EVERYBODY LIES

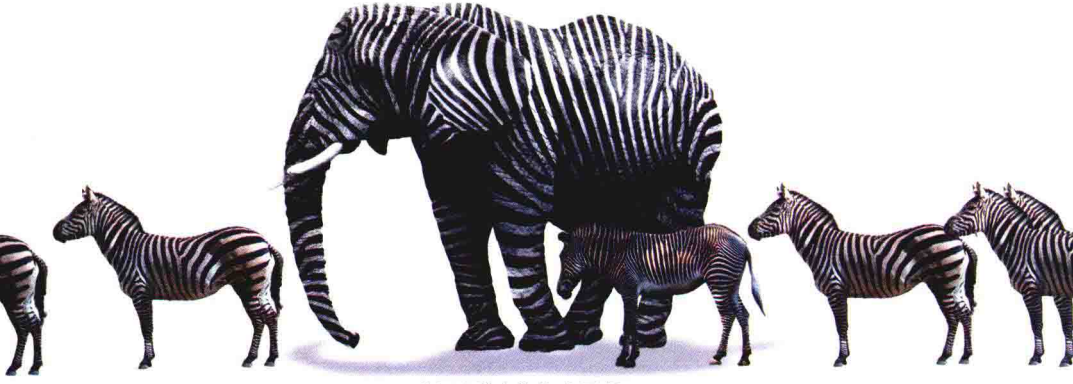
人人都在说谎

赤裸裸的数据真相

[美] 赛思·斯蒂芬斯-达维多维茨
（Seth Stephens-Davidowitz）著

胡晓姣 张晨 左润男 译

超越《魔鬼经济学》和《点石成金》，
这本无可比拟的书是大数据和智慧照亮世界并推动世界前进的一大例证。
——劳伦斯·萨默斯
美国前财政部部长、哈佛大学第27任名誉校长



《经济学人》年度图书
《企业家》商业畅销书
亚马逊年度图书
《纽约时报》畅销书

中信出版集团

一种研究思想的全新方法，一个窥视人类内心前所未有的入口，斯蒂芬斯-达维多维茨的发现一次又一次颠覆我对自己国家和人民的认知，这本书魅力无限！

——斯蒂芬·平克
《人性中的善良天使》作者

这本书列举了诸多鲜活的应用场景，证明了大数据是如何解构和重构这个世界的，又是如何大大延展人的认知边界的。越来越多的真相是反直觉的，大数据为我们揭示了一个更为真实的世界，人人都在说谎，不管你承不承认。

——李丽
太平再保险（中国）有限公司副总经理

《人人都在说谎》以大数据揭开了我们作为文明人的思维面纱。这本书引人入胜，令人震惊，有时甚至让人毛骨悚然，但最重要的是，它极具启发意义。

——吴修铭
《注意力经济》作者

《魔鬼经济学》的升级版——这本书展示了大数据如何对重要且有趣的问题做出出乎意料的全新解答。斯蒂芬斯-达维多维茨以干脆、诙谐的方式使数据分析活起来，为大数据如何塑造社会科学做了极好的阐述。

——拉杰·切蒂
斯坦福大学经济学教授

大数据揭露了我们日常生活的秘密，《人人都在说谎》一书就是对此绝顶聪明又有点儿顽皮的探寻。赛思·斯蒂芬斯-达维多维茨是我见过的最优秀的数据作者之一。

——史蒂芬·列维特
《魔鬼经济学》合著者

这是一次以搜索数据为导向的有关现代人类心理的旋风式旅程。这本书里的实证研究结果太有趣了，书中有很多有趣的案例，让人欲罢不能！

——《经济学人》

码上相逢

改变世界
从认识世界开始

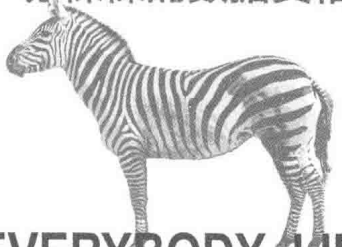
ISBN 978-7-5086-9387-3

9 787508 693873 >

定价：58.00 元

人人都在说谎

赤裸裸的数据真相



EVERYBODY LIES

[美] 赛思·斯蒂芬斯-达维多维茨
（Seth Stephens-Davidowitz）著

胡晓姣 张晨 左润男 译

中信出版集团 · 北京

图书在版编目（CIP）数据

人人都在说谎：赤裸裸的数据真相 /（美）赛思·斯蒂芬斯－达维多维茨著；胡晓姣，张晨，左润男译 . --
北京：中信出版社，2018.11
书名原文：Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are
ISBN 978-7-5086-9387-3

