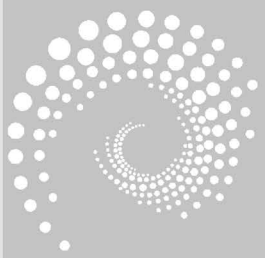
第3章

数据安全与数据伦理

CHAPTER 3



本章要点

数据自古就存在，在互联网出现并普及之后，由于数字化被记录、积累，而成为可供计算机快速提取和分析的大数据。近几年来，它被广泛地运用于人类社会的生产、生活、管理和社会治理，成为并列于资本、劳动和自然资源等新的生产要素。当前，大数据、数字经济等高频词语已成为世界各国推动经济社会可持续发展的着力点和热点。然而，当大数据让我们对世界事物的发展变化产生新的认知，人类在频繁地分析数据和挖掘数据价值的同时，伦理问题也逐步开始浮现。

哈佛大学研究显示，只需知道一个人的年龄、性别和邮编，就可以在公开的数据库中识别出此人87%的身份信息。在模拟小数据时代，一般只有政府机构才能掌握个人数据，而如今许多企业、社会组织也拥有海量数据，甚至在某些方面超过政府，这些海量数据的汇集增加了敏感数据暴露的可能性，大数据的收集、处理、保存不当更是会增加数据信息泄露的风险。

2016年，山东考生徐玉玉因被不法分子通过诈骗电话的手段骗取上大学的费用9900元，伤心欲绝，最终导致心脏骤停，经医院抢救无效不幸离世。经查，犯罪嫌疑人通过数据黑市购买了五万余条山东省2016年高考考生信息，并冒充教育局工作人员，以发放助学金名义对高考录取生实施“精准诈骗”，由此庞大的数据黑市得以浮现在公众视野。

2018年，全球最大的社交网站Facebook被曝出负面新闻：“一个名不见经传的小公司，通过不正当手段，在Facebook网站上获取了8700万用户的数据。这些数据随后被用于多个国家选举中的选民分析，美国总统特朗普就曾经雇佣这家公司，这引发了选民排山倒海式的质疑，直逼特朗普当选的合法性。”这场风波揭开了一个序幕，数据问题不仅仅涉及隐私问题，它还包含数据权益，这既是一个法律问题，也是一个公益问题，只是人们往往没有认识到这是一种权益。

无独有偶，2018 年初，中国大数据公司也被曝出一系列负面新闻。一位中国网友在微博上讲述了自己遭遇大数据“宰客”的经历。他经常通过某旅行网站预订某酒店的房间，价格常年为 380 到 400 元。偶然一次，酒店前台告诉他淡季价格为 300 元左右。他用朋友的账号查询后发现，果然是 300 元，但用自己的账号去查，还是 380 元。该微博引发了网友的强烈反响。据不完全统计，包括滴滴出行、携程、飞猪、美团、京东在内的多家互联网平台均被曝出存在“杀熟”的情况。阿里巴巴和腾讯的一些产品，如“淘宝”“微粒贷”，都将精细化和个性化做到了登峰造极，“千人千价”也深刻改变了商家和消费者之间的关系。

在大数据时代，除了众多互联网商业创新企业依赖于这些涉及个人网络行为的海量数据，国家信息安全也同样面临着严峻考验。2013 年 6 月，前美国中情局职员斯诺登将美国国家安全局一份关于“棱镜计划”的秘密文档披露给了卫报和华盛顿邮报，在国际社会引起巨大反响，称为“棱镜门”事件。该计划是一项由美国国家安全局和联邦调查局从 2007 年起开始实施的绝密电子监听计划。计划由美国情报人员实施，开展全球范围内监听活动，美国对全世界重点地区、部门、公司，甚至个人进行布控。监控范围包括信息发布、电子邮件、即时聊天信息，以及一些视频、图片、音频、备份数据、文件传输、视频会议、登录和离线时间、社交网络资料等细节，以及部门和个人的联系方式与活动等。

通过棱镜项目，美国国家安全局在未告知更不可能告知本人的情况下，一天可以获得 50 亿人次的通话记录，监听着本应属于公民个人隐私的海量网络活动。由此，情报部门既可以直接获取公民的通话和网络活动具体内容，也可以借助先进的大数据分析推断出个人性格、习惯、爱好、犯罪倾向等信息。随着“棱镜计划”细节逐渐曝光，国际舆论和许多国家政府都公开反对棱镜计划。同时，“棱镜门事件”将大数据热潮中如何尊重与保护个人隐私的伦理问题，清晰地摆在公众面前。网络信息安全受到前所未有的关注，这个事件也深刻影响网络时代的国家战略规划。

上述案例都揭示了大数据创新正面临着诸多伦理问题，除了隐私权的保护外，值得我们进一步思考和探究的问题还有很多，本章将逐一进行讨论和揭示。


微课视频

3.1 大数据时代的伦理问题

3.1.1 数据和大数据

1. 数据

数据是指对客观事件进行记录并且是可以鉴别的符号。它是对客观事物的性质、状态以及关系等进行记载的物理符号，或者是这些物理符号的组合，是可识别的抽象的符号。

数据的类型有很多种，比如数字、文字、图像、音频和视频等。数据这个概念自古有之，古人用“结绳计数”记录数量，用文字记录历史，这些统称为数据。这些符号和文字是人类早期对自身活动的记录方式。到了 16 世纪前后，人类开启了大航海时代，数据出现了一个高峰。随着航海仪器的普及，欧洲对土地测量、建筑设计、矿山开采、人口统计的需求也应运而生。人类发现，只有更加精确的测量和计算，才能够满足科学和管理的需要，这引发了历史上第一次数据爆炸。进入 21 世纪后，由于计算机、互联网和智能手机的普及，语音、图片、视频数据前所未有的大规模爆发。过去 5000 年的文明看似浩如烟海，但和今天的数据相比，

其实相当有限，史书大部分都聚焦在为数不多的帝王将相身上，关于普通百姓的个体性记录少之又少。今天不一样了，智能手机的普及，如微信、抖音、电商等平台时时刻刻都在记录着每个人的数据。未来可能不仅有国家史、社会史、行业史，还会有数量惊人的“个人史”。数据将会像雪球一样越滚越大，其规模将前所未有。图灵奖获得者 JimGray 提出：每 18 个月全球新增信息量是计算机有史以来全部信息量的总和。

随着数据经济的发展，人们对数据价值的认识也由浅入深、由简单趋向复杂。总体来看，数据价值的发展分为三个阶段：第一阶段是数据资源阶段，数据是记录、反映现实世界的一种资源；第二阶段是数据资产阶段，数据不仅是一种资源，还是一种资产，是个人或企业资产的重要组成部分，是创造财富的基础；第三个阶段是数据资本阶段，数据的资源和资产的特性得到进一步发展，与价值进行结合，通过交易等各种流动方式，最终变成资本。由此也引发了当前业界的一大难题，即数据产权问题。只有解决了数据产权问题，数据交易才具备顺利开展的前提基础。

2. 大数据特点

数据和大数据往往是人们容易混淆的概念，实际上，两者是有区别的。大数据是一个抽象的概念，其重要性已经得到公认，但人们对于大数据的定义却各执己见。Apache Hadoop 定义大数据为“通过传统的计算机在可接受的范围内不能捕获、管理和处理的数据集合”。麦肯锡全球研究所定义大数据为“一种规模大到在获取、存储、管理、分析方面大大超过了传统数据库软件工具能力范围的数据集合”。高德纳(Gartner)咨询公司给出的定义如下：“大数据是需要新处理模式才能具有更强的决策力、洞察力和流程优化能力来适应海量、高增长率和多样化的信息资产。”可以看出，大数据意味着通过传统的软件或者硬件无法在有限时间内获得有意义的数据集，而经过大数据技术处理后就可以获得有意义数据。

