普通高中教科书

信息技术

必 修 1

数据与计算

上海科技教育出版社

编写人员名单

主　　编：郑　骏

副 主 编：沈富可

分册主编：郑　骏

主要编写人员（以姓氏笔画为序）：

毛黎莉　卢　源　朱晴婷

陈　凯　房爱莲　胡　杨

郭　骏　曹红霞

欢迎广大师生来电来函指出教材的差错和不足，提出宝贵意见。

上海科技教育出版社地址：上海市柳州路 218 号

邮政编码：200235

联系电话：021-64702058

邮件地址：office@sste.com

写给学生的话

亲爱的同学：

不知道你是否留意，我们几乎每天都能听到“数据”这个词。数据和每个人的生活都密切相关，它不仅是信息的载体，也是人们提取信息、作出决策的重要依据，并逐步成为社会发展的一项资源。人们合理选用技术工具处理数据，可以提高数据应用效能，发现其中隐含的信息，精准解决生活与学习中的问题。

在《数据与计算》的学习中，我们将带领你理解数据、信息和知识的相互关系，体验利用数字化工具处理数据和发现信息的过程；利用一种程序设计语言编写程序，实现简单算法，经历计算机解决问题的整个过程。你将在运用数字化工具的学习活动中，理解当今数字化世界的运转方式，提高利用信息技术解决问题的能力，发展信息意识和信息社会责任，养成数字化学习与创新的习惯。

为了让你在学习《数据与计算》的过程中获得更大的成功，请浏览本书的栏目介绍。

单元引言、学习目标和单元挑战

从生活经验出发引入本单元将要学习的内容，提出本单元学习要达成的学习目标，预告学习完本单元后要接受的单元挑战。

项目引言和学习目标

描述项目产生的背景和意义，介绍项目学习的主要内容，并提出一些具体问题，引导你带着问题探究。

项目学习指引

通过剖析真实的项目实施过程，帮助你了解学科思想方法，理解相关概念，掌握具体技能。

核心概念和小贴士

解释一些重要概念和术语，或提示相关知识和技术，帮助你抓住重点，扫除认知障碍。

思考与讨论??

提出若干问题引导你对技术背后的原理以及人、信息技术与社会的关系等进行思考和讨论。

数字化学习

引导你利用网络、数字化工具和数字资源进行学习。

活　动

提出活动任务，并引导你运用所学知识，使用信息技术工具进行探究、总结和展示。

知识链接

系统整理和归纳本项目的知识要点，方便你学习。

拓展阅读

补充更丰富的阅读材料，开阔你的视野。

单元挑战

布置面向真实情境的项目任务，希望你综合运用本单元所学的知识与技能去解决问题。

单元小结

用思维导图可视化呈现本单元的知识脉络，提供基于学科核心素养的评价表，为你的学习表现进行自我评价。

在学习过程中，希望你勤实践体验、多思考讨论，借助各种数字化工具、资源进行学习与创新，不仅要理解和掌握具体的信息技术知识与技能，还要把握用信息技术解决问题的思想方法，并思考将信息技术应用于社会时所引发的各种挑战，以开放、包容的心态与信息技术、信息社会一起进步。

编　　者

目 录

第一单元

数据与信息

在现实世界中，每个人每天会产生大量数据，如去过哪里、买过什么商品、走了多少路等。这些看似平凡的数据却蕴含了大量的信息，如果善加利用，会给社会创造意想不到的价值。例如，电商平台根据用户的浏览和购买记录，有针对性地向用户推荐商品，以提高商品销量；智能手环告诉佩戴者每天走了多少步、消耗了多少热量、深度睡眠有多长时间等，并提供保健建议，甚至推荐相应的健身产品；无人驾驶汽车使用摄像头、车载雷达、激光测距仪等设备采集数据，识别周围的交通状况，利用实时更新的地图进行自动导航，实现无人驾驶。

那么，究竟什么是数据？什么是信息？数据在计算机中是如何表示和处理的？本单元将带领大家揭开数据与信息的神秘面纱。

学习目标

- 通过具体实例，感知数据与信息，描述数据与信息的特征。
- 理解数据、信息与知识的关系，认识数据对人们日常生活的影响。
- 知道数值、文本、声音、图像等各类型数据的基本编码方式。

单元挑战

认识并制作二维码

项目一

探秘鸟类研究

——认识数据、信息与知识

为了解决各种问题，各行各业的人们都在做着采集数据、获取信息甚至构建知识的工作。例如，商店采集顾客购买的商品等数据，获取顾客购物喜好、商品畅销程度等信息，甚至构建以顾客为导向的市场营销战略知识，以更好地开展商品营销活动。又如，科学家长期在野外采集鸟类活动的数据（图 1-1），获取鸟类分布、鸟类对栖息地的选择等信息，从而构建鸟类与植物关系的知识，用于开展鸟类保护工作和生物多样性研究工作。

图 1-1　采集鸟类活动的数据

项目学习目标

在本项目中，我们将探秘一群科学家的鸟类研究活动，了解科学家是如何采集数据、获取信息的。

完成本项目学习，须回答以下问题：

1. 什么是数据？什么是信息？数据和信息的特征有哪些？
2. 数据、信息与知识的关系是什么？
3. 什么是大数据？大数据的特征有哪些？

项目学习指引

1. 采集鸟类活动的数据

空中掠过几只鸟，转瞬消失在树林里，难觅踪迹。这些美丽的精灵栖息在哪里？它们喜欢怎样的林木环境？为回答这些问题，科学家在浙江某国家森林公园的一片实验林地里设立了一个国家野外科学观测研究站，并长期在那里采集各种数据。

通过观察、测量等工作，林地里的鸟类活动数据，如鸟的种类、数量、行为等，被定期记录了下来。这些描述鸟类活动的数据，有数值、文本、图形、图像等形式。为了更好地存储、处理这些野外采集来的数据，科学家将它们录入计算机中，如表 1-1、图 1-2 和图 1-3 所示。

核心概念

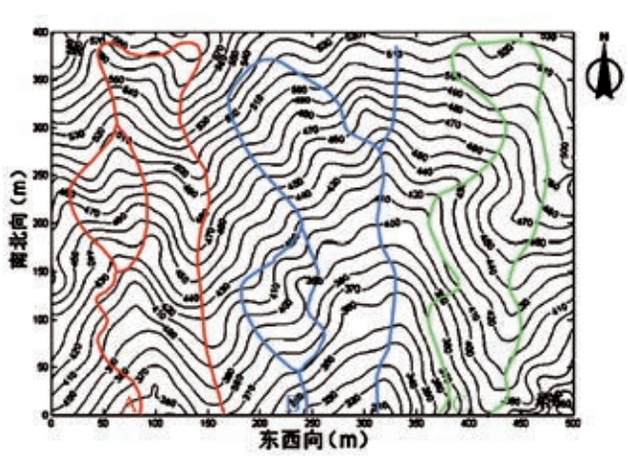
数据（data）是对客观事物属性的描述，是记录下来的某种可以识别的符号。在计算机科学中，数据是指所有能输入到计算机中并能被计算机程序处理的符号的总称。

表 1-1　林地鸟类活动调查记录表

记录人：×××　　记录时间：×××× 年 ×× 月 ×× 日

鸟的种类	数量（只）	栖息取食基层	行为	树牌号
灰眶雀鹛	5	灌丛	鸣叫	2130198
灰眶雀鹛	6	灌丛	鸣叫	2130123
白头鹎	5	冠中上	鸣叫	4080059
黄眉柳莺	2	冠中上	觅食	3080061
黑鹎	9	冠上	飞行	5060137
黄眉柳莺	1	冠中下	跳跃	1090013
红头长尾山雀	20	整个冠层	觅食，啄树干	2100030

图 1-2　用数码相机拍摄的灰眶雀鹛



（A、B、C 为林地调查线路）

图 1-3　用 GIS（地理信息系统）绘制的林地调查路线图

近年来，人们利用各种信息技术工具，实现了自动采集数据。例如，在林地里安装实时监控设备（图 1–4），利用红外摄影机全天候拍摄视频数据，利用录音设备录制声音数据，这些数据可以直接保存到信息系统中，供人们分析、研究。

图 1–4　用实时监控设备记录鸟类活动

小贴士

信息技术的发展使得人们采集和处理数据的手段不断加强，数据的内涵也逐渐丰富。在计算机发明前及发明初期，“数据”更多的是指数值型数据。随着计算机技术的发展，人们利用计算机处理的数据类型越来越丰富，涵盖了文本、声音、图形、图像、视频等非数值型数据。

活　动

1.1　近年来，随着信息技术的普及，国内外不少民间鸟类爱好者开始积极配合鸟类专业工作者，参与多项鸟类科学调查活动。一些观鸟网站和鸟类 App 都具有采集、整理、分享鸟类活动数据的功能，如图 1–5 和图 1–6 所示。查找并选择一个观鸟网站或鸟类 App，了解它向鸟类爱好者采集哪些鸟类活动数据。

图 1–5　某观鸟网站

图 1–6　某鸟类 App

2. 处理数据，获取信息

经过长年观测，这个研究站的工作人员采集了大量鸟类活动的数据。这些数据被多名科学家共享，他们对这些数据进行加工、分析，从而得出各种信息，为各自的科研服务。

现在人们越来越多地通过计算机来表示、组织和处理数据，从而可以获取并传播有价值的信息。

例如，科学家用计算机汇总 2010 年 10 月到 2012 年 10 月的数据后得知，在实验林地共观测到鸟类 44 种、4823 只次。其中，留鸟有 23 种，冬候鸟有 9 种，旅鸟有 8 种，夏候鸟有 4 种。进一步处理这些数据，能得出以下鸟类居留型种数的柱状图（图 1–7）及居留型比例的饼图（图 1–8）。

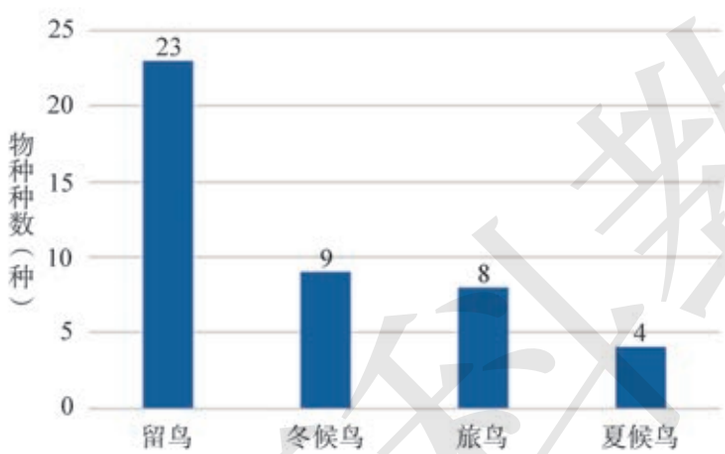


图 1–7　2010 年 10 月到 2012 年 10 月林地内鸟类居留型种数柱状图

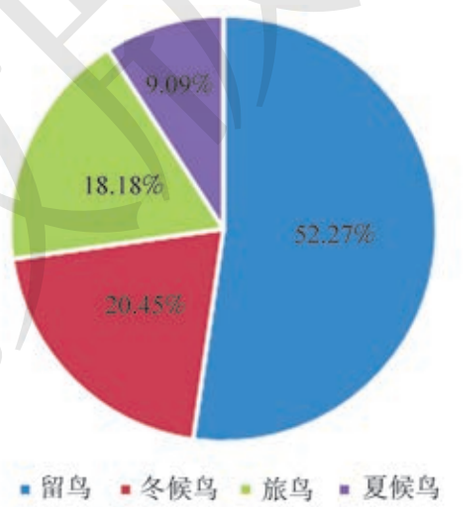


图 1–8　2010 年 10 月到 2012 年 10 月林地内鸟类居留型比例饼图

这些数据包含着怎样的意义？从中能获取哪些信息？这些信息对人们有什么价值？

信息是对数据的解释，具有主观性。对同样的客观数据，不同的人会得到不同的信息。对于一般人而言，或许能从以上数据中得到“该林地鸟类众多”的信息，还可以得到“该林地的留鸟种数比候鸟多”的信息。对于鸟类爱好者来说，或许会得到“该林地是适宜的观鸟地点”这一信息。而科学家获得的信息或许是：该地区鸟类物种多样性高，且繁殖鸟（留鸟和夏候鸟）占总种数约六成，这表明“该林地是鸟类繁衍生息的重要场所”。

核心概念

信息（information）是数据中所包含的意义，是对数据进行加工的结果。把数据有组织、有规律地采集在一起就形成了信息。数据一方面承载着信息，另一方面也产生着信息。

参见 P9 知识链接“数据和信息”

思考与讨论

1. 该林地的野外数据采集工作是由多名工作人员共同完成的，但数据大家都可以使用。这反映了数据的什么特征？

2. 同一份鸟类研究数据，鸟类学家可以从中获取鸟类的生存状况与栖息地环境改变之间联系的信息，从而指导人们开展环境保护工作；卫生防疫部门可以从中获取候鸟迁徙路线的相关信息，从而指导人们开展禽流感防护工作。这反映了数据的什么特征？这说明数据和信息之间有怎样的关系？你能列举出类似的例子吗？

数字化学习

上网查找并梳理国内外学者对信息的各种认识。

小贴士

知识（knowledge）是人们在改造世界的实践活动中所获得的可用于指导实践的认识、规律和经验，是归纳提炼出来的有价值的信息。

通过对大量数据、信息的归纳整理和反复验证，科学家完成了许多研究论文和学术著作，构建出不少关于鸟类的知识，如植物群落多样性与鸟类生存的关系等，为鸟类栖息地保护和物种多样性保护提供了理论依据（图 1-9）。

参见 P10 知识链接“数据、信息与知识的关系”

图 1-9　鸟类知识的构建

思考与讨论

在班级里介绍自己知道的鸟类知识，说说自己是从哪里获取这些知识的。

活　动

1.2（1）利用活动 1.1 中的观鸟网站或鸟类 App 的查询统计功能（图 1–10 和图 1–11），可以获取哪些数据？利用这些数据，鸟类爱好者和鸟类专业工作者可能会获取哪些信息？这些数据和信息有着怎样的价值？时效性如何？

图 1–10　某观鸟网站的鸟种统计页面

图 1–11　2016 年天津地区观测到的鸟种及次数

（2）利用思维导图，整理这个网站或 App 提供的数据和服务，并向大家介绍。

3. 利用大数据获取信息

当前，随着信息技术的飞速发展、数据采集规模的快速增长和数据处理速度的突飞猛进，大数据已深深影响了科学家开展科学研究和发现新知识的方式。

仍以鸟类研究为例，如今，摄像机、雷达乃至卫星等各种设备每天不停歇地自动获取规模大得不可想象的数据，经过计算机的高速处理，产生信息或知识。例如，利用从多个

核心概念

大数据（big data）是指无法在可承受的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。

监测天气的雷达基站下载的海量图像数据，通过计算机的高速数据处理与分析，科学家获得了鸟类对于山川地理的认知地图，获得了它们感知地球磁场、确定飞行方位的内在机理，以及关于鸟类迁徙的更多知识。

图 1-12 所示的是 2010 年 9 月 10 日晚上某国东北部数百万只鸟的迁徙轨迹（圆圈的大小表示鸟的密度，颜色表示雷达的回波强度，箭头所指的是鸟儿的迁徙方向，箭头的长度表示鸟的飞行速度）。

参见 P10 知识链接“大数据”

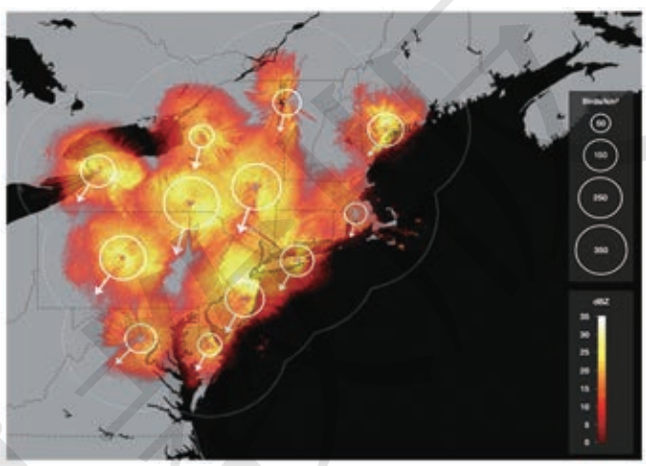


图 1-12　某夜某地上空鸟的迁徙轨迹图

思考与讨论

除了科学发现，大数据对我们的日常生活也有着越来越深刻的影响。试着交流你所了解的大数据及其应用。

活　动

1.3　地球夜间灯光分布卫星图是根据卫星获得的数据制作的测绘地图，它展示了地球上入夜区域的城市灯光分布情况。

（1）上网查找、观看中国夜间灯光分布卫星图。结合地理知识，分析我国各地夜间灯光分布情况，找出图中灯光最集中的城市，分析那里灯光强烈的原因。

（2）有一项跟踪研究指出，夜间灯光与能源消耗、人口增长、GDP 增长紧密相关。你认同这一观点吗？试着发掘图中信息，分析夜间灯光分布的规律及其原因。

（3）结合活动体会，分析数据、信息与知识的关系，并利用恰当的工具绘制这三者的关系图。

知识链接

数据和信息

数据是对客观事物属性的描述，是记录下来的某种可以识别的符号。信息是数据中所包含的意义，是对数据进行加工的结果。

数据和信息之间有着固有的联系和区别，它们具有许多共同特征，同时又存在差异，具有一些不同的特征。

1. 数据和信息的不同特征

（1）数据的载体性与信息的依附性

数据是信息的符号表示，是信息的载体；信息是数据的含义，是对数据的解释。两者密不可分。信息必须依附于某种载体，通过某种数据形式才能存储、表达和传播。相同的信息可以依附于不同的载体，其内容不会因载体形式的不同而发生变化。

例如，2017 年 7 月我国多地降暴雨，各地气象台发布了降雨量的数据，电视台播放了暴雨来袭的视频，广播台播放了语音报道，报纸做了文字和图片报道……这些文字、图像、声音、视频等都是数据，它们承载着“多地降暴雨”的信息。同时，“多地降暴雨”这一信息在存储和传播过程中依附于文字、图像、声音、视频等多种载体。虽然信息传递的渠道不同，所依附的载体形式不同，但信息的内容是相同的。

（2）数据的孤立性与信息的联系性

数据是最原始的记录，与其他数据之间没有建立联系之前，是分散和孤立的。只有通过对数据进行加工处理，与其他数据之间建立联系，才能形成针对某个特定问题的信息。例如，1、3、5、7、9、11、13、15 是一组数值数据，孤立地看每一个数，无法知晓它代表什么。但如果将这组数据联系起来，可以发现这是一个公差为 2 的等差数列的开头一段，据此可以推断其后面的数应该是 17、19、21……这样通过分析得出的结论便是信息。

（3）数据的客观性与信息的主观性

数据是记录下来可以被识别的符号，是原始事实，具有客观性；信息是对数据的解释，是数据处理的结果，具有主观性。数据本身没有意义，只有经过加工和解释，才具有意义，从而转化为信息。例如，用粉笔在黑板上画一个圆圈，请被测试者回答这是什么，会得到许多答案，如“数字 0”“英文字母 O”“句号”“月亮”……这里，黑板上的圆圈是数据，是客观存在的一个符号，没有确定的含义，而“数字 0”“英文字母 O”等是人们解读这一数据得到的信息。同一数据，具有不同知识、经验的人从不同的角度解读，会得到不同的信息。

