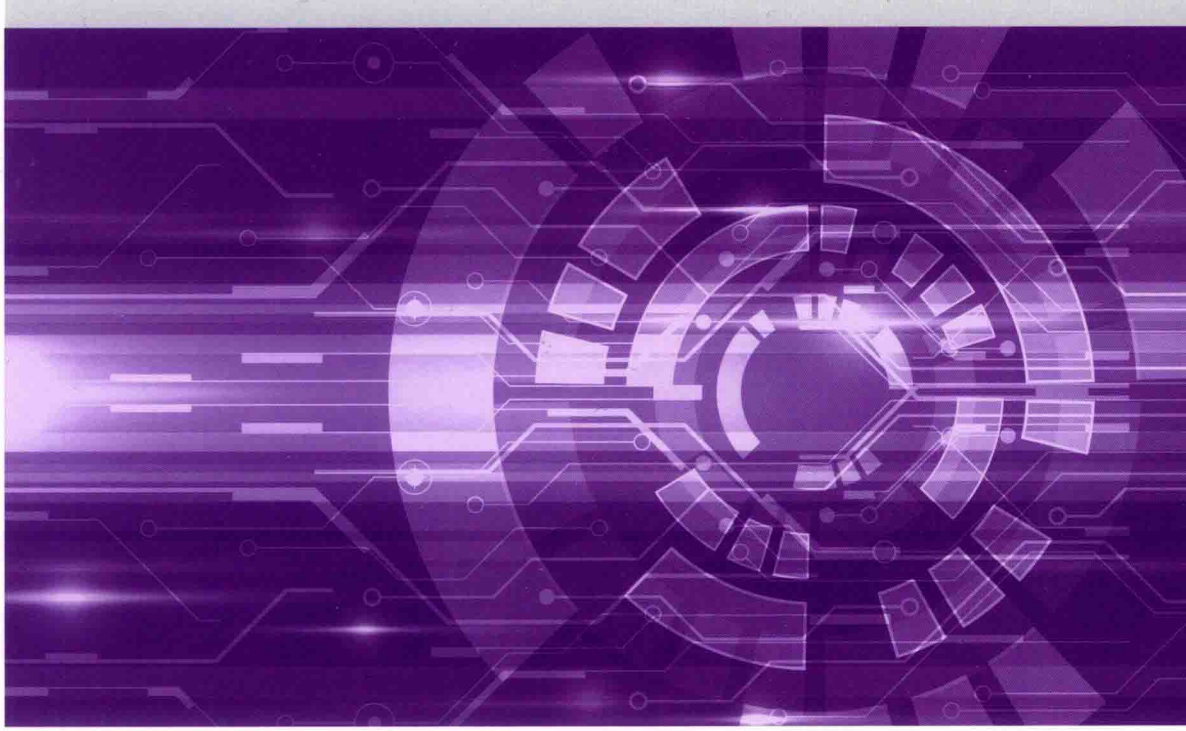
• 大数据应用人才培养系列教材 •

数据标注工程

■总主编◎刘　鹏　张　燕　■主编◎刘　鹏



清华大学出版社

大数据应用人才培养系列教材

数据标注工程

总主编　刘　鹏　张　燕
主　编　刘　鹏

清华大学出版社
北　京

内 容 简 介

本书是由中国大数据应用联盟人工智能专家委员会主任刘鹏教授主编的一本系统学习数据标注技术的教材。本书使用浅显易懂的语言，系统地介绍了数据标注的基本概念、分类、流程、质量检验、管理和应用等。通过理论与实战相结合的方式，帮助读者由浅入深进行学习，从而真正掌握数据标注的核心技术、实施和管理方法。本书既可以作为培养应用型人才的课程教材，也适用于初学者，以及广大的数据标注行业从业者。数据标注行业正迅速成长，目前正缺乏一本权威教材，希望本书能够填补这个空白。

图书在版编目（CIP）数据

数据标注工程 / 刘鹏主编. —北京：清华大学出版社，2019
（大数据应用人才培养系列教材）
ISBN 978-7-302-52844-9

I. ①数… II. ①刘… III. ①数据处理－教材 IV. ① TP274

中国版本图书馆 CIP 数据核字（2019）第 082682 号

责任编辑： 贾小红
封面设计： 刘 超
版式设计： 王凤杰
责任校对： 马军令
责任印制： 杨 艳

出版发行： 清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社 总 机：010-62770175 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者： 北京密云胶印厂
经 销： 全国新华书店
开 本： 185mm × 260mm **印 张：** 9 **字 数：** 152 千字
版 次： 2019 年 6 月第 1 版 **印 次：** 2019 年 6 月第 1 次印刷
定 价： 46.00 元

