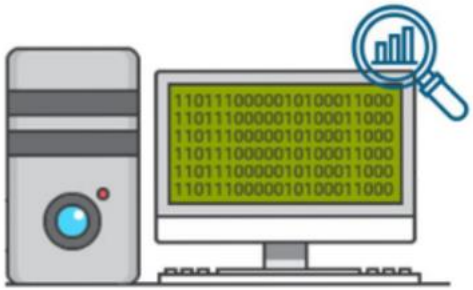
什么是机器学习

- 学习：

观察（E） → 学习 → 技能

- 机器学习：

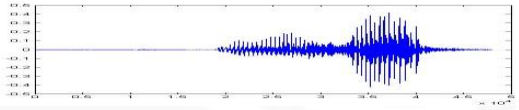
数据（Data） → 机器学习 → 技能



狗

机器学习 ≈ 构建一个映射函数

- 语音识别

  $f($  $)=$ “你好”

- 图像识别

  $f($  $)=$ “猫”

- 围棋

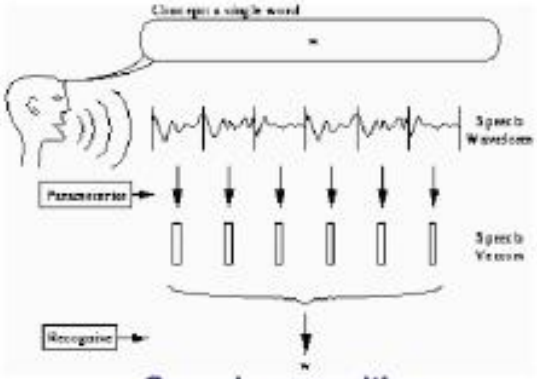
  $f($  $)=$ “5-5” (落子位置)

- 对话系统

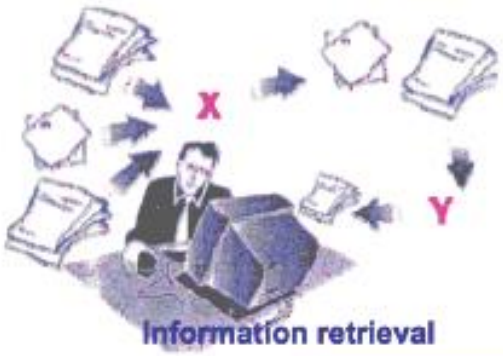
  $f($ “你好” $)=$ “今天天气真不错”

