数字原来会说谎

对很多平台来说，
或许这是一个流量
比事实更重要的时代

数字图表？
有图不代表有真相！

后真相时代与谎言媒体！
是谎言媒体导致我们进入了后真相时代？
还是后真相时代孕育出了媒体谎言？

扣小米——著

化学工业出版社

数字原来会说谎

扣小米——著

化学工业出版社
·北 京·

在大数据时代，数字被看作是巨大的金矿，变得前所未有地重要。人们可以通过一串串数字刻画整个世界，甚至预测未来。但是数字却永远无法代替真实，现在数字和数据被滥用的现象越来越常见，特别是新技术的运用更是使数据从收集到处理，从可视化到信息表达，每个环节都存在用数字做手脚的机会，让人防不胜防。不过数字永远都是那些数字，说谎的并不是数字本身，而是使用数字的人，是数字使用者把数字变成了“任人打扮的小姑娘”。

本书将用简单易懂的语言分析常见的利用数字说谎的情况，并结合一些常见的例子，对现有的一些“数字陷阱”现象进行解析。

图书在版编目（CIP）数据

数字原来会说谎 / 扣小米著. — 北京：化学工业出版社，2017.11

ISBN 978-7-122-30722-4

Ⅰ. ①数… Ⅱ. ①扣… Ⅲ. ①经济统计－统计数据－普及读物 Ⅳ. ①F222-49

中国版本图书馆 CIP 数据核字（2017）第 243709 号

责任编辑：罗 琨　　装帧设计：韩 飞

责任校对：王素芹

出版发行：化学工业出版社（北京市东城区青年湖南街13号 邮政编码100011）

印　　装：三河市双峰印刷装订有限公司

710mm×1000mm 1/16 印张13½ 字数169千字 2018年2月北京第1版第1次印刷

购书咨询：010-64518888（传真：010-64519686） 售后服务：010-64518899

网　　址：http://www.cip.com.cn

凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：39.80元

前言

谎言有三种：谎言、该死的谎言、统计数字

——本杰明·迪斯雷利

数字是我们日常生活中每天都会碰到的符号，是一种全世界每个国家都通行的语言。以数字为基础的数学更是所有自然科学的基础。有了数学的表达，一门科学才算是能够精确推导、理论致臻完善。

但是数字在改变世界的同时，也蕴藏着巨大的陷阱。美国大文豪马克·吐温曾经在《我的自传》一书中写道："数字经常欺骗我，特别是我自己整理它们时。针对这一情况，本杰明·迪斯雷利的说法十分准确：'世界上有三种谎言：谎言、该死的谎言、统计数字。'"（There are three kinds of lies: lies, damned lies, and statistics.）

1955年，美国作家达莱尔·哈夫（Darrell Huff）出版了《统计数字会撒谎》（*How to Lie with Statistics*）一书，该书用大量生动有趣的实例揭露了当时美国社会中一些利用数字和数据造假的现象，引起了极大反响，并且后来被

翻译成多种语言。书中提出的统计陷阱的例子，比如样本选择偏差、平均数的选择以及相关性的滥用，在现今的生活中仍然十分常见。

达莱尔·哈夫的这本经典著作给了我们很多启示，统计数字看似客观公正，但背后其实隐藏着很多秘密。不过，由于《统计数字会撒谎》一书出版时间较早，书中描述的数据造假现象如今已经出现了新的变化。同时，随着技术的进步、学科的发展，以及各学科之间的融合，最近几十年出现了不少新的研究领域，比如大数据、机器学习等。在如今数据量膨胀的时代，数据收集和数据分析的方法越来越多样，而数据中蕴含的陷阱也越来越多。

这是一个变革的时代，变革不仅体现在政治领域，还体现在我们的观念和行为上。2016年一个个重大事件中，“情绪”代替了“真相”，“感性”代替了“理性”，事实被谎言包裹得越来越严密。《牛津英语词典》将2016年的年度词汇授予了“后真相”（post-truth），而德国人也十分巧合地把年度德语词汇给了postfaktisch，同样是“后真相”。整个世界似乎都进入了“后真相时代”，在这个时代，人们为了引导公众情绪，可以罔顾真相，扭曲事实。

2017年1月10日和11日，即将离任的美国总统奥巴马和新任总统特朗普分别在芝加哥和纽约与公众见面。奥巴马回到了他政治生涯开始的地方，发表了告别演讲。奥巴马在演讲中感谢了自己的妻子和孩子，说到动情之处不禁流下了泪水，此情此景既感人又温情。而到了特朗普那边，完全是画风大变。那是特朗普竞选成功以来的首次新闻发布会，可在这次发布会上，特朗普本色尽显，猛烈攻击军工、汽车制造等行业，场面剑拔弩张。更让人惊讶的是，当CNN记者想要提问时，特朗普直接拒绝，双方发生了语言上的冲突，特朗普更是直接斥责对方：“你们（报道的）是虚假新闻！”（You are fake news!）