2. 数据和信息的共同特征

（1）普遍性

数据是对客观事物属性的描述。事物是普遍存在的，因此，数据也无处不在，无时不有。考试的成绩、上课的铃声是数据，人们阅读的文章、观看的影片也是数据……有了数据，人们就会感知其中的意义，自觉或不自觉地获取信息。因此，信息也是普遍存在的。

（2）可处理性

对数据可以进行加工处理，生成新的数据。信息是数据加工的结果，同一数据经过不同的加工可以得到不同的信息。同时，对信息进行分析处理，可以得到更多的信息。例如，从某个人的身份证号码中提取第 7~12 位，得到数据 199006。根据身份证号码的编码规则，从中可以解读出信息——这个人的出生年月为 1990 年 6 月。作进一步加工处理，还可以从中解读出更多信息，如这个人的年龄、属相等。

（3）传递性与共享性

数据和信息是可以传递和共享的，同一数据或信息可以通过复制、传播，被多人重复使用。在传递和共享的过程中，数据和信息本身不会像物质和能源那样产生损耗。例如，同一新闻，可以通过报纸、电视、网络等多种渠道传播。在这一过程中，新闻通过各种渠道传递给多人，而新闻本身不会因传递和共享而有任何损失。

（4）价值相对性与时效性

数据和信息是有价值的，但其价值只有当数据或信息被利用时才能体现出来。数据和信息的价值具有相对性，是否有价值及价值的大小取决于使用者的需求，以及使用者对数据和信息的认知、理解和应用能力。例如，两家鞋厂分别派一位推销员到一个岛上推销鞋，他们上岛后共同感知的数据是“岛上居民一年四季都光着脚”。然而这两位推销员从中获取了不同的信息：第一位推销员认为“岛上无人穿鞋，没有市场”；第二位推销员认为“岛上无人穿鞋，市场潜力很大”。对数据和信息的不同理解，使他们做出了不同的选择，采取了不同的行动，从而获得了不同的结果。第一位推销员失望而归，第二位推销员请鞋厂速寄来 100 双鞋，把鞋送给岛上的居民，最终为鞋厂赢得了销售市场。数据和信息的价值也与个人需求有关。例如，“岛上居民一年四季都光着脚”这一数据及它承载的信息，对于想到岛上推销鞋的推销员是有价值的，但对于与此事不相关的人来说可能并没有什么价值。

数据和信息的价值还与时间有关，即具有时效性。例如，某商场今年 10 月 1 日至 7 日举办店庆活动，商品打折促销，如果消费者恰好想去该商场购物，并在 10 月 7 日之前获得了这个信息，那么，该信息对其是有价值的。但过了 10 月 7 日，该信息就无效了，其价值就降低了。

数据、信息与知识的关系

从数据到信息，再到知识，是一个从低级到高级的认知过程。数据是信息和知识的来源。无论信息还是知识，都来自于数据，都是以数据为载体而存在的。信息是经过加工的数据，知识是经过人类归纳整理和反复验证后沉淀下来而呈现的规律。同时，相应的知识又是加工数据、提炼信息的基础，能帮助人们理解信息。由此可见，数据、信息与知识之间不存在绝对的界限，三者有着千丝万缕的联系。

大数据

信息技术与经济社会的交汇融合引发了数据量的迅猛增长，数据已成为国家基础性战略资源。大数据正日益对全球生产、流通、分配、消费活动，乃至经济运行机制、社会生活方式和国家治理能力产生重要影响。

大数据是指无法在可承受的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。大数据通常具有 4V 特征，也就是 Volume（数据量）、Velocity（处理速度）、Variety（多样性）、Veracity（真实性）。

（1）数据量：大数据的体量很大，且数据集合的规模还在不断扩大。随着信息技术的大规模普及和应用，教育、商业、工业、科学研究、医疗等各行各业所产生的数据量都呈现出指数增长的趋势。

（2）处理速度：由于数据量增长速度快，大数据处理速度也必须快，且时效性要求高。大数据往往以数据流的形式动态地、快速地产生，需要在一定的时间限度下得到及时处理。

（3）多样性：大数据来自多种数据源，数据种类和格式非常丰富。随着智能设备、社交网络等的流行，机器和传感器数据（如设备日志、地理位置数据）、社交数据（如网站用户行为记录数据等）等各种新类型数据越来越多。

（4）真实性：大数据的真实性主要包括数据的可信性、真伪性、来源和信誉、有效性等。

拓展阅读

数据——信息社会的重要资源

随着人类跃进到大数据时代，数据不仅是新知识的来源，还是记录历史的最重要、最可靠、最好的方式。从今以后，人类所有的历史记录，无论是数字、文档、图片，还是音频和视频，都将以数据的形式存在，数据就是静态的历史，历史就是动态的数据。历史的碎片，就是游离的数据；历史的迷雾，就是模糊的数据；历史的盲点，就是缺失的数据。用数据构建的历史，因为精确的细节而永远鲜活，数据越丰富，后世的历史学家也就越能经由数据更好地再现当时的社会。

除了发现知识、记录历史，人类使用数据的巅峰形式，是通过数据训练机器，让机器获得智能，在不远的将来，无处不在的计算设备和网络将像有智商的人一样，为人类工作和服务。这意味着我们在向智能型社会迈进，在这个新的社会形态下，由于精准的计算和预测，整个社会的各个部分可以像无数个大大小小的轴承和齿轮一样，环环相扣，齿齿吻合。日常管理将通过数据得到优化，各种任务、合作可以无缝对接，社会运行的成本可大幅度降低。更重要的是，越来越多的工作将被计算机或者机器人代替。这既是进步，又是挑战。回望农业时代和工业时代，人类不断地开发我们赖以生存的自然环境，从地表到地下，物理性的资源终有耗尽的一天。而大数据将成为人类取之不尽、用之不竭的新资源，在这片资源之上，再通过软件和算法，人类将建设一个智能型世界。

数据，正在成为这个世界最重要的土壤和基础。

——摘自《数据之巅　大数据革命历史、现实与未来》

项目一

探究计算机中的数据表示
——认识数据编码

在鸟类研究过程中，科学家采集了各种各样、丰富多彩的数据。为了有效存储和处理这些数据，需要将它们数字化后存入计算机。

计算机是由逻辑电路组成的，逻辑电路通常只有高低两种电位状态，正好可以表示“0”与“1”，所以计算机采用二进制来存储和表示数据（生活中人们常用十进制数，二进制数和十进制数转换如图 1-13 所示）。因此，要想用计算机存储和处理现实中的数值、文本、图形、图像、声音和视频等数据，必须对数据进行二进制编码，即将其转化为由“0”和“1”组成的代码。数据的类型不同，编码的方法也不同。

二进制数	0	1	10	11	100	101	110	111	1000	1001
十进制数	0	1	2	3	4	5	6	7	8	9
二进制数	1010	1011	1100	1101	1110	1111	10000	10001	10010	10011
十进制数	10	11	12	13	14	15	16	17	18	19
二进制数	10100	10101	10110	10111	11000	11001	11010	11011	11100	11101
十进制数	20	21	22	23	24	25	26	27	28	29

图 1-13　二进制数和十进制数

项目学习目标

在本项目中，我们将通过探究一些鸟类活动数据的编码，了解数值数据和文本数据的编码方法，以及声音和图像的数字化过程。

完成本项目学习，须回答以下问题：

1. 编码的意义和作用是什么？
2. 数值数据编码的基本方法是怎样的？
3. 常用的文本数据编码方式是怎样的？
4. 声音数字化的基本方法是怎样的？
5. 图像数字化的基本方法是怎样的？

项目学习指引

1. 从树牌号认识编码

在项目一的林地鸟类活动调查中，科研人员以树为单位观察并记录每棵树上鸟的活动数据。为了清楚无误地区分和表示每一棵树，方便识别和交流，科研人员为林地中的每一棵树都设置了一个编号——树牌号，如图 1-14 所示。

树牌号
2130198
2130123
4080059

图 1-14　树牌号示例

给树编号的过程其实是一个编码的过程。为了给林地里的每一棵树设置一个唯一的树牌号，需要制定相应的编码规则。图 1-15 中的树牌号编码规则如下：每个树牌号由 7 位数字组成，第一位数字为一级区域编码（0~9，分别代表林地划分的一个一级区域），第二位和第三位数字为二级区域编码（01~20，分别代表该一级区域中的一个子区），第四位至第七位数字为树木编码（0001~9999，分别代表每个子区中的一棵树）。例如，树牌号 2130198，就代表 2 区 13 子区的第 198 棵树。

图 1-15　某树牌号的编码

生活中编码无处不在，如身份证号、银行卡号、邮政编码、学籍号、车牌号及条形码、二维码等，都是按照一定的规则产生的编码。

思考与讨论

如果 2 区 11 子区中有 10023 棵树，以上的编码规则是否适用？

核心概念

编码（encoding）是指用预先规定的方法将文字、数字或其他对象转换成规定的符号组合，或将信息、数据转换为规定的脉冲电信号。在计算机中，编码一般是指用预先规定的方法将数字、文字、图像、声音、视频等对象编成二进制代码的过程。

参见 P18 知识链接“编码”

活　动

2.1　了解生活中的编码。

（1）了解身份证号的编码规则，分析一代身份证号与二代身份证号的区别，思考启用二代身份证号的原因。

（2）根据本校实际情况，设计适用的学籍号编码规则，保证每位学生拥有一个唯一的学籍号。

（3）在班级内分享自己的学籍号编码方案，说明如何保证无重码，以及在什么情况下需要修改编码规则、如何修改。

2. 了解数值数据和文本数据的编码

要想用计算机存储和处理数据，必须先对它们进行编码，将它们转换成由“0”和“1”组成的二进制代码。对不同类型的数据，应采用不同的编码方法。

（1）数值数据的编码

数值数据是一类常见数据，是可用于算术运算的具体数值。例如，鸟的数量是 21 只，这个数值数据在计算机中是如何表示的呢？

计算机中的数值数据是以补码的方式表示的，以十进制数 +21 和 −21 的 8 位编码为例，它们的二进制数、原码、反码和补码分别如下。

$(+21)_{10}=(+10101)_2$	$(-21)_{10}=(-10101)_2$
$[+10101]_{原}=00010101$	$[-10101]_{原}=10010101$
$[+10101]_{反}=00010101$	$[-10101]_{反}=11101010$
$[+10101]_{补}=00010101$	$[-10101]_{补}=11101011$

（2）文本数据的编码

记录鸟类活动时需要记录鸟的名称，例如灰眶雀鹛的学名是 Alcippe Morrisonia。对这些由字母构成的数据，计算机是如何存储和表示的呢？

字母、数字、标点符号等，称为西文字符。计算机在存储和处理这些西文字符时，需要为每个字符规定一个由 0 和 1 组成的代码。目前，国际上普遍采用的西文字符编码标准是 ASCII 码（American Standard Code for Information Interchange，美国标准信息交换代码）。

小贴士

十进制（decimal system）是生活中常用的数制，二进制（binary system）是计算技术中广泛采用的数制。

参见 P19 知识链接“数值数据的编码”

标准 ASCII 码表如图 1-16 所示。

	000	001	010	011	100	101	110	111
0000	NUL	DLE	SP	0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

图 1-16　标准 ASCII 码表

小贴士

标准 ASCII 码用 7 个二进制位表示 1 个字符，如，字母 A 的 ASCII 码是 1000001，符号 # 的 ASCII 码是 0100011。

由于标准 ASCII 码只能表示 128 个字符，无法满足西文字符编码的需要，后来又扩充了 128 个字符，称为扩展 ASCII 码。

思考与讨论

常用汉字有近 5000 个，一个汉字的编码要用 2 个字节表示，而不是 1 个字节，这是为什么？

参见 P20 知识链接“文本数据的编码”

活　动

2.2 加密解密游戏。

（1）以标准 ASCII 码表作为密码本，选出 4 位学生配合完成加密解密游戏。

模拟保密电文的发送和接收过程，角色分配及建议流程如下：

• 首长 1：拟电文（设计一段由字母、数字或符号组成的明文），传递给发报员。

• 发报员：对电文进行加密（将字母、数字或符号转换为 ASCII 码，成为密文），传递给接报员。

• 接报员：接收密文，进行解密（将 ASCII 码转换为字母、数字或符号），解出明文，并传递给首长 2。

• 首长 2：向首长 1 核对解密后的电文与原电文是否一致。若不一致，组织小组成员查找问题，并改正。

（2）各小组自己设计编码方案和密码本，再玩一次加密解密游戏。

3. 了解声音和图像的数字化

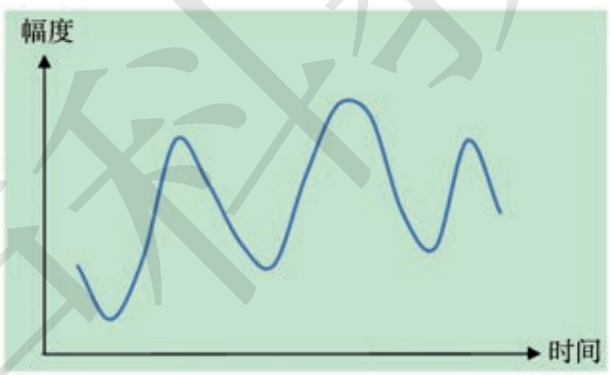
把自然界的鸟鸣声录制下来并转换为音频文件，经历了什么样的转换过程呢？自然界的鸟鸣声是一种连续的声波，为了用计算机存储和处理这些声音数据，需要将它们数字化，并记录成为音频文件。将模拟声音信号转换成数字声音信号，需要经历采样、量化和编码三个步骤，如图 1-17 所示。

图 1-17　声音数字化的过程

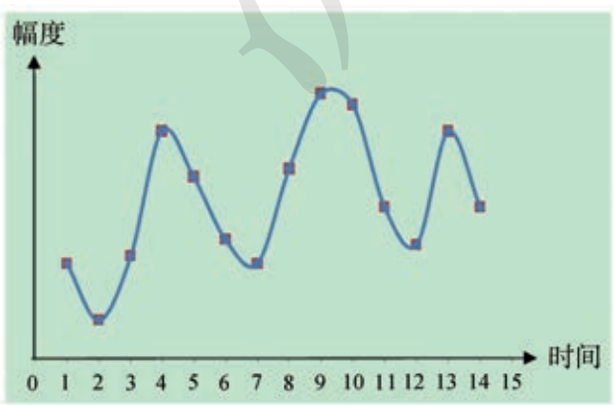
参见 P21 知识链接“声音数字化”

（1）采样

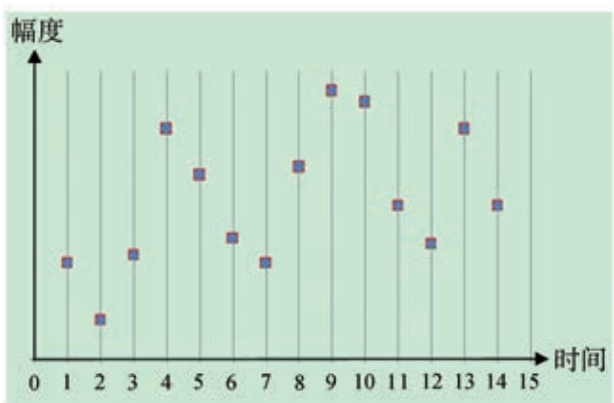
采样（sampling）即每隔一段时间在模拟声音信号的波形上采集一个幅度值。图 1-18（a）是一段鸟鸣声的模拟声音信号，对其采样时，在波形信号上按时间维度等距离地选取若干个离散的点，如图 1-18（b）所示。这些采样得到的幅度值被记录下来，如图 1-18（c）所示。



（a）



（b）



（c）

图 1-18　声音数字化的采样过程

（2）量化

采样之后，要用二进制数将采样得到的幅度值表示出来，这就是量化（quantization）。例如，取量化位数为 4，量化过程如下：

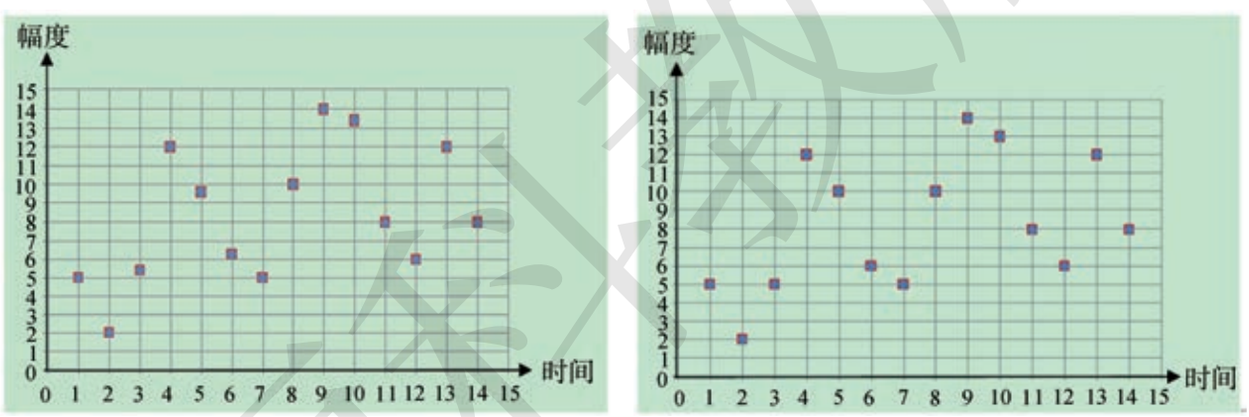
首先，确定量化位数为 4。

然后，将声音信号的幅度值范围划分为 2^4（16）个量化级数。

第三，确定采样点的量化值。若采样得到的幅度值不在这些级数之内，则按照一定的规则将它近似到某个级数值上。如图 1-19（a）中，第 3 个采样点的真实幅度值约为 5.4，将其四舍五入近似到级数值 5；第 5 个采样点的真实幅度值约为 9.8，将其四舍五入近似到级数值 10。同理可将第 6、第 10 个采样点的真实幅度值近似到相应的级数值。量化结果如图 1-19（b）所示。

小贴士

量化位数：存储、记录声音幅度值所使用的二进制位数。



（a）　　（b）

图 1-19　声音数字化的量化过程

最后，用二进制数表示这些采样点的量化值。例如，用 4 位二进制数来表示，第 1 个采样点的量化值为 0101，第 2 个采样点的量化值为 0010，第 3 个采样点的量化值为 0101……

（3）编码

经过采样和量化，模拟声音信号转化为一组二进制数序列，再通过编码将其按照一定的规则记录下来。采用不同的编码方法，会形成不同格式的音频文件，如 WAV 格式、MP3 格式等。

通过手机、数码相机、数码摄像机等数字设备，可以拍摄鸟类的照片，得到图像文件。图像数字化的过程和声音数字化类似，都会经历采样、量化和编码三个步骤。

参见 P22 知识链接“图像数字化”

活　动

2.3 探究声音数字化参数对音频文件的影响。

（1）探究采样频率对音频文件大小与音质的影响。

① 选择一种音频编辑软件，新建一个 WAV 文件，设置采样频率为 44.1kHz，量化位数为 32 位，声道数为 2（立体声），录制一段声音，并将其保存为“录音 1.WAV”。