产品编号：082269-01

编写委员会

总主编　刘　鹏　张　燕

主　编　刘　鹏

编　委　张　燕　梁　南　武郑浩　李燕祥

总 序

短短几年间，大数据就以一日千里的发展速度快速实现了从概念到落地，直接带动了相关产业的井喷式发展。数据采集、数据存储、数据挖掘、数据分析等大数据技术在越来越多的行业中得到应用，随之而来的就是大数据人才缺口问题的凸显。根据《人民日报》的报道，未来 3 ~ 5 年，中国需要 180 万名数据人才，但目前只有约 30 万人，人才缺口达到 150 万名之多。

大数据是一门实践性很强的学科，在其呈现金字塔型的人才资源模型中，数据科学家居于塔尖位置，然而该领域对于经验丰富的数据科学家需求相对有限，反而是对大数据底层设计、数据清洗、数据挖掘及大数据安全等相关人才的需求急剧上升，可以说占据了大数据人才需求的 80% 以上。比如数据清洗、数据挖掘等相关职位，需要源源不断的大量专业人才。

巨大的人才需求直接催热了相应的大数据应用专业。2018 年 1 月 18 日，教育部公布“大数据技术与应用”专业备案和审批结果，已有 270 所高职院校申报开设“大数据技术与应用”专业，其中共有 208 所职业院校获批“大数据技术与应用”专业。随着大数据的深入发展，未来几年申请与获批该专业的职业院校数量仍将持续走高。同时，对于国家教育部正式设立的“数据科学与大数据技术”本科新专业，在已获批的 35 所大学之外，2017 年申请院校也高达 263 所。

即使如此，就目前而言，在大数据人才培养和大数据课程建设方面，大部分专科院校仍然处于起步阶段，需要探索的问题还有很多。首先，大数据是个新生事物，懂大数据的老师少之又少，院校缺“人”；其次，院校尚未形成完善的大数据人才培养和课程体系，缺乏“机制”；再次，大数据实验需要为每位学生提供集群计算机，院校缺“机器”；最后，院校没有海量数据，开展大数据教学实验工作缺少“原材料”。

对于注重实操的“大数据技术与应用”专业专科建设而言，需要重点

面向网络爬虫、大数据分析、大数据开发、大数据可视化、大数据运维工程师的工作岗位，帮助学生掌握大数据技术与应用专业必备知识，使其具备大数据采集、存储、清洗、分析、开发及系统维护的专业能力和技能，成为能够服务区域经济的发展型、创新型或复合型技术技能人才。无论是缺“人”、缺“机制”、缺“机器”，还是缺少“原材料”，最终都难以培养出合格的大数据人才。

其实，早在网格计算和云计算兴起时，我国科技工作者就曾遇到过类似的挑战，我有幸参与了这些问题的解决过程。为了解决网格计算问题，我在清华大学读博期间，于2001年创办了中国网格信息中转站网站，每天花几个小时收集和分享有价值的资料给学术界，此后我也多次筹办和主持全国性的网格计算学术会议，进行信息传递与知识分享。2002年，我与其他专家合作的《网格计算》教材正式面世。

2008年，当云计算开始萌芽之时，我创办了中国云计算网站（在各大搜索引擎“云计算”关键词中排名第一），2010年出版了《云计算》，2011年出版了《云计算》（第2版），2015年出版了《云计算》（第3版），每一版都花费了大量成本制作并免费分享了对应的几十个教学PPT。目前，这些PPT的下载总量达到了几百万次之多。同时，《云计算》一书也成为国内高校的优秀教材，在中国知网公布的高被引图书名单中，《云计算》在自动化和计算机领域排名全国第一。

除了资料分享，在2010年，我们也在南京组织了全国高校云计算师资培训班，培养了国内第一批云计算老师，并通过与华为、中兴和360等知名企业合作，输出云计算技术，培养云计算研发人才。这些工作获得了大家的认可与好评，此后我接连担任了工信部云计算研究中心专家、中国云计算专家委员会云存储组组长、中国大数据应用联盟人工智能专家委员会主任等。

