

普通高等教育经管类专业系列教材

大数据分析基础

李艺 宋琪 主编

清华大学出版社
北 京

内 容 简 介

《大数据分析基础》从大数据基础知识、大数据分析过程、大数据应用场景三个方面勾勒大数据基础知识图谱,全书共有16个章节。通过对大数据时代背景、大数据基本概念、大数据思维以及大数据与新一代数字技术的关系等大数据通识知识的介绍让读者了解大数据基础知识构成;从大数据分析的基本框架、数据采集与预处理、数据存储与管理、数据挖掘与分析、数据可视化等角度全面介绍大数据分析过程,全方位介绍大数据应用场景;从大数据安全、互联网大数据、物流大数据、城市管理大数据、零售大数据、大数据产业、大数据未来等方面,进行实景案例分析,拓展对大数据分析的认识和理解。

本书适合作为高等院校新商科和软件信息等专业本科生、研究生和MBA(工商管理硕士)学习大数据分析基础知识的教材,也可以供从事管理实践的中高层管理人员、信息系统管理和企业大数据管理部门人员以及从事大数据分析相关研究工作人员参考使用。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。举报:010-62782989, beiqinquan@tup.tsinghua.edu.cn。

图书在版编目(CIP)数据

大数据分析基础 / 李艺, 宋琪主编. -- 北京: 清华大学出版社, 2026. 7. -- (普通高等教育经管类专业系列教材). -- ISBN 978-7-302-72020-1

I. TP274

中国国家版本馆 CIP 数据核字第 2026L0A755 号

责任编辑: 施 猛 张 敏

封面设计: 马筱琨

版式设计: 熊仁丹

责任校对: 马遥遥

责任印制: 沈 露

出版发行: 清华大学出版社

网 址: <https://www.tup.com.cn>, <https://www.wqxuetang.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-83470000 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: 三河市东方印刷有限公司

经 销: 全国新华书店

开 本: 185mm×260mm 印 张: 16.25 字 数: 325 千字

版 次: 2026 年 7 月第 1 版 印 次: 2026 年 7 月第 1 次印刷

定 价: 59.00 元

产品编号: 106886-01

前 言

党的二十大提出：“加快发展数字经济，促进数字经济和实体经济深度融合，打造具有国际竞争力的数字产业集群。”在数字化发展的浪潮下，数据已成为推动社会进步、经济繁荣与科技创新的核心要素，是加快建设制造强国、质量强国、交通强国、网络强国、数字中国的基础，引发各领域巨大变革。数据应用场景存在于社会的方方面面，数据的力量无处不在，驱动着社会经济蓬勃发展。大数据分析能力培养成为高等院校新商科人才培养的重要知识模块，而了解大数据时代背景、把握大数据与新一代信息技术的关系、掌握大数据分析思维和基本过程以及在不同场景下的应用知识，成为数据时代人才培养的基本内容。

本书紧跟科教兴国、人才强国战略步伐，立足新商科人才对大数据基础知识及应用能力培养的需求，从大数据基础知识、大数据分析过程、大数据应用场景三个方面构建大数据基础知识图谱，帮助学生理解大数据创造经济和社会价值的基本活动。本书分为三部分，共有16章内容。大数据基础知识部分主要包括大数据时代背景、大数据基本概念、大数据思维以及大数据与新一代数字技术的关系，共4个章节内容；大数据分析过程部分主要介绍大数据分析的基本框架、数据采集与预处理、数据存储与管理、数据挖掘与分析、数据可视化，共5个章节内容；大数据应用场景部分是大数据知识的拓展，介绍了大数据安全、互联网大数据、物流大数据、城市管理大数据、零售大数据、大数据产业、大数据的未来，共7个章节内容，通过案例介绍帮助学生了解大数据在不同场景下的应用情况，提升学生对大数据基础知识的掌握。

本书的一大亮点是聚焦前沿理论，全面介绍大数据分析的基础知识，以案例为引领，帮助学生深度了解大数据时代的数据价值，分析中国本土案例，讲述中国大数据发展故事。本书另一大亮点是将正确的数据观和价值取向融入教材中，突出实事求是、合法合规、重视隐私、诚信为人等理念，引导学生合理、合法、合规地使用技术和数据，助力其创造商业价值和社会价值。

本书适合作为高等院校新商科和软件信息等专业本科生、硕士研究生和MBA（工商管理硕士）学习大数据分析基础知识的教材，也可以供从事管理实践的中高层管理人员、信息系统管理和企业大数据管理部门人员以及从事大数据分析相关研究工作人员参考使

用。本书配有数字化资源（扫描下方二维码获取），方便教师开展混合式教学，助力学生自主学习。



本书凝结了作者近20年一线教学经验和数据分析领域研究实践，是作者在分析、探索 and 整合中外研究前沿和企业最佳实践基础上取得的成果，更是团队成员通力合作的集体智慧结晶。感谢沈阳工业大学管理学院的硕士研究生吴宇森、宋雪、刘佳琪、申仔艳、张永娜、景文迪、张浩楠在书稿撰写和整理资料过程中承担的大量工作和做出的贡献。本书在撰写过程中还得到了多方面的帮助和支持，在此一并深表感谢！大数据技术发展日新月异，本书编写团队在内容选取和编排上力求科学与前瞻，但因水平有限，书中难免有不足之处，恳请广大师生和读者提出宝贵意见，以便后续修订完善，不断提高。反馈邮箱：wkservice@vip.163.com。

李艺 宋琪

2026年1月15日

目 录 / CONTENTS

第一部分

大数据基础知识

第 1 章 大数据时代 001

1.1 大数据的背景 004

1.2 大数据的发展历程 004

1.3 世界各国的大数据战略 007

1.4 大数据的行业应用 009

1.5 大数据的产业链 011

■ 本章小结 012

■ 思考题 012

第 2 章 洞悉大数据 013

2.1 大数据的定义 014

2.2 大数据的特征 015

2.3 大数据的类型 017

2.4 大数据标准化 018

2.5 大数据的价值 019

2.6 大数据的影响 021

■ 本章小结 025

■ 思考题 025

第 3 章 大数据思维 026

3.1 传统的思维方式 028

3.2 大数据时代需要新的思维方式 029

3.3 大数据的思维方式 030

■ 本章小结 035

■ 思考题 035

第 4 章 大数据与新一代数字技术的关系 036

4.1 大数据与云计算 037

4.2 大数据与物联网 045

4.3 大数据与云计算、物联网的关系 052

4.4 大数据与人工智能 054

■ 本章小结 066

■ 思考题 066

第5章 大数据分析的基本框架 068

5.1 从大数据集成到大数据分析 069

5.2 大数据分析的概念 070

5.3 大数据分析的过程 071

5.4 大数据分析的技术与难点 074

5.5 大数据分析的价值 082

■ 本章小结 087

■ 思考题 088

第6章 数据采集与预处理 089

6.1 数据采集 090

6.2 数据清洗 098

6.3 数据转换 101

6.4 数据脱敏 103

■ 本章小结 105

■ 思考题 105

第7章 数据存储与管理 106

7.1 数据存储与管理概述 106

7.2 数据存储模式 107

7.3 传统的数据存储与管理技术 112

7.4 大数据时代的存储管理系统 116

■ 本章小结 126

■ 思考题 126

第8章 数据挖掘与分析 127

8.1 数据挖掘的概念 127

8.2 数据挖掘的流程 129

8.3 数据挖掘典型算法 131

8.4 数据挖掘技术 137

8.5 数据挖掘与大数据 142

■ 本章小结 142

■ 思考题 143

第9章 数据可视化 144

9.1 大数据可视化概述 145

9.2 图形组合	153
9.3 可视化的类型	155
9.4 数据分组聚合与透视	158
■ 本章小结	162
■ 思考题	162

第三部分 大数据应用场景

第 10 章 大数据安全	164
10.1 大数据法律法规	165
10.2 传统数据安全的威胁	166
10.3 大数据安全与传统数据安全的不同	167
10.4 大数据安全问题	168
10.5 大数据安全保障措施	170
■ 本章小结	173
■ 思考题	173
第 11 章 互联网大数据	174
11.1 互联网大数据概述	175
11.2 互联网大数据的应用	178
11.3 “互联网+”大数据的发展	190
11.4 互联网大数据的案例	193
11.5 互联网大数据展望	195
■ 本章小结	196
■ 思考题	196
第 12 章 物流大数据	197
12.1 物流大数据概述	198
12.2 物流大数据的应用	200
12.3 物流大数据的发展现状	207
12.4 物流大数据的应用案例	210
12.5 物流大数据展望	211
■ 本章小结	213
■ 思考题	213
第 13 章 城市管理大数据	214
13.1 城市管理大数据概述	216
13.2 城市管理大数据应用	216

