普通高等教育“十二五”规划教材

INTRODUCTION TO DATA SCIENCE

数据科学导论

ODUCTION TO DATA SCIENCE

杨 旭 汤海京 丁刚毅 主编

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

普通高等教育“十二五”规划教材

数据科学导论

杨　旭　汤海京　丁刚毅　主编

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

内容简介

数据科学，作为一门正在蓬勃发展的新学科，关注的是如何在大数据时代背景下运用各门与数据相关的技术和理论来服务社会。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，从各个方面展示数据科学的运用方式，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，并且在其中穿插了数据科学研究方式下新的思维模式的讲解。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，展望数据科学的未来发展。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。

图书在版编目（CIP）数据

数据科学导论 / 杨旭，汤海京，丁刚毅主编 . —北京：北京理工大学出版社，2014. 3
ISBN 978 - 7 - 5640 - 6384 - 9

Ⅰ. ①数…　Ⅱ. ①杨… ②汤… ③丁…　Ⅲ. ①数据管理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2014）第 011575 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）
　　82562903（教材售后服务热线）
　　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 / 保定市中画美凯印刷有限公司
开　　本 / 787 毫米 × 1092 毫米　1/16
印　　张 / 10
字　　数 / 210 千字
版　　次 / 2014 年 3 月第 1 版　2014 年 3 月第 1 次印刷
定　　价 / 29. 00 元

责任编辑 / 刘　娟
文案编辑 / 王晓莉
责任校对 / 周瑞红
责任印制 / 李志强

前 言

我们已经身处于一个数据爆炸的时代，在新的时代背景下，需要运用新的科学研究方式去应对新的挑战。数据科学，作为一门正在蓬勃发展的新学科，所关注的正是如何在大数据时代背景下，运用各门与数据相关的技术和理论，服务社会，让我们可以更好地利用身边的数据，使生活变得更加美好。

本教材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。本书将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。本教材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。

本教材由北京理工大学软件学院“数据科学与技术”课题组的杨旭老师、汤海京老师，以及北京理工大学软件学院院长丁刚毅老师担任主编。其中第 2、3、4 章由杨旭老师编写；第 1、5、6、7 章由汤海京老师编写；第 8 章由丁刚毅老师编写，并由丁刚毅老师负责全书的统稿。对书中存在的错误及不妥之处，恳请各位读者、同行批评指正。

编 者

图目录

目 录

第 1 章

引　论

1.1 序　言

“数据科学”从出现到现在已经有三十多年的历史了，其中涉及了很多方面的内容，涵盖数学、统计学、数据工程、模式识别、机器学习、高性能计算、可视化、数据仓库以及数据建模等多个领域的技术和理论。数据科学的最终目的就是从数据中挖掘出有用的信息，让数据增值。

虽然已有三十年历史了，但数据科学仍然是一门新兴的学科，尽管之前运用较多的是在计算智能或者是商业分析方面，但已经慢慢地深入到了人类社会的各个方面。之所以要开设数据科学这门学科，就是为了培养这门学科的专业人员，使他们运用所有可以得到的数据，寻找其背后的故事，从而找到办法让这些数据所蕴含的意义可以轻易地被人们所理解，即便他们不具备数据科学的相关知识。

本章将会首先从数据开始谈起，向读者讲述数据的发展概况和现状，从而感受学习数据科学的重要性。

1.2 数　据

1.2.1 数据的概念

数据科学这门学科研究的核心内容就是数据，那究竟什么是数据呢？一提到数据，我们首先想到的会是数字。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，我们对数据给出如下的定义：

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，是通过观察或者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，称为模拟数据；也可以是不连续（离散）的值，例如成绩，称为数字数据。

1.2.2 数据的发展史

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，以石器或者骨器刻画来记录数据。这些被记录的数据，或者是简单的记录日期的刻痕，或者是形象化地记载一些日常发生事件的壁画。

