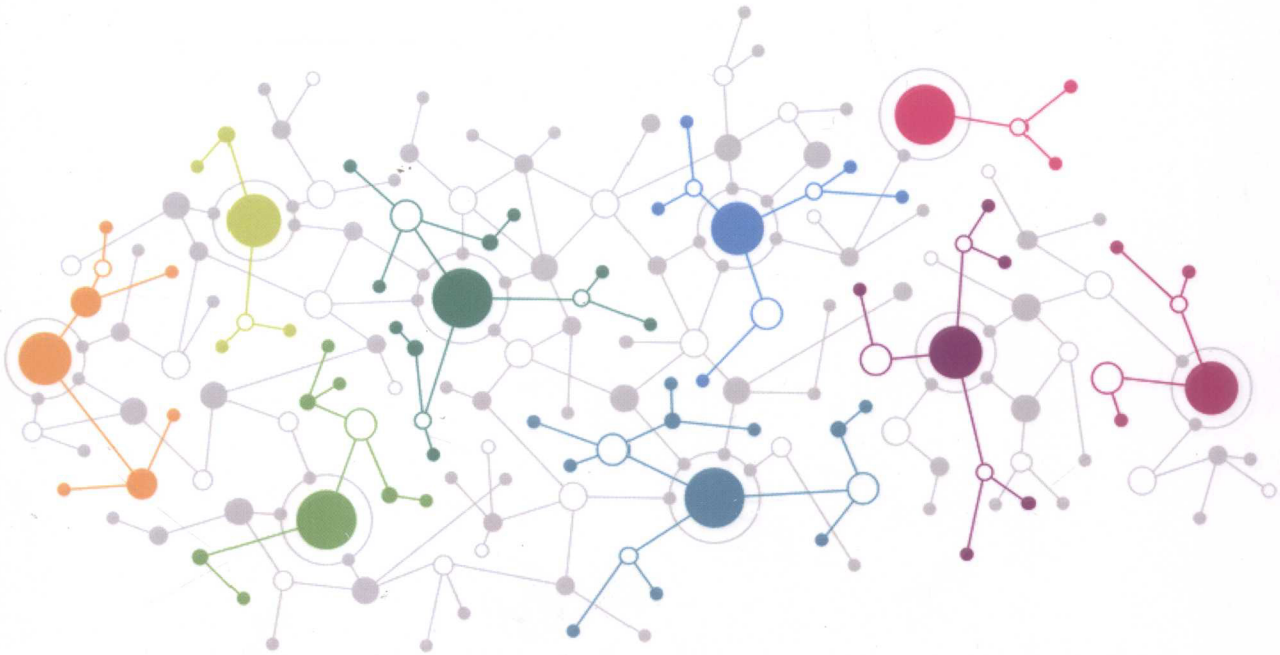
数据科学 Data Science

朝乐门 编著



清华大学出版社

数据科学 Data Science

朝乐门　编著

清华大学出版社
北京

内 容 简 介

数据科学是一门新兴的热门科学，国外一流大学纷纷设立同名课程，相应的专业、课程及书籍也深受欢迎。本书是国内第一部系统阐述数据科学的重要专著，填补了国内此领域的空白。本书在结构设计和内容选择上不仅充分借鉴了国外著名大学设立的相关课程以及全球畅销的外文专著，也考虑到了国内相关课程定位与专业人才的培养需求。

本书共包括8个部分(基础知识、数据预处理、数据统计、机器学习、数据可视化、数据计算、数据管理以及R编程)，既涵盖了数据科学的基本内容，又避免了与相关课程的低级重复。每章设有综合例题，做到理论学习与动手操作相结合。例题均采用R语言完成数据科学的特定任务。每章的首尾配有“导读”与“小结”，便于教师的教学和学生的自学。“习题”部分以主动数据收集和分析的开放题目为主，旨在帮助学生提高自我学习能力。书后附有R语言语法，便于入门的教学与学习。

本书可以满足数据科学、计算机科学与技术、管理学、数据统计、数据分析、图情档类等多个专业的老师、学生(含硕士生和博士生)的教学与自学需要。

图书在版编目(CIP)数据

数据科学/朝乐门编著. --北京：清华大学出版社，2016
ISBN 978-7-302-43699-7

Ⅰ. ①数…　Ⅱ. ①朝…　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第084834号

责任编辑：刘向威　薛　阳
封面设计：文　静
责任校对：焦丽丽
责任印制：刘海龙

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
课 件 下 载：http://www.tup.com.cn，010-62795954
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：21.25　　字　　数：420千字
版　　次：2016年8月第1版　　印　　次：2016年8月第1次印刷
印　　数：1～2000
定　　价：49.00元