近几年，面对日益突出的大数据发展难题，我们也正在尝试使用此前类似的办法去应对这些挑战。为了解决大数据技术资料缺乏和交流不够通透的问题，我们于2013年创办了中国大数据网站（thebigdata.cn），投入大量的人力进行日常维护，该网站目前已经在各大搜索引擎的“大数据”关

键词排名中位居第一；为了解决大数据师资匮乏的问题，我们面向全国院校陆续举办多期大数据师资培训班，致力于解决“缺人”的问题。

2016 年年末至今，我们在南京多次举办全国高校 / 高职 / 中职大数据免费培训班，基于《大数据》《大数据实验手册》以及云创大数据提供的大数据实验平台，帮助到场老师们跑通了 Hadoop、Spark 等多个大数据实验，使他们跨过了“从理论到实践，从知道到用过”的门槛。

其中，为了解决大数据实验难的问题而开发的大数据实验平台，正在为越来越多高校的教学科研带去方便，帮助解决“缺机器”与“缺原材料”的问题：2016 年，我带领云创大数据（股票代码：835305）的科研人员，应用 Docker 容器技术，成功开发了 BDRack 大数据实验一体机，它打破虚拟化技术的性能瓶颈，可以为每一位参加实验的人员虚拟出 Hadoop 集群、Spark 集群、Storm 集群等，自带实验所需数据，并准备了详细的实验手册（包含 42 个大数据实验）、PPT 和实验过程视频，可以开展大数据管理、大数据挖掘等各类实验，并可进行精确营销、信用分析等多种实战演练。

目前，大数据实验平台已经在郑州大学、成都理工大学、金陵科技学院、天津农学院、西京学院、郑州升达经贸管理学院、信阳师范学院、镇江高等职业技术学校等多所院校部署应用，并广受校方好评。该平台也以云服务的方式在线提供（大数据实验平台，https://bd.cstor.cn），实验更是增至 85 个，师生通过自学，可用一个月时间成为大数据实验动手的高手。此外，面对席卷而来的人工智能浪潮，我们团队推出的 AIRack 人工智能实验平台、DeepRack 深度学习一体机以及 dServer 人工智能服务器等系列应用，一举解决了人工智能实验环境搭建困难、缺乏实验指导与实验数据等问题，目前已经在清华大学、南京大学、南京农业大学、西安科技大学等高校投入使用。

在大数据教学中，本科院校的实践教学应更加系统性，偏向新技术的应用，且对工程实践能力要求更高。而高职高专院校则更偏向于技术性和技能训练，理论以够用为主，学生将主要从事数据清洗和运维方面的工作。基于此，我们联合多家高职院校专家准备了《云计算导论》《大数据导论》《数据挖掘基础》《R 语言》《数据清洗》《大数据系统运维》《大数据实践》系

列教材，帮助解决“机制”欠缺的问题。

此外，我们也将继续在中国大数据和中国云计算等网站免费提供配套PPT和其他资料。同时，持续开放大数据实验平台、免费的物联网大数据托管平台万物云和环境大数据免费分享平台环境云，使资源与数据随手可得，让大数据学习变得更加轻松。

在此，特别感谢我的硕士导师谢希仁教授和博士导师李三立院士。谢希仁教授所著的《计算机网络》已经更新到第7版，与时俱进日臻完美，时时提醒学生要以这样的标准来写书。李三立院士是留苏博士，为我国计算机事业做出了杰出贡献，曾任国家攀登计划项目首席科学家，他治学严谨，带出了一大批杰出的学生。

本丛书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。

刘　鹏

于南京大数据研究院

2018年5月

前 言

“有多少智能，就有多少人工”。随着人工智能技术突飞猛进地发展，数据标注行业也随之异军突起。经过短短几年的发展，我国专职从事数据标注行业的人员已经突破 20 万，兼职人员的数量突破 100 万。在未来 5 年，专职数据标注工程师的缺口将高达 100 万。人工智能行业巨头纷纷寻找专业的数据标注工程师，但目前接受过系统培训的数据标注工程师少之又少。

早期的数据标注工作是由专门研究人工智能算法的工程师进行小规模的数据标注，但在人工智能第三次浪潮之下，小规模的数据标注已经不能满足人工智能的发展需求，所以在 2011 年开始出现专门从事数据标注工作的团队，并且慢慢形成了数据标注行业。从 2017 年开始，人工智能的应用开始呈爆炸式增长，大规模的数据标注需求涌入，让数据标注行业迎来真正的爆发，正式进入人们的视野。

