DATA FOR THE PEOPLE

HOW TO MAKE OUR POST-PRIVACY ECONOMY WORK FOR YOU

大数据和我们

如何更好地从后隐私经济中获益？

[美] 安德雷斯·韦思岸（Andreas Weigend）著

胡小锐 李凯平 译

数据是未来的新石油

风靡斯坦福大学的社交数据革命课

亚马逊前首席科学家、大数据专家心血力作

简体中文版全球首发上市

中信出版集团 · CHINACITICPRESS

大数据和我们

如何更好地从后隐私经济中获益?

[美] 安德雷斯·韦思岸 (Andreas Weigend) 著
胡小锐 李凯平 译

DATA FOR THE PEOPLE

HOW TO MAKE OUR
POST-PRIVACY
ECONOMY WORK FOR YOU

中信出版集团 · CHINACITICPRESS · 北京

图书在版编目（CIP）数据

大数据和我们 /（美）安德雷斯·韦思岸著；胡小锐，李凯平译. -- 北京：中信出版社，2016.11
书名原文：Data for the People
ISBN 978-7-5086-6969-4

I. ①大… II. ①安…②胡…③李… III. ①经济管理－数据管理－通俗读物 IV. ①F2-39

中国版本图书馆CIP数据核字（2016）第 261031 号

大数据和我们

著　　者：[美]安德雷斯·韦思岸
译　　者：胡小锐　李凯平
策划推广：中信出版社（China CITIC Press）
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）
（CITIC Publishing Group）
承 印 者：北京诚信伟业印刷有限公司

开　　本：787mm×1092mm　1/16　　印　　张：18.25　　字　　数：220 千字
版　　次：2016 年 11 月第 1 版　　印　　次：2016 年 11 月第 1 次印刷
京权图字：01-2016-8336　　广告经营许可证：京朝工商广字第 8087 号
书　　号：ISBN 978-7-5086-6969-4
定　　价：59.00 元

当你的一切都被记录在案时

信息本身已成为世界上最大的一桩生意，人们对自己的了解还不如数据银行。数据银行记录的个人数据越多，我们就越缺少存在感。

——马歇尔·麦克卢汉（Marshall McLuhan）

1949 年，我父亲还是一个 23 岁的小伙子，在民主德国当老师。他刚到执教学校所在的小镇时，需要找一个能跟他共住一间宿舍的室友。幸运的是，他在火车站遇到了一个也在寻找住所的人，于是，他们成了室友。但他们搬进住所几天后，父亲就发现他的室友失踪了。这令他十分错愕，之后就开始为他的室友担心。

不久后的一天早晨，父亲正在吃早餐，这时响起了敲门声。也许是室友回来了！父亲打开门，发现来了几个陌生人，他们告诉父亲他获得了教学奖。这是一个十分特殊的奖项，需要亲自授予获奖者本人，他们是来陪我父亲一起去领奖的。父亲对颁奖邀请十分怀疑：当时的场面很奇怪，这些人都阴沉着脸，穿着统一的军用风衣。但是，父亲别无选择，

只能跟着他们坐进一辆轿车。上车后，他发现车窗无法从车内打开，他猛然意识到自己被苏联占领军逮捕了。

苏联人指控父亲是美国间谍，证据是他会讲英语。亲朋好友们都不知道父亲身处何地，仿佛他从人间蒸发了。父亲被关进苏联政府管理的一座荒僻的监狱中，在那里遭受了6年的折磨。他根本不知道自己为何被捕，也不知道自己为何获得释放。

透露个人信息会带来杀身之祸，因为数据可以被用来对付我们。实际上，一想起这样的危险，就让我不寒而栗、头脑清醒，因为我知道当时的苏联人是怎样收集数据并用它来对付我父亲的。

