

高等学校大数据技术与应用规划教材○

大数据 技术概论

陈 明 编著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

高等学校大数据技术与应用规划教材

大数据技术概论

陈 明 编著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

大数据技术是一个面向实际应用的技术。从大数据中获取有价值信息是大数据技术的精髓。本书概括性介绍了数据科学与大数据技术的主要内容。全书分为9章,主要包括概述、大数据处理平台、大数据获取与存储管理技术、大数据抽取技术、大数据清洗技术、大数据去噪与标准化、大数据约简与集成技术、大数据分析 with 挖掘技术、大数据分析结果解释与展现。

本书在内容上,注重基本概念、基本方法介绍,实例丰富、语言精练、逻辑层次清晰,适合作为大学“数据科学与大数据技术”专业和相近专业的教材,也可以作为科技人员的参考书。

图书在版编目(CIP)数据

大数据技术概论 / 陈明编著. —北京: 中国铁道出版社, 2019. 1

高等学校大数据技术与应用规划教材
ISBN 978-7-113-24818-5

I. ①大… II. ①陈… III. ①数据处理 - 高等学校 - 教材 IV. ① TP274

中国版本图书馆 CIP 数据核字 (2018) 第 178257 号

书 名: 大数据技术概论
作 者: 陈 明 编著

策 划: 秦绪好
责任编辑: 秦绪好
封面设计: 郑春鹏
责任校对: 张玉华
责任印制: 郭向伟

读者热线: (010) 63550836

出版发行: 中国铁道出版社 (100054, 北京市西城区右安门西街 8 号)

网 址: <http://www.tdpress.com/51eds/>

印 刷: 北京铭成印刷有限公司

版 次: 2019 年 1 月第 1 版 2019 年 1 月第 1 次印刷

开 本: 850 mm×1 168 mm 1/16 印张: 17.25 字数: 456 千

书 号: ISBN 978-7-113-24818-5

定 价: 52.00 元

版权所有 侵权必究

凡购买铁道版图书, 如有印制质量问题, 请与本社教材图书营销部联系调换。电话: (010) 63550836

打击盗版举报电话: (010) 51873659

大数据技术与应用展现出锐不可当的强大生命力，科学界与企业界对其寄予厚望。

大数据成为继 20 世纪末、21 世纪初互联网蓬勃发展以来的又一轮 IT 工业革命。

大数据技术是指经过数据获取、清洗、集成、挖掘、分析与结果解释，从各种类型的巨量数据中快速获得有价值信息的全部技术。大数据技术的精髓是从大数据中产生新见解、识别复杂关系和做出越来越精准的预测。

大数据技术是现代科学与技术发展，尤其是计算机科学技术发展的重要成果和结晶，是科学发展史的又一个新的里程碑。大数据的出现对计算机许多领域提出了挑战与冲击，推动了计算机科学技术的发展。

大数据技术的出现凝集了多学科的研究成果，是一门多学科的交叉融合技术，随着科学技术的发展，大数据技术发展更为迅速，应用更为深入与广泛，并凸显其巨大潜力和应用价值。

本书系统地介绍了大数据技术的核心内容，对大数据处理周期的各部分的模型和方法做了概括性介绍，而且基于应用的角度介绍了当下流行的 Hadoop、Storm 和 Spark 大数据处理平台，为将大数据处理周期中的处理方法在这些平台上实现建立了基础。本书主要内容说明如下。

第 1 章为概述，主要包括数据科学、大数据的生态环境、大数据的概念、大数据的性质、大数据处理周期和科学研究范式；第 2 章为大数据处理平台，主要包括 Hadoop 大数据处理平台、Storm 大数据处理平台和 Spark 大数据处理平台；第 3 章为大数据获取与存储管理技术，主要包括大数据获取、领域数据、网站数据、网络爬虫、大数据存储、大数据的存储管理技术、NewSQL 和 NoSQL、分布式文件系统、虚拟存储技术和云存储技术；第 4 章为大数据抽取技术，主要包括大数据抽取技术概述、增量数据抽取技术、

