第 1 章 绪论

在这一章里面，我们首先审视正在高歌猛进的数据科学，了解使用 GPU 进行机器学习计算的重要性。接下来，我们回顾机器学习的发展历程，并且检阅当前机器学习技术的几项最高成就：人工智能围棋（AlphaGo）、深度神经网络图像识别（ImageNet）和 IBM Waston 人工智能系统，从而领略机器学习技术震撼世界的脚步。第三部分，我们对机器学习算法进行概略分类，并且根据分类结果介绍本书内容。

1.1 概述

我们生活在一个伟大的时代，人类文明史上最卓越的心智成就以前所未有的深度、广度和速度交汇融合，催生出潜力无限的数据科学（data science）。数据科学是在人类社会数字化程度充分发展的前提下，综合计算机科学、数学和神经科学等领域的理论和技术成果，以数据挖掘作为应用形式，通过对数据进行存储、分析和可视化等各种处理，从中提炼信息并形成知识，从而引导优化决策的科学。简单说来，数据科学就是针对大数据的理论和方法。

当前，数据科学已经深度融入我们的日常生活，我们可以从一天的平凡生活中检查一下数据科学在怎样发挥作用的：上下班路上，导航系统会分析数据告诉我们不同路线的拥堵情况并且实时预测预计行程时间，如果乘坐公车的话，还可以通过历史数据和实时路况预报公车到站时间；我们打电话时，电信运营商会通过采集我们打电话的模式，诸如地点、时间段和服务套餐情况（但是不能使用时频、语音和个人帐号信息），推断我们的身份、生活习惯和经济状况，从而确定相应的推送内容；我们上网冲浪时，搜索引擎提供的内容当然是对海量网页进行分析处理的结果，而且也会我把我们的搜索内容拿去分析，从中提炼热点搜索趋势，并且对我们的行为进行推断；购物时，无论是电商还是传统商户，都可能分析我们的购物历史决定向我们推荐商品，而在付款之中或之后，银行的数据分析系统会判断这是一次正常消费还是一次欺诈；工作时，即使我们不直接使用数据分析工具，也几乎不可避免地在产生或者消费数据，有些公司（例如惠普）甚至使用预测软件分析每个雇员辞职的可能性[1]；除此之外，还有更多的数据分析系统在暗中“琢磨”我们，比如说医疗保险公司在算计我们未来的健康趋势，由此

[1] 有趣的是，数据分析师自己经常被判别为潜在离职风险较高的雇员，因为社会需求极为迫切。

决定保费应该怎样变化，社交网络公司在计算是否发现了你的同学或者熟人，或者怎样让你的社交圈通过最短路径和其它群落连接起来，还有基金公司会分析社交网络上大家的情感趋势，以此作为预测证券价格涨落的依据，如果你是单身而且在征婚网站登记的话，还会有数据分析引擎根据你的资料进行分类和匹配，为你寻找合适的另一半。

数据科学向社会生活的渗透正在以不可阻挡的势头在更大范围上更加深化。表 1-1 是远不完全（实际上完整枚举数据应用已经成为不可能完成的任务）的典型数据科学应用的清单。

表 1.1 典型数据应用

公司/组织	代表性数据应用	亮点
谷歌 Google	对全球 35 万亿个网页进行索引，并形成 1 亿 G 字节的索引记录	全部 Internet 搜索服务的 89%由 Google 提供
亚马逊 Amazon	采集并分析其 7.5 亿顾客的购物行为（包括购物和浏览），分析顾客的收入和偏好，从而为顾客进行商品推荐	Amazon 的推荐系统是其成为美国最大线上零售商（年产值 900 亿美元）的主要助力，也是其品牌的重要标志
网飞 Netflix	根据电影内容进行分类，并根据用户观看电影的历史进行喜好分析并推荐电影	非结构化数据学习的经典技术，是 Netflix 用户和流量继续加速增长的主要动力
沃尔玛 Walmart	利用购物篮分析推荐商品，使用社会和环境数据预测购买需求	沃尔玛自行开发的 Data Café 数据分析系统处理一个拥有 2000 亿组交易数据的数据库，能够把销售问题平均解决时间从 2~3 周降低至 20 分钟左右
欧洲核子研究组织 CERN	分析数据中的特殊能量特征，从中确定是否发现特定粒子	每年产生 30PB 数据，主要是粒子对撞机中粒子碰撞时产生的光信号，2013 年通过分析数据发现了希格斯玻色子
罗尔斯-罗伊斯 Rolls-Royce	分析发动机实时监控数据，确定优化维护和修理方案	支撑全球 500 家以上航空公司和 150 多支空军的航空发动机，大数据技术显著降低了运维成本
壳牌石油 Shell	分析地址数据发现油田	大幅度提高了勘探精度
莲花 F1 车队 Lotus F1 Team	分析赛场数据实时调整塞车参数，利用数据建立仿真模型优化赛车设计	把青年车手 Marlon Stockinger 的赛季总成绩从 2013 年的全球第 18 名提高到 2014 年的第 9 名
脸书 Facebook	分析用户数据推送广告	2014 年占据美国 24%的在线广告份额，创收 53 亿美元；预计 2017 年市

		场份额达到 27%，创收 100 亿美元
皇家苏格兰银行 Royal Bank of Scotland	分析交易数据最大化客户盈利以及支撑各种客户关系管理需求	通过海量数据挖掘支撑金融个性化服务
目标超市 Target	分解消费者行为预测怀孕可能性并据此推送产品推荐	能够比以往多发现 30%以上孕妇
匹兹堡大学医疗中心	出院前预测病人未来 30 天再次住院的可能性	降低治疗风险
伦敦股票交易所	分析数据决定投资方案	约 40%的股票交易由数据应用自行驱动
大陆航空公司	分析航班数据	有效降低航班延误和航线利用率
奥巴马竞选团队	分析选民数据推测哪些选民更容易被竞选活动影响	取得了惊人的程序
惠普 HP	分析全球 35 万名员工的辞职风险	预计收益 3 亿美元
美国国税局	分析纳税人数据发现水手欺诈	在不增加工作人时的前提下提升发现逃税率 25 倍

随着人类社会数字化程度的迅速提升，目前全球数据规模已经达到 44 万亿 GB。数据增长的速度更是惊人，我们可以从图 1-1 中看看当前各大网站一分钟的数据量。读者可以想象一下，在阅读这一页的过程中，全球数据又增加了多少。数据产生的来源和数量增长之快，以至于 2013 年的一份分析报告指出全球数据的 90%是在此前两年中产生的[1]，也就是说每两年产生的数据是此前全部数据的 10 倍，而且我们可以大胆的猜测到本书出版之时，95%甚至更多的数据实在过去三年内产生的。

数据规模是如此之大，种类又是如此之多，以至于一般认为当前我们能够分析的数据只是全部数据的 0.5%。那么我们怎样才能充分利用海量数据，而不是“湮没在数据中却饥渴于无法获得知识（Drowning in Data yet Starving for Knowledge）”呢？答案是显然的，机器学习算法必须借助更强劲[2]的计算硬件和更加灵活的程序设计技术。

[2] 严格讲应该是能效比更高的硬件。

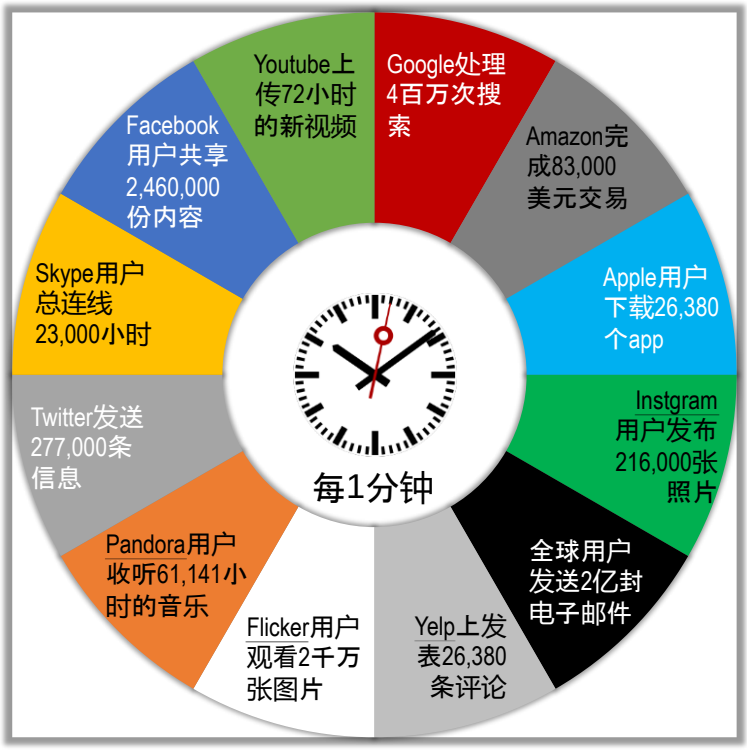

图 1-1 全球 1 分钟内产生的数据

然而，我们手中并没有一种硬件能够同时在上述两项要求上都能表现最佳。图 1-2 是对常见计算平台的比较。在图 1-2 的左侧，是执行顺序程序的 CPU，其编程模式符合人类的思维方式，编程工具完备而成熟，然而性能相对有限。特别是自从 2000 年以后，传统上以增加时钟频率提升 CPU 性能的方法已经遇到瓶颈，继续提高频率提升性能有限，反而带来功耗的大幅度增加。数字信号处理器是对 CPU 进行订制，针对特定应用引入专用指令和硬件从而提高性能的处理器，其编程灵活性有所下降，但是能够提高相应应用的性能。数字信号处理器曾经是高性能的标志，但是随着多核 CPU 的出现，已经逐渐退出高性能计算市场，主要用于嵌入式产品。多核 CPU 是在集成电路工艺的集成能力继续提升而单核性能饱和的产物，通过引入多个并行执行指令的 CPU 内核保证整体性能的增加。多核 CPU 必须使用并行程序才能获得更好的性能，其编程灵活性有所限制。在图 1-2 的右端是专用集成电路，即针对特定应用采用特定算法而设计的硬件平台，完全不具备编程能力，但是性能可以达到极致。在当前市场需求多元化并且高速变化的背景下，缺乏可编程能力是严重的缺陷，因此专用集成电路只有在用量极大的前提下才具有竞争力，越来越多的电子产品使用系统芯片，即集成专用集成电路和嵌入式处理器的芯片。以 FPGA 为代表的可编程硬件比专用集成电路性能