② 保持其他参数不变，修改采样频率为 11.025kHz，并将文件另存为“录音 2.WAV”，比较两个文件的大小及音质，分析原因。

（2）模仿上述做法，分别探究量化位数和声道数对音频文件大小及音质的影响。

2.4 探究图像数字化。

（1）开展数字化学习，了解图像数字化的知识，对比声音数字化与图像数字化的过程。

（2）利用手机、数码相机等工具采集鸟类活动图片，选择一种图像处理工具，将鸟类活动的图像文件统一处理为相同分辨率和颜色深度的 BMP 文件。

（3）参考活动 2.3 设计实验方案，探究图像分辨率和颜色深度对图像呈现效果和文件大小的影响，并设计表格记录实验数据及结论。

（4）选择一种图像处理工具，探究不同文件格式对图像文件大小及质量的影响，并设计表格记录实验数据及结论。

知识链接

编码

编码是指用预先规定的方法将数字、文字或其他对象转换成规定的符号组合，或将信息、数据转换为规定的脉冲电信号。

编码一般具备以下功能和意义。

- 鉴别：编码是对象的唯一标识。通过辨识编码可以找到其唯一对应的对象。例如，邮政编码对应的地区是唯一的，身份证号码对应的人是唯一的，包裹单上的条形码对应的包裹也是唯一的。
- 排序：编码的符号都具有一定的顺序，比较容易进行排序。
- 专用含义：编码一般都会包含一定的含义，例如，本项目的树牌号中包含着所在区和子区的信息，身份证号码中包含着出生日期的信息。

在计算机中，编码一般是指用预先规定的方法将数字、文字、图像、声音、视频等对象编成二进制代码的过程。

数值数据的编码

数值数据，又称为数字数据，是可用于算术运算的具体的数值。

1. 数制

数值数据通常采用数制来表达，如，1 打等于 12 个，用的是十二进制；1 小时等于 60 分钟，用的是六十进制；1 米等于 10 分米，用的是十进制。

生活中常用的是十进制数，它的基数为 10，由 10 个基本数码（0、1、2、3、4、5、6、7、8、9）组成，逢 10 进 1。例如，十进制数 328.56 中，3、2、8、5、6 所代表的数值大小分别如图 1-20 所示。其中，10^0、10^1 等称为位权，以小数点为界，向左（整数部分）各位的位权依次为 10^0、10^1、10^2……向右（小数部分）各位的位权依次为 10^{-1}、10^{-2}……

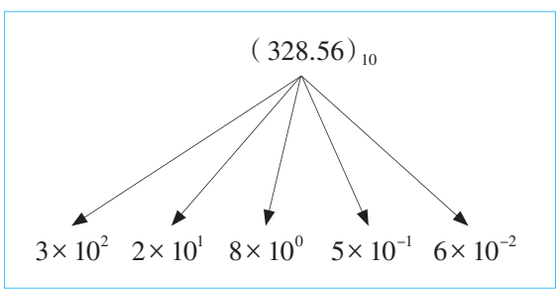


图 1-20　十进制数各位的位权

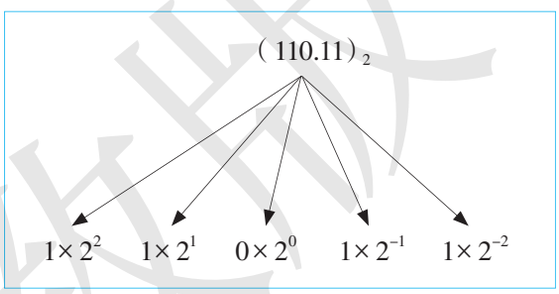


图 1-21　二进制数各位的位权

二进制是计算技术中广泛采用的一种数制，它的基数为 2。同样，以小数点为界，向左（整数部分）各位的位权依次为 2^0、2^1、2^2……向右（小数部分）各位的位权依次为 2^{-1}、2^{-2}……例如，二进制数 110.11 中，各位数字所代表的数值大小分别如图 1-21 所示。

计算技术中常用的数制还有八进制和十六进制，见表 1-2。

表 1-2　常用数制

数制	基数	可用符号	位权	进位规则
十进制数	10	0、1、2、3、4、5、6、7、8、9	10^{n-1}、10^{n-2}……10^0、10^{-1}、10^{-2}……	逢 10 进 1
二进制数	2	0、1	2^{n-1}、2^{n-2}……2^0、2^{-1}、2^{-2}……	逢 2 进 1
八进制数	8	0、1、2、3、4、5、6、7	8^{n-1}、8^{n-2}……8^0、8^{-1}、8^{-2}……	逢 8 进 1
十六进制数	16	0、1、2、3、4、5、6、7、8、9、A、B、C、D、E、F	16^{n-1}、16^{n-2}……16^0、16^{-1}、16^{-2}……	逢 16 进 1
R 进制数	R	0、1……R-1	R^{n-1}、R^{n-2}……R^0、R^{-1}、R^{-2}……	逢 R 进 1

数值数据可用于算术运算，每种数制都有其运算规则。二进制数的算术运算规则如下。

加运算：0+0=0，0+1=1，1+0=1，1+1=10（逢 2 进 1）

减运算：1-1=0，1-0=1，0-0=0，10-1=1（向高位借 1 当 2）

乘运算：0×0=0，0×1=0，1×0=0，1×1=1

除运算：0÷1=0，1÷1=1

2. 数值数据的编码

数值数据的编码过程如图 1-22 所示。

图 1-22　数值数据的编码过程

（1）转换

要用计算机存储和处理数值数据，首先要将其转换为二进制数。十进制数转换为二进制数，整数部分的转换方法是除 2 反向取余，小数部分的转换方法是乘 2 正向取整。如图 1-23 和图 1-24 所示，将十进制数 37.375 转换为二进制数，首先将其整数部分和小数部分分别转换为二进制数，然后再合并，得到$(37.375)_{10}=(100101.011)_{2}$。

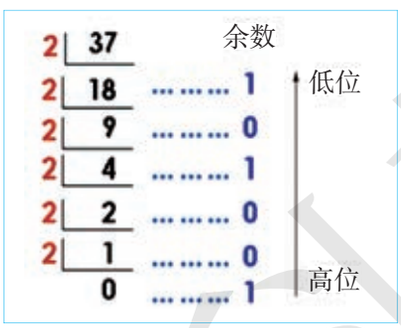


图 1-23　整数部分的转换

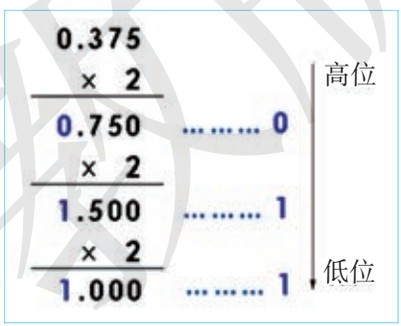


图 1-24　小数部分的转换

（2）编码

计算机中数值数据的编码分为原码、反码和补码，其编码规则见表 1-3。通常情况下，计算机用一个数的最高位存放符号，即用 0、1 表示正负符号，正数为 0，负数为 1，这个二进制位称为符号位。

表 1-3　原码、反码和补码的编码规则

	原码	反码	补码
正数	符号位（0）+ 数字部分 （如果原数不足 n-1 位，在高位补 0）	同原码	同原码
负数	符号位（1）+ 数字部分 （如果原数不足 n-1 位，在高位补 0）	在原码的基础上，符号位不变，其余各位取反	在反码的基础上 +1

注：n 为编码总位数

文本数据的编码

1. 西文字符的编码

应用最广泛的西文字符编码方案是 ASCII 码。ASCII 码是由美国国家标准学会（American National Standard Institute，ANSI）制定的通用单字节字符编码方案。

ASCII 码分为标准 ASCII 码和扩展 ASCII 码。标准 ASCII 码也叫基础 ASCII 码，使用 7 个二进制位来表示西文字符，包括所有的大写和小写字母、数字 0 到 9、标点符号，以及在美式英语中使用的特殊控制字符。扩展 ASCII 码用 8 个二进制位来表示字符，第 8 位用于确定附加的 128 个特殊符号字符、外来语字母和图形符号。

2. 汉字的编码

常用汉字有近 5000 个，这种信息容量要用 2 个字节长即 16 位二进制编码才能满足。1980 年，中国国家标准总局发布了中华人民共和国国家标准 GB2312—1980《信息交换用汉字编码字符集——基本集》，又称为国标码。国标码用 2 个字节表示一个汉字，其中每个字节的最高位为 0。例如，“大”字的国标码为 0011010001110011。

国标码在计算机内部存储和处理时会与 ASCII 码发生冲突，例如“4s”的 ASCII 码在计算机中的表示也是 0011010001110011。为了解决这个问题，汉字编码在计算机内的表示在国标码基础上稍做改动，将每个字节的最高位设为 1，这被称为机内码（简称内码）。例如，“大”字的机内码为 1011010011110011。机内码是用最高位均为 1 的 2 个字节表示一个汉字，是计算机内部存储、处理汉字所使用的统一编码。

3. Unicode

全世界有上百种语言，人们希望有一种编码，能将世界上所有的符号都纳入其中，每一个符号都给予一个独一无二的编码。Unicode 应运而生。

Unicode 是国际组织制定的可以容纳世界上所有文字和符号的字符编码方案。1990 年开始研发，1994 年正式公布。它为每种语言中的每个字符设定了统一且唯一的二进制编码，以满足跨语言、跨平台进行文本转换和处理的要求。目前的 Unicode 字符分为 17 组编排，每组称为 Plane（平面），每个 Plane 拥有 65536 个码位，共 1114112 个码位。

Unicode 一般用 2 个字节表示一个字符（非常偏僻的字符用 4 个字节）。但是，一篇英文文章，用 Unicode 编码比用 ASCII 编码需要多一倍的存储空间。于是，又出现了把 Unicode 编码转化为“可变长编码”的 UTF-8 编码。UTF-8 编码把一个 Unicode 字符根据不同的数字大小编码成 1~6 个字节，常用的英文字母被编码成 1 个字节，汉字通常是 3 个字节，只有很生僻的字符才会被编码成 4~6 个字节。

Unicode 的实现方式还有 UTF-16 和 UTF-32 等。

声音数字化

现实世界的声音是一种连续的波，称为声波。声音有两个参数：幅度和频率。要用计算机处理声音数据，必须把连续变化的波形信号转换成为离散的数字信号，将幅度和频率以 0 和 1 编码的形式表示出来，这一过程称为声音数字化。声音数字化的过程包含采样、量化、编码三个步骤。

1. 采样

声音的采样是指每隔一段时间在模拟声音信号的波形上取一个幅度值。相隔时间相等的采样为均匀采样（又称为线性采样），相隔时间不相等的采样为不均匀采样（又称为非线性采样）。

计算机每秒钟在模拟声音信号的波形上采样的次数称为采样频率。常见的采样频率有44.1kHz、22.05kHz、11.025kHz 等。采样频率越高，即采样的时间间隔越短，则在单位时间内得到的声音样本数据越多，对声音信号波形的表示越精确，声音的保真度越高。

2. 量化

声音的量化是用二进制数表示采样所得到的幅度值的过程。首先将幅度值范围划分为 2^n 个级数，每个级数对应一个幅度值，然后将采样得到的各个幅度值按一定的规则近似到某个级数值，并用二进制数表示，从而形成一组二进制数序列。这里的 n 称为量化位数。量化位数越大，划分的级数越多，采样结果近似到某个级数值时产生的误差就越小。因此，量化位数越多，数字化精度越高，声音就越保真。

3. 编码

声音的编码就是按照一定格式把经过采样和量化得到的离散数据记录下来，并在其基础上加入用于纠错、同步和控制的数据，最终转换成数字音频信号。不同的编码方法形成了不同格式的音频文件，如 WAV 格式、MP3 格式等。

采样频率、量化位数和声道数是数字化音频的技术指标，被称为声音数字化的三要素。它们直接影响数字化后音频的质量及其数据量的大小。一般情况下，未经压缩的音频文件的数据量可以按如下方法计算：

数据量（单位：字节）= 数据率 × 持续时间

=（采样频率 × 量化位数 × 声道数）÷ 8 × 持续时间

例如，一张 CD-ROM 中存放了 1 小时的数字音乐（未经压缩），则其数据量可按以下方法计算：

$$\begin{aligned}\text{数据量} &= (44100 \times 16 \times 2) \div 8 \times 60 \times 60\text{B} \\ &= 635040000\text{B} \\ &= 620156.25\text{KB} \\ &\approx 606\text{MB}\end{aligned}$$

标准 CD 格式的采样频率为 44.1kHz，量化位数为 16 位，声道数为 2（双声道）。数据量计算公式中的“÷ 8”是将位数转换成字节，一个字节由 8 个二进制位组成。1MB=1024KB，1KB=1024B。

图像数字化

从小小的商标到大型宣传海报，从各式各样的照片到风格迥异的图画，从茶杯上的图案到教材中的插图……凡此种种，从信息技术的角度看，都属于模拟图像，运用扫描技术或数字摄像技术可以将空间上连续的模拟图像转换成用 0、1 表示的数字图像，这一过程称为图像数字化。图像数字化的过程包含采样、量化、编码三个步骤。

1. 采样

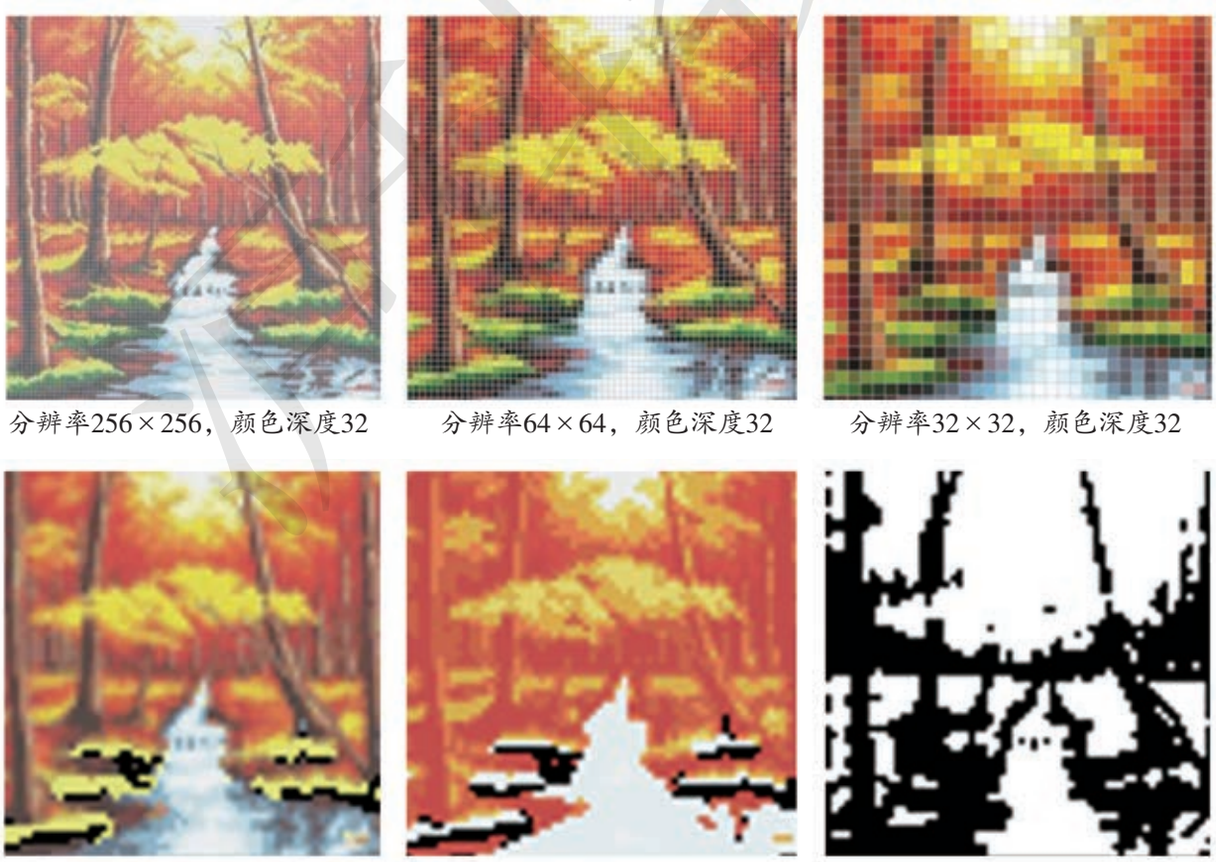
图像的采样是按一定的空间间隔自左到右、自上而下提取画面信息，将一幅连续的模拟图像在空间上转换成若干个离散的像素点，每个像素点呈现不同的颜色（彩色图像）或亮度（灰度图像）。

一幅图像所包含的横向和纵向的像素点的数目称为图像分辨率。例如，一幅图像的分辨率为 640×480，表示该图像由横向 640 个像素点、纵向 480 个像素点，共 640×480=307200 个像素点组成。如果不考虑其他因素的影响，图像分辨率越高，采样的精度就越高，数字化后的图像越清晰，同时图像所占的存储空间也越大。

2. 量化

图像的量化是将采样得到的每个像素点的颜色或亮度用若干位二进制数表示出来，其方法与声音数据量化的方法类似。首先确定颜色或亮度的取值范围，然后将近似的颜色划分成同一种颜色，每种颜色用一个二进制数来表示。例如，一幅黑白图像只有两种颜色，则每个像素点只需 1 个二进制位即可表示：1 表示黑色，0 表示白色。一幅 256 级灰度图像，每个像素点需要用 8 个二进制位来表示，表示 2^8=256 个亮度层次；一幅 24 位真彩色的 RGB 图像，其三原色红、绿、蓝的每种光的强度被分成 256 个级别（0～255），需要用 8 个二进制位来表示，每个像素点有三种颜色，所以共需要用 24 个二进制位来表示，表示 2^{24}=16777216 种颜色。

记录每个像素点的颜色或亮度所需的二进制位数，称为颜色深度（也称色彩位数）。对于彩色图像来说，颜色深度决定了该图像可以使用的最多颜色数目，颜色深度越大，显示的图像色彩越丰富，画面越自然、逼真。但要注意，在图像分辨率相同的情况下，颜色深度越大，图像所占的存储空间也越大。



分辨率256×256，颜色深度32　　分辨率64×64，颜色深度32　　分辨率32×32，颜色深度32

分辨率64×64，颜色深度8　　分辨率64×64，颜色深度4　　分辨率64×64，颜色深度1

图 1-25　不同分辨率和颜色深度的位图比较

这种由纵横排列的像素点组成的图像称为位图（又称点阵图）。位图的质量主要由图像分辨率和颜色深度决定。图 1–25 呈现的是采用不同分辨率和颜色深度数字化后的位图。

未经压缩的位图图像的数据量（单位：字节）= 图像分辨率 × 颜色深度 ÷ 8。

3. 编码

图像的编码就是按照一定的格式将位图上各个像素点的量化数据记录下来的过程。由于位图的数据量大，并且含有大量的重复数据，编码时一般采用数据压缩技术进行压缩和还原处理。不同的编码方法形成了不同格式的图像文件，如 BMP 格式、JPEG 格式等。

拓展阅读

计算机采用二进制的原因

1. 技术实现简单

人类早期设计的机械计算装置中主要用的是十进制。十进制数有 10 个基本符号，要用 10 种状态才能表示。使用电子器件的状态来表示 10 个基本符号过于复杂，而用电子器件的高电位与低电位或逻辑器件的开与关两种状态来表示两个基本符号就比较容易，所以二进制就成为电子计算机的数制基础。也就是说，电子器件的两种状态决定了电子计算机采用二进制来表示数据。

