BIG DATA

BIG DATA FOR DUMMIES

认识大数据

大数据的历史、内涵、哲学与技术，你了解多少？

张玉宏 ◆ 著

百位大牛 PK 大数据，听谁的？

正方 第 1 位 《数据之巅》作者涂子沛说：“本质上，数据表征的是已发生的事实，而其核心的作用则是对未来的预测。从远古的小数据，到当今的大数据，数据的作用，莫不如此。”

反方 第 2 位 《黑天鹅：如何应对不可预知的未来》作者塔勒布说：“人类文明之路不断向前，途中不断发现未知，不断出现新情况。因此，我们压根就没有足够的所谓大数据，去发现和预测新场景、新状况、新未知。如果非要说我们能够实现预测人类行为这一远大图景，那也仅仅是在经验主义、在过去数据的基础上，得出一个大概均值罢了。”

……

正方 第 99 位 《大数据时代》作者舍恩伯格说：“大数据就是全数据（n=all），无需采样，也不再有采样偏差的问题，因为采样已经包含了所有数据。”

反方 第 100 位 《你的数字感：走出大数据分析与解读的误区》作者、美国纽约大学统计学教授冯启思（Kaiser Fung）说：“不要简单地假定自己掌握了所有有关的数据。‘n=all’仅仅是对数据的一种假设，而非现实。”

本书博采众长，从大数据的历史、内涵、哲学与技术 4 个维度进行了系统的阐述，

终极目的是让读者形成自己的大数据认知体系！

北京大学出版社
PEKING UNIVERSITY PRESS

大数据

大数据的历史、内涵、哲学与技术，你了解多少？

张玉宏 ◆ 著

北京大学出版社
PEKING UNIVERSITY PRESS

内容提要

当下，大数据是一个热门的话题，很多领域的学者，从不同的角度进行了深入讨论。本书从大数据的历史、内涵、哲学和技术四个角度，全面解析大数据，让读者对大数据有更深入的理解。

全书共 11 章，大致分为 4 大块：第 1 ~ 3 章主要漫谈了大数据有趣的历史，包括数据的启蒙、信息载体的演变和数据管理的发展脉络。第 4 ~ 6 章主要聊聊大数据的内涵，包括大数据与哲学及第四科学范氏的关联。第 7 ~ 9 章是大数据的杂谈，包括大数据的用途、可能面临的陷阱以及通过小故事对大数据进行一些反思，第 10 ~ 11 章主要涉及大数据的技术，包括 100 余篇大数据论文的漫读及 Hadoop 的初级实战篇。

图书结构完整，行文幽默，并以图文并茂、通俗易懂的方式力图让读者心有余力地品味大数据。图书援引了数以百计大家牛人的观点，或褒或贬，高手过招，精彩纷呈，是一本不容读者错过的大数据图书。

图书在版编目 (CIP) 数据

品味大数据 / 张玉宏著 . — 北京：北京大学出版社，2016.10
ISBN 978-7-301-27609-9

Ⅰ . ①品… Ⅱ . ①张… Ⅲ . ①数据处理—研究 Ⅳ . ① TP274

中国版本图书馆 CIP 数据核字 (2016) 第 231945 号

书　　名	品味大数据 PINWEI DA SHUJU
著作责任者	张玉宏　著
责任编辑	尹毅
标准书号	ISBN 978-7-301-27609-9
出版发行	北京大学出版社
地　　址	北京市海淀区成府路 205 号　100871
网　　址	http：//www. pup. cn　　新浪微博：@ 北京大学出版社
电子信箱	pup7@ pup. cn
电　　话	邮购部 62752015　发行部 62750672　编辑部 62580653
印 刷 者	北京大学印刷厂
经 销 者	新华书店
	787 毫米 ×1092 毫米　16 开本　26.25 印张　620 千字
	2016 年 10 月第 1 版　2016 年 10 月第 1 次印刷
印　　数	1–3000 册
定　　价	59.00 元

Preface

序

在路上，学而时习之

当下，大数据（Big Data）被炒得如此之火，以至于很多人都在谈论它。有人逢人说项，甘做它的布道者；也有人对大数据的炙手可热现象嗤之以鼻。然而，到底什么才是大数据，却没有太多人能真正说得透彻。