低一个档次，但是具有硬件编程能力，因此也成为一种重要的计算平台。专用处理器也是折衷可编程性和性能的产物，其思想是针对特定应用设计指令集，其中某些指令可以通过专用硬件直接执行，从而在保持一定编程灵活性的基础上改善性能。然而，专用处理器的应用范围比较窄，因此编程工具极为有限、使用人群较小，因而也限制了灵活性。图 1-2 的中央是图形处理器（Graphics Processing Unit，简称 GPU），其前身是为图形渲染应用而设计的专用处理器，但是经过 30 年的发展，随着图形应用的复杂度越来越高、性能要求越来越突出，已经演变为具有高度计算能力和高度可编程能力的计算平台。在各种计算硬件中，GPU 比较完美地折衷了性能和灵活性。注意以上讨论中，我们所说的性能其实指特定制造工艺下单位面积提供的性能，不同制造工艺下的不同类硬件平台的性能错综复杂。由于 GPU 拥有图形渲染市场的支持，能够保证其出货量，因此能够使用最先进的制造工艺并且制造较大的芯片，从而能够提供极高的单片性能，在较低工艺下制造的专用集成电路和 FPGA 反而不容易达到使用最新工艺的 GPU 的性能。从 2006 年开始，NVIDIA 和 AMD 等 GPU 制造商意识到 GPU 可以成为一种与 CPU 互补的通用计算平台，相继退出一系列编程工具，从而极大地开阔了 GPU 的应用。从 2010 年开始，机器学习成为全球化热点，众多企业、科研和政府机构开始在日常工作中大量使用数据挖掘工具，而机器学习算法普遍具有计算密集特点，特别适合 GPU 硬件执行，因此，图形处理器几乎一夜之间成为机器学习最重要的应用平台。

图 1-2 常见计算平台的计算能力和可编程性

1.2 机器学习简史

机器学习是同计算机科学一起诞生的，图灵、冯•诺依曼、赫伯特•西蒙等计算机科学先驱同时也是机器学习理论的教父。本节回顾机器学习理论和技术发展的历史，从中领略人类心智激动人心的前进脚步。

在回顾之前，我们首先梳理相关概念。作为正在高速发展的跨领域学科，数据科学的术语体系比较混乱，特别是『数据挖掘』和『机器学习』两个属于经常被混用。作为数据科学的核心手段，数据挖掘以数据库技术为依托，以机器学习作为核心算法，以统计、分类、预测模型和数据可视化作为典型数据输出方式，借助现代计算机硬件特别是各种并行处理器的强大处理能力以及程序设计技术，从海量数据中精炼出信息、知识直至智慧。图 1-3 是对数据挖掘和其它相关学科的关系。按照米切尔（Tom M. Mitchell）的定义，机器学习是能够针对某种任务从经验学习的计算机程序，在执行该任务时的性能随经验增加而提高[]。 作为数据挖掘的算法和手段，机器学习属于人工智能的一部分，整合了计算机科学、数学（特别是统计学）和神经科学等领域的最新成果，图 1-4 是数据挖掘与其相关领域的关系图。机器学习技术的焦点是预测，即通过对未知概率分布的历史观测数据进行处理，推断由该分布产生的观测数据的未来趋势。早期的机器学习更多遵循传统算法研究的方法，而过去的 20 年中，统计学理论已经在很大程度上改造了机器学习理论和方法，使得机器学习获得了更为严格的理论基础；同时，神经科学也为机器学习注入了新的灵感，通过借鉴人脑神经原理，深度神经网络在近年中取得了巨大的成功。

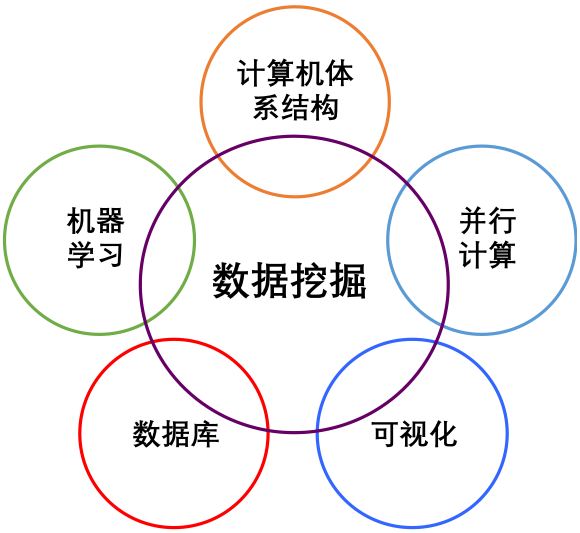

图 1-3 数据挖掘技术要素图

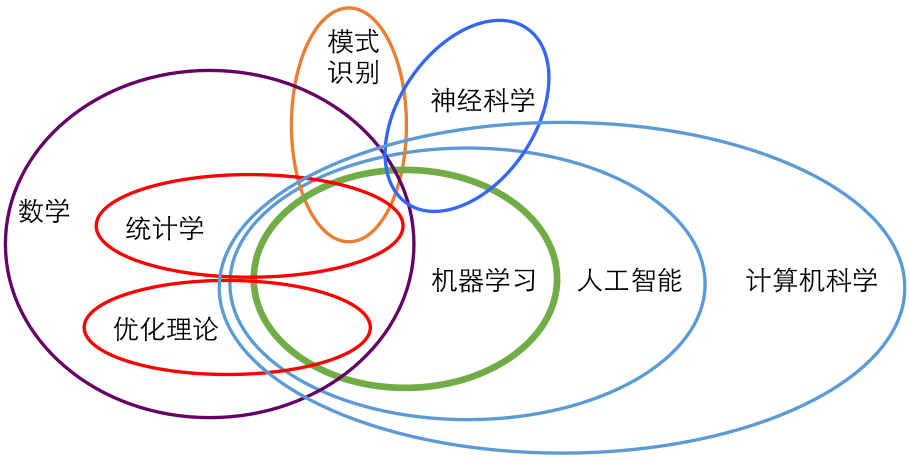

图 1-4. 机器学习及其相关学科关系

人工智能是人类长时间的梦想。为了形成人工智能，很自然地希望机器能够像人类一样，通过一个学习的过程（当然我们希望这个过程尽量缩短）掌握相关知识，并且形成自动推理工具。1950 年，伟大的计算机科学家阿兰•图灵提出了『图灵测验』的概念，宣告人类第一次以严格而形式化的方式定义人工智能的测试准则。图灵测验要求人类测试者和计算机进行交互，如果测试者不能分辨交互对象是人还是计算机，那么就可以认为计算机（至少在某一方面）具有和人类相似的智能。

1952 年，当时正在 IBM 工作的阿瑟•塞缪尔（Arthur Samuel）设计了第一个计算机学习程序。这个程序针对西洋跳棋（checker）进行学习，已经具备了现代机器学习的很多特征，通过使用搜索树（search tree）存储对弈策略，能够通过不断对弈提高自身的能力。由于当时计算机内存极少，因此塞缪尔还设计了α-β减枝（α-β pruning）策略，通过打分机制减少需要完全探索的分支，这个打分机制能够根据对弈盘面判断每一步骤对整体形式的影响。

1956 年，由约翰•麦卡锡（John McCarth）发起，马文•明斯基（Marvin Minsky）、奥利佛•塞尔福里奇（Oliver Selfridge）、克劳德·香农（Claude Shannon）、阿兰•纽厄尔（Allen Newell）、司马贺（Herbert Simon，音译为赫伯特•西蒙，他自己为自己取了中文名字）、雷•所罗门诺夫（Ray Solomonoff）、阿瑟•塞缪尔、内森尼尔•罗切斯特（Nathaniel Rochester）和特兰查德•摩尔（Trenchard More）等学术巨擘[3]共同参加了达特茅斯回忆（Dartmouth Conferences），为未来人工智能研

[3] 其中香农是老一辈的任务，已经发表了他著名的信息论论文。司马贺是百科全

究提出了纲领，认为人工智能(『人工智能』这个提法要到后来 1965 年才能提出）研究应该在 七个方向上开展：1.可编程计算机，2.编程语言，3.神经网络，4.计算复杂性理论，5.机器学习，6.抽象，7.随机性和创见性。这次堪与物理学史上著名的索末菲会议标志着人工智能科学的正式诞生。

1957 年，佛兰克•罗森布拉特（Frank Rosenblatt）根据赫布学习法则（Hebb's Rules）设计了第一个能够被计算机执行的人工神经网，当时叫做感知器（perceptron）。从今天的观点看，感知器是单层神经网，把多个输入值通过经过训练的函数处理后形成输出。罗森布拉特同时构造了最小二乘法和梯度下降法等训练方法，直至今天仍然是深度神经网络训练的基本机制。然而，马文•明斯基和西摩•派珀特（Seymour Papert）详细分析了感知器的局限性，证明感知器不能处理所谓『线性不可分问题』，其中最著名的是异或计算，我们无法用一个线性函数把两变量的异或计算结果区分开。明斯基和派珀特的工作主要是为了说明单层感知器的能力有限、而多层感知器具有更加强大的学习能力。但是，也许异或计算给人的印象过于鲜明，这篇文章带来了人工设计网络的冬天，直至二十世纪末杰佛里•辛顿（Geoffrey Everest Hinton）以极度的坚韧开创性地证明了深度神经网络的强大威力，神经网络的春天才再次降临。

另一方面，基于搜索和逻辑等传统计算机科学技术的机器学习仍然在继续发展。1957 年，贝尔实验室的斯图尔特•劳依德（ Stuart Lloyd）提出 K-means 算法，这是最早的非监督学习算法。1967 年，最近邻（Nearest Neighbor）算法被构造出来，这实际上是一种基本的模式识别算法。1960 年代也诞生了决策树算法（Decision Tree），能够通过学习历史数据形成支撑分类决策的树形结构，1970 年代后期，决策树算法演变为 ID3 算法，后来又改进为 ID4.5 算法。这些算法直到今天仍然被广泛用于各种数据分类问题。

1980 年代是专家系统的时代，学术研究的热点是对各种领域知识进行整理。因此，各种基于规则的专家系统或知识库如雨后春笋般诞生，佛雷德里克•海伊

书式的大师，未来将与纽厄尔一起创立了卡内基•梅隆大学计算机系，还讲一起获得图灵奖，还讲获得诺贝尔经济学奖。明斯基、麦卡锡、纽厄尔都将获得图灵奖。雷•所罗门诺夫未来是算法概率论的发明人，罗切斯特将设计 IBM 701 计算机和世界上第一个汇编程序，塞尔福里奇将被称为『机器感知之父』（他的祖父是伦敦著名的塞尔福里奇百货公司创始人），塞缪尔前文已经介绍，是人工智能游戏的奠基人之一。

斯-罗斯（Frederick Hayes-Roth）在他的专著中把这些系统分为解释、预测、诊断、设计、规划、监控、调试、维修、教学和控制十大类[2]。然而，这些基于知识的技术并未取得深刻的成功。一方面，此时对知识的处理很大程度上是把知识编码为规则，典型形式为“if…else…”的条件判断，而不是真正意义的学习，因此专家系统实际上很难做到“发人所未发”；另一方面，人类知识绝大多数不具备数据的严格性，在实际的工程和商业知识中，总是存在大量的不确定性和不完备性，专家系统只能从编码的知识出发进行处理，对于不能编码的知识则无能为力。

由于基于知识的机器学习理论和技术局限性，1990 年代人类进入了数据驱动学习的时代。这段时期，分类、回归、搜索和关联关系挖掘等各种学习算法飞快走向成熟，核心思想是分析数据从而提取知识。1997 年，人工智能的突破终于到来，IBM“深蓝”计算机第一次在国际象棋比赛中击败了代表人类最高水平的国家象棋大师卡斯帕罗夫。当然应该指出，“深蓝”的突出能力在高速搜索最佳对弈策略，而国际象棋策略相对容易编码，因此“深蓝”的胜利还不能完全等同于机器学习的胜利。