产品编号：067987-01

FOREWORD 前言

数据科学已成为领域专家必备的知识和能力之一。如今，几乎所有的专家都在谈论大数据，但是部分“专家”并不是真正懂得大数据及其背后的科学——数据科学。在国内，数据科学的系统性研究仍属空白，人们只知道需要学习这门新兴科学，但并不知道如何学习。为此，本书：

(1) 是我国最早的系统阐述数据科学的专著之一；

(2) 以“经典理论×最佳实践”为编写思路，吸收了国内外重要的研究进展与实践经验；

(3) 提出了数据科学的理论体系，而不是现有文献的简单汇编；

(4) 加入了作者的创造性研究工作；

(5) 利用三年时间精心撰写的图书。

但是，学习数据科学确实存在一定的“难度”。就数据科学的理论基础——统计学、机器学习和可视化分析学而言，很多读者尤其是社会科学领域的，很容易对其产生“恐惧感”或“距离感”。为此，本书：

(6) 以“最深奥的理论÷最简单的逻辑”为编写思路，深入浅出，力争做到“阅读障碍的最小化”；

(7) 以“导读”形式给出学习建议；

(8) 以“图表”形式揭示数据科学中的重点知识；

(9) 以“脚注”形式解释读者容易曲解或需要深入了解的难点；

(10) 以“实用性”为选择内容的重要标准，不断给读者带来学习的“成就感”；

(11) 以“培养兴趣和信心”为撰写宗旨，并非停留在介绍知识和信息层次。

学习数据科学需要注意理论与实践相结合。数据科学与其他传统科学的重要区别之一在于与实践经验的耦合度高，读者不仅需要具备扎实的理论功底，而且应具有熟练的操作能力。为此，本书：

(12) 以“理论精讲+R 编程”为编写思路,协助读者提升理论联系实践的能力;

(13) 在“案例分析”部分提供了两个不同的案例——最佳实践和 R 编程,供读者选择性阅读;

(14) 以“独特的 R 代码注解”和“R 编程中的常见问题解答”的方式,帮助读者快速掌握 R 编程;

(15) 在“习题”中给出的问题不是课程内容的低级重复,而是帮助读者提升理论联系实践能力与自助学习能力;

(16) 在“参考文献与扩展阅读文献”中列出了相关领域的核心文献。

学习数据科学还需要注意与读者自己的领域知识相结合。数据科学与其他传统科学的另一个主要区别在于对领域知识的依赖度高。因此,学习应以“掌握面向领域的数据科学”或“发现领域中的数据科学”为主要目的,脱离具体领域的方式学习数据科学必将导致学习行为的空洞化和学习动力不足。为此,本书:

(17) 以“全集知识-领域差异性知识”为编写思路,在基本内容的设计与选择上,力争做到领域共性;

(18) 在学习建议和习题安排上,尽量体现出不同领域的差异性;

(19) 在内容细节的编写思路上,鼓励读者在自己的领域中使用数据科学的知识;

(20) 从数据科学视角介绍机器学习、统计学、数据可视化等基础知识,而不是简单重复这些课程。

数据科学是一门快速发展的新兴学科。目前,数据科学仍处于快速发展和不断演变的过程之中。为此,本书:

(21) 充分考虑到其未来发展中“变与不变”的问题,并重点描述“不变”部分;

(22) 以“本章小结”方式给出相关理论的发展趋势,即未来可能“如何变”的问题;

(23) 以“习题”方式给出未来理论的变化趋势及新理论的获得方法,即“如何跟踪最新变化”的问题;

(24) 在第 1 章理论基础中给出数据科学领域的主要期刊、会议、课程、学位项目、代表人物等,以便读者跟踪学习,也是属于“如何跟踪最新变化”的范畴。

在本书的撰写过程中,参阅了大量国内外教材、专著、论文、原始数据和相关资料,虽然书中对参考文献多有标注,但也难免挂一漏万,敬希相关作者鉴谅,笔者在此谨表示诚挚的谢意。同时,特别感谢:

(25) 中国人民大学原常务副校长冯惠玲教授、中国人民大学数据工程与知识工程教育部重点实验室主任、信息学院院长杜小勇教授,信息资源管理学院院长张斌教授为本书的

出版给予的大量指导与关心；

（26）中国人民大学路海娟、杨倩倩、马广惠、张莹等学生参与了部分章节的校对和PPT制作工作；

（27）清华大学出版社领导及编辑，尤其是刘向威博士和薛阳编辑为本书的出版做出的大量工作；

（28）国家自然科学基金项目（71103020）、国家社会科学基金项目（15BTQ054，12&ZD220）对本专著相关研究提供的资金支持；

（29）长期以来，亲人的理解与支持。本人从事基础研究，淡泊名利，他们却从不抱怨；

（30）即将为本书提出宝贵意见的您。书中必有不足之处，希望不吝赐教，让我们共同为数据科学的发展做出贡献！

朝乐门

于人民大学
2016年5月

CONTENTS 目录

第 1 章　基础理论 …… 1

1.1　数据 …… 3

1.1.1　数据模型 …… 5

1.1.2　数据维度 …… 7

1.2　大数据 …… 9

1.2.1　内涵与特征 …… 10

1.2.2　大数据时代的新理念 …… 12

1.2.3　大数据时代的新术语 …… 16

1.3　数据科学概述 …… 20

1.3.1　研究目的 …… 22

1.3.2　理论基础 …… 24

1.3.3　研究内容 …… 25

1.3.4　基本流程 …… 27

1.3.5　主要原则 …… 28

1.3.6　典型应用 …… 29

1.4　数据科学家 …… 30

1.4.1　主要任务 …… 30

1.4.2　能力要求 …… 30

1.4.3　常用工具 …… 31

1.4.4　团队工作 …… 32

1.5　数据科学项目 …… 33

1.5.1　角色定义 …… 33

1.5.2 基本流程 …… 34
1.6 应用案例 …… 35
小结 …… 41
习题 …… 42
参考文献及扩展阅读资料 …… 42
第 2 章 数据预处理 …… 45
2.1 数据质量 …… 48
2.1.1 统计学规律 …… 50
2.1.2 语言学规律 …… 51
2.1.3 数据连续性理论 …… 52
2.1.4 数据鉴别技术 …… 54
2.1.5 探索性数据分析 …… 57
2.2 数据审计 …… 60
2.2.1 预定义审计 …… 60
2.2.2 自定义审计 …… 61
2.2.3 可视化审计 …… 61
2.3 数据清洗 …… 62
2.3.1 缺失数据处理 …… 63
2.3.2 冗余数据处理 …… 64
2.3.3 噪声数据处理 …… 65
2.4 数据变换 …… 68
2.4.1 大小变换 …… 69
2.4.2 类型变换 …… 70
2.5 数据集成 …… 71
2.5.1 基本类型 …… 71
2.5.2 主要问题 …… 72
2.6 其他预处理方法 …… 74
2.6.1 数据脱敏 …… 74
2.6.2 数据归约 …… 75
2.6.3 数据标注 …… 76
2.7 应用案例 …… 76

小结 …… 87
习题 …… 87
参考文献及扩展阅读资料 …… 88
第 3 章　数据统计 …… 90
3.1　概率分布 …… 93
3.1.1　正态分布 …… 95
3.1.2　卡方分布 …… 97
3.1.3　t 分布 …… 97
3.1.4　F 分布 …… 98
3.2　参数估计 …… 98
3.2.1　点估计 …… 99
3.2.2　区间估计 …… 99
3.3　假设检验 …… 101
3.3.1　参数检验 …… 103
3.3.2　非参数检验 …… 104
3.4　基本分析方法 …… 105
3.4.1　相关分析 …… 106
3.4.2　回归分析 …… 108
3.4.3　方差分析 …… 111
3.4.4　分类分析 …… 112
3.4.5　聚类分析 …… 114
3.4.6　时间序列分析 …… 115
3.4.7　其他方法 …… 116
3.5　元分析方法 …… 118
3.5.1　加权平均法 …… 118
3.5.2　优化方法 …… 119
3.6　应用案例 …… 120
小结 …… 126
习题 …… 127
参考文献及扩展阅读资料 …… 128