在我父亲被关押期间和出狱后，民主德国国家安全部（又称斯塔西）收集了他的很多信息。两德统一10年后，我请求查阅这些信息。并非只有我一人想知道斯塔西对我的家人所做的一切，柏林墙倒塌后，近300万人都要求查阅本人或其亲属的档案。不幸的是，负责公布斯塔西档案的部门来信告诉我，有关我父亲的所有档案似乎都已被销毁。

但是，这封信中附了一张照片，照片中是斯塔西为我建立的档案。我感到十分诧异，斯塔西竟然有我的档案？我当时只是一个物理学专业的学生。秘密警察早在1979年就开始收集我的信息了，那时候我还是一个懵懂少年。在我搬到美国之后的第二年，也就是1987年，斯塔西更新了我的档案。照片中的我的档案只剩下了一张封皮，我不知道斯塔西收集了我的哪些信息，他们为什么要这么做，要用这些信息做什么。

在斯塔西的势力猖獗之时，收集其“重点关注的公民”的信息是一件非常困难的工作。斯塔西会对这些人跟踪、拍照、截获他们的信件、走访他们的朋友，还会在他们家中安装窃听器，只有这样做才能收集到数据，之后再以纯手工方式对这些数据进行分析。需要收集的数据非常

多，在民主德国政府垮台时，全国劳动人口中有1%的人全职服务于秘密警察。但是，这些人还远远不够。德国联邦政府称，民主德国政府在临近垮台之时，有大约20万人为其收集信息。

图 0–1　斯塔西为我建立的档案封皮

如今，数据收集人员的工作则简单得多。让我们看几个知名案例。针对美国国家安全局对电话的监听行为，隐私保护活动家经过数月的抗议和法庭交锋后取得了一场小小的胜利。但即便知道自己通话的元数据会受到美国国家安全局或其他部门的监控，也很少有人会因此退订自己的手机服务。美国加利福尼亚州的一名女销售员提出索赔时称自己因卸载了一款手机应用而遭解雇，因为这款应用无论在上班时间还是闲暇时间都会跟踪定位她的地理位置，并分享给她的主管。新闻曝光脸谱网正在研究人与人之间的情绪影响时，公众愤怒地质问该公司是否在“操控”用户的情感。但是，用户们一如既往地使用脸谱网，而脸谱网也继续在未经用户许可的情况下进行实验，原因很简单，即实验是在线平台设计

的重要组成部分。2015 年，商业巨头阿里巴巴公司旗下的蚂蚁金服公司在中国推出了一项试点服务——芝麻信用，它通过分析个人交易数据来评估用户的信誉度，这类似于对用户在亚马逊上的购买记录进行评估，以判断该用户是否具备信贷资格。芝麻信用很快在其他领域得以运用，包括在中国一家知名交友网站中作为个人档案的选填项目之一而大受欢迎。没有人会呼吁我们停止使用手机、电子邮箱、导航软件、社交媒体账户、零售网站及其他数字化服务，因为它们让我们的生活更加方便。

我之所以热衷于隐私保护，是因为我发现斯塔西为我建立了档案吗？绝非如此。实际上，斯塔西档案与我每天自愿与他人分享的个人信息相比，不值一提。

从 2006 年开始，我将自己计划发表的每一场讲座和演讲，以及即将乘坐的每一趟航班信息都发布在我的个人网页上，甚至包括我的航班座位号。我这样做的原因是，我相信通过分享自己的数据所获得的实际价值要高于这样做的风险。数据带来了探索和优化的机会，所以关键问题在于要找到好办法，确保数据使用者的利益与我们的自身利益一致。

我们如何实现这一目标呢？我们可以了解我们分享的是什么数据（和不久的将来可能分享的数据），以及数据公司会如何分析并使用我们的数据。在充分尊重麦克卢汉观点的前提下，我要告诉大家的是，数据公司对我们每个人的数据记录得越多，我们的存在感就越强，我们对自己的了解也越透彻。真正的问题在于，如何确保数据公司对我们的透明性和我们对数据公司的透明性是对等的，还要确保我们对本人数据的使用方式有一定的主导权。本书将告诉我们如何才能实现这两个目标。

第 7 章 把未来创造出来

社交数据革命

如何确保数据会为我们服务?

