第2章
数　据

2.1　数据的定义

2.1.1　数据的定义

数据科学的核心就是研究数据，获取知识。那么究竟什么是数据呢？一提到数据，人们首先想到的是数字。但数据并不局限于数字，文本、音频、图像、视频都可以是数据。在本书里，对数据给出如下定义：

数据是指以定性或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。

简单说来，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的地方、事件、对象或概念的描述和反映。

数据可以是连续的值，例如声音，它们称为模拟数据；也可以是不连续（离散）的值，例如成绩，它们称为数字数据。

2.1.2　其他相关概念

数据、信息与知识，这三个概念，在后面的学习中会多次出现。这三个概念往往会存在一些交叠，容易混淆，在这里先做一下区分。

这三个概念之间最主要的区别是所考虑的抽象层次不同。数据是最低层次的抽象，信息次之，知识则是最高层次的抽象。数据是原始的、零散的，数据本身是没有意义的，数据经过了处理依然是数据，只有经过解释和理解才有意义。从数据抽象到信息的过程，就是对数据解读和释义的过程。

人们只有对数据进行解释和理解之后，才可以从数据中提取出有用的信息。只有对信息进行整合和呈现，才能够获得知识。例如，世界第一高峰珠穆朗玛峰的高度为 8 844.43 m，这可以认为是“数据”；而一本关于珠穆朗玛峰地质特性的书籍，则包含了“信息”；而一份包含了攀上珠穆朗玛峰最佳路径信息的报告，就是“知识”了。所以，数据是信息的载体，是形成知识的源泉，是智慧、决策以及价值创造的基石。

信息所涉及的范畴是非常广泛的，从日常生活到技术细节都可以涵盖其中。通常而言，信息这个概念，一般是与约束、形式、指示、含义、样式、表达等紧密关联的。数据是一些符号的组合，而当这些符号被用来指示某个事物或者某件事情时，则成了信息。

而知识则是人们对某件物品或某种现象的理论性或实践性的理解，知识一般是形式化的

或系统化的。知识的获取，一般是通过传授或亲身经历。

比如一碗热汤摆在人们的面前，要认识到它很烫的这个特性，或说获取到它很烫的这个知识，既可以经由长辈告诉我们，也可以通过把手伸进去自我感受得到。

数据科学所研究的正是从“数据”整合成“信息”进而组织成“知识”的整个过程，其中包含了对数据进行采集、分类、录入、储存、处理、统计、分析、整合、呈现等一系列活动。

在数据科学中还有一些概念是人们需要用到的，包括：

- 元数据——即数据的数据；
- 元信息——即信息的信息；
- 数据文件——指信息与元数据的集成，用以描述数据的各个方面。

为了加深对这些概念的理解，这里举两个例子来予以说明。

以岩石样本为例，数据就是所收集的岩石样本的重量、形状、尺寸等；信息则是关于这些岩石样本的成分分析图像；知识指的则是由分析这些岩石样本所获取到的地质活动的相关证据；元数据则指的是这些岩石样本收集的时间、收集的地点等；数据文件则是一本出版了的实验室报告。

以天气为例，数据就是当天的风速、风向、温度等；信息则是由这些数据制成的气象云图；知识则是从中获取到的如高气压系统的分布情况、天气的稳定性等；元数据则指这些天气数据获取时所用的雷达类型、传感器类型等。

2.2 数据简史

人类历史上最早的有记录的数据，可以追溯到穴居的原始人时期。当时的人类，会在作为居处的洞穴墙壁上，以石器或骨器刻画来记录数据。这些被记录的数据，或是简单地记录日期的刻痕，或是形象化地记载一些日常发生事件的壁画（图 2－1）。

