深度学习

《深度学习》

绪论

概览

1 课程介绍

1.1~1.3 章节结构、课程定位

1.4~1.7 相关材料、课程要求

2 深度学习简介

2.1 深度学习的定位

2.2 何谓智能

2.3 人工智能的研究领域

2.4 发展历史

2.5 人工智能的流派

3 从“浅”到“深”

3.1 机器学习

3.2 表示学习

3.2.1 语义鸿沟

3.2.2 分布式表示

3.3 深度学习

4 神经网络

4.1 仿生学特点

4.2学习过程

4.3 常用框架

4.4 神经网络的性质

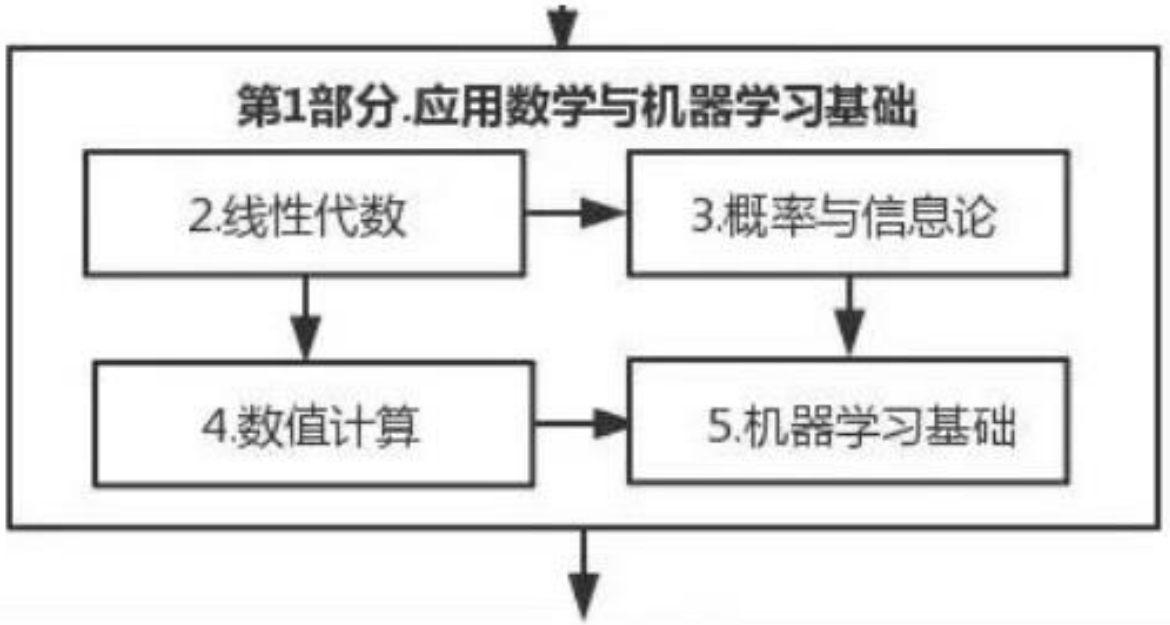
1.1 前导课程

- 数学类

  线性代数、微积分、概率论、信息论、数学优化

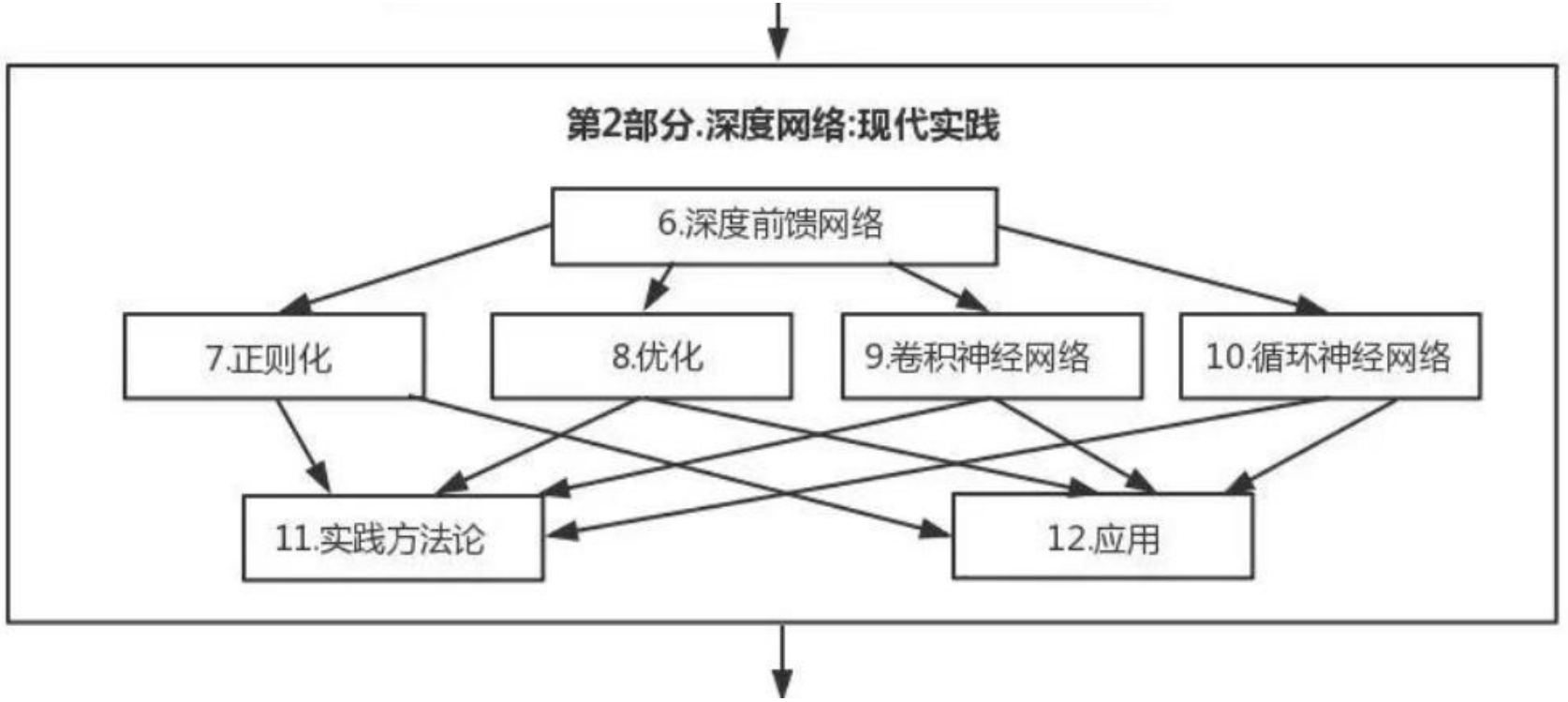
- 编程类

- 专业课程类



1.2 课程结构及要求

章节关系：如右图



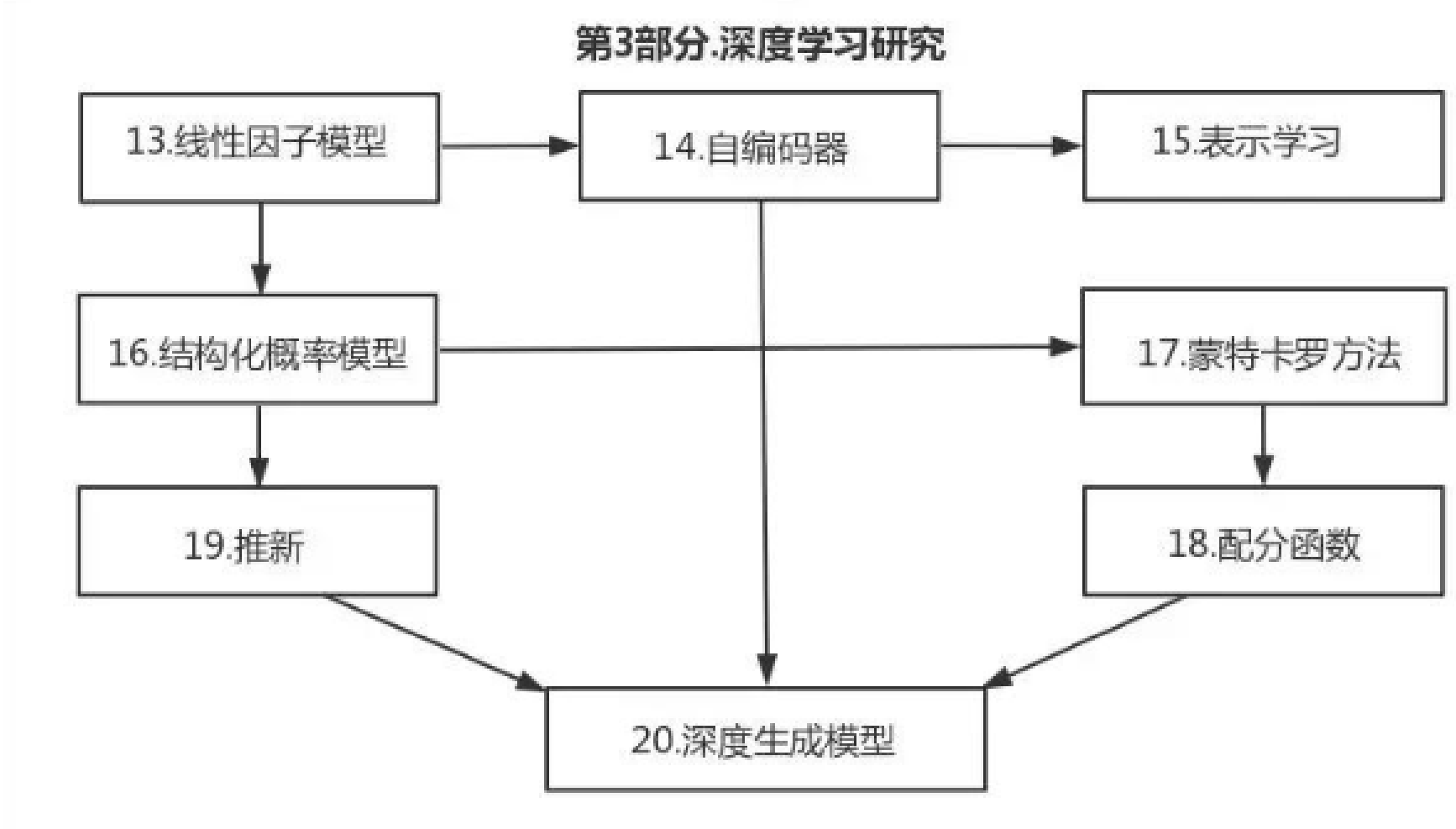
课程定位

课程要求：

深度学习基础理论（几个层次）

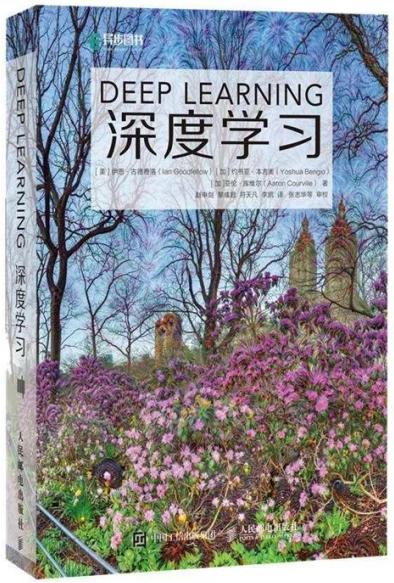
实践能力

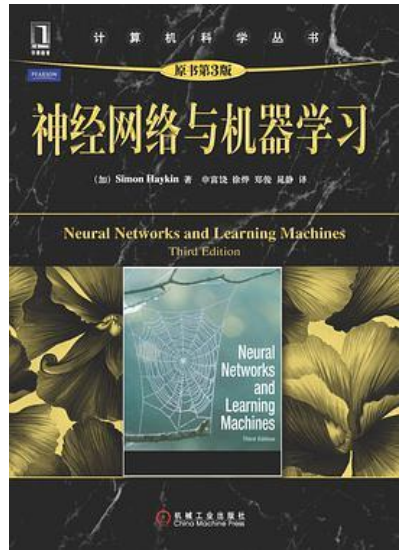
1.3 进阶学习


第3部分.深度学习研究
13.线性因子模型
14.自编码器
15.表示学习
16.结构化概率模型
17.蒙特卡罗方法
19.推新
18.配分函数
20.深度生成模型