  用户输入　　机器

机器学习的应用



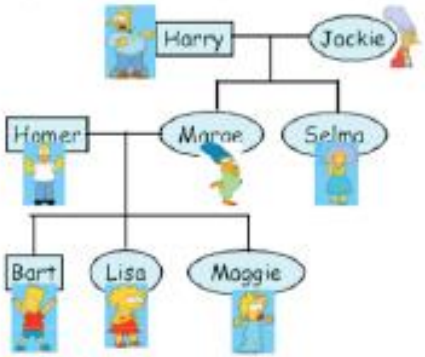
Speech recognition



Information retrieval

Computer vision

Games



Pedigree

Evolution

Robotic control

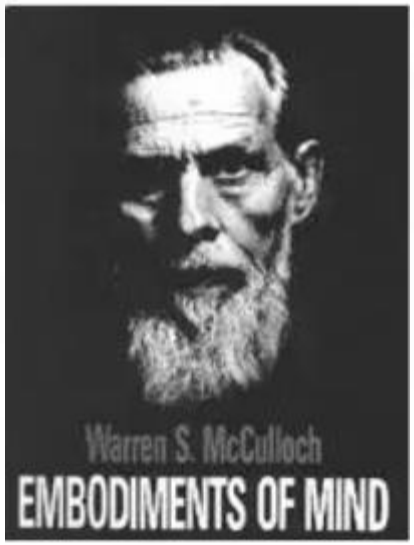
Planning

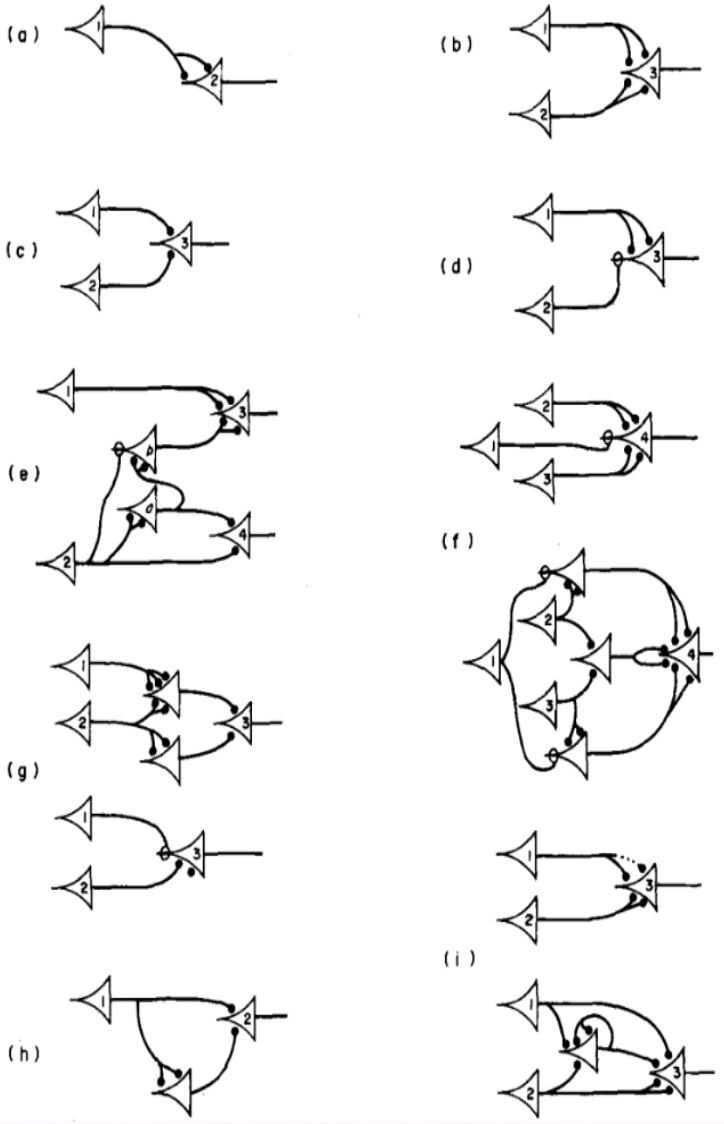
机器学习：实现人工智能的一种方法

- 人工智能: 机器展现的人类智能
- 机器学习：实现人工智能的一种方法
- 深度学习：**实现机器学习的一种技术**

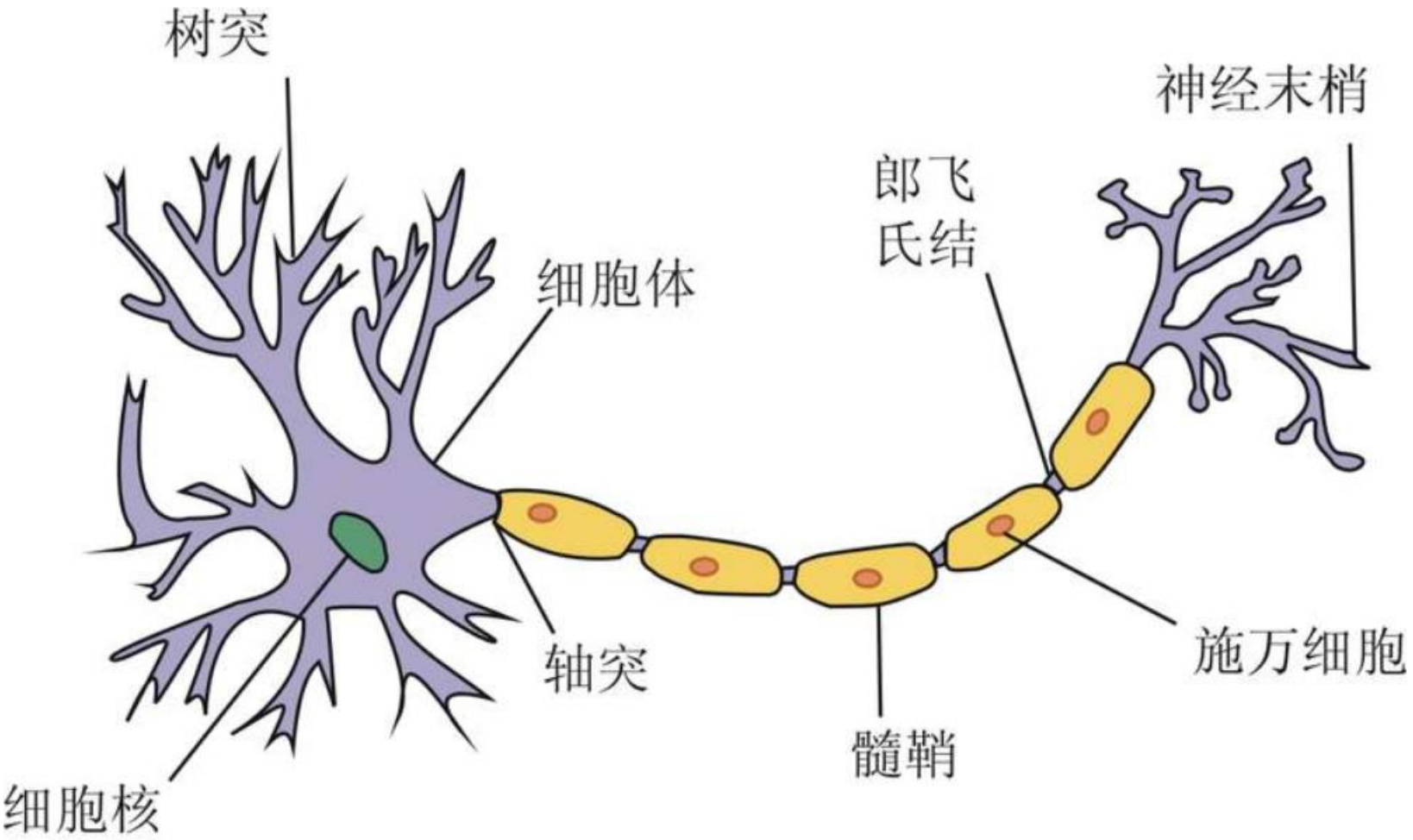
机器学习历史

- 1943年，心理学家Warren McCulloch和数理逻辑学家Walter Pitts提出了神经网络层次结构模型，确立了神经网络的计算模型理论，从而为机器学习的发展奠定了基础