虽然大数据的定义没有统一，但是由国际知名咨询公司 IDC 定义的大数据的四个特征却受到业界的广泛接受，也就是 4V 特征——数据量大(Volume)、数据种类多(Variety)、数据价值巨大(Value) 以及数据产生和处理速度快(Velocity)。

1) 数据量大

当前，以大数据、物联网、人工智能为核心特征的数字化浪潮正席卷全球，全世界每时每刻都在产生大量的数据。衡量数据大小的单位从 MB 到 GB，到 TB，再到 PB、EB，相信未来还会不断地出现新的存储容量单位。根据著名的咨询机构 IDC 做出的估测，人类社会产生的数据一直都以每年 50%的速度增长，也就是大约每两年就会增加一倍，这个被称为“大数据摩尔定律”。随着数据量的不断增加，数据所蕴含的价值会从量变发展到质变，同时，传统的数据存储、分析和计算方法及技术势必不能满足现实需求，迫切需要更智能的算法、更加强大的处理平台和更新的数据处理技术来挖掘数据价值。

2) 数据种类多

随着传感器、智能设备以及移动互联网的飞速发展，数据类型也变得更加复杂多样。除了传统的关系型数据，还出现了来自网页、日志文件、图片、视频、语音、地理信息、电子邮件、社交论坛、各类传感器数据等半结构化和非结构化数据。大数据技术为处理这些不同来源、不同格式的多元化数据提供了可能。

3）数据价值巨大

大数据最大的特点在于通过各种数据分析和挖掘方法，发现诸多看似无关的数据之间暗含的规律和关联。例如阿里巴巴集团每天拥有几亿人的购物数据，通过分析这些数据就可以知道各种产品和市场发展的走势，也可以知道不同用户的爱好和需求，从而进行针对性的推荐，以提高平台的交易量。

不过，我们也要看到，虽然大数据的价值巨大，但价值密度是比较低的，很多有价值的信息都是分散在海量数据中的。例如城市的各个角落布满了监控摄像头，这些摄像头每时每刻都在生成相关的数据并构成视频大数据，一般来讲，这些数据是没有价值的。但当小区发生抢劫或盗窃，马路上发生交通事故等案件时，两三秒的视频可能就会帮助我们破案，这时数据是非常有用的，单点的数据价值是非常高的。

4）数据产生和处理速度快

与传统的报纸、图书等数据载体不同，当下，数据产生和传播的速度非常快，这个“快”体现在两个方面：一是数据产生速度快，例如欧洲核子研究中心的大型强子对撞机在工作状态下每秒产生 PB 级的数据，这种数据呈现爆发式产生，还有的数据是涓涓细流式产生，例如网页点击率、日志文件、GPS 位置信息等，由于用户量巨大，短时间产生的数据量依然非常庞大；二是数据处理的快，随着数据智能化和实时性的要求越来越高，越来越多的数据处理服务趋于前端化，大数据的处理时间需要符合秒级定律，一般要在秒级时间范围内给出对数据的分析结果，响应时间过长，数据就会降低或失去价值。

3.1.2 大数据伦理

大数据技术是一场新的信息技术革命，它给数据收集、存储、管理和利用带来了重大变革，也由此给世界存在方式和人们的生产生活方式带来了大变革。正如有学者所说“如同显微镜使我们观测到深邃的微观世界，望远镜让我们认识到浩瀚的宇宙，大数据技术正在改变我们的生活习惯以及理解世界的方式”。然而，大数据技术同其他信息技术一样，是一把双刃剑，给人类社会带来福祉的同时，也造成更为严峻的数据伦理困境。

大数据伦理（Big Data Ethics）正在成为新的应用伦理方向。然而，它并没有完整、公开和达成共识的定义。清华大学张佐教授认为，与大数据伦理相关的内容包括以下几个方面。

（1）鉴别数据的获取、处理、存储、分发（发布）过程中涉及哪些不同利益主体；

（2）发现大数据实践中对相关利益主体的安全、责任、自由、平等、公平、正义、节俭、环保等伦理原则造成威胁的风险类别、程度大小；

（3）确定数据伦理的价值准则和哲学依据；

（4）指导形成正当行为的行为规范。

厦门大学林子雨教授认为，大数据伦理属于科技伦理的范畴，指的是由于大数据技术的产生和使用而引发的社会问题，是集体和人与人之间关系的行为准则问题。作为一种新的技术，大数据技术像其他所有技术一样，其本身是无所谓好坏的，而它的“善”与“恶”全然在于大数据技术的使用者想要通过大数据技术达到怎样的目的。一般而言，使用大数据技术的个人、公司都有着不同的目的和动机，由此导致大数据技术的应用会产生积极影响和消极影响。

3.2　个人信息保护


微课视频

大数据时代，人们经常有一种“被扒光”和“被操控”的无力感，因为数据比我们更清楚地了解我们自身。凯文·凯利(Kevin Kelly)在《必然》一书中列出了美国对公民进行常规追踪的清单。

> 汽车活动——从2006年开始，每辆车都有一块芯片。当你发动汽车时，它就开始记录车速、刹车、转弯、里程及事故等状况。
>
> 邮政信件——你寄出或收到的每封信的表面信息都被扫描并数字化了。
>
> 手机位置和通话记录——你通话的时间、地点和对象(元数据)会被储存数月，有些手机供应商通常会把信息和电话的内容存储几天到几十天不等。
>
> 信用卡——所有购买行为都被追踪。信用卡和复杂的人工智能相结合形成某种模式，揭示了你的人格、种族、癖好、政治观点和爱好。

通过各种细枝末节的数据拼接，现代互联网技术完全能勾勒出一个人的信息形象。这些信息形象包括外貌、性格甚至人格，一旦遭到泄露、滥用和侵害、后果将不堪设想。

作为信息时代新兴课题，个人信息保护问题较早地在西方被得到了关注。1970年，德国黑森州制定并颁布了《数据保护法》，这是全球第一部保护公民个人信息的国内立法。1977年，德国联邦政府颁布了《联邦数据保护法》。在美国，理论界通过对隐私权理论的改造提出了信息隐私权的概念，以此对信息主体的个人信息提供保护。自此，众多国家和地区纷纷制定了与个人信息保护有关的法律。在中国，由于受传统儒家文化的影响，长久以来本着团体本位的模式构建个人与家庭、个人与社会以及个人与国家的关系，故而对公民隐私以及个人信息的保护一直不够重视，对个人信息保护问题的研究起步较晚。直到2021年8月20日，正式通过了《中华人民共和国个人信息保护法》(以下简称《个人信息保护法》)，并于2021年11月1日正式施行，该法律是一部专门保护公民个人信息的法律。

3.2.1　个人信息和隐私权

《个人信息保护法》对个人信息和隐私信息(敏感个人信息)有了明确的定义：个人信息指以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息；而敏感个人信息是个人信息被特殊保护的部分，指一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息，包括生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等信息，以及不满十四周岁未成年人的个人信息。