随着技术的发展，计算机的电子器件由电子管逐步变为晶体管、集成电路、大规模集成电路乃至超大规模集成电路，但电子器件的工作特点并没有改变。计算机是由逻辑电路组成的，逻辑电路通常只有两个状态，这两种状态可以表示“1”和“0”。这样的电路设计简单，而且只具有两种状态的电子器件容易找到，如继电器开关、灯泡、二极管、三极管等。所以至今，现代计算机仍然采用二进制来存储和表示数据。

2. 运算简单

二进制数的算术运算规则简单，有利于简化计算机内部结构，提高运算速度。同时，二进制只有两个数码，正好与逻辑代数中的“真”和“假”相吻合。因此，采用二进制可以简单方便地进行算术运算和逻辑运算。

另外，二进制数具有抗干扰能力强、可靠性高等优点，因为每位数字不是“0”就是“1”，当受到一定程度的干扰时，仍能可靠地分辨出它是“0”还是“1”。

单元挑战　认识并制作二维码

一、项目任务

二维码是近几年来移动设备上流行的一种编码方式。作为一种全新的数据存储、传递和识别技术，二维码的应用已渗透到人们生活的各个方面，如地铁广告、报纸、火车票、快餐店、电影院、团购网站及各类商品外包装上都能见到二维码，见图 1–26 和图 1–27。二维码被誉为“线上线下的一个关键入口”。

图 1–26　火车票上的二维码

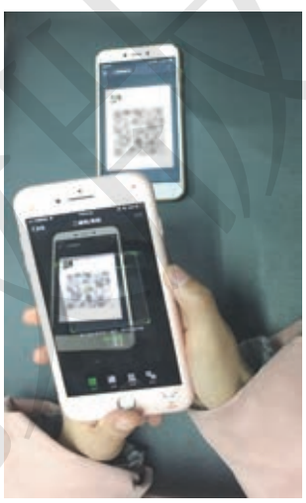


图 1–27　手机扫描二维码

以小组为单位，了解二维码的组成结构、编码原理、功能、分类、特点及应用，学习二维码的制作方法，设计并制作小组的二维码。

二、项目指引

1. 以小组为单位开展数字化学习，收集资料，了解二维码的相关知识，并加以梳理和归纳，用思维导图呈现小组的学习成果。

2. 收集资料，了解二维码的制作工具和方法、二维码的类型及可以承载的内容；小组成员共同规划设计小组的二维码，确定本小组的设计目标和通过二维码承载的信息（如小组名、小组图标、电子邮箱地址等）；选择合适的工具和方法制作小组的二维码。

三、交流评价与反思

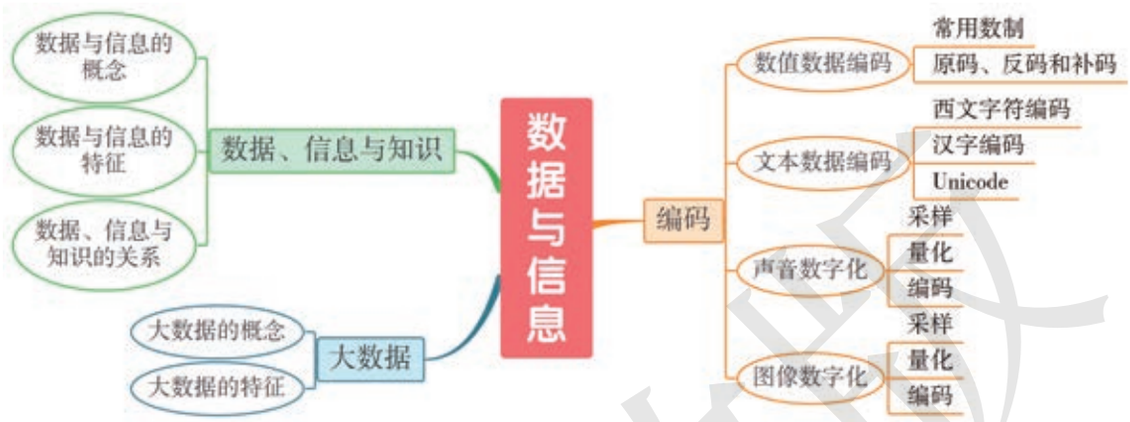
1. 各小组展示自己制作的二维码及介绍二维码的思维导图，分享学习体会。

2. 互相扫描各小组的二维码，对各小组制作的二维码、思维导图进行评价。

3. 小组成员共同回顾项目活动的过程，反思并交流本小组的收获与不足。

单元小结

一、主要内容梳理



二、单元练习

1. 某校正筹备运动会，需要为运动员编号。该校有高一、高二、高三共三个年级，每个年级 20 个班，每个班 45 ~ 55 名学生。请为该校设计运动员编码规则，保证每位运动员拥有一个唯一的编号，并能体现所在班级和性别。

2. 图书馆有一批纸质照片要转换成 JPG 格式的电子照片，刘楠小组承担了该任务。但刘楠发现大家提交的图像文件的大小和质量差距很大。排除摄影技术和环境因素，哪些参数会影响图像文件的大小和质量？为什么？请结合图像数字化的过程加以分析。

三、单元评价

评价内容	达成情况
能够通过实例分析，描述数据与信息的概念与特征（A、T、R）	
能够通过对日常生活情景的分析，阐述数据对人们日常生活的影响（A、I、R）	
在运用数字化工具的学习活动中，理解数据、信息与知识的相互关系（A、T、I、R）	
知道大数据的概念、特征和常见应用领域（A、T、R）	
理解编码的意义和作用，知道数据编码的基本方式（T）	

说明：A—信息意识，T—计算思维，I—数字化学习与创新，R—信息社会责任

第二单元

数据处理与应用

信息社会里，人们的生产、生活越来越依赖于数据的处理与应用。企业管理者面向消费者开展市场调研，利用调研得到的数据和信息提高决策的科学性，降低企业经营风险；足球赛场上，教练组根据数据分析师的实时数据分析制定最佳应对策略，及时调整攻防模式；驾车出行时，人们会查询交通线路，获取当前道路的实时车流量数据，确定恰当的出行方案……现代社会中无处不在的数据处理与应用帮助人们了解现状，预测未来，制订措施和方案。而能否准确、灵活地使用信息技术，熟练地进行数据的采集、分析和可视化等，很大程度上会影响人们工作、学习的效率和质量。

在本单元中，我们将一起探究身边的数据处理工作，了解数据的采集、分析和可视化的基本方法，经历利用软件工具或平台处理数据的过程。

单元挑战

采集与分析气象数据

学习目标

- 了解数据处理及其作用，能够根据任务需求，选用恰当的软件工具或平台处理数据，完成分析报告。
- 了解采集数据的基本方法，掌握通过公式、函数进行数据计算的方法。
- 掌握数据分析的基本方法，学习数据可视化的表示方法，能够读懂图表传递的信息。
- 理解对数据进行保护的意义。

项目二

调查中学生移动学习现状

——经历数据处理的一般过程

在信息技术飞速发展的时代，信息、知识的数量以指数增长的迅猛趋势，给传统的学习带来了机遇与挑战，让人们强烈地意识到终身学习的必要性与紧迫性。随着科技的不断发展与完善，移动互联技术渐趋成熟，各类移动设备层出不穷，让人们有了紧跟信息时代的“神器”。借助无线网络，利用移动设备随时随地获取学习资源，已成为一种新型的学习模式——移动学习（图 2-1）。移动学习代表着许多新的学习理念，如自主化学习、个性化学习、终身学习等。专家预言，移动学习必将成为未来学习的一个重要方式。

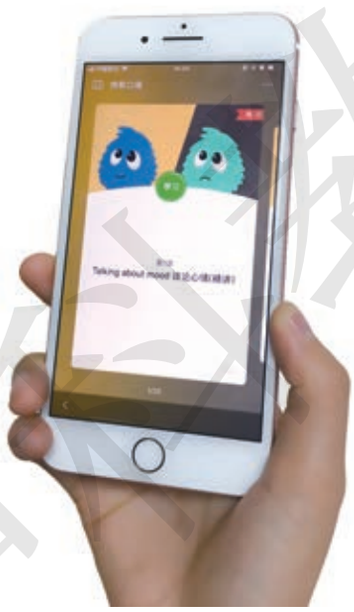


图 2-1　移动学习

当代中学生移动学习的现状如何？中学生对移动学习抱着怎样的认识和态度？中学生需要怎样的移动学习资源和移动学习环境？中学生移动学习的效率如何？中学生的移动学习存在哪些困难、哪些误区……这些是很多教育工作者、移动学习资源开发者和平台提供者迫切想要了解的问题。

项目学习目标

在本项目中，我们将围绕中学生移动学习现状，通过在线问卷调查采集数据，并借助软件工具完成数据分析和可视化，撰写调查报告，经历数据处理的全过程。

完成本项目学习，须回答以下问题：

1. 什么是数据处理？数据处理的过程是怎样的？
2. 如何认识数据处理的应用价值？
3. 什么是数据采集？什么是数据分析？数据分析的方法有哪些？
4. 什么是数据的可视化？它有哪些呈现方式？

项目学习指引

1. 明确数据需求

中学生移动学习现状调查是一项社会调查研究。社会调查研究是利用特定的方法和技术，从社会现实中采集研究需要的数据，通过数据的加工、分析和可视化，对社会现实进行描述或解释的认知活动。从技术角度看，社会调查研究就是围绕研究目的、需要研究问题而进行的一项数据处理工作（图 2-2）。

图 2-2　数据处理工作是调查研究的核心

调查正式开始前，需要根据调查的目的和研究的问题来明确数据需求。不同的调查目的，需要不同的数据来支持。如果研究者想通过“中学生移动学习现状调查”了解中学生使用移动学习的基本情况，则只需要采集使用时间、使用频率、使用资源类型等方面的数据；如果想了解哪些因素影响中学生的移动学习，则需要采集中学生对移动学习的态度、认识、周围环境因素（如家长和教师是否支持、移动设备的拥有情况）等方面的数据；如果还想通过调研，为移动学习资源、产品的开发者提供决策依据，则需要采集中学生对现有移动学习资源、产品的需求数据，以及中学生期待的移动学习资源类型和设计特点等方面的数据。

思考与讨论

日常生活中，你知道哪些利用数据帮助判断、决策、解决问题的例子？

核心概念

数据处理（data processing）是对数据进行采集、存储、加工、分析和表达的过程。

参见 P38 知识链接“数据处理及其作用”

可以利用思维导图软件，进行头脑风暴，帮助自己理清研究思路，如图 2-3 所示。

图 2-3　利用思维导图软件工具梳理数据需求

思考与讨论

面对中学生开展移动学习这一社会现象，不同的组织或个人，如教育部门、移动学习资源开发企业、家长，会面临或思考哪些问题？需要收集哪些数据帮助自己形成判断和决策？

活　动

3.1　班级同学分成若干小组，分别扮演不同的角色，如移动学习资源开发企业主、学校教师、家长。

（1）针对“中学生移动学习现状”，各小组分别确定各自的调查目标、想要研究的问题以及希望通过调查收集的数据。

（2）各小组绘制思维导图，梳理本小组的调查目的及需要采集的数据。

参见 P38 知识链接“数据处理的一般过程”

2. 采集数据

（1）选择合适的数据采集方法和工具

传统的社会调查方法主要有发放纸质问卷开展调查、面

对面访谈调查、实地考察等。信息技术的飞速发展大大丰富了人们采集数据的方法和手段。随着互联网、移动用户的增多，通过网络开展在线调查已成为很多研究者的选择。人们推出了很多在线调查平台、网站及在线调查系统（图 2-4），提供从在线问卷设计、数据采集（data acquisition）到数据加工分析的“一站式服务”。研究者要根据自己的需求，选择最适合自己的数据采集方法和工具。

图 2-4　各种在线调查平台

思考与讨论

1. 你了解并尝试使用了哪些在线调查平台？你认为这些平台各自有什么特色？

2. 如何判断在线调查平台的设计是否专业、是否可信赖？

（2）在线编辑发放问卷，采集数据

问卷是为了达到调研目的和采集必要数据而设计的一系列问题。问卷设计的好坏，直接关系到数据采集工作质量的高低。

小贴士

“一站式服务”实质就是利用信息技术对服务进行集成、整合，因此也被认为是“提供整体解决方案”。它在提高服务效率的同时，也提高了服务提供者的竞争力。

小贴士

移动学习软件可以自动采集用户基础数据和行为数据，如图 2-5 所示。

图 2-5　某移动学习软件用户随时可以查看自己的学习数据

活　动

3.2 寻找适合自己研究的数据采集方法和工具。

（1）小组成员各自上网查找、了解可以利用的在线调查平台，并尝试使用。

（2）小组交流讨论，对各种在线调查平台进行评估：

- 是有偿使用还是免费使用？
- 提供的服务有哪些？能否满足需要？
- 问卷编辑是否方便？能否方便地导入？
- 问卷发放有哪些渠道？数据导出格式有哪些？是否支持自选分析工具？
- 使用过该平台的人如何评价？

（3）小组成员协商、确定本小组选用的在线调查平台。

数字化学习

请利用网络开展学习，了解问卷的组成，明确问卷的设计原则及注意事项。分析几份调查问卷实例，掌握问卷的各种题型及其使用场合。

小贴士

在很多在线调查平台上编辑问卷时，可在屏幕上直观地得到即将打印到纸张上的效果，故也称可视化操作。这是一种“所见即所得”技术。

当前许多在线调查平台都提供多种问卷创建方式，一般有三种模式，一种是设计好纸质问卷并直接导入，另一种是在平台上根据提示自由创建问卷（图 2-6），还有一种是直接使用平台提供的模板通过简单编辑生成问卷（图 2-7）。

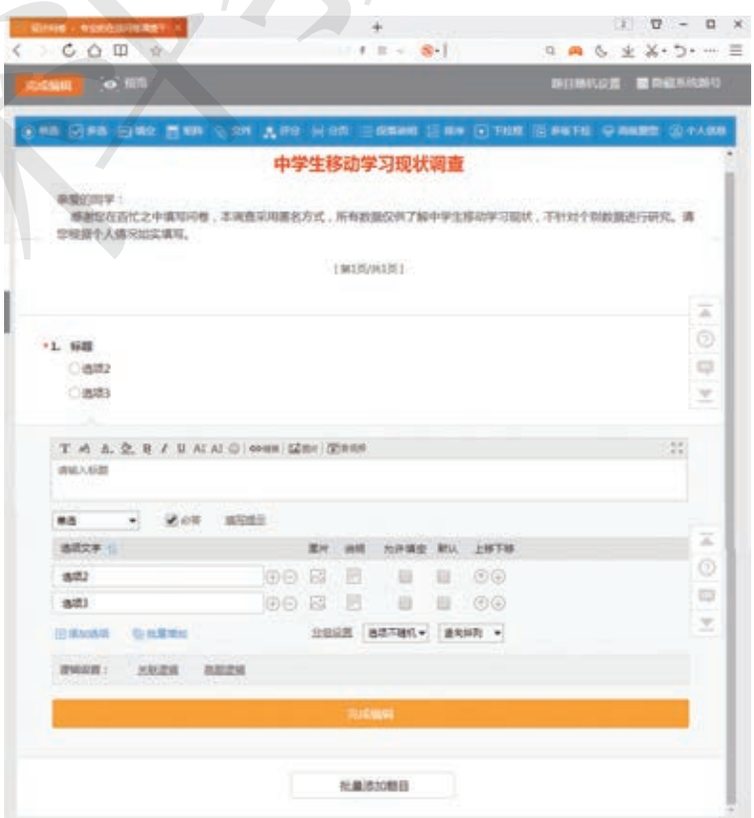


图 2-6　某在线调查平台的问卷编辑页面

思考与讨论

在线调查平台在编辑问卷时，要求先明确每个问题的类型，如单选题、多选题、排序题等。这是为什么？

图 2-7　某在线调查平台提供的模板

在线问卷可以利用平台中的相关设置来控制数据采集的目标人群和问卷发放数量等（图 2-8）。

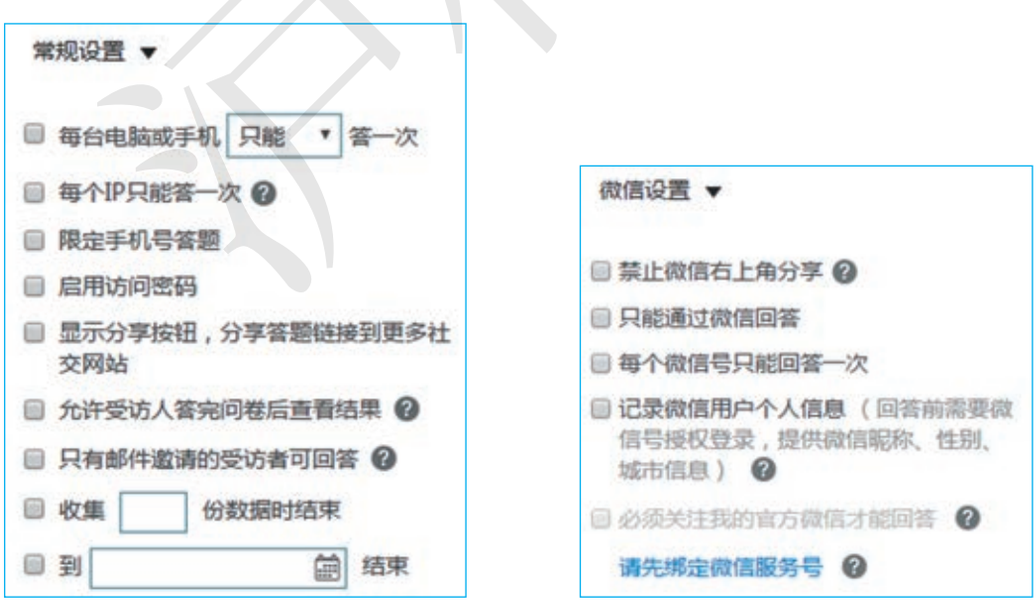


图 2-8　某在线调查平台的数据采集设置

思考与讨论

1. 小组选择的平台提供了哪些数据采集设置？每条设置会对调查取样产生怎样的影响？

2. 如果只想将问卷链接发送给本地区几所指定中学的学生，有哪些方法？采用这些方法可能会采集到大量非目标人群的数据吗？

（3）获取表格数据

传统纸质问卷回收后，需要研究者设计表格，手工录入纸质答卷上的数据。在线问卷则可以利用调查平台的相关功能，如“导出数据”，直接下载本次调查的答卷文件、表格数据（图 2-9）。