第 4 章 机器学习 …… 129
4.1 基本概念 …… 133
4.1.1 定义 …… 133
4.1.2 应用 …… 134
4.2 机器学习活动 …… 135
4.2.1 训练经验的选择 …… 135
4.2.2 目标函数的选择 …… 136
4.2.3 目标函数的表示 …… 138
4.2.4 函数逼近算法的选择 …… 139
4.3 机器学习系统 …… 141
4.3.1 执行器 …… 141
4.3.2 评价器 …… 142
4.3.3 泛化器 …… 143
4.3.4 实验生成器 …… 143
4.4 主要类型 …… 143
4.4.1 基于实例学习 …… 144
4.4.2 概念学习 …… 144
4.4.3 决策树学习 …… 147
4.4.4 人工神经网络学习 …… 148
4.4.5 贝叶斯学习 …… 151
4.4.6 遗传算法 …… 152
4.4.7 分析学习 …… 154
4.4.8 增强学习 …… 159
4.5 典型算法 …… 160
4.5.1 *K*-Means 算法 …… 161
4.5.2 KNN 算法 …… 162
4.5.3 ID3 算法 …… 164
4.6 应用案例 …… 167
小结 …… 176
习题 …… 177
参考文献及扩展阅读资料 …… 178

第 5 章　数据可视化 …… 179
5.1　主要类型 …… 184
5.1.1　科学可视化 …… 184
5.1.2　信息可视化 …… 185
5.1.3　可视分析学 …… 186
5.2　基本模型 …… 187
5.2.1　顺序模型 …… 187
5.2.2　循环模型 …… 187
5.2.3　分析模型 …… 188
5.3　常用方法 …… 190
5.3.1　视觉编码 …… 191
5.3.2　统计图表 …… 193
5.3.3　图论方法 …… 198
5.3.4　视觉隐喻 …… 200
5.3.5　图形符号学 …… 202
5.3.6　面向领域的方法 …… 203
5.4　视觉编码 …… 205
5.4.1　视觉感知 …… 205
5.4.2　数据类型 …… 206
5.4.3　视觉通道 …… 207
5.4.4　视觉假象 …… 210
5.5　评价与改进 …… 211
5.5.1　测评原则 …… 211
5.5.2　测评流程 …… 212
5.5.3　测评方法 …… 213
5.6　应用案例 …… 213
小结 …… 218
习题 …… 220
参考文献及扩展阅读资料 …… 220
第 6 章　数据计算 …… 222
6.1　计算模式的演变 …… 224

6.1.1 集中式计算 …… 225
6.1.2 分布式计算 …… 225
6.1.3 网格计算 …… 227
6.1.4 云计算 …… 227
6.2 主流计算框架——MapReduce …… 229
6.2.1 基本思想 …… 230
6.2.2 实现过程 …… 232
6.2.3 主要特征 …… 233
6.2.4 关键技术 …… 236
6.5.5 下一代 MapReduce …… 238
6.3 主流计算平台——Hadoop MapReduce …… 240
6.3.1 数据流 …… 240
6.3.2 任务处理 …… 242
6.3.3 技术实现 …… 244
6.3.4 YARN …… 247
6.4 其他相关计算系统——Hadoop 生态系统 …… 249
6.4.1 HDFS …… 251
6.4.2 Hive …… 251
6.4.3 Pig …… 252
6.4.4 Mahout …… 253
6.4.5 HBase …… 254
6.4.6 ZooKeeper …… 254
6.4.7 Flume …… 256
6.4.8 Sqoop …… 257
6.5 应用案例 …… 258
小结 …… 261
习题 …… 262
参考文献及扩展阅读资料 …… 262
第 7 章 数据管理 …… 264
7.1 基本类型 …… 267
7.1.1 关系数据库 …… 268

7.1.2 NoSQL …… 271
7.1.3 关系云 …… 273
7.2 体系结构 …… 273
7.2.1 Master-Slave 结构 …… 275
7.2.2 P2P 结构 …… 276
7.3 关键技术 …… 278
7.3.1 数据模型 …… 278
7.3.2 数据分布 …… 282
7.3.3 数据一致性 …… 285
7.3.4 CAP 理论与 BASE 原则 …… 287
7.3.5 视图与物化视图 …… 288
7.3.6 事务与版本戳 …… 289
7.4 典型系统 …… 291
7.4.1 Memcached …… 291
7.4.2 MongoDB …… 294
7.4.3 Cassandra …… 296
7.4.4 HBase …… 298
7.5 应用案例 …… 301
小结 …… 304
习题 …… 307
参考文献及扩展阅读资料 …… 307
附录 A　R 语言与 R 软件 …… 309
附录 B　术语索引 …… 318

