A REVOLUT L TRANSFORM HOW WORK,AND THINK

大数据时代

生活、工作与思维的大变革

[英]维克托·迈尔-舍恩伯格 (Viktor Mayer-Schönberger) 肯尼思·库克耶 (Kenneth Cukier) ◎著 盛杨燕 周涛◎译

浙江人民出版社
ZHEJIANG PEOPLE'S PUBLISHING HOUSE

BIG DATA

A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK

大数据时代

生活、工作与思维的大变革

[英] 维克托·迈尔-舍恩伯格（Viktor Mayer-Schönberger）◎著
肯尼思·库克耶（Kenneth Cukier）
盛杨燕 周涛◎译

浙江人民出版社
ZHEJIANG PEOPLE'S PUBLISHING HOUSE

图书在版编目（CIP）数据

大数据时代 /（英）迈尔－舍恩伯格，（英）库克耶著；盛杨燕，周涛译 . —杭州：浙江人民出版社，2013.1

ISBN 978-7-213-05254-5

Ⅰ.①大… Ⅱ.①迈… ②库… ③盛… ④周… Ⅲ.①网络经济–通俗读物 Ⅳ.①F062.5–49

中国版本图书馆 CIP 数据核字（2013）第 288357 号

浙江省版权局
著作权合同登记章
图字:11-2012-222 号

大数据时代

作　　者：维克托·迈尔–舍恩伯格　肯尼思·库克耶　著

译　　者：盛杨燕　周　涛

出版发行：浙江人民出版社（杭州体育场路347号　邮编　310006）

市场部电话：（0571）85061682　85176516

集团网址：浙江出版联合集团　http://www.zjcb.com

责任编辑：金　纪　王方玲

责任校对：朱志萍

印　　刷：北京京北印刷有限公司

开　　本：720 mm × 965 mm　1/16　　**印　　张**：18

字　　数：21.5 万　　**插　　页**：4

版　　次：2013 年 1 月第 1 版　　**印　　次**：2013 年 1 月第 1 次印刷

书　　号：ISBN 978-7-213-05254-5

定　　价：49.90 元

拥抱“大数据时代”

宽带资本董事长　田溯宁

从硅谷到北京，大数据的话题正在被传播。随着智能手机以及“可佩带”计算设备的出现，我们的行为、位置，甚至身体生理数据等每一点变化都成为了可被记录和分析的数据。以此为基础，“反馈经济”（feedback economy）等新经济、新商业模式也正在开始形成。**维克托·迈尔－舍恩伯格教授这本《大数据时代》，是我看到的最好的大数据著作，不管对于产业实践者，还是对于政府和公众机构，都是非常具有价值的。**

如今，一个大规模生产、分享和应用数据的时代正在开启。正如维克托教授所说，大数据的真实价值就像漂浮在海洋中的冰山，第一眼只能看到冰山的一角，绝大部分都隐藏在表面之下。而发掘数据价值、征服数据海洋的“动力”就是云计算。互联网时代，尤其是社交网络、电子商务与移动通信把人类社会带入了一个以“PB”（1 024TB）为单位的结构与非结构数据信息的新时代。在云计算出现之前，传统的计算机是无法处理如此量大、并且不规则的“非结构数据”的。

以云计算为基础的信息存储、分享和挖掘手段，可以便宜、有效地将这些大量、高速、多变化的终端数据存储下来，并随时进行分析与计算。大数据与云计算是一个问题的两面：一个是问题，一个是解决问题的方法。通过云计算对大数据进行分析、预测，会使得决策更为精准，释放出更多数据的隐藏价值。**数据，这个21世纪人类探索的新边疆，正在被云计算发现、征服。**

《大数据时代》列举了众多在公共卫生、商业服务领域大数据变革的例子。一旦“不再追求精确度，不再追求因果关系，而是承认混杂性，探索相关关系”，“思维转变过来，数据就能被巧妙地用来激发新产品和新型服务”。数据正成为巨大的经济资产，成为新世纪的矿产与石油，将带来全新的创业方向、商业模式和投资机会。

庞大的人群和应用市场，复杂性高、充满变化，使得中国成为世界上最复杂的大数据国家。**解决这种由大规模数据引发的问题，探索以大数据为基础的解决方案，是中国产业升级、效率提高的重要手段。**数据挖掘不仅能够成为公司竞争力的来源，也将成为国家竞争力的一部分。联系到我国现代化所面临的种种问题以及教育、交通、医疗保健等各方面挑战，通过大数据这种创新方式来解决问题，创建新的产业群，实现“中国制造到中国创造”的改变，意义就更大。

