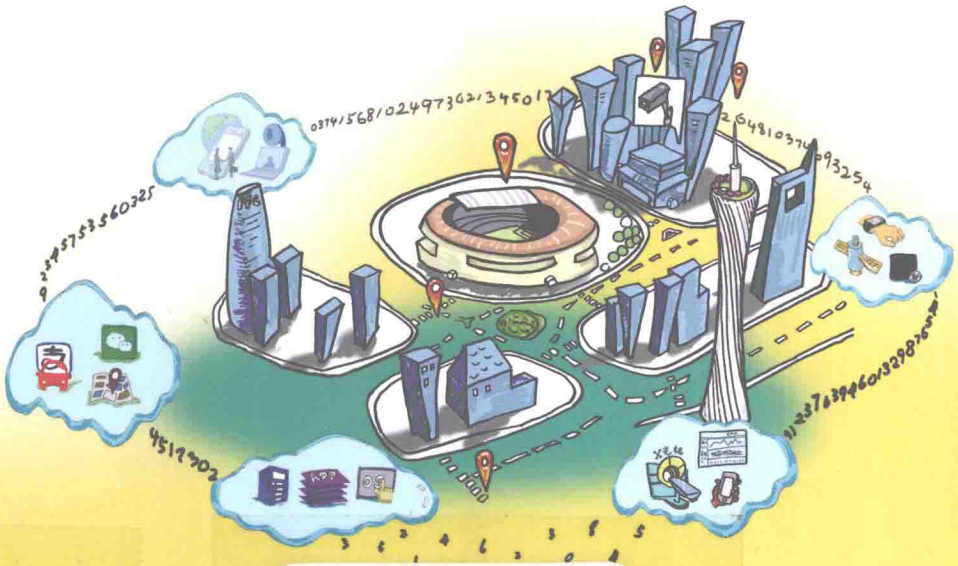
高新技术科普丛书
（第3辑）

主编　汪疆平

数据如海可淘金

——大数据技术及其在智慧城市的应用

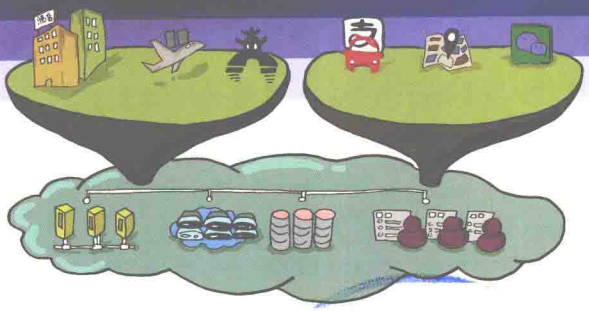


南方出版传媒
广东科技出版社 | 全国优秀出版社

高新技术科普丛书（第3辑）

数据如海可淘金

——大数据技术及其在智慧城市的应用



主编　汪疆平

SPM 南方出版传媒
广东科技出版社 | 全国优秀出版社
· 广　州 ·

图书在版编目（CIP）数据

数据如海可淘金：大数据技术及其在智慧城市的应用 / 汪疆平主编. —广州：广东科技出版社，2015.7

（高新技术科普丛书. 第3辑）

ISBN 978-7-5359-6069-6

Ⅰ. ①数… Ⅱ. ①汪… Ⅲ. ①数据处理—普及读物 Ⅳ. ① TP274-39

中国版本图书馆 CIP 数据核字（2015）第 051018 号

数据如海可淘金——大数据技术及其在智慧城市的应用

Shujuruhai Ketaojin——Dashujujishu Jiqi zai Zhihuichengshi de Yingyong

丛书策划：崔坚志

责任编辑：林　旸　区燕宜

装帧设计：柳国雄

责任校对：黄慧怡

责任印制：罗华之

出版发行：广东科技出版社

（广州市环市东路水荫路 11 号　邮政编码：510075）

http：//www.gdstp.com.cn

E-mail：gdkjyxb@gdstp.com.cn（营销中心）

E-mail：gdkjzbb@gdstp.com.cn（总编办）

经　　销：广东新华发行集团股份有限公司

印　　刷：广州市岭美彩印有限公司

（广州市荔湾区花地大道南海南工商贸易区 A 幢　邮政编码：510385）

规　　格：889mm×1 194mm　1/32　印张 5　字数 120 千

版　　次：2015 年 7 月第 1 版

2015 年 7 月第 1 次印刷

定　　价：23.80 元

序一

PREFACE

精彩绝伦的广州亚运会开幕式，流光溢彩、美轮美奂的广州灯光夜景，令广州一夜成名，也充分展示了广州在高新技术发展中取得的成就。这种高新科技与艺术的完美结合，在受到世界各国传媒和亚运会来宾的热烈赞扬的同时，也使广州人民倍感自豪，并唤起了公众科技创新的意识和对科技创新的关注。