在快速膨胀的需求与国家扶持政策的推动下，全国高职、中职院校纷纷启动数据标注应用型人才培养计划。然而，数据标注专业建设却面临重重困难。首先，数据标注是一个新生事物，懂数据标注的教师少之又少，院校缺“人”；其次，尚未形成完善的数据标注人才培养和课程体系，院校缺“机制”；最后，院校没有数据标注项目，开展数据标注教学实践工作缺“原材料”。

为了能够更系统地培养数据标注工程师，我们的团队经过大量的市场考察与调研，深入了解数据标注行业，对数据标注各个环节进行调查整理，推出了这本教材。本书先从数据标注基本概念开始，介绍数据标注的前世今生以及发展趋势，然后系统地梳理了数据标注分类及数据标注流程，再对数据标注质量检验和数据标注管理进行详细介绍，最后分析学习热门行业数据标注应用，对四大重点行业进行数据标注实战。本书致力于将理论与实践结合在一起，让读者真正掌握数据标注的核心技术。

本书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。我的邮箱：gloud@126.com，微信公众号：刘鹏看未来（lpoutlook）。

刘鹏　教授
于南京大数据研究院
2019 年 1 月 1 日

目 录

第 1 章

数据标注概述

无人驾驶、人脸识别、语音交互……在人工智能（Artificial Intelligence，AI）第三次浪潮之下，在计算力、算法与数据的合力推动下，人工智能技术的突破与行业落地如雨后春笋，焕发源源不断的生机。尤为令人瞩目的是，在灼热的人工智能发展背后，为其发展提供数据燃料的数据标注正在成为一门新兴产业。

1.1 数据标注的起源与发展

由于数据标注与人工智能相伴相生，在研究数据标注的同时，首先需要对人工智能追本溯源。人工智能的概念最早由约翰·麦卡锡于 1956 年达特茅斯会议上提出，意指让机器具有像人一般的智能行为。

在其提出以来的 60 多年中，人工智能的发展并非坦途，而是经历了沉沉浮浮、三起三落。人工智能在达特茅斯会议上经过了两个多月的讨论，并在会后推出了第一款聊天软件，让人直呼“人工智能来了”，并掀起了此后为期 20 年的第一次人工智能浪潮。

当时主要以注重演算与推理的符号主义以及深度学习的“前身”——连接主义为代表。对于此次人工智能的兴起，专家学者尤为看好，甚至指出，未来十年机器人就能够超越人类了。然而，就在大家期盼人工智能春天到

来之际，在 20 世纪 70 年代后期，人们却逐渐发现过去的理论与模型只能用于解决一些简单的问题，同时运算能力不足，人工智能的第一次浪潮偃旗息鼓，迎来了突如其来的冬天。

此后，经过短暂的消沉后，随着 20 世纪 80 年代两层神经元网络（BP 网络）的兴起，人工智能开始焕发出新的生机，迎来了第二次发展浪潮。其间，语音识别、语音翻译以及感知机模式成了典型代表。然而，这些现在看来再寻常不过的应用，彼时离人们的实际生活仍然较为遥远，人工智能也随之进入了第二次沉寂的低潮，人工智能发展历史如图 1-1 所示。

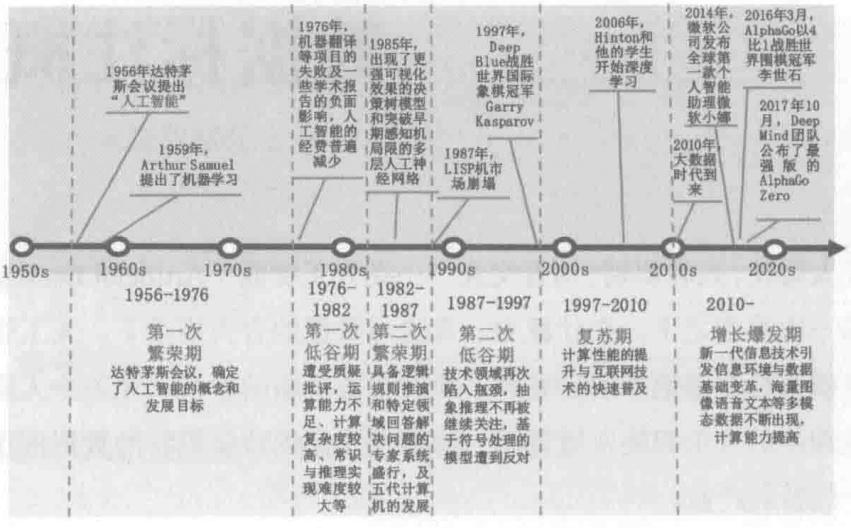


图 1-1 人工智能发展历史[1]