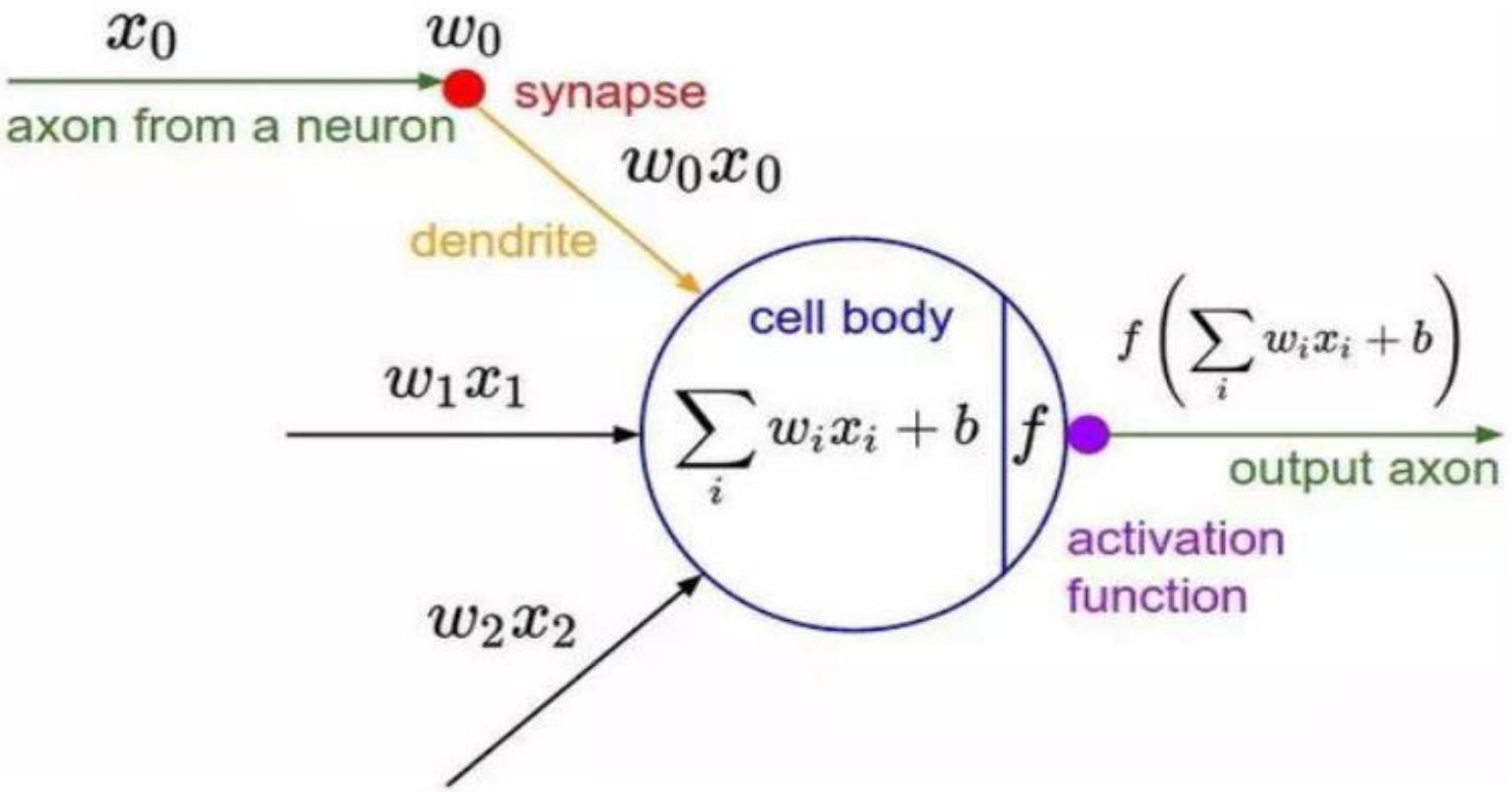




机器学习历史



机器学习历史

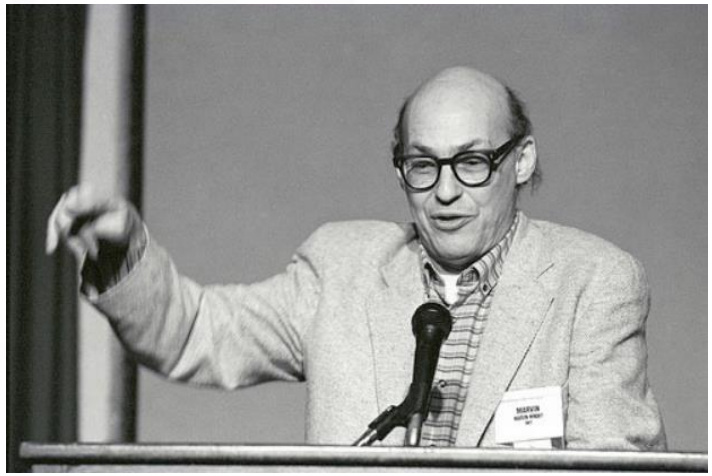

x_0
w_0
axon from a neuron
synapse
$w_0 x_0$
dendrite
cell body
$w_1 x_1$
$\sum_i w_i x_i + b$
f
$f\left(\sum_i w_i x_i + b\right)$
output axon
activation function
$w_2 x_2$

机器学习历史

生物神经元	MP神经元模型
神经元	j
输入信号	x_i
权值	w_{ij}
输出信号	y_j
总和	$\sum$
膜电位	$\sum_{i=1}^{n} w_{ij}x_i$
阈值	θ_j