I. ①人… II. ①赛… ②胡… ③张… ④左… III. ①经济学 IV. ① F0

中国版本图书馆 CIP 数据核字（2018）第 195883 号

人人都在说谎——赤裸裸的数据真相

著　　者：［美］赛思·斯蒂芬斯－达维多维茨
译　　者：胡晓姣　张　晨　左润男
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲 4 号富盛大厦 2 座　邮编　100029）
承 印 者：北京楠萍印刷有限公司

开　　本：880mm × 1230mm　1/32　　印　　张：10.75　　字　　数：250 千字
版　　次：2018 年 11 月第 1 版　　印　　次：2018 年 11 月第 1 次印刷
京权图字：01-2018-6339　　广告经营许可证：京朝工商广字第 8087 号
书　　号：ISBN 978-7-5086-9387-3
定　　价：58.00 元

谨以此书献给

我的父母双亲

序言

许多哲学家曾经猜想有一种“大脑可视仪”，一种可以将一个人的想法呈现在屏幕上的虚构工具。自那时起，社会科学家就一直在寻找可以解读人性的工具。在我作为实验心理学家的职业生涯中，形形色色的工具一时兴起，消失淡去，每一种工具我都尝试过，如评定量表、反应时间、瞳孔放大、功能性神经影像等。我甚至还研究过植入电极的癫痫患者，在两次病发的间歇期，他们很乐意参与语言实验来消磨时光。

然而，这其中却没有任何一种工具能提供大脑活动的直接影像，问题在于这涉及错综复杂的多方面荣誉的许多命题，这些命题即便对一名科学家而言也是难解之题，他要做出一种残忍的取舍。人类思维是一个复杂的命题，与伍迪·艾伦（Woody Allen）速读《战争与和平》不同，我们不仅仅认为“这本书讲述了一些俄国人的故事”。当然，在人们倾吐心声的时候，我们可以察觉到其意识流的丰富性，但独白并不是检验假设的理想数据集。另外，虽然专注于容易量化的方法，比如人们对文字的反应时间或看到图片时皮

肤的反应，我们可以进行统计，但这样我们就将认知的复杂结构简化成单纯的数字了。即便是最复杂的神经影像学方法也只能告诉我们一个想法在三维空间中是如何展现出来的，却无法告诉我们这个想法是由什么组成的。

仿佛这个易操作性和丰富性之间的取舍还不够糟糕一般，许多研究人性问题的科学家当下正备受“小数定律”（Law of Small Numbers）的困扰。小数定律是阿莫斯·特沃斯基（Amos Tversky）和丹尼尔·卡尼曼（Daniel Kahneman）用来指代思维错误的术语，即人们错误地认为无论样本数量多么少，都能反映群族的特征。最具科学素养的科学家也有直觉出现严重偏误的时候。他们凭直觉推断完成一项研究需要有多少门学科的加入，可此前他们压根儿没有从一堆随机数据中排除异常及相悖的因素，也没有采集能代表全体美国人的数据，更不用说代表全人类了。若样本是以便利抽样的方式（比如给参与我们项目的大二学生发点儿零花钱）搜集而来的话，其可信度就更低了。

这本书介绍了一种研究思维的全新方式。尽管来自网络搜索和其他在线回应的大数据并非“大脑可视仪”，但赛思·斯蒂芬斯－达维多维茨表示，这些数据为窥视人类心灵提供了前所未有的机会。凭借网络对隐私的保护，人们敢于坦陈最奇怪的事，有时候（比如浏览婚恋网站或寻求专业咨询时）是因为这些事会对现实生活产生一定的影响，但更多的时候正是因为在网上坦陈这些事不会产生什

么影响：人们可以卸下心防，放下些许希望或恐惧，同时也不会有人因此感到沮丧或不适。无论采取哪种方式，人们所做的都不仅仅是按下一个按钮或旋转一个把手那么简单，而是键入数万亿个字符来表达自己心中不吐不快、情绪万千的许多想法。更方便的是，人们以一种方便汇总和分析的形式规定了这些数据痕迹。这些人来自各行各业，可以参与那些不那么引人注目的实验，这些实验可以使刺激因素多样化，并且将多种反应制作成实时表格。这些人很乐意提供这些规模庞大的数据。

《人人都在说谎》一书远不止证明一个概念。斯蒂芬斯－达维多维茨的发现一次又一次地颠覆了我对自己的国家和族群的认知。特朗普意想不到的支持从何而来？1976 年，安·兰德斯（Ann Landers）曾经问过她的读者有没有后悔生孩子，她震惊地发现多数人确实后悔过。当时她是否被不具代表性的自选样本误导了呢？互联网应该对 21 世纪第二个 10 年后期出现的那次多余起名的危机（即“过滤气泡”[①]危机）负责吗？是什么激发了仇恨犯罪？人们会通过笑话振作精神吗？尽管我一心认为没有什么可以令我吃惊，却还是被互联网披露的人类性需求（包括一定数量的女性每个月都会在网上搜索“和毛绒玩具滚床单”这一发现）惊到了。采用反应时间、瞳孔放大或功能性神经影像等工具的实验中没有一项能反映这样的事实。