非结构化数据抽取和基于 Hadoop 平台的数据抽取；第 5 章为大数据清洗技术，主要包括数据质量与数据清洗、不完整数据清洗、异常数据清洗、重复数据清洗、文本清洗和数据清洗的实现；第 6 章为大数据去噪与标准化，主要包括基本的数据转换方法、数据平滑技术、数据规范化和数据泛化；第 7 章为大数据约简与集成技术，主要包括数据约简概述、特征约简、样本约简、数据立方体聚集、维约简属性子集选择算法、数据压缩、数值约简、数据集成的概念与相关问题、数据迁移、数据集成模式、数据集成系统和数据聚类集成；第 8 章为大数据分析挖掘技术，主要包括大数据分析概述、统计分析方法、数据挖掘理论基础、关联规则挖掘、分类方法、聚类方法、序列模式挖掘、非结构化文本数据挖掘和基于 MapReduce 的分析与挖掘实例；第 9 章为大数据分析结果解释与展现，主要包括数据分析结果解释、数据的基本展现方式、大数据可视化、大数据可视分析和数据可视化实现。

本书在结构上为积木状，各章内容均为独立、注重概念性与方法性论述。出于篇幅考虑，书中所提及理论结果没有给出证明，如需要可以查阅相关文献。由于作者水平有限，书中不足之处在所难免，敬请读者批评指正。



2018 年 10 月

第 1 章 概述	1
1.1 数据科学	2
1.1.1 数据科学的产生与发展	2
1.1.2 数据科学的相关术语	2
1.1.3 数据科学的主要内容	3
1.1.4 数据科学的研究过程与体系 框架	4
1.1.5 数据科学、数据技术与数据 工程	6
1.1.6 大数据问题	6
1.2 大数据的生态环境	7
1.2.1 互联网世界	7
1.2.2 物理世界	9
1.3 大数据的概念	9
1.3.1 数据容量	10
1.3.2 数据类型	11
1.3.3 价值密度	11
1.3.4 速度	11
1.3.5 真实性	11
1.4 大数据的性质	11
1.4.1 非结构性	12
1.4.2 不完备性	13
1.4.3 时效性	13
1.4.4 安全性	13
1.4.5 可靠性	13
1.5 大数据处理周期	13

1.5.1 大数据处理全过程	14
1.5.2 大数据技术的特征	15
1.5.3 大数据的一些热点技术	17
1.6 科学研究范式	18
1.6.1 科学研究范式的产生与发展	18
1.6.2 数据密集型科学研究第四范式	19
小结	23
第 2 章 大数据处理平台	24
2.1 Hadoop 大数据处理平台	24
2.1.1 离线计算概述	25
2.1.2 MapReduce 分布编程模型	26
2.1.3 基于 Hadoop 的分布计算	27
2.1.4 MapReduce 程序设计分析	31
2.1.5 Hadoop 环境部署与程序运行	36
2.2 Storm 大数据处理平台	42
2.2.1 流式数据的概念与特征	43
2.2.2 大数据的计算模式	45
2.2.3 流式数据处理技术的应用	46
2.2.4 流式计算的系统架构	50
2.2.5 Storm 流处理过程	52
2.2.6 单词计数 topology	56
2.3 Spark 大数据处理平台	58
2.3.1 交互式处理系统的问题	58
2.3.2 Scala 编程语言简介	58
2.3.3 Spark 的主要特点	59