“虚假新闻”（fake news）这个词是美国大选以来一直备受关注的话题。根据美国情报部门的调查，在刚刚过去的美国大选中，大量虚假新闻，特别是网络新闻，扰乱选举秩序，误导民众，影响选举结果，并且情报机构怀疑俄罗斯

政府参与其中。特朗普之所以斥责CNN，正是因为后者曾撰文披露俄罗斯掌握了特朗普不可见人的黑料。

许多人把矛头对准了社交平台facebook，指责其纵容虚假新闻在网站上传播而不加干涉。作为世界上使用范围最广的社交网站，一条新闻在这里的传播速度和影响范围要高于任何纸质媒体。

不过，facebook本身并不创造新闻，那些影响美国大选的虚假新闻，居然很大一部分来自东欧国家马其顿的一座小城Veles。在这里，编造假新闻已经成了一门生意，很多人的工作就是每天在网上凭空写一些假新闻，然后传播出去。他们当然不是怀有操纵政治的阴谋，而只是赚些养家糊口的钱。网站依靠这些耸人听闻的假新闻赚取流量，增加广告收入。就是这么简单的逻辑链条，却一不小心间接地改变了世界。

不仅政府批评媒体的不实报道，不少民众也指责媒体受到政府的控制，有选择性地播报新闻。在德国，这类为政府宣传虚假信息的媒体被称为“谎言媒体”，这个词最早可以追溯到19世纪，近些年来开始被极右民粹组织频繁使用。

媒体被看作社会公器，记者更是被看作社会的良心，可如今，他们却遭到了政府和（极右）民众，传统政党和民粹政党的同时攻击。2014年德国出版了《被收买的记者》（*Gekaufte Journalisten*）一书，作者在书中披露了大量关于德国记者如何被当事人收买而报道不实新闻，以及政客、情报机构和商人如何操纵媒体的内幕。作者在书中明确点出了上百个涉嫌操纵舆论的人员，以及多个知名媒体机构和国际组织的名字。这本书还爆料，情报机构的人员居然可以堂而皇之地在一家德国主流报纸的编辑部内撰写他们所需要的文章，然后以某位知名记者的名义刊登见报。作者更是断言，三分之二的记者都是可以被收买的。

很难讲到底是先出现了谎言媒体，才导致我们进入了后真相时代，还是后真相时代孕育出了谎言媒体，但毫无疑问的是，这二者的结合如此“契合”，

以至于真相离我们越来越远。我们正处在一个情绪比真相更重要的时代，人们没有耐心去探究事情的原委，不管是2016年的英国退欧公投还是美国总统大选，各种虚假新闻无时无刻不在撩拨着选民的情绪："退出欧盟可以每周为英国节省数亿英镑的开支""调查希拉里的FBI探员被爆死于公寓内"。投票选举期间，这类新闻在网络上流传甚广。然而，这些新闻事后都被证明要么数字是杜撰的，要么是以偏概全。

很多人把假新闻泛滥怪罪在互联网身上，认为正是互联网的发展，使得制造新闻和传播新闻的成本变得越来越低。在享受网络带来的丰富信息的同时，我们却渐渐发现，自己离真相越来越远。一个公共事件发生后，手机上立马能接到数条推送，过不了几个小时，上百篇"深度分析"就会把我们包裹起来：公众号、朋友圈、微博……

现在IT技术越来越发达，能够获取的数据量越来越大，拥有了用户信息，媒体可以很精准地推送给用户他们最感兴趣的新闻，网站可以为每位用户提供定制的消息。然而可悲的是，我们却总是迷失在信息的洪流之中，假新闻配合假数字，我们根本抓不住真相的尾巴。

我们正处在一个情绪比真相更重要的时代，越来越多的人不愿意探究事情的原委和证据，而是先通过指责和谩骂宣泄情绪。这也是一个流量比事实更重要的时代，本来一件还存在诸多疑问的事情，媒体却很"乐意"把它写成非黑即白，直接告诉读者谁是好人谁是坏人，因为这样才能成为热点话题，才能带来点击率。几个大V随手"转发"，民意立刻汹涌而来，你要么就红了，要么就是被钉在了耻辱柱上。可是在"后真相时代"，真相到底是什么？数字到底是精确的表达，还是谎言的制造者？

或许这些媒体与那些炮制虚假新闻的马其顿人一样，并没有去颠覆社会的野心，这么做只不过是为了挣钱养家糊口。这些媒体平台或许也和facebook有同样的理由：我们只是平台，只负责分发，虚假新闻又不是我们写的。可是，

