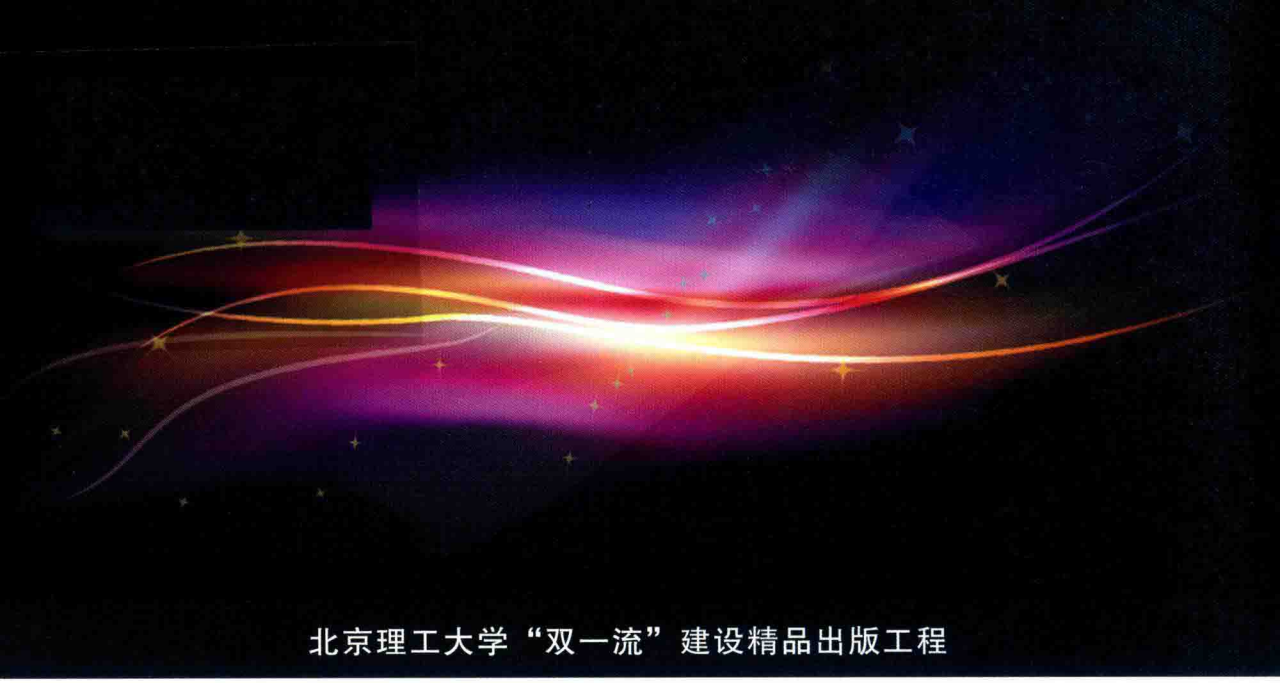


北京理工大学“双一流”建设精品出版工程

Introduction to Data Science
（3rd Edition）

数据科学导论
（第3版）

杨 旭 丁刚毅◎编著

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

Introduction to Data Science
(3rd Edition)
数据科学导论
(第3版)

杨 旭 丁刚毅◎编著

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

内 容 简 介

本书系统性地讲述了与数据科学相关的各方面知识，着重培养数据工程师所需要的技能与思维。本书从与数据科学相关的概念出发，通过丰富、翔实的案例，从各方面展示数据科学的运用方式，并且在其中穿插数据科学研究方式下新的思维模式的讲解，让读者有一个更为直观的认识，也可以从中感受到运用数据科学处理各个领域问题的方法和流程。本书还从工程概论的流程角度来讲述数据科学的工程体系架构，并展望数据科学的未来发展。同时本书特意加强了对数据预处理的理论和技术的讲解。本书可作为计算机相关专业的本科生教材，也可供相关专业技术人员阅读参考。

版权专有　侵权必究

图书在版编目（CIP）数据

数据科学导论 / 杨旭，丁刚毅编著. —3 版. —北京：北京理工大学出版社，2021.1
ISBN 978-7-5682-9235-1

Ⅰ. ①数…　Ⅱ. ①杨…②丁…　Ⅲ. ①数据处理-高等学校-教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2020）第 222936 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）68948351（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 / 三河市华骏印务包装有限公司
开　　本 / 787 毫米×1092 毫米　1/16
印　　张 / 15
字　　数 / 352 千字
版　　次 / 2021 年 1 月第 3 版　2021 年 1 月第 1 次印刷
定　　价 / 49.00 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