每一场革命最初都是一个人头脑中的一种思想，一旦同一种思想在另一个人的头脑中出现，它对于这个时代就变得至关重要了。

——拉尔夫·沃尔多·爱默生（Ralph Waldo Emerson）

早晨6点45分，手机闹钟将我叫醒。于是，我拿起手机，一边浏览电子邮件与脸谱网信息，一边走进厨房，我美好的一天就此开始。手机上的全球定位系统应用软件会记录我的位置变化，并显示出我向东、向北移动了几米。我给自己倒了一杯咖啡，然后走出厨房。这时，手机上的加速计会给出我的行走速度，气压计会记录我何时上楼。由于我在手机上安装了谷歌的应用程序，因此谷歌公司拥有我的这些数据的记录。

吃完早饭后，我要去斯坦福大学上班。在我关灯并拔下移动设备的电源插头后，电力公司安装的“智能”电表就会知道我的用电量开始下降了。当我打开车库门时，电表会探测到与之相匹配的使用签名。当我

开车上路时，电力公司已拥有足够的数据断定我已不在家中。当我的手机从另一个基站接收信号时，通信公司也知道我出门了。

驾车行驶在路上时，如果我闯了红灯，安装在街道拐角处的摄像头就会拍下我的车牌号。谢天谢地，我今天遵纪守法，不会收到交通罚单。但在行驶过程中，我的车牌会多次被拍摄。有些摄像头属于当地政府，有些则属于私营公司，它们通过分析数据了解人们的驾驶习惯，并将此作为产品出售给警方、开发商及其他利益群体。

我到达斯坦福大学时，会使用手机上的“无忧停车”应用支付停车费。停车费自动记入我的银行账户，同时学校的停车管理小组会收到我的付款通知，这样一来，校方与我的开户银行都知道我在上午 9 点 03 分到达校园。由于我的手机不再以汽车的行驶速度移动，谷歌公司会推断出我已停车并记录下我的位置，以便我日后查询当时的位置记录。我也可以通过美国车险服务商 Metromile 公司的保险应用查询我当时所在的位置，这款应用通过我的车载诊断系统实时记录我的驾驶数据。这让我可以立刻发现今天的汽车燃油效率较低——每加仑[①]汽油行驶了 19 英里[②]，我此次通勤花了 2.05 美元。

上完课后，我打算和旧金山的新朋友见个面。我们在“虚拟世界”中见过面，当时我们共同的朋友在脸谱网上发了帖子，我们都对它进行了评论，也很赞赏对方的看法。之后，又发现我们在脸谱网上有 30 多个共同好友，所以我们确实应该见一面。

谷歌地图预计我将在晚上 7 点 12 分到达目的地。与往常一样，它的预测误差只有几分钟。这位朋友居住公寓的一层是一家销售烟草产品和

① 1 加仑≈ 3.8 升。——编者注

② 1 英里≈ 1.6 千米。——编者注

吸食大麻器具的商店，而我的智能手机上的全球定位系统应用软件无法区分公寓和商铺。我的车载导航与谷歌导航都告诉我，我今天晚上去了一趟毒品商店——这是我上床前查阅第二天的天气预报时，谷歌广告推送告诉我的。

这不只是一场社交数据革命。

将欲取之，必先予之

每天都有 10 多亿人像我这样产生和分享社交数据。社交数据是有关你本人的信息，例如你的运动、行为、兴趣，以及你和其他人、地点、产品，甚至意识形态之间的关系。其中有些数据是在你本人知情的前提下自愿分享的，例如在使用谷歌地图时登录并键入目的地；其他数据则并非如此，你经常会在不经意间就分享了自己的数据，这是享受互联网与移动设备所带来的便捷性过程的重要部分。显然，在某些情况下，分享数据是你获取服务的必要条件：如果你不向应用软件提供你当前所在的位置和目的地，谷歌公司就无法为你找出最佳的行车路线。在某些情况下，你可能很乐意提供信息，例如你给某个朋友在脸谱网上的发帖点赞或在领英网上对同事的工作表示肯定，以表明你愿意以某种方式鼓励和支持他。