2.3.4 软件栈	60	3.4.7 分布式爬虫的系统结构	82
2.3.5 核心概念	61	3.5 大数据存储	84
2.3.6 弹性分布式数据集 RDD	62	3.5.1 大数据存储模型	84
2.3.7 单词计数实例分析	66	3.5.2 大数据存储问题	84
小结	67	3.5.3 大数据存储方式	85
第3章 大数据获取与存储管理技术	68	3.6 大数据存储管理技术	86
3.1 大数据获取	69	3.6.1 数据容量问题	86
3.1.1 大数据获取的挑战	69	3.6.2 大图数据	87
3.1.2 传统的数据获取与大数据获取的 区别	69	3.6.3 数据存储管理	88
3.2 领域数据	70	3.7 NoSQL 和 NewSQL	89
3.2.1 文本数据	70	3.7.1 NoSQL	90
3.2.2 语音数据	71	3.7.2 NewSQL 数据库	93
3.2.3 图片数据	71	3.7.3 不同数据库架构混合应用模式	93
3.2.4 摄像头视频数据	71	3.8 分布式文件系统	94
3.2.5 图像数字化数据	71	3.8.1 评价指标	94
3.2.6 图形数字化数据	71	3.8.2 HDFS 文件系统	95
3.2.7 空间数据	72	3.8.3 NFS 文件系统	96
3.3 网站数据	72	3.9 虚拟存储技术	97
3.3.1 网站内部数据	73	3.9.1 虚拟存储的特点	98
3.3.2 网站外部数据	73	3.9.2 虚拟存储的应用	98
3.3.3 移动网站数据	73	3.10 云存储技术	99
3.4 网络爬虫	73	3.10.1 云存储原理	99
3.4.1 网络爬虫的工作过程	74	3.10.2 网络结构	99
3.4.2 通用网络爬虫	74	3.10.3 云的分类	99
3.4.3 聚焦网络爬虫	77	小结	100
3.4.4 数据抓取目标的定义	79	第4章 大数据抽取技术	101
3.4.5 网页分析算法	80	4.1 大数据抽取技术概述	101
3.4.6 更新策略	81	4.1.1 数据抽取的定义	102
		4.1.2 数据映射与数据迁移	102

4.1.3 数据抽取程序	103	5.2.1 基本方法	121
4.1.4 抽取、转换和加载	103	5.2.2 基于 k -NN 近邻缺失数据的 填充算法.....	123
4.1.5 数据抽取方式	104	5.3 异常数据清洗	124
4.2 增量数据抽取技术	104	5.3.1 异常值产生的原因	125
4.2.1 增量抽取的特点与策略	105	5.3.2 统计方法	125
4.2.2 基于触发器的增量抽取 方式.....	105	5.3.3 基于邻近度的离群点检测	126
4.2.3 基于时间戳的增量抽取 方式.....	105	5.4 重复数据清洗	127
4.2.4 全表删除插入方式	106	5.4.1 使用字段相似度识别重复值 算法.....	127
4.2.5 全表比对抽取方式	106	5.4.2 搜索引擎快速去重算法	128
4.2.6 日志表方式	107	5.5 文本清洗	128
4.2.7 系统日志分析方式	107	5.5.1 字符串匹配算法	129
4.2.8 各种数据抽取机制的比较 与分析.....	107	5.5.2 文本相似度度量	131
4.3 非结构化数据抽取	109	5.5.3 文档去重算法	135
4.3.1 非结构化数据类型	109	5.6 数据清洗的实现	135
4.3.2 非结构化数据模型	109	5.6.1 数据清洗的步骤	135
4.3.3 非结构化数据组织	111	5.6.2 基于 MapReduce 的大数据 去重.....	136
4.3.4 纯文本抽取通用程序库	112	小结	138
4.4 基于 Hadoop 平台的数据抽取	113	第 6 章 大数据去噪与标准化.....	139
小结	114	6.1 基本的数据转换方法	139
第 5 章 大数据清洗技术	115	6.1.1 对数转换	139
5.1 数据质量与数据清洗	115	6.1.2 平方根转换	140
5.1.1 数据质量	116	6.1.3 平方转换	140
5.1.2 数据质量提高技术	118	6.1.4 倒数变换	140
5.1.3 数据清洗算法的标准	120	6.2 数据平滑技术	140
5.1.4 数据清洗的过程与模型	120	6.2.1 移动平均法	141
5.2 不完整数据清洗	121	6.2.2 指数平滑法	142

