北京理工大学“双一流”建设精品出版工程

Big Data Analytics Theory and Technology

大数据分析理论与技术

罗森林　潘丽敏 ◎ 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

北京理工大学“双一流”建设精品出版工程

Big Data Analytics Theory and Technology

大数据分析理论与技术

罗森林　潘丽敏 ◎ 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

本书系统、全面地研究和论述大数据分析理论与技术，主要内容包括大数据分析基础认知、大数据分析核心架构、大数据分析计算模式、大数据与网络空间安全、大数据与自然语言处理、大数据与医学信息处理。

本书可满足各类高校多样化人才长期培养的需求，可供从事网络空间安全、计算机科学与技术、软件工程、人工智能、信息与通信工程等相关学科专业的教学、科研、应用人员阅读和使用，对从事大数据分析相关研究的人员具有重要的实用和参考价值。此外，本书也可供其他非专业及相关研究人员参考使用，具有重要的指导意义。

图书在版编目（CIP）数据

大数据分析理论与技术 / 罗森林，潘丽敏著. --北京：北京理工大学出版社，2022.2
ISBN 978-7-5763-0920-1

Ⅰ. ①大…　Ⅱ. ①罗…　②潘…　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2022）第 016496 号

出版发行 / 北京理工大学出版社有限责任公司
社　　址 / 北京市海淀区中关村南大街 5 号
邮　　编 / 100081
电　　话 / （010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68944723（其他图书服务热线）
网　　址 / http://www.bitpress.com.cn
经　　销 / 全国各地新华书店
印　　刷 / 保定市中画美凯印刷有限公司
开　　本 / 787 毫米×1092 毫米　1/16
印　　张 / 18.25
字　　数 / 426 千字
版　　次 / 2022 年 2 月第 1 版　2022 年 2 月第 1 次印刷
定　　价 / 78.00 元

责任编辑 / 王晓莉
文案编辑 / 王晓莉
责任校对 / 周瑞红
责任印制 / 李志强

PREFACE

前言

针对规模巨大的数据，大数据分析可以从数据资源中揭示内在规律、挖掘有用信息和帮助人们科学决策，其战略意义不仅在于数据资源的累积，更在于对其进行分析处理。大数据分析已经成数据科学的新常态，具有明显的多学科交叉特征，目前无论是高校、研究所还是企业，均要对其进行深入研究和应用。大数据分析成为计算机科学与技术、网络空间安全、信息与通信工程、软件工程、人工智能、数据科学等学科专业的必修内容，几乎所有的重点院校均有这些学科专业。

本书系统讨论了大数据理论与技术的体系框架及其核心知识图谱，融入著者的最新研究成果，主要内容包括大数据分析基础认知、大数据分析核心架构、大数据分析计算模式、大数据与信息安全对抗、大数据与自然语言处理、大数据与生物信息处理等。本书编写的目的是全面培养学生的大数据理论技术能力，加强其对数据科学基础理论和应用的理解，使其争取成为有竞争力的数据科学家。通过学习本书，可以增强学生的如下能力：综合运用计算机和数学知识分析处理大规模数据集的能力、从复杂数据中快速得到信息和发现关系的能力、现实世界具体问题的建模能力等。

体系结构方面，强调知识的系统性、层次性，突出重点既见树木又见森林，内容全面但不厚重。纵向上分为理论技术与工程实践，横向上强调网络空间安全、自然语言处理、生命信息处理等多学科应用。在抓住其精要的同时知识点尽量全面，涵盖多门单项技术的同时保持知识结构的系统性，有利于核心知识的快速理解与掌握。

内容范围方面，注重内容的深入性、先进性、时效性，强调交叉学科理论与实践的有机结合。适应大数据理论技术的动态发展，保证其理论技术核心知识的基础性和长时有效性；引入新理论、新技术、新方法、新案例，融入作者研究成果，保证其理论技术的先进性和前瞻性。

灵活使用方面，基于研究型教学思想注重读者的兴趣和学习的灵活性，支持学习的可持续发展，关注学习的间接效果，可满足各类高校多样化人才长期培养的需求。同时，融入人文素养的培养，讨论大数据分析中的非技术和工程类问题，涉及算法的可用性、隐私和社会影响等。