该条定义将个人信息分为两类：一类是能够识别公民个人身份的电子信息，另一类是涉及个人隐私的电子信息。换言之，个人信息包括隐私信息与非隐私信息。个人信息是个人隐私信息的上位概念，个人隐私信息是个人信息的一部分。

美国将个人信息的保护纳入隐私权的范畴。给隐私权下定义是一件非常困难的事，直到现在，美国对隐私权都没有一个统一的界定。学者米勒(Arthur Miller)认为，隐私权的含义是模糊的，其内涵容易随着时间的变化而变化。在欧洲，虽然制定了统一适用的个人信息

保护法规，但该法规仍然没有对个人信息与隐私信息之间进行严格区分。如欧盟 1995 年《数据保护指令》在确立个人信息保护价值时，将个人信息保护价值确定为“基本权利”“自由”及“隐私”。由此可推知，欧盟《数据保护指令》中保护的个人信息包含个人隐私权的保护。

鉴于立法与司法很难在个人信息与隐私信息之间划出一条泾渭分明的界线，本书对个人信息、隐私概念的使用未作区分，将它们作为统一范畴展开讨论。另外，考虑到信息技术的迅猛发展给公民个人信息保护带来的危机，本书主要集中在能够识别个体身份与涉及公民个人隐私的电子信息领域进行介绍。

3.2.2 个人信息泄露问题

大数据时代，越来越多的人习惯在线解决问题，而互联网强大的记忆和存储功能让一切在线行为可以被快速捕捉。这让个人信息保护呈现出三大特点：数据隐私泄露更常见、侵犯隐私手段更隐蔽、侵犯隐私后果更严重。

目前，隐私权被侵犯的现象实际上是非常严重的。我们经常会收到各种各样的骚扰电话，这是由于我们很多本应该被保护的信息被泄露、被出卖或者被交易所致。有些是因为数据安全的问题被泄露，收集了他人数据的机构由于管理漏洞或者管理不当，数据被别人窃取；有些是数据被收集后，收集者缺乏管理和保护数据的意识，导致数据的流失或者被滥用；还有的是数据收集者为了牟取不法利益而将他们收集的数据出卖给第三方。

2014 年，全球最大的互联网娱乐社区日本索尼的 PlayStation Network 遭到黑客的攻击，导致 7700 万用户数据被外泄，当时黑客曝光了 7700 万用户的个人信息，包括用户名、出生日期、电子邮件地址和信用卡详细信息等，迫使索尼关闭整个 PlayStation Network 系统近一个月时间。

2014 年，支付宝前技术员工将多达 20GB 的用户数据非法贩卖给电商公司和数据公司。2017 年，京东试用期员工与黑客勾结盗取了大量个人信息，大约有 51 亿的数据，在网络黑市进行贩卖，这些信息包括交通、物流、医疗等信息；2018 年，华住酒店旗下多个连锁酒店入住信息数据被售卖，数据涉及 5 亿条的用户个人信息及入住记录，其中包含不少隐私信息，比如身份证号、家庭住址、银行卡号等；12306 网站 470 万余条用户数据在网络上被贩卖。据 2018 年 8 月中国消费者协会发布的《App 个人信息泄露情况调查报告》显示，遇到过个人信息泄露情况的人数占比 85.2%，没有遇到过的仅占 14.8%。

中国互联网协会《中国网民权益保护调查报告(2021)》显示，63.4%的网民个人网上活动信息被泄露(如通话记录、网购记录、网站浏览痕迹、IP 地址、软件使用痕迹及地理位置等)，82.3%的网民亲身感受到了由于个人信息泄露导致对个人日常生活造成的影响，49.7%的网民认为个人信息泄露情况非常严重。

与此同时，与个人信息相关的犯罪也屡见不鲜，上市公司“数据堂”涉嫌侵犯公民个人信息罪被查，简历大数据公司“巧达科技”非法交易个人信息达数亿条，引起社会的广泛关注，被有关部门查封。

3.2.3　个人信息泄露根源及途径

1. 用户画像

在日常生活中，我们经常会收到诸如骚扰电话、短信、邮件或软件推送信息，究其原因“用户画像”功不可没。用户画像是根据用户在互联网留下的痕迹加工形成的一系列用户标签，是通过收集、汇聚、分析个人爱好，对某特定自然人的个人特征，如职业、经济、健康、教育、个人喜好、信用及行为等方面作出分析或预测，形成其个人特征模型的过程。

当前，用户画像的主要应用场景有两个：一是精准营销，即互联网广告业。和传统的广告业相比，今天的互联网和智能手机通过记录消费者不断产生的数据，对用户个体进行“画像”，向用户推送个性化的广告，这种方式可以使广告发布方精确地找到目标用户，提高行业效率，这也是大数据革命在商业领域的起源；二是商业和社会信用，其主体是金融机构，除了精准营销，这是利用大数据赚钱的第二个法门，也是我们看到互联网企业陆续进入金融领域的原因。互联网企业通过用户的消费记录评估用户的信誉，对用户进行“画像”，从后续的金融服务中盈利。例如阿里巴巴旗下的“芝麻信用”和腾讯旗下的“微粒贷”，他们在给用户打信用分的基础上，向单个用户提供贷款等金融服务。这两种商业模式都需要通过数据监控用户在互联网上的一举一动，用户个体因此成为被观察、被分析、被监测的对象，这就必然带来个人隐私保护问题。

那么，为了构建全面的用户“画像”，企业要给用户打多少标签呢？腾讯高管透露，用户人均被腾讯标记的标签高达 2000 多个，通过这些标签，腾讯可以不断分析“我们是谁、我们要干什么”。移动支付习惯、购买力、常用出行方式、关注新闻类型、性别、年龄段等都是腾讯给我们每个人做标记的标签。这些数据主要分为两类——静态数据和动态数据。静态数据包括年龄、性别、商圈、职业、婚姻状况、生育情况、消费等级等，这些数据一旦收集完毕即形成用户标签。动态数据需要持续不断地收集，如网页点击、视频观看、出行路径等。总的来说，用户画像是由若干标签组成，标签越多，则用户画像越完整。形成用户画像后，厂商可据此实现精准营销和产品开发。

用户画像是在商业利润的驱动下，成为各大互联网企业青睐的对象，也成为当前个人信息泄露的“重灾区”。2013 年，南京一用户发现在百度上搜索“减肥”关键字后，在其他特定网站上出现与关键字相关的广告。这位用户认为，百度在未告知的情况下，将其兴趣爱好和个人需求显露在相关网站上，并利用记录的关键词，对其浏览的网页进行广告投放，侵犯了其隐私权，便向法院提起诉讼。而法院认为，虽然搜索记录具有隐私性质，但不属于个人信息，因为百度公司是个性化推荐服务收集和推送信息的浏览器，没有定向识别使用该浏览器的网络用户身份。同时，互联网领域普遍采用 cookie 技术，基于此技术而产生的个性化推荐服务仅涉及匿名信息的收集、利用，网络服务提供者对此依法明示告知即可。百度在《使用百度前必读》中已经予以说明并为用户提供了退出机制，在此情况下，用户仍然使用百度搜索引擎服务，应视为默认许可。因此，法院最终判其败诉。这是以数据权属不明确而做出的判决，实际上也是对用户画像的一种默许，必然助长行业内以基于个人信息的用户画像手段而开展的各种盈利活动，愈发增加个人信息保护的难度。