6.2.3 分箱平滑法	146	7.6.1 逐步向前选择属性	159
6.3 数据规范化	148	7.6.2 逐步向后删除属性	159
6.3.1 最小-最大规范化方法	148	7.6.3 混合式选择	159
6.3.2 z 分数规范化方法	149	7.6.4 判定树归纳	159
6.3.3 小数定标规范化方法	149	7.7 数据压缩	160
6.4 数据泛化处理	150	7.7.1 离散小波变换方法	160
6.4.1 空间数据支配泛化算法	150	7.7.2 主成分分析压缩方法	161
6.4.2 非空间数据支配泛化方法	151	7.8 数值约简	162
6.4.3 统计信息网格方法	151	7.8.1 有参数值约简	162
小结	151	7.8.2 无参数值约简	162
第 7 章 大数据约简与集成技术	152	7.9 数据集成的概念与相关问题	163
7.1 数据约简概述	153	7.9.1 数据集成的核心问题	165
7.1.1 数据约简定义	153	7.9.2 数据集成的分类	166
7.1.2 数据约简方式	153	7.10 数据迁移	168
7.2 特征约简	153	7.10.1 在组织内部移动数据	169
7.2.1 特征提取	154	7.10.2 非结构化数据集成	169
7.2.2 特征选择	154	7.10.3 将处理移动到数据端	170
7.3 样本约简	155	7.11 数据集成模式	171
7.3.1 简单随机抽样	155	7.11.1 联邦数据库集成模式	171
7.3.2 系统抽样	155	7.11.2 中间件集成模式	172
7.3.3 分层抽样	155	7.11.3 数据仓库集成模式	173
7.4 数据立方体聚集	155	7.12 数据集成系统	173
7.4.1 多维性	156	7.12.1 全局模式	174
7.4.2 数据聚集	156	7.12.2 语义映射	175
7.5 维约简	157	7.12.3 查询重写	175
7.5.1 维约简的目的	157	7.13 数据聚类集成	175
7.5.2 维约简的基本策略	158	7.13.1 数据聚类集成概述	175
7.5.3 维约简的分类	158	7.13.2 高维数据聚类集成	176
7.6 属性子集选择算法	159	小结	178