广州，这座南中国最具活力的现代化城市，诞生了中国第一家免费电子邮局；拥有全国城市中位列第一的网民数量；广州的装备制造、生物医药、电子信息等高新技术产业发展迅猛。将这些高新技术知识普及给公众，以提高公众的科学素养，具有现实和深远的意义，也是我们科学工作者责无旁贷的历史使命。为此，广州市科技创新委员会与广州市科技进步基金会资助推出《高新技术科普丛书》。这又是广州一件有重大意义的科普盛事，这将为人们提供打开科学大门、了解高新技术的“金钥匙”。

丛书内容包括生物医学、电子信息以及新能源、新材料等三大板块，有《量体裁药不是梦——从基因到个体化用药》《网事真不如烟——互联网的现在与未来》《上天入地觅“新能”——新能源和可再生能源》《探“显”之旅——近代平板显示技术》《七彩霓裳新光源——LED 与现代生活》以及关

于干细胞、生物导弹、分子诊断、基因药物、软件、物联网、数字家庭、新材料、电动汽车等多方面的图书。

我长期从事医学科研和临床医学工作，深深了解生物医学对于今后医学发展的划时代意义，深知医学是与人文科学联系最密切的一门学科。因此，在宣传高新科技知识的同时，要注意与人文思想相结合。传播科学知识，不能视为单纯的自然科学，必须融汇人文科学的知识。这些科普图书正是秉持这样的理念，把人文科学融汇于全书的字里行间，让读者爱不释手。

丛书采用了吸收新闻元素、流行元素并予以创新的写法，充分体现了海纳百川、兼收并蓄的岭南文化特色。并按照当今“读图时代”的理念，加插了大量故事化、生活化的生动活泼的插图，把复杂的科技原理变成浅显易懂的图解，使整套丛书集科学性、通俗性、趣味性、艺术性于一体，美不胜收。

我一向认为，科技知识深奥广博，又与千家万户息息相关。因此科普工作与科研工作一样重要，唯有用科研的精神和态度来对待科普创作，才有可能出精品。用准确生动、深入浅出的形式，把深奥的科技知识和精邃的科学方法向大众传播，使大众读得懂、喜欢读，并有所感悟，这是我本人多年来一直最想做的事情之一。

我欣喜地看到，广东省科普作家协会的专家们与来自广州地区研发单位的作者们一道，在这方面成功地开创了一条科普创作新路。我衷心祝愿广州市的科普工作和科普创作不断取得更大的成就！

中国工程院院士 

序二
PREFACE

让高新科学技术星火燎原

21 世纪第二个十年伊始，广州就迎来喜事连连。广州亚运会成功举办，这是亚洲体育界的盛事；《高新技术科普丛书》面世，这是广州科普界的喜事。

改革开放 30 多年来，广州在经济、科技、文化等各方面都取得了惊人的飞跃发展，城市面貌也变得越来越美。手机、电脑、互联网、液晶电视大屏幕、风光互补路灯等高新技术产品遍布广州，让广大人民群众的生活变得越来越美好，学习和工作越来越方便；同时，也激发了人们，特别是青少年对科学的向往和对高新技术的好奇心。所有这些都使广州形成了关注科技进步的社会氛围。

然而，如果仅限于以上对高新技术产品的感性认识，那还是远远不够的。广州要在 21 世纪继续保持和发挥全国领先的作用，最重要的是要培养出在科学领域敢于突破、敢于独创的领军人才，以及在高新技术研究开发领域勇于创新的尖端人才。

那么，怎样才能培养出拔尖的优秀人才呢？我想，著名科学家爱因斯坦在他的“自传”里写的一段话就很有启发意义：“在 12 ~ 16 岁的时候，我熟悉了基础数学，包括微积

分原理。这时，我幸运地接触到一些书，它们在逻辑严密性方面并不太严格，但是能够简单明了地突出基本思想。”他还明确地点出了其中的一本书：“我还幸运地从一部卓越的通俗读物（伯恩斯坦的《自然科学通俗读本》）中知道了整个自然领域里的主要成果和方法，这部著作几乎完全局限于定性的叙述，这是一部我聚精会神地阅读了的著作。”——实际上，除了爱因斯坦以外，有许多著名科学家（以至社会科学家、文学家等），也都曾满怀感激地回忆过令他们的人生轨迹指向杰出和伟大的科普图书。

由此可见，广州市科技创新委员会与广州市科技进步基金会，联袂组织奋斗在科研与开发一线的科技人员创作本专业的科普图书，并邀请广东科普作家指导创作，这对广州今后的科技创新和人才培养，是一件具有深远战略意义的大事。