	A	B	C	D	E	F	G	H	I
1	序号	提交答卷时间	所用时间	来源	来源详情	来自IP	1、性别：	2、你的年级：	3、你平时住校还是走读？
2	1	2017/2/14 20:19:45	66秒	手机提交	直接访问	1.189.139.58(黑龙江-哈尔滨	女	高二	住校
3	2	2017/2/14 20:20:49	116秒	手机提交	微信	101.85.178.104(上海-上海)	男	高三	住校
4	3	2017/2/14 20:28:03	163秒	手机提交	直接访问	222.69.88.7(上海-上海)	女	高二	住校
5	4	2017/2/14 20:28:03	91秒	手机提交	微信	1.189.139.60(黑龙江-哈尔滨	男	高一	住校
6	5	2017/2/14 20:32:07	87秒	手机提交	直接访问	117.136.8.73(上海-上海)	男	高一	住校
7	6	2017/2/14 20:33:26	68秒	链接	http://192.168.	1.189.139.59(黑龙江-哈尔滨	男	高三	走读
8	7	2017/2/14 20:33:26	59秒	手机提交	直接访问	114.89.185.164(上海-上海)	男	高二	住校
9	8	2017/2/14 20:34:47	84秒	手机提交	直接访问	114.89.248.170(上海-上海)	男	高二	住校
10	9	2017/2/14 20:35:25	97秒	链接	http://192.168.	1.189.139.59(黑龙江-哈尔滨	女	高二	住校
11	10	2017/2/14 20:35:45	92秒	手机提交	直接访问	101.90.127.227(上海-上海)	男	高一	住校
12	11	2017/2/14 20:38:32	174秒	手机提交	微信	101.85.152.58(上海-上海)	女	高一	住校
13	12	2017/2/14 20:38:39	70秒	链接	http://192.168.	1.189.139.60(黑龙江-哈尔滨	男	高三	住校
14	13	2017/2/14 20:44:53	101秒	手机提交	微信	114.81.254.163(上海-上海)	女	高二	住校
15	14	2017/2/14 20:55:10	87秒	手机提交	直接访问	58.40.172.92(上海-上海)	男	高三	走读
16	15	2017/2/14 21:08:27	126秒	手机提交	直接访问	180.160.55.47(上海-上海)	男	高三	住校

图 2-9　某在线调查平台提供的答卷文件

思考与讨论

1. 传统纸质问卷通过“不记名”达到保护个人隐私的目的，但在线调查通过 IP 地址就可以跟踪数据来源。在线调查应该如何保护个人隐私？如何对数据进行保密？

2. 某同学在公交车站等车时，一位市场调查人员请他用手机扫一个二维码，说是在网上简单回答一些问题，即可获得一份小礼品。这位同学是否应该接受这位调查人员的提议呢？

活　动

3.3　各小组设计并发放自己的“中学生移动学习现状调查”问卷，采集数据。

（1）小组同学合作设计一份问卷。

（2）在小组选择的平台注册，并在平台上完成问卷的创建和编辑。问卷设计好后，各组自评或互评问卷，根据所提意见修改完善问卷。

（3）确定问卷的发放范围、途径、数量及调查结束时间，发放问卷，将链接发送给目标人群。

（4）问卷收集结束后，下载答卷文件，获取表格数据。

3. 加工、分析和可视化数据

调查问卷被收回意味着数据采集工作结束。接下来就要进行数据加工、数据分析和数据可视化了。

在线调查平台会在与用户的交互中，完成数据的加工、分析和可视化，用户可以直接查看结果，选择想要的结果呈现类型（图 2-10）。

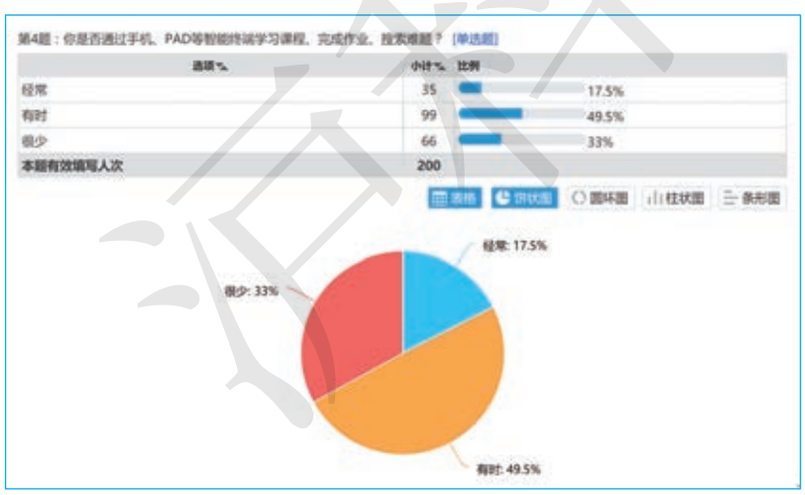


图 2-10　单选题统计结果表格和饼状图示例

在观察了关于每个问题的统计数据后，如果发现有些数据还需要进一步挖掘，可以利用平台提供的分析工具，如分类统计、交叉分析等，进一步完成一些比较复杂的数据分析工作。例如，若调查结果显示，75% 的中学生每周移动学习时间小于 1 小时，那么可进一步问，是否住校、不同性别所得

核心概念

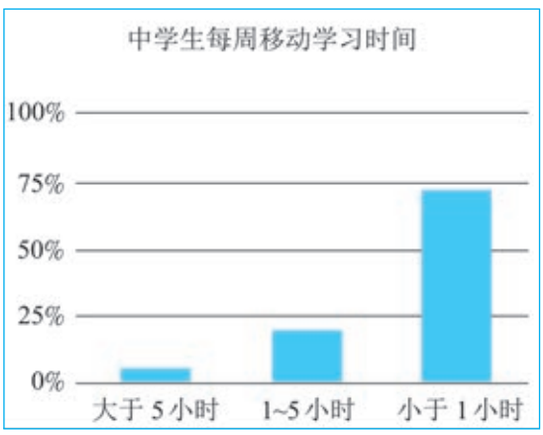
数据分析（data analysis）是指用适当的分析方法与工具，对采集的数据进行分类整理，提取与发现其中有价值的信息，以形成结论的过程。

数据可视化（data visualization）是指将数据分析的结果通过表格、图表、图形等形式显示出来。

数字化学习

请利用网络开展学习，了解利用电子表格软件制作图表的方法。

到的结果有没有差异？为了验证自己的假设，再对数据做交叉分析（图 2-11）。可以发现，是否住校对中学生移动学习的时间有比较大的影响。



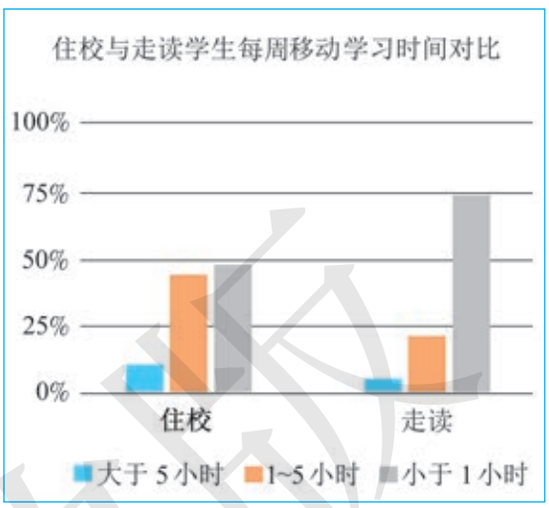


图 2-11　交叉分析及可视化示例

如果调查者认为平台提供的数据分析功能无法满足某些个性化或复杂的需求，还可以利用下载的答卷文件，运用专业的统计分析软件进行数据加工、分析和可视化。

思考与讨论

1. 问卷调查中常见的单选题、多选题、排序题，分别适合哪些统计分析方法？分别用什么类型的图表呈现效果较好？

2. 调查采集到的数据需要备份吗？为什么？如果打算备份数据，则应采用何种方法？

活　动

3.4 分析和可视化“中学生移动学习现状调查”数据。

（1）查看在线调查平台中每个问题的统计结果，选择合适的统计方法和呈现方式。

（2）观察数据及其初步统计的结果，展开小组讨论，提出新的假设，使用在线调查平台的交叉分析工具对问卷中的一些数据进行交叉分析，挖掘更多的信息。

（3）选择合适的可视化工具，将数据分析结果用图表等形式表达出来。

4. 撰写报告，提出数据应用建议

调查报告要清楚、准确地报告研究者为解决所研究的问题而做的一切工作。调查报告一般包括五个部分：研究背景和研究目的、调查对象和调查方法、调查结果、调查结论、意见和建议。进行汇报时要包括以下内容：

- 数据需求的产生（研究背景和研究目的）。
- 数据来源和采集数据的方法（调查对象和调查方法）。
- 数据的分析和可视化结果（调查结果）。
- 数据背后隐藏的信息（调查结论）。
- 数据应用（意见和建议）。

社会调查研究的最终目的是应用研究结果，为相关人员提供信息，帮助他们更好地决策。因此，“中学生移动学习现状调查”项目最后的工作就是选择合适的受众，向他们宣传自己的研究（图 2–12）。研究者也可以将自己的研究通过互联网与更多人共享。

小贴士

研究人员应该坚持实事求是、认真严谨的态度。研究报告应该全面而诚实地报告研究过程和研究结果，不能编造数据来支持自己的观点，更不能因为商业利益而在报告中宣传某些产品。

图 2–12　通过演示与他人分享研究成果

思考与讨论

1. 与传统的纸质问卷调查相比，在线问卷调查有哪些优势和劣势？

2. 对于现有的在线问卷调查平台，你有哪些改进意见和建议？

活　动

3.5　撰写“中学生移动学习现状调查报告”。

（1）了解报告的结构、数据的显示方式、结论与建议等。

（2）各小组撰写调查报告。

（3）全班互评调查报告。

（4）从数据来源和问卷设计等方面，反思本次问卷调查的数据对于本小组观点的支持度和有关结论的可信度。

知识链接

数据处理及其作用

数据并不是一堆枯燥无味的、单纯的数字，深挖数据背后的价值，可获取更多有用的信息。数据处理是从大量的原始数据中抽取出有价值信息的过程，即数据转换成信息的过程。它是对输入的各种形式的数据进行加工整理，这一过程包含对数据的采集、存储、加工、分析和表达。

数据处理的作用体现在三个方面：现状分析、原因分析和预测分析。

数据处理的应用无处不在，举不胜举。例如，某高校对学生在食堂刷卡吃饭的数据进行分析，确定受扶助学生的名单和资助金额，“偷偷”给这些学生的饭卡充钱，精准地扶贫；基于对用户搜索行为、浏览行为、评论历史和个人资料等数据的分析，某互联网企业向用户推荐他们可能喜欢的书籍、电影、美食或近期可能要购买的商品；某互联网公司大数据部上线的“疾病预测”，利用用户的搜索数据，并结合气温、湿度变化等因素建立预测模型，实时提供流感等疾病的活跃度、流行指数；交通部门基于用户和车辆的定位数据，分析道路拥堵的原因，从而针对不同时间点、不同道路的车流量进行智能车辆调度或采用潮汐车道。

数据处理的一般过程

数据处理的一般过程如图 2-13 所示。

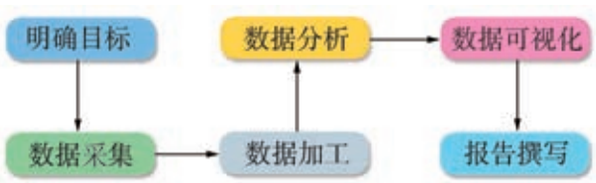


图 2-13　数据处理的一般过程

1. 明确目标

明确目标是指明确数据处理的目的，确立分析思路。首先，要思考开展数据处理的原因，即要解决什么问题。然后，要梳理数据分析的思路，搭建分析框架，确定使用哪些分析方法和工具。

2. 数据采集

数据采集是指人们根据需要获取数据，它是确保数据处理过程有效的基础。技术工具的发展使得数据采集方式日趋多样。目前数据采集的来源主要有以下几种：

- 人工输入的观察、调研数据；
- 利用技术工具（例如传感器）直接采集的数据；
- 各种数据库中的数据；
- 利用搜索引擎工具在网络上快速获取的数据；
- 通过网络调查问卷采集的数据。

3. 数据加工

数据加工是指通过数据编码、数据清洗、数据重组等一系列过程，使采集到的数据符合数据分析的需求。

在本项目中，直接使用平台提供的功能对采集到的问卷数据进行统计分析，省略了数据加工环节。但在现实工作中，采集到的数据大多不能立即用于数据分析，还须使用恰当的工具和方法进行加工。以问卷调查为例，须剔除出现答案残缺不全、重复填写、数据错误等问题的答卷；还有些数据须进行编码，如性别、年级、是否住校等选择题答案（图2–14）；而开放式问题得到的回答会非常多，显得杂乱无章，须分类并确定代码。

G	H	I
1、性别：	2、你的年级：	3、你平时住校还是走读？
女	高二	住校
男	高三	住校
女	高二	住校
男	高一	住校
男	高一	住校
男	高三	走读
男	高二	住校

→

G	H	I
1、性别：	2、你的年级：	3、你平时住校还是走读？
2	2	1
1	3	1
2	2	1
1	1	1
1	1	1
1	3	2
1	2	1

图 2–14　未经编码的数据表（左）和经过编码的数据表（右）

4. 数据分析

数据分析是指用适当的分析方法与工具，对采集到的数据进行分类整理，提取与发现其中有价值的信息，形成结论的过程。数据分析的目的是从描述研究对象的数据中，发现其内在特征和规律。数据分析有对比、细分和预测三大类，它们又各自对应不同的具体分析方法。

在日常工作和现状研究中，运用最多的是描述性分析方法，如对比分析法、平均分析法和交叉分析法。

（1）对比分析法

对比是人们认识客观世界的基本方法。通过将两个或两个以上的数据进行对比，分析它们的差异，可以分辨数据的性质、变化、发展等个性特征。对比分析法可以分为横向比较和纵向比较（图 2–15）。横向比较是同一时间不同总体指标的对比，纵向比较是不同时间同一总体指标的对比。

（2）平均分析法

利用计算平均数的方法，可以反映总体在一定时间、地点下数据特征的一般水平。平均分析法可以分为位置平均数和数值平均数（图 2–16）。其中，运用得最多的是算术平均数。

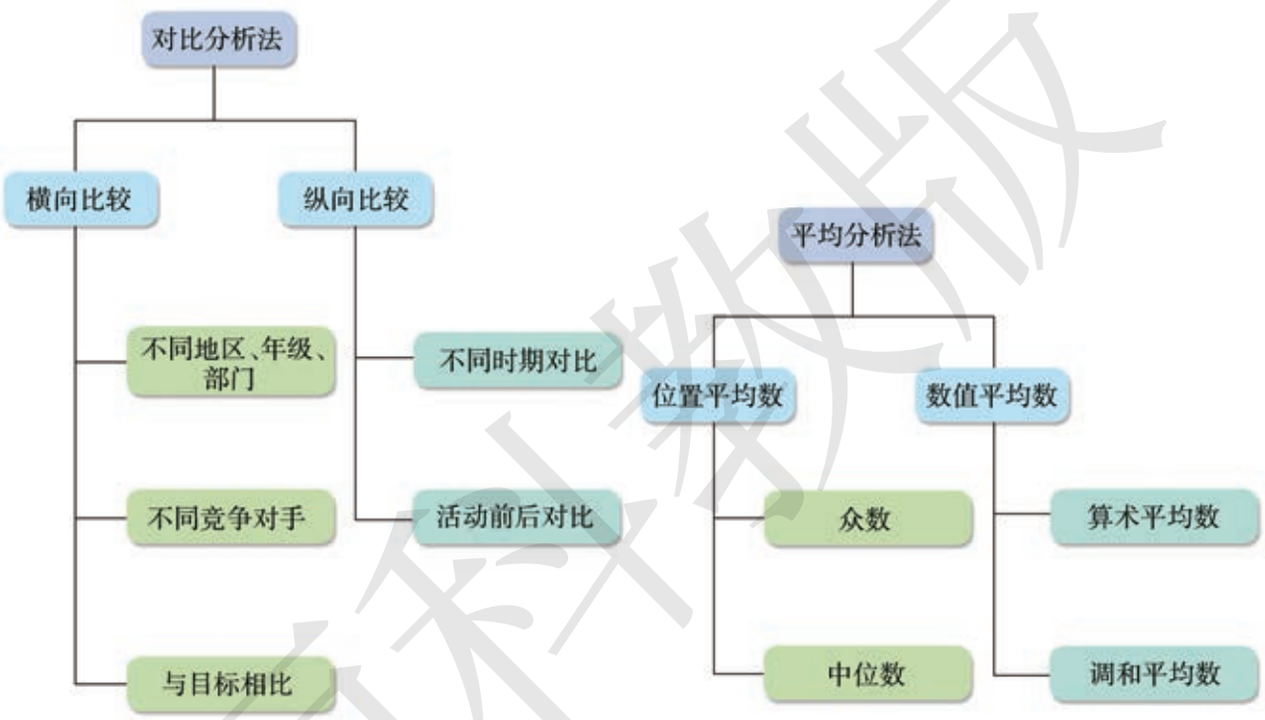


图 2–15　对比分析法

图 2–16　平均分析法

（3）交叉分析法

这是一种立体分析法，它从横向和纵向两个方向来计算两个或多个有联系的变量在交叉点的统计值。

5. 数据可视化

数据可视化是指将数据分析的结果通过表格、图表、图形等形式显示出来，还可以通过这些形式对分析结果进行一些交互处理。利用人对形状、颜色等特性的感官敏感性，数据可视化能更清晰、有效地帮助人们发现数据之间的关系、规律和趋势，传递数据背后的信息，如图 2–17 所示。

常见的数据图表包括条形图（图 2–18）、折线图、饼图（图 2–19）、柱状图（图 2–20）、面积图、散点图、雷达图等，使用图表工具还可以得到交互图表（图 2–21）、漏斗图、帕累托图、旋风图、矩阵图等，数据可视化图形则包括地图、词云（图 2–22）、热力图（图 2–23）、树图、网络图等，甚至可以是动图、动画。

图 2-17　利用数据可视化工具呈现的某游乐园游客实时数据

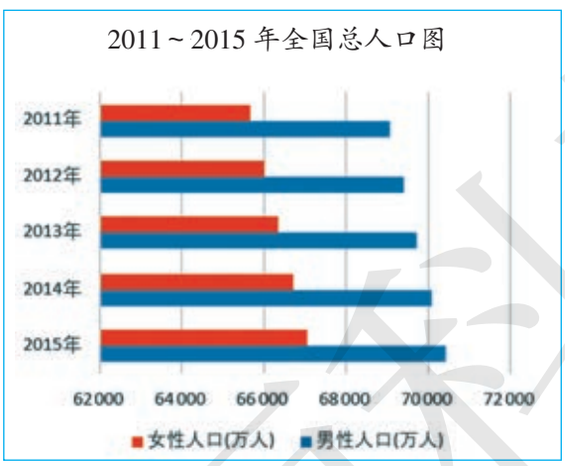


图 2-18　条形图示例

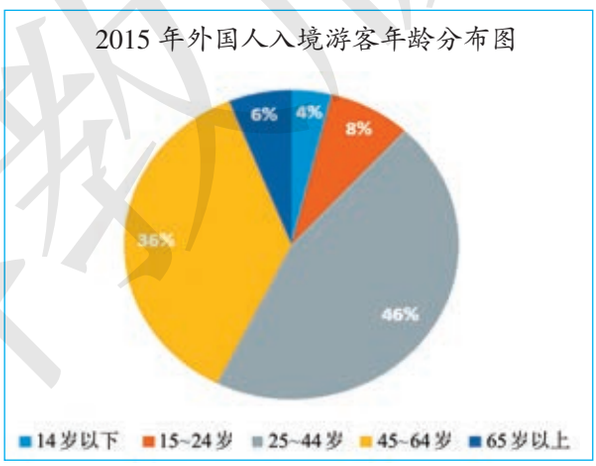


图 2-19　饼图示例

图 2-20　柱状图示例

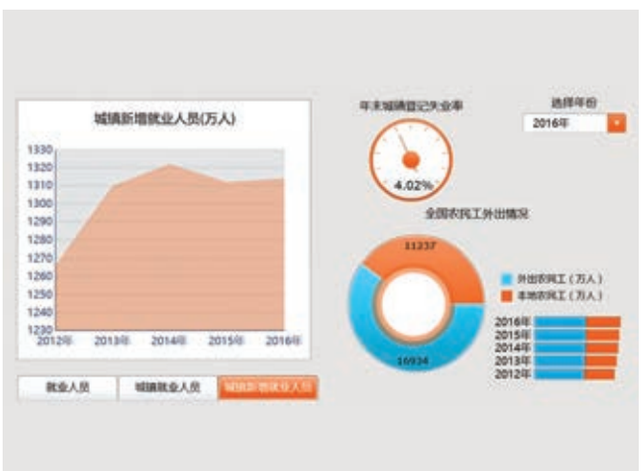


图 2-21　交互图表示例

图 2-22　词云示例

图 2-23　热力图示例

最常用的图表工具是 WPS 表格或 Excel 等电子表格软件，在互联网上有大量优秀的数据可视化工具，如 RAW、Infogram、Tableau 等。除了利用现成的工具，也可以按自己需要编程实现。

