理想国
ginist

The Big Data Revolution

涂子沛 著

除了上帝，任何人都必须用数据来说话。

大数据

正在到来的数据革命，
以及它如何改变政府、
商业与我们的生活

GUANGXI NORMAL UNIVERSITY PRESS
广西师范大学出版社

涂子沛 著

The Big Data Revolution

大数据

正在到来的数据革命，
以及它如何改变政府、
商业与我们的生活

广西师范大学出版社
·桂林·

图书在版编目 (CIP) 数据

大数据：正在到来的数据革命，以及它如何改变政府、商业与我们的生活
/ 涂子沛著. —桂林：广西师范大学出版社，2012.7

ISBN 978-7-5495-1837-1

Ⅰ. ①大… Ⅱ. ①涂… Ⅲ. ①信息产业－产业经济－通俗读物
Ⅳ. ① F49-49

中国版本图书馆 CIP 数据核字 (2012) 第 105215 号

广西师范大学出版社出版发行

桂林市中华路22号　邮政编码：541001
网址：www.bbtpress.com

出 版 人：何林夏

全国新华书店经销

发行热线：010-64284815

山东临沂新华印刷物流集团印刷

开本：700mm × 1000mm　1/16

印张：22.25　字数：210千字　图表：145幅

2012年7月第1版　2012年7月第1次印刷

定价：45.00元

序言一　大数据：为华文世界提出一个重要话题

许倬云

涂子沛先生的新著《大数据》，已经完成，是一部300多页的大作。最近他将这本书的打样稿送来给我看，并且要求我作序文。可惜的是，我将要远行，而他索序甚急，我只能在浏览一遍之后，发抒自己一些感想。至于细细咀嚼和消化这本有趣作品的内容，必须要在一两个月以后，也就赶不上涂先生大作的出版时间。我事先声明，这些只是我浏览他大作以后的一些感想，非常立即的直觉，也一时谈不上深刻的见解。

这本书的内容，强调今天是一个大量数据公开于大众的时代。自从资讯革命以来，资讯工具、硬件、软件，平行发展、与时俱进。数据在我们生活之中，日常接触，已是处处可见的现象。收集数据，当然是靠电脑的快速分类和记录，然而更重要的是搜索引擎的进步，与网络之间网际的交流。到今天，一个一个网，不但可以串连在一起，互相沟通，而且“云端”（云计算）的设计，可以将每一个地区个别数据，储成一个大的数据库，有助于我们更迅速广泛地搜索。

这些现象，20 世纪最后四分之一以来，已经呈现加速度发展的新事物，在 21 世纪进展速度之快，更是铺天盖地、无所不在。涂先生在这本书里强调：不仅数据经过管理而大量地存在，而且，在现代的国家，尤其是民主社会，开放的社会与政府之间，经由数据，彼此一目了然、无所隐瞒。一个自由的社会，掌握公权力的政府，跟任何其他政府一样，有压倒社会的庞大力量，因为他们手上掌握了人生需要的许多数据。不过，这些数据，在自由的社会，公民也可以一样取得，使政府所作所为，可以摊开在天地之间，让我们检验。过去封建专制和集权各种政体，其执政者能压迫老百姓，而老百姓没有办法回制公权力的压迫。涂先生特别标榜，美国奥巴马接任以后，尽力将数据开放于大众，固然奥巴马是一个有开放心胸的政治人物，如此将政府掌握的数据，大量地开放于群众，也是拜时代之赐，有如此的机缘，才能将数据公开。

涂先生引用胡适之先生与黄仁宇先生的话。胡先生说中国人习惯于“差不多先生”，凡事马马虎虎、不求精确。黄仁宇先生认为，中国不懂得用数字来管理国家。涂先生引用这两位先生的名言，当然是要彰显传统中国和今天美国之间的巨大差异。不过我必须有所说明：胡先生和黄先生的话语，都是“爱之深而责之切”的心态，他们身经当时中国的混乱，激愤而出此感言。

从历史上看看，不论中国和西方，任何国家发展到可以有一个复杂文官系统管理以后，没有不依照数据来治国的。人口、资源、土地、财产种种的统计数字，在中国历史上，自从战国时代形成列国的国家体制以后，没有一个朝代不具有一定的数据库；只是以今天的标准来讲，粗糙和细密之间，古今有很多的差别而已。以汉代为例，汉简所显示的家户统计，每一户中的人口，男女老小，以及拥有的资产数目字，都详细统计，而且不论是居延边塞，或是荆州内郡，格式一致。汉简各种家户统计，与唐代西域州府的记录对比，其内容格式也是相当一致。这种基本的数据，在列朝的会典中，都见到其大概。当然，各个朝代的数据，有做得好的，也有做得差的。大致讲起来，外族侵犯中国建立的朝代，以武装力量强制建立政权，也往往依靠暴力的掠夺，取得他们所需要的资源。一个上轨道的朝代，其数据还是相当完整。