2006 年，机器学习大师杰佛里•辛顿定义了“深度学习”概念，用来包装他和合作者们多年开发的一组基于神经网的学习算法。此时的神经网不仅是多层神经网，而是借鉴人脑神经通路结构形成的“深度”神经网，能够让计算机“看到”并且分辨复杂的图形和文字。实际上，深度神经网最早的成功应用是对支票的识别，美国银行几乎一夜之间全都采用了基于卷积神经网络的自动识别系统。

2010 年以后，以深度神经网络为代表的深度学习理论和技术取得了惊人的成功。2010 年微软推出的 Kinect 系统能够实时（允许相关内容以每秒 30 帧速率刷新）跟踪 20 个人体特征，根据机器学习形成的模型识别人体行为，从而使得计算机游戏（或其它应用）和人通过运动和姿态进行实时交互。2011 年，IBM 开发的深度问题-答案（DeepQA）系统集成于 Waston 计算机，在著名的问答类智力游戏“危险边缘”中打败了人类参赛者。谷歌 2011 年开发了谷歌大脑系统，能够对图片进行分析，从中识别物体并进行分类，基本达到猫的识别和分类水平，同期谷歌 Xlab 开发了能够自动浏览 YouTube 视频并从中发现含有猫的内容。2010 年开始的大规模视觉识别竞赛（Large Scale Visual Recognition Challenge）更是极

大促进了以图像识别为目的的机器学习技术，到 2014 年，深度神经网识别人脸的精确的已经超越人类水平。如果使用人脸识别作为图灵测验的话，计算机已经需要假装认不出某些人脸了。机器学习最新的成就，是谷歌的 AlphaGo 在 2016 年击败世界围棋冠军李世乭，攻克了普遍认为是可能性最多、局势最难判断因而最困难的对弈游戏。不同于“深蓝”计算机，AlphaGo 的的确确在进行学习，不仅使用深度神经网络学习某一盘面下最好的下法，也通过增强式学习某一下法对全局的影响。

机器学习技术是如此之成功，以至于有识之士们已经开始担心人工智能技术被用于智能化武器，从而对人类本身造成巨大威胁。2015 年，包括斯蒂芬•霍金（Stephen Hawking）、埃隆 •马斯克（Elon Musk）和斯蒂夫 •沃兹尼亚克（Stephen Wozniak）等在内的 3000 多位科学家和工程师签名发表了公开信号召国际社会采取措施，防止智能武器在不受人干涉的情况下选择目标和发起攻击。

虽然机器学习取得了惊人的成就，我们还是要说这仅仅是一个伟大时代的开端。伟大的十八世纪物理学家和数学家拉普拉斯认为只要我们有足够的初始条件和物理知识，那么依靠牛顿力学就足以计算出宇宙的过去、现在和未来。著名科技杂志《连线》（Wired）在 2008 年发表当时主编克里斯•安德森（Chris Anderson）的一篇文章，则认为“数量庞大的数据会使人们不再需要理论，甚至不再需要科学的方法”。安德森的意思是说，随着数据越来越多，机器学习手段越来越发达，我们可以通过研究历史数据直接揭示事物之间的因果关系，而不再需要专门的理论和方法。这个结论可以说是拉普拉斯决定论的“机器学习版本”，虽然可能过于宏伟，但是机器学习的确可以作为一种科学发现的重要工具。近年来，因果性挖掘（Causal Analysis）、隐变量分析（Latent Variable Modeling）和贝叶斯学习（Bayesian Learning）理论吸引了大量研究者，这些理论必将成为自动知识发现的利器。

1.3 机器学习改变世界：基于 GPU 的机器学习实例

机器学习技术正在不断取得举世瞩目的成就，这一节会介绍三个机器学习的成功案例，让大家体会机器学习技术怎样解决极度挑战性的实际问题。

1.3.1 基于深度神经网络的视觉识别

在过去的五年中，深度学习成为最为热门的名词，机器学习的一切概念只要加上“深度”两个字仿佛就拥有了魔力。深度学习现在泛指一切基于深度神经网

络的机器学习方法和应用。严格说来，深度神经网络的本质还是人工神经网络，只不过其规模极大（2010 年获得 ImageNet 大规模视觉识别竞赛最高奖的卷积神经网络拥有 65 万个神经元节点和六千万个参数），层数极多（例如谷歌的 GoogLeNet 有 22 层），而且训练数据集合规模惊人（ImageNet 数据集由超过千万张图片）。类似于人脑中数以千亿计神经元的协同就产生了认知、智慧和感情一样，深度神经网络在高效学习算法的帮助下，确实体现出惊人的威力。

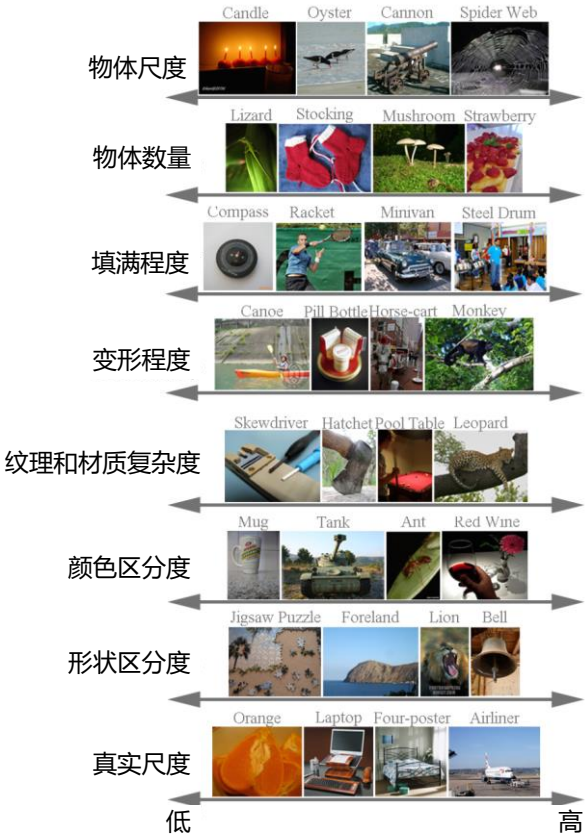

图 1-5. ImageNet 大规模视觉识别竞赛（ImageNet Large Scale Visual Recognition Challenge）的图片

目前，深度学习应用以基于深度卷积神经网络的计算机视觉识别最为成功。而这个成功在相当程度上依靠 ImageNet 大规模视觉识别竞赛（ImageNet Large Scale Visual Recognition Challenge）[5]，该竞赛提供了丰富的训练数据样本，并且精细设计了多个竞赛科目，自从 2010 年开始举办以来，已经成为深度学习的试金石，极大带动了深度神经网络技术。ImageNet 数据集拥有大量张图片，这些图片通过众包方式由人工标注，标注体现为两种形式：1.图片级标注以“有”或“没有”方式说明图片中是否包含某一类对象；2.对象级标注为图片中的物体定义一个“紧”包围盒（即尽量小但是能够包围对象的长方形）并且说明对象在图片

中的核心信息，例如“图片中有一个中心在（20,25）、宽度为 50 像素、高度为 30 像素的螺丝刀”[5]。

2010 年第一届的时候，ImageNet 大规模视觉识别竞赛只有一个项目即图像分类（Image Classification），后续又开设了物体定位（Single-Object Localization）和检测（Object Detection）两个项目。竞赛形式都是由 ImageNet 提供公开的训练数据集合和不公开的测试数据集合，参赛团队提交程序时对测试数据集合进行识别，以识别精度决定最终名次。ImageNet 的图片分类数据集包含 14,197,122 张图片，内容和复杂度极为多样。图 1-5 描绘了数据集的多个复杂度维度，例如图中物体的尺度、物体数量、填满程度、物体变形程度、纹理复杂度等等。图 1-6 给出了容易分类和难于分类的图片的例子。分类竞赛的目标是对图像进行判别，把每一张图片中的物体分到 1000 个类别（Category）。如果使用卷积神经网进行识别的话（目前成绩最好的参赛程序几乎都使用深度卷积神经网），对每一张测试图片，神经网都会输出该图形属于每一类别的概率（所有 1000 个概率之和为 1）。竞赛采用 top-1 分数和 top-5 分数，前者指参赛程序为测试图片报告概率最高的类别，如果与实现人工标注结果吻合，则表示分类正确，后者则是参赛程序为测试图片报告概率最高的五个类别，只要其中一个与人工标注结果吻合则为正确[4]。物体定位竞赛使用与分类竞赛同样的数据集，同样针对 1000 个目标类别，机器学习程序需要列出图片中包含的每一类别的一个物体，并为识别出的物体给出一个相应的“紧”包围盒。只有在物体清单包含人工标注的物体，并且包围盒大小合适时，才能算是正确识别。图 1.7 列出了一些定位难度很高的图片，例如其中从左向右第四张，人类的标注结果是吊车（Crane），但是吊车只有部分出现在图片中，而且不在中心部分，再例如第六张香蕉（Banana），图中大量香蕉堆积在一起呈现出多种奇怪的形状。在目标检测竞赛中，机器学习程序列出图片中所有属于上述 1000 个类别的物体，并为每一物体定义一个“紧”包围盒，识别结果由识别数量和准确性综合衡量。

[4] Top-5 分数是很有意义的，因为人在辨识图片时也经常会任务有多种潜在可能性。

（a） 容易完成分类的图片

（b） 很难完成分类的图片

图 1-6 分类竞赛的图片[5]

图 1-7. ImageNet 数据集中难以识别的图像[5]

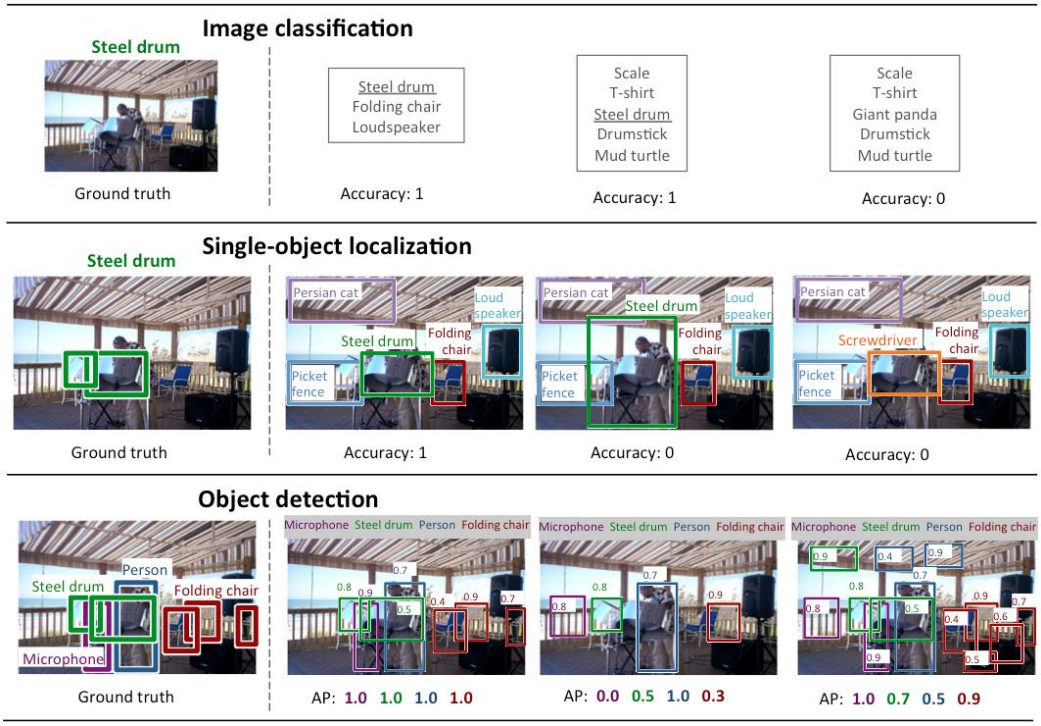

图 1-8. ImageNet 大规模视觉识别竞赛[5]

