/军队“2110工程”三期建设教材/

作战数据管理

包磊 黄亮 罗兵 杨乐 李玉江 编著

国防工业出版社
National Defense Industry Press

军队“2110工程”三期建设教材

作战数据管理

包磊 黄亮 罗兵 杨乐 李玉江 编著

国防工业出版社
·北京·

内容简介

作战数据是现代指挥信息系统的“血液”，它将计算机硬件和作战软件有机地连接在一起，而数据资源建设是目前制约作战系统效能发挥的主要瓶颈。本书立足于作战数据管理的实际需求，对作战数据管理导论、数据建模与设计、数据加工、数据保存、数据中心建设与管理等五大部分作战数据管理的相关内容进行了全面、系统的梳理，以期让读者建立起数据的工程化管理概念，对数据管理中涉及的问题有一个较全面的理解和认识，帮助读者解决实际问题。

本书适合于相关领域的科研工作者和工程技术人员阅读，也可作为高等院校相关专业的教学用书和学习参考书。

图书在版编目(CIP)数据

作战数据管理/包磊等编著. —北京：国防工业出版社，2015.6

ISBN 978-7-118-10114-0

Ⅰ.①作… Ⅱ.①包… Ⅲ.①作战-数据管理 Ⅳ.①E83

中国版本图书馆 CIP 数据核字(2015)第 109606 号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号 邮政编码 100048)

北京奥鑫印刷厂印刷

新华书店经售

*

开本 787×1092 1/16 印张 17½ 字数 399 千字

2015 年 6 月第 1 版第 1 次印刷 印数 1—2500 册 定价 48.00 元

(本书如有印装错误，我社负责调换)

国防书店：(010)88540777 发行邮购：(010)88540776

发行传真：(010)88540755 发行业务：(010)88540717

前　言

硬件、软件和数据是作战系统的主要组成元素，其中数据是系统的“血液”，它将计算机硬件和作战软件有机地连接在一起。近年来，我军指挥信息系统的硬件和软件建设获得快速发展，但数据资源建设是目前制约作战系统效能发挥的主要瓶颈。面对系统中日益增多的海量数据资源，作战人员提出了越来越多的数据需求，尽管数据总量在不断增加，但在解决实际问题时却常缺少有效的数据支撑。由于涉及多个领域、多个层级，因此作战数据建设是一项极为庞大的系统工程。由于作战数据牵涉机密，其需求论证、采集方法、处理程序及利用模式等众多环节都无法借鉴国外。另外，由于作战数据在使用上的透明性，数据建设的成效目前也无明确的评判标准，易被人忽视。且由于使用者、生产者的能力水平问题，数据管理盲目混乱、多头管理、数据割据等问题长期未能得到有效解决。当前，各军事强国借助大数据争夺信息制高点的战略正在全面升级。面对挑战，我们必须反思如何善用数据说话，在破解作战数据建设的重重困境中，努力探寻实践途径。

在发达国家军事组织纷纷推行大数据战略的背景下，我们应清醒地意识到“数据制胜”的信息化战争发展趋势，更应立足于实际，正视数据浪费与数据缺乏的矛盾已严重制约基于信息系统的体系作战能力，探寻破解困境的途径。在大数据引领信息化战争发展方向的今天，我们不可能永远将作战数据定位于保障层次。考虑到我军实际，应努力探索以数据为中心的新技术、新装备和新战法，加大作战数据专业人才的培养力度。

数据科学、数据工程与数据管理是目前的新兴学科，本书的目的在于抛砖引玉，让读者初步建立数据工程化管理概念，对数据管理中涉及的问题有一个较全面的理解和认识，进而帮助读者解决实际问题。

由于相关领域内技术发展日新月异，作者能力所限，书中难免有错漏之处，恳请读者批评指正。

作　者

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第 1 章　数据管理概述

我们身处在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。对于数据的管理构成数据科学的主要内容，关注的是在数据时代的背景下，运用各门与数据相关的技术和理论服务社会，让我们可以更好地对数据资源进行开发和利用，推动社会发展和进步。本书主要介绍数据时代的背景与趋势，介绍数据的基本概念，阐述军事数据工程产生的背景、内涵及研究对象，介绍作战数据管理的现状与发展。

