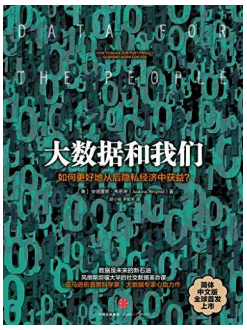


**您为此书添加的 KINDLE 笔记：**

# 大数据和我们

作者：安德雷斯・韦思岸、 胡小锐、 李凯平

免费 Kindle 极速预览：http://z.cn/7B8CtGA

## 70 条标注 | 2 条笔记

标注（黄） | 位置 121

> 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

标注（黄） | 位置 248

> 无论何时，只要我听到客户服务代表说通话可能会被录音时，我就会对他说，我也可能会对此次通话录音，以保证我所获得的服务质量。

标注（黄） | 位置 269

> 如果我们能促使数据公司同意提供一系列有意义的权利与工具，就能产生我所说的“关系反转”，即对个人与机构之间的传统关系予以逆转。

---

标注（黄） | 位置 284

> 信息是权力的中心。

---

标注（黄）和笔记 | 位置 329

> 作为21世纪最重要的原材料，数据就是新石油。一个多世纪以来，油田的发现与石油开采技术的进步对经济与社会产生了深远的影响。人们通过提取、储存和精炼等环节，把石油变成为人类服务的各种产品。现在，由原始数据转变而成的产品和服务，正在改变我们的生活，其影响力足以与工业革命相媲美

**数据是信息革命的石油**

---

标注（黄）和笔记 | 位置 350

> 曾任普林斯顿大学心理学教授的乔治·米勒在其作品中谈到判断一个人是否有文化的现代标准。他为众多毕业生在阅读、数学和科学等方面的素养不够而深感不安，担心他们在一个知识产业占主导地位的经济体里找不到工作。如今，我认为人们同样迫切需要培养一种新素养——数据素养，

**有文化的现代素养:阅读,数学,科学,数据**

---

标注（黄） | 位置 406

> 就其本身而言，数据存储算不上革命性举措。令亚马逊脱颖而出的是，它始终致力于数据挖掘，根据顾客的兴趣、偏好与当前状况向他们推荐商品

---

标注（黄） | 位置 411

> 亚马逊利用顾客与网站互动时产生的所有数据，改变了营销活动。它还通过商品评论，赋予了顾客创造数据的权利

---

标注（黄） | 位置 416

> 亚马逊裁撤了网站编辑人员，安排人手开发算法，并将最有用的顾客评论展示在产品页面的顶端。加大对技术与数据的投资以改进顾客的购物体验，这种做法的效果优于在综合管理上增加投入

---

标注（黄） | 位置 418

> 亚马逊的数据挖掘改变了10亿人的购物习惯。2015年，美国零售业有近半数的购买活动是从登录亚马逊网站搜索产品开始的（不管这名顾客最终是在哪里购买了该产品）

---

标注（黄） | 位置 477

> 培养数据素养，你还需要知道何时你的数据独立存在，何时与总体数据融为一体

---

标注（黄） | 位置 489

> 拉尼尔笔下的那些做出贡献的人，其实已经得到了回报。这些人很有可能也使用了谷歌的文本翻译服务，因此他们已经得到了回报。只不过他们得到的不是金钱，而是更完善的数据产品和服务

---

标注（黄） | 位置 516

> 社交数据可以为你的决策提供帮助，而不只是帮助某个特大企业发起一场目标更明确的广告运动。我相信，你创建的数据必然符合你的特点，你做出的决策也必然符合你的特点。这就是你的数据的价值

---

标注（黄） | 位置 629

> 微软在这个领域有一个名叫英瑞克斯（Inrix）的子公司。该公司每天分析一亿多部手机的地理位置数据，了解它们的目的地（更重要的是，了解它们不会去哪些地方），以推断人与数字产品的运动趋势。英瑞克斯从通信运营商那里获取数据，了解这一亿部手机正在连接哪些蜂窝基站。英瑞克斯完成数据挖掘后，就会将其出售给那些为驾驶员提供导航和路线规划服务的公司，包括佳明（Garmin）、MapQuest网站、福特、宝马等。英瑞克斯还在城市规划方面为政府出谋划策，为修建新桥、加装红绿灯、建设新的公立医院等公共设施选择合适的地点