正是这一层一层的“无意”，纵容了虚假新闻和数字陷阱的盛行。

不过，虚假信息也绝不是今天才出现，任何垄断信息源的媒介或平台，都有控制新闻和舆论的能力。如今，只不过是换了个“信息的守门人”，因为互联网来了。我们期待着互联网带来“去中心化”，期待信息不再只掌握在少数人手里。可是未曾想，那些掌握流量和信息分发渠道的互联网平台，却拥有了让当年传统媒体都难以企及的“权力”，他们影响甚至决定了我们平时看什么、听什么、聊什么，甚至是想什么，而且他们的影响跨越边界，不分种族。在这个时代，数字不仅是向外传递出的信息，同时还是各行各业迫切追求的“宝藏”。“大数据时代”，数字不再只是一个个枯燥的计算工具，而是能描述世间万事万物的神奇符号。信息流和数字流，它们不仅影响着我们的生活，也在深刻地改变着我们的思想。可是，当数字变得越重要，人们的生活都被数字连接起来，数字所带来的风险也就越大，用数字来说谎所造成的后果也就越严重。

人们的个人信息源源不断地汇入到互联网企业的手中，这些信息虽然只是由一些数据流组成，但却可以准确地反映出用户的各种形象、喜好，企业可以极其精准地为每位用户提供个性化服务。这个时代的人是幸福的，可以体验到以前从未有过的便捷生活，但是另一方面，信息安全也被提高到了前所未有的高度，我们身边经历的谎言事件越来越多，骗子的骗术也越来越高超，原因正是我们的个人信息太容易被泄露，所有这些被非法掌握的信息，最后串起来就能够勾勒出一个个个体。

在各种各样的社交媒体上，我们每天都会收到各式各样的文章，其中许多都起着耸人听闻的标题，然而内容却多以编造的谣言为主。在一个依靠流量的时代，文章的质量已经是次要的了，吸引眼球才是最关键的，所以我们就能够看到各式各样生拉硬拽的“因果关系”，添加一些看似合情合理的数字辅助，或者是直接对数字做手脚。数字被布下了各种陷阱，数据成了欺骗的手段。

但其实，数字本身不会说谎，数字只是一个信息载体，说谎的其实是使用数字的人。数字既可以拿来解释客观世界，也可以用来曲解事实真相。用数字撒谎的方法多种多样，不一而足，本书将尝试用简单易懂的语言分析常见的利用数字说谎的现象，同时结合一些常见的例子进行解析。

在大数据时代，数字和数字组成的数据成为最有价值的信息，人类拥有了以往从未有过的巨大信息量，一方面，信息量的猛增帮助人们更加深刻地理解自然世界和人类社会；但另一方面，以数字化为基础的各类新技术又在不断地改变人类社会，有的时候这种改变让人类生活变得更加美好，但有的时候却让人猝不及防。人类的工作岗位不断地被机器所取代，信息安全不断地受到威胁，这种数字化带来的挑战也成为“数字陷阱”的另一种表现形式。

历代科幻电影都乐意将拥有人工智能的机器人反叛人类当作主题，想象着这些被“数字”制造出来的机器是如何摧毁人类的；美国中央情报局前雇员爱德华·斯诺登曝光的美国国家安全局的“棱镜计划”监听项目更是让公众了解到，人类是如何利用数字为自己埋下“陷阱”的。

相比于用数字造假，这样改变社会结构的数字陷阱更加隐晦，但影响却更加深远。数字不仅影响我们的认知，也在影响整个人类历史文明。

目录

第1章

数字、数据与统计

我们已经进入了大数据时代，数据的体量已经大大高于历史上的任何时期，这为我们创造出了无限地可能，但同时也存在着不小的隐患。这一章将首先介绍有关数字和数据的基本信息，以及关于统计学的一些常用概念和应用场景。

●1.1 数字与统计学

数字被看作是一套严谨的表达体系，历史上人类发现了众多神奇的数字，它们看似巧合，但往往包含着深刻的道理，对数字的研究也是对人类社会和客观世界的研究。统计学就是一门大量依赖数字的学科，人们的生活中包含了大量的统计学知识，统计学能够帮助人们打开新的视野。

❑ 1.1.1
数字不仅仅是算算术

13世纪，意大利数学家斐波那契发现了一组对世界影响深远的神奇数字，这组数字为0、1、1、2、3、5、8、13、21、34、55、89、144、233、377、610、987……这组数字暗含着很多神奇而有趣的规律，吸引着后来的人们不断地挖掘，比如：1）从第三个数字开始，后一个数字都等于前两个数字之和，如1+1=2，8+13=21，55+89=144；2）随着数列项数的增加，每一个数字与其之后的数字的比值无限接近于0.618，如3/5=0.6，34/55≈0.6182，233/377≈0.618。

而0.618这个数字，因为具有严格的比例性、艺术性，蕴藏了很深的美学价值，在《蒙娜丽莎》等艺术作品以及人体结构上都有表现，因此这也被称作黄金分割点。

1679年，德国哲学家、数学家莱布尼茨发明了二进制计数系统，用简简单单的0和1两个数字来进行计数。而如今功能强大的计算机的运算正是以二进制为基础的。0到9这十个数字，可以描绘出各种能够想象到的和无法想象的图景。