6. 报告撰写

报告撰写是对整个数据处理过程的总结。通过报告，将数据处理的目的、过程、结果及方案完整地呈现出来，为决策提供参考或依据。报告的种类很多，但不管采用怎样的呈现方式，都应做到清晰可读，尤其要注重数据可视化，以便于阅读者正确、迅速地理解报告内容。此外，报告不仅要发现问题，更要有依据科学、严谨的数据分析过程推导出来的结论和建议。

项目四

认识智能停车场中的数据处理
——体验数据处理的方法和工具

在面积较大、空闲率低的停车场，尤其是层数较多的地下停车场，找到空闲车位往往是让驾驶员头疼的事。为了帮助驾驶员快速找到空闲车位，避免因盲目找车位造成停车场内部通道的堵塞，近年来，人们利用信息技术设计研发了智能停车场管理系统，如图2-24所示。在智能停车场中，停车引导和车辆收费是两项重要的工作。不同于传统的由人工登记进出车辆相关数据的做法，智能停车场能自动记录停车位的使用数据和车辆的进出数据，并自动计算车辆的停车时间和费用。有些智能停车场，还会对多年积累的车位使用数据、缴费数据等进行分析，为停车场的科学管理提供依据。在有些城市，市中心不少智能停车场的数据都被接入城市智能交通系统，再通过街上的引导屏，告知周边停车场的空闲车位数，引导车辆找到停车位。

图2-24　某智能停车场

项目学习目标

在本项目中，我们首先通过探究智能停车场的车辆引导和停车费计算这两项工作，了解数据采集、组织和计算的基本方法。然后学习使用一种数据处理工具，对智能停车场数据库中记录的停车位数据进行处理，获取隐藏在这些数据中的信息。

完成本项目学习，须回答以下问题：

1. 采集数据的具体方法和工具有哪些？
2. 数据的组织方式是怎样的？
3. 数据的存储方式有哪些？
4. 表格数据的加工方法有哪些？

图 2-25　户外车位引导屏

项目学习指引

1. 探究停车引导中的数据处理

智能停车场往往在停车场入口、各层各区域的交叉路口设立引导屏，显示行车路线指引信息和空闲车位数（图 2-25）。同时，在每个车位的上方，根据车位的使用情况控制车位指示灯显示不同的颜色——绿色为“空闲”，红色为“占用”。驾驶员在几十米外即可看到指示灯，方便他们快速找到空闲车位。

思考与讨论

要统计出图 2-26 户外车位引导屏中的空闲车位数，需要采集每一个车位的哪些数据？与一个停车场中的车位相关的数据还有哪些？

根据户外车位引导屏指示进入车库

根据区域引导屏指示进入有空闲车位的区域

根据车位指示灯快速找到空闲车位

图 2-26　停车引导过程

小贴士

数据的自动采集是指利用技术手段（如传感器、摄像头）从系统外部直接采集数据并输送到系统内部接口的过程。

要实现停车引导，首先应该了解车位占用情况，再将采集到的数据保存下来，并进行分析。

（1）车位占用情况数据的自动采集

大型停车场通常采用多层多区域的组织方式，车位占用情况数据（空闲或占用）不是用人工记录，而是通过传感装置自动采集。图 2-27 是使用超声波传感器采集车位占用情况的示意图。

图 2-27　超声波传感器采集车位占用情况

参见 P56 知识链接“数据采集的方法和工具”

安装在车位上方的超声波传感器自上而下发出超声波，并据此探测、分析物体或地面的反射波，精确测量出反射面到传感器的距离，由此准确地检测出每个车位的占用情况，从而实时采集到车位占用情况数据，同时将该数据用车位指示灯直观呈现。

思考与讨论

某居民小区拟在出入口安装车牌识别摄像头，但部分车主担心自己的个人信息、车牌信息、停车数据等隐私被泄露。你觉得他们的顾虑有道理吗？

参见 P57 知识链接“数据的保护”

活　动

4.1　调查某个居民小区停车管理的数据采集方式。

有些大型居民小区已对小区的停车实现了智能化管理。小区停车位有在地面的，有在地下的；有出售给业主的，有租赁给业主的，还有供临时停放的。有固定车位的业主不用在每次进出小区时缴费，临时停放在小区的车辆离开时需要按规定缴费。智能停车管理系统会在小区入口实时采集进入车辆的数据，在出口对离开车辆进行识别，并对临时停放的车辆根据车辆出入时间计算停车费。

（1）选择一个居民小区开展调查，了解其停车收费管理中需要用到哪些数据，以及这些数据是如何采集的。

（2）了解该居民小区在入口采集车辆数据的方式、采集数据所使用的硬件设备，以及获取的数据。

方式 1：_____　数据采集设备：_____　获取的数据：_____

方式 2：_____　数据采集设备：_____　获取的数据：_____

（2）停车位数据的组织

智能停车场引导屏上的空闲车位数是按停车场的层、区域甚至方位汇总得到的。停车引导系统采集好车位占用情况数据后，除了改变空闲车位数，还要输送车位的所在层、所在区域等数据。

停车位数据到底包括哪些内容？这个问题关系到如何有效构建数据。任何一个事物都包含许多属性，事物的全部属性可按需求选择性地呈现。要解决的问题不同，对同一事物要呈现的属性也不同。在应用系统中，由若干属性构成的数据称为结构化数据（structured data）。

停车引导工作的需求是汇总不同层、不同区域的空闲车位数，统计时往往要用到车位占用情况、层、区域等属性，而无须关心车位尺寸等属性。例如，某停车场的停车位使用实时数据包括序号、采集时间、层、区域、车位占用情况、状态（对应该停车场的管理情况，如将车位分为开放、关闭、内部），其结构图如图 2-28 所示。

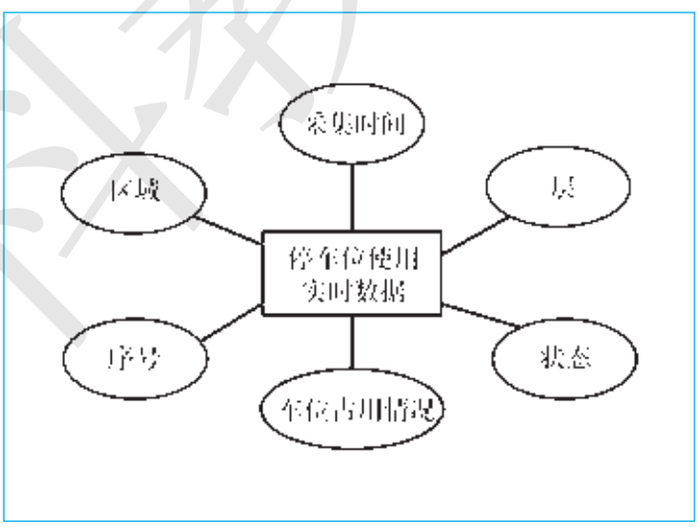


图 2-28　停车位使用实时数据结构图

小贴士

关系表（relational table）：一种规范化的二维表，符合关系模型的要求。

在大多数的数据处理中，通常以关系表的形式组织数据，如图 2-29 所示的停车位使用实时数据表。表格的第一行是标题，列出“序号”“采集时间”等属性名称。从第二行开始，每一行是一条记录（record）。每一列是一个属性，称为表格的字段（field）。行列交叉处是一个单元格，存放一条记录的某个字段值。

序号	采集时间	层	区域	编号	车位占用情况	状态
1	2016-12-8 14:30:00	B1	A	10	0	内部
2	2016-12-8 14:30:00	B1	A	11	0	内部
3	2016-12-8 14:30:00	B1	B	1	1	开放
4	2016-12-8 14:30:00	B1	C	1	1	开放
5	2016-12-8 14:30:00	B1	D	1	1	开放
6	2016-12-8 14:30:00	B1	D	2	0	开放
7	2016-12-8 14:30:00	B2	A	1	0	关闭
8	2016-12-8 14:30:00	B2	A	2	0	关闭

字段

记录

图 2-29　停车位使用实时数据表

（3）停车位数据的存储

如图 2-30 所示，停车引导工作本质上是一个数据处理过程。停车位使用实时数据表就存储在停车场服务器的数据库中，供停车场管理者查询实时数据或分析历史数据。

要将数据存储在数据库中，首先要创建关系表的结构，然后将记录添加到关系表中。创建关系表结构时要确定关系表中每一个属性的数据类型。为了保证数据的有效性，还要设置属性值的数据约束，包括属性值是否唯一、是否可以为空、是否要在一定的数值范围内等。

小贴士

数据库（database）是按照特定的数据结构（data structure）来组织、存储和管理数据的、建立在计算机存储设备上的仓库。不同类型数据在计算机内存储和处理的方式不相同，因此数据库中的数据必须明确其数据类型。

参见 P58 知识链接“数据的组织和存储”

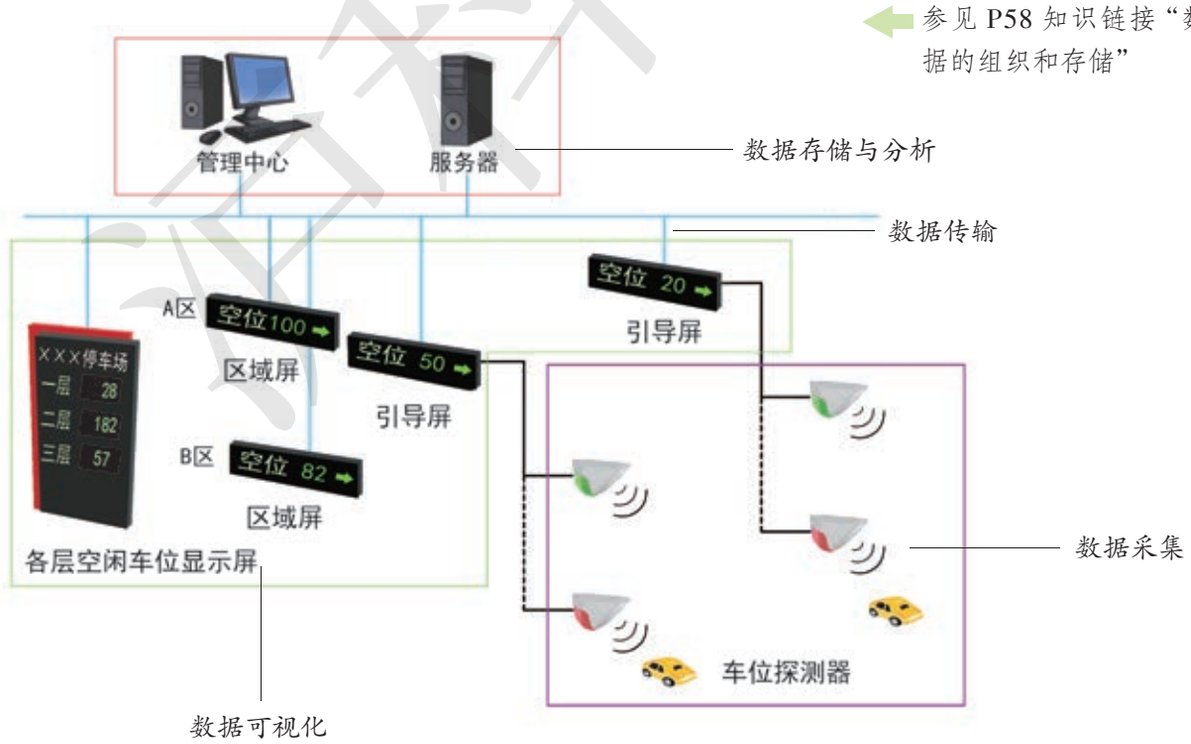


图 2-30　停车引导中的数据处理

活　动

4.2 设计居民小区停车位管理表。

居民小区的停车管理是小区物业的重要工作。请为活动 4.1 中所调查的居民小区设计一张停车位管理表，方便物业管理人员了解停车位的分配情况（已出售、已租赁、待租赁、临时），以及已租赁停车位的月租费缴纳情况。

（1）设计停车位管理表。

（2）使用一种电子表格软件创建停车位管理表。

（3）分析表中的各项属性值可以从哪里采集。

4.3 表 2-1 是某停车场的停车位使用实时数据表的字段构成。请用“数值”“文本”“日期 / 时间”或“逻辑”填写各字段的数据类型，用“唯一”或“非空”填写数据约束，也可以设定属性的合理数值范围，便于出错时进行检查。

表 2-1　停车位使用实时数据表的字段构成

属性	序号	采集时间	层	区域	编号	车位占用情况	状态
数据类型							
数据约束							

2. 计算停车费

图 2-31　某停车场收费规定

在停车场管理中，除了停车引导，计算停车费也是一项重要工作。为了计算停车费，停车场大多在入口处采集车辆的驶入时间，在出口处采集车辆的驶出时间。根据驶入时间和驶出时间，计算车辆在停车场的停留时间，再根据停车收费规定计算停车费。

在智能停车场中，停车费的计算工作是由系统自动完成的。例如，按照图 2-31 的停车收费规定，利用电子表格软件这一常见的数据处理工具，可以模拟停车费的计算工作。

电子表格软件一般以工作表来组织数据，一张工作表由若干个单元格构成。每个单元格可以存储一个数据，单元格的值可以直接手工输入，也可以通过公式计算得到。利用查询等功能，可以从停车场数据库中导出数据，再在电子表格软件中导入数据，得到类似图 2-32 所示的表格。

	A	B	C	D
1	序号	车牌	驶入时间	驶出时间
2	1	沪A8735	2016/02/28 11:00	2016/02/28 11:25
3	2	沪A3949	2016/02/28 07:50	2016/02/28 09:45
4	3	沪C0381	2016/02/28 22:00	2016/02/29 05:30
5	4	沪A4398	2016/02/28 10:23	2016/03/03 13:43

图 2-32　从停车场数据库中导出的表格数据

每个单元格都有一个引用名称，例如 A1 表示第 1 行第 1 列的单元格。多个单元格在行列方向上连续排列而构成的矩形称为区域，引用名称由左上角单元格和右下角单元格构成。例如 B3：C4 表示从第 3 行第 2 列到第 4 行第 3 列的矩形区域，包含 4 个单元格。

参见 P58 知识链接“表格数据的加工方法”

思考与讨论

1. 在上表的 C 列和 D 列中新插入一列后，序号 1 的驶出时间单元格的引用名称仍是 D2 吗？为什么？

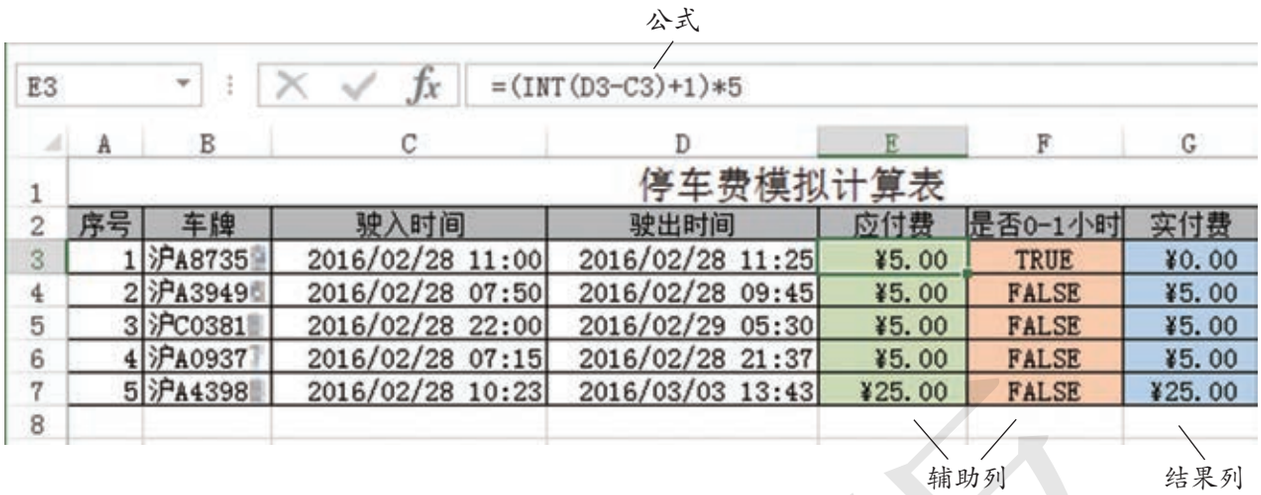
2. 讨论上表中各个字段的计算特性，确定每个字段的数据类型。

根据停车收费规定，可以在以上导出表格的基础上添加辅助列和结果列，并利用软件内置的公式和函数计算停车费用，如图 2-33 所示。

电子表格软件中的公式（formula）是由常量、单元格、运算符、函数等构成的表达式，由“=”引出。

函数（function）则是软件内置的一些常用功能模块，在使用的时候可以代入参数，获得相应的计算结果。图 2-30 的公式中使用的 INT 函数的作用是取整数。

时间 / 日期数据在计算机中表示为一个数值型的编码，其中日期是长整型，时间则是小于 1 的实数。时间 / 日期数据可以进行日期、时间的加减运算。两个时间 / 日期数据做减法，得到的差值单位为天；若要以小时进行计算，可以再乘以 24。



	A	B	C	D	E	F	G
1	停车费模拟计算表						
2	序号	车牌	驶入时间	驶出时间	应付费	是否0-1小时	实付费
3	1	沪A8735	2016/02/28 11:00	2016/02/28 11:25	¥5.00	TRUE	¥0.00
4	2	沪A3949	2016/02/28 07:50	2016/02/28 09:45	¥5.00	FALSE	¥5.00
5	3	沪C0381	2016/02/28 22:00	2016/02/29 05:30	¥5.00	FALSE	¥5.00
6	4	沪A0937	2016/02/28 07:15	2016/02/28 21:37	¥5.00	FALSE	¥5.00
7	5	沪A4398	2016/02/28 10:23	2016/03/03 13:43	¥25.00	FALSE	¥25.00
8							

图 2-33　停车费模拟计算表

思考与讨论

1. 图 2-30 的公式中为什么会出现“+1”？

2. 在 E3 单元格中输入公式后，如何将公式应用于其他单元格（E4 至 E7）？你能想到几种方法？

3. 若在“驶入时间”列前插入一空列，则原来的“驶入时间”列的单元格 C3：C7 变成了 D3：D7，这会影响辅助列和结果列的值吗？为什么？

活　动

4.4 为某停车场计算停车费。

某停车场收费规则为：半小时以内免费，超过半小时不到 1 小时收费 5 元，之后每小时加收 2 元，但 24 小时内最高收费限额为 20 元，超过 24 小时则重新按上述规定计费。