13.3 城市管理大数据的发展及现状	220
13.4 城市管理大数据的案例	221
13.5 城市管理大数据展望	222
■ 本章小结	222
■ 思考题	222
第 14 章 零售大数据	223
14.1 零售大数据概述	225
14.2 零售大数据的应用	226
14.3 零售大数据的案例	227
■ 本章小结	230
■ 思考题	230
第 15 章 大数据产业	231
15.1 大数据产业概述	232
15.2 数字经济	235
15.3 新基建背景下的数据中心发展	239
■ 本章小结	241
■ 思考题	241
第 16 章 大数据的未来	242
16.1 “大”字当头	243
16.2 经济发展趋势	244
16.3 科技发展趋势	245
16.4 产业发展趋势	245
■ 本章小结	247
■ 思考题	247

参考文献	248
------	-----



第一部分

大数据基础知识

第1章 大数据时代

学习目标及要求

1. 了解大数据的崛起过程。
2. 掌握大数据的概念。
3. 掌握大数据的特点。
4. 了解国外国内的大数据战略。
5. 了解在大数据时代，我们面临的机遇和挑战。

引导案例

2013年5月28日，阿里巴巴集团联合银泰集团、复星集团、富春控股、顺丰、“三通一达”（申通、圆通、中通、韵达）、宅急送、汇通以及相关金融机构共同宣布，开始联手共建“中国智能物流骨干网（China Smart Logistics Network, CSN）”，又名“菜鸟”（见图1-1）。菜鸟第一期将投入1000亿元人民币，计划在10年内，建立一张能支撑日均300亿元（年度约10万亿元）网络零售额的智能物流骨干网络，助力数千万家新型企业成长发展，实现全国范围内任意地区做到24小时内物流送达。不仅如此，菜鸟还提供充分满足个性化需求的物流服务。例如，用户在网购下单时，可选择“时效最快”“成本最低”“最安全”“服务最好”等多个快递组合类型。



图1-1 菜鸟

菜鸟网络由物流仓储平台和物流信息系统构成。物流仓储平台由多个大仓储节点、若干重要节点和更多城市节点组成。大仓储节点分为东北、华北、华东、华南、华中、

西南和西北等七大区域，选择中心位置进行仓储投资。物流信息系统整合了所有服务商的信息系统，实现了骨干网内部的信息统一。同时，该系统向所有的制造商、网商、快递公司、第三方物流公司完全开放，有利于物流生态系统内各参与方利用信息系统开展活动。

菜鸟是阿里巴巴创始人马云整合各方力量实施的“天网+地网”计划的重要组成部分。所谓“地网”，即阿里巴巴打造的中国智能物流骨干网，旨在建成覆盖全国的超级物流体系。所谓“天网”，是指以阿里巴巴集团旗下多个电商平台为核心的大数据平台。阿里巴巴集团的电商业务在中国占据绝对垄断地位，在这个平台上聚集了众多的商家、用户、物流企业，每天都会产生大量的在线交易，因此，这个平台掌握了网络购物物流需求数据、电商货源数据、货流量与分布数据以及消费者长期购买习惯数据，物流公司可以对这些数据进行大数据分析，科学优化仓储选址、干线物流基础设施建设以及物流体系建设，并根据商品需求分析结果提前把货物配送到需求较为集中的区域，做到“买家没有下单，货就已经在路上”，最终实现“以天网数据优化地网效率”的目标。有了天网数据的支撑，阿里巴巴可以充分利用大数据技术为用户提供个性化的电子商务和物流服务。用户从“时效最快”“成本最低”“最安全”“服务最好”等服务选项中选择快递组合类型后，阿里巴巴会根据以往快递公司的服务情况、各个分段的报价情况、即时运力资源情况、该流向的即时件量等信息，甚至可以融合天气预测、交通预测等数据进行相关的大数据分析，从而得到贴合用户需求的最优线路方案，并最终把相关数据分发给各个物流公司，完成物流配送工作。

可以说，菜鸟网络的关键在于信息整合，而不是资金和技术的整合。阿里巴巴的“天网”和“地网”，能够把供应商、电商企业、物流公司、金融企业、消费者的各种数据全方位、透明化地加以整合，并转化为电子商务和物流系统的行动方案。

一年一度的“双11购物狂欢节”是中国网民的一大盛事，也是对智能物流网络的一大考验。在每年的“双11”活动中，阿里巴巴都会结合历史数据，根据进入“双11”的商家名单、备货量等信息进行分析，并提前对“双11”订单量做出预测，精确到每个区域、网点的收发量，将所有信息与快递公司共享，这使得快递公司运力布局的调整更加精准。菜鸟网络还将数据向电商开放，如果某个区域的快递压力明显增大，那么菜鸟网络会通知电商错峰发货，或是提早与消费者沟通，以便快递公司及时调配运力。2014年天猫“双11”购物节的数据显示，大数据已经开始全面发力，阿里巴巴搭建的规模庞大的IT基础设施，可平稳支撑购物节当天571亿元交易额，实现6小时处理100PB数据、每秒处理7万笔交易订单。同时，以大数据为驱动，借助智能物流体系——菜鸟网络，天猫已经实现预发货，达成订单未下达、货品已提前调拨的高效物流运作模式。

1.1 大数据的背景

“大数据”指代一种数据集合，其规模已远超传统信息技术管理能力边界，突破常规数据库软件工具的处理极限。自2011年起，“大数据”概念席卷全球，成为继“云计算”“物联网”后又一科技焦点，相关报道与著作不断涌现。事实上，早在1980年，未来学家阿尔文·托夫勒在《第三次浪潮》中便将其誉为“第三次浪潮的华彩乐章”。2008年，《自然》杂志推出“Big Data”专题封面。2011年6月，麦肯锡全球研究院发布《大数据：下一个创新、竞争和生产力的前沿》报告，系统剖析其影响、关键技术及应用领域，预言其将成为未来生产力革新、消费需求增长的核心驱动力。2012年3月，美国联邦政府启动“大数据的研究和发展计划”，由国家科学基金会、卫生与公共服务部、能源部、国防部等六大部门联合投资2亿美元推进技术研发，引发全球关注。同年7月，联合国发布《大数据促发展：挑战与机遇》政务白皮书，总结各国政府利用大数据服务民生、指导经济的实践，并倡议成员国建立“脉搏实验室”挖掘数据价值。2014年，欧盟出台《数据驱动经济战略》，推动大数据成为独立经济行业，助力经济增长与就业扩张。

我国自2015年正式将大数据提升至国家发展战略。2015年5月，国务院在《中国制造2025》中明确“建设重点领域制造业工程数据中心，推动创新知识与工程数据开放共享”；同年8月，《促进大数据发展行动纲要》进一步要求全面推进大数据应用，加速数据强国建设。

全球范围内，随着计算机技术普及与行业系统规模扩张，企业级大数据规模已达数百TB至数百PB量级，远超传统计算技术与信息系统的处理能力。据互联网数据中心监测，全球数据量呈指数级增长，约每两年翻一番，且这一增速在2020年前持续保持，这意味着近两年产生的数据量等同于此前人类历史数据总和。

现如今，大数据正引发一场重大时代转型，其发展将深刻变革生活、工作与思维方式，重塑商业模式，并对经济、政治、科技、社会等多维度产生深远影响。

1.2 大数据的发展历程

1. 大数据的技术前提

大数据发展的技术前提在于三大核心要素的协同演进，其本质是数据存储、采集与处理能力的革命性突破。

第一大前提是存储技术的突破性进步与成本体系的重构。自1965年戈登·摩尔提出摩尔定律以来，半导体存储技术便开启了指数量级进化通道——集成电路上可容纳的晶体管

数量每18至24个月翻一番，性能同步提升。以物理存储为例，1955年IBM首台商用硬盘需6000美元存储1MB数据，而到2010年，同等存储容量成本骤降至0.005美分，且单位体积存储密度激增，单根发丝截面可集成数万个晶体管。这种存储密度与成本的颠覆性变化，配合数据中心规模化建设的规模效应，使海量数据存储从“贵族专属”变为“平民化”服务，为大数据时代的到来奠定了坚实的物质基础。