引用 TED[①] 的创始人丹·艾瑞里（Dan Ariely）写在自己脸谱网（Facebook）上的一句玩笑话，“大数据就像青少年谈性，每个人都津津有味地谈论它，却没有人真的知道如何来做，而每个人又认为其他人在做，于是每个人都声称自己在做。”[②]

艾瑞里对大数据的理解，无厘头中不乏有趣，他形象地道明了，很多人在谈论“大数据”时，其实都有这么一种似是而非的感觉。

虽然对大数据并未涉猎多深，可能是所学专业为计算机的缘故，每每有友人问我这个“专业人士”有关大数据的问题，心里不免发虚，甚至“两股战战，几欲先走”。迫于这种压力，我决定一探究竟，大数据到底是个什么东西？为何几乎身边每个人都在谈？就这样，我踏上了学习大数据之路。

随着对大数据的不断学习，我逐渐明白：对于大数据，真想要弄个清清楚楚、明明白白，并非易事！学习的道路也并不平坦。为了避免“偏信则暗”，就得集众家之所长，花费大量时间和精力，去阅读大量文献和书籍，“兼听则明”。

为了让自己更清楚地理解大数据，我就一些有关大数据的核心观点、关键技术，以及一些读书体会、些许感悟和随想等落笔成文，形成了一系列的科技随笔。

这些科技随笔，如“来自大数据的反思：需要你读懂的 10 个小故事”“大数据，小数据，哪道才是你的菜？”“大数据专家 Bernard Marr：大数据是如何对抗癌症的？”及“PayPal 高级工程总监：读完这 100 篇论文就能成大数据高手”（编译）等，先后在知名中文 IT 社区 CSDN 上作为头条发表，并被很多大数据网站及微信公众号平台上转载，受到了读者的普遍好评，甚感欣慰。

---

① TED（Technology，Entertainment，Design，即技术、娱乐、设计）是美国一家著名的非营利性机构，致力于传播创意，以组织 TED 大会著称，这个会议的宗旨是“用思想的力量来改变世界”。

② 对应的英文原文是：Big data is like teenage sex：everyone talks about it，nobody really knows how to do it，everyone thinks everyone else is doing it，so everyone claims they are doing it...

后来，北京大学出版社和龙马高新教育的编辑老师鼓励我，让我把大数据的科技随笔和学习笔记集结成书，写一本半学术化的大数据科普图书。编辑老师“盛情难却”，但我又“诚惶诚恐”。

这是因为，在中国，“文人相轻”由来已久。或许有的读者会很“犀利”地质疑，现在有关大数据的书如此之多，何差你这一本?

这里，我不妨用哈佛大学著名统计学家孟晓犁（Xiao-Li Meng）教授的一句格言，聊以自慰：“你不需要先成为一名酿酒师，才能品酒。”① 因此拙著的名称恰恰就叫《品味大数据》。

在这本书中，我把很多学者、大家的观点汇集、梳理后呈现出来，他们的观点虽角度不同，却自成一家；观点之间甚至可能相左，在书中我并没定论观点的正确与否，而是交给读者来判断。犹如上了一桌菜，请君品尝，然后，你说哪道菜好，那它就好，是为书名取“品味”的寓意。

此外，在写作过程中，我也常用法国作家安德烈·纪德（Andre Gide）的话来给自己打气：“所有值得拿出来说的事情，早就已经被人说过了。但是，由于以前根本没有人在听，所以必须拿出来再说一遍。”

著名作家、曾经的优秀程序员王小波先生曾戏言②，写作其实就是个“减熵”的过程。熵（Entropy），代表的是信息的一种不确定性，一种未知的程度。对于王小波先生的写作而言，“减熵”过程，其实就是将不确定的、不靠谱的情节，尘埃落定，让它确定下来。这个过程并非易事，很可能是“出力不讨好”的。

王小波先生对写作的评论，让我想起我在本书第三章写下的两段话：

“今天之大数据，之所以再次吸引众人的眼球，就是因为当下的数据体积之庞大、种类之繁多、呈现之迅速，再次超过了当前秩序的容量，于是混沌重现。

“但大数据的价值之大，也吸引着人们不得不接纳这种‘混沌’。然而‘混沌无序’的大数据，是不能给我们创造价值的。因此，目前所有对大数据的研究，本质上都在干一件事，就是将这个无序的大数据时代，变得更加有序、可控、能为我所用。”

