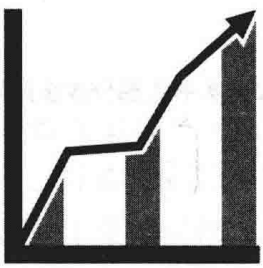
Broadview®
www.broadview.com.cn

从1开始
数据分析师成长之路

张旭东 著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn



从1开始
数据分析师成长之路

张旭东 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

数据分析行业就像所有新兴行业初期一样，伴随着混乱和盲目，一方面市场上培训机构巧立名目颁发证书，另一方面也有许多国外的著作被生搬硬套过来供自学者学习。本书是第一本结合国内公司实际状况和作者多年数据分析经验，系统而又详尽地介绍数据分析工作的作品。相较于使用 Excel 进行数据统计工作更加专业化、系统化，相较于数据挖掘与编程算法更加易于理解和贴合业务。从简单的制作报表开始和大家一起学习数据分析的五大模块：报表 BI 系统、异常数据分析、解决数据需求、项目性数据分析以及数据建模，为大家全方位、体系化地呈现数据分析到底是什么。

图书在版编目（CIP）数据

从 1 开始：数据分析师成长之路 / 张旭东著. —北京：电子工业出版社，2017.1
ISBN 978-7-121-30679-2

Ⅰ. ①从… Ⅱ. ①张… Ⅲ. ①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2016)第 311433 号

策划编辑：石　倩
责任编辑：石　倩
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：720×1000　1/16　印张：12.75　字数：204 千字
版　　次：2017 年 1 月第 1 版
印　　次：2017 年 1 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

序言

20 世纪 80 年代，伴随着微型智能计算机的发展，第三次工业革命进入了一个崭新的时代，计算机科学伴随着摩尔定律一路高歌猛进冲进了每个人的生活。从工业化时代转换到互联网时代一个最为突出的特征就是“信息爆炸”，近 30 年来人类生产的信息已超过过去 5000 年信息生产的总和。而当下信息的主要载体是数据库，庞大的信息量对应着庞大的数据量，那么这些承载着庞大信息量的数据处理就显得尤为重要，数据分析作为一门新兴的行业也变得越来越受人瞩目。

就像在计算机行业刚刚火爆的那些年，由于没有现成的体系化的知识，几乎所有人都在摸索中前进。大家的知识一方面来源于相互探讨交流，另一方面借鉴西方发达国家的教材资料。数据分析现在同样没有体系化的知识结构，没有成熟的经验教训，数据分析从业者中一部分是从计算机编程开始做数据挖掘，另一部分是从统计学开始做数据分析，还有一小部分人是凭借着自己的兴趣爱好自己探索着前进。国内对于数据分析的解读一方面偏向于基于 Excel 可视化报表，另一方面偏向于数据挖掘与编程算法，前者太过流于表面，后者又十分晦涩难懂。张旭东的这本《从 1 开始——数据分析师成长之路》算是国内第一本详尽而又系统的介绍数据分析前因后果的书籍了，在保证通俗易懂的同时又有数据分析的深度，作为数据分析的入门书籍的确是相当不错。

数据分析行业一定会伴随着大数据时代的到来逐渐被大家重视和认可，如果你想把握住机会成为大数据时代的弄潮儿，这本书值得一看。

卢斌

中国人民大学高礼研究院执行院长

前言

随着大数据这个概念被越来越多的人提起，数据分析与数据挖掘这两个词汇频繁地出现在人们的视野中，越来越得到大家的重视和青睐。从事数据分析工作的这些年，身边不断有人问起数据分析如何入门或是如何做好数据分析，市场也有各类“速成数据分析”或是“零基础数据分析”等培训课程，颇有当年人人都去做产品经理的势头。与此同时在一些问答类网站上出现了许多诸如这样的问题：

“文科生如何转行数据分析？”

“数学基础不好能做数据分析吗？”

“听了某某专家的演讲觉得数据分析很棒，如何入门？”

……

问题下面往往有很多因各种各样的原因推荐的书籍、教程、公众号……内容乏善可陈的同时太容易误导新人，看着着实心痛。

与此同时，通过这些年来的了解和熟悉，身边有太多“盲目”的数据分析从业人员，只是了解了 Excel 中相关图表与统计的功能，在从事分析工作时也有许多的不严谨和漏洞。在一些社区或是平台经常遇到一些人把原始数据直接挂在网上，问该怎么分析数据甚至是通过这些数据能得出什么结论。现在想一想，他们真的适合做数据分析吗？数据保密性的职业素养不说，不经大脑思考地贴数据要