社交数据有时可以做到比较精准，能将你的位置精确到 1 米之内。但是，在通常情况下，社交数据都很粗略，有时也不够完整。例如，除非我登录可以显示家中智能电表读数的某个应用（比如，为了查看我在去机场之前是否将家中所有的灯都关上了），电力公司才能知道我何时离家，但也仅限于此。这种数据过于粗略，也许对我没有太大的帮助。与

此相似，我在拜访旧金山的那位新朋友时，虽然社交数据可以准确地显示出我所在位置的经度和纬度，但对我当晚活动的推测却是完全错误的。有时候，虽然数据看似十分精确，但在很大程度上这是数据解读的结果。实际上，社交数据本身是非常粗略的。粗略的数据很可能不完整、易出错，有时其中还会掺杂欺诈数据。

无论是被动还是主动分享的数据、强制还是自愿分享的数据、精确还是粗略的数据，社交数据的总量呈指数增长趋势。如今，社交数据总量翻一番所需的时间只有 18 个月。在未来 5 年内，社交数据总量将增长约 10 倍，或者说增长一个数量级；在未来 10 年内，社交数据总量将增长约 100 倍。换言之，2000 年全年产生的数据总量目前只需要 1 天即可完成。以这样的增长速度计算，预计到 2020 年，不到 1 个小时就能产生等量的数据。

要知道，“社交数据”并非仅适用于社交媒体的流行词汇，这一点很关键。许多社交媒体平台的设计旨在进行播报，以推特为例，沟通几乎总是单向进行的，由名人、权威人士或营销人士向公众传播信息。社交数据更加民主化，你可以通过推特或脸谱网分享你的信息、所在公司的信息、你的成果、你的看法，但你的电子踪迹比这些更深远。根据你在谷歌网站上的搜索记录、你在亚马逊网站上的购买记录、你在讯佳普（Skype）上的通话记录、你手机的实时定位，再将这些信息与其他多种渠道相结合，就能得出有关某个人的一幅独特的“肖像画”。

此外，社交数据不会止于你本人。在你展示自己通过与亲朋好友、工作同事的沟通建立起的亲密关系时，你便创建并分享了数据。你所创建的社交数据不仅涉及友人，也会涉及陌生人，例如你在评价某件商品或在照片墙（Instagram）上传照片时。空中食宿（Airbnb）是一个租用

房间或套房的应用平台，你若要注册账户就需要验证身份——不仅要使用政府核发的身份证，还要使用你的脸谱网账户。社交数据正在嵌入你家中的智能温度计、汽车的导航系统以及职场的办公软件，并开始成为教室与医院诊疗室中的亮点。随着手机配备了越来越多的传感器和应用，它们可在我们的家中、商场或单位里跟踪我们的一举一动。你将越来越难以掌控有关你日常活动的数据，甚至包括你内心中最隐秘的愿望。数据科学家将化身为侦探与艺术家，通过人们留下的电子踪迹为他们绘制出越发清晰的行为素描画。

通过检查并提炼这些电子踪迹，可以发现人们的偏好或倾向，还能做出预测，例如人们可能会购买何种商品。在我担任亚马逊公司首席科学家期间，我与杰夫·贝索斯共同制定了该公司的数据战略和以客户为中心的文化。我们开展了一系列实验，比较网站编辑或消费者所写的商品评论中哪一种会让客户更开心，并观察依据传统的人口统计信息或个人点击情况为客户做推荐是否成功率更高。在举办厂商赞助的促销活动时，我们发现真正的沟通可以爆发出巨大的力量。我们为亚马逊开发个性化工具，使人们做出购买决定的过程及所购买的商品都产生了根本性改变，并且成为电子商务的标准。