基础理论

本章在探讨大数据新思维的基础上，介绍数据科学的基础理论——研究目的、理论基础、研究内容、基本流程、主要原则及典型应用，并进一步探讨数据科学家及数据科学项目的核心问题(图 1-1)。数据科学家(或团队)应具备数据科学的基本知识、实战经验和创造性思维。

本章的编写目的是介绍大数据时代的新理念和新术语，讲解数据科学的基本原理，并结合典型案例分析，帮助读者掌握数据科学的主要思想及典型应用。

本章基本结构及主要内容如下。

首先，探讨数据的含义、数据中存在的主要矛盾、数据模型的层次及数据分类的维度。

其次，介绍大数据的内涵、特征以及大数据时代的新思维与新术语。

再次，讲解数据科学的基础理论，包括数据科学的研究目的、理论基础、研究内容、基本流程、主要原则及典型应用。

接着，介绍数据科学家的主要角色、能力要求、常用工具和团队工作。在此基础上进一步讨论数据科学项目的角色定位和基本流程。

最后，讨论大数据环境下的数据管理技术的两个案例——贝尔实验室和 2012 年美国总统大选中数据科学的研究与应用。

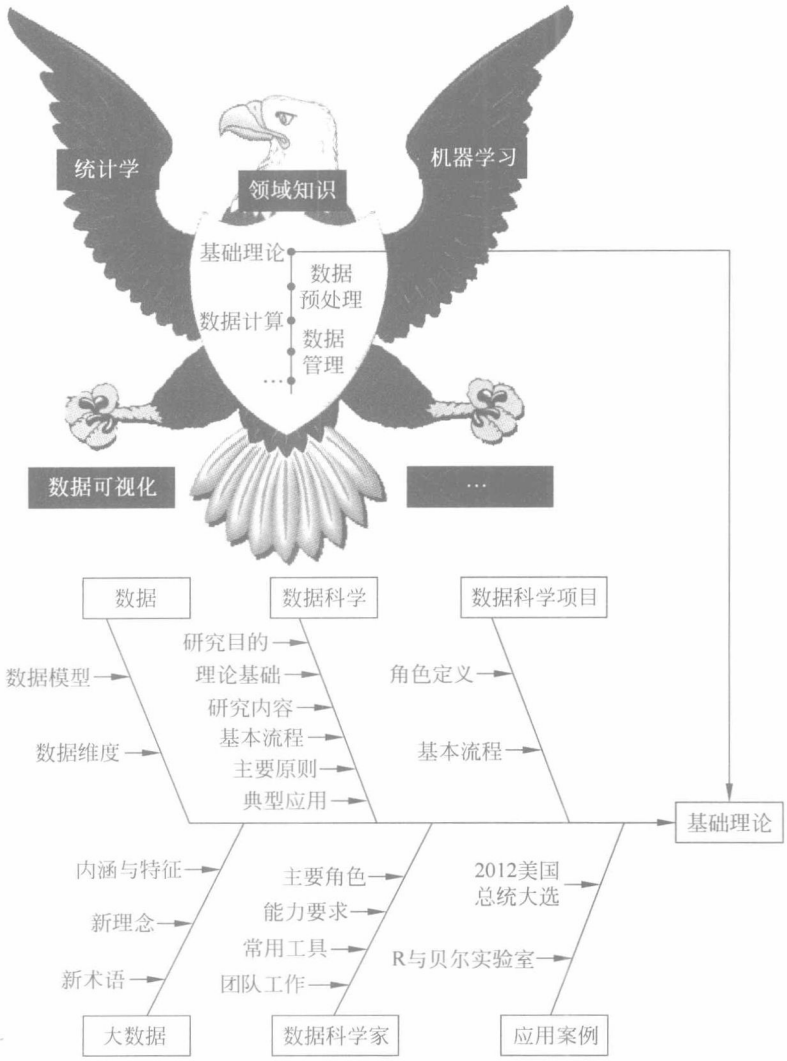


图 1-1　数据科学的基础理论

本章教学目的是帮助学生：

（1）了解——数据的含义、数据中存在的主要矛盾、数据模型的层次及数据分类的维度；数据科学项目的基本流程和主要角色。

（2）理解——大数据的内涵与特征、大数据时代的新理念与新术语、数据科学家的主要角色和常用工具。

(3) 掌握——数据科学的基础知识以及数据科学家的主要能力要求。

(4) 熟练掌握——结合读者自己所在专业领域中常用的数据科学方法、技术与工具。

本章教学重点是数据科学的基础理论——数据科学的研究目的、理论基础、研究内容、基本流程、主要原则及典型应用。

本章学习难点在于大数据时代的新理念；数据科学的研究目的、研究内容、基本流程与原则。

随着人们对数据价值的深入认识，尤其是大数据时代的到来，数据科学成为一门独立的科学，并受到学界和业界的广泛关注与热烈讨论。美国白宫、贝尔实验室、Amazon、Facebook、IBM、哈佛、MIT、斯坦福大学都在研究、开发和应用数据科学的理念、理论、方法、技术和工具。以“数据科学”命名的专著、学术期刊、国际会议、研究机构、专题网站、专门课程、学位教育也越来越多。本章的编写目的在于带领读者轻松步入一个全新的学科领域——数据科学，并帮助读者掌握数据科学的核心知识，为后续章节的学习奠定基础。