---

标注（黄） | 位置 716

> 法国亚马逊的开发人员出于某个原因，忘记在结账时计入货运成本。这个错误发生后，在短时间内订单数出现“井喷”现象，这也给了亚马逊一个灵感：商品包邮可以增加销量

---

标注（黄） | 位置 728

> 道理：“获取数据并不难，但获取可靠的数据却非常难。”我由衷地赞同这个观点。这句话也可以套用到数据挖掘上：提供推荐意见并不难，但评估推荐意见的好坏却非常难

---

标注（黄） | 位置 775

> 我上学时学的是物理学，现在从事社交数据实验的很多人也是如此。这并不奇怪，因为我们浏览网页、使用手机等行为留下来的数字痕迹与粒子探测器捕捉到的路径及计数非常相似。事实上，从事实验性粒子物理研究的经历为人类完成电子商务实验奠定了坚实的基础

---

标注（黄） | 位置 969

> 想要亚马逊送货上门的话，你就必须提供个人信息，包括姓名、送货地址等。提供正确的地址对你有利，否则你就无法收到包裹。不过，购买记录既包含你买给自己的商品，还有你买给他人的商品。如果你将某件商品标记为礼品，亚马逊在为你做产品推荐时就不会考虑这件商品。利用这些数据，个性化算

> 法在处理你备注为他人购买的商品时，就会将它与你购买的其他商品区分开。如果你为某位女性买礼物，那么你在选择衬衫尺寸时就会公开她的体型数据。如果你买这件衬衫的时间是在母亲节之前的一两周，而且收件人的姓氏与你相同，亚马逊的算法就会推断出你们之间的关系。一年后，亚马逊甚至有可能给你发电子邮件，推荐适合你的母亲节礼物

---

标注（黄） | 位置 1001

> 研究小组正在开发其他软件，用于分析照片背景和具体环境，以判断你是置身于熙熙攘攘的酒吧还是待在荒无人烟的山顶。如果你出现在其中一个场合中的次数更多，算法就可能把你归到社交蝴蝶或者孤胆探险家一类

---

标注（黄） | 位置 1049

> 此外，游戏网站据说可以根据鼠标移动特点推测孩子的年龄，而且误差仅为3~6个月。对于10岁以下或刚刚10岁的孩子而言，他们的鼠标移动特点与其运动技能的发展情况密切相关

---

标注（黄） | 位置 1121

> 能够更好地反映人们对哪一类电影感兴趣的真实信号，是他们在线观看这部电影的实际时长。换言之，对于形成推荐意见而言，检视数据比评论数据更有帮助

---

标注（黄） | 位置 1387

> 斯坦福大学的社会学家马克·格兰诺维特（Mark Granovetter）研究过人际关系中联系纽带的强度。在他于1973年发表的开创性论文“弱关系的强度”中，他对关系的强度进行了定义，即时间长短、情感强度、亲密性（相互信任的程度），以及相互帮助的程度等的综合体。网络中人们交换的不仅是情感和信息，还有影响力和帮助

---

标注（黄） | 位置 1482

> 埃伦说：“最不适合建立人际关系的时间是在你有所求的时候，需求会把交友变成交易，这两者根本不是一回事儿。建立人际关系的最好办法是帮助别人，而且没有任何不可告人的动机。

---

标注（黄） | 位置 1485

> 要从用户那里获取数据，领英就必须先给用户提供数据

---

标注（黄） | 位置 1525

> 脸谱网让我们看到朋友的点赞和评论，却不让我们知道谁看了我们的照片，尽管它有这方面的数据。脸谱网上的好友可以下载我上传的任何照片，而我却一无所知。我希望领英网资料浏览方面的对等性能被更多的数据服务商采纳，同时这种对等性可以应用到更多类型的内容上

---

标注（黄） | 位置 1739

> 空中食宿利用用户在网站上创建的数据（例如用户搜索、评分、评论、交流记录等反馈信息）与外部数据，核实用户身份，评估他们的可信度

---

标注（黄） | 位置 1853

> 这与BioCatch公司根据键盘使用和鼠标运动规律创建“操作指纹”用于鉴定用户身份的做法如出一辙

---

标注（黄） | 位置 1854

> 脸谱网的设备指纹是根据多个数据源创建的，包括操作系统的语言设定、已安装应用程序的清单和用户的联系人名单（如果用户同意脸谱网获取这些数据）。脸谱网创建设备指纹的主要用途是保护用户的账户安全