其实第一段话，概括起来，就是大数据给世人带来了“增熵”。第二段话核心思想就是，大数据价值很大，吸引世人为之折腰，为挖掘其价值，就得“减熵”。

由此看来，王小波先生所言的写作过程，其实和大数据的处理过程，在本质上有异曲同工之妙。

王小波先生的“文学创作”，天马行空，收放自如。科技（或科普）写作好像还难以做到这样。在笔者图书的创作过程中，有时感觉到写作（更确切地说，是学习）过程，其实还是个“增熵”的过程。这是因为，有时候，想把一件事情（或一个概念、一个技术细节）弄明白、写清楚，就得去查阅很多资料，结果查得越多，感觉自己不知道的就越多，“熵”的水平就上来了，信心随之降下去了。有时候，写作也非常摧残人的自信，阅读大量的文献，常常让自己感到渺小和无助。

在笔者心中，时常有两个小人在纠结斗争。一个富有阿Q精神的乐观小人骄傲地说：写吧，写吧，从CSDN发表的文章反响来看，还不错，一些已经面世的大数据图书，还不一定有你写得深入浅出呢！而另一个“横眉冷对”的悲观小人——小D则狠狠打击道：你这么说、这样写，

① 对应的英文原句是：You don't need to become a winemaker to become a wine connoisseur.
② 池建强 . MacTalk. 人生元编程 [M]. 北京：人民邮电出版社，2014.

难道不怕让那些大数据“大家们”贻笑大方吗?

感性地说，写作不仅是一个智力活，也是个自我调节的心理活，它更是个体力活，要花大量的时间、精力，投入其中。这本书的面世，花费了笔者一年有余的几乎所有休息时间，可谓累身累心。此刻,笔者算是更加深刻地体会到曹雪芹完成《红楼梦》后的自我评价:“满纸荒唐言，一把辛酸泪，都云作者痴，谁解其中味？”

在这本书里，笔者采用了 562 条注解（其中包括很多经典的论文、图书及网页资料），之所以这么做，就是为了确保文中的观点是靠谱的，是有据可查的。况且，参考文献的价值，有时大过著作（或论文）本身。因为这是，即使读者认为著作（或论文）本身不咋样，但通过参考文献的指引，相信读者也能快速找到更有价值、更高档次的文献材料。①

在这本书里，还配有 248 幅含信息量的插图。因为有时候，“一图胜千言”。书中没有复杂的公式，也没有难懂的代码。笔者尽量用通俗易懂的语言，告诉你大数据中比较晦涩难懂的概念和术语。

在这里，之所以用“562”和“248”这样量化的数字，无非是想告诉读者，这是一本很有诚意的、尽心尽力的大数据图书。

在这本书中，侃侃而谈的范围很广，小到街头巷尾的小故事、网络段子，大到《自然》《科学》高级别的学术论文。笔者会把有关大数据的一些（鼓励的、批评的及反思的）观点，加上一些注解，以图文并茂、通俗易懂的方式展现出来，力图让读者更加立体、充实地品读大数据。

本书的布局大致分为如下 4 块:第 1~3 章,主要漫谈了大数据有趣的历史,包括数据的启蒙、信息载体的演变和数据管理的发展脉络。第 4~6 章，主要聊聊大数据的内涵，包括大数据与哲学及第四科学范氏的关联。第 7~9 章是大数据的杂谈，包括大数据的用途所在、可能面临的陷阱以及通过小故事对大数据进行了一些反思。第 10~11 章主要涉及大数据的技术，包括 100 多篇大数据论文的漫读及 Hadoop 的实战篇。整体的脉络概括起来就是四个字——“顶天立地”:所谓“顶天”，是指我们先讲了一些“务虚”的大数据道理；而所谓“立地”，是指我们随后又聊了聊比较接地气的大数据技术。