第二大前提是数据生成能力的指数级扩张，这得益于物联网与互联网技术的深度融合。物联网通过传感器网络将物理世界数字化，如智能家居实时采集温湿度数据，工业设备持续传输运行参数；互联网则通过智能终端构建“人体延伸”的数据采集网，如智能手环可连续记录用户心率、步频等生理信号，移动设备网络每日产生EB级用户行为数据。此外，大型科研装置（如粒子对撞机）单次实验即生成PB级原始数据，社交媒体与“互联网+”生态更催生了全民参与的数据洪流——每个用户都成为独立的数据源，持续产生并沉淀数据，最终引发人类历史上前所未有的数据爆炸。

第三大前提是云计算架构的革命性创新，它重构了数据处理的能力边界。云计算通过超大规模集群（如谷歌百万级服务器网络）构建起弹性计算资源池，实现计算成本的“分母效应”——企业不需要自建高成本数据中心，即可按需采购专业算力服务，如同使用水电般便捷。其核心优势体现在三方面：一是资源池的规模经济效应，通过集约化部署降低单位算力成本；二是按需分配的弹性架构，用户可实时调整计算资源规模；三是分布式计算范式的突破，使普通人也能获得曾专属于科研机构的超算能力。正是这种计算模式的革新，使每秒10万亿次运算从“实验室奇迹”变为“日常可用”，支撑起核爆模拟、气候预测、基因测序等前沿领域的突破。云计算将存储规模、数据采集量与计算能力推向新高度，人类开始从海量数据中挖掘其潜在价值，“大数据”概念由此孕育并逐步成为科研与产业的焦点。

2. 大数据的雏形

2007年1月，图灵奖得主、关系数据库的鼻祖吉姆·格雷（Jim Gray）于美国加州的一次会议上，发表了人生最后一次演讲——《第四范式：数据密集型科学发现》（*The Fourth Paradigm: Data-Intensive Scientific Discovery*）。吉姆·格雷凭借对人类科学发展特征的深刻洞察，提出将科学研究分为四类范式，依次为实验归纳、模型推演、模拟仿真和数据密集型科学发现（见图1-2）。

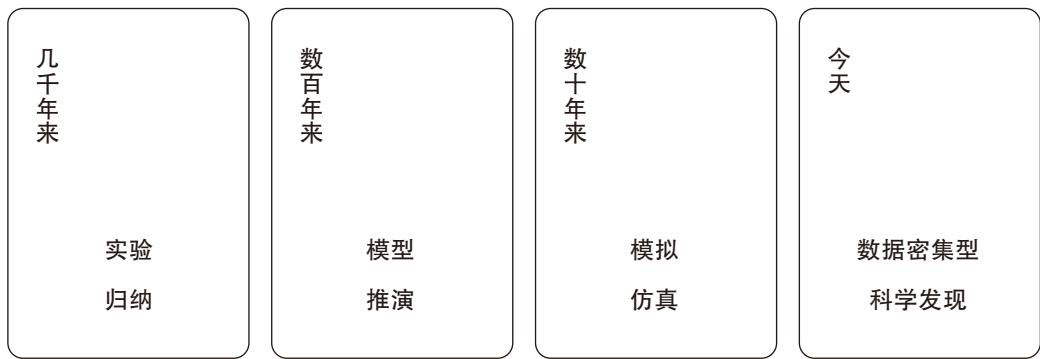


图1-2 科学研究的四类范式

让我们回溯科学范式演进的关键节点与范式革命。第一范式为经验科学阶段，涵盖18世纪前的科学实践，其核心是通过对有限对象的观察、归纳提炼规律，如伽利略通过斜面实验总结自由落体定律。第二范式是19世纪兴起的理论科学，以演绎推理构建理论体系，代表成果包括爱因斯坦相对论、麦克斯韦电磁方程组、量子力学与概率论等经典理论。第三范式始于20世纪中期，随着复杂现象研究需求激增，人类开始借助计算机超强算力开展建模预测，在天气模拟、地震分析、核试验仿真、原子运动轨迹计算等领域取得了突破。

然而，近年来数据量爆炸式增长正突破第三范式的能力边界——摩尔定律加速下的数据洪流已让传统计算科学范式力不从心。正如图灵奖得主吉姆·格雷在演讲中所说：“科研人员通过传感器、影像设备、超级计算机、粒子对撞机等手段获取海量数据，但当这些数据涌入硬盘后，如何有效利用却成难题。常有人向我求助：数据在手却无所适从，电子表格已失控！当需要处理万级电子表格而每个包含50个工作表时，传统工具必然崩溃。”欧洲大型强子对撞机、斯隆数字巡天项目等每天产生PB级数据，已远超第三范式的处理极限，传统方法难以挖掘其价值。

在此背景下，以海量数据处理为核心的第四范式——数据密集型科学发现应运而生。为适应数据爆发式增长这一趋势，亟待在数据捕获、存储管理、分析挖掘与可视化四大技术维度实现根本性突破。吉姆·格雷强调，现有工具尚未完全适配这一新范式需求，必须通过研究工具革新迎接数据社会挑战。他提出的“超大规模”“微细规模”等概念，精准刻画了数据时代的特征，揭示了数据爆炸对传统科研工具的根本性颠覆。至此，大数据革命的理论框架已清晰成型。这一思想的前驱吉姆·格雷提出的第四范式构成了大数据革命的思想雏形，标志着科学发现正式进入数据驱动的新纪元。

3. 大数据正式问世

2008年9月，美国《自然》杂志以专刊形式正式将“大数据”概念推向全球学术与产

业前沿，这期题为“下一个谷歌”的专题报道的发表标志着大数据正式进入主流视野并迅速成为科技创新领域的核心议题。该专刊通过跨领域专家访谈与趋势研判提出关键命题：正如多年前谷歌开创搜索引擎引发社会巨变一样，未来十年最具颠覆性的创新可能源于构建覆盖万物信息的超级数据库——通过打破虚拟与现实的边界，实现数据驱动的全球性变革。这种变革的潜力被描述为“下一个谷歌级”的创新范式，可能重新定义人类社会的信息交互与价值创造方式。随着专刊的发布，“大数据”作为描述数据爆炸性增长的新术语开始频繁出现在政策报告、学术论坛与产业白皮书中，其内涵从最初的技术概念逐步延伸为涉及经济、社会、科技等多维度的系统性变革理念，而《自然》杂志的预言也随着技术发展逐步从理论猜想转化为现实应用场景。

4. 大数据席卷全球

大数据浪潮席卷全球之际，专业机构与政策制定者率先构建起认知与实践框架。继《自然》杂志提出大数据基础概念后，麦肯锡研究院于2011年5月发布里程碑式报告《大数据：下一个创新竞争和生产力的前沿》，首次系统解构大数据革命的内在逻辑，并通过定量分析揭示其颠覆性影响力。该报告的发布引发了企业界、政府与学术界的空前关注，推动了全球范围内基于大数据的创新案例呈指数级增长。

2014年5月，美国白宫发布具有政策指导意义的《大数据：抓住机遇、守护价值》白皮书，提出双维并进的发展理念：一方面倡导通过数据驱动推动社会进步，特别是在市场与现有机制难以覆盖的领域释放数据价值；另一方面强调构建配套框架与研究体系，以捍卫美国社会对个人隐私保护、公平正义及反歧视的核心价值诉求。

在这场全球性变革中，任何企业与政府都无法置身事外。企业需要将自有数据资产提升至战略资源层级进行系统管理，通过培育数据科学家、数据工程师等专业人才，甚至设立首席数据官等专职岗位，构建数据驱动型业务增长模式。政府则需要主动推动公共数据开放共享，强化数据基础设施建设，同时引导社会各界调整发展战略，在大数据环境中保持创新力与竞争力。这一系列举措的核心目标是确保各类组织在大数据革命中把握战略转型机遇，有效应对全局性冲击，最终在数据要素市场化配置与价值挖掘的赛道上占据先机。

1.3 世界各国的大数据战略

随着大数据时代的到来，全球主要经济体均将大数据视为核心战略资源，并围绕其展开系统性布局。瑞士洛桑国际管理学院2017年发布的《世界数字竞争力排名》揭示了一个关键规律：国家数字竞争力与整体综合竞争力呈现显著正相关关系——数字竞争力领先

的国度，其整体竞争力同样位居前列，且更易孕育颠覆性创新成果。以美、英、日、韩等为代表的发达国家率先行动，将大数据发展提升至国家战略层面，视其为驱动经济转型升级、推动社会变革、强化国家综合竞争力的关键引擎，如表1-1所示。这些国家通过制定专项战略规划、加大研发投入、培育产业生态等举措，全力抢占大数据技术与产业发展的先发优势，同时积极构建数据主权防护体系，力求在数据要素市场化配置、数据价值挖掘利用等前沿领域占据制高点，从而在大数据时代的全球竞争中赢得战略主动。