2. 默认授权

近年来，默认勾选已成为业内信息收集、捆绑销售的“通用”做法。

2018年1月，支付宝引发年度账单事件，支付宝在年度账单的首页左下方使用小字体、接近背景色和默认勾选同意，让相当多的用户在不知情的情况下“被同意”接受芝麻信用。3月，支付宝又由于默认勾选“授权淘宝获取你线下交易信息并展示”的选项，并将交易信息提供给支付宝、淘宝、天猫等平台，而被用户起诉，索赔2元。

手机互联时代，使用App前在用户协议上“打个勾”，就意味着用户已同意让出自己的部分权利。像支付宝年度账单这样的做法（提供同意勾选框但默认同意），只是“默认勾选”的“初级套路”，还有比“初级套路”设计更隐蔽、情节更严重的“其他套路”。比如有的不给选项，用户点击“注册”“登录”“下一步”就意味着同意的“中级套路”；以诱导用户同意（如提供所谓“优惠套餐”），结果用户不是不知不觉中多花了钱，就是由此泄露更多隐私的“高级套路”；还有设计协议捆绑，用户看似只同意了一个协议，实际上同意了一连串协议的“顶级套路”……据南都个人信息保护研究中心发布的《关于收集个人信息“明示同意”的测评报告与建议》显示，在实测了100款常用App后发现，仅有11%的App做到了合乎法规及规范的“明示同意”。

用户稍不注意，就可能掉入“套路的陷阱”，稀里糊涂中多花钞票不说，个人隐私面临着更大的被泄露的风险。支付宝年度账单“默认勾选”受到舆论质疑后，蚂蚁金服等三家企业被监管部门约谈，监管部门及时回应舆论关切，对三家企业提出批评警示，但目前，监管部门直接关注并予以批评警示的，主要是支付宝账单那样的“初级套路”，其他一些企业在默认勾选上进行的中级、高级和顶级套路，尚未引起监管部门的足够重视，更没有受到应有的严查处理。这种失衡局面，显然不利于有效遏制企业App“默认勾选”乱象，不利于保护用户个人信息安全。

3. 霸王条款

安装App时，我们都遇到过这样的情况：“请求获取位置权限”“请求获取通信录内容”“请求获取设备信息”……App要求获取的权限一个接一个弹出，依次等待我们予以放权。App对获取权限的行为进行明示是对用户知情同意权的尊重。然而频繁弹出的权限获取申请使我们不禁要问，这些权限都是App运行必需的吗？如果我们不同意App权限申请会怎么样？

2018年，有网络安全机构在对中国电信App进行检测时发现，用户在初次安装使用该App时仅有存储、电话、通信录和位置信息4项权限提示，但随后App还要获取其他70项子权限。而且，修改通信录、读取联系人、录音、修改通话记录、拨打电话、发送短信等敏感项并不显示通知用户。如果用户点击“不同意”，则应用自动退出。2018年4月，知乎App也被曝出隐私政策存在霸王条款，用户不同意就不能使用。网友直呼不如将“不同意”按钮直接改为“一键卸载”！

2019年中央广播电视总台“3.15”晚会上，央视曝光了手机App通过不平等、不合理条款的授权协议，强制索取用户个人信息的乱象。晚会现场，主持人用真实的社保信息对“社保掌上通”App进行了测试。在该App上输入身份证号、社保账号、手机号等个人信息，完成注册后，电脑就能远程截取用户输入的几乎全部信息。为了使截取用户信息的操作“合法合规”，App一般会通过隐秘的隐私条款获得用户授权。例如，“社保掌上通”App在隐私条款中写“您在此充分地、有效地、不可撤销地明示同意并授权我们使用您的社保账户密码为您提供查询服务。”这意味着用户的这些信息可以向第三方大数据公司开放。不光是查社

保、查违章的 App 存在着此类问题，目前在下载安装大部分应用软件时都会遇到类似的“霸王条款”，除了签订不合理条款，还会有强制获取权限的行为。这些行为都会使用户在和软件发行商的博弈中处于下风，甚至毫无话语权。

实事求是地说，部分 App 公开隐私政策，明示需获取的各种权限和信息，是基于尊重用户知情权，在用户清楚情况并同意的情况下合理、合法地获取信息。然而，技术运用不合理、用户体验考虑不周全等因素，使用户的使用感受与 App 的设计初衷大相径庭，甚至南辕北辙。App 通过过度索取、强制授权两套组合拳，能轻易迫使用户开通 App 要求的各种权限是不争的事实。

4. 未经授权

除了设置霸王条款和强制获取用户的隐私以外，未经用户许可，私自获取用户数据，也是部分 App 的惯用伎俩。某些 App 表面一套背地一套，窃取用户个人信息的行为屡禁不止，这种现象在使用安卓系统手机用户中尤为严重。2017 年 12 月，国家计算机病毒应急处理中心通过对互联网监测，发现“欢聊”“仿 Iphone 来电闪光”两款应用软件都存在窃取用户隐私信息、造成用户隐私泄露及资费消耗的情况。

App 越界获取手机隐私权限的情况一直存在，但呈现逐渐递减的趋势。根据腾讯社会研究中心和 DCCI 互联网数据中心联合发布的《2018 年度网络隐私及网络欺诈行为研究分析报告》显示，2017 年上半年，25.3% 的安卓系统 App 存在越界获取用户隐私的行为；2018 年上半年，该比例降低到 5.1%；2018 年下半年，仅有 2%的 App 存在越界获取手机隐私权限的行为。获取位置信息、读取联系人和读取短信是安卓系统最常获取的三大核心隐私权限。

不只在主动使用手机 App 时会被未经授权收集用户的隐私，还有企业专门研制信息窃取工具，在公众场合下也能获取用户的隐私。

2019 年，“315”晚会曝光了商场使用探针盒子获取顾客隐私，这款由声牙科技有限公司研发的产品，当用户手机无线局域网处于打开状态时，会向周围发出寻找无线网络的信号，探针盒子发现这个信号后，就能迅速识别出用户手机的 MAC 地址，转换成 IMEI 码，再转换成手机号码。一些公司将这种小盒子放在商场、超市、便利店、写字楼等地，在用户毫不知情的情况下，搜集个人信息，甚至包括婚姻、教育程度、收入情况等个人信息，对用户进行精准画像。这些被窃取的信息通常会被用于房地产、汽车、金融、教育等行业的精准营销。

3.2.4　个人信息保护相关法律

2021 年 11 月 1 日实施的《个人信息保护法》是一部保护公民个人信息的专门法律。这部法律与《民法典》《刑法》《未成年人保护法》《电商法》《网络安全法》《广告法》《消费者权益保护法》《数据安全法》《反垄断法》一起，共同编织成一张个人信息“保护网”。

之前对于公民个人信息的保护虽然也被写入了立法，但主要散落于《民法典》《刑法》《网络安全法》《消费者权益保护法》《电子商务法》《数据安全法》等法律中，且缺乏保护的基本原则。《个人信息保护法》的正式实施，强化了对公民个人信息的系统性保护。尤其重要的是，《个人信息保护法》明确将“告知—同意”原则作为个人信息保护的基本规则，当作为个人信息处理者处理用户信息的规范前提，赋予了公民对个人信息处理的知情权、决定权，也为界定“合法”与“违法”划出了分水岭。