图书出现印装质量问题，请拨打售后服务热线，本社负责调换

PREFACE

前言

中国古代《易传·系辞上传》有言："是故，易有太极，是生两仪，两仪生四象，四象生八卦，八卦定吉凶，吉凶生大业。"从无诞生了有，有分阴阳，阴阳化四方，四方演八极，八极而定万物之吉凶，知过去与未来。似乎很玄，却不无道理。阴阳好比 0、1，四方代表二维平面，八极就是三维空间，加上时间，所有一切就都有迹可循，如何不可通晓？这其中的核心就是数据。有形的、无形的，天上的、地下的，过去的、未来的，何物何事何时，俱可归于数据。只要数据在手，无物可逃过你法眼。

谷歌的首席经济学家哈尔·瓦里安也说过："与数据相关的能力——包括获取数据、理解数据、处理数据、从数据中提取价值、用可视化方式展现数据、交流数据——将成为未来数十年至关重要的一项能力。不仅专业技术人员应该掌握它，即使是在我们的小学、中学和大学，也都应该传授相关的技巧。因为我们已经进入了大数据时代，数据无处不在、无孔不入。"

本书由"数"开篇，"说"古道今，推陈出"新"，结"语"收章。从数据的定义入手，探讨大数据和数据科学各自的含义，纵论数据的前世今生与当前国际国内对数据的运用和研究情况，分享数据的发展带给我们的新方法、新策略、新生活，最后给出结语。

目 录

CONTENTS

第一部分 数

第 1 章 数据 ……003
1.1 序言 ……003
1.2 数据的定义 ……003
1.2.1 数据的定义 ……003
1.2.2 其他相关概念 ……005
第 2 章 大数据与数据科学 ……006
2.1 大数据的概念 ……006
2.2 大数据的 4 V 特性 ……007
2.2.1 体量（Volume） ……007
2.2.2 多样性（Variety） ……009
2.2.3 价值/真实性（Value/Veracity） ……010
2.2.4 速度（Velocity） ……011
2.2.5 对 4 V 特性的体会 ……011
2.3 数据科学简述 ……011
2.3.1 数据科学的定义 ……011
2.3.2 数据科学的由来 ……012
2.3.3 数据科学的研究范畴 ……013

第二部分 说

第 3 章 说古 ……017
3.1 数据简史 ……017
3.2 大数据简史 ……023
3.3 大数据时代来临的表现 ……025
3.4 大数据的产生根源 ……026
3.4.1 大数据时代出现的技术基础 ……027
3.4.2 大数据时代出现的数据基础 ……027

第4章　道今……029
4.1　政府篇……029
4.1.1　联合国的大数据研究……029
4.1.2　美国的大数据研究……031
4.1.3　欧盟的大数据研究……032
4.2　企业篇……034
4.2.1　谷歌……034
4.2.2　IBM……034
4.2.3　百度……036
4.2.4　阿里巴巴……036
4.2.5　腾讯……037
4.2.6　华为……038
4.2.7　网易……039

第三部分　新

第5章　新方法……043
5.1　数据密集型研究方法……043
5.1.1　范式和范式的演化过程……043
5.1.2　第四范式兴起的社会根源……046
5.1.3　对第四范式的分析……050
5.1.4　数据科学研究的一般流程……051
5.2　数据的获取和预处理……052
5.2.1　数据的类型……052
5.2.2　网络爬虫技术……055
5.2.3　数据预处理的目的……061
5.2.4　数据清洗……063
5.2.5　数据集成……067
5.2.6　数据变换……070
5.2.7　数据归约……072
5.3　数据的存储与管理……076
5.3.1　数据的存储……076
5.3.2　数据的管理……082
5.4　数据的处理……088
5.4.1　Hadoop……088
5.4.2　Spark……092
5.5　数据的可视化……095
5.5.1　概述……095