两千多年前，孔夫子就曾说过，“学而时习之，不亦说乎？”这句话有着很多精彩的解读。费孝通先生认为②，“学”是和陌生事物的最初接触，“习”是陶炼，而“不亦说乎”描写的则是熟悉之后的亲密感觉。

笔者更喜欢杨伯峻先生在《论语译注》③中对这句话的解释:“学了，然后（按一定的时间）去实习它，不也高兴吗？”对于大数据的学习，也应是这样，大数据的道理我们懂了之后，还得创造条件去实践它。这也是本书的写作初心。

在学习大数据的路上，读完这本书，读者能从中得到什么呢？在回答这个问题之前，我们先重温一下 2012 年《哈佛商业周刊》刊登的一篇文章④，文章指出，数据科学家是 21 世纪最

---

① 《亲，你看懂这句话背后的含义了吗？》，在这本书里，有目前市面上大多数大数据畅销书的核心观点。在一番了解之后，如果想深入理解，再买他们书，这样省时省力省银子，多好!

② 费孝通 . 乡土中国 [M]. 北京 : 北京大学出版社，2012.

③ 杨伯峻 . 论语译注（简体字本）[M]. 北京 : 中华书局，2006.

④ Davenport T.H., Patil D.J. *Data scientist: the sexiest job of the 21st century*.[J]. Harvard Business Review, 2012, 90 (10): 70–76, 128.

性感的一个职业（Data Scientist：The sexiest Job of 21st Century）。在这个数据日益泛滥成灾的大环境下，对于这个论断，不管你信不信，反正我是信了。

数据科学家，其实就是采用科学方法、运用数据挖掘工具，在数据中寻找有价值的新洞察的那么一群人。

而想成为这么一群人，首先要具备大数据思维，认可大数据能带来大价值。我们知道，思维影响决策，很多时候，正确的思维能起到非常重要的战略引领作用。阿里巴巴董事局主席马云先生，就是一个非常励志的例子。作为一位曾经的大学英语教师，他本人并不懂什么具体的大数据技术，却能够非常成功地带领阿里巴巴走上大数据之路，并布局未来，要从 IT（Information Technology，信息技术）向 DT（Data Technology，数据技术）战略转型。这不仅仅是技术的升级，更是思维方式的巨大变革。而这本书的前半部分，正有助于读者培养这方面的思维。对于文科背景的读者，效果可能更为明显。

大数据思维有了，还得“知行合一”，把思维落实到行动上。这就需要有专门从事大数据技术的 DT 工程师。数据科学家的主体，其实就是这类人群。对于理工科背景的读者而言，这本书除了能辅助读者培养大数据思维，还是一个通俗易懂的大数据技术“综述”，特别是本书的最后两个章节，能为读者提供大数据技术全栈的“大图”（Big Picture），并能给读者带来一个比较感性的初级技术入门指引。

从笔者把这本小书定位为科技随笔，就可以看出，它的面世真真切切地源于它是“站在巨人的肩膀上”的。这里，笔者由衷地感谢很多前辈、大家以及网络资源提供者所作的贡献，没有你们的真知灼见，就没有这本书！

最后，说句很老套但很重要的话，由于本人的能力、学力及精力有限，书中疏忽甚至错误之处，在所难免，真心欢迎读者朋友的批评和指正。

注意：本书所采用的图片、表格及模型等素材，均为所属公司、网站或个人所有，在本书引用仅作为说明之用，绝无侵权之意，特此申明。

目录

Contents

Contents
>> 目录

Contents
>> 目录

>> 第一章

Chapter 1

大数据简史漫谈之一

——数字的来源及数据思维的发展

源头茫昧虽难觅，活水奔流喜不休。

——法国著名数学家、科学哲学家
昂利·彭加莱（Henri Poincare）

本章基本上是按照从古到今的时间轴线，漫谈数据的发展简史。[①] 在了解大数据的内涵之前，有必要简要地回顾一下大数据时代前的漫长历史。数据是人类认识客观世界的标度，数据与人类相伴的历史，可谓是源远流长。

著名社会学家费孝通先生曾说，[②] 人类的“当前”，包含着从“过去”历史中拔萃出来的投影和时间选择的积累。

翻开人类的科技史，我们很快就会发现，这就是一部人类对事物进行数据化的历史。在某个领域，越是能用数据来表征，其科学化的程度就越高，人类对其认识的程度也就越深。[③]