---

标注（黄） | 位置 2024

> 他是在市区游荡，还是待在自己的卧室里？数据服务商可以通过他近期和当下的地理位置数据，针对性地为他提供搜索结果，确定他希望前往的目的地

---

标注（黄） | 位置 2072

> 还有一种完全不同的数据源，可以透露你曾经的位置信息，那就是你拍摄的照片和你被拍到的照片。首先，在网上公开发布的照片大多是用手机拍摄的，而大多数有照相功能的手机都有GPS。照片默认关联的元数据包括照片拍摄地点的经纬度，尽管你可以把这些元数据从你自己拍摄的照片中删除，但对于其他人拍摄的照片，你就无能为力了。每天，人们都会拍摄不计其数的照片，因此你的位置很有可能被记录下来

---

标注（黄） | 位置 2120

> Vigilant公司与其他私有汽车牌照数据库的主要客户是执法机构，但从理论上讲，任何人都可以接受Vigilant公司的付费服务。公司的委托人包括希望收回汽车的汽车经销商和希望了解事故详情的保险公司，私家侦探也会使用这些数据库。对你的亲朋好友而言，你的车牌号不是秘密，他只需看看车库即可了解这个信息。如果你怀疑自己的配偶是否真的每晚都在办公室加班，雇用侦探跟踪他的做法已经过时了。最有效的做法是将他的车牌号输入数据库，查询他的汽车去过哪里，以及是否有其他汽车也到过那里。如果你还想知道那些汽车的驾驶者是谁，数据库对汽车经常停放的位置都了如指掌，很可能据此找出驾驶者的居住地点与工作场所。一度成本昂贵，甚至具有危险性的数据收集工作，现在已经毫无危险可言，费用也大幅下降

---

标注（黄） | 位置 2137

> 例如苹果的Siri（语音控制功能）和微软的Cortana（微软小娜），还可以利用周围环境中的噪声分析你所在环境的特征

---

标注（黄） | 位置 2183

> 麻省理工学院教授威廉·T·弗里曼（William T. Freeman）带领同事研发的算法，可以侦测到皮肤颜色的像素级的微小变化，进而测量人们的脉搏，包括血液在人体裸露部位的分布情况。弗里曼团队演示了他们的研究成果之后，人们纷至沓来，要求他们提供这套算法。于是，他们将算法发布到网上，允许任何人用于非商业用途。一位扑克牌玩家想利用这个算法来侦测对手玩牌时是否心跳加速，以便判断他是不是在虚张声势

---

标注（黄） | 位置 2296

> 奥赛罗犯下的令人扼腕的错误说明了一个事实：检测某种情感的生理指标比较容易，而发现其背后的原因却难得多。在利用情感数据进行决策时，无论解读这些数据的是人还是机器，都必须时刻牢记奥赛罗的教训

---

标注（黄） | 位置 2319

> 在为美国国家航空航天局优化信息显示系统的过程中，埃里克·霍尔维茨和他的同事碰到了一个难题——认知负荷。所谓认知负荷，是指处理信息和解决问题时需要付出的注意力。根据丹尼尔·卡尼曼和杰克逊·比提（Jackson Beatty）的研究，瞳孔的相对直径可以反映人在完成任务过程中的认知负荷的大小。在接收新信息（例如，倾听一连串数字）时，瞳孔会放大；在报告这些信息时，瞳孔会收缩。任务的难度越大，瞳孔的变化越明显

---

标注（黄） | 位置 2335

> 不过，包括瑞士Tobii（心拓英启科技公司）在内的几个企业，成功地研发出眼球运动追踪专用设备和软件，捕捉并分析这些微小扫视运动。这些设备通常是用发光二极管（LED）发射的红外线，将设计好的图案投射到实验对象的眼睛上。尽管人眼看不见红外线，但红外摄像头可以监测到视网膜上的反射光，从而推断出眼球所在的位置和运动方向

---

标注（黄） | 位置 2362

> 在未来的几年里，机器的凝视控制系统的精准程度将不断提高，人机交互必将再次发生翻天覆地的变化

---

标注（黄） | 位置 2399

> 因此，从交谈时人们迁就某个人说话模式的情况就可以看出这个人的影响力。说话流畅者（几乎没有“嗯”、“啊”等口头语或停顿，也很少被人打断）常常是那些被奉为专家的人