这部法律将改变公民的生活。2021 年，中消协发布提示：如未经消费者同意，经营者不得向消费者推送商业信息；不能将人脸识别作为出入小区物业、经营场所的唯一验证方式。这意味着，公民的手机、计算机将不再成为商业信息轰炸的阵地；人们出入小区、经营场所时，将有更灵活的选择方式。对于公民而言，应当学法、知法、守法，掌握《个人信息保护法》相关规定，防止个人信息“裸奔”，主动拿起法律的武器维护自己的合法权益。

这部法律还将改变互联网经济生态。《个人信息保护法》既是一部公民权利保护法，也是一部企业行为约束法。在《个人信息保护法》正式实施之前，苹果公司表示，保护用户隐私、让用户掌控个人信息，是苹果设计产品和服务时一贯坚持的理念，仅在符合法律依据的情况下使用用户的个人数据，宣布尊重中国消费者在《个人信息保护法》下享有的知情、查阅、更正、转移、限制处理和删除其个人数据等。只有秉持法治精神，加强个人信息保护，企业平台才能行稳致远，互联网生态才能健康发展。


微课视频

3.3 数据滥用与数据安全

3.3.1 大数据时代下的数据安全

数据安全是指通过采取必要措施，确保数据处于有效保护和被合法利用的状态，以及具备保障持续安全状态的能力。数据安全应保证数据生产、存储、传输、访问、使用、销毁、公开等全过程的安全，并保证数据处理过程的保密性、完整性、可用性。

威胁数据安全的因素有很多，比较常见的主要有以下几种。

（1）计算机病毒：计算机感染病毒而导致被破坏，甚至造成重大的经济损失，计算机病毒的复制能力强，感染性强，特别是处于网络环境下，传播性更快。

（2）黑客攻击：入侵者借助系统漏洞、监管不力等通过网络远程入侵系统。

（3）人为错误：由于操作失误，使用者可能会误删除系统的重要文件，或者修改了影响系统运行的参数，以及没有按照规定要求或操作不当导致的系统宕机。

（4）数据信息存储介质的损坏：物理介质层次上的安全隐患大致包括两个方面：一是自然灾害、物理损坏和设备故障；二是电池辐射、磁干扰等。

随着信息技术的进一步发展，信息社会从小数据时代进入更高形态的大数据时代。在这个阶段，通过共享、交易等流通的方式，数据质量和价值得到更大程度地实现和提升，数据的动态流动逐渐走向了常态化和多元化，这使得大数据安全表现出与传统数据安全不同的特征。传统的数据安全理论重点关注的是数据作为资料的保密性、完整性和可用性等静态的安全，其主要受到的威胁在于数据泄露、篡改、灭失所导致的“三性”破坏。大数据时代的数据安全表现在以下几个方面：

第一，大数据成为网络攻击者的显著目标。大数据是更容易被发现的大目标，一方面，大数据对潜在的攻击者具有较大吸引力，因为大数据不仅数据量大，而且还包含大量复杂和敏感的数据；另一方面，当数据在某一个地方大量聚集时，安全屏障一旦被攻破，攻击者就能一次性获得较大的收益。

第二，大数据加大了隐私泄露的风险。从技术层面上来说，大数据平台采用的 NoSQL 数据库，是目前被企业广泛推崇的非关系型数据库，其发展历史较短，目前还没有形成一套

完备的安全防护机制。相对传统的关系数据库，NoSQL 数据库的安全风险更大。同时，NoSQL 对来自不同系统，不同运用程序以及不同活动的数据进行关联，也加大了隐私泄露的风险。

第三，大数据成为高级可持续攻击的载体。在大数据时代，黑客的攻击行为往往较为隐蔽，依靠传统的安全防护机制是很难被监测到的，因为传统的安全检测机制一般是基于单个时间点进行的，而且是具有威胁特征的实时匹配检测。而高级可持续攻击，它是一个实时过程，并不具备能够被实时检测出来的明显特征，因而无法被实时检测。

【案例 3-1】 2018 年，全球最大的社交网络 Facebook 被曝出负面新闻：一个名为剑桥分析的小公司通过不正当的手段在 Facebook 网站上获取了 8700 万用户的数据，并通过用户在浏览网页时的点赞频率来收集个人信息，推断用户的性别、年龄、兴趣爱好、性格特点、职业专长或政治观点等，进行心理画像，通过算法分析出不同人群的希望点、恐惧点、共鸣点、煽情点，结合分析结果向用户推送特定的新闻和广告，这些数据随后被用于多个国家选举中的选民分析环节。2016 年当选的美国总统特朗普就曾经雇佣这家公司，这引发了关于数据操纵选举的舆论批评，直逼特朗普当选的合法性，令人触目惊心。

美国 Facebook 数据泄露事件扭转了大众对大数据安全的传统认知，大数据安全的话题不再仅是个人和企业层面的问题，同时也深入涉及政治权力的崛起，直接影响社会稳定和国家的政治安全。总的来说，数据从静态安全到动态利用安全的转变，使得数据安全的内涵不再只是确保数据本身的保密性，完整性和可用性，它更承载着个人、企业、国家等多方主体的利益诉求。

3.3.2　数据安全问题

近年来，数据泄露事件屡屡发生，数据泄露数量不断增加，波及众多行业，给企业和用户带来了不可估量的严重后果。Verizon 发布《2022 年数据泄露调查报告》指出，2022 年发生了 23 896 起安全事件，其中 5212 起是数据泄露事件，这些数据泄露事件中，82％的违规行为涉及人为因素，勒索软件泄露事件增加了 13％，超过过去五年的总和。

1. 系统漏洞与黑客

系统漏洞是造成数据泄露的罪魁祸首之一。2018 年 Facebook 经历自创立以来的至暗时刻，全年被曝光的 3 次数据泄漏事件中，其中 2 次都与系统漏洞直接相关，涉及约 1 亿用户。2018 年 9 月，Facebook 在泄露 5000 万条用户信息后再次卷入数据泄露旋涡，其系统因安全漏洞遭黑客攻击，导致 3000 万条用户信息被泄露，包括 1400 万条用户的姓名、联系方式、搜索记录、登录位置等敏感信息。12 月，Facebook 再次被曝因软件漏洞可能导致 6800 万用户的私人照片被泄露。

2021 年 3 月，黑客利用微软 Exchange Server(电子邮件服务器)的 4 个重大零日漏洞，造成不需要身份验证或访问个人电子邮件账户即可从 Exchange 服务器读取其他用户的电子邮件。这使得全球数十万台 Exchange 服务器被攻击，大量企业、政府部门的系统被感染，超过 6 万家组织受影响。

由系统漏洞引发的数据泄露事件不一而足，那什么是漏洞呢？计算机系统漏洞问题是主要的网络安全隐患之一，一般是由于编程人员在进行系统设计时，由于安全技术和专业能力的限制造成了程序逻辑结构不严谨或设计错误而产生的缺陷。

漏洞伴随着系统的诞生而持续存在。目前，大型信息系统的代码动辄数百、上千万行，Windows 7 操作系统有 5000 万行代码，Windows 8 有上亿行代码，其中潜藏着成千上万的漏洞。更可怕的是，随着信息系统运行、检测、迭代升级，尽管绝大部分漏洞可以被发现并及时清除，但仍有部分漏洞如附骨之疽一样难以被发现，更不会被修复，成为持续影响系统安全的重要源头。例如，2018 年 1 月被发现的能影响几乎所有 Intel CPU、AMD CPU 和部分 ARM CPU 的 Meltdown 和 Spectre 漏洞，其产生时间可追溯至 1995 年，当时 CPU 刚刚开始使用乱序执行和预测执行等硬件设计等特性。微软自动认证漏洞、BadTunnel 漏洞、Windows 打印机漏洞、Shellshock 漏洞等在被发现并修复之前，潜藏时间均超过 20 年。