就数据的增长曲线而言，最初极小的初值，需要经历极其漫长的发展过程，才能达到人类能感知的曲线拐点。当下，“大数据”作为一个时髦的专业术语（buzz word），其历史还很短暂，但是它所依赖的基础在很久以前就建立了。[④] 人类的文明与进步，在某种意义上来说，就是通过对数据的收集、处理和总结而达成的。历史对于我们来说，并不是什么可有可无的点缀饰物，而是实用的、不可或缺的前行的基础。了解相关的历史，有助于培养我们的数据思维和基于数据的创新能力。

---

① 本章之所以说是漫谈，是因为笔者所介绍的一些历史，虽力图保证史实的正确性，但毕竟不是专业的科技史工作者，难免有不尽如人意之处。此外，既然是漫谈，内容也不见得十分扣题，信马由缰的地方也是有的，但这或许也是趣味之所在。

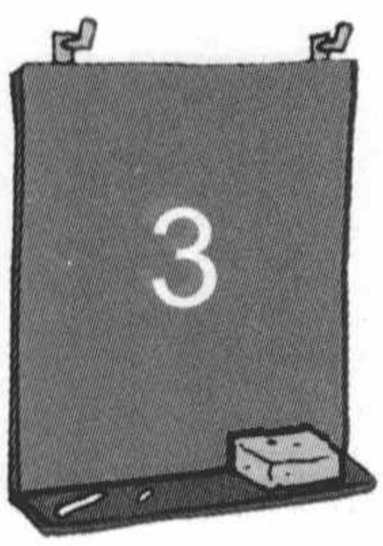
② 费孝通 . 乡土中国 [M]. 北京 : 北京大学出版社，2012.

③ 黄欣荣 . 大数据对科学认识论的发展 [J]. 自然辩证法研究，2014，（9）: 83–88.

④ Bernard Barr. *A Brief History of Big Data Everyone Should Read*. https://www.linkedin.com/pulse/brief–history–big–data–everyone–should–read–bernard–marr.

1.1 人类的“数觉”与计数系统

自从人类开始有文字和数字，数据就开始产生。数据作为一种计量工具与技术相融合，充分体现了其精确性和实用性的特征。人类文明的历程，大部分都可归属于小数据时代，甚至极小数据时代。



扫一扫，查看高清彩图

比一比谁说出的数字大，谁就赢

图 1-1　两个酋长比数数

美籍俄裔理论物理学家乔治·伽莫夫（George Gamow），在其著名科普著作《从一到无穷大》中，杜撰了这么一个小故事[①]（如图 1-1 所示）。

在非洲一个原始部族里，有两个酋长决定做一个数数游戏——比一比谁说出的数字大，谁就赢。

“好，”一个酋长说，“你先说吧！”

另一个酋长绞尽脑汁想了好几分钟，终于说出了他所能想到的最大数字：“3。”现在轮到另一个酋长动脑筋了。在苦思冥想半天后，他表示认输：“你赢啦！”

上面的小故事，其实是想说明，在远古时代，由于物质极其匮乏，人类对计数系统的认知，还处于懵懂状态。对少于 3 个的事物，人们尚能掌控，但对 3 个以上的事物，就只能称之为“很多”或“数不胜数”。在这种情况下，人类远古时代是很难出现完整的计数系统的。

人类文明的发展，存在严重的区域性不平衡。在澳大利亚的原始森林中，至今还有停滞于原始发展水平的部落。对数字的感知，普通人也就知道 1、2、3。即使是部落里的“聪明人”，也就只知道 4 和 5。数量再多，他们一概称之为“很多很多”。这是人类远古状态的无变异延续，可视作“活化石”。

> 数的概念，始于原始人采集、狩猎等生产活动，他们通过对不同类事物之间的比较，逐渐认识到事物存在某种共同的特征，然后从感性认识升华至抽象层面，于是就产生了数。

数从萌芽到诞生，经历了极其漫长的岁月。

在进化的蒙昧时期，人类已经具备一种才能，即在由同类事物组成的小样本集合中，当增加或者减少集合中的元素时，尽管我们的先祖还不能确切地知道增减多少，但仍能够

① 乔治·加莫夫 著 . 从一到无穷大 [M]. 暴永宁 译 . 北京：科学出版社，2014.