图 1-8 是对以上三个问题的概括，其中从上到下的三行分别对应三个问题，最左一列是相应的“基础事实”（Ground Truth），即人工识别后的正确结果（或者说人类观察者的平均识别结果），右面三列分别列出三种不同的识别结果及其相应的精度分数。

从 2010 年起，ImageNet 大规模视觉识别竞赛见证了深度卷积神经网络的成功，历届最好结果都由该网络实现。深度卷积神经网络我们将在第 14 章专门介绍，这里我们只扼要介绍其结构和原理。深度卷积神经网络由燕乐存（Yann LeCun）在 1998 年提出[6]，当时被称作 LeNet 5，应用目标为手写数字和字母识别。初试啼声的 LeNet 5 已经能够在 MNIST 数据集上达到 99.2%的正确识别率。

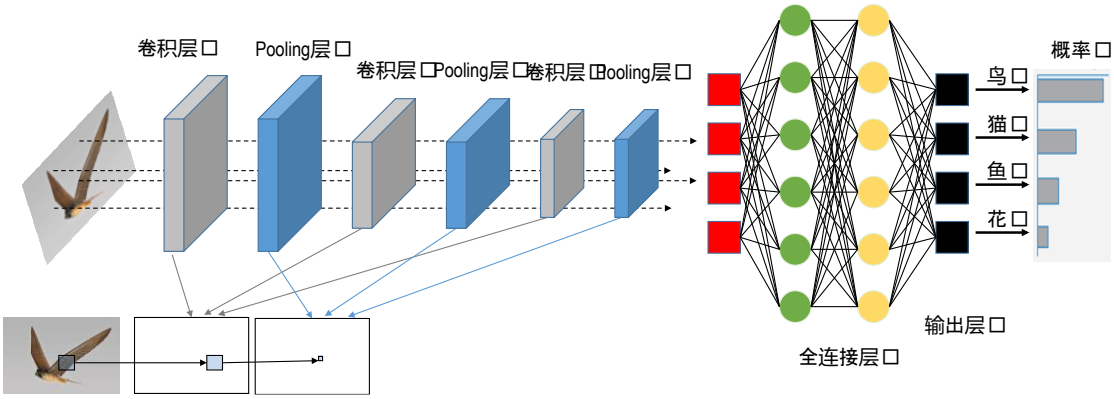

图 1-9. 深度卷积神经网络结构示意图

图 1-9 是深度卷积神经网络的典型结构。类似于人脑的视神经网络拓扑结构，深度卷积神经网络一般包含三种神经层，即卷积层（Convolutional Layer）、池化层[5]（Pooling Layer）和全连接层，其中每种均可以由多层网络实现。卷积神经网络的输入为以像素颜色和亮度表征的图片，输出为图片的分类概率或其它目标。图 1-9 的输入为一张飞鸟的图片，经过交错的三层卷积层和三层池化层提取特征后，送入全连接层，该层又包含本身的输入层、两个隐含层和输出层，最后输出 4 类潜在目标的概率，其中“鸟”的概率最高，因此将输入图片归类为“鸟”。接下来，我们看看这三种神经层分别完成什么功能。

- **卷积层**：卷积层网络严格来说不是真正意义的人工神经网络，而是用于从图像中提取特征，即对输入图片一定区域若干像素点的特定度量。具体度量方式由卷积核函数（Convolutional Kernel）决定，可以理解为图像滤波器，即从像素中过滤掉不关心的信息，并且加强关注的信息。在 Photoshop 图像处理中，这

[5] Pooling Layer 尚没有统一的中文译名，本书姑且翻译为“池化层”。

些卷积核通常被用于产生特殊效果。图 1-10 是完成“浮雕”特效的卷积核函数，该卷积核体现为一个 3×3 矩阵，包含 9 个系数，这些系数与输入图片中每块 3×3 区域的相应元素进行乘法运算，所有乘积的和输出为计算结果。对卷积神经网络而言，这个运算的效果是从图像中提取颜色变化强烈的区域。注意卷积运算的特点是对整张图片反复使用同一函数进行特征提取，因此可以抽取到分散在不同位置的相同特征。卷积层的输出结果为与输入图片相同维度的特征图（Feature Map）。

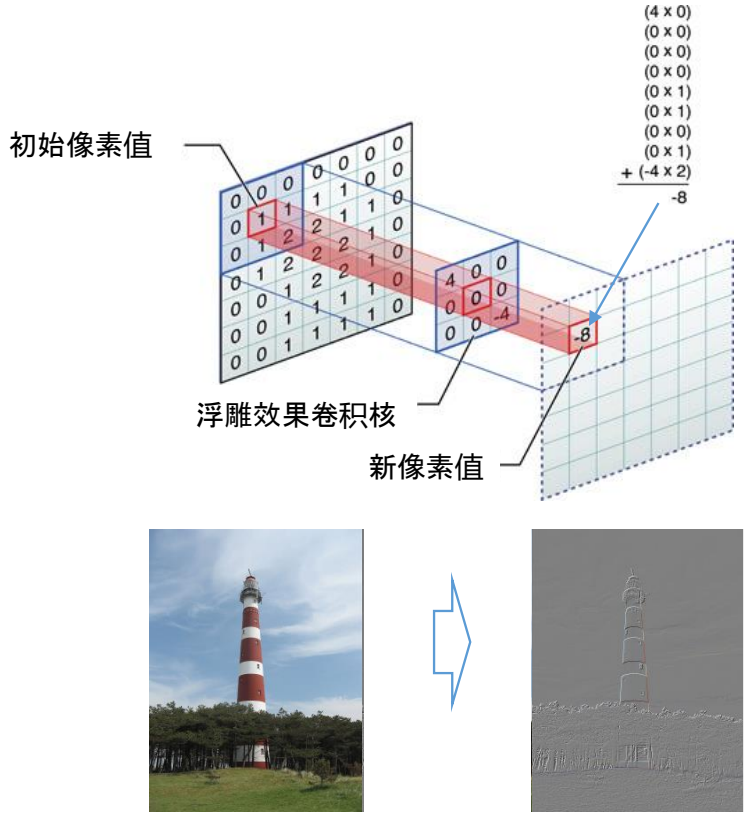

图 1-10. 实现“浮雕”效果的卷积核

- **池化层**：通常每个卷积神经层后面都会有一个池化层（Pooling Layer），该层实际上完成降采样（Subsampling）操作，即对相邻若干像素进行特定运算，过滤掉局部不重要的特征，把输入特征图的维度降低，有助于减少神经网络参数的数量。最常见的池化操作有最大池化（Max-Pooling）和平均池化（Average pooling），即计算特征图相邻区域元素的最大值和平均值。

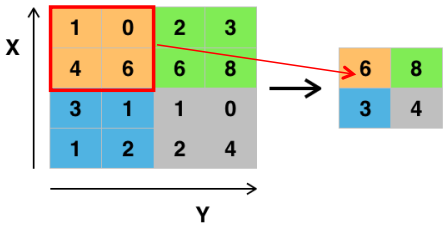

图 1-11. 最大池化运算

- **全连接层**：卷积神经网最后是若干全连接层，严格说来，这才是真正的神经网。由于每一层网络中的每一个神经元均与下一层网络的每一个神经元连接，因此被称作全连接层。第一层全连接层接受池化层的输出，然后送往若干隐藏层，直至最后的输出层，这里输出层每个神经元一般对应一定的分类结果。神经元之间的每一连接均有一个相应的权重，在训练过程中使用后向传播算法（Backpropagation）对权重值进行迭代优化。

在过去五年中，深度卷积神经网的性能也得到了惊人的提升。图 1-12 描述深度卷积神经网在历届 ImageNet 大规模视觉识别竞赛在分类、定位和检测上达到的精度。以图片分类来看，分类错误率从 2010 年的 28%下降至 2015 年的 3.6%，已经低于人类的平均错误率。与此同时，当前的卷积神经网呈现出日益复杂的趋势，其层数经常超过 20 层，参数数量达到 1 亿个左右，因此，深度卷积神经网的训练普遍采用计算机集群进行，而图形处理器更是成为计算平台的标准配置。

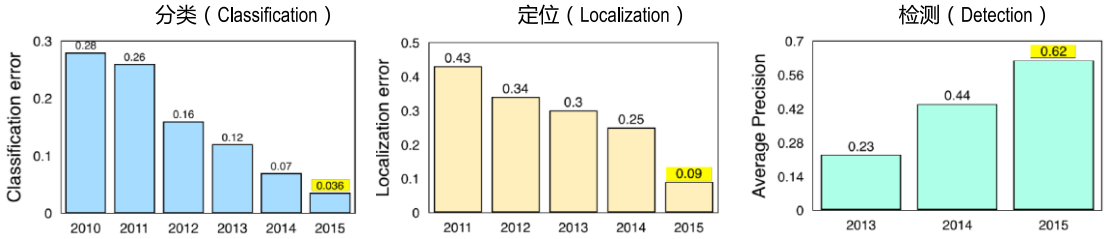

图 1-12. ImageNet 大规模视觉识别竞赛结果

1.3.2 AlphaGO

在过去的几年中，谷歌 DeepMind 团队以一系列重量级工作引起了世人关注。在收购 DeepMind 之前，谷歌本来已经在深度卷积神经网络上形成了深厚积累。DeepMind 创造性地把深度卷积网络移植到增强式学习（Reinforcement Learning）框架，用深度学习改造增强式学习方法，从而以天才的一笔而创造了新的奇迹。

DeepMind 的第一项重要工作是让人工智能学会在雅达利 2600（Atari 2600）游戏机上打游戏。雅达利 2600 游戏机是雅达利公司 1977 年推出的掌上游戏机，提供了打砖块、小蜜蜂、吃豆子、大金刚等多款经典游戏，是第二代电子游戏的