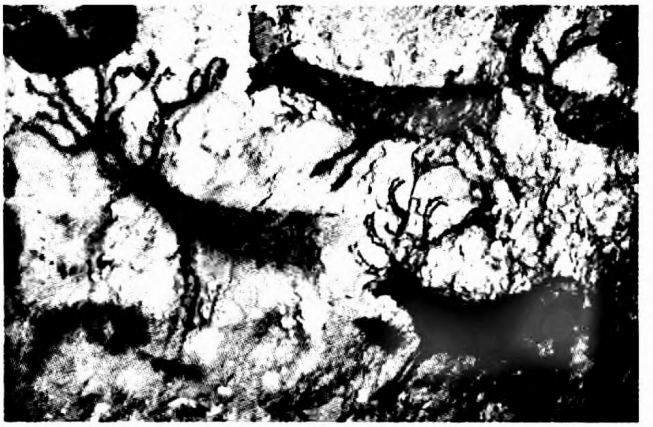

图 2－1　穴居壁画——最古老的数据记录形式

后来，人们创造了结绳记事的方式来记录数据。《周易・系辞下》中有云：“上古结绳而治，后世圣人易之以书契。”即在一条绳子上打结，用以记事。上古时期的中国及秘鲁印第安人皆有此习惯。到了近代，一些没有文字的民族，仍然将结绳记事来作为数据记录方式传递信息。古人采取的结绳方法，据古书记载为：“事大，大结其绳；事小，小结其绳，结之多少，

随物众寡。”

图 2-2 所示的是古代印加人采用的一种结绳记事的方法，用来计数或记录历史。事大，大结其绳；事小，小结其绳。不过，这种记事的方法已经失传，目前还没有人能够了解其全部含义。

随着数字和文字的出现，古人开始以更加明确的形式来记录数据。古埃及人发明了莎草纸，用来进行记录。埃及博物馆中陈列的各种莎草纸文书、图画表明，莎草纸是人类历史上最早、最便利的书写材料之一，是记录古埃及历史的主要载体（图 2-3）。

图 2-2 结绳记事

图 2-3 古埃及人用莎草记录数据

我们的祖先在汉代发明了造纸术（图 2-4），这种数据记录方式一直延续到现在。这里一定要强调，本书中所提到的数据，不仅包括数字形式的数据，以文本、图像、语音等方式所记录的数据都是数据科学所研究的内容。造纸术的发明和改进，让文本形式的数据记录方式开始盛行起来。

最早的录音机，也叫留声机，诞生于 1877 年，是发明大王——爱迪生所制造的。爱迪生发现了电话传话器里的模板随着说话声而震动的现象，于是他拿短针做了实验，从中得到了很大的启发。说话的快慢高低能使短针产生相应的颤动。那么，反过来，这种颤动也一定能发出原先的说话声音。于是，他开始研究声音重发的问题。

在 1877 年 8 月 15 日，爱迪生让助手克瑞西按图样制造出一台由大圆筒、曲柄、受话机和模板组成的怪机器。爱迪生指着这台怪机器对助手说：“这是一台会说话的机器。”他取出一张锡箔，卷在刻有螺旋槽纹的金属圆筒上，让针的一头轻擦着锡箔转动，另一头和受话机连接。爱迪生摇动曲柄，对着受话机唱起了：“玛丽有只小羊羔，雪球儿似一身毛……”唱完后，把针又放回原处，再轻悠悠地摇动曲柄。接着，机器不紧不慢、一圈又一圈地转动着，唱起了：“玛丽有只小羊羔……”与刚才爱迪生唱的一模一样。在他身旁的助手们，见到一架会说话的机器，都惊讶得说不出话来。

“会说话的机器”诞生的消息，轰动了全世界。1877 年 12 月，爱迪生公开演示了留声机的使用方法（图 2-5），外界舆论马上把他誉为“科学界的拿破仑”，留声机是 19 世纪最让人振奋的三大发明之一。当时即将开幕的巴黎世界博览会立即把它作为时新展品展出。就连当时美国总统海斯也在留声机旁转了 2 个多小时。