表1-1 世界各国的大数据发展战略

国家	战略
美国	稳定实施“三步走”战略，打造面向未来的大数据创新生态
英国	紧抓大数据产业机遇，应对脱欧后的经济挑战
法国	通过发展创新性解决方案并应用于实践来促进大数据发展
韩国	以大数据等技术为核心应对第四次工业革命
日本	开放公共数据，夯实应用开发
中国	实施国家大数据战略，建设数字中国

1. 美国战略

作为大数据战略的先行者，美国率先完成了从商业概念到国家战略的升华，通过“三步走”战略构建起全球领先优势：第一步聚焦核心技术攻关与场景试点，第二步完善隐私保护等政策法规体系，第三步强化数据驱动的国家治理能力建设。2012年《大数据研究和发展计划》的出台标志着联邦政府正式启动系统性布局，成立高级指导小组统筹跨部门协作，通过海量数据收集分析提升社会预测能力。后续发布的报告《支持数据驱动型创新的技术与政策》进一步明确了技术路径，2014年发布的白皮书则平衡了价值挖掘与隐私保护的辩证关系，形成了完整的战略闭环。

2. 英国战略

英国采取“跟跑—并跑—领跑”的渐进式战略，在借鉴美国经验的基础上融入本土特色。2010年，英国政府数据网站上线，实现类Data.gov功能；2012年，数字化战略由商业创新技能部牵头成立数据战略委员会，通过数据开放吸引社会资本，培育新兴产业；2013年，英国政府拨付1.89亿英镑专项资金，其中的1000万英镑用于建立开放数据研究所。此后，《数字战略2017》《工业战略》等规划更明确了2025年数字经济贡献2000亿英镑的目标，以应对脱欧后的经济转型挑战。

3. 法国战略

法国依托传统工业强国的底蕴，将大数据深度融入社会治理。2011年，法国开放数据项目，率先实现交通、文旅等领域移动端应用；2013年，出台数字化路线图，明确大数据为战略性高新技术；2013—2018年，政府通过“公共私营合作”模式，在巴黎等地设立大数据孵化器，投资3亿欧元支持初创企业，并结合《创新2025》《新工业法国》等规划形成协同效应，推动大数据在城市管理、公共服务中的创新应用。

4. 韩国战略

韩国凭借高智能终端普及率和5G网络优势，将大数据定位为经济增长新引擎。在“创意经济”的框架下，韩国于2011年构建英特尔综合数据库；2012年，制定未来环境战略，同年启动千家大数据企业培育计划，并通过《第五次国家信息化计划》等落地实施；2016年，发布《智能信息社会中长期对策》，瞄准第四次工业革命，强化技术底座支撑。

5. 日本战略

日本以ICT技术融合为特色，2010年发布《信息通信技术新战略》，确立电子政务目标；2012年，发布电子政务开放数据战略草案，实现政府数据公开的突破；2013年，印发新IT战略，明确2013—2020年开放数据的核心地位。在抗震救灾、核电事故等公共危机中，大数据与ICT技术的融合已展现显著效能，形成了“技术—应用—治理”的良性循环。

6. 中国战略

中国的大数据战略呈现清晰的政策演进脉络：2014年，“大数据”首入中国政府工作报告；2015年，《促进大数据发展的行动纲要》确立“三位一体”战略框架；2016年，“十三五”规划单设国家大数据战略篇章；2017年，党的十九大报告强调与实体经济深度融合；2020年，数据被确立为生产要素；2021年，《数据安全法》构建安全顶层设计，“十四五”规划专章部署数字中国建设；2022年，“数据二十条”系统布局产权、交易、分配、治理四大制度；2023年，国家数据局成立，标志数据基础制度体系从“四梁八柱”走向具体实施，开启了数字中国建设新纪元。

1.4 大数据的行业应用

“数据，正在改变甚至颠覆我们所处的整个时代”，《大数据时代》一书的作者维克托·迈尔·舍恩伯格教授发出如此感慨。发展到今天，大数据已经无处不在，大数据的应

用已经深入金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、文娱等社会各行各业。大数据在各个行业的应用情况如表1-2所示。

表1-2 大数据在各个领域的行业应用一览

领域	大数据的行业应用
制造业	利用工业大数据提升制造业水平，包括产品故障诊断与预测、分析工艺流程、改进生产工艺、优化生产过程能耗、工业供应链分析与优化、生产计划与排程
金融行业	大数据在高频交易、社交情绪分析和信贷风险分析三大金融创新领域发挥重要作用
汽车行业	利用大数据和物联网技术的无人驾驶汽车，已走入我们的生活
互联网行业	借助大数据技术，可以分析客户行为，进行商品推荐和有针对性的广告投放
餐饮行业	利用大数据实现餐饮O2O模式，彻底改变传统餐饮经营方式
电信行业	利用大数据技术实现客户离网分析，及时掌握客户离网倾向，出台客户挽留措施
能源行业	随着智能电网的发展，电力公司可以掌握海量的用户用电信息，利用大数据技术分析用户的用电模式，可以改进电网运行，合理地设计电力需求响应系统，确保电网运行安全
物流行业	利用大数据优化物流网络，提高物流效率，降低物流成本
城市管理领域	可以利用大数据实现智能交通、环保监测、城市规划和智能安防
生物医学领域	大数据可以助力流行病预测、智慧医疗、健康管理，还可以辅助解读DNA，了解更多的生命奥秘
体育和娱乐领域	大数据可助力球队科学训练、筛选影视创作题材
安全领域	政府可以利用大数据技术构建起强大的国家安全保障体系，企业可以利用大数据抵御网络攻击，警察可以借助大数据开展犯罪预防工作
个人生活	大数据可以应用于个人生活，利用与每个人相关联的“个人大数据”，分析个人生活行为习惯，从而提供更加周到的个性化服务

在企业层面，大数据正转化为实实在在的经济价值引擎。众多企业已将全业务链条植入数据驱动能力，尤其在营销领域表现尤为突出——eBay通过精密计算广告关键词的投入产出比，实现自2007年起广告费用骤降99%的同时，顶级卖家销售额占比攀升至32%；淘宝则通过挖掘用户浏览轨迹与购买历史的海量数据，构建个性化推荐系统，精准推送商品以提升转化率。更有企业依托大数据洞察市场趋势，制定经营战略，研发创新产品，快速抢占市场高地。如今，大数据已如经济血脉般渗透全球各产业领域，成为驱动增长的核心要素。

在政府治理方面，大数据正重塑决策范式与服务效能。通过构建数据驱动的决策模型，政府得以突破传统经验决策的局限，依托对社会经济人文数据的深度分析，科学制定宏观调控政策、战略规划与产业布局。在公共服务领域，利用大数据分析公众与企业行为模式，可精准优化资源配置；在城市治理中，数据技术推动管理方式从粗放向精细转型，提升了社会

治理水平。甚至在政治活动中，大数据也已悄然成为洞察民意、优化政策的重要工具。

在医疗健康领域，大数据同样展现了非凡价值。通过分析新生儿重症监护室数百万条监测数据，可捕捉到新生儿体温升高、心率加速等细微指标变化，提前识别潜在致命性感染风险——这种早期预警能力远超传统人工巡查。华盛顿中心医院就通过对患者多年匿名医疗记录的统计分析，发现患者出院后心理干预与降低再入院率之间存在显著关联，为优化诊疗方案提供了科学依据。

综观全局，大数据已深度融入社会运行各环节，持续重塑着我们的生活方式、认知模式与价值创造路径，成为数字时代不可替代的基础设施与创新源泉。

1.5 大数据的产业链

大数据产业是指一切与支撑大数据组织管理和价值发现相关的企业经济活动的集合。大数据产业链包括IT基础设施层、数据源层、数据管理层、数据分析层、数据平台层和数据应用层，具体内容见表1-3。