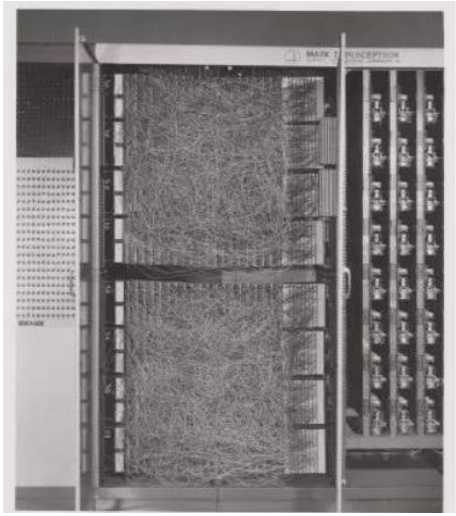
机器学习历史

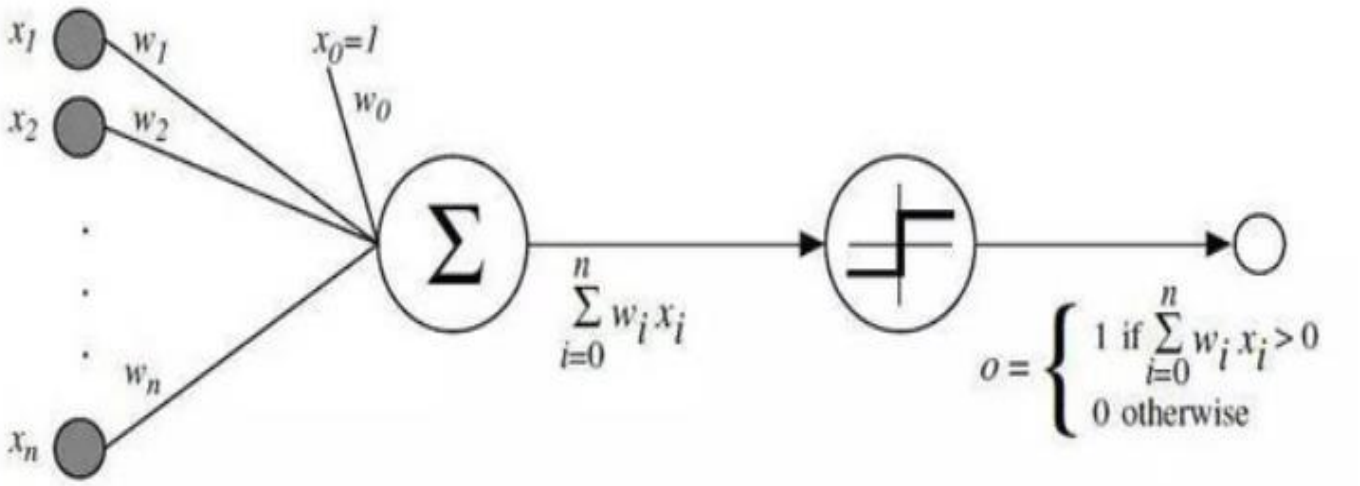
- 1949年，心理学家Donald O. Hebb在《The Organization of Behavior》中描述了神经元学习法则。
- Hebb Law(学习就是改变连接): Cells that fire together, wire together(在同一时间被激发的神经元间的联系会被强化)。
- 1951年， Marvin Minsky制造出第一台神经网络机SNARC，在只有40个神经元的小网络里，第一次模拟了神经信号的传播。