5.5.2　数据可视化的重要性……096
5.5.3　可视化的基本原则——最大化数据－墨水……100
5.5.4　可视化工具……102
第 6 章　新策略……107
6.1　关联分析……107
6.1.1　啤酒与尿布……107
6.1.2　亚马逊的个性化推荐……117
6.1.3　潘多拉音乐组计划……123
6.1.4　塔吉特的大数据营销……129
6.2　趋势预测……132
6.2.1　“搜索＋比价”……132
6.2.2　Twitter 与对冲基金……137
6.2.3　疾病预测……139
6.2.4　电影票房预测……145
6.2.5　奥斯卡预测……149
6.3　决策支持……153
6.3.1　《纸牌屋》……153
6.3.2　美国总统大选……157
6.3.3　大数据助力医疗决策……160
6.3.4　大数据辅助行车路线优化……161
6.4　模式创新……162
6.4.1　大数据与反恐……162
6.4.2　利用大数据打击犯罪……167
6.4.3　大数据与破案……169
6.4.4　大数据与城市设施维护……170
6.4.5　大数据帮助寻根问祖……171
第 7 章　新生活……175
7.1　智慧城市……175
7.1.1　概述……175
7.1.2　大数据与智慧城市……178
7.1.3　智慧城市案例……182
7.2　智慧医疗……196
7.2.1　概述……196
7.2.2　智慧医疗的范畴……196
7.2.3　大数据与智慧医疗……201
7.2.4　可穿戴技术……204

第四部分　语

第8章　结语……215
8.1　数据带来的挑战……215
8.1.1　数据规模……216
8.1.2　数据的多样性和异构性……216
8.1.3　数据的不可靠问题……216
8.1.4　数据的实时性要求……216
8.1.5　数据隐私问题……217
8.1.6　人机协作问题……217
8.1.7　数据的访问与共享……217
8.1.8　数据运用的合理性……217
8.1.9　小结……218
8.2　数据带来的机遇……218
8.2.1　数据科学家的定义……218
8.2.2　数据科学家的从业前景……218
8.3　数据带来的误区……221
8.3.1　数据不是万能的……221
8.3.2　大数据时代的伦理忧思……223
8.3.3　不可被数据战胜……223
8.3.4　提防进入数据使用的误区……225
8.4　结语……225

图 目 录

图 1-1 都是数据 ……004
图 2-1 国际数据公司对大数据的定义 ……006
图 2-2 大数据的 4 V 特性 ……007
图 2-3 大数据与数据仓库对比 ……008
图 2-4 大数据的类型繁多 ……009
图 2-5 寻找数据价值 ……010
图 2-6 科学体系 ……012
图 2-7 彼得·诺尔（图灵奖得主，丹麦人） ……012
图 2-8 吴建福（国际知名统计学家，美国国家工程院院士） ……013
图 2-9 数据科学的研究范畴 ……013
图 2-10 数据科学与其他学科的关系……014
图 3-1 穴居壁画——最古老的数据记录形式 ……017
图 3-2 结绳记事 ……018
图 3-3 古埃及人用莎草记录数据 ……018
图 3-4 造纸流程 ……019
图 3-5 最早的留声机 ……019
图 3-6 最古老的照相机 ……019
图 3-7 早期的摄影机 ……019
图 3-8 世界上第一台电子计算机 ……020
图 3-9 当前世界上最快的计算机——天河二号 ……020
图 3-10 网络与数据……021
图 3-11 物联网……022
图 3-12 用数据产生智慧之花……022
图 3-13 用数据一窥未来……023
图 3-14 穿孔卡片……024
图 3-15 IBM 的沃森计算机系统 ……025
图 3-16 大数据时代来临组图一……025
图 3-17 大数据时代来临组图二……025
图 3-18 大数据时代来临组图三……026
图 3-19 移动互联网的飞速发展……027
图 3-20 物联网天生就是大数据……028
图 4-1 全球脉搏计划 ……029
图 4-2 通过社交媒体话题监测失业高峰 ……030