① “过滤气泡”是指在算法推荐机制下，高度同质化的信息流会阻碍人们认识真实的世界。——编者注

每个人都会喜欢这本书。赛思·斯蒂芬斯–达维多维茨用他的好奇心和智慧为21世纪的社会科学指出了一条新道路。有这样一扇散发着无限魅力，能够窥视人类内心喜好的窗户，谁还需要大脑可视仪呢？

——斯蒂芬·平克（Steven Pinker），2017年

目　录

序　言 / V

绪　论 / 001

第一部分　大数据，小数据

1　你的直觉出错了 / 027

第二部分　大数据的力量

2　弗洛伊德是正确的吗 / 047

3　数据重构 / 057

以身体为数据 / 064

文字数据 / 076

图片数据 / 098

4 数字吐真剂 / 105
性的真相 / 112
憎恶与偏见的真相 / 124
互联网的真相 / 136
虐待儿童和人工流产的真相 / 141
脸谱网好友的真相 / 146
用户的真相 / 149
我们能处理真相吗 / 154

5 放大数据 / 161
我们的县、市和镇中到底在发生着什么 / 168
如何填满我们的每时每刻 / 186
我们的二重身 / 193
数据的故事 / 202

6 世界就是一个实验室 / 205
A/B 测试三两面 / 207
自然残酷而又发人深省的试验 / 219

第三部分 大数据：请小心轻放

7 大数据，大框架？其力有何不能胜 / 243
维度的诅咒 / 246
过分强调什么是可以测量的 / 252

8 数据越多，问题越多？哪些事情不可为 / 259

授权公司的危险 / 259

授权政府的危险 / 267

结 论 / 271

致 谢 / 285

注 释 / 291

绪论

改革概述

人们说，他必败无疑。

2016 年美国共和党初选时，民意调查专家断定特朗普毫无胜算，毕竟特朗普曾冒犯过不少少数群体。民意调查结果显示，几乎没有任何一个美国人赞成这样的行径。

当时，大多数民意调查专家认为特朗普会在普选环节败北。很多拟投票的选民说，考虑到特朗普的言行，他们最终放弃投票。

但当时确实有一些迹象表明特朗普有可能赢得党内初选以及普选——这些迹象源于网络。

我是一名互联网数据专家，每天都会跟踪记录人们浏览网页时留下的数据痕迹。根据人们点击的频度，我努力解读他们真正想要的、真正要做的和他们的真面目。下面我来解释一下我是如何走上这条不寻常之路的。

说来话长（这样一讲，好像是几个世纪前的事了），事情要从2008年总统大选和那个社会科学界争论已久的问题说起：在美国，种族偏见到底有多大的影响？

奥巴马当年是以美国主要政党中第一位非洲裔美国总统候选人的身份参与竞选的。他赢得非常轻松。民意调查结果显示，种族并不是影响美国人投票的因素之一。例如，盖洛普民意测验公司（Gallup）在奥巴马初选前后进行了多次民意调查，结论是什么？美国选民多半不在意奥巴马是黑人。[1] 选举结束后不久，加州大学伯克利分校的两位知名专家使用更加复杂的数据挖掘技术（data-mining techniques）研究了其他调查数据并得出了相似的结论。[2]

而且，在奥巴马任职期间，这也成了许多媒体和众多科研院所的共识。媒体和社会科学家80多年来一直用于了解这个世界的信息资源告诉我们，在判断奥巴马应不应该成为总统时，绝大多数美国人根本不在意他是黑人。

这个国家曾因奴隶制度和种族隔离法而长期备受诟病，如今貌似终于不再以肤色来评判一个人了。这似乎表明种族歧视在美国已经穷途末路了。事实上，有些专家甚至宣称我们已生活在后种族社会（post-racial society）了。[3]

2012年，当时还是一名经济学研究生的我，对生活感到十分迷茫，对经济学领域的研究也失去了热情，我自信（甚至有些自大）对世界的运作方式和人们在21世纪的所思所虑都有着深刻理解。涉

及种族偏见问题时，基于对心理学和政治科学领域的了解，我相信显性种族主义（explicit racism）仅仅局限于极少数美国人——其中大多数人是保守的共和党人，且大都居住在南方诸州。