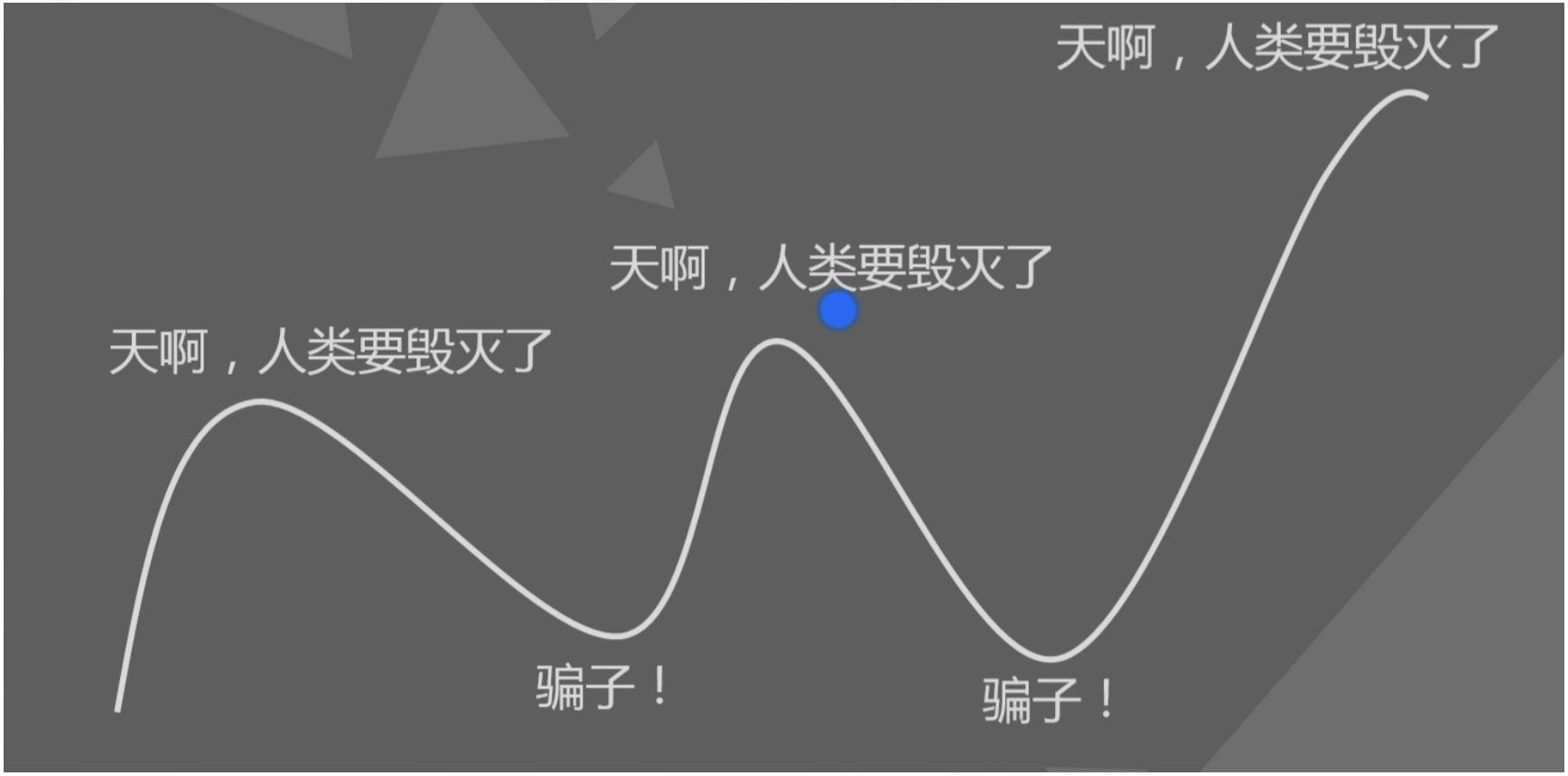
机器学习历史

- 1957年，Rosenblatt提出了Perceptron（感知器）概念，用Rosenblatt算法对Perceptron进行训练。并且首次用算法精确定义了自组织自学习的神经网络数学模型，设计出了第一个计算机神经网络（NN算法），开启了NN研究活动的第一次兴起





机器学习历史



机器学习历史

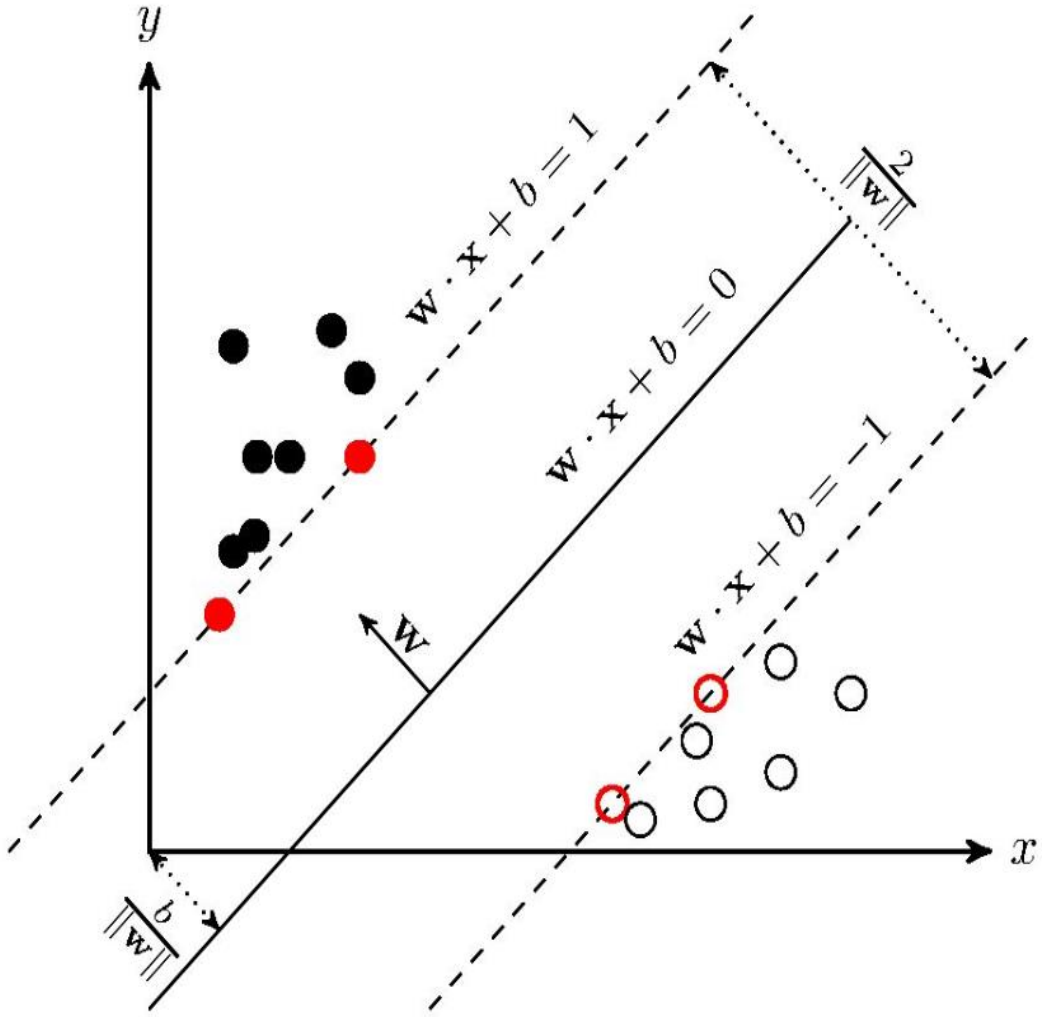
- 1958年，Cox给Logistic Regression方法正式命名，用于解决美国人口普查任务
- 1959年，Samuel设计了一个具有学习能力的跳棋程序，曾经战胜了美国保持8年不败的冠军。这个程序向人们初步展示了机器学习的能力，Samuel将机器学习定义为无需明确编程即可为计算机提供能力的研究领域
- 1960年，Widrow用delta学习法则来对Perceptron进行训练，可以比Rosenblatt算法更有效地训练出良好的线性分类器

机器学习历史

- 1962年，Hubel和Wiesel发现了猫脑皮层中独特的神经网络结构可以有效降低学习的复杂性，从而提出著名的Hubel-Wiese生物视觉模型，该模型卷积神经网络（CNN）的雏形，这之后提出的神经网络模型也均受此启迪
- 1963年伦纳德·武赫和查尔斯·瓦斯勒发表了关于模式识别的论文，描述了第一个机器学习程序