图4-3 通过Twitter话题关注妇女儿童健康……031
图4-4 data.gov网站……032
图4-5 西班牙桑坦德……033
图4-6 谷歌的大数据布局……034
图4-7 IBM的大数据平台和应用程序框架……035
图4-8 百度的大数据引擎……036
图4-9 阿里巴巴大数据竞赛……037
图4-10 腾讯大数据平台……038
图4-11 华为FusionInsight大数据平台……039
图5-1 关联数据运动……048
图5-2 数据科学的研究流程……052
图5-3 订单……053
图5-4 订单数据的数据类型区分……054
图5-5 通用的网络爬虫框架……055
图5-6 互联网网页的划分……056
图5-7 网页拓扑结构示例……056
图5-8 聚类抽样策略……058
图5-9 分布式抓取系统结构……058
图5-10 主从式基本结构……059
图5-11 对等式基本结构……060
图5-12 一致性哈希法确定服务器分工……060
图5-13 数据清洗……063
图5-14 聚类平滑噪声数据……067
图5-15 回归方法平滑噪声数据……068
图5-16 数据集成……068
图5-17 客户基本情况……068
图5-18 客户交易数据……068
图5-19 数据表……069
图5-20 写入期望值……069
图5-21 数据立方体的示例……072
图5-22 数据立方体的聚集……073
图5-23 维归约……073
图5-24 布乔的构想……076
图5-25 自动提花编织机……076
图5-26 打孔纸卡……077
图5-27 穿孔纸带……077
图5-28 计数电子管……077
图5-29 盘式磁带……078
图5-30 盒式录音磁带……078

图 5-31 磁鼓……079
图 5-32 软盘……079
图 5-33 硬盘机……079
图 5-34 日立 Deskstar 7K 500 硬盘……079
图 5-35 光盘……079
图 5-36 CD 光盘……080
图 5-37 DVD 光盘……080
图 5-38 云存储……081
图 5-39 互联网数据呈爆炸式增长……084
图 5-40 Spark 工作空间……093
图 5-41 Spark 程序运行示意图……094
图 5-42 窄依赖和宽依赖……094
图 5-43 Stage 的划分……095
图 5-44 约翰·图基……096
图 5-45 Anscombe 的四重奏……096
图 5-46 Anscombe 的四重奏的可视化呈现……097
图 5-47 一堆数值……098
图 5-48 扇形可视化下的展示效果……098
图 5-49 柱状可视化展示下的效果……099
图 5-50 散点可视化展示下的效果……100
图 5-51 2-D 形式的图表……101
图 5-52 3-D 形式的图表……101
图 5-53 不建议采用的可视化例子一……101
图 5-54 不建议采用的可视化例子二……101
图 5-55 Google Chart API……102
图 5-56 Raphaël……103
图 5-57 Visual.ly……103
图 5-58 Crossfilter……103
图 5-59 Polymaps……104
图 5-60 Kartograph……104
图 5-61 Processing 编程环境……104
图 5-62 R 语言编程……105
图 5-63 Weka 编程环境……106
图 5-64 利用 Gephi 做数据可视化……106
图 6-1 啤酒与尿布……107
图 6-2 美式购物篮分析法的代表——沃尔玛……108
图 6-3 日式购物篮分析法的代表——7-11 便利店……109
图 6-4 研究数据不应停留于表面……117
图 6-5 亚马逊的个性化推荐系统……117

图6-6 鼓励用户参与投票、书评等主观性活动……119
图6-7 推荐系统……120
图6-8 基于人口统计学的推荐……120
图6-9 基于内容的推荐……121
图6-10 潘多拉音乐盒子……124
图6-11 潘多拉网络电台……124
图6-12 私人定制的电台音乐……125
图6-13 潘多拉电台App……125
图6-14 推荐的方式……126
图6-15 Delicious网站……127
图6-16 Lastfm网站……128
图6-17 CiteULike网站……128
图6-18 Hulu网站……129
图6-19 塔吉特预测怀孕……130
图6-20 Farecast机票预测……133
图6-21 Decide电商比价网站……134
图6-22 Decide的比价无远弗届……135
图6-23 对电子产品系列更替的把握……136
图6-24 分数化评价……136
图6-25 “平静”指数和道琼斯指数对比……138
图6-26 谷歌预测美国流感趋势……139
图6-27 谷歌预测数据与真实数据的对比……141
图6-28 检索词个数的选取……141
图6-29 谷歌流感趋势预测结果……142
图6-30 大数据傲慢……143
图6-31 2012年票房收入和电影搜索量的曲线……146
图6-32 2012年票房收入和两类搜索量的曲线……147
图6-33 搜索量与首周票房收入之间的关系……147
图6-34 提前一周预测票房的效果……148
图6-35 提前一个月预测票房的效果……148
图6-36 微软研究院的戴维德·罗斯柴尔德……149
图6-37 微软研究院的戴维德·罗斯柴尔德博士预测奥斯卡获奖名单……150
图6-38 PredictWise网站……150
图6-39 奥斯卡投票预测器……152
图6-40 《纸牌屋》剧中人物……153
图6-41 大数据帮助大选……158
图6-42 大数据助力UPS……162
图6-43 美国“棱镜门”风波……163
图6-44 谁是“棱镜”计划的“帮凶”……163