本书由罗森林、潘丽敏共同撰写，其中第4、5、6章由潘丽敏负责撰写，其余部分由罗森林负责撰写，罗森林负责全书的章节设计、内容规划和统稿。

本书的编写得到了北京理工大学教务部肖煊、刘畅、朱元捷等同志以及信息安全与对抗技术研究所刘晓双、郝靖伟、李新帅、闫晗、李玉、张寒青等同学多方面的帮助，在此一并表示衷心的感谢。同时，衷心感谢北京理工大学出版社王晓莉老师对本书详细、认真的修改和热情帮助，衷心感谢北京理工大学出版社多方面的支持和帮助。

由于时间和能力所限，书中难免会有不足和疏漏之处，敬请广大读者批评指正，以便再版时更加完善。

罗森林

2021 年 10 月于北京理工大学

目 录

CONTENTS

第1章
大数据分析基础认知

1.1 引　言

大数据又称巨量资料，指的是传统数据处理应用软件不足以处理的大或复杂的数据集。[1]与常规数据相比，大数据中蕴含的隐式的模式或规律可以起到更为有效的指导作用，因此有必要通过相应的数据分析技术进行深入挖掘。现代管理学之父 Peter Drucker 在其著作《21 世纪的管理挑战》中指出，我们正经历着一场信息革命，这不是在技术上、机器设备上、软件上或是速度上的革命，而是一场“概念”上的革命。以往信息技术的重点在“技术”上，目的在于扩大信息传播范围，提升信息的传播能力和传播效率，而新的信息革命的重点将会在“信息”上。在“第七次信息革命”的浪潮中，大数据及大数据分析技术扮演着至关重要的角色。

本章主要内容包括：大数据分析知识基础，大数据分析历史与现状，大数据分析主要应用，大数据分析中存在的问题，大数据分析的发展趋势。

1.2 知 识 基 础

1.2.1 基本概念

大数据（Big Data），指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。[2] IBM 使用“5V”来归纳大数据的特征，具体内涵如下。

① Volume：海量数据。大数据中数据的采集、存储和计算的量都十分庞大，只有起始计量单位达到 PB 的数据才可以被称为大数据，因此需要强大的计算能力和优秀的计算架构。

② Variety：种类和来源多样化。包括结构化、半结构化和非结构化数据。随着互联网和物联网的发展，又扩展到网页、社交媒体、感知数据，涵盖音频、图片、视频、模拟信号等，真正诠释了数据的多样性，也对数据的处理能力提出了更高的要求。

③ Value：获取有价值的数据。如果用石油行业来类比大数据分析，那么在互联网金融领域甚至整个互联网行业中，最重要的并不是如何炼油，而是如何获得优质原油。最重要的就是挖掘更多有价值的信息。因为大数据中数据价值密度相对较低，可以说是浪里淘沙却又弥足珍贵。随着互联网以及物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低，如何结合业务逻辑并通过强大的机器算法来挖掘数据价值，是大数据时代最需要解决

的问题。

④ Velocity：数据增长速度快，处理速度也快，时效性要求高。比如搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。

⑤ Veracity：数据的准确性和可信赖度，即数据的质量。大数据中的内容是与真实世界中的发生息息相关的，要保证数据的准确性和可信赖度。研究大数据就是从庞大的网络数据中提取出能够解释和预测现实事件的过程。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。典型的加工方法可分为数据分析和数据挖掘。数据分析是指根据分析目的，用适当的统计分析方法及工具，对收集来的数据进行处理与分析，提取有价值的信息，发挥数据的作用。它主要实现三大作用：现状分析、原因分析、预测分析（定量）。数据分析的目标明确，先做假设，然后通过数据分析来验证假设是否正确，从而得到相应的结论。主要采用对比分析、分组分析、交叉分析、回归分析等常用分析方法。数据分析一般都是得到一个指标统计量结果，如总和、平均值等，这些指标数据都需要与业务结合进行解读，才能发挥出数据的价值与作用。数据挖掘是指从大量的数据中，通过统计学、人工智能、机器学习等方法，挖掘出未知的且有价值的信息和知识的过程。数据挖掘主要侧重解决四类问题：分类、聚类、关联和预测（定量、定性）。数据挖掘的重点在于寻找未知的模式与规律，如我们常说的数据挖掘案例：啤酒与尿布、安全套与巧克力等，这就是事先未知的，但又是非常有价值的信息；主要采用决策树、神经网络、关联规则、聚类分析等统计学、人工智能、机器学习等方法进行挖掘；输出模型或规则，并且可相应得到模型得分或标签，模型得分如流失概率值、总和得分、相似度、预测值等，标签如高中低价值用户、流失与非流失、信用优良中差等。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘，但其必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术才能真正实现。[3]