1.4 推荐教材

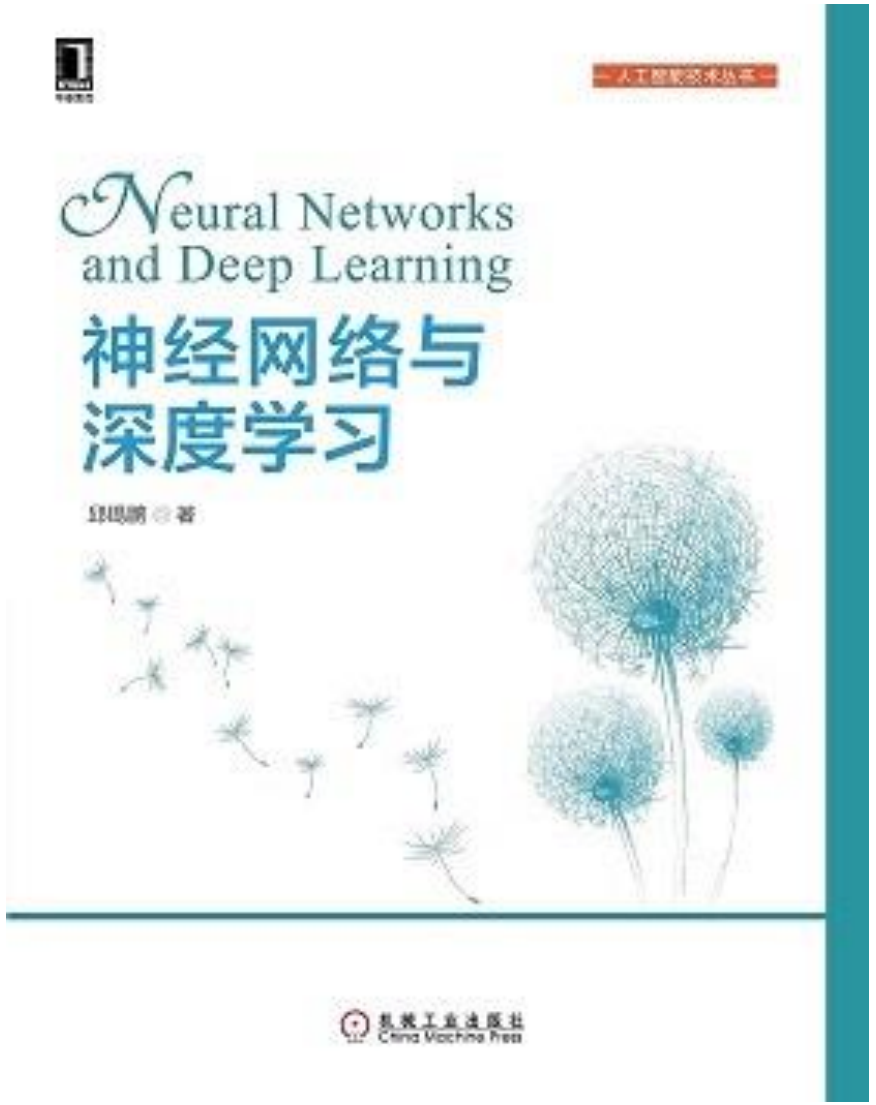
- 《深度学习》
- 作者: [美] 伊恩·古德费洛 / [加] 约书亚·本吉奥 / [加] 亚伦·库维尔
  出版社: 人民邮电出版社
  出版年: 2017-7-1
- 中文译版电子书:
  https://github.com/exacity/deeplearningbook-chinese
- 《神经网络与机器学习》
- 作者: [加] Simon Haykin
  出版社: 机械工业出版社
  出版年: 2011-3





邱锡鹏
教授，博士生导师，复旦大学计算机科学技术学院

蒲公英书:希望这本教材能够帮助更多的学生进入深度学习以及人工智能领域，他们会为人工智能领域注入新的生机与活力。



1.6 顶会论文

- 深度学习相关的学术会议主要有
  - ICLR、NeurIPS、ICML、AAAI、IJCAI
- 自然语言处理领域：
  - ACL、EMNLP
- 视觉领域：
  - CVPR、ICCV
- …

人工智能

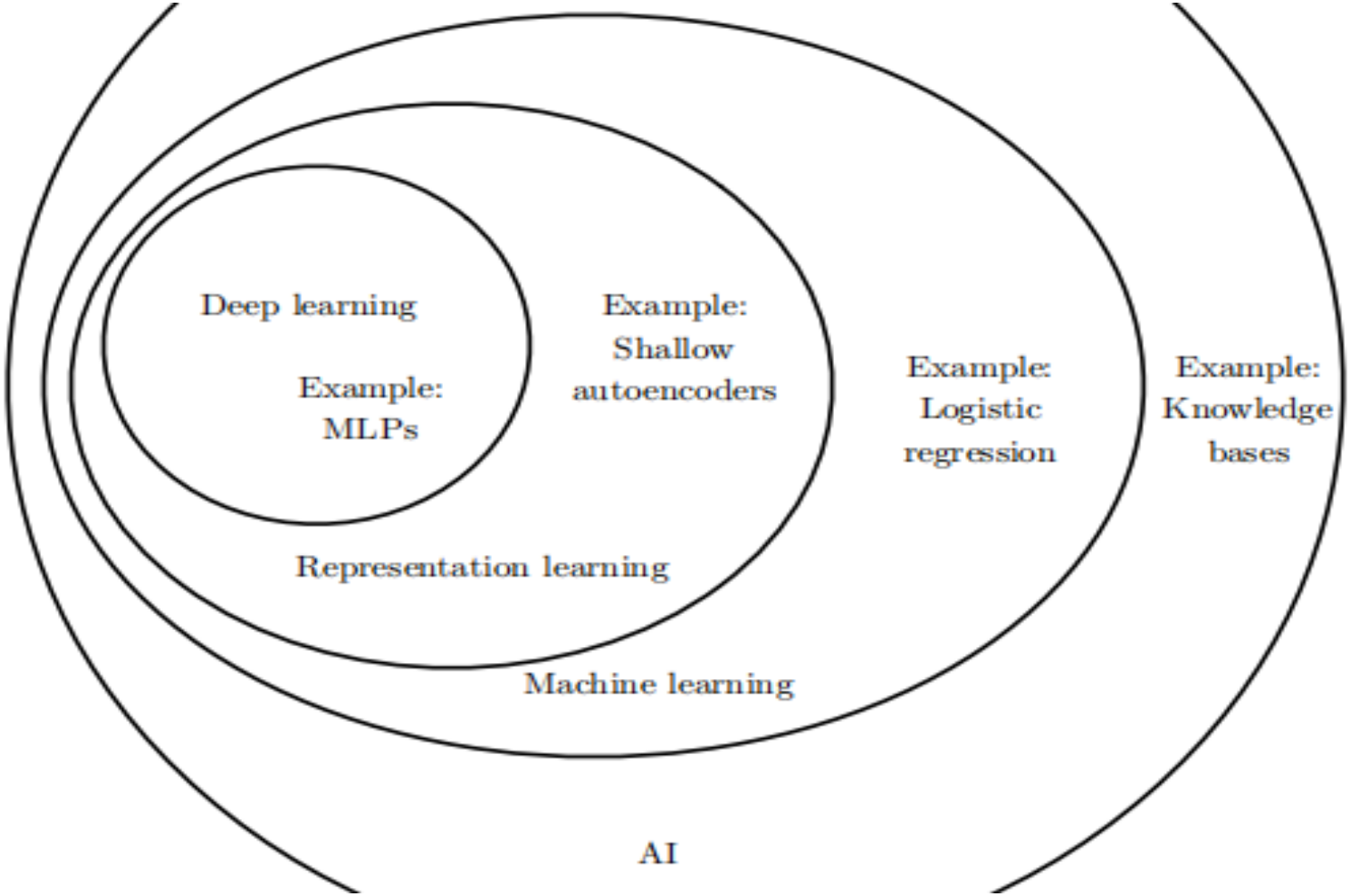
2.1 深度学习的定位

人工智能

机器学习

表示学习

深度学习



2.2 何谓“智能”

智，汉语常用字（一级字），读作zhì，最早出自甲骨文，本义是聪明，智力强。引申义有智慧、智谋、计谋、策略、有智慧的人等。

能，（拼音：néng、nài）是汉语通用规范一级字（常用字）。此字初文始见于商代甲骨文，其古字形像熊一类的野兽，这个意思后来写作“熊”。后来能假借为技能、能力的意思，又指有才能的意思。用作动词，指具备某种能力。能还表示主观上能够，其后常跟动词，如能行、能达到。由技能、能力引申，“能”在现代物理学上也指能量。

2.2.1 天工 . 开物

《天工开物》明 崇祯十年（1637） 宋应星
外国学者称它为“中国17世纪的工艺百科全书”。

作者在书中强调人类要和自然相协调、人力要与自然力相配合。

天工开物取自“天工人其代之”及“开物成务”，体现了朴素唯物主义自然观。

2.2.2 人工智能

- 人工智能（artificial intelligence，AI）就是让机器具有人类的智能。
  - “计算机控制”+“智能行为”

