BIG DATA strategy

一个**支点**，可以让**阿基米德**撬动地球
一组**大数据**，可以**让任何人**改变世界

大数据战略

个人、企业、政府的思维革命与红利洼地

徐端◎著

搜狗输入法里有什么商机
怎样的商品会在**淘宝里热卖**
警察如何利用数据预知犯罪
……

大数据战略
让93%的事情可以预测
让100%的钱变得好赚
无论是**个人、企业、政府**
都能通过它**获得巨大的提升**

新世界出版社
NEW WORLD PRESS

BIG DATA
strategy

大数据战略

个人、企业、政府的思维革命与红利洼地

徐端◎著

新世界出版社
NEW WORLD PRESS

图书在版编目（CIP）数据

大数据战略：个人、企业、政府的思维革命与红利洼地 / 徐端著.
--北京：新世界出版社，2014.3
ISBN 978-7-5104-4914-7

Ⅰ. ①大… Ⅱ. ①徐… Ⅲ. ①数据处理-研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2014）第025120号

大数据战略：个人、企业、政府的思维革命与红利洼地

作　　者：徐　端
责任编辑：黄晓林　周　璐
责任印制：李一鸣　黄厚清
出版发行：新世界出版社
社　　址：北京市西城区百万庄大街24号（100037）
发行部：（010）6899 5968　（010）6899 8733（传真）
总编室：（010）6899 5424　（010）6832 6679（传真）
http://www.nwp.cn
http://www.newworld-press.com
版权部：+86 10 6899 6306
版权部电子信箱：frank@nwp.com.cn
印刷：三河市骏杰印刷厂
经销：新华书店
开本：710×1000　1/16
字数：180千字　**印张**：12.75
版次：2014年4月第1版　2014年4月第1次印刷
书号：ISBN 978-7-5104-4914-7
定价：29.80元

序 言

2013年4月15日，波士顿马拉松比赛现场发生爆炸案。仅仅几小时内，数以千计的在场群众就通过手机、相机、平板电脑等设备将在事发现场拍摄的照片和视频放到了网上。3天后，犯罪嫌疑人便已被确认。又一天后，两名犯罪嫌疑人和警方爆发枪战，结果一死一伤。

这是历史上第一次反犯罪机构的专业能力与社会大众汇集的大数据结合起来，在与犯罪分子的时间竞赛中取胜。

在很多人还不知道“大数据”这个词的含义时，大数据时代已经悄然到来。

随着社交网络的逐渐成熟，移动带宽迅速提升，云计算、物联网应用更加丰富。更多的传感设备、移动终端接入网络，由此产生的数据及数据增长速度迅速攀升。

一项针对甲骨文公司独立用户的调查发现，90%的企业的数据量在迅速上涨，其中16%的企业的数据量每年增长率达到50%或更高。不少企业已经感受到失控的数据增长对绩效造成的冲击，其中87%的受访者将企业的应用程序性能下降归咎于不断增长的数据量。另一项调查则显示，全球数据量在2011年已达到1.8ZB，在5年里增加了5倍。

1.8ZB是什么样的概念呢？如果把所有这些数据都刻录存入普通DVD光盘里，光盘的高度将等同于从地球到月球的一个半来回，也就是大约72万英里。这相当于每位美国人每分钟写3条推特，而且还要不停地写2.6976万年。是不是很恐怖？这还不是最恐怖的，一个权威调查机构还预测全球数据量大约每两年翻一番，2015年全球数据量将达到近8ZB，到2020年，全球将达到35ZB。

这个数据不可谓不大！然而，大数据的“大”不仅仅在于数据量的庞大，还有其他的特征。某项技术要想成为大数据技术，必须满足IBM所描述的3个条件，即多样性、大容量和时效性高。

从20世纪70年代末期开始，已经实现工业化的发达国家先后开始了向信息化社会转型的过程。站在今天的角度观察，这一由工业化向信息化的转型可以分为3个时代，即计算机时代、互联网时代和大数据时代。到90年代中期，美国已经基本度过了计算机时代，计算机高度普及，解决了信息的机器可读化和数据的可计算化问题。在21世纪初，美国也基本走完了互联网时代的路程，互联网高度普及，解决了信息传递和信息服务问题。在计算机和互联网的基础上，美国正在步入一个全新的历史阶段——大数据时代。