对于数据的收集，互联网网页的搜索引擎需要将整个互联网所有的网页都下载下来，这项任务显然不可能凭一台设备完成，而是需要多台机器组成网络爬虫系统同时工作，每台机器下载一部分，才能在有限的时间内将海量数据下载完毕（如图 1.1 所示的 Nutch 搜索引擎）。

对于数据的传输，单一设备内存中的队列必定会由于数据量过于庞大而发生溢出，这时就需要基于硬盘的分布式队列发挥作用，如图 1.2 所示。分布式队列可以多台机器同时传输，只要队列的数量足够多，就不必担心数据的溢出。

对于数据的存储，也需要使用分布式文件系统来实现，使用多台机器的硬盘，使其构成统一的文件系统（如图 1.3 所示的 Hadoop 分布式文件系统，HDFS），以存储海量的数据。

再如数据的分析，单一设备的计算能力相当有限，面对海量数据往往显得力不从心。分布式计算的方法就可以很好地解决这一问题，其采用“分而治之”的思想，将大量的数据分成小份，每台机器处理一小份，多台机器并行处理，很快就能完成运算（如图 1.4 所示的 MapReduce）。

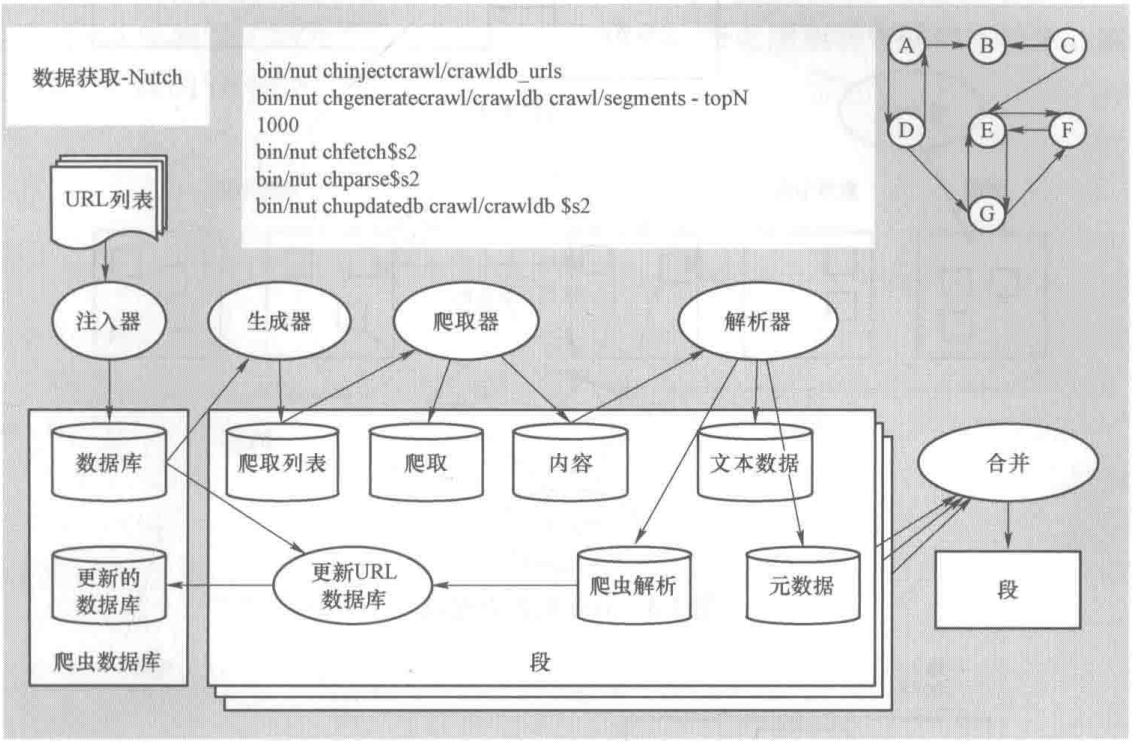

图 1.1　分布式数据获取