表1-3 大数据产业链的各个环节

产业链环节	包含内容
IT基础设施层	包括提供硬件、软件、网络等基础设施以及提供咨询、规划和系统集成服务的企业。例如，提供数据中心解决方案的IBM、惠普和戴尔等，提供存储解决方案的EMC，提供虚拟化管理软件的微软、思杰、SUN、Redhat等
数据源层	大数据生态圈里的数据提供者，生物（生物信息学领域的各类研究机构）大数据、交通（交通主管部门）大数据、医疗（各大医院、体检机构）大数据、政务（政府部门）大数据、电商（淘宝、天猫、苏宁云商、京东等）大数据、社交网络（微博、微信、人人网等）大数据、搜索引擎（百度、谷歌等）大数据等各种数据的来源
数据管理层	包括数据抽取、转换、存储和管理等服务的各类企业或产品，如分布式文件系统（如Hadoop的HDFS和谷歌的GFS）、ETL工具（Informatica、Data stage、Kettle等）、数据库和数据仓库（Oracle、MySQL、SQL Server、HBase、Green Plum等）
数据分析层	包括提供分布式计算、数据挖掘、统计分析等服务的各类企业或产品，如分布式计算框架 MapReduce、统计分析软件SPSS和SAS、数据挖掘工具Weka、数据可视化工具 Tableau、BI工具（MicroStrategy、Cognos、BO）等
数据平台层	包括提供数据分享平台、数据分析平台、数据租售平台等服务的企业或产品，如阿里巴巴、谷歌、中国电信、百度等
数据应用层	提供智能交通、智慧医疗、智能物流、智能电网等行业应用的企业、机构或政府部门，如交通主管部门、各大医疗机构、菜鸟网络、国家电网等

当前，我国大数据产业已形成五大核心区域协同发展格局，涵盖中西部、环渤海、珠三角、长三角及东北地区。在政府治理效能提升、工业数字化转型升级、金融科技创新、智慧医疗等关键领域，大数据技术正深度渗透并驱动应用创新。多地政府以“大数据产业园”为战略支点，加速培育本地产业生态——这类园区既是产业集聚载体，也是技术转化与项目孵化的核心平台，更成为企业从研发到产业化的关键过渡区。

在具体实践中，陕西西咸新区沣西新城在信息产业园中率先规划了全国首个以大数据处理与服务为特色的产业专区；重庆市通过《大数据行动计划》明确目标：打造2~3个产业示范园区，培育10家龙头企业、500家应用服务企业，引进培养千名高端人才，形成500亿元产业规模，建成国内大数据产业重要基地；天津市2013年发布的《滨海新区大数据行动方案》则提出，将天津打造为具有国际竞争力的大数据产业高地和数据资源枢纽；福建安溪龙门镇的中国国际信息技术（福建）产业园于2015年5月建成运营，作为福建省首个大数据产业园区，其以按国际最高标准建设第三方数据中心为核心，构建起涵盖云服务、数字金融、信息技术教育、国际交流等多功能于一体的绿色生态产业链，形成覆盖福建、辐射海峡西岸的国际一流信息技术产业集群。

这些产业园区通过集聚创新资源、完善配套服务、优化政策环境，正逐步释放大数据对区域经济增长的乘数效应，成为推动数字经济高质量发展的新引擎。

本章小结

在数据浪潮席卷全球的今天，我们正身处一个由0和1构建的数字新世界。信息洪流以指数级速度重塑着社会运行的底层逻辑，每个个体的行为轨迹都在被精准描绘，而智慧城市的心跳、市场需求的脉搏、社会趋势的走向，无不通过数据可视化技术跃然屏上。因此，本章首先讲述了大数据的背景及发展历程；其次，讨论了世界各国的大数据战略；最后，对大数据的行业应用及产业链进行了阐述。

思考题

1. 大数据如何影响你的生活？
2. 结合最近的新闻，思考大数据的未来发展。

| 第2章 | 洞悉大数据

学习目标及要求

1. 理解大数据的定义。
2. 掌握大数据的特征及结构。
3. 熟悉大数据标准化。
4. 了解大数据的价值及影响。

引导案例

超级商业零售连锁巨无霸沃尔玛公司（见图2-1）曾拥有世界上最大的数据仓库系统。为了能够准确了解顾客在其门店的购买习惯，沃尔玛对其顾客的购物行为进行了购物篮关联规则分析，从而知晓顾客经常一起购买的商品。沃尔玛庞大的数据仓库里集合了所有门店的详细原始交易数据，在这些原始交易数据的基础上，沃尔玛利用数据挖掘工具对这些数据进行了分析和挖掘，一个令人惊奇和意外的结果出现了：与尿不湿一起购买最多的商品竟是啤酒！这是数据挖掘技术对历史数据进行分析的结果，反映的是数据的内在规律。那么，这个结果符合现实情况吗？是否是一个有用的知识？是否有利用价值？



图2-1 沃尔玛公司

为了验证这一结果，沃尔玛派出市场调查人员和分析师对这一结果进行了调查分析。经过大量的实际调查和分析，他们揭示了一个隐藏在“纸尿裤与啤酒”背后的美国消费者的行为模式：在美国，到超市去买婴儿纸尿裤是一些年轻的父亲下班后的日常工作，而他们中有30%~40%的人同时也会为自己买一些啤酒。产生这一现象的原因是：美国的太

妈妈们常叮嘱她们的丈夫不要忘了下班后为小孩买纸尿裤，而丈夫们在买纸尿裤后又随手带回了他们喜欢的啤酒。还有一种情况是丈夫们在买啤酒时突然记起他们的责任，又去买了纸尿裤。既然纸尿裤与啤酒一起被购买的机会很多，沃尔玛就将他们所有门店里的纸尿裤与啤酒并排摆放在一起，结果纸尿裤与啤酒的销售量双双增长。按常规思维，纸尿裤与啤酒风马牛不相及，若不是借助数据挖掘技术对大量交易数据进行挖掘分析，沃尔玛不可能发现这一有价值的规律。

2.1 大数据的定义

对于大数据的定义，研究机构加特纳提出：“大数据是需要新处理模式支撑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。”

世界级领先的全球管理咨询公司麦肯锡认为：“大数据指的是所涉及的数据集规模已经超过了传统数据库软件获取、存储、管理和分析的能力。这是一个被故意设计成主观性的定义，并且是一个关于多大的数据集才能被认为是大数据的可变定义，即并不定义大于一个特定数字的TB才叫大数据。因为随着技术的不断发展，符合大数据标准的数据集容量也会增长；并且定义随不同的行业也有变化，这依赖于在一个特定行业通常使用何种软件和数据集有多大。因此，大数据在今天不同行业中的范围可以从几十TB到几PB。”

所谓大数据，狭义上可定义为：用现有的一般技术难以管理的大量数据的集合。这实际上是指用目前在企业数据库占据主流地位的关系型数据库无法进行管理的、具有复杂结构的数据。或者也可以说，大数据是指因数据量的增大，数据查询响应时间超出允许范围的数据。

广义层面上的大数据是一个综合性概念，它包括因具备规模大、类型多、价值密度低、处理速度快等特征而难以进行管理的数据，对这些数据进行存储、处理、分析的技术，以及能够通过分析这些数据获得实用意义和观点的人才和组织。其中，“存储、处理、分析的技术”指的是用于大规模数据分布式处理的框架Hadoop、具备良好扩展性的NoSQL数据库，以及机器学习和统计分析等；“能够通过分析这些数据获得实用意义和观点的人才和组织”指的是目前十分紧俏的“数据科学家”这类人才以及能够对大数据进行有效运用的组织。狭义广义大数据构成示意如图2-2。

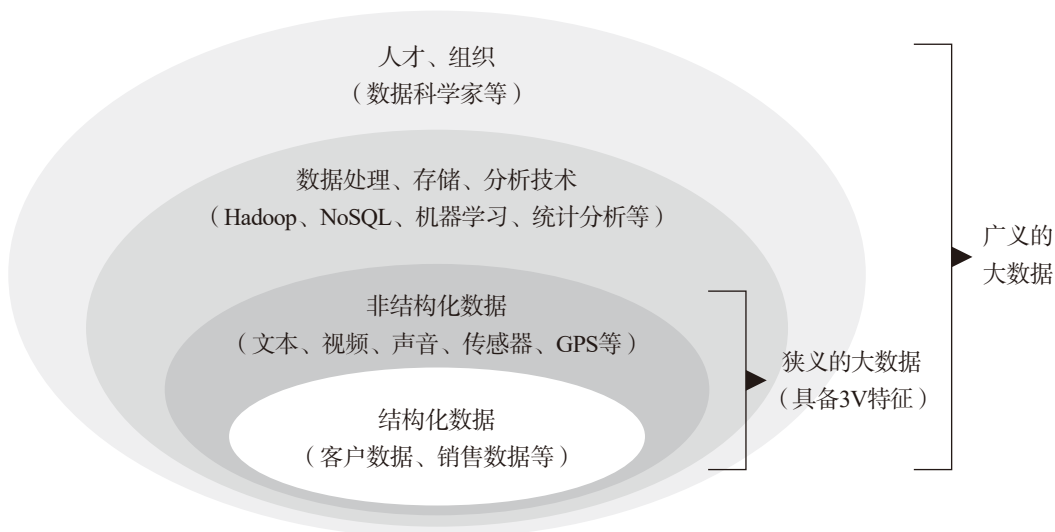


图2-2 狭义与广义大数据构成示意

2.2 大数据的特征

关于大数据的特征，IBM提出“3V”的说法，即认为大数据具备数据量大（volume）、数据类型繁多（variety）和处理速度快（velocity）3个特征。随着大数据的发展，大家比较认可的是关于大数据的“4V”说法，即大数据具有数据量大（volume）、数据类型多（variety）、处理速度快（velocity）和价值密度低（value）4个特征（见图2-3）。