---

标注（黄） | 位置 2473

> 并非所有能够量化的东西都很重要，并非所有重要的东西都能量化。——威廉·布鲁斯·卡梅隆（William Bruce Cameron

---

kindle

标注（黄） | 位置 2476

> 我们会不断地创建数据，大多数人也会不断地分享数据。想要数据服务于我们，重点不是对我们的数据实施控制，而是要求数据服务商在控制台上为我们留下一席之地

---

标注（黄） | 位置 2483

> 掌握了有关自己的特征、人脉关系、环境的所有数据之后，无论你是否清楚你自己的愿望是什么，数据服务商都能越来越准确地发现你的愿望

---

标注（黄） | 位置 2492

> 提高数据挖掘过程透明性的两项权利： 1. 访问自己数据的权利。 2. 检查数据挖掘过程的权利，包括 a. 查看数据安全审计的权利。 b. 查看隐私权效率评级的权利。 c. 查看数据回报评分的权利。 提高用户主动性的4项权利： 3. 修正数据的权利。 4. 对数据进行模糊处理的权利。 5. 开展数据挖掘实验的权利。 6. 自主导入和导出数据的权

---

标注（黄） | 位置 2508

> 我们需要深入了解用户所需的两种透明性：访问自己数据的权利与检查数据挖掘过程的权利。第一项权利有助于用户查看与解释他自己的个人数据。第二项权利能使数据挖掘过程更加透明，便于用户发现数据服务商处理与使用数据的特点

---

标注（黄） | 位置 2582

> 在行使访问你自己数据的权利时，无论数据源自何处，你都应该能够访问与你有关的所有数据

---

标注（黄） | 位置 2605

> 我们如何确定数据服务商对我们数据的挖掘，能给我们带来不错的回报且风险不大。为实现全面的透明性，你不仅需要拥有访问自己数据的权利，还应该拥有检查数据挖掘过程的权利

---

标注（黄） | 位置 2643

> 马克·古德曼（Marc Goodman）是《未来犯罪》的作者，也是联合国、北约组织、国际刑警组织的顾问，他强调安全漏洞不应当被视为发生概率极低的事件。15%~20%的世界国内生产总值（GDP）涉及有组织的犯罪活动、贩毒、人口贩卖与卖淫、窃取交易数据和侵害知识产权，互联网为这些犯罪活动提供了巨大的机会

---

标注（黄） | 位置 2752

> 隐私权是一种资源，在用用户数据创造产出的过程中被数据服务商消费

---

标注（黄） | 位置 2777

> 我们在分析备注的内容之前已将所有的用户名删除了。因此，在改进该网站服务的过程中，我们降低了隐私的消费量。我们团队中的任何人都无须知道某个用户的具体偏好，只要知道其备注的模式，并思考如何在网站中增加新的功能即可

---

标注（黄） | 位置 2815

> 为鼓励人们减少对碳的使用量，某些国家给其境内的公司规定了年度碳排放量额度。如果该公司的实际碳排放量低于其排放指标时，即可将未使用的碳排放量额度出售给碳排放量大的公司。如果某家公司的碳排放量过大且无法买到碳排放量额度时，就得支付罚款。这提高了公司的生产成本，迫使其减少碳的使用量或提供比竞争对手好得多的产品，让买方甘愿支付更多的费用和环保成本。组织与个人也可以主动为环保活动捐款，以此抵消自己的碳足迹

---

标注（黄） | 位置 2930

> 有4项主动性权利至关重要，分别是修正数据的权利、对数据进行模糊处理的权利、利用数据开展实验的权利，自主导入和导出数据的权利

---

标注（黄） | 位置 2963

> 服务商从4个方面赋予用户掌控力：修正数据的权利，对数据进行模糊处理的权利，运用数据开展实验的权利，自主导入和导出数据的权利

---

标注（黄） | 位置 3034

> 区块链概念是为虚拟货币比特币所开发的数字化分账系统

---

标注（黄） | 位置 3075

> 数据素养要求你懂得所提供数据的合适精度水平，以便从数据服务商那里获取你想要的产品或服务

---

标注（黄） | 位置 3085

> 拥有模糊处理数据的权利，还有助于人们在商业背景下更好地掌控自己的数据

---

标注（黄） | 位置 3104

> 亚马逊的金读电子阅读器（Kindle）记录了人们阅读活动的起始页码与结束页码，以及每读一页所花的时间。即便这些数据有助于老师对让学生感到困惑的课程内容进行个性化设计，学生也很有可能不希望老师看到这些数据，这取决于这些数据对他的成绩的影响程度。假设你决定将高精度的阅读数据分享给亚马逊或其他图书推荐网站，以获得个性化的图书推荐，但之后却有可能发现美国联邦调查局的特工出现在你家门前，因为你曾花大量时间阅读的一本书讲述的是波士顿马拉松炸弹袭击案犯是如何制造高压锅炸弹的。这种设想与实际情况十分接近