这套丛书的内容涵盖电子信息、新能源、新材料以及生物医学等领域，这些学科及其产业，都是近年来广州重点发展并取得较大成就的高新科技亮点。因此这套丛书不仅将普及科学知识，宣传广州高新技术研究和开发的成就，同时也将激励科技人员去抢占更高的科技制高点，为广州今后的科技、经济、社会全面发展作出更大贡献，并进一步推动广州的科技普及和科普创作事业发展，在全社会营造出有利于科技创新的良好氛围，促进优秀科技人才的茁壮成长，为广州在21世纪再创高科技辉煌打下坚实的基础！

中国科学院院士 张景中

前言

FOREWORD

大数据已经充斥到我们生活的方方面面，衣、食、住、行、社交、教育、医疗等都已经数据化，当各方面的数据汇集在一起，就形成了大数据。分析这些大数据，我们能从中发现有价值的东西，例如：预测交通状况、探究遗传病背后的原因、获知公众事件的发展趋势……通过收集数据，深入探知事物背后的原因，改变生产生活方式，精确地预知未来的发展动态，就是大数据的价值所在。

大数据将改变整个世界的运转模式，我国在信息技术方面落后于发达国家，但是差距在不断缩小，而大数据给我国提供了一个弯道超车的机会：我国的人口基数、网络用户数量、手机用户数量等都处于世界第一，这是我们在大数据领域最大的资源优势，这个优势为我国的大数据产业快速发展奠定了重要的基石。预计在不久的将来，我国会基于大数据技术产生大量创新型的模式、应用和分析结果，从而能够屹立在大数据时代的浪头。因此，了解大数据将带来的颠覆性变革，对于我们每个人都是非常重要的，特别是对于正在学习知识的青少年，掌握大数据的应用将使我们的学习生活更明确、更高效。

物联网、云计算、大数据、移动互联等新兴技术的发展，是推动城市“智慧化”的关键技术，这些新兴信息技术将给

城市带来巨大的变革，实现城市级的信息共享、协同运作的新模式。特别是大数据技术，是实现城市“智慧”的核心。大数据将带来生产力和生产方式的大变革，发展出大量创新型应用，引领世界城市进入一个全新的发展阶段。

正是基于对大数据和智慧城市的这种理解，我们编写了这本《数据如海可淘金——大数据技术及其在智慧城市的应用》，本书系统地介绍了大数据的方方面面，与一般大数据的技术书籍不同，本书用通俗易懂的语言，描述了大数据的技术奥秘和应用场景。我们力图深入浅出地介绍大数据的技术，特别是与日常生活中的场景结合在一起，描绘大数据如何在我们身边提供服务，使得只具有一般信息技术知识的读者也能清晰地理解什么是大数据。另外，由于大数据技术刚刚兴起，大部分大数据的案例发生在国外，为了大家更清晰地知道国内的大数据应用情况，本书尽可能多地采用了国内的案例，这也是在众多大数据书籍中所少见的。

更重要的是，我们将正在我国蓬勃发展的智慧城市建设与大数据结合在一起，描绘了未来智慧城市的生活，说明大数据如何推动“智慧”的实现。本书的大部分读者都可能正在经历我国高速城市化发展中的各种问题，并且在为改变城市的发展路径贡献自己的力量，我们希望通过引入大数据技术，使得我们的城市能够早日进入良性发展的轨道，早日让城市居民享受到“经济低碳、城市智慧、社会文明、生态优美、城乡一体、生活幸福”的新型城市化生活。

大数据带给我们新的发展机遇，智慧城市改变我们的生活环境，如何将美好的愿景变成现实，需要我们每个人都贡献自己的力量，哪怕是只了解大数据和智慧城市是什么，都能够帮助我们更好地采取行动。如果能够帮助读者更清晰、准确地理解这两项内容，我们也就达到了编写本书的目的。