从最早的结绳记事到后来的问卷调查，从早期巨型计算机作为唯一的电子化数据获取和处理工具到后来PC的普及，再到今天的智能手机、谷歌眼镜和穿戴型数据终端以及形形色色的数据传感装置，人类将物理界、生物界和社会界的万事万物数据化并加以存储处理的能力大幅提高，可以说无处不在，无物不读。截至2013年5月，全球具备数据获取存储处理和传输的数据终端设备已经超过100亿台，并且以每两年翻番的速度增长。互联网从早期的有线网络发展出无线网络，数据传输速度越来越快，数据传输成本越来越低。

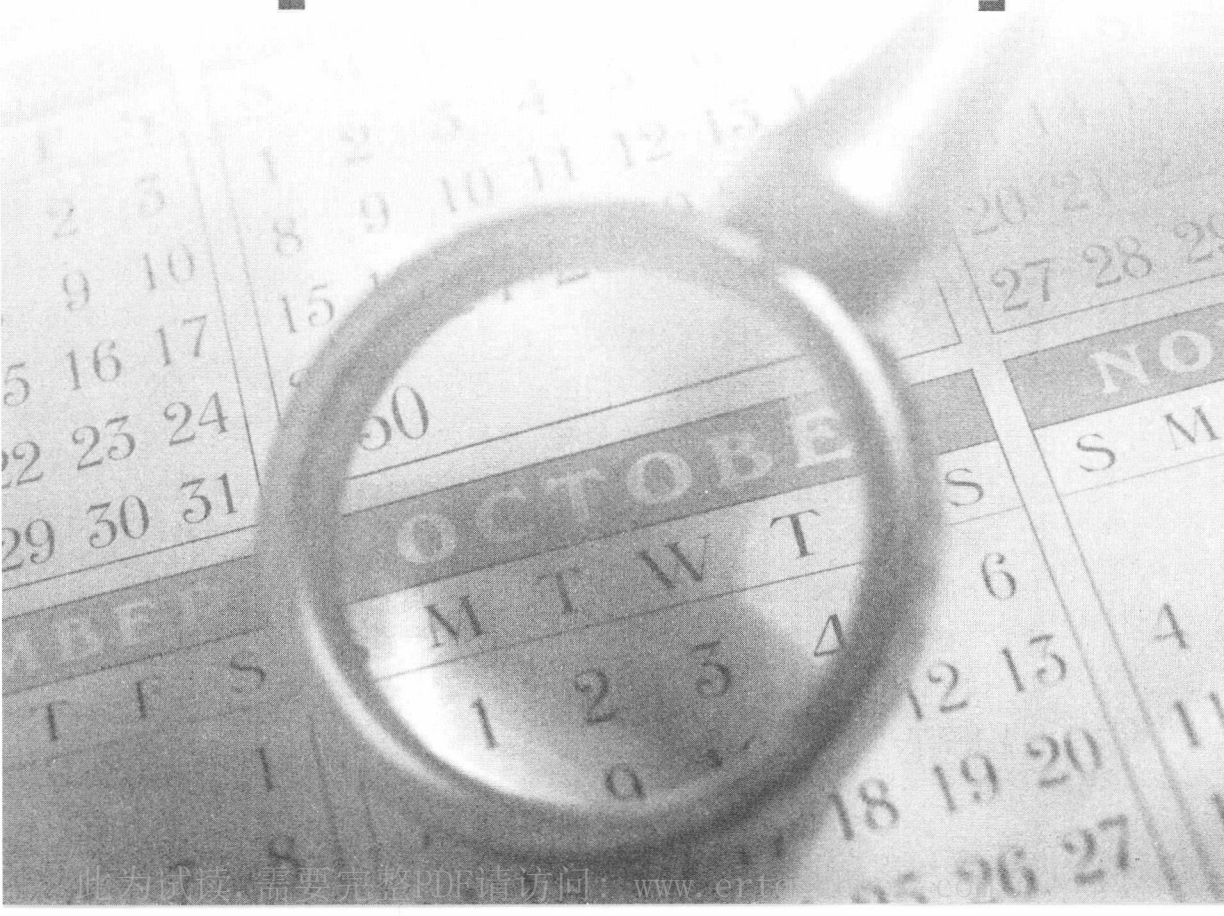
当互联网与数据终端合为一体，就开始形成一个全面深入映射现实世界的数据化世界，也就是人们所谓的大数据。获取和利用大数据，寻找过去现实世界中所没有的全新生活方式、社会治理机制和经济发展途径，开始成为社会方方面面关注的焦点，这就是人们所谓的大数据时代。当获取和利用大数据成为社会共识和社会发展的主要推动力的时刻到来，可以说人类全面进入了信息化社会。