第 8 章 大数据分析挖掘技术 179

8.1 大数据分析概述 180

8.1.1 大数据分析的类型 180

8.1.2 数字特征 182

8.1.3 统计方法论 184

8.1.4 模型与构建 186

8.1.5 R 语言 186

8.2 统计分析方法 188

8.2.1 基本方法 188

8.2.2 常用分析方法 194

8.3 数据挖掘理论基础 211

8.3.1 数据挖掘是面向应用的技术 211

8.3.2 数据挖掘的理论基础 212

8.3.3 基于数据存储方式的数据
挖掘 212

8.4 关联规则挖掘 214

8.4.1 频繁项目集生成算法 215

8.4.2 关联规则挖掘质量 217

8.5 分类方法 218

8.5.1 基于距离的分类算法 219

8.5.2 KNN 算法的 MapReduce
实现 220

8.5.3 决策树分类方法 220

8.6 聚类方法 223

8.6.1 聚类定义与分类 223

8.6.2 距离与相似性的度量 225

8.6.3 划分聚类方法 226

8.6.4 层次聚类方法 228

8.7 序列模式挖掘与文本数据挖掘 229

8.7.1 时间序列预测的常用方法 229

8.7.2 序列模式挖掘 230

8.8 非结构化文本数据挖掘 231

8.8.1 用户反馈文本 232

8.8.2 用户反馈文本挖掘的一般
过程 232

8.8.3 文本的自然语言处理 234

8.9 基于 MapReduce 的分析与挖掘

实例 235

8.9.1 大数据平均值计算 235

8.9.2 大数据排序 236

8.9.3 倒排索引 237

小结 241

第 9 章 大数据分析结果解释与展现 242

9.1 数据分析结果解释的目的与内容 243

9.1.1 解释的目的 243

9.1.2 检查和验证假设 243

9.1.3 追踪分析过程 244

9.2 数据的基本展现方式 245

9.2.1 基于时间变化的可视化展现 245

9.2.2 由大及小的可视化展现 245

9.2.3 由小及大的可视化展现 246

9.2.4 突出对比的可视化展现 246

9.2.5 地域空间可视化展现 246

9.2.6 概念可视化展现 248

9.2.7 气泡图可视化展现 248

9.2.8 注重交叉点的数据可视化
展现 248

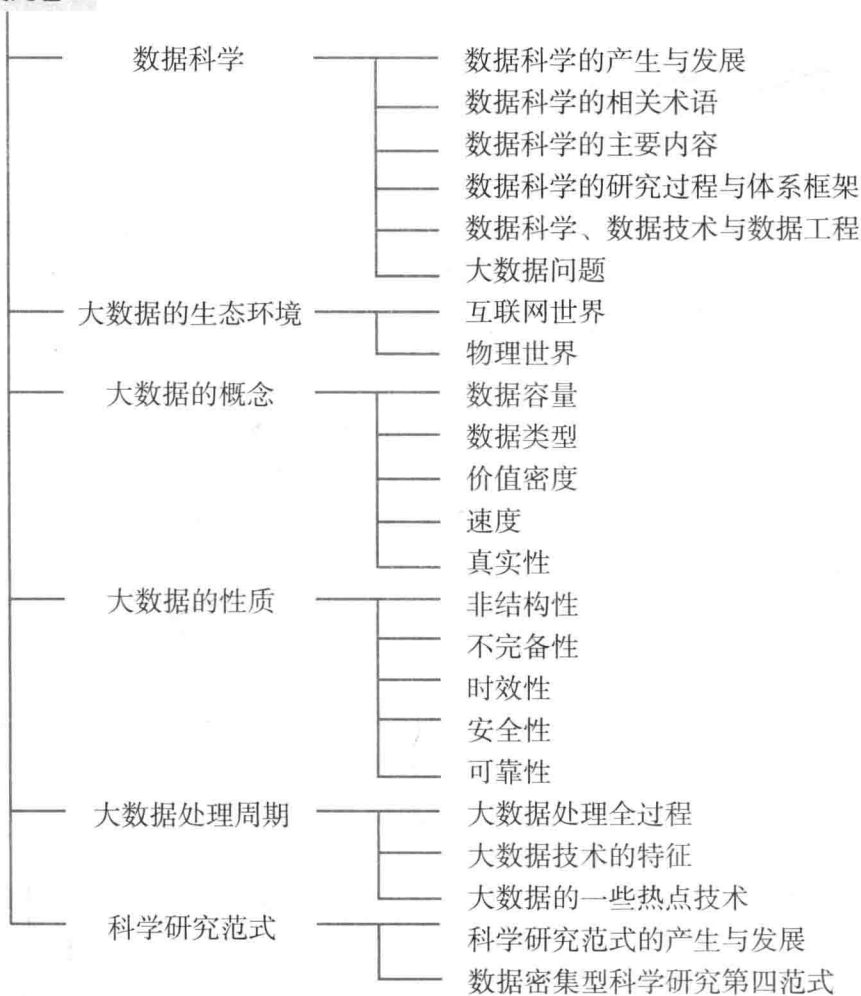
9.2.9 剖析原因的数据可视化展现 248

9.2.10 描绘出异常值的可视化 展现	249	9.4 大数据可视分析	253
9.3 大数据可视化	249	9.4.1 可视分析的理论基础	254
9.3.1 文本可视化	249	9.4.2 大数据可视分析技术	257
9.3.2 网络(图)可视化	250	9.5 数据可视化实现	260
9.3.3 时空数据可视化	252	小结	262
9.3.4 多维数据可视化	252	参考文献	263

第1章

概述

主要内容



计算机科学是算法与算法变换的科学，数据科学研究范围更为广泛。数据科学不仅可以推动数学、计算机科学、统计学、天体信息学、生物信息学、计算社会学等学科的发展，而且能够大力推动产业发展与进步。

1.1 数据科学

数据科学是关于数据的科学,基于数据的广泛性和多样性研究数据的共性。数据科学是研究探索 CYBER 空间中数据界的理论、方法和技术。

1.1.1 数据科学的产生与发展

数据科学产生于 20 世纪 60 年代。1974 年,彼得诺尔出版了《计算机方法的简明调查》,其中将数据科学定义为“处理数据的科学,一旦数据与其代表事物的关系被建立起来,将为其其他领域与科学提供借鉴”。1996 年在日本召开的“数据科学、分类和相关方法”会议上,将数据科学作为会议的主题词。2001 年美国统计学教授威廉·S·克利夫兰发表了《数据科学:拓展统计学的技术领域的行动计划》,首次将数据科学作为一个单独的学科,并把数据科学定义从统计学领域扩展到以数据作为计算对象,进而奠定了数据科学的理论基础。