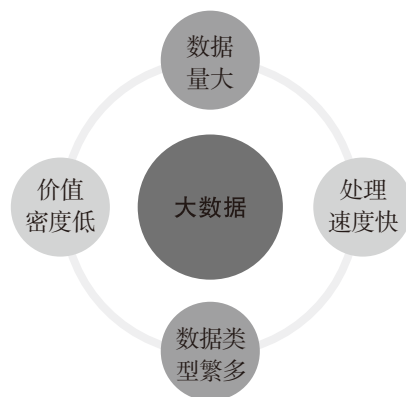


图2-3 大数据的特征

1. 数据量大

从数据量的角度来看，大数据泛指无法在可容忍的时间内用传统信息技术和软、硬件工具对其进行获取、管理和处理的巨量数据集合，需要可伸缩的计算体系结构以支持其存储、处理和分析。大数据的“体量大”指数据量级已突破传统IT系统的处理极限，达到PB（ 10^{15} 字节）、EB（ 10^{18} 字节）甚至ZB（ 10^{21} 字节）级别。例如，全球互联网每天产生超50亿GB的数据，相当于5000万部4K电影；物联网设备每秒生成的海量传感器数据、社交媒体用户每日上传的数百亿条图文内容，均需依赖分布式存储架构（如HDFS）和并行计算框架（如Spark、Flink）实现高效处理。数据处理的挑战在于如何通过数据压缩、冷热数据分层存储（如对象存储+缓存）及弹性云资源

（如AWS S3、阿里云OSS）平衡成本与性能，同时避免“数据沼泽”现象。大数据存储单位是一种计量单位，指在某一领域以一个特定量或标准为一个记录（计数）点，再以此点的某个倍数再去定义另一个点，而这个点的代名词就是计数单位或存储单位。数据存储单位之间的换算关系如表2-1所示。

表2-1 数据存储单位之间的换算关系

单位	换算关系
Byte（字节）	1 Byte=8 bit
KB（Kilobyte，千字节）	1 KB=1024 Byte
MB（Megabyte，兆字节）	1 MB=1024 KB
GB（Gigabyte，吉字节）	1 GB=1024 MB
TB（Trillionbyte，太字节）	1 TB=1024 GB
PB（Petabyte，拍字节）	1 PB=1024 TB
EB（Exabyte，艾字节）	1 EB=1024 PB
ZB（Zettabyte，泽字节）	1 ZB=1024 EB

随着数据量的不断增长，数据所蕴含的价值会从量变发展到质变。举例来说，受到照相技术的制约，早期我们只能每分钟拍1张照片，随着照相设备的不断改进，处理速度越来越快，发展到后来，就可以每秒拍1张照片，而当有一天发展到每秒可以拍10张照片以后，就产生了电影。当照片数量的增长带来质变时，照片就发展成了电影。同样的量变到质变，也会发生在数据量的增长过程之中。

2. 数据类型繁多

大数据的数据来源众多，科学研究、企业应用和Web应用等都在源源不断地生成新的类型繁多的数据。生物大数据、交通大数据、医疗大数据、电信大数据、电力大数据、金融大数据等，都呈现出“井喷式”增长，所涉及的数据量巨大，已经从TB级别跃升到PB级别。各行各业每时每刻都在生成各种不同类型的数据。

3. 处理速度快

大数据时代的数据产生速度非常快。数据生成、传输和处理的实时性要求，催生了大数据的“速度”维度。例如，金融高频交易系统要在毫秒级完成订单匹配与风险控制；智能监控摄像头要每秒处理数十帧视频流以检测异常行为（如入侵、火灾）；工业物联网中的设备状态监测要实时采集传感器数据并预警故障。技术挑战包括流计算引擎（如Kafka Streams、Apache Flink）、边缘计算（减少中心节点延迟）及内存数据库（如Redis、SAP

HANA)的应用,同时要考虑数据分片、负载均衡和容错机制,以保证高可用性。

4. 价值密度低

大数据虽然看起来很“美”,但是其数据价值密度远远低于传统关系数据库中的数据价值密度。在大数据时代,很多有价值的信息都是分散在海量数据中的。以小区监控摄像头为例(见图2-4),如果没有意外事件发生,那么连续不断产生的数据都是没有任何价值的,而当发生偷盗等意外情况时,也只有记录了事件过程的那一小段视频是有价值的。但是,为了能够获得发生偷盗等意外情况时的那一段有价值的视频,我们不得不投入大量资金购买监控设备、网络设备、存储设备,耗费大量的电能和存储空间,以保存摄像头连续不断产生的监控数据。



图2-4 小区监控摄像头

此外,有学者提出增加数据的准确性和可信赖度(veracity)这一特征,构成大数据的“5V”特性,即大数据的准确性和可信赖度高,主要指其全面、准确、细致和便利地反映现实的方式与程度。一般来说,我们将大数据的处理工作流程分为数据清理、数据收集、数据存储、数据分析和数据应用五步,其中数据清理是第一步也是极其重要的一步,它对海量原始数据中的缺失值、噪声以及人工失误所导致的异常数据等进行清理,通过一系列措施最终获得高质量的数据。在此基础上,数据收集和数据处理分析生成的结果才具有更高的准确性和可信赖度。基于真实有效的大数据,管理者可以进行有效的大数据分析,为经营管理决策和未来趋势预测提供有效的依据和保障。

2.3 大数据的类型

从数据的结构来说,一般可将大数据分为结构化数据、非结构化数据和半结构化数据(见图2-5)。

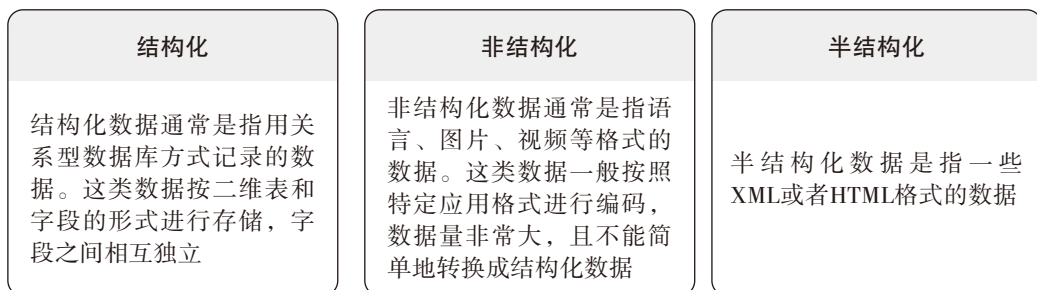


图2-5 基于结构的大数据分类

结构化数据，指可存入数据库软件、以二维表形式表达的数据。这类数据具有固定格式与有限长度，主要通过关系型数据库管理，是多数技术应用的基础支撑。

非结构化数据，指无法以二维表形式表达的数据，如图片、音频、视频等，在互联网内容中占比显著。据IDC调研，企业运营中80%的数据属于此类，其无固定结构、形式灵活且蕴含丰富信息，掌握这类数据的处理技术对大数据分析应用至关重要。