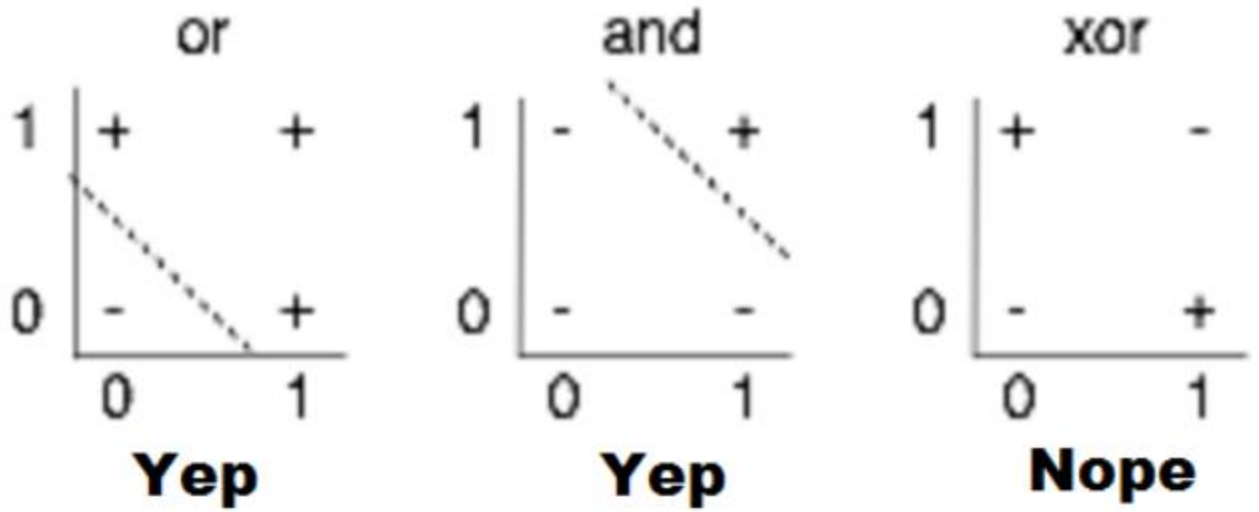
机器学习历史

- 1963年，Vapnik和Chervonenkis发明原始支持向量方法，即起决定性作用的样本为支持向量（SVM算法）



机器学习历史

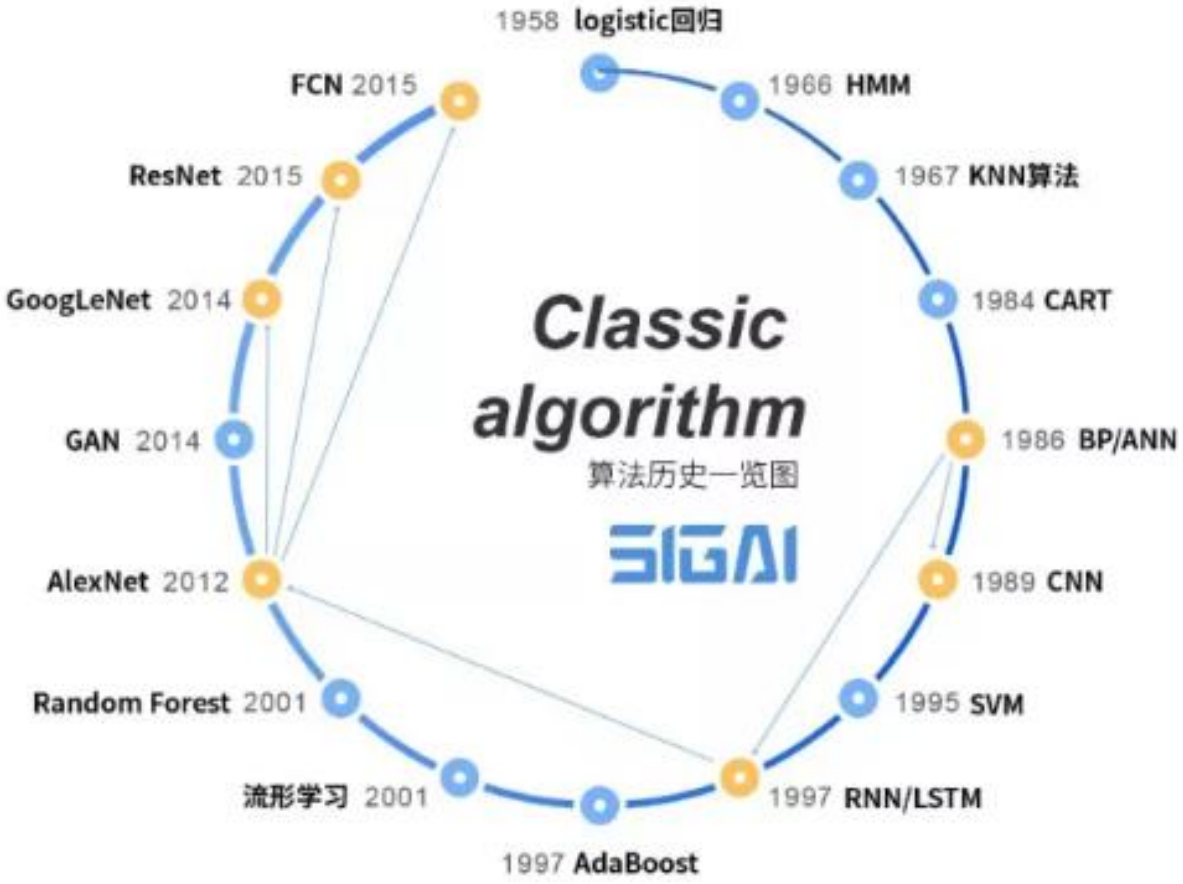
- 1969年，Minsky和Paper出版了对机器学习研究有深远影响的著作《Perceptron》，其中对于机器学习基本思想的论断：解决问题的算法能力和计算复杂性。文中提出了著名的线性感知机无法解决异或问题。