- 人工智能这个学科的诞生有着明确的标志性事件，就是1956年的达特茅斯（Dartmouth）会议。在这次会议上，“人工智能”被提出并作为本研究领域的名称。

人工智能就是要让机器的行为看起来就像是人所表现出的智能行为一样。

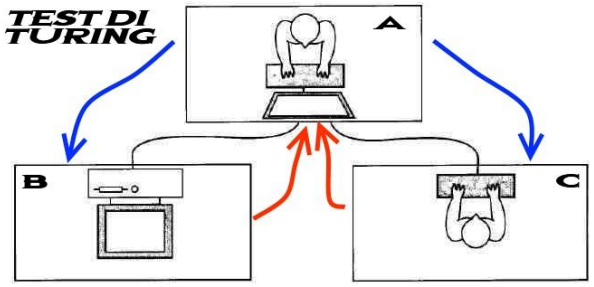
John McCarthy（1927-2011）

2.2.3 图灵测试

“一个人在不接触对方的情况下，通过一种特殊的方式，和对方进行一系列的问答。如果在相当长时间内，他无法根据这些问题判断对方是人还是计算机，那么就可以认为这个计算机是智能的”。

---Alan Turing [1950]
《Computing Machinery and Intelligence》

Alan Turing



2.3 人工智能的研究领域

- 让机器具有人类的智能
  - 机器感知（计算机视觉、语音信息处理）
  - 学习（模式识别、机器学习、强化学习）
  - 语言（自然语言处理）
  - 记忆（知识表示）
  - 决策（规划、数据挖掘）

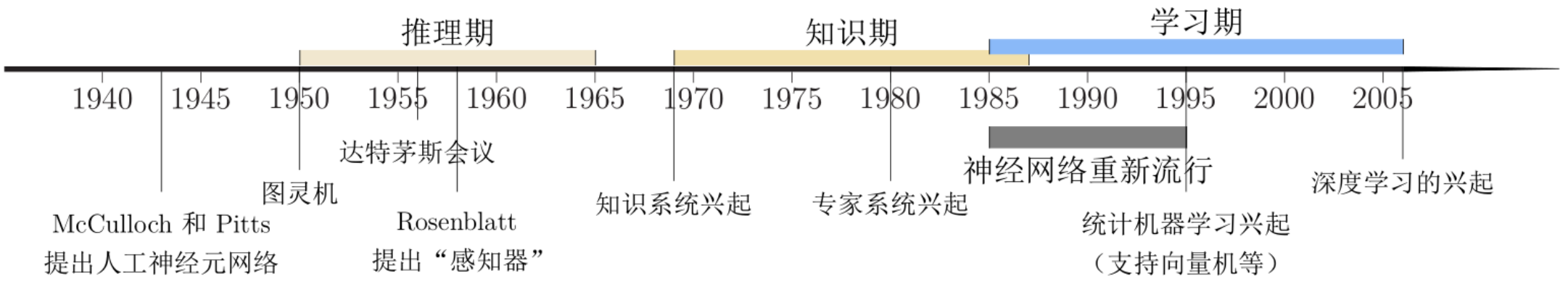
apmv：自我优化能力 预测能力 记忆能力 思维推理能力 判别能力 认知能力 感知能力

2.4.1 人工智能领域的流派

a、符号主义 ：逻辑主义、 心理学派
(推理期、知识期)
b、连接主义：仿生学派或生理学派

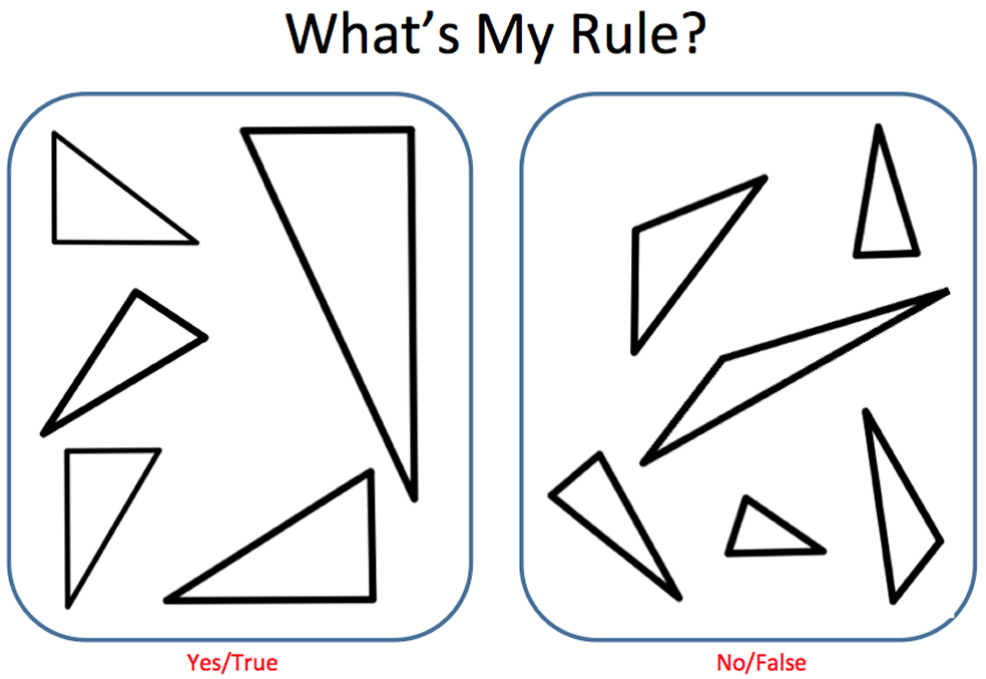
殊途同归、各有所长

2.4 人工智能的发展历史



2.4.2 如何开发一个人工智能系统?

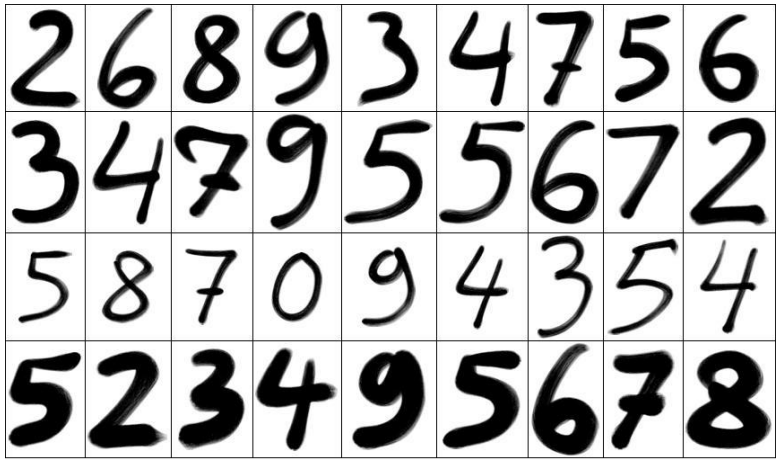
- 专家知识（人工规则）



专家系统的局限性

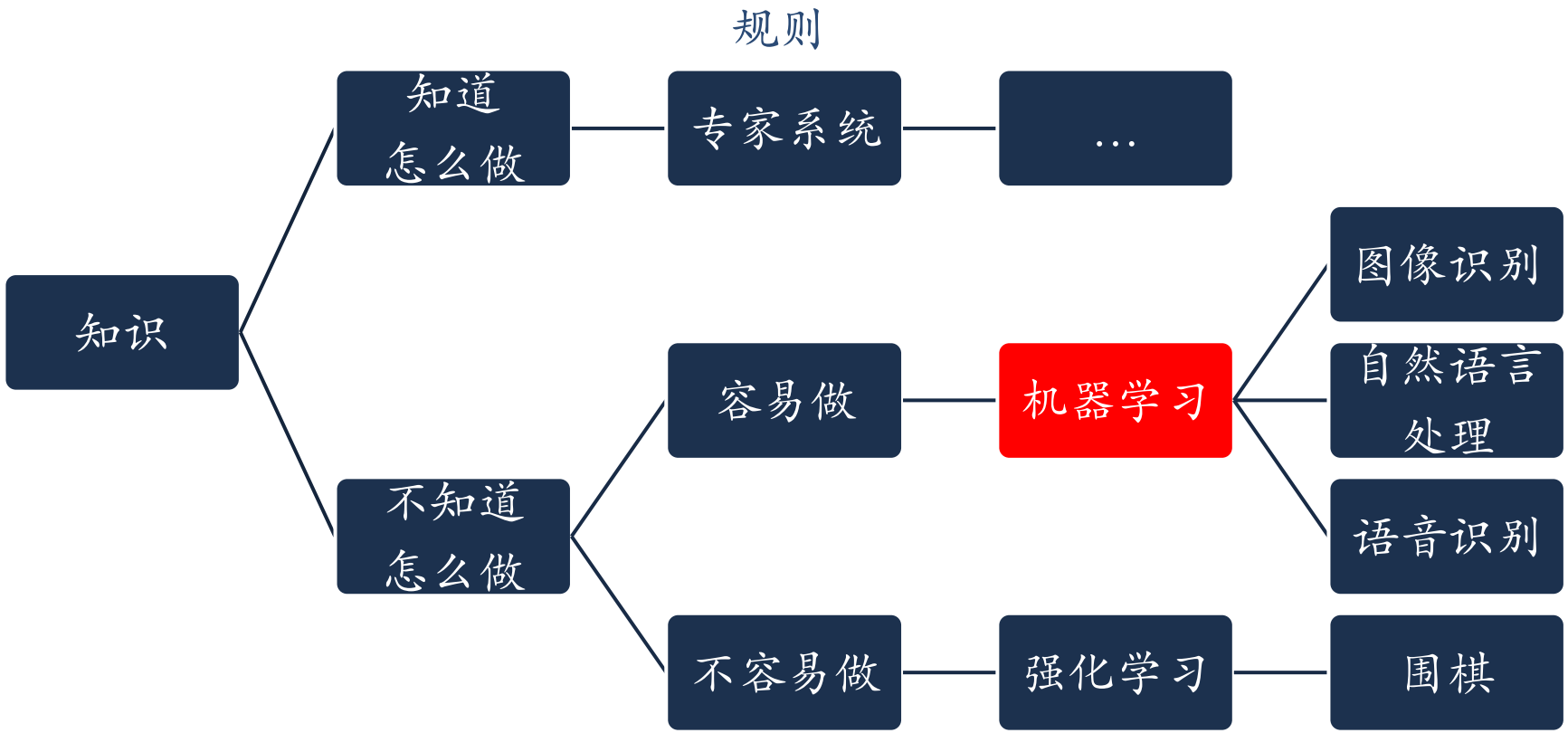
- 在特殊情况下无法做出创造性的回应
- 知识库中的错误可能导致错误的决策
- 专家系统的维护成本太高

2.4.3 What's the Rule?



机器学习

如何开发一个人工智能系统?