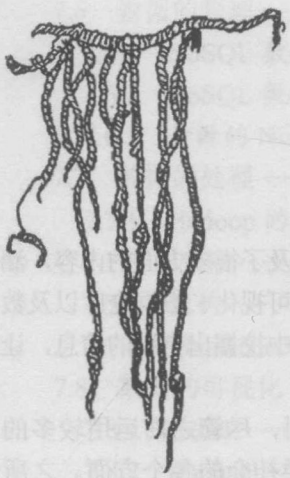

图 1–1　结绳记事

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，仍然采用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 1–1 所示的是古代印加人采用的一种结绳记事的方法，用来计数或者记录历史。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。古埃及人创造了莎草纸，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 1–2）。

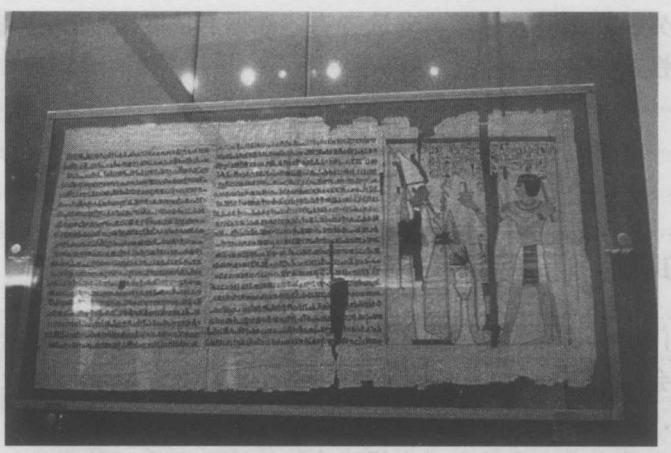

图 1–2　古埃及人用莎草记录数据

我们的祖先在汉代发明了造纸术（图 1–3），这种数据记录的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，不光是涉及数字形式的数据，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，于是他拿短针做了试验，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，爱迪生公开表演了留声机（图1–4），外界舆论马上把他誉为"科学界之拿破仑"，留声机是19世纪最让人振奋的三大发明之一。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1–3 造纸流程

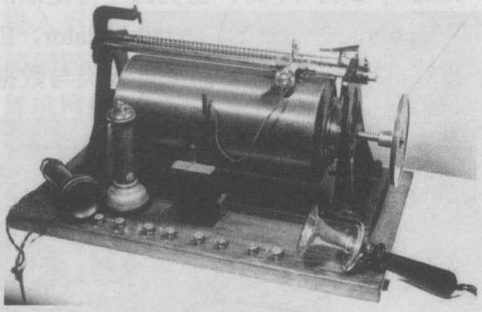

图1–4 最早的留声机

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制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。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，它是由两个木箱组成，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 1–5）。

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+Y6R 的始祖。摄影机（图 1–6）的发明，使得运动图像数据的记录成为可能。