离开亚马逊之后，我在斯坦福大学和加利福尼亚大学伯克利分校为成千上万的本科生和研究生开设了社交数据革命课程，还在中国上海的复旦大学与中欧国际商学院、北京的清华大学教授这门课程。我同时继续经营社交数据实验室，成员是我在 2011 年结识的一群数据科研人员与思想领袖。在过去 10 年里，与我合作的公司包括阿里巴巴、美国电话电报公司、沃尔玛、美国联合健康保险集团，以及一些大型航空公司、金融服务公司、交友网站。我积极倡导把数据的决策权与客户或用户分享，

他们是与你我一样的普通人。

没有人能够独自处理当下的所有数据并做出明智的决定。但在让数据服务于我们的需要和解决问题的过程中，谁能够获得必要的工具呢？从这些数据中分析得出人们的偏好、倾向和做出预测后，是将其提供给少数强大的组织，还是提供给所有人使用呢？使用社交数据所需支付的费用是多少呢？

随着我们逐渐认识到社交数据的价值，我相信我们的重点不仅是获取数据，还必须采取某些行动。我们每天都会做出很多决定，而有些决定一生中只会做一次。但是，这并不意味着今天产生的社交数据的寿命很短。我们今天的行为方式可能会影响我们今后几十年的选择，很少有人能在短期或长期内观察到自己的所有行为或分析出这些行为将如何影响自己。社交数据分析有助于我们找出各种可能性，但必须经过深思熟虑方可做出最终选择。

毕竟，这些科技无法了解我们每个人乃至整个社会对未来生活的憧憬。许多国家都出台了法律，保护个人在就业或医疗方面不受歧视。未来某一天，这些法律或许将不复存在（在某些国家，直到现在也没有这样的法律）。假设你希望获得有关减肥和锻炼的建议，于是你决定在医疗应用或网站上表达自己对胆固醇过高的担心。这样做会不会对你不利呢？如果法律规定，在医生向你告知健康风险并推荐健康的生活方式之后，你仍然不愿意放弃吃油炸食品，依旧喜欢瘫坐在沙发上，就可以依法对你收取更高的医疗费用，你怎么办？如果你的主管利用某种服务软件在网上查找有关你的信息，他可能认定你的生活方式不适合在他的公司任职，从而拒绝考虑你的求职申请，你怎么办？这些都是实实在在的风险。

如果这些数据是你独立创建并透露出去的，那么，一旦察觉到风险，你或许可以停止这种行为。这会给你带来许多不便，却是可行的。但是，人们对有关自己的许多数据并没有掌控力。由于社交数据被公司和政府用于改善结果、提高效率，因此我们更不可能掌控这些数据。

社交数据关乎社会大众，我们每个人都需要考虑怎样做才是最好的数据利用方式。科技正在飞速发展，收集和分析数据的公司主要从事信息的产出与编码，并不负责制定原则。即使它们考虑那些原则性问题，也仅仅是因为业务需要而临时为之。对人类未来会产生重大影响的原则性问题的决定权，绝不应该交到数据公司手中。

我们可以允许对所有这些数据进行收集、合并、汇聚、分析，以便能在决策过程中更好地做出取舍。取舍是任何重要决策的必要组成部分，在做取舍时，人的判断十分关键。我们的生活不应由数据来驱动，而应让数据为我们的生活服务。

后隐私时代的原则

我们已经认识到数据在生活中发挥着越来越重要的作用，也已经采取了许多措施保护自身的利益。20 世纪 70 年代，美国与欧洲针对信息的公平使用采取了大体相似的原则。人们有权知道谁在收集自己的数据以及这些数据的使用情况，当发现数据不准确时，还可以要求修正数据。然而，对于今天的新型数据来源与分析方法，这些保护措施要么过于严厉，要么过于无力。

之所以说它们过于严厉，是因为这些措施都想当然地认为可以对收集到的所有数据添加标签。亚马逊公司可能会以浅显易懂的术语，准确