再看西方历史。希腊时代，我们了解的资料不够。罗马帝国时代，全帝国包含各种不同的政治单位，并没有一个大一统的文官政府；因此，全国性的资料库似乎不存在。等到中古黑暗时期，国不成国，地方不过是大小封建领主占有领土而已，他们并没有建立详细的资料库。近代以来，列国各自组成完整的主权国家，这些数据也纷纷出现了。

这是以历史上政权掌握数字而言。一家大型的企业，例如，中国清代的票号，如果他们手上没有复杂的数据库，就不能进行汇兑、放款、存款等等活动。英国的东印度公司，手上握有丰富的资源，他们也不能不具有一个相当完整的数据库，否则无以经营这么复杂的开拓业务。

今天的数据时代，我已经在前面提过，不仅公司单位都有搜集数据的能力，而且更重要的，有搜寻引擎可以将资料迅速检索，从其中归纳出条理，有助于了解情况。举一个例说，最近我们才看到，数据资料显示，美国百分之一的人口，拥有全国财富百分之四十以上，百分之九十九的人口，拥有全国财富才过半而已。对于许多长期习惯于美国是开放社会的一般人民，这一组数据显示的现象，几乎可说是理想的破灭，使大家必须检讨：美国真是如此开放吗？还是相对地在逐渐关闭？是不是财富与权力，已经逐渐集中到社会顶端一小撮的人手中？他们以财富作为魔法师的指挥棒，安排了我们的生活、决定了我们的未来。这种现象能够暴露于众，当然就因为在美国究竟资讯是公开的。

相对而言，在极权的国家，他们手上拥有足够的资讯，足以利用这资讯，掌握每一个人的日常生活；“老大哥”的影子，可以无所不在。资讯时代，对于极权的掌权者，他们拥有无可比拟的强大工具，甚至于比坦克车和催泪弹更为有效。资讯管理、资讯控制，是无影无踪，又是无所不在。说到这里，我们不能不更多警惕。

作为一个史学工作者，看惯了世间的灾难和创伤，不能不提出警告：这个中性工具，也只有在了解到资讯工具阴暗面——双刃剑的特性，在权势独占这一工具时，可能出现的危险。有此认识，我们才能善于利用这了不起的工具，开创更好的未来，也防治不虞的灾害。

涂先生这部书，清楚地叙述了资讯时代对我们生活的影响与社会的控御力。他讨论的范围方方面面、极为广泛。我盼望有了这本书作为起头，还有很多对资讯工具有研究、也有心得的人，参加讨论，让我们更清楚地了解，这个21世纪正在坐长的新的知识工具。为此，我们要对涂先生致敬与致谢，因为他为华文世界提出一个重要的话题。

2012年4月8日 于匹兹堡

序言二　中国的雄心应该拓展到大数据领域

托马斯 · H · 达文波特[1]

无论是对中国政府，还是就中国的商业组织而言，《大数据》都是一本重要的书。大数据及其分析，将会在未来 10 年改变几乎每一个行业的业务功能。任何一个组织，如果早一点着手大数据的工作，都可以获得明显的竞争优势，正如早期在“小数据”时代脱颖而出的竞争者一样，如第一资本金融公司、前进保险公司、万豪酒店等等。时光荏苒，现在到了抓住大数据机遇的时候了。

大数据之所以产生，是因为今天无处不在的传感器和微处理器。我们正在迈进普适计算的时代。其实，所有的机械或电子设备都可以留下数据痕迹，这些痕迹表明了它的性能、位置或状态。这些设备和使用它的人，通过互联网互相交流，又形成了另外一个庞大的数据源。当这些数据和来自其他媒体、无线或有线电话、有线电视、卫星等等来源的数据相结合的时候，更加显得庞大无比。

这些数据可以被使用，这意味着我们可以把所有的商业或组织活动都视为大

1　托马斯 · H · 达文波特 (Thomas H. Davenport)，哈佛大学商学院访问教授、巴布森学院（Babson College）信息技术与管理学总统杰出奖教授，2003 年，他被《咨询》杂志评为全球“最优秀的 25 位咨询大师”之一，2005 年被《优化》杂志评为世界商业与技术分析顶级三强之一。

数据的问题。如今的制造业，大多数机器上都已经安装有一个或多个微处理器，已经进入了大数据的状态。消费营销行业，无数顾客的交易触点和网上点击的流量，也成了大数据的问题。谷歌甚至认为其无人驾驶汽车也是一个大数据的问题。

世界各国的政府也开始认识到，他们坐拥海量数据，这些数据都有待分析。在亚洲国家的政府，也出现了大数据战略以及基于数据分析的方案和倡议。去年，新加坡成立了德勤数据分析研究所（DAI），这个新的机构是由新加坡政府经济发展委员会资助成立的。德勤数据分析研究所的目标，就是引领政府和企业对于数据的研究和应用。新加坡政府还资助了几所大学开展大数据和数据分析的研究活动。

任何一个组织，要抓住大数据的机遇，就必须做好几个方面的工作。从技术角度来看，首先要收集并且开发特定的工具，来管理大规模并行服务器产生的结构化和非结构化数据，这些数据，可能是自己专有的，也可能来源于“云”。其次，每一个组织都需要选定分析软件，用它来挖掘数据的意义。但可能最重要的是，任何组织都需要人才来管理和分析大数据。这些人被称为“数据科学家”，他们集黑客和定量分析员的优势和特长于一身，非常短缺。聪明的领导人，将想方设法留住这类人才。