1.1.2 数据科学的相关术语

1. CYBER 空间

CYBER 空间意译为异次元空间、多维信息空间、计算机空间、网络空间等。其本意是指以计算机技术、现代通信网络技术、虚拟现实技术等信息技术的综合运用为基础,以知识和信息为内容的新型空间,是人类运用知识创造的人工世界,是一种用于知识交流的虚拟空间。信息化是一个数据生产的过程,是将现实世界中的事物和现象以数据的形式存储到 CYBER 空间中。数据记录了人类的行为,包括工作、生活和社会的发展,是自然和生命的一种表示形式。

2. 数据爆炸

数据快速大量地产生并存储在 CYBER 空间中的现象称为数据爆炸,数据爆炸在 CYBER 空间中形成数据自然界。数据是 CYBER 空间中的唯一存在,需要研究和探索 CYBER 空间中数据的规律和现象。探索 CYBER 空间中数据的规律和现象是探索宇宙规律、探索生命规律、寻找人类行为规律、寻找社会发展规律的一种重要手段。

3. 数据科学的定义

数据科学是关于数据的科学或者研究数据的科学,是探索 CYBER 空间中数据界奥秘的理论、方法和技术,研究的对象是数据界中的数据。与自然科学和社会科学不同,数据科学的研究对象是 CYBER 空间数据。数据科学主要包括两方面:一是研究数据本身,以科学的方法研究数据的各种类型、状态、属性及变化形式和变化规律;二是用数据的方法研究科学,为自然科学和社会科学研究提供一种新的方法,称为科学研究的数据方法,其目的在于揭示自然界和人类行为现象和规律。

4. 数据科学的方法和技术

数据科学采用收集数据的形式,进行开放式分析,不做预先假定。在许多数据科学项目中,首先要浏览原始数据,形成一个假定,然后基于假定进行调查确认。数据科学的关键概念是:数据科学是一个经验科学,直接基于数据进行科学处理。数据科学已经有一些方法和技术,例如:数据获取、数据存储与管理、数据安全、数据分析、可视化等。

数据科学不仅完成分析,而且涉及整个端到端的生命周期,数据系统本质上是用于研发真实世界理解模型的科学设备。这就表明必须深刻理解数据的来源、数据转换的适用性和准确性、

扫一扫



数据分析
应运而生

转换算法和过程之间的相互作用，以及数据存储机制。这个端到端概览的角色能够确保所有事物都能够正确执行，从而探索数据、创建并验证各项科学假设。

1.1.3 数据科学的主要内容

数据科学的主要内容包括基础理论和数据预处理、数据计算、数据管理等。其中，基础理论包括概念、理论、方法、技术和工具等。数据科学的理论基础是统计学、机器学习、数据可视化及领域实务知识与经验等，如图1-1所示。数据科学学科建立，需要完成知识结构、课程设置和专业设置等学科体系建设，探讨数据科学与自然科学和社会科学之间的关系，以及数据科学与计算机科学和信息科学之间的关系等。



图1-1 数据科学的内容

1. 基础理论

观察和逻辑推理是科学的基础，数据科学中主要采用观察方法与数据推理的理论和方法，包括数据的存在性、数据测度、时间、数据代数、数据分类、数据相似性与簇论等。

2. 实验方法与逻辑推理方法

需要建立数据科学的实验方法，需要提出科学假说和建立理论体系，并通过这些实验方法和理论体系进行数据科学的研究，从而掌握数据的各种类型、状态、属性、变化形式和变化规律，揭示自然界和人类行为现象和规律。

3. 领域数据学

将数据科学的理论和方法广泛应用，开发出专门的理论、技术和方法，从而形成专门领域的数据科学，例如：脑数据学、行为数据学、生物数据学、气象数据学、金融数据学和地理数据学等。

4. 数据资源的开发方法和技术

数据资源是重要的现代战略资源，具有巨大的价值，越来越凸显其重要性，是继石油、煤炭、矿产等传统资源之后的最重要的资源之一。人类的社会、政治和经济都将依赖于数据资源，而石油、煤炭、矿产等传统资源的勘探、开采、运输、加工、产品销售等也都依赖于数据资源，离开了数据资源，将无法开展与完成这些工作。