代表主机，相信很多资深游戏玩家还有深刻印象。雅达利 2600 的屏幕分辨率为 210×160 像素，每个像素支持 128 种颜色值。游戏控制包括一个操纵杆（上、下、左、右、上左、上右、下左、下右 8 个方向的运动）和一个游戏杆按钮（该按钮可以单独使用，也可以结合操纵杆使用）。图 1-13 是三个典型雅达利 2600 游戏的屏幕。

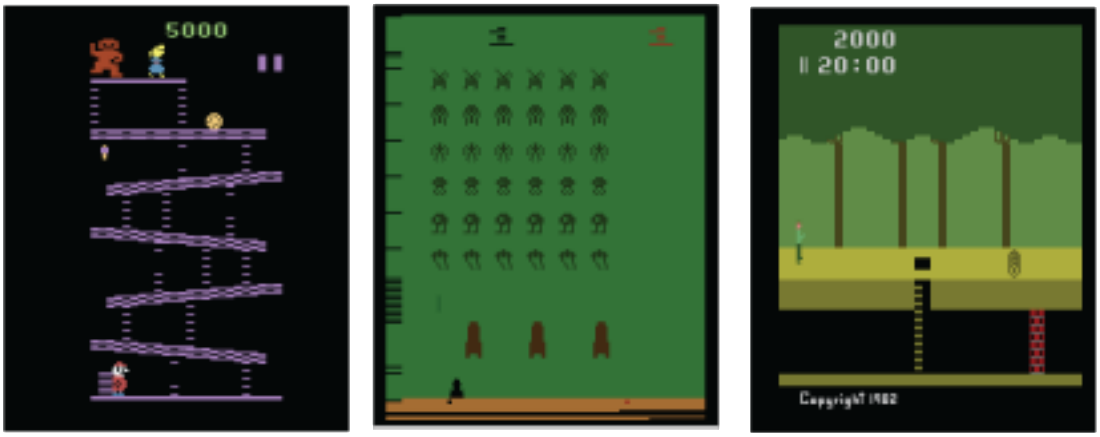

图 1-13 雅达利 2600 游戏（从左至右为大金刚（Dongkey Kong）、太空入侵者（Space Invaders）和陷阱（Pitfall））

在 DeepMind 工作之前，深度卷积神经网已经表现出惊人的威力。然而，一般来说，传统卷积神经网能够建立从此时的输入到此时输出的关系，即关注“当下”的判断；但是，在雅达利 2600 游戏中打到高分，需要找到一个连续操作序列，其中每一步操作对整体游戏的效果一般需要在游戏结束时才能获悉，也就是说需要建立现在的一步操作和一整局游戏分数的关系。有基础的读者已经可以看到，解决雅达利 2600 游戏的标准机器学习方法是增强式学习（Rinforcement Learning，也常常叫做强化学习），例如 Q-Learning 算法。Q-Learning 算法的思想是让机器自动尝试选择某个动作从一个状态跳转到下一个状态，直至目标状态（例如游戏结束或积累一定时间），然后判断每一次动作对总体目标的影响（即回报）。增强式学习的需要缺点是在稳定性较差，特别在动作和整体目标的回报关系具有非线性时容易造成不收敛的问题。

DeepMind 工作的核心创意是改造深度卷积神经网，使之能够进行强化学习。图 1-14 是该工作使用的卷积神经网的示意图。与 1.3.1 中图 1-9 比较，可以看出这里的神经网并没有本质的变化，仍然遵循卷积层+全连接层的基本样式。我们的目标是把游戏状态序列（可以理解为游戏场面的若干连续变化）和应该采取的

控制操作联系起来。这里状态指游戏整体态势，包括游戏场景、计算机操作物体的分布和动作输出、以及由玩家操作的物体状态。经过训练过程形成的深度神经网能够根据当前的状态序列（即当前状态和之前的若干状态）输出应该采取的下一步游戏操作，其中的操作是玩家控制物体的一次动作。由于卷积神经网是为处理图像而设计的，因此需要对游戏建立适当的模型，从而使得卷积神经网能够处理。

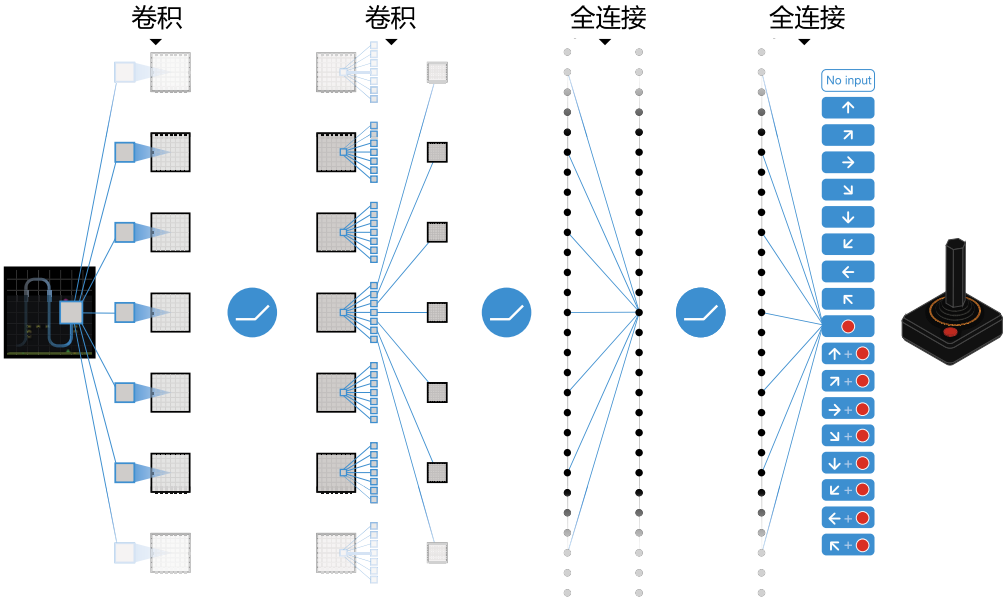

图 1-14. DeepMind 在雅达利 2600 游戏问题上使用的卷积神经网示意图（根据[6]改画）

DeepMind 的办法是以游戏输出图像序列作为神经网的输入。为了降低复杂度，采用 84×84 的降分辨率图像，每 4 帧为一个序列。这些序列被三个卷积层依次处理，分别提取相应特征。一般说来，增强式学习的结果应该是一个 Q 函数，描述操作和整体回报的映射。在图 1-14 的卷积神经网中，DeepMind 没有采用单一的 Q 函数值作为输出，而是输出每个操作对应的 Q 函数值，从而能够通过一次前向传播就找到输出结果。

训练过程被组织为若干次游戏过程。每一次过程由一系列离散时间步骤组成。在每一个实际步骤上，训练算法针对当前场景随机选取游戏操作或者从存储下来的已有操作序列中选择一个操作，观察操作执行后游戏局势的演化结果，对此次选取的操作进行评价，并根据评价结果对神经网参数进行梯度优化。图 1-15 是训练过程中得分变化的例子。在打砖块游戏中，刚开始时不同操作没有本质区别，但是一旦砖块打破最上一层砖块并在上面多次反弹，则得分会显著增加，因此深

度神经网络就会倾向于产生能够达到上层砖块的操作。DeepMind 使用单一的神经网对 49 个雅达利 2600 游戏进行训练，在不需要任何游戏知识的情况下，在一半以上的游戏上超过了人类玩家（游戏高手）。

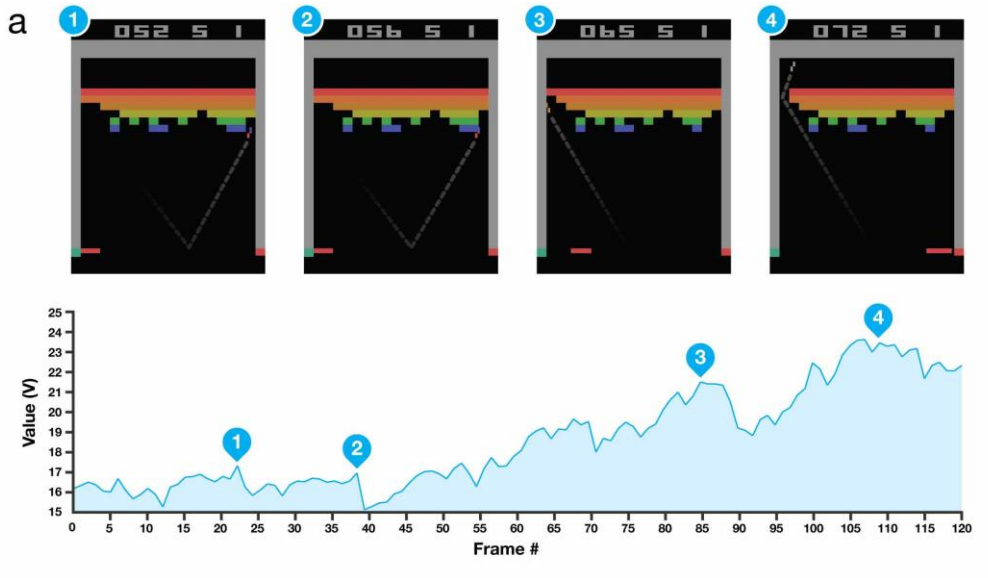

图 1-15. 有效操作价值评估（根据[6]改画）

DeepMind 再接再厉，下一个目标：冲击围棋游戏！围棋的复杂度在于两个方面。首先，弈子可能性多得惊人。一般说来，类似象棋和围棋的对弈游戏可以有 b^d 种对弈序列（即双方交替走子直至游戏结束的完整序列），其中 b 为游戏的搜索广度，即每一步有多少种可能性，对围棋来说，这个可能性的平均值为 250 [Human][6]，d 为游戏深度，即游戏结束前走过的步数，通过棋谱统计，这个数字大约是 150，因此总复杂度是惊人的 250^{150}。作为比较，国际象棋的搜索广度为 35，深度为 80。第二，总体局势难以判断。围棋棋子之间不能直接比较威力（例如国际象棋中王后显然比小兵厉害得多），而输赢要依靠全部棋子控制的区域大小决定。因此，围棋高手们需要依赖过人的天分和多年的训练才能形成判断局势的准确直觉，并且建立一手棋和最终结果之间的关系。从这个意义上讲，学棋的确是个复杂的神经网络训练问题，只不过人依靠的是脉冲神经网以及各种基于电化学反应的学习过程。

在成为围棋大师的漫漫征途上，棋手需要解决几个问题：首先，如果只考虑当前盘面，最好的应对招数是什么？其次，谁都知道与高手对弈可以快速提高自

[6] 围棋棋盘纵横各有 19 条线，因此共有 19x19 个交叉点，第一手的可能性为 361 种，第二手有 360 种，第三手 359 种，以此类推。一般来说不需要下到把棋盘填满，因此一个合理的估计是平均每手 250 种可能性。

身棋力，那么怎样获得高质量的陪练对手？第三，在可能的下法中，哪一步最可能带来胜利？ AlphaGo 提出了一套精妙的解决方案，整合了深度学习、增强式学习和基于价值网络的蒙特卡洛搜索树（Monte Carlo Search Tree）三大关键技术，通过离线学习+在线对弈的方式来解决以上问题。