感知到其中有所变化。美籍数学家托拜厄斯·丹齐克（Tobias Dantzig）将这种能力称为“数觉”（number sense）[①]。所谓数（shù）觉，就是不通过数（shǔ）数（shù），一眼就能看出物体多寡的感觉。

这种原始的数觉，在某些动物身上也有体现。例如，有些鸟类就具有数觉，但也仅局限于小数量的“数觉”。有这么一侧试验，鸟巢里原有4个蛋，可以安然地拿走1个（余下3个），“笨鸟”不会察觉其中的变化，但如果拿去2个蛋（余下2个），那这只“笨鸟”可能就要“先飞”了——因为鸟巢中蛋的数量变化，已经触发了它的“数觉”——让它意识到危险，有外物“动了它的蛋”。这表明，有些鸟类，在用某种方法来辨别2和3是不同的。

丹齐克在其科普名作《数:科学的语言》中,提供了一个更有趣的例子(如图1–2所示)。

有一只乌鸦，在一个庄园主的望楼里筑巢，庄园主不胜其扰，决心打死这只乌鸦，他尝试了多次，都没有成功，因为人一旦靠近，乌鸦就非常警惕地离开巢穴，远远地待在树上，耐心地等人离开望楼后，再飞回巢穴。

有一天，园主心生一计：决定让2个人同时走进望楼，然后留一个人潜藏里面，另一个人出来并走开。但这个乌鸦并不上当，它还是等着，直到第2个人出来。

这个实验一连做了几天：2个人，3个人，4个人，都没有成功。最后，用了5个人，也像前几天一样，先一起进望楼，然后留一人潜藏其内，其他4个人走出来。这次奏效了，乌鸦的数觉“失灵”了——也就是说，当集合变大后，乌鸦已经无法辨别4与5的差别，因此它马上飞回巢里，被留在望楼的人逮个正着。

扫一扫，查看高清彩图

图1–2　乌鸦的数觉

① [美]托拜厄斯·丹齐克 著. 数：科学的语言——为有文化而非专攻数学的人写的评论性概述（*Number: The Language of Science—A critical survey written for the cultured non-mathematician*）[M]. 苏仲湘 译. 上海：上海教育出版社，1985.

“数觉”是动物的基本心理特征。丹齐克指出，“一种比鸟类高强不了多少的原始数觉，就是产生我们数概念的核心。毫无疑问，如果人类单凭这种直接的数觉，在计算的技术上，就不会比鸟类有什么进步。但是经历了一连串的特殊的环境，人类在极为有限的数觉之外，学会了另一种技巧来给自己帮忙并注定了使他们未来的生活受到巨大的影响，这技巧就是计数。并且，正是因为有了计数，人类赢得了用数来表达我们的宇宙的惊人成就”。

需要说明的是，数觉与计数不能混为一谈。数觉是人类早已有的能力，而计数能力的出现则要晚得多，这也可能是人类独有的能力。正是有了计数，才使得具体的、表现形式各异的、用于表达多寡的概念，结合成为统一的、抽象的“数”的概念。这是数学得以蓬勃发展的重要前提。

需求是发明之母。在需求的驱动下，人类首先发明了数字。数字是计数系统的基础。很多历史学家都认为，数字最初起源于对事物的计数，例如在人数、财产（牛羊数等）或交易中的计数。知名技术作家查尔斯·佩措尔德（Charles Petzold）在其著作《编码》[①] 一书中给出了一个非常生动的例子。

在远古时期，如果有人拥有 4 只鸭子，可以用图表示为图 1–3–a 所示。后来，专门负责画鸭子的人会“偷懒”地想：为什么我非要画 4 只鸭子呢，这太麻烦了！为什么不能就画一只鸭子，再用划线的多少来表示鸭子的数量呢，于是就出现了图 1–3–b 所示的简化画法。