1.1　大数据时代的到来

麦肯锡于 2011 年 5 月发布报告《大数据：创新、竞争和生产力的下一个前沿领域》，将大数据概念从技术圈引入企业界。美国政府不久推出了《大数据研究发展计划》，将大数据上升至国家战略层面，形成国家意志。2012 年 11 月 17 日在北京召开的“数据科学与信息产业大会”上，宣告数据科学将在大数据时代焕发新生，标志学术界对大数据的重视达到了一个前所未有的新高度。正如哈佛大学量化社会科学学院院长 Gary King 所说：“这是一种革命，确实正在进行这场革命，庞大的新数据来源所带来的量化转变将在学术界、企业界和政界中迅速蔓延开来，没有哪个领域不会受到影响。”毫无疑问，上述的种种事件无不向世界传递一个信息：大数据时代已经到来！

麦肯锡报告中指出，全球数据正在呈爆炸式增长，数据已经渗透到每一个行业和业务职能领域，并成为重要的生产因素。大数据的使用将成为企业成长和竞争的关键，人们对大数据的运用将支撑新一波的生产力增长和消费者收益浪潮。麦肯锡深入研究了美国医疗卫生、欧洲公共管理部门、美国零售业、全球制造业和个人地理信息五大领域，用具体量化的方式分析研究大数据所蕴含的巨大价值。大数据的合理有效利用，为美国医疗卫生行业每年创造价值逾 3000 亿美元，为欧洲公共管理部门每年创造价值 2500 亿欧元(约 3500 亿美元)，为全球个人位置服务的服务商和最终用户分别创造至少 1000 亿美元的收入和 7000 亿美元的价值，帮助美国零售业获得 60%的净利润增长，帮助制造业在产品开发、组装方面降低成本 50%。

大数据，事关国计民生、产业兴衰、公司存亡，不可不察。信息科技经过 60 余年的发展，数据(信息)已经渗透到国家治理、国民经济运行的方方面面。经济活动中很大一部分都与数据的创造、传输和使用有关。2012 年 3 月，奥巴马公布了美国《大数据研究和发展计划》，标志着大数据已经成为国家战略，上升为国家意志。国家竞争力将部分体现为一国拥有数据的规模、活性，以及解释、运用数据的能力。国家数字主权体现为对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。

此为试读，需要完整PDF请访问：www.ertongbook.com

没有数据安全，也就没有国家安全。华为、中兴开拓美国市场受挫，就是非常明显和清晰的信号。美国政府对自家数据安全的重视程度，已经到了不能让任何外国信息基础设施产品供应商染指的地步。华为此前一直希望通过竞标和并购等方式进入北美市场，多年来未能如愿。2008 年，华为与贝恩资本联合竞购 3COM 公司，却因美国政府阻挠未能成功；2011 年，华为被迫接受美国外国投资委员会建议，撤消收购 3Leaf 公司特殊资产的申请；同样是在 2011 年，美国商务部阻止华为参与国家应急网络项目招标。

再看美国国防部立项的几个大数据项目：多尺度异常检测(ADAMS)项目，解决大规模数据集的异常检测和特征识别的问题；网络内部威胁(CINDER)计划，旨在开发新的方法来检测军事计算机网络与网络间谍活动，提高对网络威胁检测的准确性和速度；Insight 计划，主要解决目前情报、监视和侦察系统的不足，进行网络威胁的自动识别和非常规的战争行为。其他部门包括国土安全部、能源部、卫生和人类服务部、国家航天总局、美国国家科学基金会、美国国家安全局、美国地质调查局纷纷推出大数据项目。奥巴马指出："通过提高我们从大型复杂的数据集中提取知识和观点的能力，加快科学与工程前进步伐，改变教学研究，加强国家安全。"

麦肯锡给出的大数据定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。但它同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据。国际数据公司(IDC)从大数据的四个特征来定义，即海量的数据规模(Volume)、快速的数据流转和动态的数据体系(Velocity)、多样的数据类型(Variety)、巨大的数据价值(Value)。

大数据是一个宽泛的概念，见仁见智。上面几个定义，无一例外地都突出了"大"字。诚然"大"是大数据的一个重要特征，但远远不是全部。国内学者在调研多个行业后，给出了自己的定义：大数据是"在多样的或者大量数据中，迅速获取信息的能力"。这个定义更关心大数据的功用，它能帮助大家干什么。在这个定义中，重心是"能力"。大数据的核心能力，是发现规律和预测未来。