人工智能的第三次浪潮始于 Deep Blue（IBM 深蓝）的出现，其在 1997 年战胜了国际象棋冠军，而 2006 年“神经网络之父”Geoffrey Hinton 提出的深度学习技术进一步助推人工智能的发展，该技术于 2010 年大火，直接带动了人工智能的真正爆发，使其成了商界、创投界炙手可热的新星，并发展至今。不难预见，未来人工智能将实现由弱人工智能发展到强人工智能，直至超人工智能的高度。

纵览人工智能的发展脉络，在前两次发展浪潮中，人工智能发展起起伏伏，偶有爆发，却未能真正深入人们的生活。因此，当时由于量级比较小，为人工智能提供“喂养数据”的数据标注主要由研究的工程师完成，并不能称之为独立的职业。近年来，随着人工智能第三次浪潮的到来，数据标注的需求应接不暇，2011 年数据标注的外包市场开启，2017 年真正爆发，数据标注开始慢慢进入人们的视野。

1.1.1　什么是数据标注

2016 年，人工智能程序阿尔法围棋（AlphaGo）在与世界顶尖棋手的对决中奉上了令人惊艳的战绩，可谓是一战成名。此后横空出世的阿尔法零（AlphaGo Zero）作为 AlphaGo 的最新版本，自学 3 天，以 100：0 的成绩完胜此前击败李世石的 AlphaGo 版本；自学 40 天，以 89：11 的绝对优势击败阿尔法狗 Master（大师）版不同 AlphaGo 版本的棋力比较如图 1-2 所示。

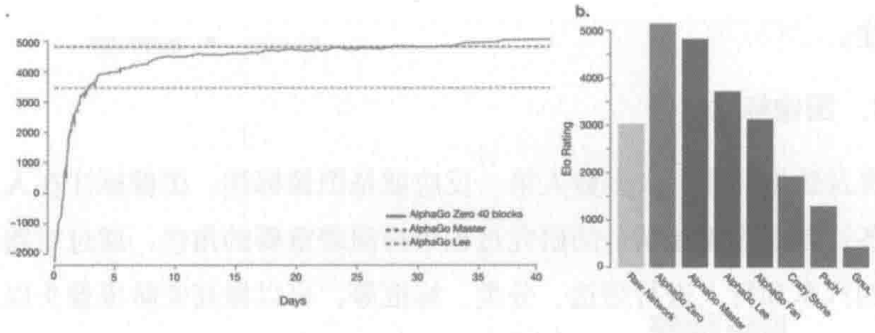


图 1-2　不同 AlphaGo 版本的棋力比较[2]

当我们感慨其成长速度时，也不能否定最初的 AlphaGo 也犹如出生的婴儿一般，对下棋一窍不通，其之所以能够快速升级成为棋坛高手，这与人类“喂养”的棋谱与数据相关，换言之，正是人类像教育小孩一样培养了 AlphaGo，才让其“学会”下棋。

举个简单的例子，当我们告诉孩子——“这是一辆汽车”，并把对应的图片展示在孩子面前，帮助他记住拥有四个轮子，可以有不同颜色的这种日常交通工具，当孩子下次在大街上遇到飞奔的汽车时，也能直呼“汽车”。

类比机器学习，如果准备让机器习得同样的认知能力，我们也需要帮助机器识得相应特征，两者不同点在于，对于人类来说，往往告诉他一次就能记住，下次遇到就能准确辨别；对于机器来说，需要我们提取有关汽车的特征，“喂”给他们大量带有汽车特征的图片，使其通过训练集反复学习，并通过测试集进行检查与巩固，最终准确识别汽车，而这些带有汽车特征的图片正是出自数据标注工程师。

简而言之，数据标注即通过分类、画框、标注、注释等，对图片、语音、文本等数据进行处理，标记对象的特征，以作为机器学习的基础素材。由