欢迎来到这个新时代，这个崭新的、神奇的大数据时代。

目 录

第一章　从小数据到大数据

过去很长时间，我们习惯采用问卷调查、现场采访等方式去采集一些有代表性的数据，借以分析我们要解决的问题。这些方法在过去取得了很不错的效果。只是，随着科技的迅猛发展，我们对数据处理和分析的要求越来越高，这些方法慢慢变得不太适合了。此时，大数据应运而生，逐渐展示出惊人的力量。



一、大数据的过去

给你一家超市

如果现在你是一家超市的经营者，你会怎么让超市的利益最大化呢？

随便想想都有很多办法，如降低进货成本、降低物流成本、精简不必要的人员、优化销售团队、策划必要的营销活动、分析其他超市的策略、分析消费者需求……这些举措中，有的可以直接降低成本、提高利润率，有的则间接地提高销售数据。而其中最重要的，肯定是分析消费者需求。那么，你会怎么分析消费者的需求？

先看一个销售的例子吧。

一位老太太走到路边的水果摊，想买一些杏子，她先到了第一个水果摊。

老太太上前，问摊主："你这个杏子怎么样？"摊主热情地说："我的杏子又大又甜，保证好吃啊。"老太太想了想，摇摇头走了。摊主一脸失望，不知道为什么老太太没买他的杏子。

老太太走到了第二个水果摊，问道："你这个杏子怎么样？"这位摊主也很热情地回答说："都挺好的啊，您想要什么样的？"老太太回答："我想要那种比较酸的。"摊主很诧异，酸杏子谁吃啊。他眼珠一转，笑眯眯地说："大妈，您这是故意套我话的吧？我这儿绝对都是个赛个的甜，保证您买了不吃亏。"老太太回答说："你这儿真没酸的？"摊主信心十足地说："有一个酸的我赔您一筐！"老太太摇摇头，叹了口气，又往前走了。自己这么保证了老太太还是没买，摊主别提多失落了。

老太太走到了第三个水果摊。摊主一见到老太太便上前问道："您想买点什么？"老太太说："我想买点酸的杏子。"摊主说："我这儿倒是有酸杏子，可是我觉得您这样年纪的人，吃得太酸了不太好呢。您保重身体啊！您要不搭配着买点儿别的，比如香蕉什么的。"老太太高兴地答复道："不是我要吃，是我儿媳妇要吃。"摊主又问："您儿媳妇要吃酸杏子啊，您这是要抱孙子了吧？"老太太高兴地说："是啊是啊，她刚怀孕没多久，就想吃点酸的。"摊主笑着回答说："酸儿辣女，您肯定能抱个大胖孙子！我这还有猕猴桃，含各种维生素，特别适合孕妇吃。您要不也来点？"就这么一句又一句，老太太特别开心，最后老太太买了摊主推荐的很多水果走了。

在这个例子里，很明显，第三个摊主是最成功的。他的成功在哪里呢？在于他问清楚了老太太的需求。在他与老太太的对话里，他获得了几个信息：老太太需要酸杏子；酸杏子是给儿媳妇吃的；儿媳妇怀孕了。由这几个信息，他便能从营养搭配等角度去推荐他的产品（水果），既满足了消费者本身的需求（酸杏子），又挖掘出消费者潜在的其他需求（各种营养）。

这样的场景，我们在生活中可能经常遇到。可是在大型超市，商家并不是一对一地跟消费者沟通。更多的情景是，商家把物品放到货架上供消费者自行选择。在这种情况下，怎么能够知道消费者潜在的需求呢？