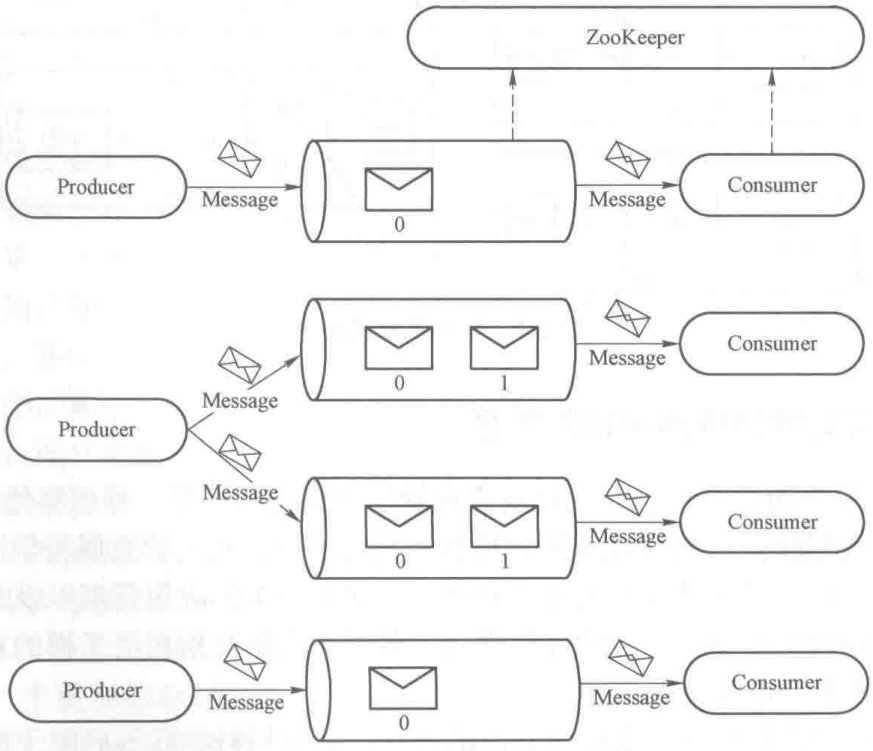

图 1.2　分布式数据传输

大数据处理的基本原理就是“众人拾柴火焰高”，通过将任务分发至大量的、分布式的节点来提升运算的性能和速度，最终完成海量的计算任务。

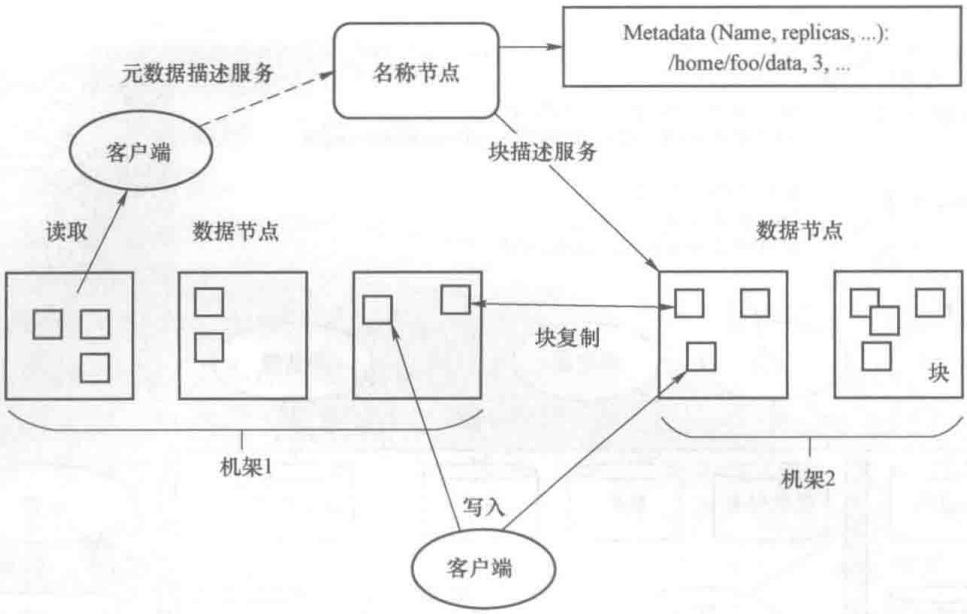

图 1.3　分布式数据存储

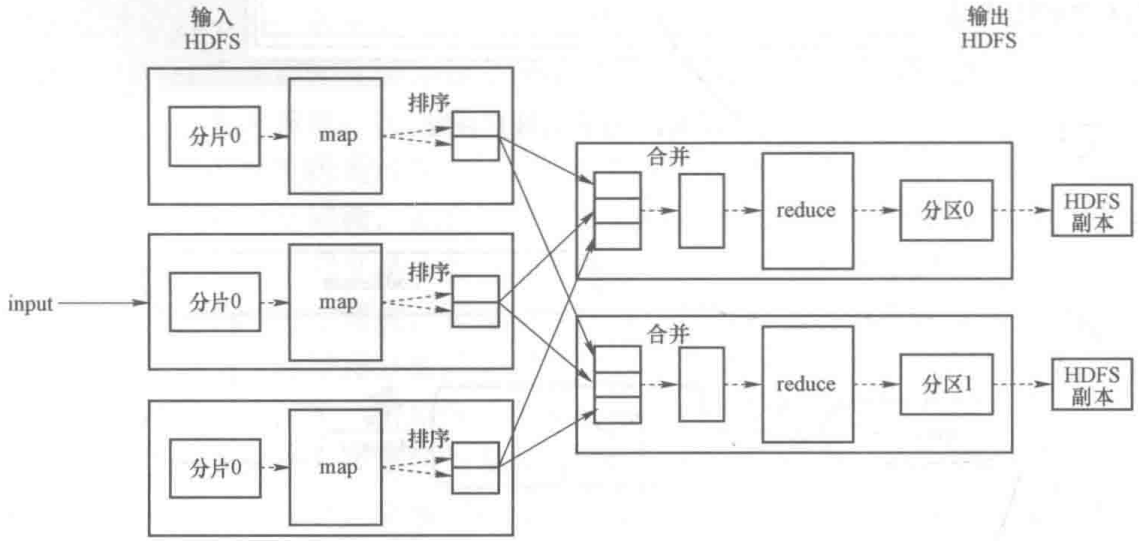

图 1.4　分布式数据分析

1.2.2　数据科学与其他学科的关系

大数据指具有“5V”特征的无法使用常规软件工具进行捕捉、管理和处理的数据集合，而研究基于海量数据的信息提取技术的学科称为数据科学。正式的数据科学通常指基于计算机科学、统计学、信息系统等学科的理论和技术，研究数据的收集整理以及从海量数据中分析处理、获得有效知识并加以应用的新兴学科；数据工程则是指利用工程的观点进行数据管理和分析以及开展系统的研发和应用。

相较之下，计算机科学学科是研究算法的科学，而数据科学远不局限于此。数据科学作为支撑大数据研究与应用的交叉学科，其理论基础来自多个不同的学科领域，包括计算机科学、统计学、人工智能、信息系统、情报科学等。