---

标注（黄） | 位置 3134

> 修正数据的权利允许用户自由地表达意见，对数据进行模糊处理的权利允许用户自行做出决定，利用数据开展实验的权利是探索的权利，它允许用户对种种可能性进行探索

---

标注（黄） | 位置 3201

> 与修正数据的权利、对数据进行模糊处理的权利、对数据开展实验的权利一样，自主导入和导出数据的权利也旨在提高用户的主动性

---

标注（黄） | 位置 3265

> 如果能将信誉数据、交易数据与其他数据从一家公司复制到另一家公司，就能帮按需经济的劳动者提高谈判筹码。自主导入与导出数据的权利可以确保用户的信誉始终跟随着用户，就像现实世界中的信誉一样

---

标注（黄） | 位置 3273

> 从用户的角度看，导入、导出数据的权利确保用户的数据不会受到某个数据服务商的绑架，即便某些数据服务商能提供此功能，用户也能找到提供此功能的其他数据服务商

---

标注（黄） | 位置 3275

> 1 000年以来，人们一直在努力争取人身自由迁徙的权利。我们现在还必须努力争取数据自由迁徙的权利，在这场数据革命中，流动性是实现人的主动性的关键

---

标注（黄） | 位置 3286

> 1978年，博世公司生产出第一款标准的防锁死刹车系统，并将其用于梅赛德斯－奔驰与宝马的顶级车型

---

标注（黄） | 位置 3325

> 人们在购物时，通常会对需要购买的产品与服务的价格、规格、评级、用户评价等货比三家。社交数据极大地减少了传统的信息不对称程度，而且客户购买模式的透明化也会改善人们的购物决策过程

---

标注（黄） | 位置 3419

> 公司利用5类以上的信息对米格尔进行金融信用评分，从而为缺乏信贷历史的人群更好地评估信贷风险。这些信息包括网络浏览行为、脸谱网与推特上的活动、手机呼叫与短信发送频率，以及手机的操作系统。该公司还考察贷款申请人在GitHub等在线社区中的活跃程度

---

标注（黄） | 位置 3426

> 公司并不完全依据对你的当前收入与开支情况的评估发放贷款，还要审核你所上的大学、所选专业、所学课程、成绩、高考分数，预测你在未来几年内的加薪趋势，从而计算出你偿还贷款的能力

---

kindle

标注（黄） | 位置 3438

> 人们在聚餐时，支付宝应用可以提供对应的支付选择，即“AA制”。这使阿里巴巴公司获得了真实世界的信息，不仅知道人们点了哪些食物，而且知道他们与谁一起用餐，以此计算芝麻信用得分

---

标注（黄） | 位置 3440

> 交易数据与社交图片数据对金融机构做出是否批准信贷申请的决定越来越重要，因此人们也应有权访问这些数据，并了解它们会如何影响你的信用分，就像了解是否按时还款在金融信用评分中所占的百分比一样

---

标注（黄） | 位置 3514

> 领英网的数据科学家注意到该网站在2008年9月14日（当天是周日）十分活跃。由于这在周末属于异常情况，他们担心网站受到黑客攻击。他们召集了安全团队，经过调查后，他们找到了数据流量的源头。所有这些行为都来自雷曼兄弟公司的雇员，他们疯狂地联系他人、更新简历、下载联系人信息。领英网怀疑这表明雷曼兄弟公司避免破产的努力宣告失败，但当时新闻尚未公开予以确认

---

标注（黄） | 位置 3602

> 马祖尔意识到，当他的学生通力合作、共同研究某个问题时，与以往他“一言堂”的授课方式相比，学生们学到了更多的知识，学习能力也变得更强。这促使马祖尔发明在线教育系统——基于云端的学习分析与管理系统Learning Catalytics

---

标注（黄） | 位置 3650

> 由于资源有限，教师经常不得不将自己的主要注意力放在某些学生的身上。他们应当将重点放在分布曲线左侧低于平均成绩的学生、分布曲线右侧高于平均成绩的学生，还是放在分布曲线中间的学生呢？我父亲在教高中理科课程时，对低于平均成绩的学生十分重视，避免他们中途辍学。我在大学教书时，优等生对我的教学最满意，因为我能激发他们的灵感，让他们创意迸发。某些教师则将重点放在提高中等生的成绩上，因为他们通常占到班级学生的大多数

---

标注（黄） | 位置 3798

> 这有点儿像自动化公司Vigilant行车记录仪组成的网络，它可以收集整个国家的车辆牌照信息

---