1.1.1 大数据的特征

1. 数据爆炸

截至 2011 年，全球拥有互联网用户数已达到 20 亿；RF 旧标签在 2005 年的保有量仅有 13 亿个，但是到 2010 年这个数字超过了 300 亿；2006 年资本市场的数据比 2003 年增长了 17.5 倍；目前新浪微博上每天上传的微博数超过 1 亿条；Facebook 每天处理 10TB 的数据；世界气象中心积累了 220TB 的 Web 数据，9PB 其他类型数据……

根据国际数据公司(IDC)的《数据宇宙》报告显示:2008 年全球数据量为 0.5ZB，2010 年为 1.2ZB，人类正式进入 ZB 时代。更为惊人的是，2020 年以前全球数据量仍将保持每年 40%多的高速增长，大约每两年就翻一倍，这与 IT 界人尽皆知的摩尔定律极为相似，姑且可以称之为"大数据爆炸定律"。预计 2015 年全球数据量将达到 7.9ZB，2020 年将突破 35ZB，是 2008 年的 70 倍、2011 年的 29 倍。

同时，根据互联网数据中心的《中国互联网市场洞见：互联网大数据技术创新研究 2012》报告显示：截至 2011 年年底，中国互联网行业持有的数据总量已达到 1.9EB，预计 2015 年该规模将增长到 8.2EB 以上。

人类社会的数据量在不断刷新一个个新的量级单位，已经从 TB、PB 级别跃升至 EB、ZB 级别。然而，35ZB、8.2EB 究竟是一个什么样的概念呢，为此，首先了解下面几组关于数据衡量单位的公式：

1B=8 bit

1KB=1024B≈1000 byte

1MB=1024KB≈1000 000 byte

1GB=1024MB≈1000 000 000 byte

1TB=1024GB≈1000 000 000 000 byte

1PB=1024TB≈1000 000 000 000 000 byte

1EB=1024PB≈1000 000 000 000 000 000 byte

1ZB=1024EB≈1000 000 000 000 000 000 000 byte

1YB=1024ZB≈1000 000 000 000 000 000 000 000 byte

一本《红楼梦》共有 87 万字(含标点)，每个汉字占两个字节，即 1 个汉字=2B，由此计算 1EB 约等于 6626 亿部红楼梦。美国国会图书馆是美国四个官方图书馆之一，也是全球最重要的图书馆之一，截至 2011 年 4 月，藏书约为 1.5 亿册，收录数据 235TB，1EB 约等于 4462 个美国国会图书馆的数据存储量。

2．非结构化数据

人们日常工作中接触的文件、照片、视频，都包含大量的数据，蕴含大量的信息。这一类数据有一个共同的特点，大小、内容、格式、用途可能都完全不一样。以最常见的 Word 文档为例，最简单的 Word 文档可能只有寥寥几行文字，但也可以混合编辑图片、音乐等内容，成为一份多媒体的文件，来增强文章的感染力。这类数据通常称为非结构化数据。

与之相对应的另一类数据，就是结构化数据。这类数据大家可以简单地理解成表格里的数据，每一条的结构都相同。大家每月都能领到工资条，每个工资条结构都是一样的，当然里面的工资和缴纳的个税、保险不同。每个人的工资条依次排列到一起，就形成了工资表。利用计算机处理结构化数据的技术比较成熟，从事会计、审计等工作的人，利用 Excel 工具很容易进行加减乘除、汇总、统计之类的运算。如果进行大量的运算，一些商业数据库软件就会派上用场，它们专门用于存储和处理这些结构化的数据。

但不幸的是，人们日常接触到的数据绝大部分都是非结构化的。有的咨询机构认为非结构化数据占企业总数据量的 80%，也有机构认为占 95%，总之，没有权威、准确的统计。如何像处理结构化数据那样，方便、快捷地处理非结构化数据，是信息产业一直以来的努力方向之一。在这个领域，信息业是走了不少弯路的。起初人们借助结构化数据处理的成果，把非结构化数据也用传统的数据库(基于关系型的数据库)来处理。非结构化数据的一大特点就是“龙生九子，各个不同”，硬要套到一个模子里面来，结果是费力不讨好。于是，人们一度认为大量的非结构化数据是难以处理的。