2.4.4 机器学习 ≈ 构建一个映射函数

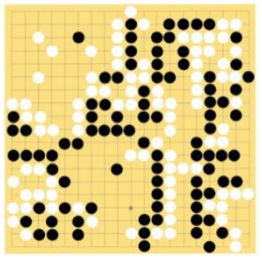
- 语音识别

$f($  $)=$ “你好”

- 图像识别

$f($  $)=$ “9”

- 围棋

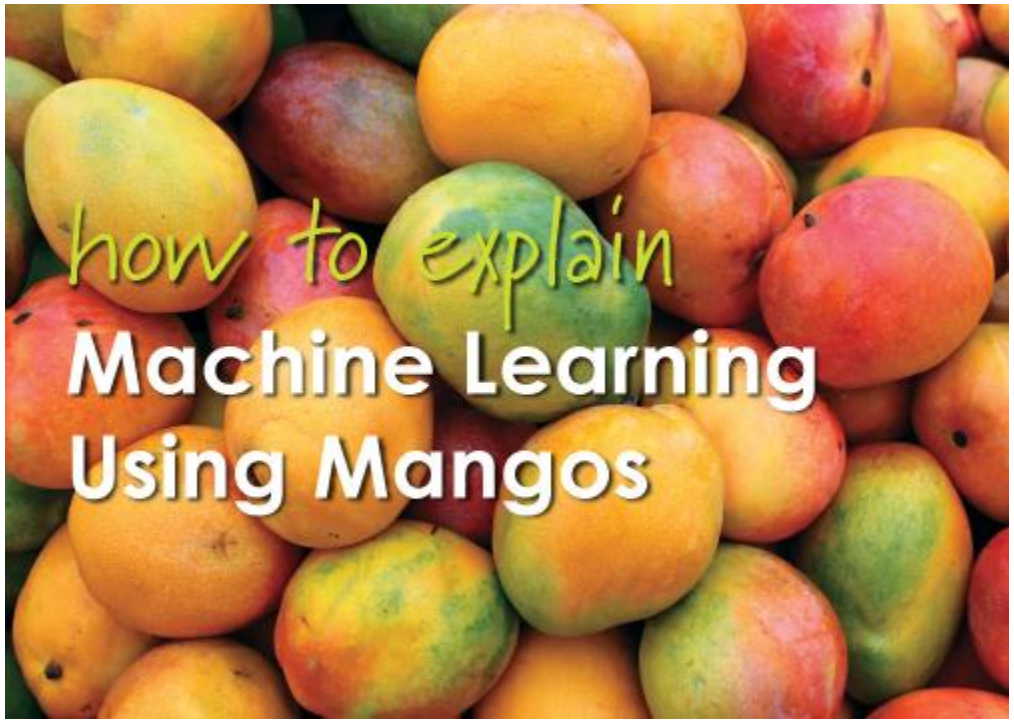
$f($  $)=$ “6-5” (落子位置)

- 机器翻译

$f($ “你好！” $)=$ “Hello!”

3.1 机器学习

如果判断芒果是否甜？



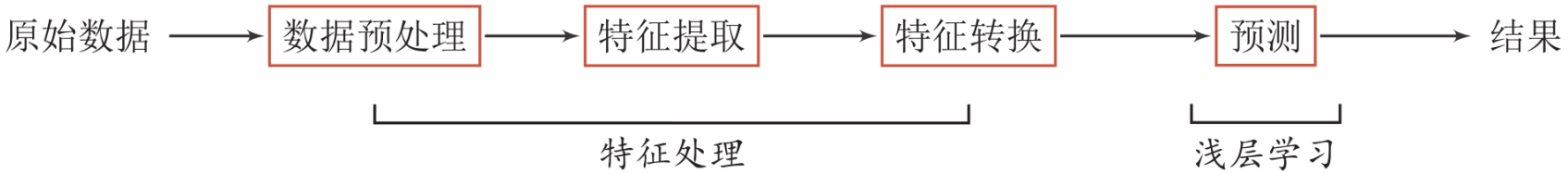
3.1.1 芒果机器学习

- 从市场上随机选取的芒果样本（训练数据），列出每个芒果的所有特征：
  - 如颜色，大小，形状，产地，品牌
- 以及芒果质量（输出变量）：
  - 甜蜜，多汁，成熟度。

- 设计一个学习算法来学习芒果的特征与输出变量之间的相关性模型。
-

3.1.2 机器学习的一般流程

- 当我们用机器学习来解决一些模式识别任务时，一般的流程包含以下几个步骤：



特征工程（Feature Engineering）

- 浅层学习（Shallow Learning ）

定义：传统的机器学习主要关注如何学习一个预测模型．一般需要首先将数据表示为一组特征（Feature），特征的表示形式可以是连续的数值、离散的符号或其他形式．然后将这些特征输入到预测模型，并输出预测结果．这类机器学习可以看作浅层学习（Shallow Learning ）

3.2 表示学习

- 数据表示是机器学习的核心问题。
  - 特征工程：需要借助人类智能
- 表示学习
  - 如何自动从数据中学习好的表示

如果有一种算法可以自动地学习出有效的特征，并提高最终机器学习模型的性能，那么这种学习就可以叫作表示学习（Representation Learning）

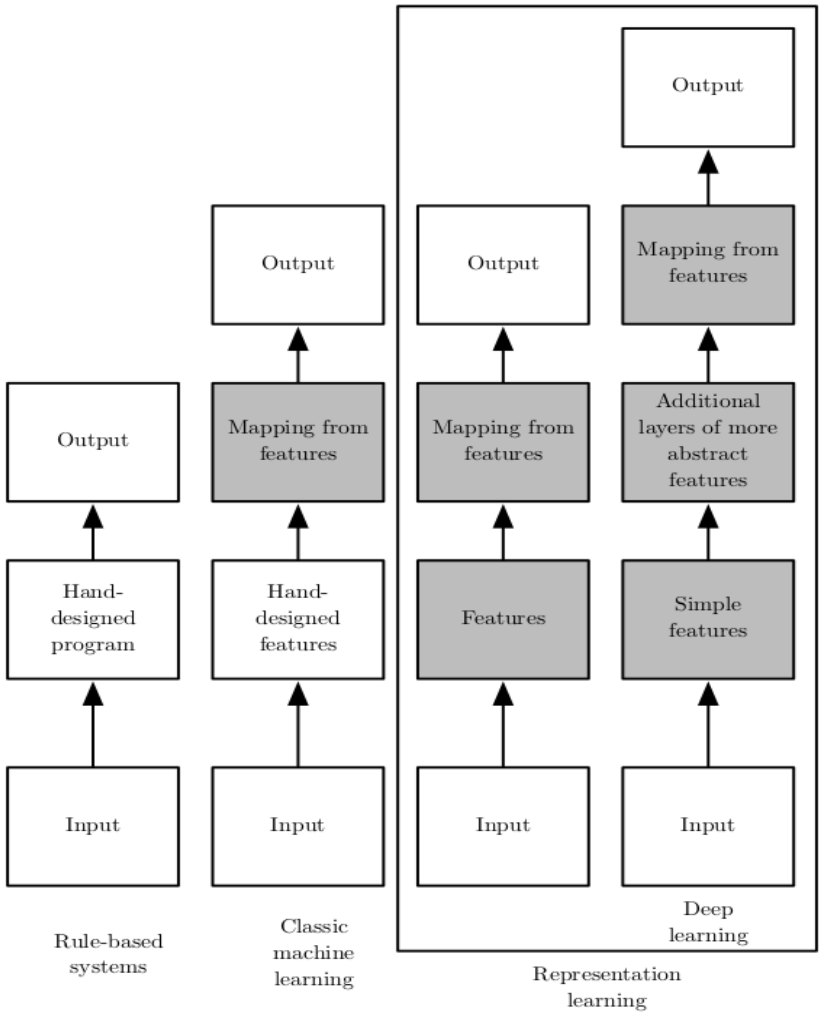


图 1.5: 流程图展示了 AI 系统的不同部分如何在不同的 AI 学科中彼此相关。阴影框表示能从数据中学习的组件。

3.2.1 语义鸿沟：人工智能的挑战之一

- 底层特征 VS 高层语义
  - 人们对文本、图像的理解无法从字符串或者图像的底层特征直接获得

床前明月光，

疑是地上霜。

举头望明月，

低头思故乡。

3.2.1 什么是好的数据表示?

- “好的表示”是一个非常主观的概念，没有一个明确的标准。

- 但一般而言，一个好的表示具有以下几个优点：
  - 应该具有很强的表示能力。
  - 应该使后续的学习任务变得简单。
  - 应该具有一般性，是任务或领域独立的。

3.2.2 语义表示

- 如何在计算机中表示语义？

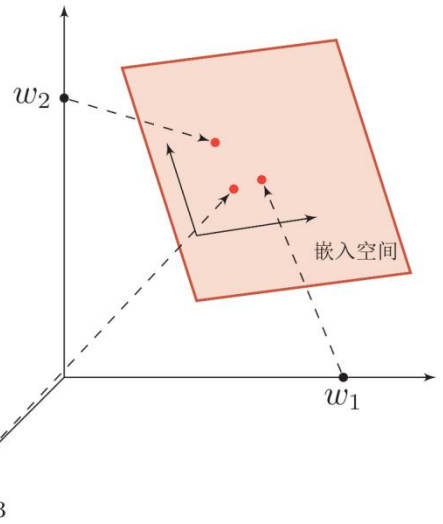
局部（符号）表示

知识库、规则

分布式表示

嵌入：压缩、低维、稠密向量

One-Hot向量



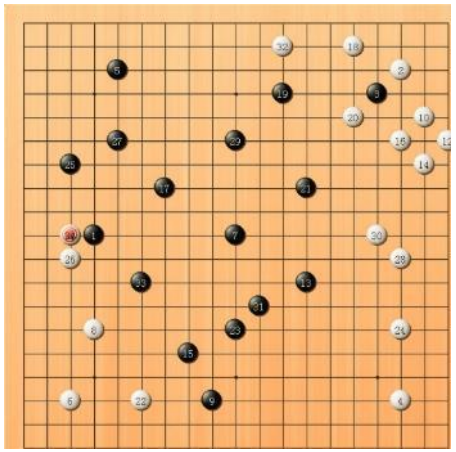
3.2.2.1 表示形式

局部表示（Local Representation）

- 离散表示、符号表示
- One-Hot向量

	局部表示	分布式表示
A	[1 0 0 0]	[0.25 0.5]
B	[0 1 0 0]	[0.2 0.9]
C	[0 0 1 0]	[0.8 0.2]
D	[0 0 0 1]	[0.9 0.1]