在一家超市中，人们发现了一个特别有趣的现象：尿布与啤酒这两种风马牛不相及的商品居然被摆在了一起。但令人不解的是，这一奇怪的举措居然使尿布和啤酒的销量大幅增加了。这可不是一个笑话，而是一直被众多商家所津津乐道的发生在美国沃尔玛连锁超市的真实案例。实际上，这不是美国人的幽默细胞所致，而是数据的魔力。这个发现为沃尔玛带来了大量的利润，但沃尔玛是如何从多如牛毛却又杂乱无章的数据中发现啤酒和尿布销售之间的联系的呢？这又给了我们什么样的启示？

沃尔玛的商品种类非常多，它有一套非常复杂的方法对所有商品的销售情况进行统计。沃尔玛通过对每件商品每天的销售数据统计发现，每到周末啤酒和尿布的销量就异样的好，这两者之间似乎有什么关联。但是，沃尔玛并没有去找这两个销售数据之间的联系，而是立即做出决定，将这两样商品摆放在一起，结果这两样商品销售量都大幅增加。显然，这个决定是正确的。那么原因是什么呢?

有人分析称，因为在美国，周末电视台一般会转播球赛，而看球赛的大部分是男人。男人们在家看球赛的时候都会拿上一罐啤酒，受到冷遇的妻子会出门逛街或和闺密小聚，照料小宝宝的重担就留给了留守的丈夫。就这样，沃尔玛把婴儿尿布放在啤酒销售区旁，男人往往会在买啤酒的时候顺手拿起尿布。

也有人说，是因为在美国家庭里，一般都是丈夫挣钱养家，妻子照顾孩子。忙于照顾孩子的妻子经常会嘱咐丈夫在下班回家的路上为孩子买尿布，而丈夫在买尿布的同时又会顺手购买自己爱喝的啤酒。

这两个原因都说得通。那么，真正的原因是什么呢?

你是不是开始思考这个问题了？打住！别忘了，我们假设的是你是一个超市的经营者，你要解决的是让超市利益最大化，达到这个目的就行了，你不是研究这些现象的科研人员，没有必要去搞清楚这些问题后面的复杂原因。如果你有一辆汽车，你更需要学习的是驾驶而不是汽车制造及修理。同样，如果你有足够多的数据并分析出了结果，你需要做的是利用结果去提高盈利而不是搞清楚结果背后的原因。

作为一名超市的管理人员，你肯定会对沃尔玛如何统计分析各类销售数据感兴趣。可是，数据到底是什么呢？我们不妨回顾一下数据的历史。

十九头牛的难题

数据是什么？

一年有365天；真空中的光速是299792458米每秒；正常人心跳每分钟大约75次（60～100次）；2013年11月15日国内汽、柴油标准品最高供应价格每吨分别为8715元和7890元；2012年度北京市职工月平均工资为5223元，比上年增长11.8%……

可以说，我们的生活里到处都是数据。

数据是对客观事物的符号化的表示，是未经加工的、用于表示客观事物的原始素材，如图形符号、数字、字母等。换句话说，数据是通过物理观察得来的事实和概念，是关于物理世界中的地方、事件、其他对象或概念的描述。在计算机科学里，数据被定义为所有能输入到计算机并被计算机程序处理的符号的介质的总称。

数据具有数值属性、物理属性，这一点和数字是不同的。很多人会把数据和数字混为一谈，其实，可以这么说，数字是一种没有物理属性的数据。

比如，1+1=？是数字计算，结果是2，这个是没有问题的。如果我们加入物理属性，1个土豆+1头牛=？由于土豆和牛的物理属性不同，我们没法求出它们的和，总不能说答案是土豆烧牛肉吧？

在计算机问世之前，人们处理的数据一般都是有关数字的数学问题，比如家喻户晓的分牛问题便是一个很经典的例子。

一位老人养了19头牛。临终前，他对3个儿子立下遗嘱：“家中有19头牛，老大可以分1/2，老二可以分得1/4，老三则只能分到1/5。牛不得杀死分肉，不得卖钱后分钱。”说完老人便去世了。3个儿子犯愁了，19头牛怎么分1/2、1/4、1/5啊？每个人都想多分一点儿，每个人又不肯吃