黑客攻击是导致数据泄露的最主要原因。根据金雅拓统计，56％的数据泄露事件是由“恶意的外部入侵者”引发的。IBM 的研究报告显示，犯罪攻击导致了 48％的数据泄露事件，漏洞攻击、病毒利用、“撞库”等是主要的获取数据方式。

【案例 3-2】 2018 年 11 月，华住酒店集团旗下酒店共计 5 亿条用户信息在暗网被售卖，涉及用户姓名、身份证号、手机号、邮箱、家庭住址、生日、入住时间、离开时间、酒店 ID 号、房间号及消费金额等敏感信息。根据调查，该事件是由于疑似华住程序员在 GitHub 上传的名为 CMS 的项目被黑客攻击所致。

【案例 3-3】 2018 年 1 月，印度的 10 亿公民身份数据库 Aadhaar 被曝遭网络攻击，姓名、电话号码、邮箱地址、指纹、虹膜记录等极度敏感的用户信息被泄露。根据调查，Aadhaar 数据库的登录和 e - Aadhaar 的下载程序存在风险，允许第三方通过白名单 IP 地址登录 Aadhaar 数据库，访问相关数据。

人为因素是导致数据泄露的重要原因。据 IBM 统计，25％的数据泄露事件由人为因素导致。人为因素分为两种情况：一种是企业内部人员或承包商实施恶意的内部攻击，导致数据泄露；另一种是企业内部人员或承包商因设备配置不当、工作疏忽，导致数据暴露在公开的互联网上。

【案例 3-4】 2018 年 6 月 19 日，一位用户在暗网兜售圆通 10 亿条快递数据，该用户表示售卖的数据为 2014 年下半年的数据，数据信息包括寄(收)件人姓名、电话、地址等信息，10 亿条数据已经经过去重处理，数据重复率低于 20％，并以 1 比特币打包出售。考虑到泄露数量之大、准确率之高，外界普遍认为数据来源为圆通内部较高级别的工作人员。

用户数据涉及大量个人隐私，其重要性对用户来说的重要性不言而喻。然而，作为数据的生产者、拥有者，用户难以掌握自身数据的流转轨迹，数据泄露后难以第一时间获知，甚至在泄露数据已经多次转手，被用于精准营销、诈骗时，都不清楚到底是哪里出了问题。另一方面，大部分企业对数据泄露等数据安全问题的认识不到位，总以为不会得到黑客的“眷顾”，并且没有建立相应的监测预警、应急响应机制和手段，不仅发现不了数据泄露，而且难以及时应对和补救。根据 IBM 统计，企业发现数据泄露的平均时间是 197 天，控制数据泄露造成的后果还额外需要 69 天。发现数据泄露的时间越长，控制数据泄露的时间越久，企业和用户的损失就越大。

2. 网络爬虫

大数据时代，企业收集数据的方式多种多样。除了直接通过用户采集之外，还包括传感器采集、网络爬虫采集等方式。其中，利用网络爬虫采集公开信息是企业数据的重要来源。相关数据显示，50％以上的互联网流量其实都是爬虫贡献的。对于某些热门网页，爬虫的访

问量甚至占总访问量的90%以上。

网络爬虫(Web Crawler)是依照一定规则主动抓取网页的程序,是搜索引擎获得信息的渠道之一。通常根据给定URL种子爬取网页,得到新的URL存放至待爬行URL中,当满足一定条件时停止爬行。本质上,网络爬虫是通过代码实现对人工访问操作的自动化。但是,网络爬虫具备的代码解析能力使其可能访问到人工不会访问或者无法访问的内容。过度使用网络爬虫可能引发一些问题:过于频繁的数据爬取操作可能加大网站负担,导致网站瘫痪,用爬取技术获取数据,可能导致数据所有者失去对数据的拥有权。如果爬取数据的企业信息和个人信息未经授权或被不正当地使用,可能导致引发商业纠纷,侵犯个人的合法权益。

为了规范网络爬虫行为,荷兰软件工程师马蒂恩·科斯特(Martijn Koster)于1994年2月起草了网络爬虫的规范——Robots协议。Robots协议又被称为爬虫协议或机器人协议,该协议是在搜索引擎诞生发展壮大的背景下应运而生的,它是互联网企业间相互博弈的结果,是在商业利益、用户个人利益和网站自身安全的基础上最终达成的一种妥协。它不具有强制性,相当于一个"君子约定"。Robots协议的作用是告诉网络爬虫哪些页面可以抓取,哪些页面不能抓取。是否遵从Robots协议,也逐步从行业规范上升为量刑的重要依据。

【案例3-5】 2017年,58同城的全国简历数据泄露事件引发轩然大波。有淘宝电商出售"58同城简历数据":一次性购买2万份以上,0.3元一条;一次性购买10万份以上,0.2元一条;同时,支付700元即可购买爬取软件。该事件是因为58同城对合作商开放的权限过度,保密措施不够完善,导致爬虫有机可乘,被盗取了用户信息。

安全专家分析,出售的数据爬取软件本质上是一款恶意爬虫工具,利用58同城系统的漏洞爬取相关信息。根据正常的商业模式,58同城、智联招聘、前程无忧等招聘网站允许企业和个人访问简历信息,网络爬虫自然也在许可范围之内。但是,无论企业、个人还是网络爬虫,都只能看到部分的简历内容,个人联系方式等敏感的简历内容需要付费才可以查看。然而,58同城系统的多个安全技术漏洞的组合使网络爬虫一步步获取到了用户的全部简历信息。具体地说,第一个漏洞允许爬虫批量获取用户的简历ID,第二个漏洞会导致用户姓名等真实信息泄露,第三个漏洞允许爬虫通过用户ID抓取用户的电话号码。在多个漏洞的叠加影响下,用户的简历信息就没有秘密可言了。那么,企业和个人应该如何使用网络爬虫这把双刃剑呢?爬取数据前,首先应识别数据性质,如果是严格禁止侵入内部系统数据,爬取数据时,避免获取个人信息、明确的著作权作品、商业秘密等。爬取数据后,严格限定数据应用场景,切忌以不劳而获、"搭便车"的方式利用他人的数据,侵害他人的商业利益。

3. 数据黑产

大数据时代,信息的高速流转和运营创造了空前巨大的价值,随之而来的是猖獗的倒卖信息数据的乱象。企业大规模数据泄露事件频发,数据安全如临深渊。根据南都大数据研究院等机构发布的《2018网络黑灰产治理研究报告》,2017年我国网络安全产业规模为450多亿元,而黑灰产已达近千亿元规模。网络黑灰产新趋势越来越多样化,以信息诈骗为主的黑灰产业链逐渐向境外转移,黑产从业人员沟通圈子扩展到更多即时通信工具、平台等,给监管治理带来巨大挑战。

目前,庞大的数据黑产已经相当完善。根据产业链内各角色分工的不同,数据黑产大致可分为上游、中游、下游三部分。数据黑产的上游以内鬼、黑客为主,他们通过访问特权或非

法入侵企业信息系统等方式获取数据，或者为黑客实施攻击提供工具支持的工具制造者、贩卖者等，均属于数据黑产的上游。数据黑产的中游以掮客、条商、代理商及处理者为主，他们负责数据的交易和流转。数据黑产的下游主要是各类数据买家和使用者，这些购买的数据往往被用于精准营销、身份认证及电信诈骗等。