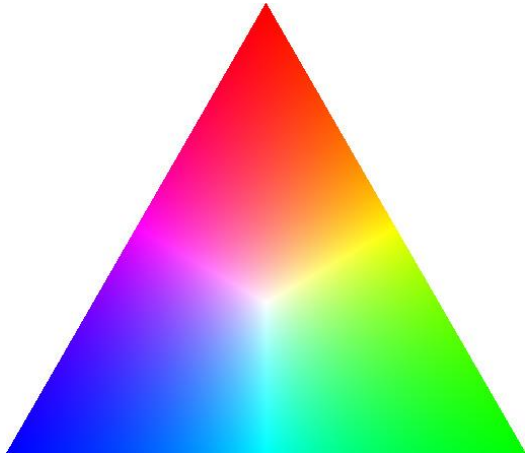
- 分布式(distributed)表示
  - 压缩、低维、稠密向量
  - 用O(N)个参数表示 $O(3^k)$区间
    - k为非0参数，k<N



分布式表示

3.2.2.2 一个生活中的例子：颜色

颜色	局部表示	分布式表示
琥珀色	$[1,\ 0,\ 0,\ 0]^{\mathrm{T}}$	$[1.00,\ 0.75,\ 0.00]^{\mathrm{T}}$
天蓝色	$[0,\ 1,\ 0,\ 0]^{\mathrm{T}}$	$[0.00,\ 0.5,\ 1.00]^{\mathrm{T}}$
中国红	$[0,\ 0,\ 1,\ 0]^{\mathrm{T}}$	$[0.67,\ 0.22,\ 0.12]^{\mathrm{T}}$
咖啡色	$[0,\ 0,\ 0,\ 1]^{\mathrm{T}}$	$[0.44,\ 0.31\ 0.22]^{\mathrm{T}}$



局部表示的特点

- 局部表示有两个优点：
  - 1）这种离散的表示方式具有很好的解释性，有利于人工归纳和总结特征，并通过特征组合进行高效的特征工程；
  - 2）通过多种特征组合得到的表示向量通常是稀疏的二值向量，当用于线性模型时计算效率非常高。
- 两个不足之处：
  - 1）one-hot向量的维数很高，且不能扩展．如果有一种新的颜色，我们就需要增加一维来表示；
  - 2）不同颜色之间的相似度都为0，即我们无法知道“红色”和“中国红”的相似度要高于“红色”和“黑色”的相似度。

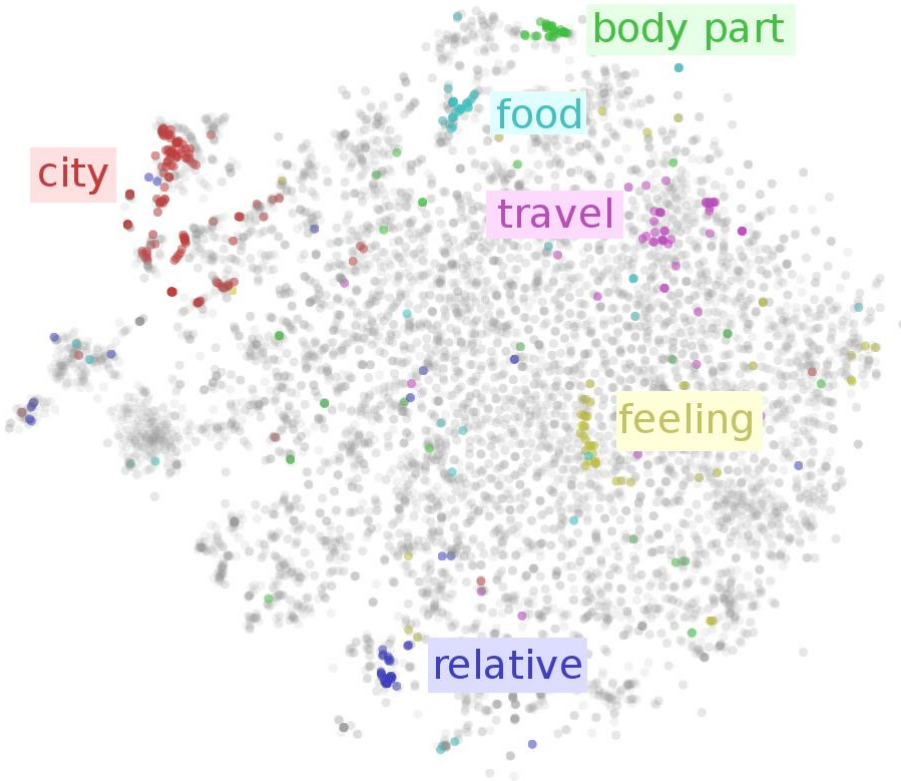
嵌入

- 嵌入（Embedding）定义：
  - 我们可以使用神经网络来将高维的局部表示空间$\mathbb{R}^{|\mathcal{V}|}$映射到一个非常低维 的分布式表示空间 $\mathbb{R}^{D}, D \ll |\mathcal{V}|$. 在这个低维空间中，每个特征不再是坐标轴上的点，而是分散在整个低维空间中．在机器学习中，这个过程也称为嵌入（Embedding）．
- 嵌入通常指将一个度量空间中的一些对象映射到另一个低维的度量空间中，并尽可能保持不同对象之间的拓扑关系。

3.2.2.3 词嵌入（Word Embeddings）

上海
北京

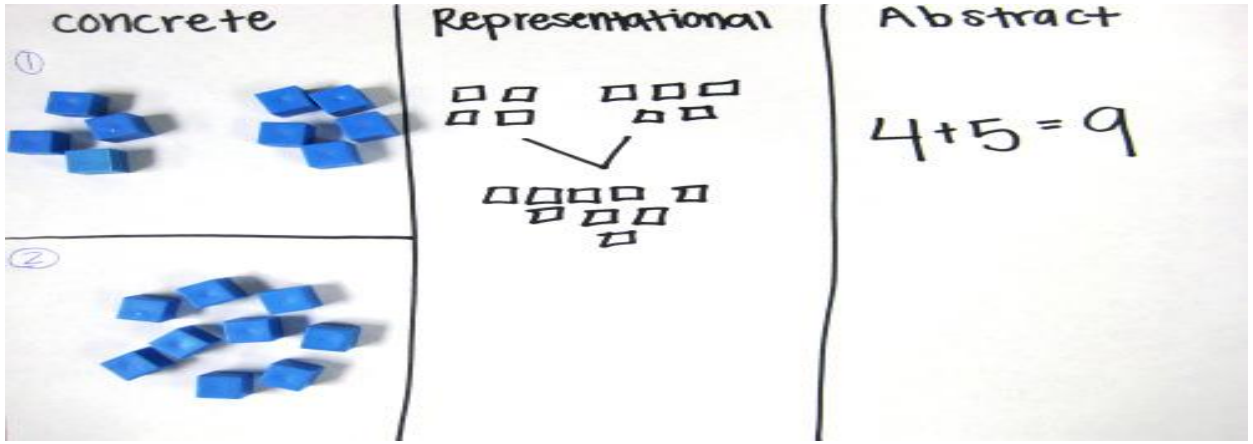
高兴
难过



https://indico.io/blog/visualizing-with-t-sne/

3.2.2.4 表示学习与深度学习

- 一个好的表示学习策略必须具备一定的深度
  - 特征重用
    - 指数级的表示能力
  - 抽象表示
    - 抽象表示需要多步的构造



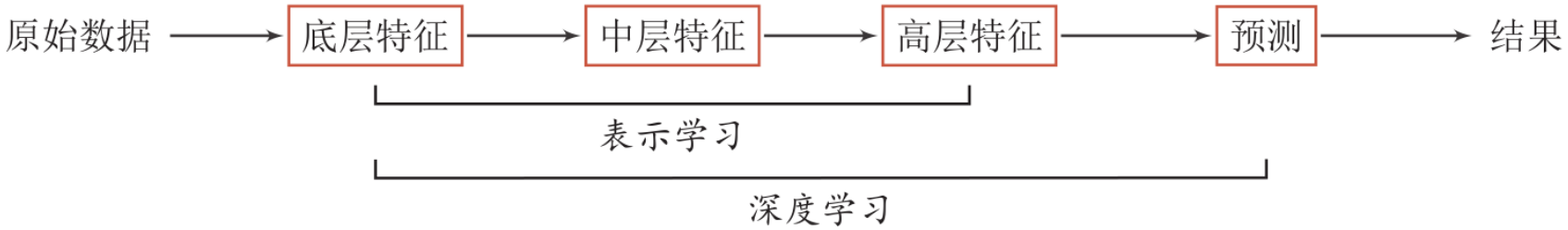
3.2.2.5 传统的特征提取

- 特征提取
  - 线性投影（子空间）
    - PCA、LDA
  - 非线性嵌入
    - LLE、Isomap、谱方法
  - 自编码器

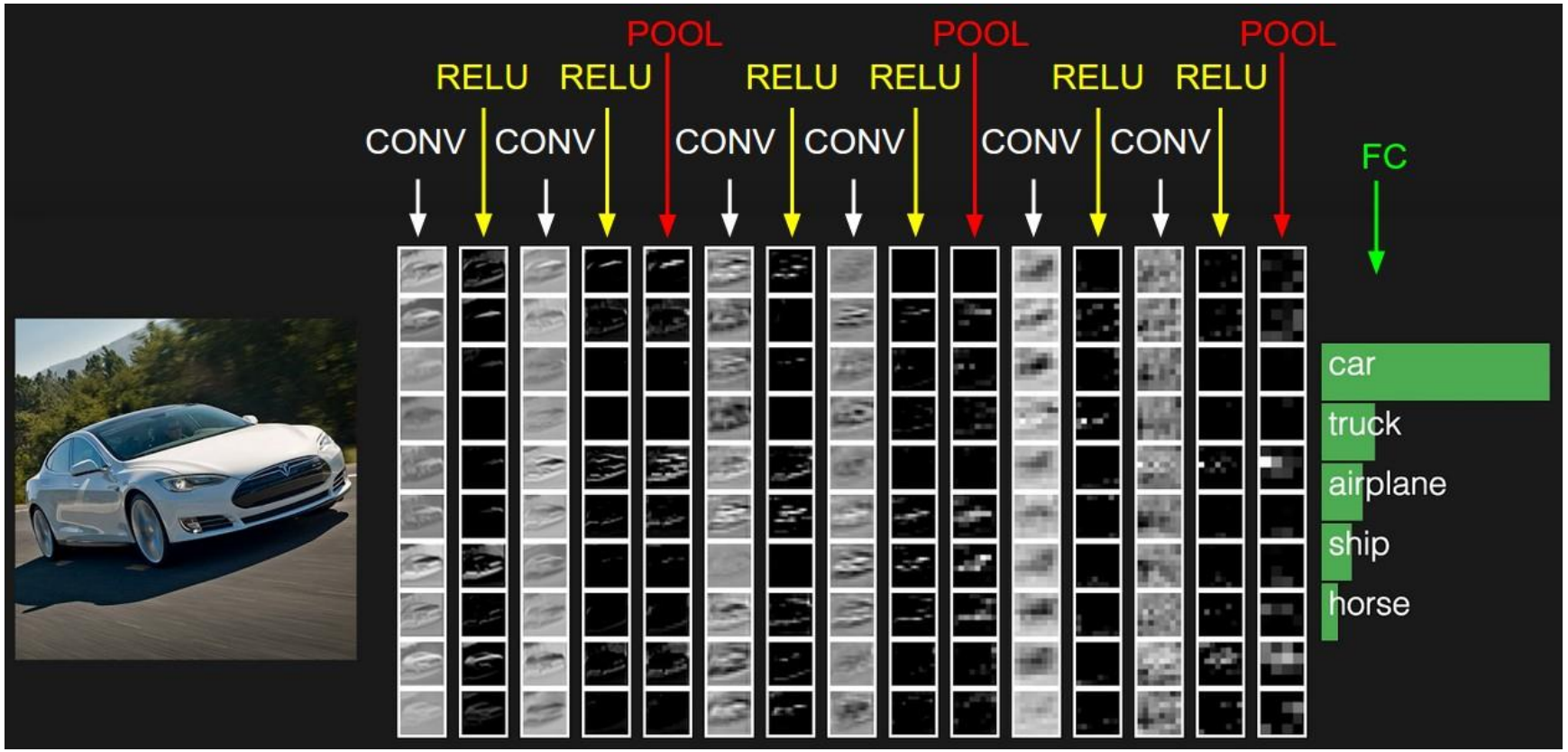
- 特征提取VS表示学习
  - 特征提取：基于任务或先验去除无用特征
  - 表示学习：通过深度模型学习高层语义特征