幸运的是，谷歌公司在为公众提供页面搜索服务的同时，顺便解决了大量网页、文档这类数据的快速访问难题，成为大数据技术的先驱。雅虎公司的一个开发小组，利用谷歌的成果成功地开发出大数据处理的一套程序框架，这就是众所周知的 Hadoop。目前，这个领域非常活跃，发展可谓日新月异。

这些公司的实践，让大家面对其他各类的非结构化数据处理难题重建信心，如高清图像、视频、音频等的处理技术都已驶入了快车道。

另外，社交网络上的表现人们情绪的数据日益丰富。例如：〔笑脸〕、〔鼓掌〕、〔握手〕、〔愤怒〕、〔纪念〕等代表人们心情的标准化图释的大量使用，无疑表达了人们对某一事件的总体情绪，可能昭示线下会发生某些行为。

3．大数据应用模式的变化

1) 数据的快速应用

这一条是传统的数据应用和大数据应用最重要的区别。过去的十几年间，金融、电信等行业都经历了核心应用系统从散落在各地市到逐步统一到总部的过程。大量数据集中后，带来的第一个问题就是大大延长了各类报表生成时间。业界一度质疑，快速地在海量数据中提取信息，是否可行。

谷歌公司在这方面的贡献，无疑是开创性的。它的搜索服务，等于向信息业界宣布，1s 之内就能检索全世界的网页，而且可以找到你想要的结果。在写作本段的时候，当用谷歌搜索关键词“大数据”时，提示“找到约 46300000 条结果(用时 0.37s)”。谷歌等于为大数据应用确立了一个标杆。如果超过 1s 的数据应用，就会给用户带来不良的使用体验。甚至在某些情况下，如果应用速度达不到“秒”级，其商业价值就会大打折扣。

下面来看一个营销的例子：价格越贵的东西，人们购买时就会越犹豫，反复掂量自己的钱包。相反，价格越便宜的东西，人们购买时更多根据一时的喜好，呈现冲动型购买的特征。京东商城根据消费者购买商品的特征，将其分为四种类型，其中冲动型购买者占 37%。冲动嘛，自然一闪即逝。所以能否在用户冲动的瞬间及时送达精准的商品信息，就成为了提高商品销售的关键所在。幸运的是，社交型互联网的应用，如美国的 Facebook、中国的微博和微信，提供了侦测人们偏好和兴趣的接口，使得这种精准的营销在大数据时代成为可能。

在以高频交易为主的股票市场，比别人快 0.02s，就可能获得惊人的超额收益。所以，有人为了抢这宝贵的 20ms，单独建了一条从西海岸到东海岸横跨美国的光纤，也有人干脆就呆在纽交所相同的街区。这种毫秒级时差造成的商业机会，也许会随着大数据的普及应用而在其他行业不断上演。

2) 孤立的数据是没有价值的

Facebook、微博为代表的社交网络应用，构建了普遍关联用户行为数据。本来大家在网络上浏览网页、购买商品，游戏休闲等，都是互不关联的。尤其是智能手机的普及，大家的网络行为更趋向于碎片化。这些碎片化数据如果没有关联，是难以进行分析并加以利用的。但是社交网络提供了统一的接口，让大家无论是玩游戏还是买商品，都能够方便轻松地分享到微博上。微博扮演了用户行为数据连接器的角色。用户在网络上的碎片化行为，经由社交网络，就能完整地勾勒出一幅生动的网络生活图景，真实地反映了用户的偏好、性格、态度等特征，这其中蕴育了大量的商业机会。

反之，孤立的数据，其价值要远远小于广泛连接的数据。然而，数据孤岛现象普遍存在。个人计算机中的文件，虽然按照目录分门别类的存放，但是之间的内容关系往往杂乱无章。企业中各部门壁垒林立，大家更倾向于尽可能地保护自己的数据。我国政府部门的数据孤岛现象更为严重，甚至可以称为“数据割据”现象。在数据孤岛的影响下，

难以发挥大数据中蕴藏的价值。

3) 活性越高价值越大

有一家公司寄来数据样本，希望帮他们评估这些数据的潜在商业价值，虽然数据量很大，但是数据更新的频率大概是每月一次。在判断数据的价值时，要看拥有数据的规模和数据活性。所谓活性，也就是数据更新的频率，更新的频率越高，数据的活性越大；更新的频率越低，数据的活性越小。一般而言，数据活性更高的数据集，蕴含更丰富的信息。所以，这家公司如果想在大数据领域有所作为的话，就需要想办法提高数据的活性。