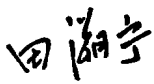
“大数据”发展的障碍，在于数据的“流动性”和“可获取性”。美国政府创建了Data.gov网站，为大数据敞开了大门；英国、印度也有“数据公开”运动。中国要赶上这样一场大数据变革，各界应该首先开始尝试公开数据、方式与方法。如同工业革命要开放物质交易、流通一样，开放、流通的数据是时代趋势的要求。《大数据时代》一书也提到了数据拥有权、隐私性保护等问题，但相比较来看，新科技可能带来的改变要远远大于其存在的问题。

推荐序一

拥抱“大数据时代”

本书的译者周涛教授是我国最年轻有为的大数据专家。这位27岁的天才型教授，数年来一直带领我国学术界在大数据研究上向国际一流看齐。更可贵的是，他不仅做研究，也关注着研究成果的商业化及传播。这部译著就是他这种努力的一个成果。

现代历史上的历次技术革命，中国均是学习者。而在这次云计算与大数据的新变革中，中国与世界的距离最小，在很多领域甚至还有着创新与领先的可能。只要我们以开放的心态、创新的勇气拥抱“大数据时代”，就一定会抓住历史赋予中国创新的机会。



实实在在大数据

中国互联网发展的重要参与者，知名IT评论人　谢文

因为我本身十分关注大数据，也写过若干关于大数据的文章，做过若干关于大数据的演讲，所以对有关这一主题的论文和书籍非常有兴趣。过去几年，在这方面我读过十几本书、上百篇论文和文章。相对而言，维克托·迈尔－舍恩伯格教授的《大数据时代》是迄今为止我读过的最好的一本专著，中英文都算上。

此书的一大贡献就是在大数据方兴未艾、众说纷纭的时刻，进一步阐述和厘清了大数据的基本概念和特点，这对许多以为大数据就是“数据大”的人来说很有帮助。

在人类历史长河中，即使是在现代社会日新月异的发展中，人们还主要是依赖抽样数据、局部数据和片面数据，甚至在无法获得实证数据的时候纯粹依赖经验、理论、假设和价值观去发现未知领域的规律。因此，人们对世界的认识往往是表面的、肤浅的、简单的、扭曲的或者是无知的。维克托指出，大数据时代的

来临使人类第一次有机会和条件，在非常多的领域和非常深入的层次获得和使用全面数据、完整数据和系统数据，深入探索现实世界的规律，获取过去不可能获取的知识，得到过去无法企及的商机。

大数据的出现，使得通过数据分析获得知识、商机和社会服务的能力从以往局限于少数象牙塔之中的学术精英圈子扩大到了普通的机构、企业和政府部门。门槛的降低直接导致了数据的容错率提高和成本的降低，但**正如维克托所强调的，最重要的是人们可以在很大程度上从对于因果关系的追求中解脱出来，转而将注意力放在相关关系的发现和使用上。**只要发现了两个现象之间存在的显著相关性，就可以创造巨大的经济或社会效益，而弄清二者为什么相关可以留待学者们慢慢研究。大数据之所以可能成为一个“时代”，在很大程度上是因为这是一个可以由社会各界广泛参与，八面出击，处处结果的社会运动，而不仅仅是少数专家学者的研究对象。

大数据将逐渐成为现代社会基础设施的一部分，就像公路、铁路、港口、水电和通信网络一样不可或缺。但就其价值特性而言，大数据却和这些物理化的基础设施不同，不会因为人们的使用而折旧和贬值。例如，一组 DNA 可能会死亡或毁灭，但数据化的 DNA 却会永存。所以，维克托赞同许多物理学家的看法，世界的本质就是数据。因此，**大数据时代的经济学、政治学、社会学和许多科学门类都会发生巨大甚至是本质上的变化和发展，进而影响人类的价值体系、知识体系和生活方式。**哲学史上争论不休的世界可知论和不可知论将会转变为实证科学中的具体问题。可知性是绝对的，无事无物不可知；不可知性是相对的，是尚未知道的意思。

对于不从事网络业、IT 业以及数据分析和使用的读者，本书的一大好处就是通俗易懂，通过具体实例说明问题，有助于人们的理解和联想。在时限上，作者概括了直到 2012 年 7 月大数据方向上的最新发展，避免了许多同类作品存在的例证过于陈旧、视野相对狭窄的毛病。

作为一位生活在欧美现代社会的学者，维克托是把民主、开放和理性作为已知前提来讨论大数据革命的。这对生活在发展中国家，社会现代化程度尚且有限的读者来说，也许是个遗憾，因为书中描述的许多已经发生的事例可能更像是神话。没有市场经济制度和法治体系作为基础支撑，大数据很可能成为发达国家在下一轮全球化竞争中的利器，而发展中国家依然处于被动依附的状态之中。整个世界可能被割裂为大数据时代、小数据时代和无数据时代。