3.3 深度学习

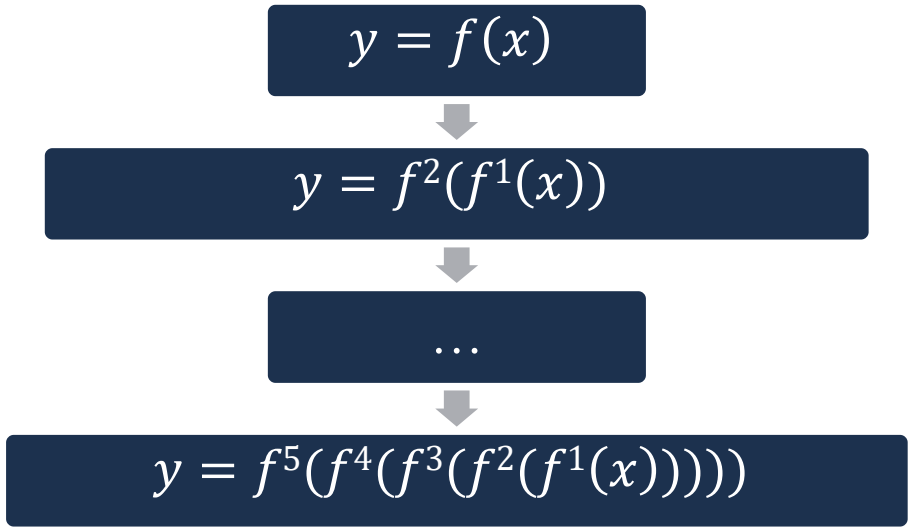
- 通过构建具有一定“深度”的模型，可以让模型来自动学习好的特征表示（从底层特征，到中层特征，再到高层特征），从而最终提升预测或识别的准确性。