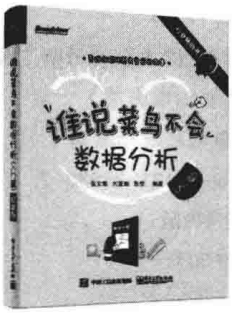
结果的分析员真的能胜任这份工作吗？

写这本书最大的愿望就是能够通过简单的描述让大家对数据分析有一个简单的了解，对自己是否适合这个职位有一个概念，不要盲目从众，能有自己的判断。市场上从零开始入门的教程鱼龙混杂，在入门之前大家首先要考虑这扇门真的适合你吗？

这本书写在数据分析入门之前，会向读者们简单地介绍究竟什么是数据分析，重点放在这个岗位有怎样的要求和特质以及如何才能达到这样的标准，也会简单介绍数据分析岗位未来的职业发展，希望对有志于从事数据分析工作的你有所帮助。

作　者

电子工业出版社好书分享

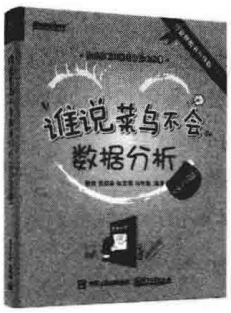


《谁说菜鸟不会数据分析（入门篇）（纪念版）（全彩）》

张文霖 刘夏 璐狄松 编著

ISBN 978-7-121-28798-5
2016年5月出版
定价：59.00元

EXCEL数据分析就像一本故事书，让你的工作更出彩（数据分析入门必读）



《谁说菜鸟不会数据分析（SPSS篇）（全彩）》

狄松祝 迎春 张文霖 马世澎 编著

ISBN 978-7-121-28801-2
2016年5月出版
定价：59.00元

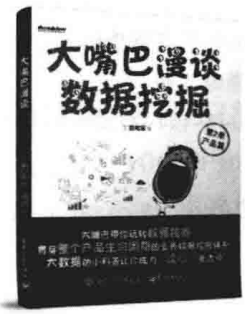
像EXCEL一样简单，一看就懂的SPSS数据分析实战（数据分析进阶必读）

《大数据时代小数据分析》

屈泽中 编著

ISBN 978-7-121-26469-6
2015年7月出版
定价：69.00元

能解决实际问题的数据分析书，大数据时代产品经理、市场营销、管理者等数据分析学习必读

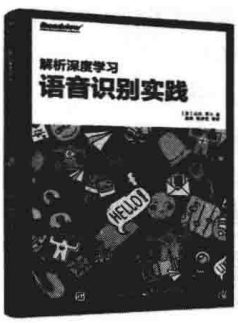


《大嘴巴漫谈数据挖掘（第2季产品篇）》

易向军 著

ISBN 978-7-121-28237-9
2016年3月出版
定价：66.00元

◎贯穿整个产品生命周期的业务数据挖掘体系，大数据的小科普让你成为“读心”魔法师！



《解析深度学习：语音识别实践》

【美】俞栋 邓力 著

俞凯 钱彦旻 译

ISBN 978-7-121-28796-1
2016年6月出版
定价：79.00元

◎了解深度学习应用实践不可错过的经典专著

《神经网络与深度学习》

吴岸城 著

ISBN 978-7-121-28869-2
2016年6月出版
定价：59.00元

◎ 从零起步了解神经网络与深度学习，AlphaGo大胜李世石的背后玄机。

电子工业出版社好书分享

《智能大数据SMART准则：数据分析方法、案例和行动纲领》

[美] Bernard Marr 著

秦磊 曹正凤 译

ISBN 978-7-121-27058-1

2015年9月出版

定价：49.00元

◎ 大数据是智能革命的核心！

◎ 接地气的案例、方法、行动纲领，适合大数据的实践者、管理者

《触手可及的大数据分析工具——Tableau案例集》

沈浩 王涛 韩朝阳 李健 著

ISBN 978-7-121-26938-7

2015年9月出版

定价：79.00元

◎ 国内首本Tableau著作，最新版本Tableau 9.0

◎ 实战经验分享，精选28个案例，涉及15个行业

《大数据的互联网思维》

段云峰 秦晓飞 著

ISBN 978-7-121-27308-7

2015年10月出版

定价：58.00元