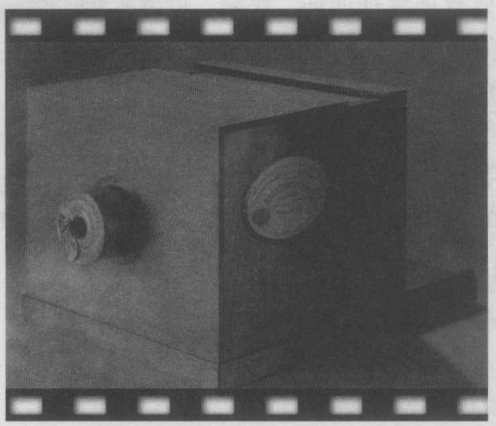

图 1–5　最古老的照相机

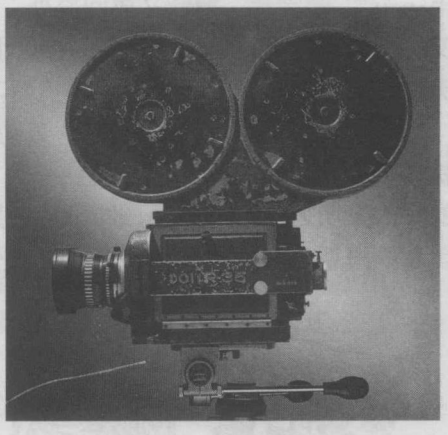

图 1–6　早期的摄影机

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（图 1–7）。

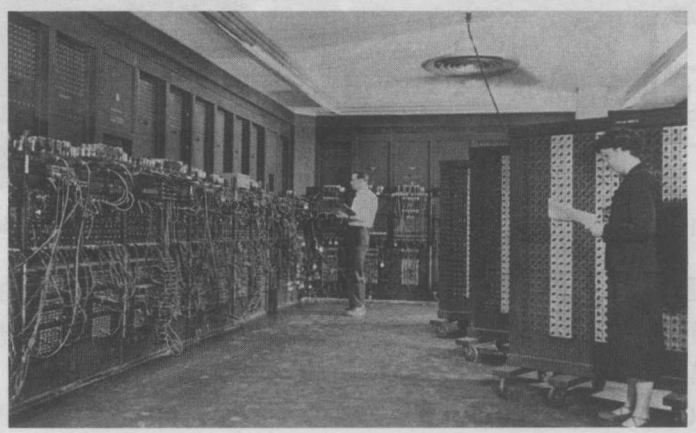

图 1–7　世界上第一台电子计算机

计算机的飞速发展，为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，一块小小的优盘就可以存储 GB 量级的数据，为我们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。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，我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 1–8），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。

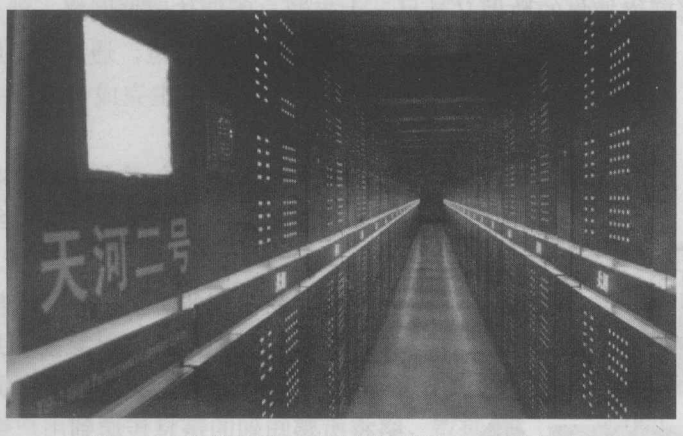

图 1–8 当前世界上最快的计算机——天河 2 号

互联网的出现，使人类与数据之间的关系进入到第三个时代的标志（图 1–9）。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。

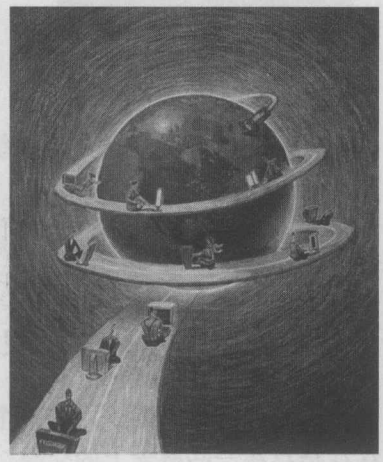

图 1–9 网络与数据

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，在联通 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，我们可以越洋对话，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 容量的玩家数据；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。让我们感受一下，这是多么庞大的数据量。

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。

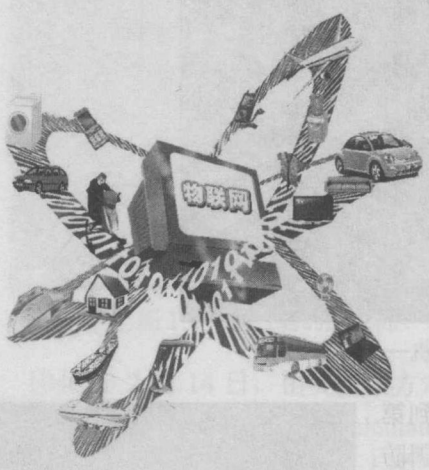

图 1–10　物联网

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 1–10）。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，正是我们数据收集来源的两大重要渠道，推动了大数据时代的来临。

1.2.3　数据、信息与知识

数据、信息与知识，这三个概念，在后面的学习中会多次出现。在使用这三者时，往往会存在一些概念上的交叠，容易混淆，在这里先做一下区分。

这三者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。

我们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。对信息进行整合和呈现，则能够获得知识。例如，世界第一高峰珠穆朗玛峰的高度 8 844.43 m，可