类似地，我们还可以用这种简化画法用于画 4 头牛、4 只羊……诸如此类。慢慢地，这个数字“4”就被抽象出来了（图 1–3–c）。

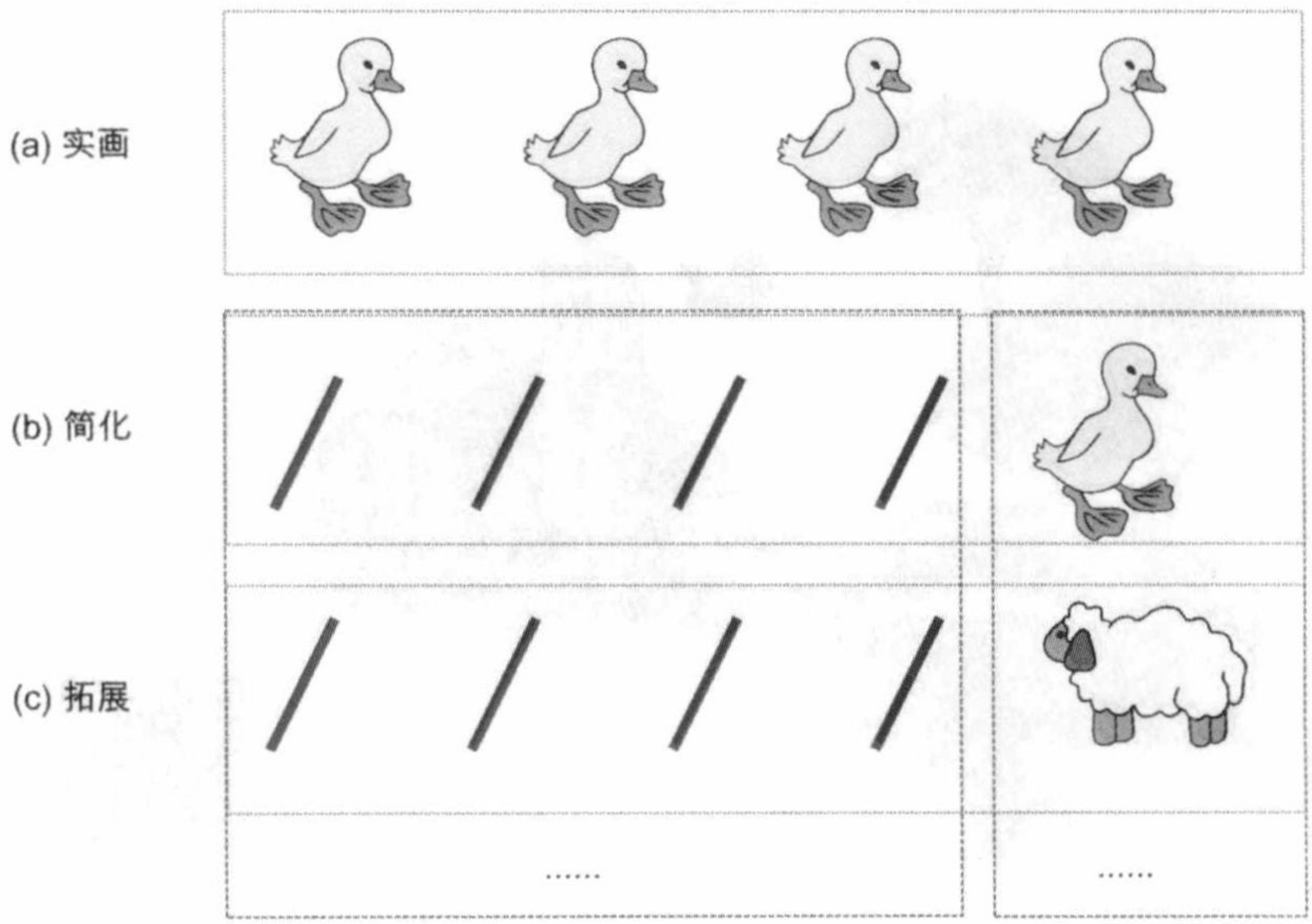


图 1–3　数字的抽象化

① 查尔斯·佩措尔德 著 . 编码：隐匿在计算机软硬件背后的语言 [M]. 左飞，薛佟佟 译 . 北京：电子工业出版社，2012.