图 6-45 美国“棱镜”计划……164
图 6-46 大数据时代的个人隐私安全……165
图 6-47 《少数派报告》……167
图 6-48 《源代码》海报……169
图 6-49 纽约沙井盖……170
图 6-50 家谱网……171
图 6-51 家谱网帮助寻根问祖……173
图 7-1 智慧城市……175
图 7-2 城市建设所面临的社会基础设施问题……178
图 7-3 智慧城市的基本特征与层次构成……178
图 7-4 智慧城市组件示意图……179
图 7-5 智慧城市建设必需的数据分析技术的发展趋势……181
图 7-6 松岛新城……183
图 7-7 韩国政府斥巨资修建的松岛新城……184
图 7-8 美国俄亥俄州的哥伦布市……185
图 7-9 My Columbu 移动应用程序……186
图 7-10 爱沙尼亚的塔林市……189
图 7-11 阿姆斯特丹的 ASC 计划……190
图 7-12 里约热内卢的城市运营中心……192
图 7-13 里约热内卢的智慧城市建设……193
图 7-14 科学家利用大数据监控脑伤病人的恢复情况……202
图 7-15 建立个性化的用药模型……203
图 7-16 利用可视化手段分析病症……204
图 7-17 可穿戴设备……205
图 7-18 Maxim 生命体征测量 T 恤……206
图 7-19 TI 公司的 Health Tech……206
图 7-20 Valencell：可随身穿戴的微型生理监测模块……207
图 7-21 Google Glass……208
图 7-22 苹果的 iWatch……209
图 7-23 BrainLink 意念头箍……209
图 8-1 大数据带来新的洞察力……215
图 8-2 分布式计算……216
图 8-3 大数据时代的隐私保护……217
图 8-4 大数据要求多领域人机协作……217
图 8-5 麦肯锡对数据科学方面人才需求空缺的预测……219
图 8-6 数据科学从业人员的未来成长性……221
图 8-7 数据的优势和不足……222
图 8-8 《千钧一发》电影海报……224

第一部分　数

第1章
数　据

1.1　序言

中国古代《易传·系辞上传》有言："是故，易有太极，是生两仪，两仪生四象，四象生八卦，八卦定吉凶，吉凶生大业。"从无诞生了有，有分阴阳，阴阳化四方，四方演八极，八极而定万物之吉凶，知过去与未来。似乎很玄，却不无道理。阴阳好比0、1，四方代表二维平面，八极就是三维空间，加上时间，所有一切就都有迹可循，如何不可通晓？这其中的核心就是数据。有形的、无形的、天上的、地下的、过去的、未来的，何物何事何时，俱可归于数据。只要数据在手，无物可逃过你法眼。

谷歌的首席经济学家哈尔·瓦里安也说过："与数据相关的能力——包括获取数据、理解数据、处理数据、从数据中提取价值、用可视化方式展现数据、交流数据——将成为未来数十年间至关重要的一项能力。不仅仅专业技术人员应该掌握它，即使是在我们的小学、中学和大学，也都应该传授相关的技巧。因为我们已经进入了大数据时代，数据无处不在、无孔不入。"

本书由"数"开篇，"说"古道今，推陈出"新"，结"语"收章。从数据的定义入手，探讨大数据和数据科学各自的含义，纵论数据的前世今生与当前国际国内对数据的运用和研究情况，分享数据的发展带给我们的新方法、新策略、新生活，最后给出结语。

1.2　数据的定义

1.2.1　数据的定义

究竟什么是数据呢？一提到数据，首先闪过人们脑海的一定是一堆数字。的确，数据最通常也最简单的表现形式就是数字。但数据可不是只包括数字。随着我国大数据战略的推广，"大数据"这个词这几年已火遍全国。但凡了解过大数据的一定知道，所谓大数据，那可是囊括了文本、音频、视频、图像、网络日志、传感器信号，等等。也就是说，这些都是数据（图1－1）。那数据究竟该怎么定义呢？