机器学习历史

- 1971年，Vapnik和Chervonenkis提出VC维概念，描述了假设空间和模型复杂度，衡量了经验误差和泛化误差的逼近程度，它给诸多机器学习方法的可学习性提供了坚实的理论基础
- 1980年，在美国卡内基梅隆大学举行了第一届机器学习国际研讨会，标志着机器学习研究在世界范围内兴起，该研讨会也是著名会议ICML的前身

机器学习历史


1958 logistic回归
1966 HMM
1967 KNN算法
1984 CART
1986 BP/ANN
1989 CNN
1995 SVM
1997 RNN/LSTM
1997 AdaBoost
流形学习 2001
Random Forest 2001
AlexNet 2012
GAN 2014
GoogLeNet 2014
ResNet 2015
FCN 2015
Classic algorithm
算法历史一览图
SIGAI

机器学习历史

- 1981年，Werbos提出多层感知机，解决了线性模型无法解决的异或问题，第二次兴起了NN研究
- 1984年，Leslie Valiant提出概率近似正确学习（Probably approximately correct learning，PAC learning），是机器学习的数学分析的框架，它将计算复杂度理论引入机器学习，描述了机器学习的有限假设空间的可学习性，无限空间的VC维相关的可学习性等问题。

机器学习历史

- 1984年，Breiman发表分类回归树（CART算法，一种决策树）
- 1986年，Quinlan提出ID3算法（一种决策树）
- 1986年，Rumelhart，Hinton和Williams联合在Nature杂志发表了著名的反向传播算法（BP算法）
- 1989年，Yann和LeCun提出了目前最为流行的卷积神经网络（CNN）计算模型，推导出基于BP算法的高效训练方法，并成功地应用于英文手写体识别

机器学习历史

- 1995年，Vapnik和Cortes发表软间隔支持向量机（SVM算法），开启了随后的机器学习领域NN和SVM两大社区的竞争
- 自1995年到随后的10年，NN研究发展缓慢，SVM在大多数任务的表现上一直压制着NN，并且Hochreiter的工作证明了NN的一个严重缺陷-梯度爆炸和梯度消失问题
- 1997年Adaboost，Freund和Schapire提出了另一种可靠的机器学习方法-Adaboost

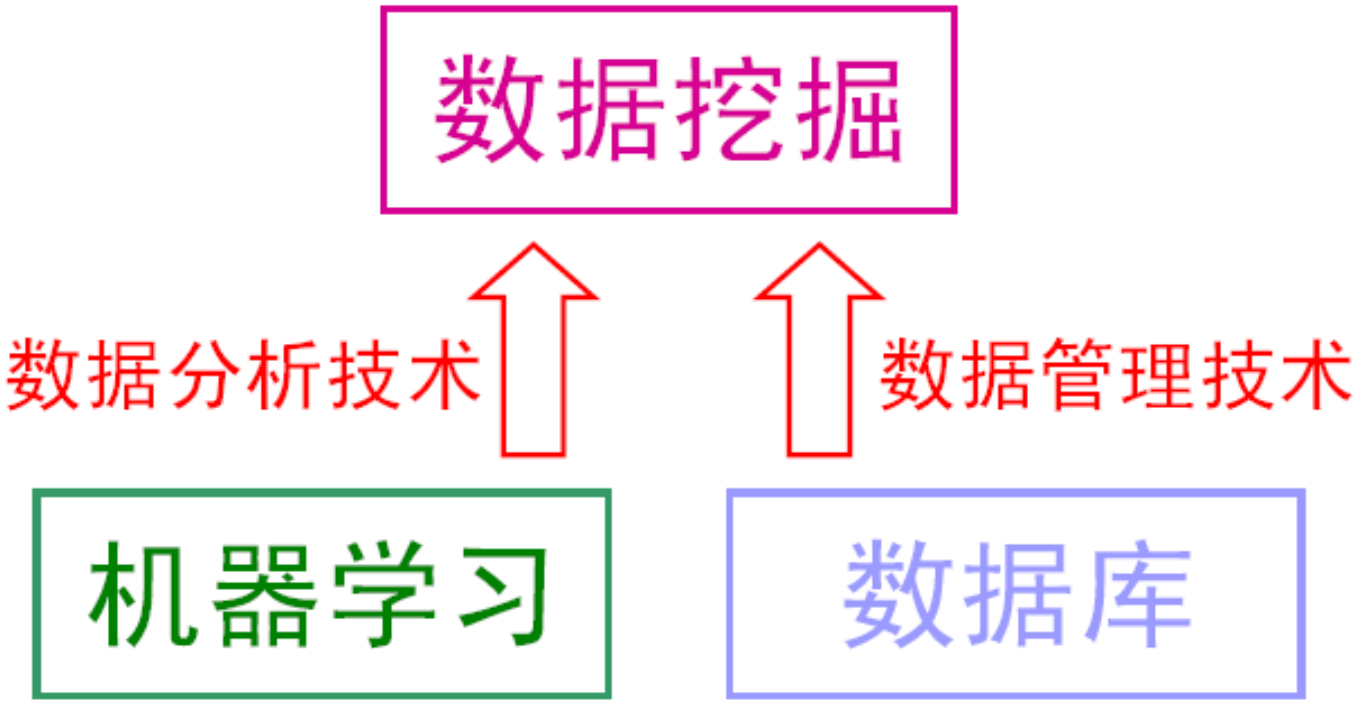
机器学习历史

- 2001年，Breiman发表随机森林方法（Random forest），Adaboost在对过拟合问题和奇异数据容忍上存在缺陷，而随机森林在这两个问题上更加鲁棒。
- 2005年，经过多年的发展，NN众多研究发现被现代NN大牛Hinton, LeCun, Bengio, Andrew Ng和其它老一辈研究者整合，NN随后开始被称为深度学习（Deep Learning），迎来了第三次崛起。