（1）按照以上收费规则，设计停车费计算表，补充表 2-2 的表头。

表 2-2　停车费计算表

序号	车牌	驶入时间	驶出时间					

（2）选择一种电子表格软件，导入本项目中从数据库导出的数据。

（3）创建停车费计算表，并利用公式和函数计算停车费。

3. 分析停车位使用数据

停车引导和车辆收费等数据处理工作是智能停车场的日常业务活动。日积月累，数据库中会存储大量的数据。例如，按一定时间间隔（半分钟甚至更短）采集每个车位的占用情况数据，一天下来，数据库中可能存储上万或几十万条记录，一周、一个月、一年的数据累计量更是庞大。而对这些数据进行分析和挖掘，可以得到新的信息，为停车场的决策提供支持服务。

（1）分析某一时刻各停车位实时使用情况

例如，某停车场接到通知：下周四 14:30，停车场附近有临时性的展览活动，预计会有大量车辆驶入。停车场管理人员想了解：该时间段停车场的接纳量大概是多少？哪些区域空闲车位多？该时间段如何开展停车引导？

① 针对此任务，数据分析人员从数据库中导出类似时刻（如采集时间为 2016 年 12 月 8 日周四 14:30）的所有记录，得到相应的数据表，并在电子表格软件中打开，如图 2-34 所示。

序号	采集时间	层	区域	编号	车位占用情况	状态
1	2016-12-8 14:30:00	B1	A	10	0	内部
2	2016-12-8 14:30:00	B1	A	11	0	内部
3	2016-12-8 14:30:00	B1	B	1	1	开放
4	2016-12-8 14:30:00	B1	C	1	1	开放
5	2016-12-8 14:30:00	B1	D	1	1	开放
6	2016-12-8 14:30:00	B1	D	2	0	开放
7	2016-12-8 14:30:00	B2	A	1	0	关闭
8	2016-12-8 14:30:00	B2	A	2	0	关闭

图 2-34　从数据库中导出的某时刻停车位使用实时数据表

② 对数据表中的数据进行汇总统计。数据透视表工具是一种快速汇总大量数据的交叉分类统计工具。利用数据透视表工具可以对数据表中的数据在行、列方向上重新布局分类字段，在行列交叉处按照设定的统计方法（如求和、计数、最大值、最小值、方差）计算统计值。使用数据透视表工具进行交叉分类统计，选择图 2-35 中的行、列、数值组合及值字段设置，就可得到某时刻停车位数据透视表，如图 2-36 所示。

小贴士

为更好地反映某时刻的车位占用情况，一般还需要导出更多数据来减少误差，如导出近三个月每周四 14:30 的数据，或者近一个月每个工作日 14:30 的数据等。

小贴士

分类字段的数值应有区分性，一个值表示一个类别，且数值范围是一个有限集合。如停车位使用实时数据表中的字段“层”“区域”“状态”都可作为分类字段。

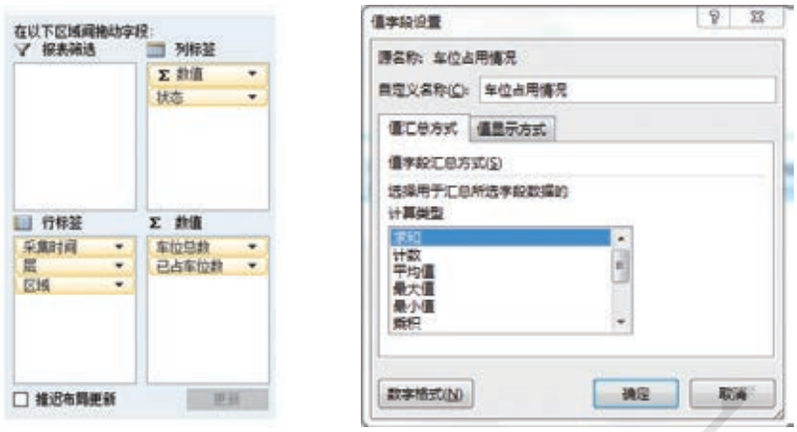


图 2-35　某电子表格软件的数据透视表设置

采集时间	层	区域	车位总数 关闭	开放	内部	已占车位数 关闭	开放	内部	车位总数汇总	已占车位数汇总
2016-12-8 14:30:00	B1	A			25			11	25	11
2016-12-8 14:30:00	B1	B		30			20		30	20
2016-12-8 14:30:00	B1	C		36			23		36	23
2016-12-8 14:30:00	B1	D		30			14		30	14
2016-12-8 14:30:00	B1	E		25			23		25	23
2016-12-8 14:30:00	B1 汇总			121	25		80	11	146	91
2016-12-8 14:30:00	B2	A		30			5		30	5
2016-12-8 14:30:00	B2	B		31			11		31	11
2016-12-8 14:30:00	B2	C	30			0			30	0
2016-12-8 14:30:00	B2	D	32			0			32	0
2016-12-8 14:30:00	B2	E	32			0			32	0
2016-12-8 14:30:00	B2 汇总		94	61		0	16		155	16
2016-12-8 14:30:00 汇总			94	182	25	0	96	11	301	107
总计			94	182	25	0	96	11	301	107

图 2-36　某时刻停车位数据透视表

从图 2-36 所示的数据透视表中，可以读到该时刻每一层每一区域的车位总数和已占用车位数，以及在不同状态（关闭、开放、内部）的统计数据。

思考与讨论

1. 在图 2-36 所示的数据透视表中，整个停车场分为几个区域？如何准确描述一个区域？

2. B1 层 A 区域内部车位的使用情况如何？B2 层 B 域区所有车位的使用情况如何？

3. B1 层有多少空闲车位？整个停车场有多少空闲车位？有多少内部车位？

小贴士

排序是对表中的一列或多列数据按指定顺序（升序或降序）重新显示。多关键字排序时，对表中数据先按主要关键字排序，对主要关键字值相同的数据再按次关键字排序，以此类推。

③ 对图 2-36 所示的数据透视表，使用排序工具按已占用车位数汇总的数据顺序进行重新排列（图 2-37），可以清楚地查看某一时刻各个停车区域的车位饱和情况，并找出各层中相对空闲的区域，如图 2-38 和图 2-39 所示。

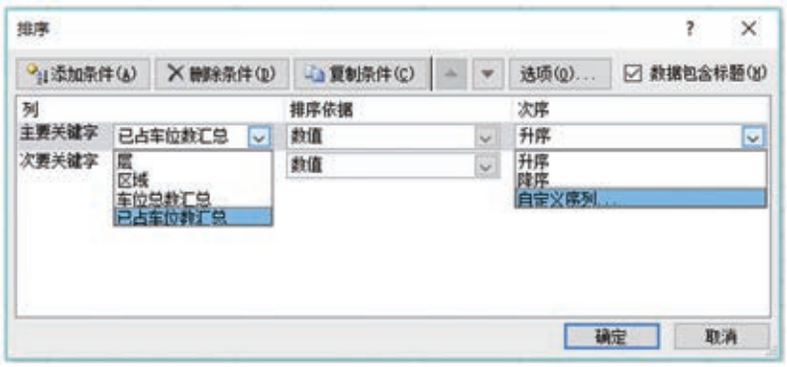


图 2-37　排序工具

层	区域	车位总数汇总	已占车位数汇总
B1	A	25	11
B1	B	30	20
B1	C	36	23
B1	D	30	14
B1	E	25	23

图 2-38　数据透视表中 B1 层的汇总数据（源数据表）

层	区域	车位总数汇总	已占车位数汇总
B1	E	25	23
B1	C	36	23
B1	B	30	20
B1	D	30	14
B1	A	25	11

图 2-39　排序后的目标数据表

④ 对图 2-33 所示的数据透视表，使用筛选工具查看某一时刻各层关闭的车位总数，可供管理人员对活动日当天是否将关闭的保留车位对外开放进行决策，如图 2-40 和图 2-41 所示。

小贴士

筛选就是留下符合条件的数据。很多数据处理软件都提供了筛选工具，数据透视表工具的每一列（除汇总数据）也都支持筛选操作。

图 2-40　筛选工具

采集时间	层	区域	车位总数 关闭	已占车位数 关闭
2016-12-8 14:30:00	B2	C	30	0
2016-12-8 14:30:00	B2	D	32	0
2016-12-8 14:30:00	B2	E	32	0
2016-12-8 14:30:00	B2 汇总		94	0
2016-12-8 14:30:00 汇总			94	0
总计			94	0

图 2-41　某一时刻 B2 层关闭的车位总数

（2）挖掘停车位历史数据

例如，某停车场接到通知：下周四 6:30 到 22:00，在停车场附近将安排一个一小时左右的活动，预计会有大量车辆驶入。为解决停车问题，希望停车场管理人员给出建议：下周四哪个时间段开展活动较好？为此，停车场管理人员需要了解平时工作日 6:30 到 22:00 各层、各区域的使用率，进而预测活动当天停车场的使用低谷时段。

① 针对此任务，数据分析人员从停车场数据库中导出类似的某天 6:30 到 22:00 的停车位使用实时数据表，筛选出以 10 分钟为间隔的记录，并按时间、层、区域分类统计已占用车位数，再计算出车位占用率（已占用车位数 / 该区域车位总数），得出各层各区域车位占用率数据表，如图 2-42 所示。

② 利用各层各区域车位占用率数据表中的数据，可以绘制出多种折线图，如图 2-43 和图 2-44 所示。

时间	层	区域	已占车位数	车位占用率
6:30	B1	A	0	0%
6:40	B1	A	0	0%
6:50	B1	A	0	0%
7:00	B1	A	2	8%
7:10	B1	A	4	16%
7:20	B1	A	5	20%
7:30	B1	A	5	20%
7:40	B1	A	5	20%
7:50	B1	A	5	20%
8:00	B1	A	5	20%
8:10	B1	A	6	24%
8:20	B1	A	6	24%
8:30	B1	A	7	28%
8:40	B1	A	7	28%
8:50	B1	A	10	40%
9:00	B1	A	10	40%
9:10	B1	A	12	48%
9:20	B1	A	12	48%
9:30	B1	A	13	52%
9:40	B1	A	13	52%
9:50	B1	A	13	52%
10:00	B1	A	13	52%

图 2-42　各层各区域车位占用率数据表（部分）

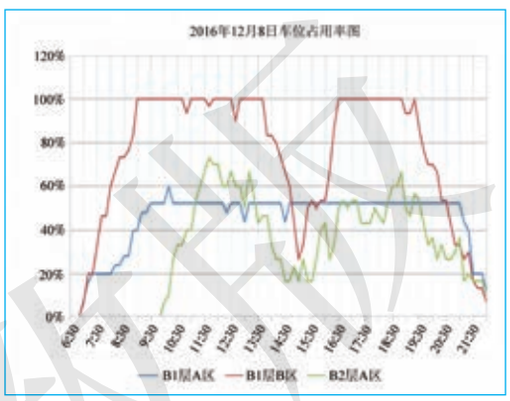


图 2-43　部分区域车位占用率图

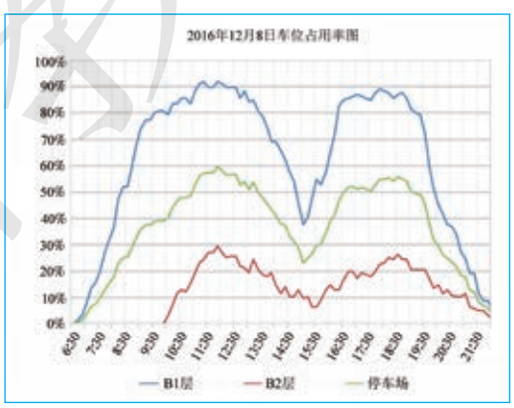


图 2-44　各层车位占用率图

从图 2-43 和图 2-44 中，可以明显看出各层各区域车位占用率的高低，以及车位占用高峰时间段，不同区域的占用率呈现出不同的趋势状态。

思考与讨论

1. 如果要得到如图 2-44 所示的各层车位占用率图，该如何处理数据表中的数据？

2. 能不能从图 2-44 中得出结论“9:00~13:30、16:30~19:30 是该停车场的高峰时期，建议 9:00 开放 B2 层”？

通过分析各层各区域车位占用率数据表，停车场管理人员可以考虑实施动态优惠停车方案（在停车场的空闲时段实施车费优惠，车辆在该时段进入并离开停车场即可享受该优惠）等进一步的决策。

此外，生活中还有许多类似的数据处理工作，如图书馆的新书入库、借书、还书等日常业务活动，以及读者借书趋势、特定图书借阅频度等数据分析工作。图书馆管理人员也可以运用类似的数据分析方式来工作，为相关图书馆业务决策提供参考。

活　动

4.5　帮助停车场管理人员统计某一时刻停车场的空闲车位，计算停车场的实时车位占用率。

（1）利用以上项目中从数据库导出的某时刻停车位使用实时数据表，制作数据透视表，计算每层、每个区域的空闲车位数，统计整个停车场的空闲车位数，计算停车场的实时车位占用率（图 2–45）。

层	区域	总车位数	空闲车位数			车位占用率
			关闭	开放	总计	
B1	A					
B1	B					
B1	C					
B1	D					
B1	E					
B1小计						
B2	A					
B2	B					
B2	C					
B2	D					
B2	E					
B2小计						
总计						

图 2–45　停车位使用实时统计表

（2）使用一种图表工具，选择合适的图表类型作图：①显示 B1 层的各个区域的车位占用率，②显示 B1 层、B2 层和整个停车场的车位占用率。

（3）从管理员的角度观察上面的图表数据，对是否开放或关闭某停车区域或停车层作出决策，实施分流措施。在小组内交流自己的决策及依据。

知识链接

数据采集的方法和工具

数据采集是数据处理工作的前提和基础。采集数据时，须运用适当的方法和工具。数据采集的常用方法和工具如表 2-3 所示。

表 2-3　数据采集的常用方法和工具

采集方法	人工获取					自动采集		
	调查	访谈	观察	实验	文献调研	物联感知	视频监控	网络平台
采集工具	问卷		表格		文献检索工具	传感器	摄像机	采集软件 网络爬虫 移动 App

1. 人工获取数据

人工获取数据指人直接从社会现象、自然现象或文献中获得数据。

（1）社会科学研究经常通过调查或访谈的方法获得一手数据。例如关于中学生移动学习现状的研究，可以从中学生的个体特质、移动学习的特点、家庭影响等角度展开探讨，设计相关问题，制作并发布问卷，从群体或个人获得一手数据。问卷是调查、访谈时常见的数据采集工具。相比传统纸质问卷，很多网络工具都可以帮助调查者更快捷、更精准地获得调查数据。

（2）在观察自然现象和进行科学实验时，需要设计各种表格，科学地记录通过观察或实验得到的各种数据。

（3）文献数据既包括正式出版、发行的纸质书刊、报表、年鉴，也包括政府机构、职能部门网站定期发布的公报、统计信息、研究报告等，还包括企业、机构网站上免费或有偿提供的数据库数据。文献数据可以通过文献检索工具获取，如利用搜索引擎工具在网络上搜索专业的数据库。

2. 自动采集数据

信息社会中，在信息技术的支持下，各种终端设备、网络数据库中记录存储着日益增长的海量数据。自动采集数据的方法通常有物联感知采集、视频监控采集、网络平台采集以及从已有数据库中采集等。

（1）物联感知采集是指，对于在物联网中使用电子标签或无线终端标识的智能化物体，通过传感器感知它们的数据变化，采集相关数据，并利用各种通信技术上传至网络信息中心存储。例如，通过佩戴相关设备，可以实时记录佩戴者的运动状态、呼吸量、血压、运动量、睡眠质量等生理状态数据，再利用无线或蓝牙技术，就可以将数据传送到网络信息中心或个人智能移动终端。

（2）视频监控采集是指借助不同监控点的摄像机采集监控区域的数据。例如，高清电子警察系统利用动态视频检测触发技术对车辆违规行为进行抓拍并完成车牌识别，清晰、完整地记录车辆违章过程，以及违章车辆的车型、车身颜色、车牌号码等数据。

（3）网络平台采集主要是指用户在访问网站或使用 App 时，网站服务器上安装的采集软件自动采集用户的各种行为数据。如一个学习平台可以采集学生浏览了哪些视频或课件，看了多长时间，重复观看了哪些课件，是否快进观看，以及观看课件的顺序等，这些行为都被完整地记录在系统日志文件中。通过日志搜索分析技术，可以筛选出有用的数据，用于判断学生的学习行为模式。

采集互联网数据的工具还有很多，如网络爬虫、移动 App 等，而且这样的工具还在不断地发展中。网络爬虫是一个自动下载网页的计算机程序或自动化脚本，是搜索引擎的重要组成部分。网络爬虫类产品如八爪鱼采集器、网络矿工采集器等，在数据采集领域有着广泛的应用，可以定期实时采集各大门户网站的数据。近年来，随着移动终端和通信技术的发展，移动 App 技术逐渐成为移动过程中数据采集的主导技术，采集方式更加灵活、多样。例如，学生可以通过无线网络，使用移动终端与云端学习平台进行互动。结合移动终端的定位技术，利用传感器、视频监控等设备，通过网络平台实时采集学习者的学习地点、学习时间、学习内容及学习状态等数据，可以让教师实时了解学生的学习情况，进而实现个性化智能辅导。

数据的保护

数据在采集、存储、管理与使用的过程中面临诸多安全风险。大数据时代，人们对数据的依赖性不断增强，数据安全与隐私保护问题更加突出。

1. 数据备份

数据在传输、存储、交换的过程中会面临导致丢失或损坏的各种风险因素，如自然灾害、信息攻击、设备故障、误操作等。为避免风险，通常需要进行数据备份。数据备份是周期性地将数据以某种方式制作一个或多个备份，并将其存放在专门设备上加以保护，以便在数据丢失或损坏时能够有效地进行数据恢复。

个人数据的备份主要通过文件的复制完成，由用户对重要的数据文件在不同的存储介质上归档保存。

企业数据是企业的重要资产，如果缺失数据备份措施，数据的安全性就得不到保障，可能导致数据丢失或损坏，对企业产生无法弥补的损失，甚至带来灾难性后果。企业要制定数据备份策略，明确数据备份内容、数据备份时间和数据备份方式等。企业的信息管理系统一般都包含数据备份的功能，以自动、全面、高效地在服务器上进行数据备份。

2. 数据的隐私保护

在大数据的背景下，人们在互联网上的一言一行都会被自动记录：在网上阅读电子图书，阅读习惯会被记录；在网上聊天，与好友的联络情况会被记录；在网上购物，购物喜好会被记录；发送电子邮件，联络方式会被记录；在网上搜索，搜索习惯会被记录……多项案例说明，即使看似无害的数据，被大量采集分析后，也会暴露个人隐私。

不少网络企业既是数据的生产者，又是数据的存储者、管理者和使用者，如果对用户数据的采集、存储、管理与使用等缺乏规范和监管，用户就无法确保自己隐私数据的安全。

每个人都应有权决定自己的数据如何被利用，决定自己的数据何时以何种形式披露，或者何时被销毁。

数据的隐私保护需要从立法、技术、管理等多方面给予保障。对个人来说，需要不断地提高和加强自身的隐私保护意识和防范能力。如在使用某一种系统和服务时，要考虑对方要求自己提供的数据是否与服务相关；在被要求提供身份证号码、电话号码等相关敏感数据时，要考虑对方是否正规机构；在一些社交媒体公布自己的生活照时，要注意是否涉及敏感数据。