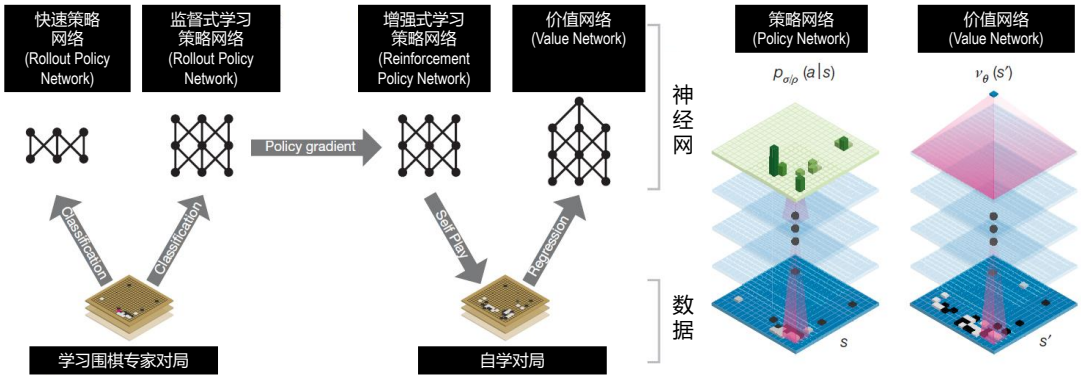

图 1-16. AlphaGo 训练过程示意图（根据[7]改画）

图 1-16 是 AlphaGo 团队在《自然》杂志发表的论文[7]中提出的 AlphaGo 训练过程示意图。首先，AlphaGo 使用专业棋手的棋谱作为输入，训练由深度神经网构造的策略网络，该网络以当前盘面作为输入，输出下一步落在所有棋盘交叉点上的概率。表征棋盘局面看似复杂，实际上可以很简单地实现：AlphaGo 使用棋盘照片作为输入，把盘面问题转化为图形模式识别问题。这一步骤实际上训练两个策略网络，一个速度快、精度低，另一个精度高、速度低。由于以上网络的训练是一个监督式学习的过程，所以这个网络被称为监督式学习策略网络。经过 3 千万多个盘面的训练后，该策略网络能够以 57%的精度根据盘面预测下一手落子。当然，优化的当前落子和最终的胜利是两码事，高手们会走出损失当前利益而优化最终胜利概率的妙手，怎样达到最终的优化，则要依赖增强式学习。第二，AlphaGo 训练一个增强式学习策略网络，该网络以随机结果作为起点，与前面训练的策略网络对弈，通过对弈结果不断修改网络结果，该网络以输入序列为输入，同样输出下一步落在所有棋盘交叉点上的概率。第三步的训练需要解决最困难的问题，即根据当前盘面判断最终结果。AlphaGo 首先利用监督式学习策略网络产生一个随机步数的棋局，以该局面作为神经网输入，然后使用增强式学习策略网络进行自我对弈，以胜负结果作为标签训练一个价值网络，该网络反应的针对当前盘面情况下一手棋的潜在收益。

经过训练，我们现在有两个监督式学习策略网络和一个价值网络。AlphaGo

用蒙特卡洛搜索树实现在线对弈。蒙特卡洛搜索树在本质上就是决策树，也就是说每一手棋都会产生若干可能的分支。为了避免盲目的遍历式搜索，蒙特卡洛搜索树利用有针对性的采样方法减少不必要的搜索空间。AlphaGo 针对具体盘面形势，使用快速监督式学习策略网络和价值网络输出的加权和决定应该向哪个方向进行深度搜索（即探索下一步棋的可能结果），最终选择优化的落子决策。

以上的训练和对弈过程计算需求极高。AlphaGo 团队设计了精巧的并行计算引擎，单机版使用 48 个 CPU 和 8 个 GPU 进行计算，而分布式的计算引擎则使用 1202 个 CPU 和 176 个 GPU。

AlphaGo 的战绩大家都已经熟悉，2015 年 AlphaGo 打败职业围棋选手，2016 年 3 月在五番棋比赛中以 4:1 击败围棋世界冠军李世乭，从而被韩国棋院授予名誉职业九段，2016 年 7 月 AlphaGo 在围棋高手中的积分已经名列世界第一。

AlphaGo 的成就应该从几个方面来看。首先，AlphaGo 并没有从围棋知识本身出发进行训练，也就是说没有直接学习先验领域知识，而是通过图像方式直接判断盘面，就能够取得如此成功，的确是深度机器学习有效性的最强证据。很多评论认为 AlphaGo 并不会思维，恐怕是对人类思维过于自负了。思维可能也只不过是大脑神经网中大量脉冲的组合和集成，事实上，人类棋手的学习过程也是让自己的脉冲神经网能够形成有效的围棋直觉，恐怕在这个层次上与 AlphaGo 并没有本质的区别。第二，人类也不必妄自菲薄。我们大脑的计算能力远不如 AlphaGo 的计算机，能够使用的功耗在 20 瓦以下，只有一个 GPU 的十分之一。顶尖棋手毕生看过的盘面，也远远没有达到 3000 万盘之多。AlphaGo 还应该继续向人类学习，特别是基于小样本甚至单样本的训练。比如说，人类棋手从大师的著名对局（例如吴清源 1933 年对战秀哉名人、推出新布局的第一盘）中可以体会无穷，而单纯的一盘的深度学习引擎的意义极为有限。第三，即使只看围棋，AlphaGo 也并非无懈可击。虽然 AlphaGo 从 3000 万个盘面学习，然而这与围棋 250^{150} 的可能性来比实在只是沧海一粟而已。同时，卷积神经网机制决定其更有效于局部模式的提取[7]，而蒙特卡洛搜索树的搜索时间严重依赖于搜索空间的大小，因此如果在对弈中尽量走出对多块棋子有影响的招法，则会严重影响 AlphaGo 决策的准确性。图 1-17 中标有三角形的白子是李世乭在与 AlphaGo 第

[7]因此，在精细的局部搏杀上，人类很可能从此无法有效抗衡 AlphaGo。

4 局对弈中第 78 步走出的妙手。这一步之前，人们普遍认为局势对 AlphaGo 有利或者基本相当。第 78 手对四片白棋都有深刻的潜在影响，AlphaGo 的价值网络未能充分判断这一步棋的价值，最终导致失败。因此，AlphaGo 的成功，也会促使人类棋手设计新的对局思想，使用更为华丽、更具有大局观的对弈策略，为古老的围棋游戏注入全新的活力。

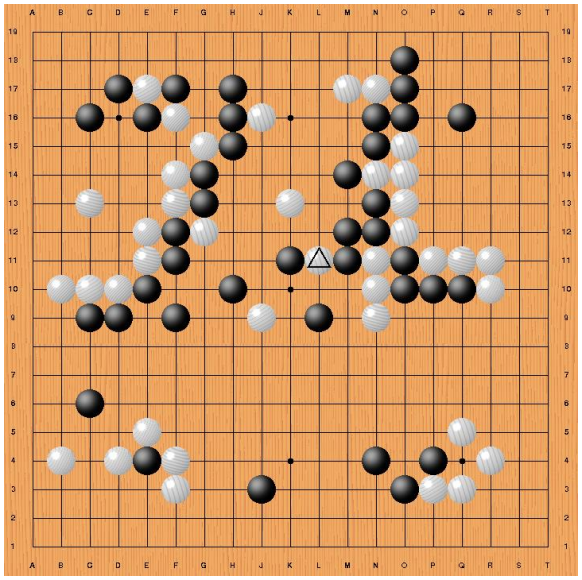

图 1-17. 李世乭在与 AlphaGo 第 4 局对弈中走出的妙手

1.3.3 IBM Waston

前面两个小节介绍的是机器学习图像和游戏的成就，还没有直接处理人类知识。2010 年，IBM “沃森”系统在人类知识竞赛中同样战胜了人类对手。IBM 选择的突破口是美国著名的电视智力竞赛节目《危险边缘》（Jeopardy!）。《危险边缘》从 1984 年开始播出，其形式如图 1-18 所示。每次竞赛有三名参赛者，以抢答形式参赛。游戏直播现场的大屏幕被分为 6×5 的网格，每列对应一个主题（如历史、科学和政治等），每行对应一定的奖金数量。每次选中一个网格后，其中的文字（英语）显示出来，节目主持人念完内容后，参赛者可以开始抢答。这里的题目是最为独特的部分，大屏幕的网格显示的内容是题目的线索，而不是问题。参赛者看到线索后，按下抢答器后要回答出相应的问题。举例来说，大屏幕上显示的线索可以是：“他曾经被 12 道金牌召回，后来被秦桧陷害而死”，那么正确的答案是“岳飞是谁”。如果抢答正确，则相应参赛者奖金增加，否则要扣除一定的奖金。显然，《危险边缘》这种独特的竞赛形式对计算机来说是相等困难的。试想，提问题说答案的形式对计算机是很容易的，只需要进行一次搜索即可，甚至不需要真正理解问题；而目前的形式要求计算机不仅能够理解线索，还要找到

线索之间的内在关系，从而找到答案。同时，在不能完全确信答案正确性的时候，计算机需要判断是冒险抢答还是保守求稳，从而最大化最后的奖金总额。

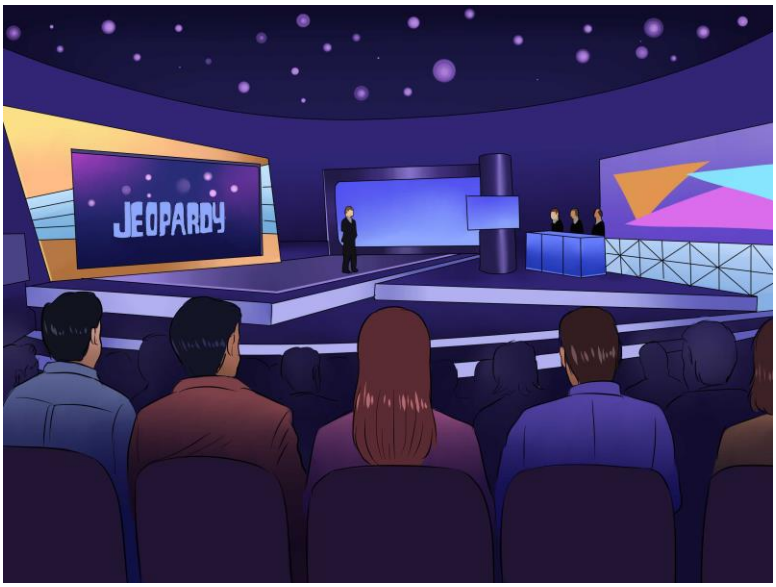

图 1-18. 《危险边缘》节目示意图

为了证明机器学习能够掌握职业水准的问题-回答能力并且能够基于该能力进行关键决策，IBM 成立了 DeepQA 团队，用三年时间研发了“沃森”计算机系统。该系统由 90 台 IBM 服务器组成，拥有 360 个 Power 7 系列处理器（由 45nm 工艺制造，每个处理器拥有 8 个内核、支持 32 个线程，主频最高可达 4.1GHz），存储容量 15TB，体积大致相当于 10 台冰箱那么大的计算机系统。“沃森”存储了大量图书、新闻和电影剧本资料、辞海、文选和《世界图书百科全书》(World Book Encyclopedia)等数百万份资料，全部资料长达 2 亿页。在参加《危险边缘》竞赛时，题面的问题线索以文字流的方式送给“沃森”，没有使用语音识别。参赛过程中，“沃森”没有上网，和人类一样只依靠自身的知识库。