数据泄露是数据黑产的源头。泄露的信息在黑市中被反复倒手，直至被榨干价值。我们总以为黑客、网络攻击及病毒等因素是造成数据泄露的主要原因，然而，相关调查显示，80%的数据泄露是企业内部人员所为，黑客和其他方式仅占 20%。泄露组织机密信息的人往往被称为“内鬼”或“细作”，而内鬼越来越成为数据泄露的罪魁祸首。

随着产业链上中下游的分工逐步明确和细化，第三方服务机构成为数据泄露的新主体。2018 年 8 月，浙江警方破获了一起上市公司非法窃取用户数据案，堪称“史上最大规模数据窃取案”。据悉，上市公司瑞智华胜借助为国内电信运营商提供精准广告投放系统的开发、维护的机会，将自主编写的恶意程序部署到运营商内部的服务器上，非法从运营商流量池中窃取搜索记录、出行记录、开房记录及交易记录等 30 亿条用户数据，导致百度、腾讯、阿里巴巴、今日头条等全国 96 家互联网公司的用户数据被窃取，国内几乎所有的大型互联网公司均被“雁过拔毛”。

整个数据黑产链条中，中游的数据清洗是数据黑产的关键步骤，“拖库与撞库”则是数据清洗的关键步骤，“拖库与撞库”的全过程如图 3-1 所示。其中，“拖库”指黑客以 txt、xls 等格式从数据库中导出数据的行为。通常，“拖库”窃取到的邮箱、社交软件等账号及密码信息大多是单一、无效的，或者有些数据库中存储的密码是经过加密的，难以直接使用，这时就需要使用“撞库”的办法对获得的数据进行清洗。“撞库”是指黑客利用工具收集整理已泄漏的用户名、密码等信息，将其整理成账号字典，通过尝试对其他网站进行批量登录的方式，得到可登录的有效用户名和密码等信息的过程。用户为了方便，经常在多个网站设置同样的用

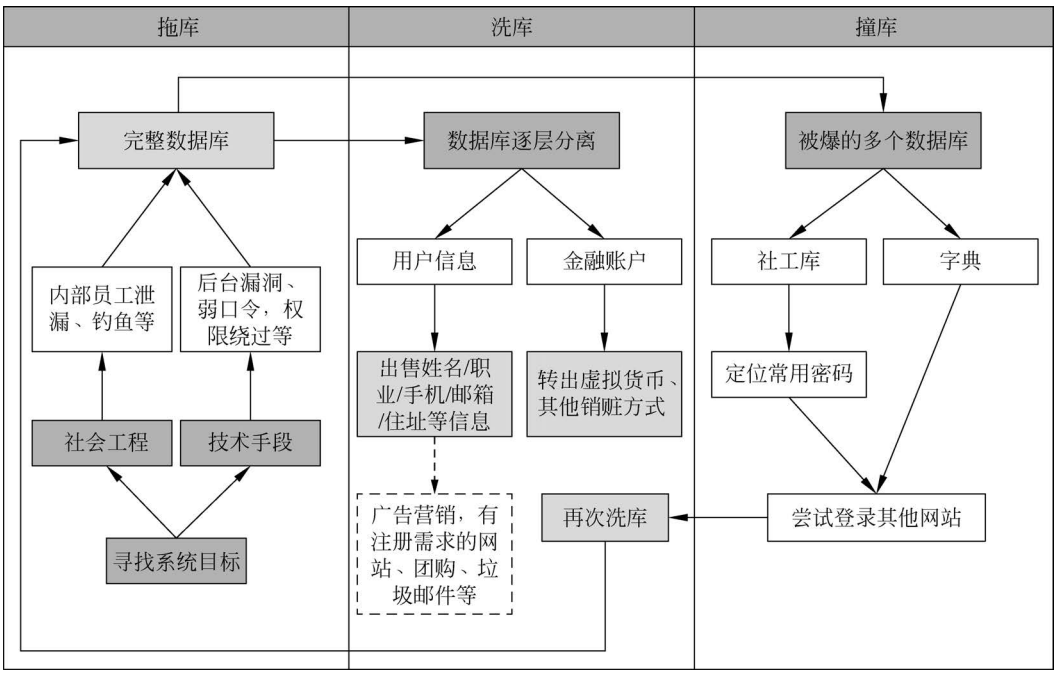


图 3-1 “拖库与撞库”的全过程

户名和密码，一旦其中一个网站的信息遭到泄露，就很容易被黑客通过“撞库”攻击的方式顺藤摸瓜，被获取手机号、身份证号、家庭住址及银行账户等敏感信息。2016 年 10 月，网易遭遇“撞库”攻击，导致网易 163、126 邮箱过亿条数据被泄露，包括用户名、密码、密码保护信息、登录 IP 以及用户生日等信息。经过“撞库”清洗后，账号、密码的有效性更强，可以精准获取用户多平台的相关注册信息，数据内容更丰富。这在犯罪分子眼中极具价值，价格也水涨船高。

在数据黑产中，数据交易是数据变现的重要方式之一。腾讯安全发布的《信息泄露：2018 企业信息安全头号威胁报告》，账号邮箱类数据、个人信息、网购物流数据是黑客交易最受欢迎的产品，交易量占比分别为 19.78%、12.19%、9.69%。数据交易在具备变现属性的同时，也是数据清洗的关键一环。据地下数据产业资深人士透露，随着数据需求的持续放大，非法数据交易等数据黑产有公开化的趋势。部分大数据初创企业通过购买各种渠道的数据，其中不乏黑客、内鬼甚至暗网出售的数据，整合数据资源，降低数据成本，提供更全面的数据服务。在这样的商业模式下，不同出身的各种数据实现了合法流通，无疑更刺激了数据非法交易现象的猖獗。

经过数据交易、数据清洗等环节的复杂运作，泄露的涉及姓名、电话、身份证、银行卡及家庭住址等真实信息的各种数据最终流入各类数据买家和使用者手中，充分展现了数据的“价值”。电信诈骗、精准营销是数据变现的最终环节。中新网 PC 端与微信端均有超过 70%的网友表示，诈骗电话、诈骗短信是自己信息被泄露后最困扰自己的事情。银联数据显示，90%的电信诈骗案、盗窃银行卡、非法套现、冒用他人银行卡及网络消费诈骗等都是由于个人数据泄露引发的。

【案例 3-6】 2016 年 8 月 21 日，山东女大学生徐玉玉被诈骗分子以发放助学金的名义骗走学费 9900 元，在报警回家的路上猝死。究其原因，就在于徐玉玉准确的录取信息、手机号码等个人信息被窃取、贩卖，进而引发了精准的电信诈骗。

3.3.3　数据滥用

1. 算法歧视

算法歧视是以算法为手段实施的歧视行为，主要指在大数据背景下、依靠机器计算的自动决策系统在对数据主体作出决策分析时，由于数据和算法本身不具有中立性或者隐含错误、被人为操控等原因，对数据主体进行差别对待，造成歧视性后果。

其实，这种数据算法的不公平甚至歧视现象出现在生活各个场景之中。例如，谷歌的线上招聘平台向女性用户推送高收入岗位信息的频率远低于男性用户，尽管这些岗位实际上并没有限制性别；购物或服务平台将同质的产品向不同的消费水平的用户收取不同的价格；美国联合航空对所有男性和女性用户分别推出不同的折扣活动等现象。