一点儿亏，于是争吵了起来。

一位智者想到了办法，他笑眯眯地对老人的3个儿子说："我有办法。"然后他把自己家的一头牛牵来，和19头牛放到一起，又对他们说："现在这里有20头牛，老大分1/2，也就是10头；老二分1/4，也就是5头；老三分1/5，也就是4头。剩下还有一头是我牵来的，我牵回去好了。"3个儿子终于解决了这个问题，喜笑颜开，重归于好。

$$20\times\frac{1}{2}=10$$
$$20\times\frac{1}{4}=5$$
$$20\times\frac{1}{5}=10$$
$$10+5+4=19$$

分牛问题

这是一个很小的有关数据的故事，日常生活中，我们经常会遇到各种数据。一般来说，我们都是通过数学来解决这些问题的：五险一金的计算问题；话费套餐的计算问题；银行利息的计算问题……我们每天都在和各式各样的数据打交道，也许我们对此已经习以为常、熟视无睹。

最开始，我们的生活里都是很小的数据：部落里20头猎物如何分给50个人；采集的200颗浆果一半给部落首领家族后其他人怎么分；两个部落间的土地如何平分……这一类的问题，随着人类数学水平的提高，慢慢地得到了解决。同时，人们也遇到了越来越棘手的问题：一个人每周买一注彩票，20年内中500万的概率有多大；一个人父母都是A型血，孩子是O型血的可能性有多大；一块完全不规则的土地，如何划分成5等份，等等。中国古代也有一些很经典的数学题：

1. 八万三千短竹竿，将来要把笔头安，管三套五为期定，问君多少能完成？

用现代的话说就是：有83000根短竹竿，每根短竹竿可制成3个笔管或者5个笔套。怎样安排制笔管和制笔套的短竹的数量，使制成的笔管和笔套正好数量匹配。

2. 有井不知深，先将绳三折入井，井外绳长四尺，后将绳四折入井，井外绳长一尺。问井和绳长各几何？

3. 今有门厅一座，不知门广高低，长杆横进使归室，无奈门狭四尺，随即竖杆过去，也长二尺无疑，对角斜进恰好齐。请问高宽各几？

4. 100个大人和小孩共吃100个馒头，已知大人每人吃3个，小孩3人合吃一个。大人和小孩各有多少？

5. 今有蒲生一日，长三尺；莞生一日，长一尺。蒲生日自半，莞生日自倍。问几何日而长等？

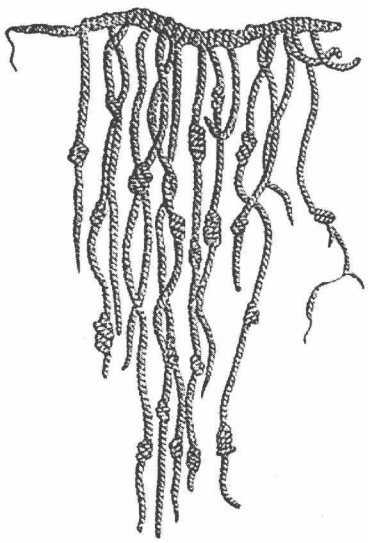
6. 甲赶群羊逐草茂，乙拽肥羊一只随其后；戏问甲及一百否？甲云所说无差谬，若得这般一群凑，再添半群小半群，得你一只来方凑，玄机奥妙谁猜透？

7. 远望巍巍塔七层，红光点点倍加增。共灯三百八十一，请问各层几盏灯？

可以说，这些数学题很多都是古人遇到的各种问题的再现。

我们不仅发明了数字用以记录储存数据，还发明了不同的计量单位、不同的进制。一分钟有60秒，一天有24小时，一秒等于100毫秒，等等，在不同的领域里，二进制、八进制、十进制、十六进制、六十进制等发挥着不同的作用。

在十进制的世界里，人们用以记录数据的数字符号有10个，分别是从0到9，数数的方式是0、1、2、3、4、5、6、7、8、9、10……而在计算机里使用的是二进制，记录数据的符号只有0和1，数数的方式是0、1、10、11、100、101、110、111、1000……再比如，中国有个成语叫作“半斤八两”，用以表示旗鼓相当，水平差不多，这是因为中国古代的秤采用的是十六进制，一斤等于十六两。半斤和八两，确实是旗鼓相当的。



印加帝国奇普

从古至今，数学一直伴随着数据处理的问题发展着。

数学，起源于人类早期的生产活动，为中国古代六艺之一，亦被古希腊学者视为哲学的起点。史前的人类除了学会以数字统计物品的数量