图 1-19. IBM“沃森”超级计算机

“沃森”在接受问题后，首先使用自然语言处理技术对问题进行语法语义分析，从中提取出关键词和核心语义。语义分析需要识别各种微妙的语言结构，例如讽刺、谜语、诗词以及特定文化现象。接下来，“沃森”把问题分解进行大规模并行计算。这里的分解包含若干层次，既有多种解题思路或角度的分解，也有基于同一思路使用不同算法的分解，还有把一套思路分解若干步骤的分解。每一个并行任务中，“沃森”根据关键词和语义从其知识库中查找线索并提取相关证据，由此产生针对答案的假设，然后对假设-证据组合进行评分。多个并行任务的答案汇总后，“沃森”再一次对解答的置信度进行评估，决定是否抢答。“沃森”能够象人类一样跳过自身不擅长的题目，甚至可以模仿开玩笑。图 1-20 是“沃森”计算机上运行的 DeepQA 深度问答系统的体系结构框图。

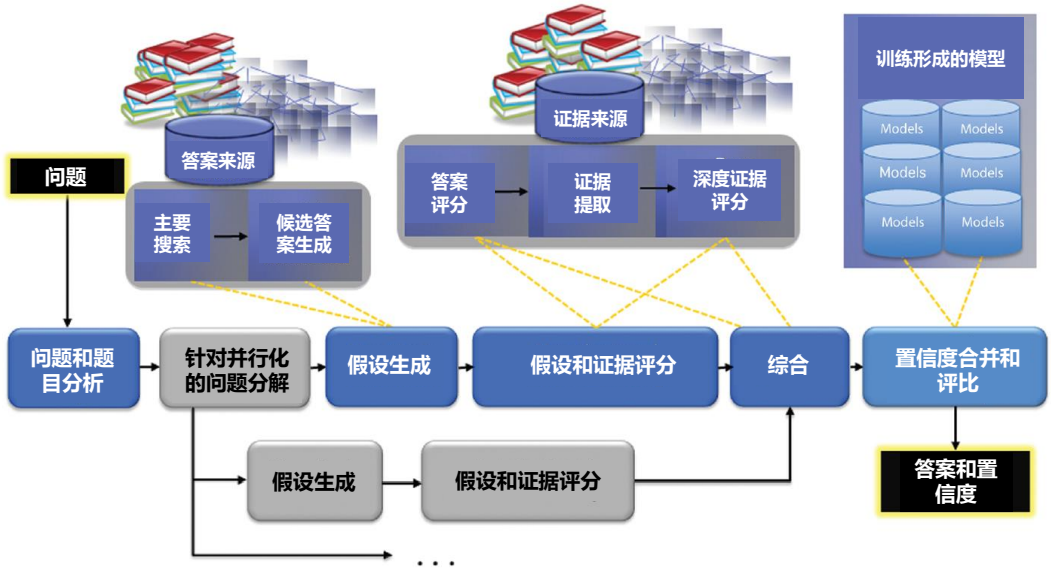

图 1-20. DeepQA 顶层架构图（根据[11]改画）

IBM“沃森”系统于 2011 年 2 月 14 日至 16 日与《危险边缘》历史上两位最成功的选手肯·詹宁斯（Ken Jennings，最长连胜记录的保持者，2004 年连续获得 74 场的胜利，共赢得 2,520,700 美元）和布拉德·鲁特（Brad Rutter，2005 年终极冠军）展开对决，图 1-22 是比赛场景的照片。最后成绩以比赛过程获得的奖金金额决定，实际冠军奖金为 100 万美元，亚军为 30 万美元，季军为 20 万美元。比赛过程略有波折：第一天，“沃森”与两位人类选手胜负难分，最终分别取得 5000 美元、5000 美元和 2000 美元的成绩；第二天，“沃森”开始发力，以 35734

美元的成绩遥遥领先，而詹宁斯和拉特分别只获得 4800 美元及 10400 美元；第三天，“沃森”势如破竹，以 41413 美元的分数击败对手，使得两位人类选手仅获得 19200 美元和 11200 美元。我们可以进一步看看“沃森”回答的一些题目：

1.“There are about 50 species of the hedgehog type of this plant, so named for its spiny fruit.”（这种形似刺猬的植物有 50 个左右品种，根据其多针的果实命名），“沃森”首先抢答，正确地猜到答案应该是“cactus”（仙人掌），显然“沃森”在这道题目的优势在于其速度；

2. “Wanted for killing Sir Danvers Carew; appearance--pale & dwarfish; seems to have a split personality.”（意图谋杀丹佛斯·卡鲁爵士，外表苍白而侏儒化，似乎有分裂人格），这道题目对人来说很容易，看过《化身博士》的参赛者立刻指导答案，“沃森”则需要确定这个内容在什么文本里面以及描述的是哪个任务，它也的确正确推导出上面描述的是“Hyde”（海德）；

3.“It was the anatomical oddity of US gymnist George Eyser who won a gold medal on the parallel bars in 1904. ”（这是生理解剖学的奇迹，美国体操运动员乔治•艾瑟尔在 1904 年赢得了一枚双杠金牌），这道题目相当困难, 乔治•艾瑟尔的奇迹在于他只有一条腿，对于“沃森”来说，知道乔治•艾瑟尔缺一条腿不难，但是推理出“缺一条腿还能得金牌是奇迹”就很难，此时“沃森”需要理解什么可以称之为奇迹，遗憾的是“沃森”给出的答案是“腿”，然而正确答案是“缺一条腿”；

4.“Its largest airport is named for a World War II hero; its second largest, for a World War II battle.”（该城市最大机场以一位二战英雄命名，第二大机场以二战的一场战役命名），这道题目必须考虑题目范畴“美国城市”，而且需要把“二战”、“城市”和“机场”等概念放在一起考虑，“沃森”错误的回答多伦多（正确答案应为芝加哥），显然是没有考虑题目范畴。

图 1-22. “沃森”与两位人类对手

在取得《危险边缘》的胜利后，IBM 为“沃森”规划了 4 条商业化道路，分别是医疗、金融、呼叫中心和政府公共事业。比如，“沃森”已经在美国克利夫兰医学中心找到了一份工作，参与医生培训工作，并与克利夫兰医学中心的临床医生和师生一起工作不断在医学领域的理解和分析能力。此外，“沃森”也为美国最大的百货公司——梅西百货处理顾客问题，为著名的“芝麻街”节目设计节目等。根据著名的市场调查公司 IDC 的报告，到 2018 年，以“沃森”为代表的 IBM 数据服务将贡献 415 亿美元的年收入。

1.4 机器学习方法分类和本书组织

本章前面几节介绍了机器学习的概念、历史和代表成就，从中可以看到这样一些核心趋势：1.机器学习已经成为人类最重要的工具并且在未来还将发生越来越重要的作用；2.机器学习需要强大的计算能力，当前图形处理器已经成为机器学习应用的关键计算平台。目前，机器学习的教材已经汗牛充栋，然而把机器学习算法转变为高性能图形处理器代码仍然是一个及其困难的过程，需要结合机器学习、并行计算和编程实现等多方面的能力。因此，本书定位于使用图形处理器实现高性能机器学习算法，帮助读者跨越从有效思想到高效代码的跨越。为了实现这一目的，本书选择了一系列经典机器学习算法，详细介绍算法流程并在此基础上介绍设计基于图形处理器的并行程序设计。

表 1-2. 机器学习重要术语

机器学习术语	定义	在统计学和相关学科的等价术语
输入（Input）	用于进行训练和分类、判别和预测的输入数据	变量、独立变量（Independent Variable）、（数据库的）列
特征（Feature）	输入数据的特定度量方式	
案例（case）、实例（Instance）、例子（Example）	一组独立观测的数据特征	观测（Observation）、记录、（数据库的）行
标签（Label）	机器学习程序对于给定数据的理想输出结果	依赖变量（dependent Variable）、目标
训练（Train）	通过对数据进行学习操作、提高机器学习程序性能的过程	拟合（fit）、学习（Learning）

本书不是机器学习的教程，不会系统介绍相关的概念。另一方面，由于机器学习理论从计算机科学发展而来，同时在近年来又高速整合了统计学和其它学科的理论和方法，因此术语有一些混乱。为了帮助读者厘清概念，表 1-2 列出常见机器学习术语的定义及其在统计学和相关学科的常见叫法。

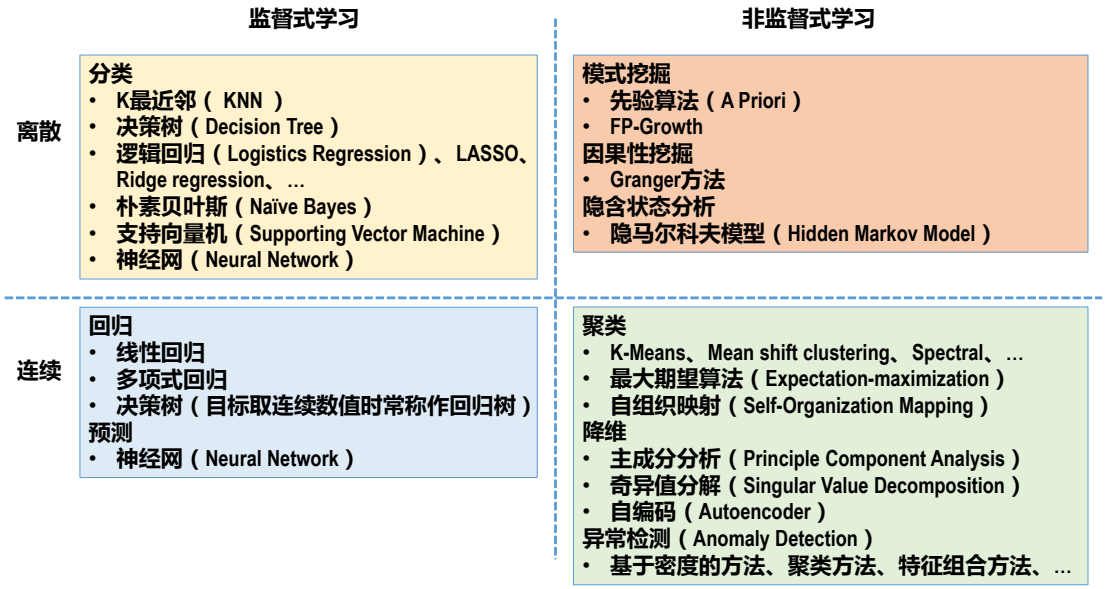

图 1-23. 机器学习算法分类

由于目标应用和算法结构的复杂性，机器学习算法种类及其繁杂。为了能够让读者建立一个宏观的认识。机器学习算法总是要根据目标对现有数据进行学习或训练形成一定的模型，然后用模型对当前数据进行分析，从而实现识别、分类或其它目标。因此，怎样进行训练是对机器学习算法最基本的分类手段。实际上，按照训练方法，机器学习算法可以分为三个大类：监督式学习、非监督式学习和