处于发展中国家前列的中国，目前正面临着一个重大的历史抉择关口。应该说，在过去的三十余年时间里，中国在快速走向工业化、信息化、网络化方面交出了一份不错的成绩单。如今适逢世界走向数据化，迈入大数据时代的时刻，无论对个人、企业还是对社会和国家，都有认真理解、严肃决策的必要性和紧迫性。哪怕仅从这一点考虑，读一读这本书也是很值得的。

在路上·晃晃悠悠

电子科技大学教授，互联网科学中心主任　周　涛

接下翻译这本《大数据时代》的任务时，我的目标是做到110%的好。因为作者维克托·迈尔－舍恩伯格毕竟不像我们每天在一线与数据厮杀搏斗，其爱其恨都更深刻。特别地，我们可以为中文的读者补充很多中国的例子和参考资料。很遗憾，我们最终只做到了90%，应该补充的一些材料还没有整理好，遣词造句也多有生硬疏忽之处。如果再给我一个月的时间，就可以达到我预想的110%甚至120%。

为什么现在把这个版本呈现给诸位呢？一是因为我们的努力使得本书中译本的出版和英文原版完全同步，单从获取知识的角度讲，我们一点儿不比美国的读者慢！二是我相信作者在书中的一个重要观点，就是大数据时代，要允许一点点的错误和不完美，因为效率可能更加重要！留下一些可供提高的地方，也使得我们的每一次印刷，都能够与以前有所不同。亲，这不是建议你等到某个更好的版本才去购买，而是说，其实你应该每个版本都买一本：）

《大数据时代》这本书是200%的好，因此90%的译本也绝对值得一读。首

先，作者抛出了大数据时代处理数据理念上的三大转变：要全体不要抽样，要效率不要绝对精确，要相关不要因果；接着，从万事万物数据化和数据交叉复用的巨大价值两个方面，讲述驱动大数据战车在材质和智力方面向前滚动的最根本动力；最后，作者冷静描绘了大数据帝国前夜的脆弱和不安，包括产业生态环境、数据安全隐私、信息公正公开等问题。

国内最近也出版了一些大数据方面的著作，可以和本书互为补充。郑毅的《证析》对于数据通过交叉复用体现的新价值、大数据战略在企业与政府执行层面的流程和大数据科学家这一新职位，以及围绕这个职位的能力和责任给出了最深刻、最具体的描述；子沛的《大数据》对于数据的公正性、公平性以及信息和数据管理等方面理念、政策和执行的变化，特别是美国在这方面的进展，给出了完整的介绍；苏萌、林森和我合著的《个性化：商业的未来》则对大数据时代最重要的技术、个性化技术，以及与之相关的新商业模式给出了从理念到技术细节的全景工笔。**总的来说，这三本书都针对本书的某一局部给出了更深刻的介绍和洞见，也各有明显超出本书的优点，但三本之和也无法囊括本书的菁华，亦缺乏本书的宏大视野。**

简单地说，这本书好在三个地方：

一是观点掷地有声，绝非主流媒体上若干讨论的简单汇总和平均，更不是一个宏大概念面前暧昧的叫好声。读者可能对其中一些观点并不认同，但是读完之后不可能一个都记不住。

二是观念高屋建瓴，作者试图从很多实例和经验，包括历史事件中萃取出普适性的观念，而不仅仅是适用于几个特定情况的案例分析。

三是例子丰富翔实，不长的篇幅包括了上百个学术和商业的实例。

三点近乎完美地结合起来，体现了作者驾驭大问题的能力和丰富的知识，以及，可能更为重要地，作者渴求立言立说的野心！所以说，这本书绝对不是一堆枯燥的纲要，更不是一本巨厚的杂志。

我在这里拼命叫好，是为了这本书卖得更多，但不代表作者的所有观点都是绝对真理。举个例子，我本人对于大数据时代“相关关系比因果关系更重要”这个观点就不认同。有了机器学习，特别是集成学习，我们解决问题的方式变成了训练所有可能的模型和拟合所有可能的参数——问题从一个端口进去，答案从另一个端口出来，中间则是一个黑匣子，因为没有人能够从成千上万的参数拟合值里面读到“科学”，我们读到的只是“计算机工程”。与其说大数据让我们重视相关胜于因果，不如说机器学习和以结果为导向的研究思路让我们变成这样。