1.1　数据

在数据科学中，各种符号(如字符、数字等)的组合、语音、图形、图像、动画、视频、多媒体和富媒体等统称为数据(Data)。我们无法也没有必要给出“数据”唯一的权威定义，但至少需要注意以下两点。

(1) **“数据”与“数值”是两个不同的概念**。“数值”仅仅是“数据”的一种存在形式而已。除了“数值”，数据科学中所说的“数据”还包括文字、图形、图像、动画、文本、语音、视频、多媒体和富媒体等多种类型，如图 1-2 所示。

(2) **“数据”与“信息”、“知识”和“智慧”等概念之间存在一定的区别与联系**，如图 1-3 所示的 DIKW 金字塔(DIKW Pyramid)。从图 1-3 可看出，从“数据”到“智慧”的认识转变过程，同时也是“从认识部分到理解整体、从描述过去(或现在)到预测未来”的过程。

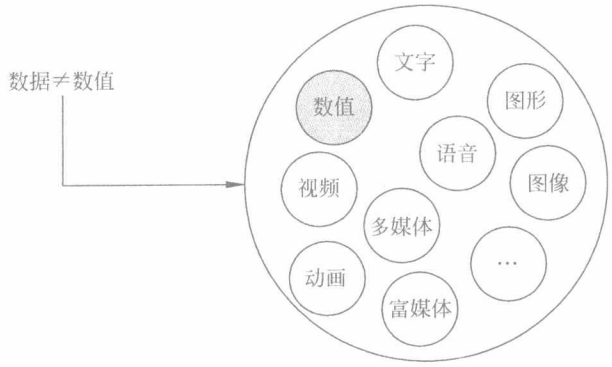


图 1-2 “数据”不等同于“数值”

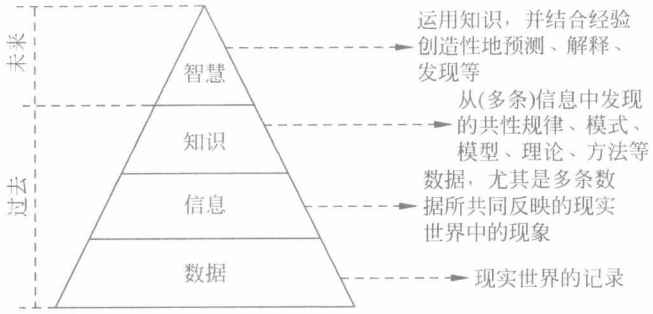


图 1-3 DIKW 金字塔

近年来，数据规模与利用率之间的矛盾日益凸显。一方面，数据规模的“存量”和“增量”在快速增长。2013 年全球数据总量大约为 4.4ZB，IDC 曾估计 2020 年将增长至 40ZB，人均达到 5.2TB(图 1-4)。在人们的生活与生产中，正在生成、捕获和积累着海量数据。

(1) 纽约证券交易所(The New York Stock Exchange，NYSE)每天生成大约 4～5TB 的数据。

(2) Illumina 的 HiSeq 2000 测序仪(Illumina HiSeq 2000 Sequencer)每天可以产生 1TB 的数据，大型实验室拥有几十台类似 LSST 望远镜(Large Synoptic Survey Telescope)的机器，每天可以生成 40TB 的数据。

(3) Facebook 每个月数据增长达到 7PB。