3.3.1 表示学习与深度学习



3.3.2 深度学习的数学描述



浅层学习

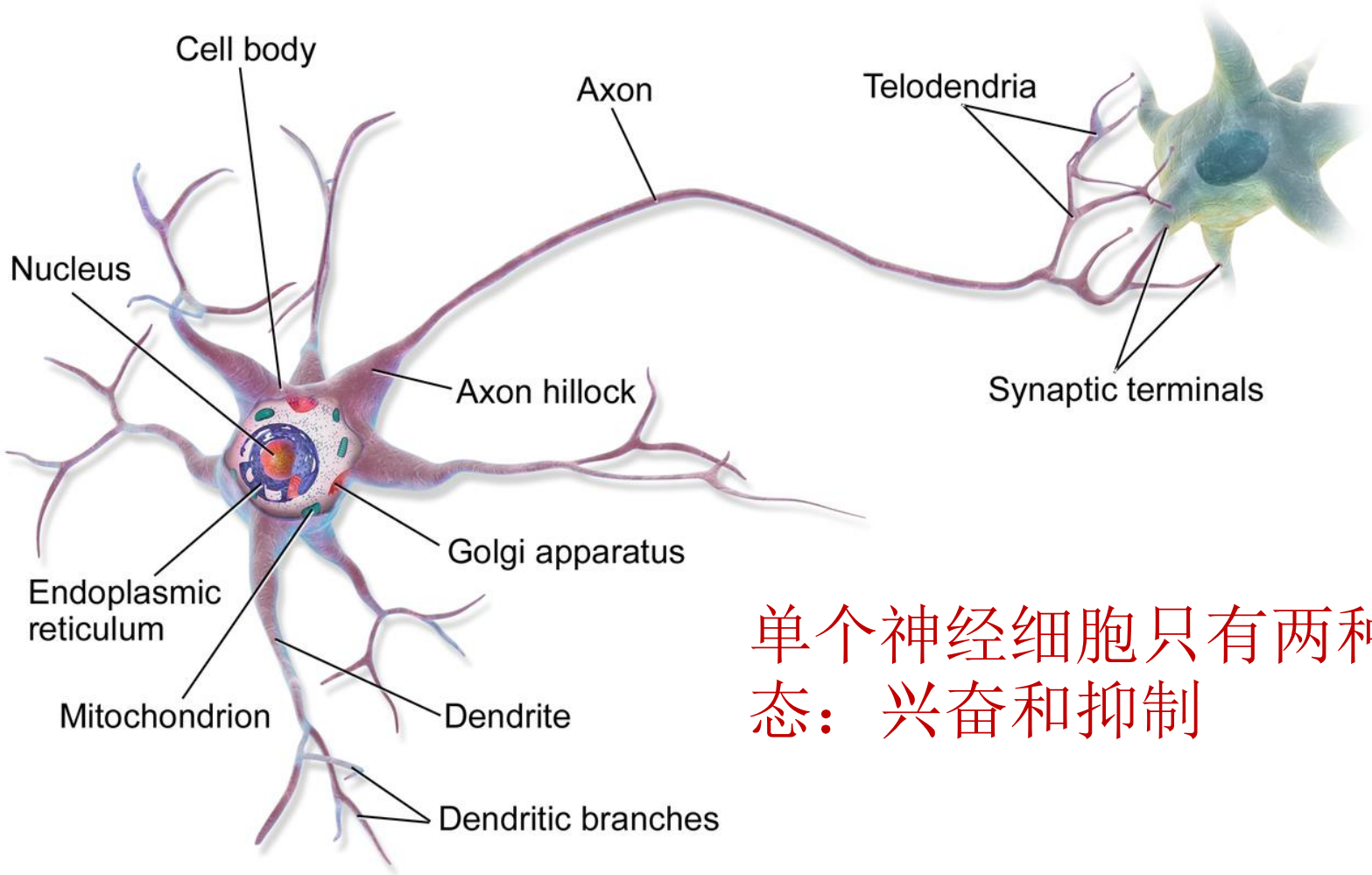
深度学习

当$f^l(x)$连续时，　比如$f^l(x) = \sigma\left(W^l f^{l-1}(x)\right)$，这个复合函数称为神经网络。

神经网络

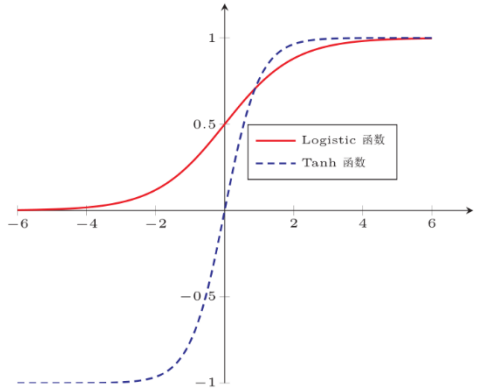
4.1 生物神经元

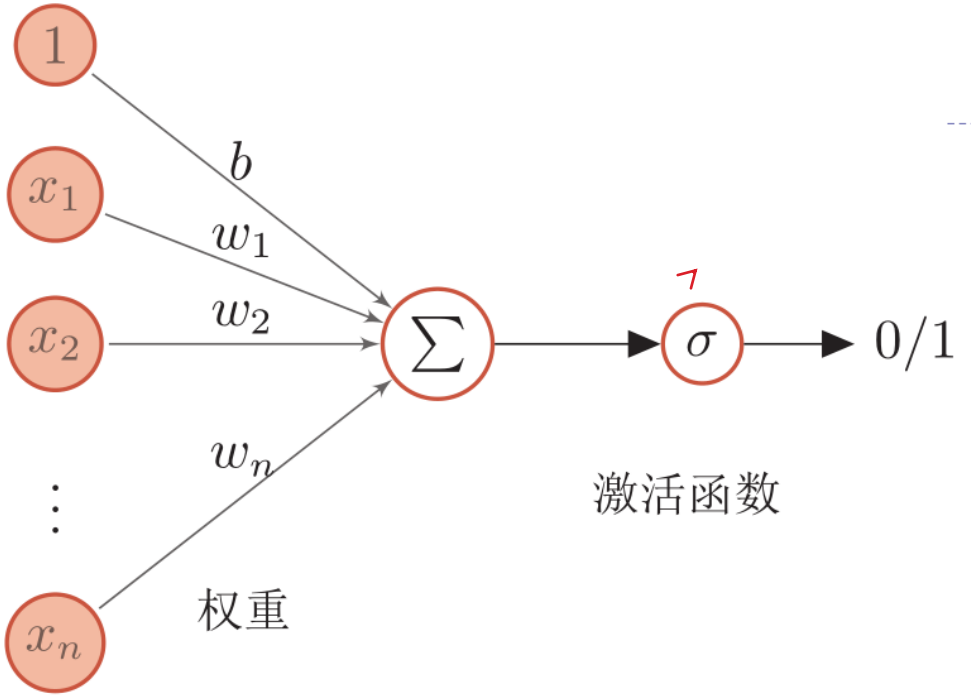
人脑有860亿个神经元　　video: structure of brain



单个神经细胞只有两种状态：兴奋和抑制

4.1.2 人工神经元


1
0.5
−0.5
−1
−6
−4
−2
2
4
6
Logistic 函数
Tanh 函数


1
x_1
x_2
x_n
b
w_1
w_2
w_n
$\sum$
σ
0/1
激活函数
权重

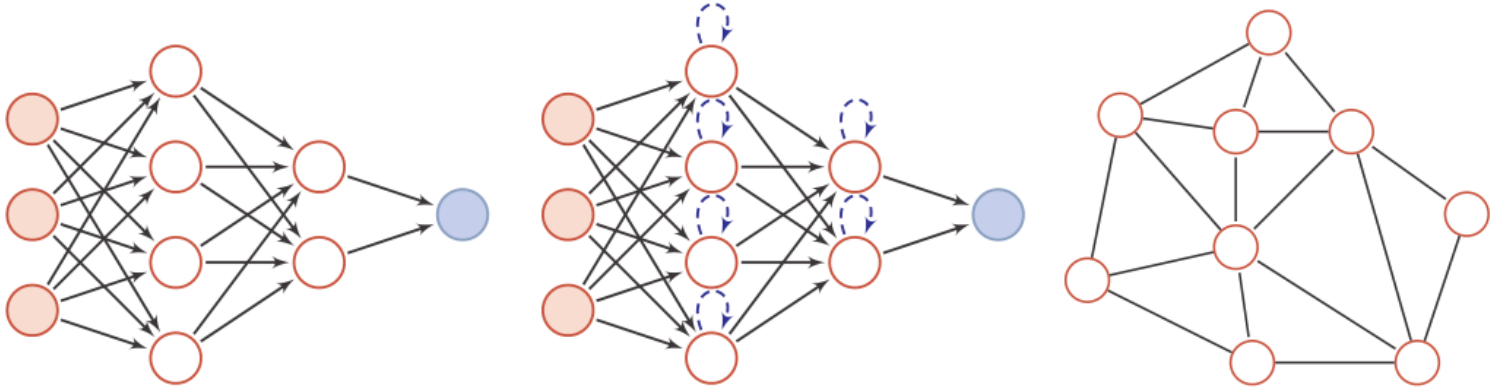
4.1.3 人工神经网络

- 人工神经网络主要由大量的神经元以及它们之间的有向连接构成。因此考虑三方面：

- 神经元的激活规则
  - 主要是指神经元输入到输出之间的映射关系，一般为非线性函数。
- 网络的拓扑结构
  - 不同神经元之间的连接关系。
- 学习算法
  - 通过训练数据来学习神经网络的参数。

4.1.4 人工神经网络

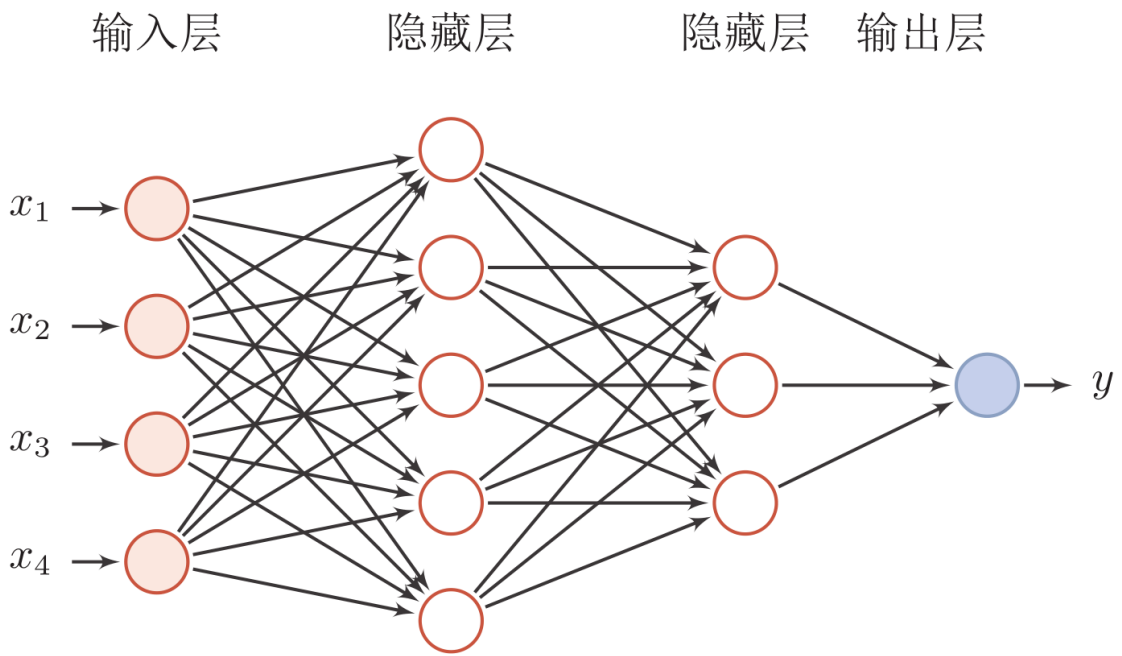
- 人工神经网络由神经元模型构成，这种由许多神经元组成的信息处理网络具有并行分布结构。
  - 虽然这里将 神经网络结构大体上分为三种类型，但是大多数网络都是复合型结构，即一个神经网络中包括多种网络结构。



(a) 前馈网络　(b) 记忆网络　(c) 图网络

4.1.5 神经网络

$$f_l(x) = \sigma(W_l f_{l-1}(x))$$



学习过程

神经网络学习过程主要有：

监督学习（supervised learning） → **有教师学习**

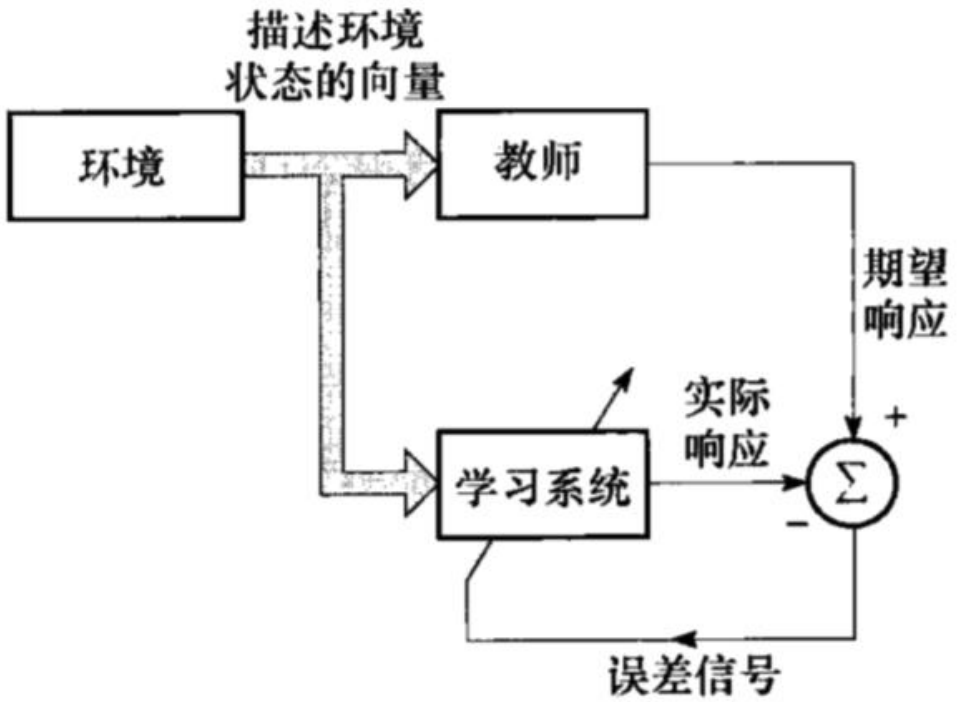
无监督学习（unsupervised learning）

强化学习（reinforcement learning） → **无教师学习**

Author:WangQi
Email:wangqikaixin@gmail.com

监督学习

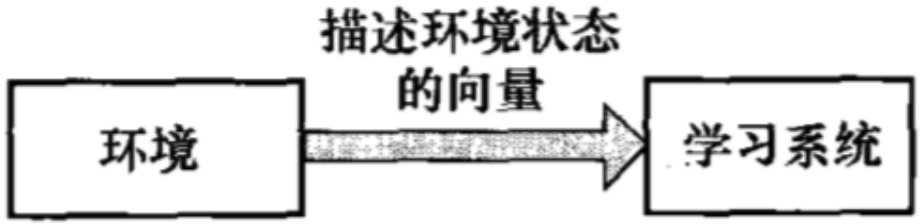
也称为“有教师学习”，即在“外界”帮助下改正自己的错误。



Author:WangQi
Email:wangqikaixin@gmail.com

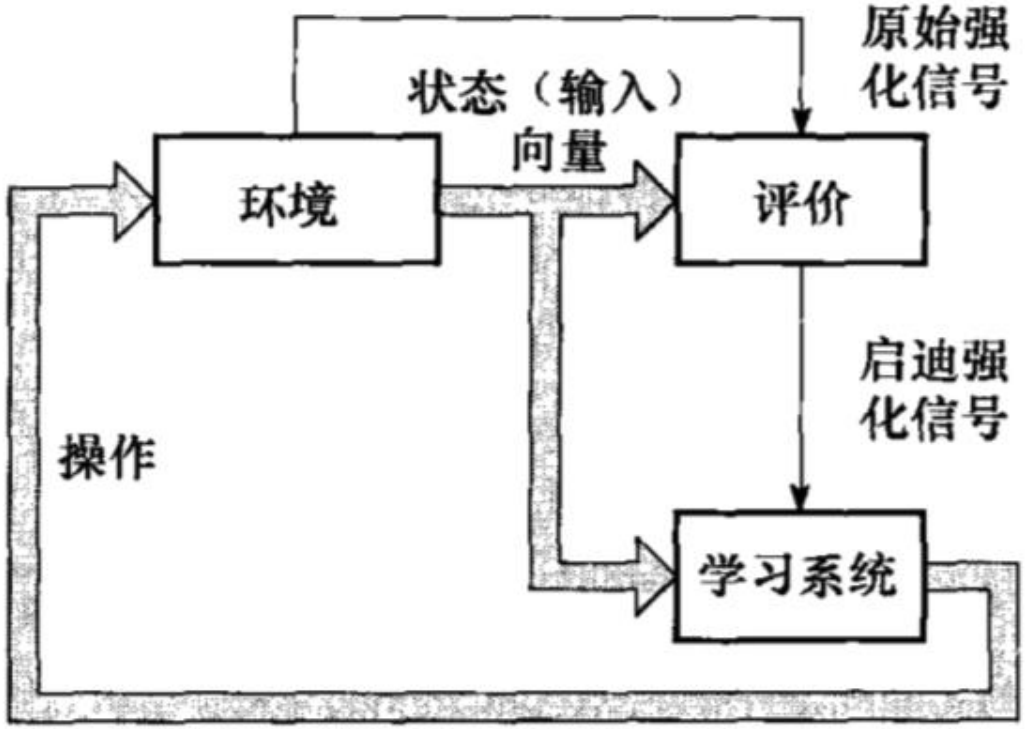
无监督学习

无需“外界”提供标准答案，自己总结数据中的规律



强化学习

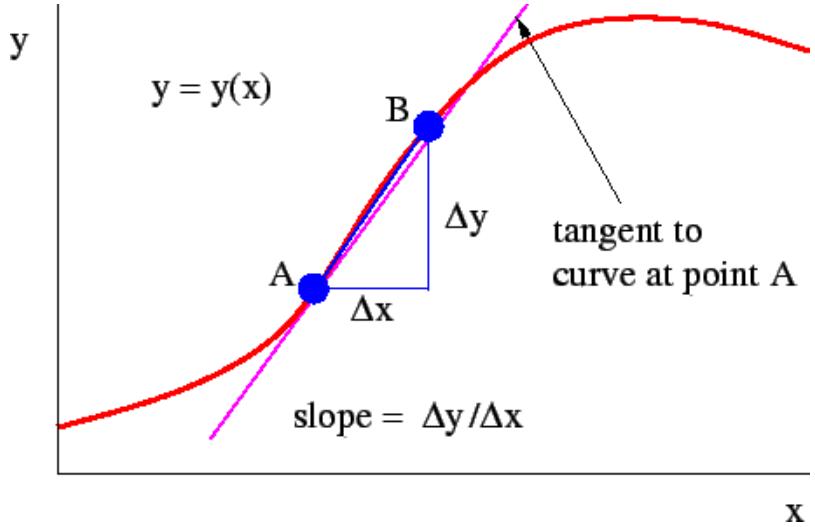
基于环境而行动，以取得最大化的预期收益。



Author:WangQi
Email:wangqikaixin@gmail.com

4.1.6 如何解决贡献度分配问题?