然后我发现了谷歌趋势（Google Trends）。

2009 年，谷歌隆重推出一款数据挖掘工具——谷歌趋势，它可以告诉使用者任何一个词语或短语在不同时间、不同地点的使用频率。谷歌趋势的宣传定位是一种有趣的工具——也许是因为它可以让朋友之间讨论哪位明星最受欢迎，什么样的潮流一下子火了起来。这一工具最初的几个版本还包括一句幽默的警告：人们应该“不想借助这一数据撰写博士学位论文”。这句话立刻激发了我依靠这些数据完成学位论文的积极性。[①]

当时，对“正派”学术研究来说，谷歌搜索数据似乎并不是恰当的信息来源。与调查不同，谷歌搜索数据的创建并非用于帮助我们了解人类的心灵。人们发明谷歌，是为了了解世界，而非让研究

① 我的数据有很大一部分都源于谷歌趋势。然而，由于该方法只允许比较不同搜索的相对频率，无法报告任何特定搜索的绝对数量，因此我常使用谷歌广告关键词（Google Adwords）加以辅助，这一搜索方式能准确报告每种搜索的频率。在大多数情况下，我也能够利用自己基于谷歌趋势的算法来锐化图片。有关这一点，我在我的博士学位论文《使用谷歌数据的论文》（Essays Using Google Data）和在《公共经济学》（*Journal of Public Economics*）上发表的论文《种族敌意对黑人候选人的影响：使用谷歌搜索数据的证据》（The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data）中做过论述。我的博士学位论文、论文链接以及对本书提及的所有原始研究使用的数据和代码的完整解释都可参阅我的个人网站 sethsd.com。——作者注

人员了解人类，不过最终结果却是我们上网探求知识时留下的痕迹遭到了很大程度的暴露。

换句话说，人们搜寻信息这一行为本身就是信息。事实证明，他们何时何地搜寻真相、格言、笑话、地点、人物、事件或帮助，可以在很大程度上反映他们真实的想法、欲望、恐惧和职业，其程度之高是任何人都想象不到的。尤其是人们向谷歌坦陈“我恨我的老板”“我喝醉了”“我爸爸打了我”等心境时更是如此。

把词语或短语输入一个小小的白色长方形对话框这一日常行为总会留下关于真相的蛛丝马迹：这个细微的痕迹重复出现数百万次，最终一定会揭示许多深刻的现实问题。我在谷歌趋势输入的第一个词语是“上帝”，我了解到，使用谷歌搜索提及“上帝”一词最多的州有亚拉巴马州、密西西比州和阿肯色州，即《圣经》地带（the Bible Belt），而那些搜索大多发生在周日。这都不足为奇，但有趣的是，搜索数据可以揭示这样一种清晰的模式。我试着搜索了“尼克斯队”，结果显示搜索次数最多的地区是纽约市。这也是毫无疑问的。接着，我又输入了自己的名字，谷歌趋势提示我“很抱歉，搜索量不足”，无法显示结果。因此，我了解到只有在很多人做过相同的搜索之后，谷歌趋势才会提供数据。

谷歌搜索的功能不是告诉我们上帝在南方很受欢迎，不是尼克斯队在纽约市很受欢迎，也不是我在哪儿都不招人待见。任何一项调查都可以反映上述事实。谷歌数据的功能在于，人们会向这个巨

大的搜索引擎倾吐他们不会告诉任何人的事情。

就以性（在本书后文中会深入探讨这个话题）为例。那些调查并不足以反映人们性生活的实际状况。我分析过综合社会调查的数据，这项调查被视为反映美国人行为的最具影响力和权威性的信息来源之一。[4] 根据这项调查，谈及异性性行为时，女性会说她们平均每年有 55 次性行为，其中 16% 的情况下使用安全套。据此，每年安全套的使用量会多达 11 亿个。有异性性行为的男性则说，他们每年一共使用 16 亿个安全套。这两个数字本应是一致的。那么，谁说的是实话呢，男性还是女性？

结果显示他们都没有说实话。根据追踪消费者行为的全球信息与计量公司尼尔森市场调查公司（Nielsen）的数据，每年安全套的销量尚不足 6 亿个。[5] 因此，人人都在说谎，唯一的不同就是说谎程度的大小。

事实上，说谎是一种普遍行为。未婚男性称他们每年人均使用 29 个安全套，这一数据合计要比美国已婚人士和单身人士这两个人群每年人均购买的安全套总和还要多。已婚人士也夸大了他们的性行为次数。平均而言，65 岁以下的已婚男性告诉调查人员他们每周有一次性行为，只有 1% 的人说他们去年一整年都没有发生性行为。已婚女性称她们的性行为次数要少一点，但也不会少很多。

谷歌搜索为我们呈现了一幅婚内性行为的图像，虽然没那么生动，但我认为其更加确切。在谷歌上，网民对婚姻抱怨最多的就是