数据科学的目的在于系统深入地探索大数据应用中遇到的各类科学问题、技术问题和工程实现问题，包括数据全生命周期管理、数据管理和分析技术和算法、数据系统基础设施建设以及大数据应用实施和推广。[4] 因此，多学

科交叉融合是数据科学的一个特点。图 1.5 是第一张关于“数据科学”概念的韦恩图，由 Drew Conway 在 2010 年制作。图中的中心部分是数据科学，韦恩图表明它是黑客技术、数学、统计学和其他实质性的专业知识的组合。

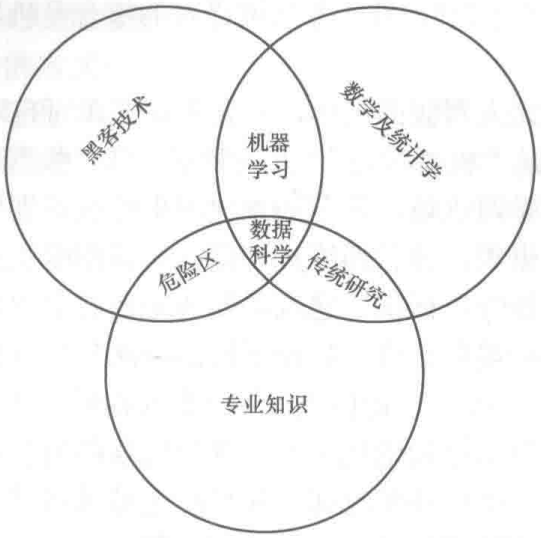

图 1.5 “数据科学”概念的韦恩图（Drew Conway 制作）

（一）数据科学与计算机科学

计算机科学是系统研究信息与计算理论基础以及它们在计算机系统中如何实现与应用的实用技术的学科。它通常被形容为对那些创造、描述以及转换信息的算法处理的系统研究。计算机科学和数据科学有重叠之处，两个领域中都使用了计算过程，且同样需要对编程语言和算法的有效理解，而基于这种理解去做什么则是这两个领域之间的主要区别。具体而言，计算机科学关注的是“如何”（How），而数据科学则关注“为什么”（Why）。计算机科学是一门基础学科，而数据科学则是一门应用学科。