1.1.2 数据是关键资产

长期以来，经济学著作中，土地、资本和人力并称为企业的生产要素。人类进入工业时代以来，技术成为独立的生产要素之一。在信息时代，数据将成为独立的生产要素。有人把“数据”比喻为工业时代的石油，事实上“数据”和农耕时代“土地”的属性更加接近。

在互联网领域，令人称道的谷歌、亚马逊和 Facebook，分别拥有不同的数据资源。谷歌之所以能打破微软垄断的铁幕，依仗的就是世界上最大的网页数据库，并建立了充分发挥这些数据资产潜在价值的数字媒体商业模式。许多公司开始把谷歌当作竞争对手，依葫芦画瓢推出和谷歌类似的搜索引擎，但是，包括微软公司在内没有一家可以撼动谷歌的根基，直到 Facebook 推出 graph search 引擎，才让谷歌感到真正的威胁。原因很简单，Facebook 拥有谷歌缺乏的一类数据资产——人们的关系数据，这是 Facebook 区别于所有竞争对手的关键因素。当谷歌和 Facebook 打得不可开交的时候，亚马逊却乐得坐山观虎斗。因为无论是谷歌还是 Facebook 都可以帮助亚马逊卖出更多的商品。亚马逊拥有世界上最大的商品电子目录。当所有的公司对苹果的平板电脑横扫世界束手无策的时候，亚马逊庞大的商品帮了大忙，人们愿意购买亚马逊的平板电脑，因为可以免费获得海量的图书。和亚马逊相比，缺少电子图书，恰巧是苹果的弱项。所以没有独一无二的数据资产，几乎无法参与巨人间的游戏。

从表 1-1 中，我们可以看出谷歌是如何通过收购来丰富其数据资产的。

表 1-1 谷歌公司的典型收购

收购日期	公司	性质	改造/整合对象	数据资产
2003 年 2 月	Pyra Labs	博客软件	Bloger	收集博客数据
2003 年 4 月	Applied Semantics	网络广告	AdSense、AdWords	
2004 年 7 月	Picasa	图像管理工具		收集图像数据
2004 年	ZipDash、Where2、keyhole	地图	谷歌地图	这三家丰富的地图数据，获得分析技术
2005 年 7 月	Current	宽带因特网连接	网络骨干	基础电信数据
2005 年 8 月	Android	移动设备操作系统		获得移动设备使用数据
2006 年 3 月	Upstartle	文字处理器	Google Docs	获得文档数据
2006 年 10 月	YouTube	视频分享网站		获得视频数据

(续)

收购日期	公司	性质	改造/整合对象	数据资产
2006 年 8 月	Neven Vision	人脸识别	Picasa	图像挖掘技术
2007 年 4 月	DoubleClick	网络广告	Adsense	广告技术
2007 年 6 月	Panoramio	照片分享		获得图像数据
2010 年 5 月	Simplify	音乐同步	Android	获取音乐数据
2010 年 5 月	Ruba	旅行向导		获取出行数据
2010 年 8 月	Slide.com	社交游戏	Google+	获取娱乐数据
2010 年 7 月	ITA	航班信息	GoogleFlight	获取出行数据
2011 年 8 月	摩托罗拉	智能手机		控制产业链

中国的互联网市场也是硝烟弥漫。阿里巴巴旗下的一淘网，抓取京东商城的客户评论数据；京东则采取技术手段屏蔽一淘的“爬虫”。另一方面，电商则纷纷抓取竞争对手的各类商品的实时价格，作为评估对手战略动向、促销战术的重要依据。这还只是在表面现象，事实上互联网平台型的公司，都在围绕数据资产为核心整合产业生态。它们推出新的产品、新的服务，就会收集更多类型的数据。数据越多，不同类型数据之间的关联性、实时性越强，就会提炼出更有价值的信息指导它们开展各类精准的广告业务、金融业务。马云在 2012 年网商大会上，鲜明地提出阿里巴巴未来的战略是围绕三大方向即平台、金融、数据展开。平台汇聚数据，数据衍生金融，金融反哺平台。可见互联网公司对于数据资产的战略价值认知最为深刻，行动最为果断。京东商城也已经启动供应链金融服务。表面上看，电子商务公司和金融机构井水不犯河水，其实电商凭借数据积累，已经侵入到金融行业的腹地。