图 2－4　造纸流程

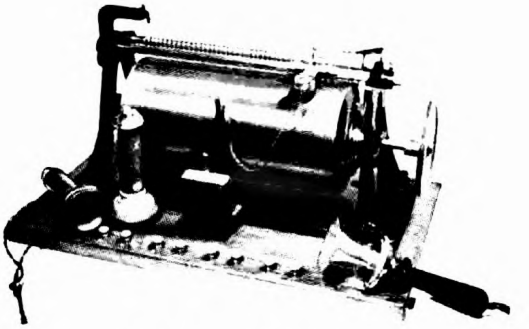

图 2－5　最早的留声机

10 年后，爱迪生又把留声机上的大圆筒和小曲柄改成类似时钟发条的装置，由发动机带动一个薄薄的蜡制大圆盘转动，从此以后留声机才广为普及。留声机的发明，让音频数据的记录成为可能。

在公元前 400 年前，墨子所著《墨经》中已有针孔成像的记载；在 13 世纪，欧洲也出现了利用针孔成像原理制成的映像暗箱，人们可以走进暗箱观赏映像或描绘景物。

但直到 1822 年，法国的涅普斯才在感光材料上拍摄出了世界上的第一张照片，不过当时成像不太清晰，而且需要 8 个小时的曝光时间。1826 年，他又在涂有感光性沥青的锡基底版上，通过暗箱拍摄了一张照片。

1839 年，法国的达盖尔制成了第一台实用的银版照相机，它由两个木箱组成，把一个木箱插入另一个木箱中进行调焦，用镜头盖作为快门，来控制长达三十分钟的曝光时间，从而拍摄出清晰的图像，最终实现了静止图像数据的记录（图 2－6）。

1874 年，法国的朱尔·让桑发明了一种摄影机。他将感光胶片卷绕在带齿的供片盘上，在一个钟摆机构的控制下，供片盘在圆形供片盒内做间歇供片运动，同时钟摆机构带动快门旋转，每当胶片停下时，快门开启曝光。让桑将这种相机与一架望远镜相接，能以每秒一张的速度拍下行星运动的一组照片。让桑将其命名为摄影枪，这就是现代电影摄影机的始祖。摄影机（图 2－7）的发明，使运动图像数据的记录成为可能。

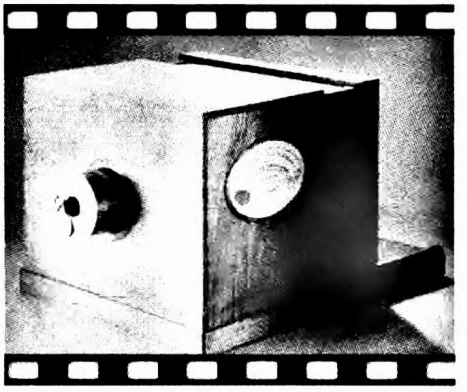

图 2－6　最古老的照相机

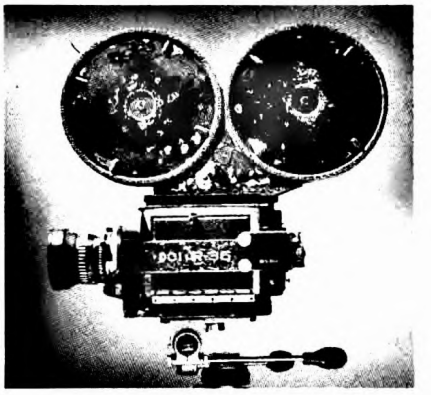

图 2－7　早期的摄影机

1946 年 2 月 14 日，由美国军方定制的世界第一台电子计算机“电子数字积分计算机”（Electronic Numerical and Calculator，ENIAC）在美国宾夕法尼亚大学问世，这表明电子计算机时代的到来。从此，人类与数据的关系进入了第二个时代，带来了一场数据存储方式的革命（图 2－8）。