其中，理论基础是在数据科学的边界之外。

1.1.4 数据科学的研究过程与体系框架

1. 数据科学的研究过程

① 数据集获取与存储。常用的数据类型有表格、点集、时间序列、图像、视频、网页和网络数据等。获取的数据存于数据库系统中。

② 数据的预处理。通过数据抽取、清洗、去噪与标准化、约简和集成，获得达到一定质量要求的数据。

③ 数据分析与挖掘。以科学的方法进行数据分析，进而发现整体特性。数据分析的基本假设是观察到的数据都是基于某个模型产生，通过数据分析找出这个模型。数据分析的主要困难是数据维数高，为此，需要降低算法的复杂度和应用分布式计算。通过数据分析与挖掘，发现数据规律。

④ 感知化与可视化数据分析结果。

2. 数据科学的构成

数据科学主要是计算机科学、数学与统计学知识以及行业经验的交集。

将数据科学进一步细化为如图 1-2 所示 12 个主要领域。

3. 数据科学的体系框架

数据科学的体系框架如图 1-3 所示。图 1-3 的上部分描述了数据的内容，下部分是数据科学基础描述。

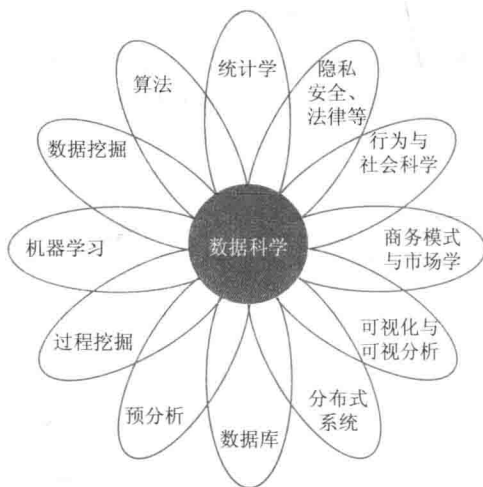


图1-2 数据科学的主要领域



图1-3 数据科学的体系框架

数据科学主要研究从数据中获取信息与知识、认识自然和行为，促进了科学与产业之间关



系的发展。

(1) 从数据中获取信息与知识

数据科学的研究对象、研究目的与研究方法等不同于计算科学、信息科学。数据存于CYBER空间中,信息是自然界、人类社会及人类思维活动中存在和发生的现象,知识是人们在实践中所获得的认识和经验。数据可以作为信息和知识符号的表示或载体,但数据本身并不是信息或知识。数据科学的研究对象是数据,而不是信息,也不是知识。数据科学通过以科学的方法从数据中获取对自然、生命和行为的认识,进而获得重要信息与知识。数据科学与其他学科的关系如图1-4所示。

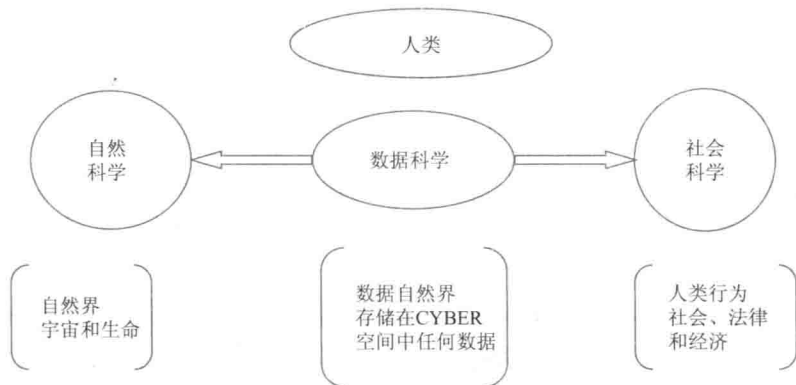


图1-4 数据科学与其他学科的关系

(2) 通过认识与探索数据来认识自然和行为

自然科学研究自然现象和规律,认识的对象是整个自然界物质的各种类型、状态、属性及运动形式。行为科学是研究自然和社会环境中人的行为以及低级动物行为的科学,已经确认的学科包括心理学、社会学、社会人类学等。数据科学支持自然科学和行为科学的研究。随着数据科学的进展,越来越多的科学研究可以直接针对数据进行,这将使人类通过数据研究科学,进而认识与探索数据来认识自然和行为。