◎ 作者十多年大数据相关从业经验的积累

◎ 累计投入 100 多亿元买来的大数据产品及运营的实战经验

《电商数据分析，淘宝实战》

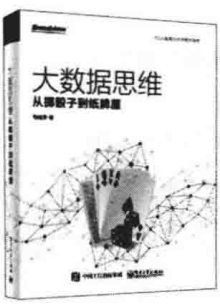
零一 著

ISBN 978-7-121-28117-4

2016年2月出版

定价 59.00元

◎解密数据背后的真相，给网店运营、淘宝店长、数据分析者的书



《大数据思维——从掷骰子到纸牌屋》

马继华 著

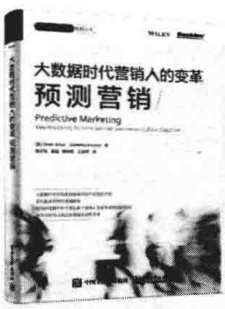
ISBN 978-7-121-29407-5

2016年8月出版

定价：55.00元

编辑推荐：

读者不需要任何统计学知识，也没必要掌握复杂的公式与算法，在最通俗易懂的案例介绍和娓娓道来中就可以轻松理解大数据分析的基本模式与方法。



《大数据时代营销人的变革：预测营销》

[美] ÖmerArtun

Dominique Levin 著

曹正凤 秦磊 谢邦昌 王淑燕 译

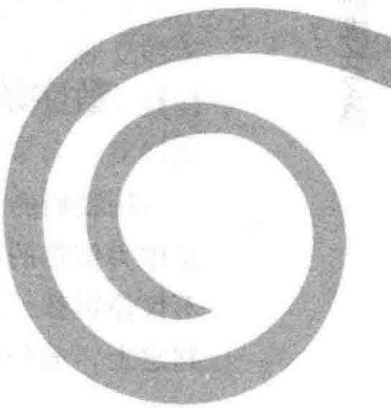
ISBN 978-7-121-29125-8

2016年6月出版

定价：59.00元

◎ 一套数据驱动的营销框架，适合实战营销人、大数据分析师阅读

目录



第1章 数字、数据、数学

数据作为数据分析的目标与对象无时无刻不充斥在我们的生活中，用数据说话或是数据为王的准则被人们时时刻刻地挂在嘴边，那么数字、数据与数学到底是怎样的存在呢？

1.1 数字的起源

从我们的祖先发明结绳计数以来，数字和文字、自然语言就一起作为信息的载体融入生活之中了。数字让人们对日常生活的认识不仅停留在“多”或是“少”，“够”或是“不够”这一简单的逻辑判定上，而是让人们开始有了量化的认识。

在数字还未出现的原始社会，原始人类的生存活动主要停留在觅食中，努力生存下去，“是否”有充足的实物以及“是否”有安全的庇护这些“True or False”的逻辑判断几乎是日常生活的全部内容，人们不需要数字。就好像人们在满足基本生存需求之前不会思考“我从哪里来？要到哪里去？”的哲学问题，原始人类在生存渴望的驱动下面临的问题只是是否有食物？是否有住所？是否存在危险？……

然而随着旧石器时代的到来，祖先开始使用工具来狩猎、种植、生产，这些生产条件的改善使得原始人类的猎物出现了剩余，原先能否生存下去的问题逐渐淡出视线，人们开始思考如何对剩余财产进行储存与分配，数字在这个时候也就应运而生了。数字给了财产以量化的准则和分配的标准，每个人应该分配到多少就有了标准，仓库存储剩余就有了准则，以这样的形式，数字作为承载财产的量化信息出现在人们的视野中。

数字作为一种符号，现存的在使用的就有好多种，诸如阿拉伯数字、罗马数字、中文数字等不同的表现方式，也有诸如二进制、十进制、十六进制等不同的计量方式。但是不管以怎样的形式，数字都是一种符号，如果抛去数字所承载的信息它只能算是一种工具。

只有当数字承载着计量财产信息，计算着天文历法、农忙耕种，估算着投入与产出的比重时，数字才具有价值。

数字不仅是一种符号，数字是一种规范的符号。在我们对数字制定诸如加减乘除这样的数学规则之前数字像文字一样是独立的符号。如果我们用 1、2、3、4 对应吃、穿、住、行，当我们饿了我们就说“1”——饭，当我们冷了我们就需要“2”——穿，当我们需要睡觉就说“3”——住，当我们要走路就说“4”——行（图 1-1）。