不少公司都意识到了这种难得的机遇，现在已经采取了行动。例如，通用电气将投资 15 亿美元在旧金山湾区建立一个全球软件和分析中心，作为其全球研发机构的一部分。这个中心拟雇用至少 400 名数据科学家，现在已经有 180 名各就其位了。通用电气在全球拥有超过 1 万名工程师从事软件开发和数据分析工作，通过共同的分析平台、训练、领导力培训以及创新，他们的努力得以协调合作。通用电气对于大数据的研究活动，相当一部分集中在工业产品上，例如机车、涡轮机、喷气发动机以及大型能源发电设施。

对任何一个试图通过大数据获得成功的组织来说，通用电气的投资规模和雄心都是一个榜样。在很多领域，中国政府和中国的企业都有雄心勃勃的计划，这引起了全世界的关注，这些雄心和计划，现在应该拓展到大数据的领域。涂子沛先生的这本书，将在这个重要的领域，为中国政府和企业的努力提供引导和帮助。

Foreword

[达文波特序言英文原文]

This book is an important one for Chinese government and business organizations. Big data and analytics based on it promise to change virtually every industry and business function over the next decade. Any organization that gets started early with big data can gain a significant competitive edge. Just as early analytical competitors in the "small data" era (including Capital One bank, Progressive Insurance, and Marriott hotels) moved out ahead of their competitors and built a sizable competitive edge, the time is now for firms to seize the big data opportunity.

The pervasive future of big data is enabled by the pervasive nature of sensors and microprocessors today. We are entering into the ubiquitous computing age now. Virtually every mechanical or electronic device can leave a trail that describes its performance, location, or state. These devices, and the people who use them, communicate through the Internet—which leads to another vast data source. When all these bits are combined with those from other media—wireless and wired telephony, cable, satellite, and so forth—the future of data appears even bigger.

The availability of all this data means that virtually every business or organizational activity can be viewed as a big data problem or initiative. Manufacturing, in which most machines already have one or more microprocessors, is already a big data situation. Consumer marketing, with myriad customer touchpoints and clickstreams, is already a big data problem. Google has even described the self-driving car as a big data problem.

Governments have begun to recognize that they sit on enormous collections of data

that wait to be analyzed. We can see big data and analytics initiatives among governments in Asia. Last year, Singapore helped to launch the Deloitte Analytics Institute (DAI). This new institute is sponsored in part by the Economic Development Board of the Singapore government. The DAI's goal is to do research and thought leadership on the application of analytics to government and business. Singapore has also sponsored several university-based research initiatives on analytics and big data.

Organizations that want to pursue big data opportunities need to begin working along several fronts. From a technology standpoint, they need to acquire and develop tools to manage both structured and unstructured data in massively parallel server environments, either on premise or in the cloud. They need to select analytical software to make sense of the data. Perhaps most importantly, they need to hire or develop the human talent to manage and analyze big data. These people are typically known as "data scientists"—hybrids of hacker and quantitative analyst—and they are in extremely short supply. The wise executive will develop approaches to securing the best people.

Some companies are beginning to realize the extent of the opportunity, and to act upon it now. GE, for example, has committed to spend more than $1.5 billion to develop its Global Software and Analytics Center in the San Francisco Bay Area as a part of its Global Research organization. The company plans to hire at least 400 computer and data scientists at this location, and has already hired 180. Globally GE has over 10,000 engineers engaged in developing software and analytics products and services, and their efforts will be coordinated through common analytics platforms, training and leadership education, and innovative offerings. A significant portion of big data activities at GE will be focused on industrial products, such as locomotives, turbines, jet engines, and large energy generation facilities.

The size and ambition of GE's commitment should set the tone for other organizations that want to succeed with big data. Chinese government agencies and firms are noted worldwide for their ambitious plans in other domains, and these should be extended to big data. Zipei Tu's book will help to guide government and business organization's efforts in this important area.

Thomas H. Davenport

目 录

序幕　新总统的第一天

你们每个人，都可以拿了毕业证、走下这个讲台，然后去追求锦衣玉食等等这个金钱社会视为理所当然的东西。你可以选择只关心自己的喜怒哀乐，把你的生活和国家的发展割裂开来。

但我不希望你这样做。这不仅仅是因为你对那些没有你幸运的人负有责任，尽管你确实负有责任；也不仅仅是因为你对帮助你走到今天的人欠有债，尽管你确实欠下了债。

这是因为：你对你自己负有使命和责任。这是因为：我们个人的命运依赖于群体的命运。这是因为：如果你仅仅考虑你自己、满足眼前的需要，这是一种贫乏。这是因为：只有你把你自己的战车和其他一些更伟大的东西绑定到一起的时候，你才能发现你真正的能量，才能发现你为美国这个国家继续书写历史时所能扮演的角色。[01]

——奥巴马，在卫斯理大学毕业典礼上的演讲，2008 年 5 月 27 日