机器学习和数据挖掘的关系

- 机器学习是数据挖掘的重要工具。
- 数据挖掘不仅仅要研究、拓展、应用一些机器学习方法，还要通过许多非机器学习技术解决数据仓储、大规模数据、数据噪音等等更为实际的问题。
- 机器学习的涉及面更宽，常用在数据挖掘上的方法通常只是“从数据学习”，然则机器学习不仅仅可以用在数据挖掘上，一些机器学习的子领域甚至与数据挖掘关系不大，例如增强学习与自动控制等等。
- 数据挖掘试图从海量数据中找出有用的知识。
- 大体上看，数据挖掘可以视为机器学习和数据库的交叉，它主要利用**机器学习界提供的技术来分析海量数据，利用数据库界提供的技术来管理海量数据**。

机器学习和数据挖掘的关系



机器学习和统计学习

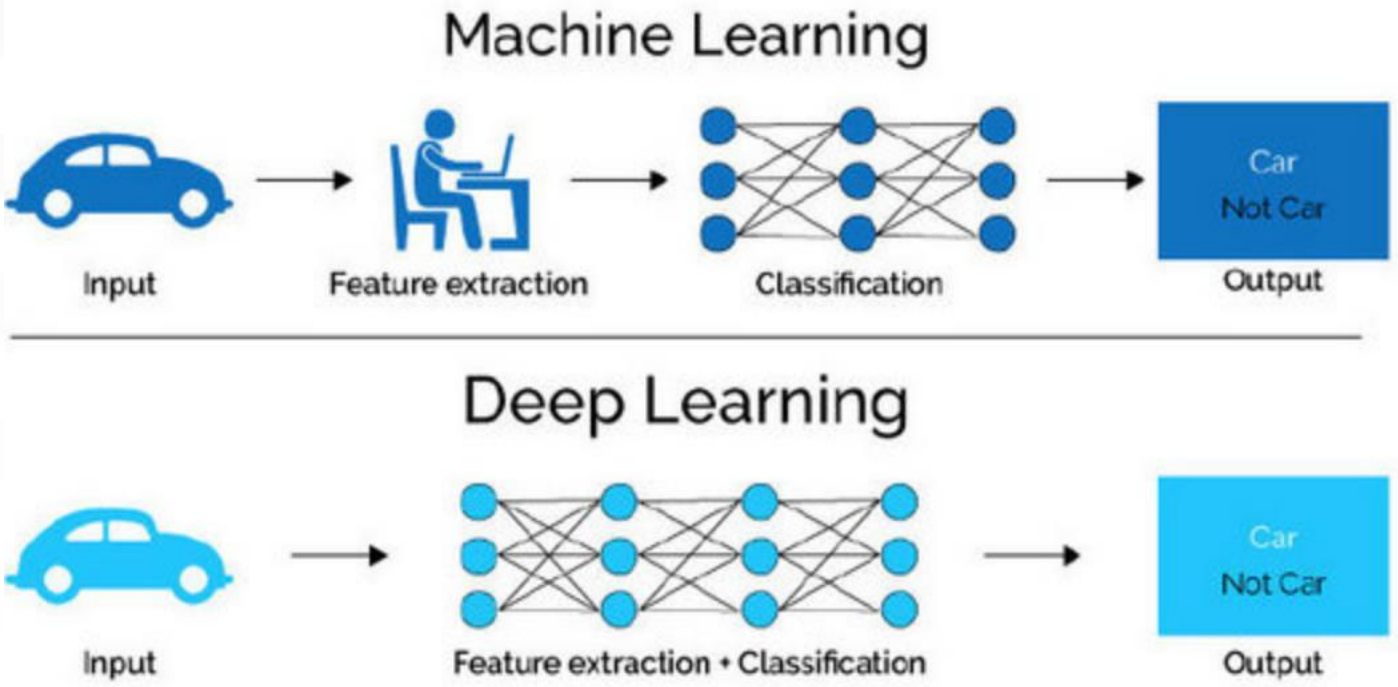
- 统计学习是theory-driven，对数据分布进行假设，以强大的数学理论支撑解释因果，注重参数推断（Inference）
- 机器学习是data-driven，依赖于大数据规模预测未来，弱化了收敛性问题，注重模型预测（Prediction）；

理解和预测

- 解释因果：统计学习（theory drive）
  - ✓ 回归和假设检验
- 预测未来：机器学习（data drive）
  - ✓ 优化问题

机器学习和深度学习

- 深度学习是机器学习的一个子领域，特征提取更依赖于隐层模型，解释性弱，趋于黑盒子，对数据依赖性更强，更擅长处理高维度大数据。



机器学习与传统编程

机器学习通过程序让计算机来模拟人的学习过程

例：通过身高x，预测体重y

传统编程：（1）确定输入x，输出y

（2）[根据已有数据集，]通过人的经验或者查询资料，确定x和y的关系：y=0.9x-90

机器学习(2a) 设计模型为y=ax+b，编写学习算法，对已有数据集进行训练，得到预测模型y=0.8x-100

机器学习的适用条件

- 适用条件
  - 事物本身存在某种潜在规律
  - 某些问题难以使用普通编程解决（图像识别、语音识别）
  - 有大量的数据样本可供使用
- 大数据
  - Web：Google index 包括大约450亿页面
  - Click-stream data: 10-100TB/天
  - Transaction data：5-50TB/天
  - TV：2TB/天/频道；YouTube 4TB/天 的上传量
  - Photos：15亿张/周的上传量
  - 数字电话：100 PB/天

机器学习的适用条件

练习题：

Which of the following is best suited for machine learning?

1. predicting whether the next cry of the baby girl happens at an even-numbered minute or not
2. determining whether a given graph contains a cycle
3. deciding whether to approve credit card to some customer
4. guessing whether the earth will be destroyed by the misuse of nuclear power in the next ten years

THE END