人们生活在现实自然界和数据自然界两个世界中,人、社会和宇宙的历史将变为数据的历史。人类可以通过探索数据自然界来探索自然界,人类还需要探索数据自然界特有的现象和规律,这是数据科学的任务。可以期望与看出,目前的所有的科学研究领域都可能形成相应的数据科学。

数据科学的最终目的不仅是回答若干问题,还要生产数据产品。数据产品能够让其他人也利用上数据,并在此基础上进行数据分析。

(3) 促进了科学与产业之间关系的发展

数据科学不仅将给科研和教学体制带来大幅度的变革,也会给科学与产业之间及科学与社会之间的关系带来大幅度的变革。信息时代,万物数字化,许多学科已经和信息科技深度融合,形成新的研究领域,如生物信息学、天体信息学、数字地球、计算社会学等。一方面,用数据来研究科学与技术已经是科学研究的主要手段之一。另一方面,大量的、非结构化的数据,同样需要科学的手段研究数据。产业界在生产经营中积累了丰富的数据,学术界则有待于实践检验的模型和算法。数据科学为学术界和产业界的紧密衔接提供了纽带和桥梁,促进了产、学、研的深度融合与协作。

1.1.5 数据科学、数据技术与数据工程

科学是对客观世界本质规律的探索与认识,其发展的主要形态是发现,主要手段是研究,主要成果是学术论文与专著。技术是科学与工程之间的桥梁。其发展的主要形态是发明,主要手段是研发,主要成果是专利,也包括论文和专著。工程是科学与技术的应用和归宿,是以创新思想对现实世界发展的新问题进行求解,其主要发展形态是综合集成,主要手段是设计、制造、应用与服务,主要成果是产品、作品、工程实现与产业。科学家的工作是发现,工程师的工作是创造。

1. 数据科学

数据科学是对大数据世界的本质规律进行探索与认识,是基于计算机科学、统计学、信息系统等学科的理论,甚至发展出新的理论。它研究数据从产生与感知到分析与利用整个数据处理周期的本质规律,是一门新兴的学科。

2. 数据技术

数据技术是数据科学与数据工程之间的桥梁,包括数据的采集与感知技术、数据的存储技术、数据的计算与分析技术、数据的可视化技术等。

3. 数据工程

数据工程是数据科学与数据技术的应用,是以创新思想对现实世界的的数据问题进行求解,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,包括数据系统的设计、数据的应用、数据的服务等。

数据科学和工程可以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。数据科学与工程学科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。因此,多学科交叉融合是数据科学与工程学科的一个特点。与传统计算机和软件工程等学科相比,数据科学与工程学科具备独特的学科基础和内涵。数据科学与工程学科的理论基础涉及统计分析、商务智能以及数据处理基础。

数据科学随着计算机应用从以计算为中心逐渐向以数据为中心迁移,数据科学与工程学科的内涵和外延更加宽泛。软件工程学科中的相关技术提供了数据分析处理的工具以及具体开发的范式。数据处理技术是数据研究领域的一种重要的研究方法,用于研究和发现数据本身的现象和规律。

1.1.6 大数据问题

大数据是指传统数据处理应用软件不足以处理的大的或复杂的数据集。大数据表达理论方面主要包括大数据的处理周期、演化与传播规律,数据科学与社会学、经济学等之间的互动机制,以及大数据的结构与效能的规律性。在大数据计算理论方面主要研究大数据的表示以及大数据的计算模型及其复杂性;在大数据应用基础理论方面主要研究大数据与知识发现,大数据环境下的实验与验证方法以及大数据的安全与隐私。

大数据可分成大数据技术、大数据工程和大数据应用等领域。从解决问题角度出发,目前关注最多的是大数据技术和大数据应用。大数据工程指大数据的规划建设运营管理的系统工程。