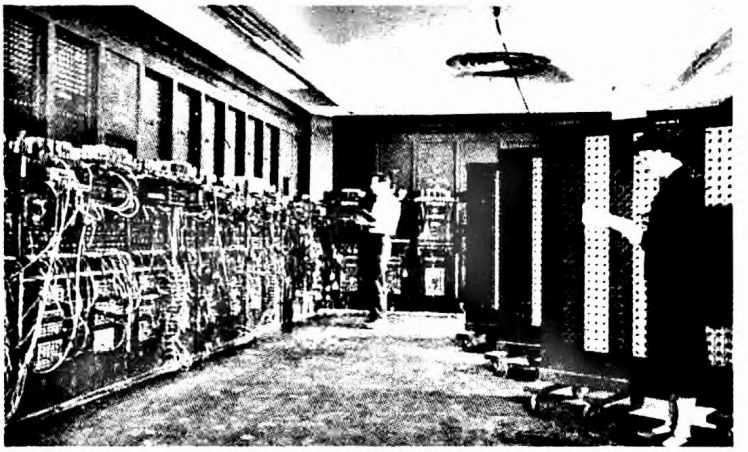

图 2－8　世界上第一台电子计算机

计算机的飞速发展，使数据的存储和处理模式带来了巨大的变革。以往的数据需要存储在纸张、胶片、磁带等介质上，一方面，数据的存储无法进行压缩，另一方面，介质的存储需要占用大量的空间。而计算机的发明，从本质上改变了这一点。数据可以通过多种算法进行压缩。而且随着半导体工业的发展，存储能力不断增强，数据所需要的存储实体空间也在不断缩小。如今，一块小小的 U 盘就可以存储 GB 量级的数据，节约了大量的数据存储空间。

随着计算机技术的发展，数据的处理能力也在不断提升。在计算机发明以前，数据都是通过人工的方式来进行处理的。而有了计算机的帮助，通过各种各样的计算方式和统计软件，人们可以快速地处理数据。根据最新的统计，目前世界上最快的计算机——中国制造的天河 2 号（图 2－9），处理速率已经达到了每秒钟进行 22.86 千万亿次浮点操作的水平。

图 2－9　当前世界上最快的计算机——天河 2 号

互联网的出现，是人类与数据之间的关系进入到第三个时代的标志，带来了一场数据产

图 2－10　网络与数据

生和传播的革命（图 2－10）。最早的网络，是由美国国防部高级研究计划局（ARPA）建立的。现代计算机网络中的很多概念和方法，如分组交换技术都来自于 ARPANET。ARPANET 不仅进行了租用线互联的分组交换技术研究，而且做了无线、卫星网的分组交换技术研究，其结果就是加速了 TCP/IP 的问世。

1977—1979 年，Arpanet 推出了 TCP/IP 体系结构和协议。1980 年前后，Arpanet 上的所有计算机开始了 TCP/IP 协议的转换工作，并以 Arpanet 为主干网建立了初期的 Internet。到 1983 年时，Arpanet 的全部计算机完成了向 TCP/IP 的转换，并在 UNIX（BSD 4.1）上实现了 TCP/IP。到 1984 年时，美国国家科学基金会 NSF 规划建立了 13 个国家超级计算中心及国家教育科技网，随之替代了 Arpanet 的骨干地位。1988 年，Internet 开始对外开放。到了 1991 年 6 月，在接入 Internet 的计算机中，商业用户首次超过了学术界用户，这是 Internet 发展史上的一个里程碑，从此 Internet 的成长速度一发不可收拾。

互联网的精神就在于“开放、分享、平等、合作”。网络的出现，让人与人之间的距离变得越来越短，地球村的概念也随之产生。通过网络，人们可以越洋对话，可以浏览海量的数据，可以实时地关注国际上最新的事件。网络让数据的产生和共享进入了一个崭新的时代。