那么，大数据是不是都这样了？其实很多时候恰恰相反。想想瑞士日内瓦的强子对撞机，我们在上面捕获了人类有史以来最大规模的单位时间数据。我们是希望找到或者验证某种相关关系吗？不是！我们试图回答的，正是人类所能问出的关于因果关系最伟大的问题：希格斯玻色子是否存在，我们的宇宙是否有可能用标准模型刻画。这个问题的最终答案，将打破人和神的界限！认为相关重于因果，是某些有代表性的大数据分析手段（譬如机器学习）里面内禀的实用主义的魅影，绝非大数据自身的诉求。从小处讲，作者试图避免的“数据的独裁”和“错误的前提导致错误的结论”，其解决之道恰在于挖掘因果逻辑而非相关性；从大处讲，放弃对因果性的追求，就是放弃了人类凌驾于计算机之上的智力优势，是人类自身的放纵和堕落。如果未来某一天机器和计算完全接管了这个世界，那么这种放弃就是末日之始。

苏珊·朗格（Susan Langer）在《哲学新视野》一书中说：

> 某些观念有时会以惊人的力量给知识状况带来巨大的冲击。由于这些观念能一下子解决许多问题，所以，它们似乎将有希望解决所有基本问题，澄清所有不明了的疑点。每个人都想迅速地抓住它们，作为进入某种新实证科学的法宝，作为可以用来建构一个综合分析体系的概念轴心。这种‘宏大概念’突然流行起来，一时间把几乎所有的东西都挤到了一边。

这段话通常被认为是对当时“存在主义”和“精神分析法”这类万能概念的善意批评，而如今特别适合作为一盆冷水泼在那些没有任何深刻理解，却月月日日分分秒秒穿行于各种“大数据嘉年华”的投资人、媒体人和创业者身上。

希望《大数据时代》给予各位的是一些实实在在的知识和思考，并且唤起各位安静思索相关问题的心境。大数据是一个很重要的概念，代表了很重要的趋势，但我不希望它成为一种放之四海皆准的万能概念——因为越是万能的，就越是空洞的！人类学家克利福德·吉尔兹（Clifford Geertz）在其著作《文化的解释》中曾给出了一个朴素而冷静的劝说：“努力在可以应用、可以拓展的地方，应用它、拓展它；在不能应用、不能拓展的地方，就停下来。”我想，这应该是所有人面对一个新领域或新概念时应有的态度。

大数据的道路上没有戈多，我们已经在路上，晃晃悠悠。人类的自由意志和诸神之下的尊严，会在这条道路上异化甚至消逝吗？极目远眺，不知道世界的尽头，是否是一个冷酷的仙境！诸位为之奋斗吧，而我只想，做一个，麦田里的守望者。

以为序。

BIG DATA

A Revolution That Will Transform How We Live, Work, and Think

目 录

第一部分　大数据时代的思维变革

01 更多

当数据处理技术已经发生了翻天覆地的变化时，在大数据时代进行抽样分析就像在汽车时代骑马一样。一切都改变了，我们需要的是所有的数据，“样本＝总体”。

让数据“发声”

小数据时代的随机采样，最少的数据获得最多的信息

全数据模式，样本＝总体

02 更杂

执迷于精确性是信息缺乏时代和模拟时代的产物。只有5%的数据是有框架且能适用于传统数据库的。如果不接受混乱，剩下95%的非框架数据都无法被利用，只有接受不精确性，我们才能打开一扇从未涉足的世界的窗户。

允许不精确

大数据的简单算法比小数据的复杂算法更有效

纷繁的数据越多越好

混杂性，不是竭力避免，而是标准途径

新的数据库设计的诞生

03 更好

知道“是什么”就够了，没必要知道“为什么”。在大数据时代，我们不必非得知道现象背后的原因，而是要让数据自己“发声”。

第二部分 大数据时代的商业变革

04 数据化

大数据发展的核心动力来源于人类测量、记录和分析世界的渴望。信息技术变革随处可见，但是如今信息技术变革的重点在“T”（技术）上，而不是在“I”（信息）上。现在，我们是时候把聚关灯打向“I”，开始关注信息本身了。

05 价值

数据就像一个神奇的钻石矿，当它的首要价值被发掘后仍能不断给予。它的真实价值就像漂浮在海洋中的冰山，第一眼只能看到冰山的一角，而绝大部分都隐藏在表面之下。

微软以 1.1 亿美元的价格购买了大数据公司 Farecast，而两年后谷歌则以 7 亿美元的价格购买了给 Farecast 提供数据的 ITA Software 公司。如今，我们正处在大数据时代的早期，思维和技术是最有价值的，但是最终大部分的价值还是必须从数据本身来挖掘。

我们时刻都暴露在“第三只眼”之下：亚马逊监视着我们的购物习惯，谷歌着监视我们的网页浏览习惯，而微