增强式学习。所谓监督式学习，指的是学习过程中的数据样本拥有已知结果，即正确的输入/输出对，机器学习程序根据已知结果进行训练，然后可以处理未知结果的数据。该类方法目前最为成熟，回归分析和神经网属于典型的监督式学习方法。非监督式学习方法不需要已知结果的样本，机器学习程序自行从数据中寻找关联性或相似性。聚类分析是典型的非监督式学习方法。非监督式学习也经常作为监督式学习的一个预处理步骤，从数据中寻找恰当的特征。增强式学习解决的实际上是规划问题，重点在于根据当前历史和当前状态产生一个动作序列，从而最大化的预期利益。前述的 AlphaGo 是标准的增强式学习案例。增强式学习强调的是过程，在每一个步骤上，经常采用监督式学习和非监督式学习。因此，本书主要考虑监督式学习和非监督式学习方法。按照学习目标的属性来说，可以分为离散和连续两类。所谓离散目标，指机器学习程序的输出是若干事先确定的数值或范畴，例如对图像分类时，可以输出“猫”、“狗”、“鸟”、“鱼”和“未知”；而连续目标指机器学习程序的输出是一个连续的变量，例如是由回归分析产生的函数，或者是一个概率分布。图 1-23 给出一个简化的机器学习算法分类图。此外，机器学习方法的一个重要分支是使用组合算法（Ensemble Algorithm），即使用若干相对较弱的学习模型独立进行训练，然后把训练结果整合起来进行整体预测。这是一类非常强大的算法，例如 Boosting、Bagging 和 AdaBoos 都是非常成功的范例。

从以上分类出发，同时考虑图形处理器计算，本书分成四个部分：基础、监督式学习方法、非监督式学习方法以及图模型和神经网。基础部分首先介绍机器学习的概念、历史和成功案例，然后介绍 GPU 计算和矩阵计算的基本知识；第二部分介绍经典的监督式学习方法及其 GPU 编程方法，包括 KNN 分类算法、决策树分类算法、支持向量机、逻辑回归和 Boosting 方法；第三部分侧重于非监督式学习方法及其 GPU 编程方法，包括聚类算法、降维方法和关联规则挖掘；第四部分介绍比较深入的内容，包括图模型（贝伊斯网络）、神经网和深度神经网。最后一部分中，神经网和深度神经网其实属于监督式学习，由于其特殊性放在这里讨论。以上内容当然只能涵盖机器学习领域的一小部分，但是具有足够的代表性，能够覆盖机器学习计算过程的绝大多数核心模式，从而帮助读者建立针对机器学习计算的并行思维方式和基本并行化方法。

文献导读

数据科学家和分析师的工作极具挑战性。首先，她（他）们必须具有数学家特别是统计学家的技巧，能够驾驭复杂的数学工具；其次，数据科学家和分析师需要福尔摩斯一般的推理能力，从纷繁复杂的数据中看到隐藏的规律；第三，她（他）们需要培养自然科学家的直觉，能够用足够简单然而又不是过分简化的模型去综合数据；第四，数据科学家和分析师需要拥有一点艺术天赋，能够把数据用美好的方式程序出来，使得背后的规律更加容易揭示；最后，她（他）们必须向机会主义者一样愿意尝试任何可能性，并且永远保持乐观，坚信数据中必然存在可以发现的规律。

机器学习的教材和相关书籍极其丰富，用汗牛充栋形容绝不过分。同时，经典机器学习教材的共同特点是知识密集、篇幅浩繁。那么，初学者怎样快速而有效地入门呢？

笔者的推荐是首先阅读科普级文献，逐步摸清机器学习的整体脉络和目标应用。其中，笔者特别推荐《The LION Way: Machine Leaning plus Intelligent Optimization》（Roberto Battiti 和 Mauro Brunato 著，CreateSpace Independent Publishing Platform; 1 edition, 2014），这本书篇幅不大，内容却相当全面，覆盖了机器学习理论的各条主线，但是没有涉及到过多的理论和算法细节，同时穿插了各种相关背景知识，读来趣味十足。如果想了解进行应用方面的科普，可以看看各种以“Data Mining”和“Data Science”为题目的著作，例如《Learning to Love Data Science》（Mike Barlow，O’Reilly Media, Inc., 2015），这本书介绍数据挖掘的背景，对机器学习技术在各个领域的应用和前景有有简明扼要的介绍。类似的著作还有《The Art of Data Science》（Roger D. Peng 和 Elizabeth Matsui 著，lulu.com，2016）和《Data Science from Scratch》（Joel Grus 著，O’Reilly Media，2015）。由 Mohamed Medhat Gaber 主编的《Journeys to Data Mining: Experiences from 15 Renowned Researchers》（Mohamed Medhat Gaber 编著，Springer，2012）提供了 15 位著名机器学习科学家的研究历程，介绍他们的动机、成果、研究工具和预期目标以及对初学者的建议，非常值得初学者阅读。此外，Nate Silver 的名著《The Signal and the Noise: Why So Many Predictions Fail--but Some Don't》（Penguin Books; 1 edition 2015）是一本应用统计学的科普著作，由于统计学已经成为机器学习的核心工具，所以该书适合作为机器学习启蒙读本，其中文版已经由中信出版社出版（西尔弗著，《“信号与噪声》，2015）。类似的统计学科普名著还有《The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century》（David Salsburg 著，Holt Paperbacks, 2002，中文版是《女士品茶》，中国统计出版社，2004）和 Sharon B. McGrayne 的科普名著《The Theory

that Would Not Die》（Yale University Press，2011），前者对于现代统计学发展史有一个清楚的梳理，后者对统计学的贝叶斯理论有着极为精到的论述，这两本书也提供了众多统计学和统计学家的八卦。另一本有趣的科普著作是 Philip E. Tetlock and Dan Gardner 所著的《Superforecasting: The Art and Science of Prediction》侧重于基于数据的预测，颇有思想深度，该书的中文版也已经由中信出版社出版（《超预测：遇见未来的艺术和科学》，2015）。另一本有趣的科普著作是著名复杂学家 Albert-Laszlo Barabasi 撰写的《Bursts: The Hidden Patterns Behind Everything We Do, from Your E-mail to Bloody Crusades》（Plume，Reprint edition, 2011），这本书非常巧妙地把大数据和复杂学理论编织在一起，为机器学习和数据科学引入了新的视角。

接下来，读者需要确定自己的目标，是成为机器学习研究人员，机器学习应用开发人员，还是数据分析师？

如果目标是机器学习研究人员，系统化的理论学习必不可少，那么总要读一下大部头著作了。早期的机器学习算法以计算机科学为主要工具，集大成的著作是 Stephen Marsland 的 Machine Learning: An Algorithmic Perspective（Chapman and Hall/CRC, 2009。近来，统计学理论一已经改造了整个机器学习领域，从该视角入手并以百科全书方式讲解机器学习理论的两本经典著作是《The Elements of Statistical Learning: Data Mining, Inference, and Prediction》（Trevor Hastie, Robert Tibshirani 和 Jerome Friedman 著，Springer，2nd ed. 2009. Corr. 7th printing 2013 edition，2011，即著名的“ESL”）和《Pattern Recognition and Machine Learning》（Christopher Bishop 著，Springer，2007）。不过，上述两边书稍嫌艰涩，能够兼顾深度和广度、并且比较适合拥有计算机科学或其它相关领域基础的初学者的教材是《Machine Learning: A Probabilistic Perspective》（Kevin P. Murphy，The MIT Press，2012）。此外《The Elements of Statistical Learning: Data Mining, Inference, and Prediction》的作者们也另行撰写了一部更适合初学者的教材《An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)》（Gareth James，Daniela Witten，Trevor Hastie 和 Robert Tibshirani 著，Springer, 2013）。更加具有时代感、从贝叶斯理论和优化理论入手介绍机器学习的作品为《Machine Learning: A Bayesian and Optimization Perspective》（Sergios Theodoridis 著）。深度学习的专门教材还不多见，目前系统化的教材只有 Ian Goodfellow、Yoshua Bengio 和 Aaron Courville 合著的《Deep Learning》，即将由 MIT 出版社出版，目前的电子版可以在 http://www.deeplearningbook.org/看到。

对于机器学习应用开发人员来说，John D. Kelleher、Brian Mac Namee 和 Aoife D'Arcy 合著的《Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies》（The MIT Press，2015）是很好的入门选择，更多从应用角度介绍针对预测的数据分析方法。如果嫌上本书篇幅太大（1000页以上），可以考虑《Machine Learning》（Peter Flach 著，Cambridge University Press，2012）和《Applied Predictive Modeling》（Max Kuhn 和 Kjell Johnson 著，Springer，2013），论述非常清晰。Jason Bell 的《Machine Learning: Hands-On for Developers and Technical Professionals》（Wiley，2014）适合已经具有信息科学训练和实践的程序员作为入门书。Michael Nielsen 的电子书《Neural Networks and Deep Learning》（http://neuralnetworksanddeeplearning.com/）是深度学习的入门书。

从数据分析师角度看，选择一种适合机器学习任务的脚本式编程语言作为入门手段是最好的方式，相关著作包括 Sebastian Raschka 的《Python Machine Learning》（Packt Publishing，2015）、Brett Lantz 的《Machine Learning with R》（Packt Publishing，2015）和 Giancarlo Zaccone 的《Getting Started with TensorFlow》（Packt Publishing，2016）等。

当然，以上我们只介绍相关教材和著作，要洞悉机器学习的奥秘，更加重要的是动手练习，最好的手段当然是带着问题去寻找工具、寻找方法。

参考文献

[1] SINTEF. "Big Data, for better or worse: 90% of world's data generated over last two years." Science Daily, 22 May 2013.

[2] Hayes-Roth, Frederick; Waterman, Donald; Lenat, Douglas (1983). Building Expert Systems. Addison-Wesley. ISBN 0-201-10686-8.

[3] Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7, p.2.

[4] Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired. Aug. 2008.

[5] Olga Russakovsky, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis (2015) 115:211–252.

[6] LeCun, Yann, et al. “Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.

[7] Jürgen Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117.

[8] UNC Vision Lab ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015), 2015.

[9] Mnih, Volodymyr. Human-level control through deep reinforcement learning. Nature, Vol. 518, Feb. 2015, 529–533

[10] Silver, David, et al. Mastering the game of Go with deep neural networks and tree search, Nature, Vol. 529, Jan. 2016, 484–489.

[11] Ferrucci, David, et al. Building Watson: An Overview of the DeepQA Project. IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 10 MAY/JULY 2012.
[12] D. A. Ferrucci, Introduction to “his is Watson”, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1 MAY/JULY 2012.
[13] Kalyanpur, A. et al. Structured data and inference in DeepQA, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 10 MAY/JULY 2012.