网络时代的来临，造就了数据的大爆炸。据统计，2012 年年底，有超过 6 000 万用户，通过社交网站 Facebook 发布了超过 300 亿条的新内容；游戏商 Zynga 每天要处理超过 1 PB 的玩家数据；每天通过视频网站 Youtube 被浏览的视频量大约为 20 亿次；每个月通过微博 Twitter 所进行的搜索量会达到 320 亿次。

通过传感器网络搜集的数据又是另一大来源。所谓传感器网络，就是由大量部署在作用区域内的、具有无线通信与计算能力的微小传感器节点，通过自组织的方式所构成的，能根据环境自主完成指定任务的分布式智能化网络系统。

传感器网络综合了多种先进技术，如传感器技术、嵌入式计算技术、现代网络及无线通信技术、分布式信息处理技术等。它能够通过各类集成化的微型传感器协作，来实时监测、感知和采集各种环境或监测对象的信息，并可通过嵌入式系统对信息进行处理，并且通过无线通信网络将所感知到的信息传送到用户终端。

利用传感器网络，通过感知识别技术，让物品“开口说话、发布信息”，融合物理世界和信息世界，便可以建立物联网（图 2－11）。物联网的“触手”是位于感知识别层的大量信息生成设备，包括 RFID、传感器网络、定位系统等。传感器网络所感知的数据正是物联网海量信息的重要来源之一。