1.1.3 建立数据思维模式

数据思维的重要性远远超过数据资产，大数据时代最重要的是建立数据思维，而非仅仅盯住数据资产，具备数据思维，才能够积累数据资产。因此，大数据时代最稀缺的资源是人才，大数据人才的招募、培养与使用将是大数据创新与创业所面临的最大挑战。通过合理的模式释放大数据人才价值的过程同时也是释放大数据价值的过程，而数据的积累、挖掘、分析、归纳、整理，是新时代人才所必须具备的基本素养。

1.2 数据科学的兴起

在大数据时代，科学领域里的表现是数据科学的兴起。常常有人问：多大才算是“大数据”？“大数据”和“海量数据”有什么区别？其实根本没有必要为“大数据”这个名词的确切含义而纠结。“大数据”是一个热点名词，它代表的是一种潮流、一个时代，它可以有多方面的含义。“海量数据”是一个技术名词，它强调数据量之大。而数据科学则是一门新兴的学科。为什么要强调数据科学？它和已有的信息科学、统计学、机器学习等学科有什么不太一样？

数据科学主要包括两个方面：用数据的方法来研究科学和用科学的方法来研究数据。

前者包括生物信息学、天体信息学、数字地球等领域，后者包括统计学、机器学习、数据挖掘、数据库等领域。这些学科是数据科学的重要组成部分，但只有把它们有机地放在一起，才能形成整个数据科学的全貌。

1.2.1 用数据的方法来研究科学

用数据的方法来研究科学，最典型的例子是开普勒关于行星运动的三大定律。

开普勒的三大定律是根据他的前任，一位名叫第谷的天文学家留给他的观察数据总结出来的。表 1-2 是一个典型的例子，这里列出的数据是行星绕太阳一周所需要的时间(以年为单位)和行星离太阳的平均距离(以地球与太阳的平均距离为单位)。从这组数据可以看出，行星绕太阳运行的周期的二次方和行星离太阳的平均距离的二次方成正比。这就是开普勒的第三定律。

表 1-2 太阳系八大行星绕太阳运动的数据

行星	周期/年	平均距离	周期^2/距离^2	行星	周期/年	平均距离	周期^2/距离^2
水星	0.241	0.39	0.98	木星	11.8	5.20	0.99
金星	0.615	0.72	1.01	土星	29.5	9.54	1.00
地球	1.00	1.00	1.00	天王星	84.0	19.18	1.00
火星	1.88	1.52	1.01	海王星	165	30.06	1.00

开普勒虽然总结出行星运动的三大定律，但他并不理解其内涵。牛顿则不然，牛顿用他的第二定律和万有引力定律把行星运动归结成一个纯粹的数学问题，即一个常微分方程组。如果忽略行星之间的相互作用，那么这就成了一个两体问题。因此很容易求出这个常微分方程组的解，并由此推出开普勒的三大定律。

牛顿运用的是寻求基本原理的方法，远比开普勒的方法深刻。牛顿不仅知其然，而且知其所以然。所以牛顿开创的寻求基本原理的方法成了科学研究的首选模式。这种方法在 20 世纪初期达到了顶峰：在它的指导下，物理学家们发现了量子力学。从原则上来讲，日常生活中的自然现象都可以从量子力学的角度来解释。量子力学提供了研究化学、材料科学、工程科学、生命科学等几乎所有自然和工程学科的基本原理。这应该说是很成功的，但事情远非这么简单。正如狄拉克指出的那样，如果以量子力学的基本原理为出发点去解决这些问题，那么其中的数学问题太难了。所以如果要想有进展，还是必须做妥协，也就是说要对基本原理作近似。

用数据的方法来研究科学问题，并不意味着就不需要模型了。只是模型的出发点不一样，不是从基本原理的角度去找模型。就拿图像处理的例子来说，基于基本原理的模型需要描述人的视觉系统以及它与图像之间的关系，而通常的方法则可以是基于更为简单的数学模型如函数逼近的模型。

1.2.2 用科学的方法来研究数据

怎样用科学的方法来研究数据？这包括以下几个方面的内容：数据的获取、存储和数据的分析。这也就是数据工程的基本出发点。数据工程是以数据作为研究对象、以数

此为试读，需要完整PDF请访问：www.ertongbook.com