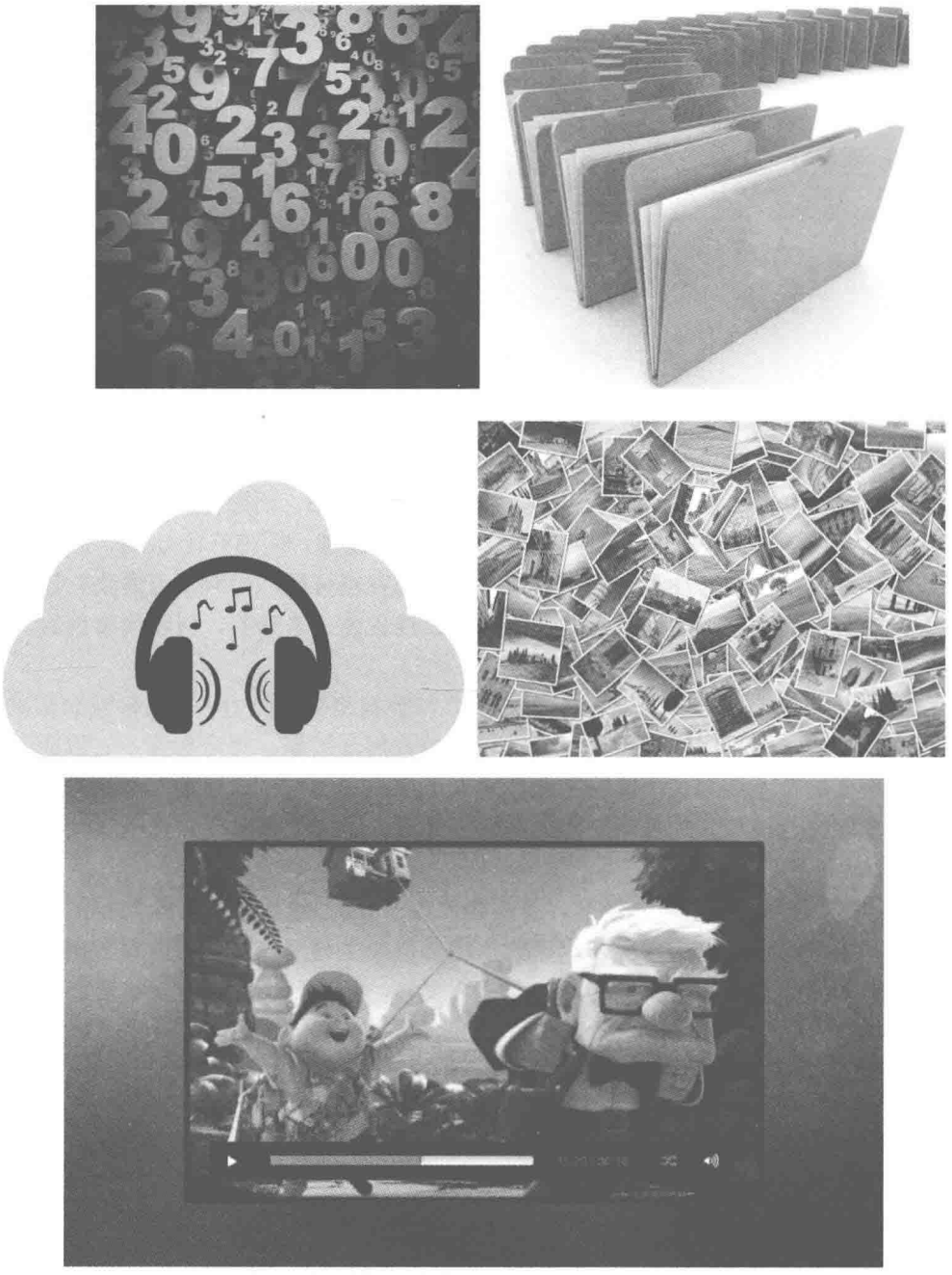


图1-1　都是数据

本书中，我们认为数据是指以定性或者定量的方式来描述事物的符号记录，是可定义为意义的实体，它涉及事物的存在形式。

简单说，数据就是人为创造的一种对事物的表示方式，是通过观察或实验得来的对现实世界中地方、事件、对象或概念的描述和反映。

这里有两个核心的点：

1）数据是人创造的，不管是通过观察还是实验得到的，它一定是人的主观意志的表现，