- 偏导数



- 贡献度

$$\frac{\partial y}{\partial W_l} = \frac{y(W_l + \Delta W_l) - y(W_l)}{\Delta W_l}$$

4.2.1 神经网络发展史

- 神经网络的发展大致经过五个阶段。
- 第一阶段：模型提出
  - 在1943年，心理学家Warren McCulloch和数学家Walter Pitts和最早描述了一种理想化的人工神经网络，并构建了一种基于简单逻辑运算的计算机制。他们提出的神经网络模型称为MP模型。
  - 阿兰·图灵在1948年的论文中描述了一种“B型图灵机”。(赫布型学习)
  - 1951年，McCulloch和Pitts的学生Marvin Minsky建造了第一台神经网络机，称为SNARC。
  - Rosenblatt [1958]最早提出可以模拟人类感知能力的神经网络模型，并称之为感知器（Perceptron），并提出了一种接近于人类学习过程（迭代、试错）的学习算法。

4.2.2 神经网络发展史

- 第二阶段：冰河期
  - 1969年，Marvin Minsky出版《感知器》一书，书中论断直接将神经网络打入冷宫，导致神经网络十多年的“冰河期”。他们发现了神经网络的两个关键问题：
    - 1）基本感知器无法处理异或回路。
    - 2）电脑没有足够的能力来处理大型神经网络所需要的很长的计算时间。
  - 1974年，哈佛大学的Paul Webos发明反向传播算法，但当时未受到应有的重视。
  - 1980年，Kunihiko Fukushima（福岛邦彦）提出了一种带卷积和子采样操作的多层神经网络：新知机（Neocognitron）

4.2.3 神经网络发展史

- 第三阶段：反向传播算法引起的复兴
  - 1983年，物理学家John Hopfield对神经网络引入能量函数的概念，并提出了用于联想记忆和优化计算的网络（称为Hopfield网络），在旅行商问题上获得当时最好结果，引起轰动。
  - 1984年，Geoffrey Hinton提出一种随机化版本的Hopfield网络，即玻尔兹曼机。
  - 1986年， David Rumelhart和James McClelland对于联结主义在计算机模拟神经活动中的应用提供了全面的论述，并重新发明了反向传播算法。
  - 1986年，Geoffrey Hinton等人将引入反向传播算法到多层感知器
  - 1989 年，LeCun等人将反向传播算法引入了卷积神经网络，并在手写体数字识别上取得了很大的成功。

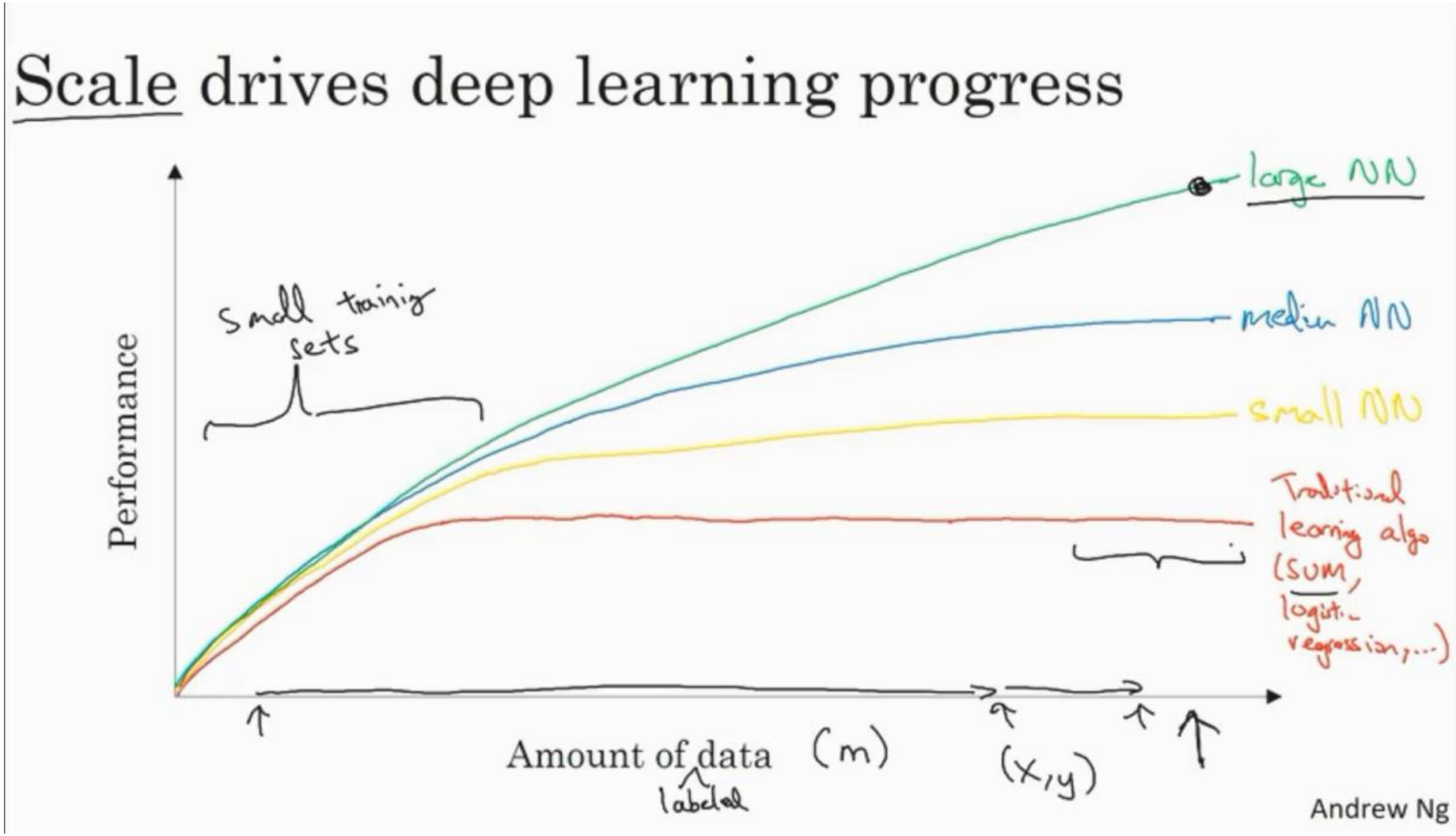
4.2.4 神经网络发展史

- 第四阶段：流行度降低
  - 在20世纪90年代中期，统计学习理论和以支持向量机为代表的机器学习模型开始兴起。
  - 相比之下，神经网络的理论基础不清晰、优化困难、可解释性差等缺点更加凸显，神经网络的研究又一次陷入低潮。

4.2.5 神经网络发展史

- 第五阶段：深度学习的崛起
  - 2006年，Hinton等人发现多层前馈神经网络可以先通过逐层预训练，再用反向传播算法进行精调的方式进行有效学习。
    - 深度神经网络在语音识别和图像分类等任务上的巨大成功。
  - 2013年，AlexNet：第一个现代深度卷积网络模型，是深度学习技术在图像分类上取得真正突破的开端。
    - AlexNet不用预训练和逐层训练，首次使用了很多现代深度网络的技术
  - 随着大规模并行计算以及GPU设备的普及，计算机的计算能力得以大幅提高。此外，可供机器学习的数据规模也越来越大。在计算能力和数据规模的支持下，计算机已经可以训练大规模的人工神经网络。

三驾马车


Scale drives deep learning progress
Performance
Small training sets
large NN
medium NN
small NN
Traditional learning algo (SVM, logistic regression,...)
Amount of data (m)
labeled
(x,y)
Andrew Ng

4.3 常用的深度学习框架

- 简易和快速的原型设计
- 自动梯度计算
- 无缝CPU和GPU切换

1. 非线性（**nonlinearity**）：人工神经元可以是线性的，也可以是非线性的。非线性是一个非常重要的特性，特别是当产生输入信号的内部物理机制是天生非线性的时候。

2. 输入输出映射（**input-output mapping**）：即可进行监督学习。

3. 自适应性（**adaptivity**）：神经网络具有调整自身突触权值以适应外界环境变化的固有能力。

4. 证据响应（**evidential response**）：在模式分类问题中，神经网络可以设计成不仅提供选择哪一个特定模型的信息，还提供关于决策的置信度信息。后者可以用来拒判那些可能出现的过于模糊的模式，从而进一步改善网络的分类性能。

5. 上下文信息（**contextual information**）：神经网络的特定结构和激发状态代表知识。网络中每一个神经元都受网络中所有其他神经元全局活动的潜在影响。因此，神经网络很自然地能够处理上下文信息。

6. 容错性（**fault tolerance**）：神经网络在不利的运行条件下的性能是逐步下降的。比如一个神经元或其连接坏了，存储模式的记忆性在质量上会被削弱。但由于网络信息存储的分布特性，部分神经元的损坏不会造成灾难性的后果。

7. 超大规模集成实现（**VLSI implementability**）：神经网络的大规模并行性使它具有快速处理某些任务的潜在能力。

8. 分析和设计的一致性：神经网络作为信息处理器具有通用性。神经元，不管形式如何，在所有的神经网络中都代表一种相同成分。这种共性使得在不同应用中的神经网络共享相同的理论和学习算法成为可能。模块化网络可以用模块的无缝集成来实现。

4.5 深度学习发展困境

- 对数据的改变过于敏感，对抗样本带来了模型输出结果的不确定性；
- 依赖巨大的计算资源和数据资源；
- CNN、RNN等模型与人的思维方式差异极大
- 难以解决逻辑推理、因果关系等问题。