半结构化数据，指介于结构化与非结构化之间，格式不固定却具备一定结构性（如XML、JSON文件）的数据。这类数据比纯文本更灵活，又比关系型数据库更自由。

在企业大数据架构中，数据需经历从生产到应用的全流程闭环，且各环节紧密关联。通常在算法预测前，数据需完成以下处理。

（1）数据采集：将应用程序产生的业务数据、日志文件等同步至大数据系统，完成原始数据归集。

（2）数据存储：海量数据通过分布式存储系统实现有序存放，以确保高效查询与调用。

（3）数据处理：原始数据经清洗、整合、转换等环节方可应用，包含离线批量处理与实时在线分析两种模式，前者处理历史数据，后者响应实时需求。

（4）数据应用：处理后的数据可对外输出价值，如生成可视化报表辅助决策、支撑交互式数据分析、为推荐系统提供模型训练素材等，最终实现数据到业务价值的转化。

这种分类与流程设计，既体现了数据结构的多样性特征，又构建了从原始数据到业务应用的完整链路，为大数据驱动的企业数字化转型提供基础框架。

2.4 大数据标准化

大数据标准化是通过制定统一的技术规范、数据格式、质量评估及安全准则，实现数据全生命周期（采集、存储、处理、分析、共享、应用）的规范化管理。其核心目标是解决“数据孤岛”问题，提升数据一致性、可比性和可操作性，支撑跨领域、跨系统的数据融合与价值挖掘。例如，通过统一数据编码、元数据定义及接口协议，确保不同来源的医疗影像、基因数据、电子病历无缝整合，为精准医疗提供基础。

2023年发布的《大数据标准化白皮书（2023版）》提出了大数据标准体系的7个类别标准，具体如下。

基础标准：定义术语、参考架构（如ISO/IEC 19845《大数据参考架构》），为行业共识提供基准。

数据标准：规范数据格式（如JSON、XML）、元数据管理、主数据定义及数据字典，确保数据可解析、可追溯。

技术标准：涵盖数据采集（传感器协议）、存储（分布式文件系统）、处理（流计算引擎）、分析（机器学习算法）及可视化（交互式图表）等全流程技术规范。

管理标准：包括数据治理框架、质量评估模型（如DQM）、资产管理及生命周期管理，确保数据可控、可信。

安全标准：聚焦数据加密、访问控制、隐私保护（如GDPR合规）、区块链存证及安全审计，防范数据泄露与滥用。

行业应用标准：针对金融、医疗、交通、工业等垂直领域制定专用标准，如金融反欺诈模型、工业物联网数据采集规范。

平台与工具标准：规范大数据平台（如Hadoop、Spark）及工具（如ETL工具FineDataLink）的性能指标、接口兼容性测试方法。

2.5 大数据的价值

大数据的价值体现在多个方面，具体如下。

1. 企业战略决策的智能引擎

大数据通过全链路数据采集与实时分析，助力企业构建“数据驱动型决策体系”。例如，零售企业可利用用户画像与消费轨迹分析实现精准选品、动态定价及库存优化，将传统经验决策升级为基于数据洞察的科学决策，提升市场响应速度与资源配置效率，形成“预测—决策—验证”的闭环迭代机制。

2. 政府治理现代化的数字基石

在政务场景中，大数据推动“数字政府”向“智慧政府”转型。政府通过整合人口、交通、环境等多维度数据，实现跨部门协同治理与智能决策。例如，城市大脑系统通过实时交通流量分析优化信号灯配时，提升通行效率；在公共卫生事件中，大数据追踪可快速锁定传播链，支撑精准防控策略，凸显“数据治域”的社会治理新范式。

3. 个性化服务的价值共创

对个人用户而言，大数据重构了“人—信息—服务”的连接方式。智能推荐系统通过分析用户行为、偏好及情境数据，提供千人千面的内容推送与服务匹配，如音乐平台的个性化歌单、电商平台的“猜你喜欢”。同时，隐私计算技术（如联邦学习、差分隐私）在保障数据安全的前提下，实现“数据可用不可见”的价值挖掘，平衡个性化服务与隐私保护需求。

4. 垂直行业的变革催化剂

在金融领域，大数据风控模型通过多源异构数据融合（如征信、交易、社交数据）实现信贷风险精准评估，降低坏账率；在医疗领域，电子病历与基因组数据结合推动精准医疗，如癌症早筛与个性化用药方案；零售与电商通过用户行为数据实现动态定价与智能营销，提升转化率与客户黏性，形成“数据—算法—业务”的深度融合。

5. 科学研究的范式革命

大数据打破了传统“假设驱动”的研究局限，催生“数据密集型科学发现”。在天文学领域，海量天文数据支持黑洞成像与星系演化模拟；在生物学领域，单细胞测序数据推动细胞图谱构建与疾病机制解析；在社会科学中，社交媒体数据与网络行为分析揭示群体行为规律，为政策制定提供实证依据，形成“数据—模型—验证”的科研新路径。

6. 国家安全与反恐的智能防线

在国防与安全领域，大数据通过多源情报融合与模式识别，构建自动化预警系统。例如，通过分析网络流量、地理信息与社交媒体数据，识别潜在安全威胁；在反恐场景中，大数据追踪可快速锁定异常行为模式，支撑精准打击与预防策略，提升国家安全防御的智能化水平。

7. 新兴产业的创新驱动力

作为“数字新基建”的核心要素，大数据与云计算、人工智能、物联网形成技术协同，催生新产业与新业态。例如，工业互联网通过设备数据采集与分析实现生产流程优化；自动驾驶依赖高精地图与实时路况数据实现路径规划；智能城市通过能源、交通、环境数据融合推动绿色低碳发展，形成“数据—技术—产业”的创新生态。

8. 文化遗产与社会进步的数字桥梁

大数据跨越时空限制，推动文化资源的数字化保存与全球化共享。例如，数字博物馆通过3D扫描与虚拟现实技术实现文物“活态传承”；跨语言数据平台促进不同文化背景下的知识交流与理解；在文化遗产保护中，大数据支持非遗项目的数字化记录与传播，促进文化多样性的保护与传承，最终形成“数据—文化—价值”的良性循环，推动社会进步与文明互鉴。

2.6 大数据的影响

大数据对科学研究、社会发展、就业市场和人才培养都具有重要而深远的影响。在科学研究方面，大数据使得人类科学研究在经历了实验、理论、计算三种范式之后，迎来了第四种范式——数据。在社会发展方面，大数据决策逐渐成为一种新的决策方式，大数据应用有力地促进了信息技术与各行业的深度融合，大数据开发大大推动了新技术和新应用的不断涌现；在就业市场方面，大数据的兴起使得数据科学家成为热门职业；在人才培养方面，大数据的兴起将在很大程度上改变中国高校信息技术相关专业的现有教学和科研体制。

2.6.1 大数据对科学研究的影响

大数据最根本的价值在于为人类提供了认识复杂系统的新思维和新手段。图灵奖获得者、著名数据库专家吉姆·格雷博士观察并总结，人类自古以来在科学研究上先后历经实验科学、理论科学、计算科学和数据密集型科学4种范式（见图2-6）。

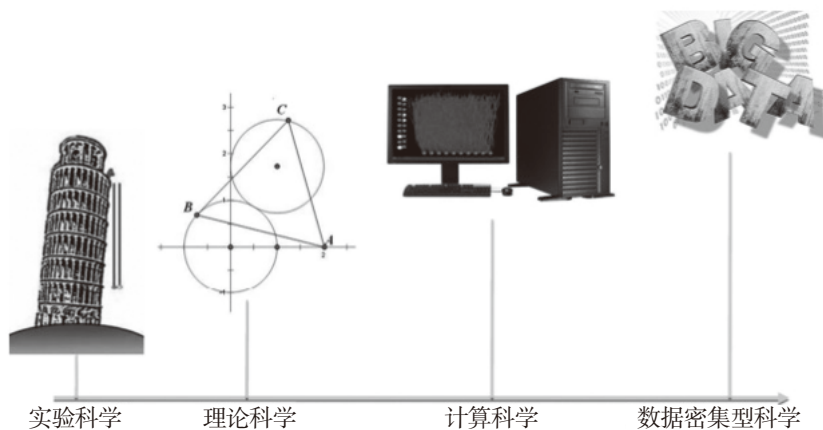


图2-6 科学研究的4种范式

1. 实验科学：以观测验证为核心的探索起点

在科学萌芽期，人类主要通过直接实验与观测来验证自然现象。1590年，伽利略在比萨斜塔完成的“自由落体实验”便是典型代表：不同重量的铁球同时落地，直接推翻了亚里士多德“重物下落更快”的千年谬误。这一研究范式，强调通过可重复的实验操作获取实证数据，厘清现象与结果的因果关系，是人类科学认知的原始积累阶段。

2. 理论科学：以数学建模为特征的体系构建

随着科学的进步，实验数据的局限性逐渐显现，人类开始运用数学、几何等抽象工具构建理论体系。牛顿经典力学三定律的提出堪称里程碑：将物体运动规律抽象为普适定律，不仅解释了苹果落地、行星运动等具体现象，更建立起完整的力学理论框架。这一研究范式将实验观察提升为逻辑严密的数学推导，推动科学从解释具体现象走向总结普遍规律。