目录
CONTENTS

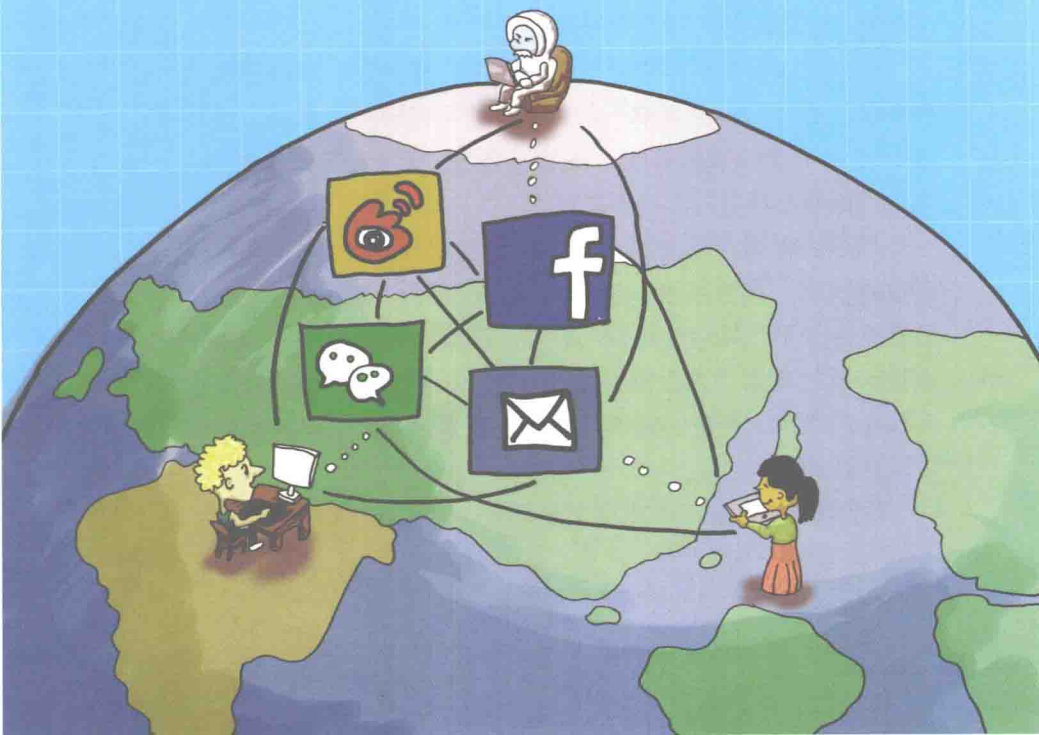
一 信息爆炸的硝烟

二 “捕捉”大数据

三 大数据的“藏宝洞”

四 在大数据金山里“淘金”

一　信息爆炸的硝烟



奥斯卡奖的预测

2014 年第 83 届奥斯卡结果揭晓前，你觉得谁最有可能获奖？最佳男主角是屡战屡败的“小李”莱奥纳多 · 迪卡普里奥，还是突破自我的马修 · 麦康纳？最佳导演该是展现了科技进步与宇宙思索的阿方索 · 卡隆，还是展现了人性复杂的马丁 · 西科塞斯？且慢，我们来看其中一个预测——

最佳影片：《为奴十二年》88.7%；

最佳导演：阿方索 · 卡隆（《地心引力》）97.6%；

最佳男主角：马修 · 麦康纳（《达拉斯买家俱乐部》）90.9%。

结果可想而知，这个猜测全部应验。是谁做出了这个预测？答案是 David Rothschild，一位来自微软纽约研究院的经济学家。这一届的奥斯卡，他竟猜中了 24 个奖项中的 21 项！而早在 2013 年，他就做过类似尝试，结果猜中了奥斯卡全部 24 个奖项中 19 个的归属。

Rothschild 并不是任何一位提名者的拥护者，他的预测跟明星、影迷们的预测都不同，没有掺杂任何私人趣味。过去从来没有谁的预测有这么高的准确度。那么，Rothschild 是如何做到如此高的准确率的？

原来，他有个撒手锏，名叫“大数据”，他的预测纯粹以数据说话。具体来说，Rothschild 先设好一个看似简单的数据聚合模型，然后去寻找与各位入围者相关的数据，再做调查。最后运用“大数据技术”，给各位入围者都设定好一个获奖的概率，有第一概率、第二概率之分。第一概率者最终就是获奖者。

David Rothschild 不止预测奥斯卡。在 2012 年的美国总

统大选中，他成功猜对了 51 个选区中的 50 个区的结果，准确率高达 98%。现在，他在网站上主要发布体育和政治方面的预测。

1 信息爆炸催生大数据时代

数据如何从“小”变“大”？

数据最早是怎么被记录的?

在远古时代，人们记事依靠的是一根绳子。在绳子上打

一个结，便是记一件事，如果要记住两件事，就打两个结，如此类推。这是最为原始的记录方式，虽简单却不可靠。因为，当人们在绳子上打了太多的结，恐怕也会记不清谁是谁了。可以说，在人类社会发展之初，数据的存留量是十分有限的。

随着人类文明的发展，文字的出现无疑具有划时代的意义，犹如曙光照耀一方大地。在早期，文字是直接具体的表意文字（象形文字），具有很强的图画特质，如山、日、月等文字至今仍然能够看出其原始的符号形态。到后来，文字的书写越来越规范，也越来越便于书写。正是这些不起眼的笔画勾勒出的符号，穿越了无尽的时空，将前人的思想保留了下来，使文明得以生生不息。在文明社会中，文字作为高效的信息传播工具，大大提高了文化、思想、艺术、技术等人类文明的传播速度和效率。

直到电脑的出现，人类记录数据的方式才有了本质上的改变，也寻找到一种可以代替人的计算工具。早在 17 世纪，