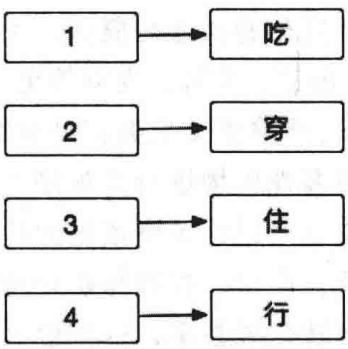


图 1-1

这个时候数字作为一种符号与其他符号相比是不具有特殊性的，但是当我们用数字来量化多少时就把数字赋予了单调性，这里的单调性指的是

$$1<2<3<4<5<6<7\cdots\cdots$$

但是我们不能说

吃>穿>住>行

吃、穿、住、行这样几个符号就不具有单调性。同时我们再赋予数字以加法和减法

$$1+2=3$$

$$4-1=3$$

通过这样的方式，我们的量化符号不再是静态的表示多与少，

它还能表示动态的变化。今年的收成比去年少了，少了多少呢？这就是一个减法的过程，数字对这样一个过程进行了量化和描述，数字的意义就更加丰富了。

1.2 数据

什么是数据呢？

数据是对信息一种量化的描述与概括，不仅包含数字所承载的信息，还包含了文字、图片、声音、视频等更多形式所承载的信息，这些都可以统称为数据。严格意义上数据是对客观事物的逻辑归纳，用符号、字母等方式对客观事物进行直观描述。数据是进行各种统计、计算、科学研究或技术设计等所依据的数值，是表达知识的字符的集合。文字、图片、声音、视频等媒介承载的信息可以通过技术手段进行量化，进而转化成数字，这些数字承载了媒介所传达的信息，构成了数据的一部分。

数据作为信息的载体，承载着信息的内容；信息通过数据来表现，让信息变得易识别（图 1-2）。

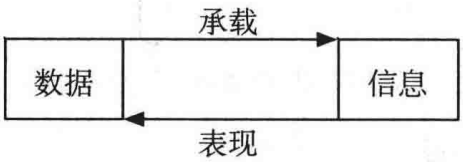


图 1-2

数字、文字、声音、视频、气味等所有承载的信息的事物都可以理解为数据（图 1-3）。

所以说数据相对于数字是一个更广阔的概念，一切生产活动产生的信息都可以被称之为数据。

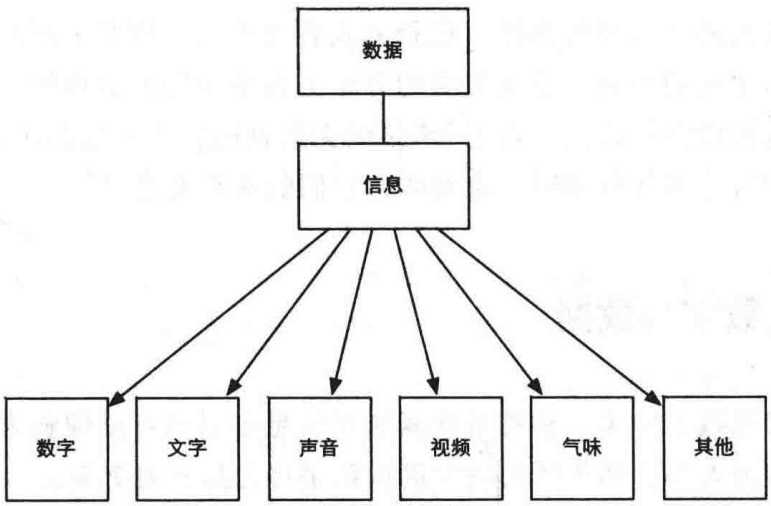


图 1-3

- 按照数据的性质来分，数据可以分为：

a．定位的，如坐标类的数据；

b．定性的，如表示事物属性的数据（居民地、河流、道路等）；

c．定量的，反映事物数量特征的数据，如长度、面积、体积等几何量或重量、速度等物理量；

d．定时的，反映事物时间特性的数据，如年、月、日、时、分、秒等。

- 按照数据的表现形式可以分为：

a．数字数据，如各种统计或量测数据。数字数据在某个区间内是离散的值；

b．模拟数据，由连续函数组成，是指在某个区间连续变化的物理量，又可以分为图形数据（如点、线、面）、符号数据、文字数据和图像数据等，如声音的大小和温度的变化等。

- 按照记录方式又可以分为：

地图、表格、影像、磁带、纸带等。