3. 计算科学：以算法模拟为手段的复杂问题求解

1946年，ENIAC计算机的诞生开启了计算科学时代。借助计算机的强大算力，科学家能够对复杂系统进行高精度模拟，小至天气预报的流体动力学计算，大至量子力学的薛定谔方程求解。这一范式的核心特征是“先理论后计算”：先建立数学模型，再通过算法编程和计算机运算验证或优化理论。计算机的高速度、大存储特性使科学探索突破了人力计算的局限，解决了大量传统方法难以处理的复杂问题。

4. 数据密集型科学：以数据驱动为本质的发现革命

进入大数据时代，科学范式迎来根本性变革。物联网、云计算等技术催生了海量数据的积累，使得科学研究转向“数据驱动”模式。与计算科学“理论—数据—验证”的路径不同，数据密集型科学遵循“数据—分析—发现”的逆向逻辑：通过机器学习、数据挖掘等技术，直接从PB级甚至EB级数据中提取模式、发现未知规律，甚至直接生成理论。例如，天文学通过射电望远镜阵列数据直接发现快速射电暴的周期性规律，生物学通过单细胞测序数据重构细胞发育轨迹，真正实现了“让数据自己说话”的科学发现新范式。

这四种范式并非简单的替代关系，而是层层递进、相互补充的探索阶梯。从实验的实证基础到理论的抽象升华，从计算的算法支撑到数据驱动的智能发现，每个阶段都拓展了人类认知自然的维度，共同推动着科学从“已知”向“未知”持续突破。

2.6.2 大数据对社会发展的影响

大数据正以前所未有的广度与深度重塑社会运行逻辑，其社会价值可从四大维度诠释，具体如下。

1. 决策范式革新：从经验驱动到数据智能

传统决策依赖数据仓库与商业智能工具，虽能整合历史数据实现周期性分析，但受限于关系型数据库的容量与结构化数据边界。如今，大数据决策突破这一局限，可处理非

结构化、多模态的海量数据，形成“全量数据洞察—实时规则触发—战略战术协同”的新型决策体系。例如，政府部门通过整合论坛、博客、社交媒体等多源异构数据开展舆情分析，不仅能捕捉显性民意走向，更能通过关联模式挖掘揭示隐性社会情绪与潜在风险，为突发事件应对提供精准情报支撑，实现从“被动响应”到“主动预判”的治理转型。

2. 治理能力升级：数据赋能的国家治理现代化

大数据已成为提升国家治理能力的核心杠杆。在政治层面，跨部门数据融合可揭示传统方法难以捕捉的政策效应关联网络，如财政政策与民生指标的动态响应关系；在经济层面，大数据与实体经济深度融合催生智能制造、数字农业等新业态，推动传统产业提质增效；在社会层面，打通“政府—公共服务”数据壁垒后，行政审批实现“数据多跑路、群众少跑腿”，公共服务效率显著提升，群众获得感与幸福感持续增强，构建起“数据治理”的现代治理新范式。

3. 产业融合加速：信息技术与实体经济的深度耦合

大数据正引发全行业变革浪潮，未来十年几乎所有行业的业务逻辑都将被重构。在物流领域，企业通过路况、天气、货量等多维数据实时优化配送路线，降低运输成本；在金融领域，投资者基于海量市场数据构建量化模型，实现更优资产配置；在零售领域，用户画像与消费轨迹分析助力精准营销与目标客群定位；在能源领域，电网负荷预测与智能调度确保电力供需平衡与系统安全。这种深度融合使每个行业都衍生出新的业务方向与发展维度，推动社会生产生活方式发生根本性变革。

4. 技术生态进化：需求驱动的创新引擎

大数据应用需求持续催生技术创新，形成“应用—技术—应用”的良性循环。车联网数据的深度挖掘就是典型例证：传统车险仅依据出险次数进行简单分级定价，而基于车辆行驶数据、驾驶习惯、路况信息的多维分析可构建更精细的客户风险模型，实现“一人一价”的个性化保费方案。这种基于大数据的商业模式创新，不仅重塑了保险业竞争格局，更推动了传感器技术、边缘计算、隐私计算等配套技术的突破性发展，最终实现“数据能量释放—技术迭代升级—应用场景拓展”的螺旋式上升，使原本依赖人工经验判断的领域逐步转向智能化决策，释放出巨大的社会生产力。

2.6.3 大数据对就业市场的影响

大数据的兴起使得数据科学家成为热门职业。2010年，在高科技劳动力市场上还很难见到数据科学家的头衔，但此后，数据科学家逐渐发展为市场上最热门的职位之一，具有广阔的发展前景，并代表着未来的发展方向。互联网、零售、金融等各类企业都在积极争夺大数据人才，数据科学家成为大数据时代最紧缺的人才。据麦肯锡报告统计，到2018年，仅美国本土就缺少14万~19万具备数据深入分析能力的专业人才，而能够通过分析大数据支撑企业做出有效决策的数据管理人员和分析师，缺口更是高达约150万人。国内大数据专家估算，目前国内的大数据人才缺口达到130万人。

现阶段，我国大数据应用仍以结构化数据分析为主，尚未进入依托半结构化和非结构化数据分析挖掘新兴市场空间的阶段。但是，大数据中包含了大量的非结构化数据，未来将会产生大量针对非结构化数据进行分析的市场需求，未来中国市场对掌握大数据分析专业技能的数据科学家的需求会逐年递增。尽管少数人认为，未来更多的数据可采用自动化处理，会逐步降低对数据科学家的需求，但是更多业内人士认为，随着数据科学家给企业所带来的商业价值的日益体现，市场对数据科学家的需求会日益增加。

2.6.4 大数据对人才培养的影响

大数据的兴起将在很大程度上改变中国高校信息技术相关专业的现有教学和科研体制。一方面，数据科学家是需要掌握统计学、数学、机器学习、数据可视化、编程等多方面知识的复合型人才。在中国高校现有的学科和专业设置中，上述专业知识分布在数学、统计学和计算机等多个学科中，通常任何一个学科都只能培养某个方向的专业人才，无法培养全面掌握数据科学相关知识的复合型人才。另一方面，数据科学家需要大数据应用实战环境，在真正的大数据环境中不断学习、实践并融会贯通，将自身专业背景与所在行业业务需求进行深度融合，从数据中发现有价值的信息，但是，目前大多高校还不具备这种培养环境，不仅缺乏大规模基础数据，也缺乏对领域业务需求的理解。鉴于上述两个原因，目前国内的数据科学人才并不是由高校培养的，而主要是在企业实际应用环境中通过边工作边学习的方式不断成长起来的，其中，互联网领域集中了大多数数据科学人才。

为破局人才缺口，许多高校已启动多维改革：2014年起，中国科学院大学开设大数据技术与应用专业，清华大学成立数据科学研究院推出交叉学科硕士项目，复旦大学大数据学院融合数学、计算机、医学、经济学等多学科培养复合人才；2016年，北京大学等三校率先获批“数据科学与大数据技术”本科专业，至2020年全国超千所高校设立相关专业。

此外，许多高校还采取“引进来+走出去”双轨策略：“引进来”即引入企业真实数据与实战人才共建教学环境，“走出去”则推动学生赴企业实践、教师参与产学研项目，实现理论与实践的深度融合。在课程体系上，许多高校打破学科壁垒，设置跨院系组合课程，由计算机、数学、统计学等院系教师联合授课，培养学生系统掌握数据分析、商业洞察、自然语言处理等核心能力，并强化实践创新素养。

这种人才培养模式的革新，不仅旨在填补当前百万级人才缺口，更着眼于构建数据驱动的未来社会所需的新型智识体系，为经济社会发展注入持久动力。

本章小结

人类已经步入大数据时代，我们的生活被数据所“环绕”，并因数据而发生了深刻改变。作为大数据时代的公民，我们应该接近数据，了解数据，并利用好数据。因此，本章首先讲述了大数据的概念，讨论了大数据的“4V”特性；其次，对大数据的标准化政策进行了举例；最后，阐述了大数据的价值及其对科学研究、社会发展、就业市场和人才培养的影响。

思考题

1. 请阐述大数据的类型。
2. 请阐述大数据标准化的分类。
3. 请阐述大数据的“4V”特性。
4. 请阐述大数据对科学研究有什么影响。
5. 请举例说明大数据的价值。