计算机科学着眼于算法原理，致力于研究计算过程的具体细节，而不去过分关心功能实现的特定逻辑结果。计算机科学家可以开发应用程序，编写新的编程语言，或者设计一个生成和排序数据流的系统。但是对于计算机科学家来说，这些过程通常是建立在电压到比特的符号逻辑基础上，其结果是可预测的。

在数据科学中，算法原理被应用于更大的不确定领域，通常会给出关于商业等跨学科问题的概率性答案。现代数据科学家通常精通计算机科学，但他们可以有着数学、统计甚至商业背景。数据科学家可以设计算法，精练数据集，并通过数学模型解析大量数据，从而挖掘出可操作的知识。为实现此过程，数据科学家必须采取跨学科的方法，接受并处理不确定性。

（二）数据科学与软件工程

软件工程是软件开发领域里对工程方法的系统应用。1993 年，电气电子工程师学会（IEEE）给出了一个更加综合的定义：“将系统化的、规范的、可度量的方法用于软件的开发、运行和维护的过程，即将工程化应用于软件开发中。”数据科学通常需要对无法使用常规软件工具进行捕捉、管理和处理的数据集合进行处理，而设计并实现可以处理海量数据的系统及架构则是软件工程的目标。[5]

（三）数据科学与人工智能

人工智能是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统

的一门新的技术科学。人工智能的主要实现方式是通过大量数据的训练来实现它们的目标，这意味着人工智能往往需要一个巨大的数据集。然而，虽然数据科学和人工智能的主要实现方法都以大数据作为基础，但是二者的目标存在一定的差异。具体而言，数据科学旨在产生“见解”，人工智能旨在产生“行为”。另一个常被提及的概念是机器学习，通常机器学习旨在产生“预测”。[6]

假设我们正在制造一辆无人驾驶的汽车，并正在研究车可自动停靠在有停车标识的位置这个特定的问题。我们需要从“机器学习”“人工智能”和“数据科学”三个领域分别提取自己所需的知识技能。在机器学习领域，汽车必须使用摄像头识别停车标识。我们构建了包含数百万个街景标识图像的数据集，并且训练一个算法来预测哪里会有停车标识。在人工智能领域，一旦我们的车可以识别停车标识，它就需要决定何时采取刹车这个行为。过早或过晚刹车都是很危险的，并且我们需要它可以处理不同的道路状况（例如，识别一条光滑道路，它并不能很快减速），这是一个控制理论问题。[7] 而在数据科学领域，在街道测试中，我们发现汽车的表现并不足够好，停车标识出现了不少导致错误的消极因素。在分析街道测试数据之后，我们得到的结论是漏判率取决于时间：在日出之前或日落之后，更有可能错过停车标识。我们意识到，大部分训练数据仅包含白天时的对象，因此我们构建了包含夜间图像的更好的数据集，并返回到机器学习步骤。[8]

1.3 历史现状

1.3.1 发展历史

（一）“大数据”出现阶段（1980—2008 年）

1997 年，美国宇航局研究员 Michael Cox 和 David Ellswerth 首次使用“大数据”这一术语来描述 20 世纪 90 年代的挑战：模拟飞机周围的气流——是不能被处理和可视化的。[9] 数据集通常之大，超出了主存储器、本地磁盘，甚至远程磁盘的承载能力。这一问题被称为“大数据问题”。

2002 年在“9 • 11”袭击后，美国政府为阻止恐怖主义已经涉足大规模数据挖掘。前国家安全顾问 John Marlan Poindexter 领导国防部整合现有政府的数据集，组建一个用于筛选通信、犯罪、教育、金融、医疗和旅行等记录来识别可疑人的大数据库。一年后国会因担忧公民自由权而停止了这一项目。