为对数据隐私做好保护，一些技术便应运而生。数据隐私保护技术包括：数据采集时的隐私保护，如数据精度处理；数据共享、发布时的隐私保护，如数据的匿名处理、人工加扰等；数据分析时的隐私保护；数据生命周期的隐私保护；以及隐私数据的可信销毁。

数据的组织和存储

数据以文件或数据库的形式永久存储在外存储器中。按照数据的组织和编码方式，文件可以分为不同的类型，可以由文件的后缀名加以区分。数据库按照特定的数据结构来组织、存储和管理数据，它相当于建立在计算机存储设备上的仓库。数据库有很多种类型，从最简单的存储各种数据的表格，到能够进行海量数据存储的大型数据库系统，都有着十分广泛的应用。

结构化数据通常存储在关系型数据库或表格文件中。关系型数据库是现代信息系统中最流行的一种数据存储结构。关系型数据库中的关系也称表（table），一个关系型数据库由若干个二维表组成。非结构化数据（unstructured data）主要以多媒体格式文件存储，例如各种格式的视频文件、音频文件、图像文件、文本文件等。

随着大数据时代的到来，数据量急速增长。为了满足大数据的海量存储、快速查询、安全兼容的需求，一些新型的非关系型数据库应运而生，以应对大规模数据集合和多重数据种类带来的挑战，尤其是解决一些大数据应用难题。

表格数据的加工方法

数据分析中最常遇见的数据是表格数据，表格数据的加工方法主要包括数据的计算、排序、筛选和分类汇总。

1. 数据的计算

（1）数值数据

数值数据一般由阿拉伯数字、小数点和正负号构成，一些数据处理软件还提供分数、百分比、货币、科学记数法等表示形式，其写法如表 2–4 所示。

数值数据的计算包括算术运算和关系运算。算术运算包括加法（+）、减法（–）、乘法（*）、除法（/）、乘方（^）。关系运算也称为比较运算，包括等于（=）、大于（>）、小于（<）、

表 2-4　数值数据表示形式示例

正数	负数	小数	分数	百分比	货币	科学记数法
1233	-534	23.5	2/3	78%	¥12.00	1.21212E+11
123	-9012	0.909	10 5/7	78.00%	$566.00	2.31E-11
213.534	-90.32	212.64	2 1/11	0.90%	€ 34.00	3.23E+11

大于等于（>=）、小于等于（<=）、不等于（<>）。关系运算的结果为 TRUE 或者 FALSE，属于逻辑数据。

常见的数值数据函数主要有：用于数值统计的 SUM（求和）、COUNT（计数）、AVERAGE（求平均值）、MAX（求最大值）、MIN（求最小值）等；用于计算的 SQRT（求平方根）、MOD（求余数）、POWER（求乘幂）等；以及三角函数、数值舍入取整函数、随机数函数等。

例如，区域 A24:A132 中存放着一组实数，求它们总和的公式为 SUM(A24:A132)，求它们平均值的公式为 AVERAGE(A24:A132)，求它们的最大值和最小值之差的公式为 MAX(A24:A132)-MIN(A24:A132)。

（2）文本数据

文本数据一般是字母、汉字等字符，但也可以是完全由数字构成的文本数据，例如邮政编码、身份证号码、工号等。

文本数据一般只有连接（&）运算，即将两个操作数据连接在一起，构成新的文本数据。例如，如果 B2 单元格存储的数据是“沈萧”，那么公式 B2&“同学”，就是将 B2 单元格的数据与文本数据“同学”连接，得到“沈萧同学”。

文本数据函数用于对文本数据进行操作，主要有求字符串长的 LEN，求子串的 LEFT（从左边截取字符串）、RIGHT（从右边截取字符串）、MID（截取指定子串），查找字符串的 FIND（返回字符串的位置），删除空格的 TRIM 等。

例如，要取出身份证号码中表示性别的第 17 位字符，可使用 MID 函数，截取源字符串中从指定位置开始的指定长度的子串。

（3）日期 / 时间数据

日期 / 时间数据包括日期和时间两部分，输入方式为 YYYY-MM-DD HH:MM:SS。时间和日期的显示方式非常丰富，不同地区、不同应用场合都有所不同。日期 / 时间数据是通过数值编码的，日期以 1 天为单位编码，时间以 $\frac{1}{24\times60\times60}$ 天（1 秒）为单位编码。系统规定 1900/1/1 是第一天，编码为 1，其他的日期按照递增方式编码，所以日期 / 时间数据支持有效范围内的加减运算。例如，求两个日期相距的天数可以用减法完成，求一个日期后的第 n 天是哪一天可以用加法完成。

常见的日期 / 时间数据函数有 TODAY（返回当天日期）、NOW（返回当前日期和时间）、DATEDIF（返回两日期间相差的实足年数、月数和天数）、DATE（构造一个日期）、TIME（构造一个时间）等，还有从日期中提取年、月、日信息的函数——YEAR、MONTH、DAY，以及从时间中提取小时、分、秒信息的函数——HOUR、MINUTE、SECOND。

日期 / 时间数据在公式中不能直接书写，要用 DATE 函数构造，用法为：DATE(year, month, day)，返回的是日期数据。例如，求现在距离 2029 年国庆节还有多少天的公式为：DATE(2029,10,1) – TODAY()。

（4）逻辑数据

逻辑数据只有两个：TRUE（真）、FALSE（假）。逻辑数据没有运算，但是关系运算和逻辑函数都会产生逻辑数据。

最典型的逻辑数据函数是 IF 函数，它可以根据条件是否满足返回不同的结果。例如，IF(A1>60, " 合格 ", " 不合格 ")。

IF 语句支持多个条件判断，可以解决复合判断问题。例如，身高体重指数 BMI（BMI= 体重 ÷ 身高 ^{2}）可以从一个方面反映人的健康情况，如表 2–5 所示。

表 2–5　BMI 指数

BMI 指数范围	评价
BMI <18.5	体重轻
18.5<=BMI<24	健康
24<=BMI<28	超重
BMI>=28	肥胖

可以设计如表 2–6 所示的数据表，并计算出对每组数据的评价。

表 2–6　体检表

序号	体检号	身高（米）	体重（千克）	BMI	评价
1	10100510228	1.56	63.1	25.93	超重
2	10100720214	1.60	48.7	19.02	健康

例如，序号 1 的评价由嵌套的 IF 公式给出（假设序号 1 的“BMI”单元格为 E2）：

IF(E2<18.5, “体重轻” ,IF(E2<24," 健康 ",IF(E2<28," 超重 "," 肥胖 ")))

2. 数据的排序

排序工具可以对表格中的一列或多列数据按指定顺序重新显示。排序有助于快速地组织和查找所需的数据，也有助于更好地理解数据，是数据处理中不可缺少的常见操作。

排序有升序和降序两种基本方式。所谓升序，就是从小到大排列数据，降序则正好相反。数值数据按数值的大小排序；文本数据按 ASCII 码值的大小排序；逻辑数据的 FALSE 相当于 0，TRUE 相当于 1；汉字有两种排序方式，一是按拼音的字典顺序排序，二是按笔画的多少逐字排序。

3. 数据的筛选

当需要从表格中找出满足一定条件的几行或几列数据时，就需要用到数据筛选功能。数据筛选仅仅是将不符合条件的数据隐藏起来，只显示那些满足条件的数据。筛选条件通常是针对某一列进行设定。筛选可以累加，进行多列筛选时，后一次的筛选是在前一次的基础上完成的。

筛选条件的具体设定方法随着数据类型不同而不同，见图 2-46。数值数据的筛选可针对数据的值域范围、平均值、最大值、最小值等来设定；文本数据的筛选可针对所包含的字符或字符串实现模糊查找；日期 / 时间数据的筛选支持按年、月、日分级选择，筛选条件可根据日期的大小和范围来设定；逻辑数据只有两个值，只需选择相应的值进行筛选。

图 2-46　筛选条件的设定示例

4. 数据的分类汇总

表格数据可以按照不同的类别进行汇总统计，汇总统计包括求总和、计数、求平均值、取最大值、取最小值、求偏差、求方差等。

数据处理软件通常会提供两种汇总工具，一种是单方向分类汇总，另一种是交叉分类汇总。

（1）单方向分类汇总

如图 2-47 所示，可以对表格“某地 2013 年天气数据”进行分类汇总，分类字段为“天气类别”，汇总各类天气的天数。

日期	最高气温	最低气温	风向	风力	天气类别
2013/1/1	7	1	北风	3-4级转4-5级	晴
2013/1/2	3	1	西北风	4-5级	阴
2013/1/3	3	-2	北风	4-5级转3-4级	雪
2013/1/4	5	0	北风	3-4级	雪
2013/1/5	6	2	北风	3-4级	雪
2013/1/6	5	1	西北风	3-4级	雪
2013/1/7	6	1	西风	3-4级	雪
2013/1/8	6	2	东风	4-5级	阴
2013/1/9	6	0	西北风	3-4级	晴
2013/1/10	9	1	北风	3-4级	晴
2013/1/11	10	2	西风	3-4级	阴
2013/1/12	8	4	西北风	3-4级转4-5级	中雨
2013/1/13	10	3	西北风	3-4级	多云

图 2-47　某地 2013 年天气数据（部分）

首先将表格中的数据按分类字段排序，如图 2-48 所示，然后利用“分类汇总”工具对汇总方式、汇总项等进行设定，如图 2-49 所示。

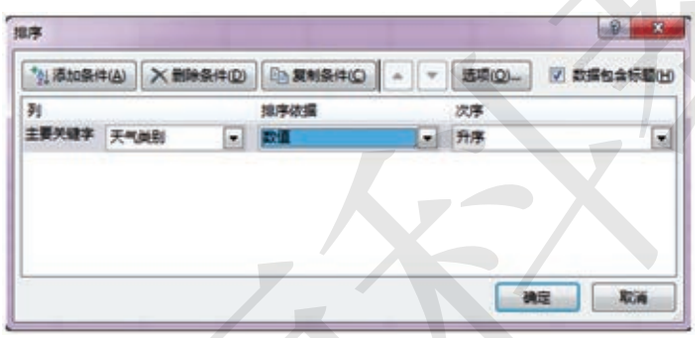


图 2-48　按分类字段排序

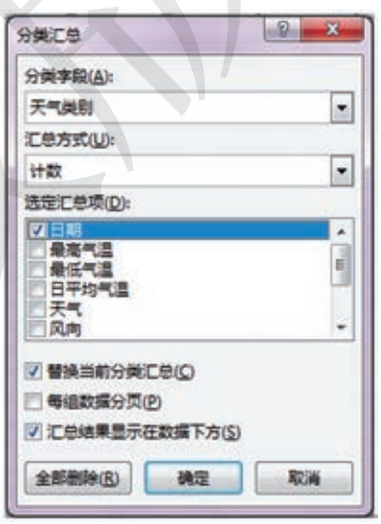


图 2-49　单方向分类汇总设置

设定完成后，每一个部分的结束行后面会出现汇总行，汇总出每个天气类别的计数，如图 2-50 所示。

1	日期	最高气温	最低气温	风向	风力	天气类别
126	124					晴 计数
194	67					多云 计数
254	59					阴 计数
361	106					雨 计数
371	9					雪 计数
372	365					总计数

图 2-50　单方向分类汇总结果

单方向分类汇总支持多级分类汇总，例如，可以对表格“某地 2013 年天气数据”先按“天气类别”再按“风向”进行多级汇总，汇总前需要将表格先按主关键字“天气类别”再按次关键字“风向”进行排序。

（2）交叉分类汇总

数据处理软件提供的交叉分类汇总工具通常是数据透视表。数据透视表是一种功能强大、操作简单的数据分析工具。进行交叉分类汇总时，选择表格中不同的行列组合，可以得到不同的统计数据。

表 2-7　学生档案数据表

学号	年级	班级	姓名	性别	出生日期	中考成绩

以表 2-7 所示的某高中的学生档案数据表为例，按图 2-51 的行列字段组合进行设置，得到的数据透视表是统计各年级各班的男生人数和女生人数，行汇总可以得到每个班的人数，列汇总可以得到全校的男生人数、女生人数和全校人数。若在行标签中删去“班级”字段，则得到的统计数据是各年级的男生人数和女生人数。

图 2-51　交叉分类汇总设置

拓展阅读

“手机导航 + 智能停车”服务

身在都市的有车族一般都有被寻找停车位、缴纳停车费困扰的经历。某地图 App 厂家，瞄准用户的这一痛点，与两家智能停车企业开展合作，为用户提供一体化的停车场电子支付服务。用户在手机地图 App 的导航下到达目的地时，该 App 会给用户推荐周边的智能停车场，并引导用户进行实时停车费查询和手机支付停车费等智能停车服务。

通过手机寻找停车场、缴纳停车费，有两个便利：一是可以通过导航到达停车场，这对不熟悉道路的车主很有用；二是可以直接在线完成支付，省去找零钱的麻烦。在手机地图 App 中，进入“发现周边服务”，可以在“车主服务”中找到“停车场”功能。“停车场”功能可以显示附近停车场的地图和列表，选择其中一个停车场便可以得到该停车场的地理位置。车主进入停车场后，可以直接在手机地图 App 里完成“找车位—进场—在线支付—快速出场”一系列动作。此外，车主还可以实时查询停车账单和预存停车费。

许多移动互联网用户已经习惯将手机地图作为生活服务的入口，充分使用“位置服务 + 生活服务”带来的便利。“手机导航 + 智能停车”将会为众多车主提供智能、便利的“行 + 停”无缝出行体验。

单元挑战　采集与分析气象数据

一、项目任务

气象数据是反映天气的一组数据，气象站采集的地面气象观测数据如温度、湿度、气压和风力、风向等，是气象数据的重要组成部分。随着技术的发展，人工气象站正逐渐被自动气象站所取代（图 2–52）。

查找资料，了解并比较人工气象站和自动气象站采集气象数据的方法和过程。重点关注自动气象站是如何自动采集数据的，借助了哪些技术手段或工具。

上网采集当地或某地近几年 7 月和 8 月的历史气温数据，并对气温数据进行统计和分析，了解近几年该地区 7 月和 8 月高温天气的走势。

图 2–52　自动气象站

二、项目指引

1. 以小组为单位，查找资料或访问气象站，了解常见的地面气象观测数据及其人工采集方法和自动采集方法。

2. 每个小组确定一个城市，上网查找该城市近几年 7 月和 8 月的历史气温数据，并选择一种电子表格软件设计、创建表格，输入数据。

3. 制作合适的图表，对比不同年份 7 月和 8 月气温的走势。

4. 使用函数计算或数据分析工具等方法，参考高温预警信号的等级分类定义（图 2–53），统计该城市近几年 7 月和 8 月各级高温天气的天数，进行对比分析，并选择适当的工具可视化数据。

5. 根据以上分析数据，了解该城市 7 月和 8 月的高温天气呈现怎样的趋势，思考背后的原因，并撰写报告。

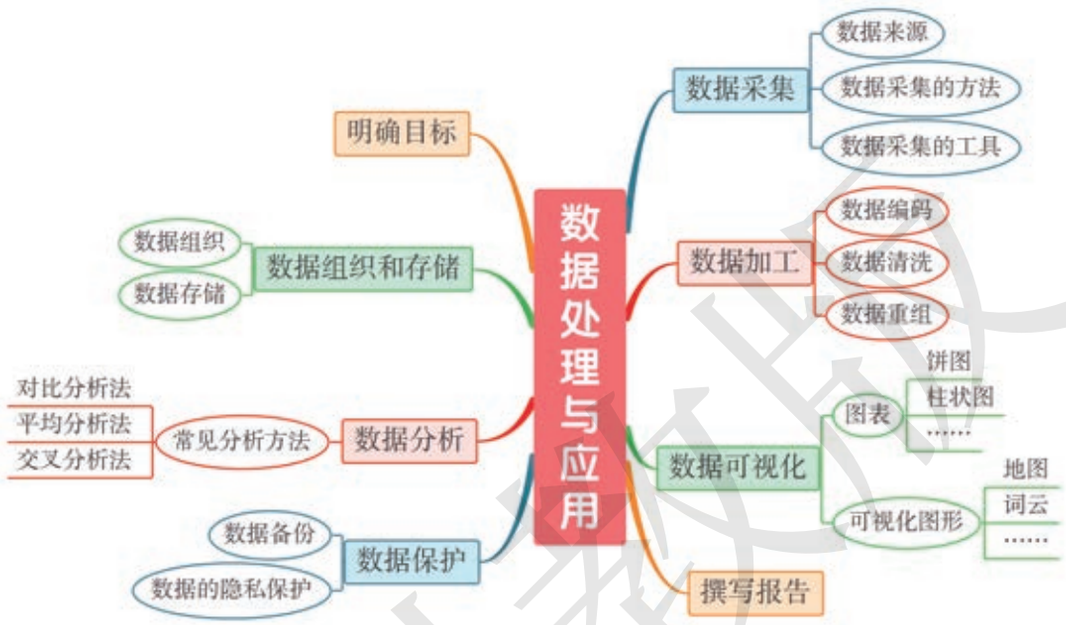
图 2–53　高温预警信号

三、交流评价与反思

各小组派出代表，用自己熟悉的信息表达工具（如演示文稿等）制作电子作品，通过网络或课堂，展示交流小组的报告，并对其他小组的报告进行评价。

单元小结

一、主要内容梳理



二、单元练习

1. 某学校拟举办校园歌手大赛，有 20 名选手报名参加。比赛时共有 6 位评委评分（0 ~ 10 分），每位选手的最终成绩为 6 位评委所评分数的平均分。

（1）设计选手成绩管理表，并使用电子表格软件创建表，然后，将一些模拟数据输入表中。

（2）利用电子表格软件计算各选手的最终成绩，并按成绩从大到小排序。

2. 为了解全班学生的身体健康和每周锻炼情况，请以小组为单位，采用多种方式和工具采集数据，并对数据进行分析。

（1）利用多种数据采集方式和工具，获取全班学生的身体健康数据和每周锻炼情况的数据。

（2）对采集到的数据进行加工，分析班级学生的身体健康状况与每周锻炼情况。

（3）利用可视化的方式展示分析结果，撰写分析报告。

3. 教材配套资源中的“订单表”记录了某网上书店过去一年的订单数据。尝试利用电子表格软件对该订单表进行分析。

（1）筛选出过去一年中销售量最高的 10 本图书。

（2）分析一年中不同月份图书销售量的变化。

三、单元评价

评价内容	达成情况
了解数据处理的概念和过程（A、T）	
能够认识数据处理的作用和应用价值（A、R）	
了解数据采集的方法和工具（A、T）	
能够根据实际情况，选择合适的数据采集方法（A、T）	
了解数据分析的概念和基本方法（A、T）	
知道数据的组织方式（A、T）	
知道数据的存储方式（A、T）	
能够根据给定的任务，使用并设计二维表进行数据存储（A、T）	
能够掌握通过公式、函数进行数据加工的方法（T）	
能够根据给定的任务需求，选用恰当的软件工具或平台处理数据，完成分析报告（A、T、I）	
能够使用合适的数据可视化的方法，表示数据的含义（A、T、I）	
能够根据完成的数据分析报告，读懂数据背后隐藏的信息（A、T、I、R）	
能够理解对数据进行保护的意义（A、R）	

说明：A—信息意识，T—计算思维，I—数字化学习与创新，R—信息社会责任