互联网和物联网，正是数据收集的两大重要渠道，推动了大数据时代的来临。

大数据时代的来临，标志着人类和数据的关系进入了第四个时代，带来的是一场数据运用的革命。

一方面，大量的数据被产生和积存下来，只有想办法努力地汲取其中的养分，才能让这些数据产生价值，开出智慧的花朵（图 2－12）。

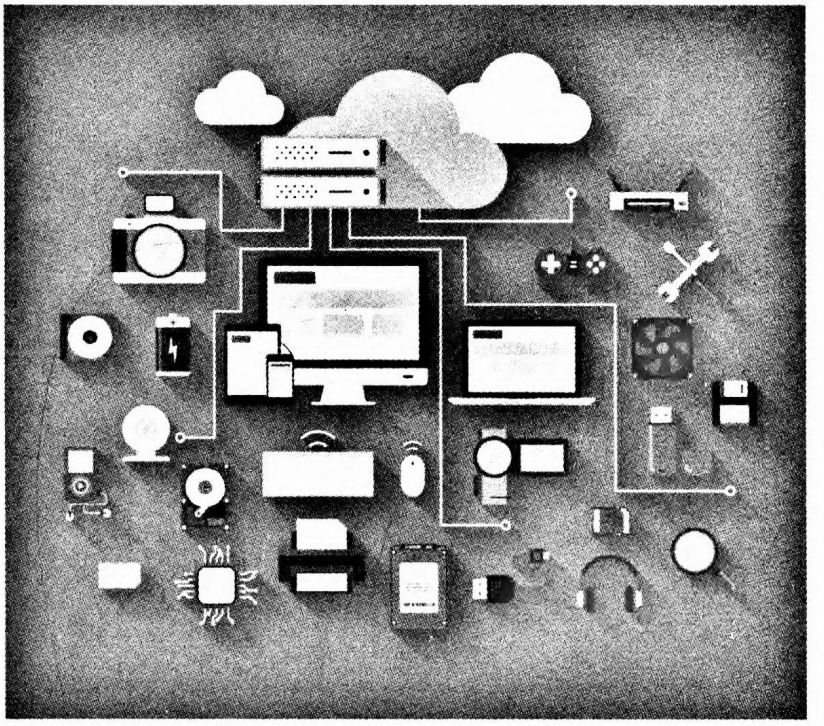

图 2－11　物联网

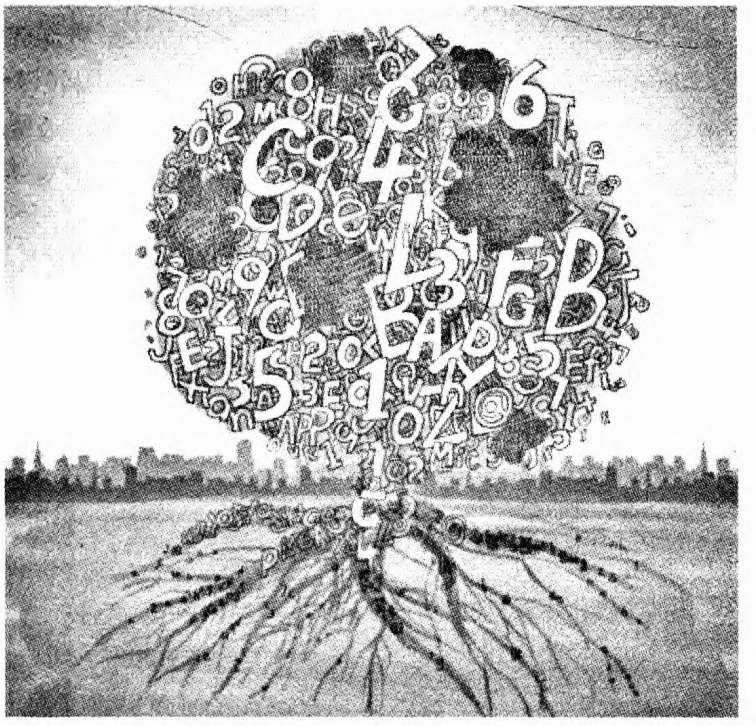

图 2－12　用数据产生智慧之花

另一方面，大量的数据也提供了更好地了解这个世界运行方式的渠道，让人们通过对数据的分析可以一窥未来（图 2－13）。

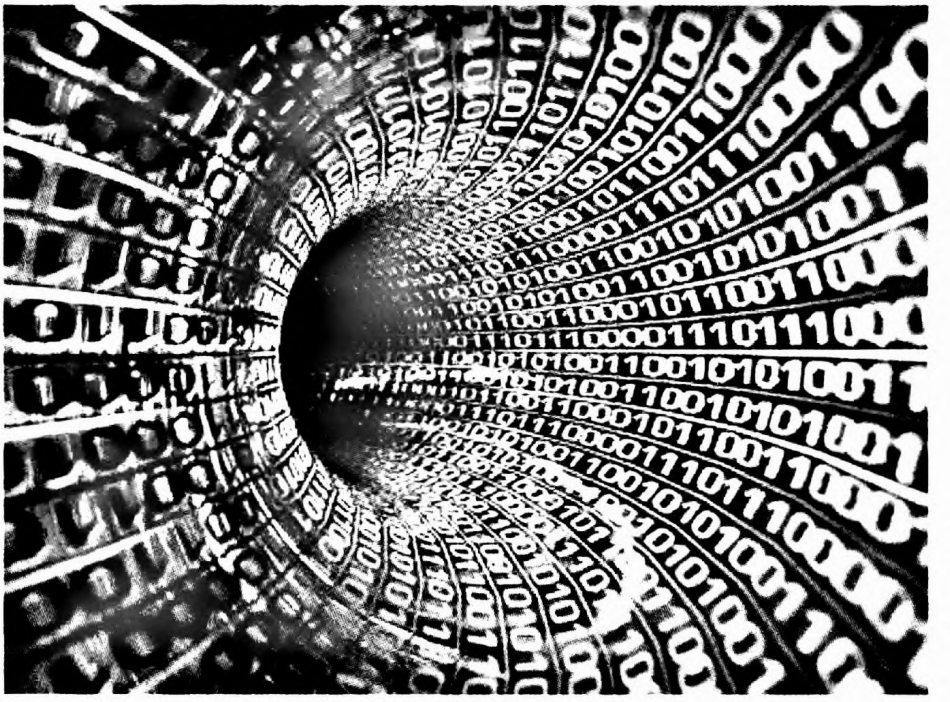

图 2－13　用数据来一窥未来