2004 年，“9 • 11”委员会呼吁反恐机构应统一组建“一个基于网络的信息共享系统”，以便能快速处理应接不暇的数据。

2006 年，Google 首先提出云计算的概念，“大数据”在云计算出现之后才凸显其真正价值。

2007—2008 年，随着社交网络的激增，技术博客和专业人士为“大数据”概念注入新的生机。“当前世界范围内已有的一些其他工具将被大量数据和应用算法取代”。《连线》的 Chris Anderson 认为当时处于一个“理论终结时代”。一些政府机构和美国的顶尖计算机科学家声称，“应该深入参与大数据计算的开发和部署工作，因为它将直接有利于许多任务的实现”。2008 年 9 月，《自然》杂志推出了名为“大数据”的封面专栏。

（二）“大数据”热门阶段（2009—2011 年）

2009—2010 年，“大数据”成为互联网技术行业中的热门词汇。

2009 年，印度建立了用于身份识别管理的生物识别数据库；联合国全球脉冲项目已研究了对如何利用手机和社交网站的数据源来分析预测从螺旋价格到疾病爆发之类的问题；美国政府通过启动 Data.gov 网站的方式进一步开放了数据的大门，该网站超过 4.45 万个数据集被应用，以便保证一些网站和智能手机应用程序能够跟踪信息——这一行动激发了从肯尼亚到英国范围内的政府，他们相继推出了类似举措；欧洲一些领先的研究型图书馆和科技信息研究机构建立了伙伴关系，致力于改善在互联网上获取科学数据的简易性。

2010 年，Kenneth Cukier 发表大数据专题报告《数据，无所不在的数据》。

2011 年，扫描 2 亿年的页面信息，或 4 MB 磁盘存储，只需几秒即可完成；IBM 的沃森计算机系统在智力竞赛节目《危险边缘》中打败了两名人类挑战者，《纽约时报》称这一刻为“大数据计算的胜利”。同年 6 月，Mckinsey 发布了关于“大数据”的名为《大数据时代已经到来》的报告，正式定义了大数据的概念，并逐渐受到各行各业关注。12 月，工业和信息化部发布的物联网“十二五”规划，把信息处理技术作为四项关键技术创新工程之一提了出来，其中包括海量数据存储、数据挖掘、图像视频智能分析，这些是大数据的重要组成部分。

（三）“大数据”成为时代特征（2012—2016 年）

2012 年，Vikter Mayer－Schönberger（最早洞见大数据时代发展趋势的数据科学家之一）及 Kenneth Cookyer 在其著作《大数据时代》中，将大数据的影响分成了三个不同的层面，分别是思维变革、商业变革和管理变革。“大数据”这一概念乘着互联网的浪潮在各行各业中扮演着举足轻重的角色。“大数据”一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。数据正在迅速膨胀变大，它决定着人们的未来发展。随着时间的推移，人们将越发意识到数据的重要性。

2012 年 1 月，于瑞士达沃斯召开的世界经济论坛上发布的报告《大数据，大影响》宣称，数据已经成为一种新的经济资产类别。

2012 年，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，该倡议标志着大数据已经成为重要的时代特征；3 月，奥巴马政府宣布将 2 亿美元投资大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭。

2012 年，美国颁布了《大数据的研究和发展计划》；英国发布了《英国数据能力发展战略规划》；日本发布了《创建最尖端 IT 国家宣言》；韩国提出了“大数据中心战略”；世界上其他的一些国家也制定了相应的战略和规划。

2012 年 7 月，联合国发布了一份关于大数据政务的白皮书《大数据促发展，挑战与机遇》，总结了各国政府如何利用大数据更好地服务和保护人民。

2013 年是中国的“大数据元年”。虽说大数据概念存在已有时日，却因为互联网和信息行业的发展而引起人们关注，这一年大数据开始在我国逐渐展开，以势不可当的姿态进入人们的思想意识，并在社会的各个领域探索与落地实践。阿里巴巴 2013 年 1 月 1 日转型重塑平台、金融和数据三大业务，成为最早提出通过数据进行企业数据化运营的企业。

2014 年，“大数据”首次出现在当年的《政府工作报告》中。《政府工作报告》指出，要设立新兴产业创业创新平台，在大数据等方面赶超先进，引领未来产业发展。国务院通过《企业信息公示暂行条例（草案）》，要求在企业部门间建立互联共享信息平台，运用大数据等手