人们希望的是数据挖掘技术能够为个体与社会决策提供一个更加安全、公平、具有包容性的保障，这是对数据挖掘技术的美好愿望。然而，近年来，各个领域的学者纷纷指出，人们所盼望的这种公平性和客观性并不能够得到保证。相反，数据挖掘还可能会加剧生活中原本的不公甚至增加新的偏见。

算法歧视问题不仅使算法无法充分发挥其正向效用，也成为大数据科学及人工智能技术推广中不可忽视的障碍。对于用户而言，算法歧视问题侵害用户个人权益及尊严感。对

于企业而言，一方面，算法歧视可能会导致企业的直接经济损失，比如信息推送不精确、广告投放对象偏差、人才招聘选择范围过窄等问题；另一方面，算法歧视问题会通过影响用户满意度而间接影响企业的收益及声誉。

2. 大数据杀熟

大数据“杀熟”指互联网平台利用大数据挖掘算法获取用户信息并对用户进行“画像”分析，进而对不同消费者群体提供差别性报价，以达到销售额最大化或吸引新用户等目的的行为。这种企业“杀熟”现象的本质是通过一定的算法筛查，对用户群体进行分类，形成一套端口配多套服务的模式。而这里的“熟”指的是那些已经被大数据挖掘算法充分掌握信息的用户。

【案例 3-7】 2018 年，一位网友自述他经常通过某旅行服务网站预订一家出差常住的酒店每间房每天的常年价格在 380～400 元。他偶然通过前台了解到，相同房型每天在淡季的价格在 300 元左右，他通过朋友账号查询后发现，果然显示的是 300 元，但自己的账号却显示的是 380 元。无独有偶，另一位网友自述某外卖平台 App 在同一时间同一家店进行点餐，会员的配送费反而要比非会员的配送费多 4 元。随后又查看了附近的其他外卖商家，发现开通会员的账号普遍比不开通会员的账号需要多支付 1 元～5 元。

事实上，大数据“杀熟”现象不仅仅出现在旅游类、外卖类平台上。调查显示购物类、打车类等平台 App 均存在着大数据“杀熟”现象。这些互联网电商、在线旅游平台等对某用户越了解，就越知道该用户的价格承受度和是否有比价习惯，一旦认定该用户是价格不敏感用户，就会很容易对其“杀熟”。反而对于新用户，这些平台为了留存还会给予一定的优惠，留出一段“养熟期”。

2022 年 3 月，北京市消费者协会发布互联网消费大数据“杀熟”问题调查结果，对于 16 个平台的 32 个模拟消费体验样本，竟有 14 个样本的新老账户价格不一致。其中，网络购物、在线旅游和外卖是大数据“杀熟”的重灾区。有八成多(82.44%)受访者表示在网络购物过程中遭遇过大数据“杀熟”，七成多(76.85%)受访者在在线旅游消费中遭遇过大数据“杀熟”，反映在网络外卖(66.96%)和网络打车(63.00%)消费过程中遭遇大数据“杀熟”的受访者均达到六成多。

究其大数据“杀熟”的原因，一方面是利益驱动使然；另一方面是由于平台在技术、信息等方面，对消费者拥有压倒性优势。因此，消费者遭遇“杀熟”，面临着举证不易、维权困难等问题。

此外，大数据“杀熟”在于平台对用户数据的保护和利用不当。基于便利，用户在平台使用各项功能时让渡了自己的部分数据权利。例如，让平台获取自己的消费习惯、消费能力、商品偏好、价格敏感等信息。然而，这并不意味着平台可以随意使用这些用户数据，或者利用信息不对称进行牟利。

大数据“杀熟”严重侵害消费者权益，如果任由其发展，将不利于电商行业的持续健康发展。大数据“杀熟”暴露出大数据产业发展过程中的非对称以及不透明。平台根据搜集用户的个人资料、流量轨迹、购买习惯等行为信息，通过平台大数据模型建立用户画像，然后根据这个画像来给用户推荐相应的产品、服务并相应定价。这种大数据“杀熟”，同一平台针对不同的消费者制定不同的价格违反了《消费者权益保护法》中规定的公平诚实信用原则，侵犯了消费者的知情权。

信息时代，大数据给生活带来了更多可能，算法、用户画像、精准推送等技术日新月异，但都不应脱离法律和道德的约束，不能损害公众的利益。从这个意义出发，必须依法治理，及时规制负面因素，确保技术更好造福社会。

3. 动态定价

动态定价是根据产品或服务的供需关系对价格进行动态调整，是企业利润最大化的方法，是很多行业常用的运营手段。对于机票销售等产品库存固定的行业，动态定价由来已久。但是融合大数据技术、挖掘算法和人工智能技术后，动态定价的作用就开始变得“惊艳”。

2000 年 9 月，亚马逊网站进行了一项动态定价实验，它根据客户的购买历史以不同的价格（最多相差 40%）把 DVD 出售给客户。当该实验的消息被曝光后，消费者隐私团体对其提出了严厉批评。而亚马逊公司也公开道歉，并向 6896 名客户退回相关款项。虽然 DVD 的动态定价在媒体上得到了更多的关注，但亚马逊的动态定价实际上是在 2000 年 5 月首次被发现的，当时它被发现在一个流行的 MP3 播放器上向一些客户提供超过 20%的折扣。除了动态定价外，企业还利用消费者的资料数据来定位广告和产品推荐。事实上，使用消费者档案数据的定制广告的售价是无目标广告的十倍。

20 多年过去了，这种动态定价进展如何？随着互联网从 PC 端向移动端的转变，动态定价问题并未得到解决，相反，还出现了个性化动态定价。由于互联网技术不断创新，技术层面有较大突破，其中受到各大互联网平台追捧的技术莫过于“千人千面”技术。淘宝早在 2013 年就提出“千人千面”算法，依靠淘宝大数据及云计算能力，能从细分类目中抓取那些特征与买家兴趣点匹配的宝贝，展现在目标客户浏览的网页上，从而帮助卖家锁定真正的潜在买家，实现精准营销。

这种模式在拼多多、京东等购物平台中都会出现，由于“千人千面”的硬件基础是移动互联网带来的“一人一屏”。商家针对不同用户制定不同价格，“千人千价”油然而生。简单地说，通过数据区分穷人和富人、新人和旧人、价格敏感人群和价格不敏感人群，甚至是安卓手机用户与苹果手机用户。同一时间、同一起点、同一目的地，不同手机使用同一个打车软件跳出的价格却不同，甚至是在购物软件，使用不同手机选择同样的商品也会出现不同的价格。

“千人千面”可以理解为更高阶的“杀熟”手段。“初代杀熟”的差异化定价较明显，“二代杀熟”却更为隐蔽，也更能被合理解释。“千人千面”技术利用大数据对用户进行分析，给出对客户最有诱惑力的解决方案。例如，上海市消费者权益保护委员会曾在不同平台测试了订房、买菜等业务，不同账号的价格差异比以前更大。不同的是，现在的“千人千价”由原价与各种优惠券组成，券并不是账号钱包里原有的，而是算法临时生成的。虽然算法为优惠券安上各种理由，比如“上个月你打过车”“前天你买过菜”等，但如果优惠券是浏览时才临时产生的，就是一种“二代杀熟”。当然这种方案可能是最划算的，也可能是最贵的。“千人千价”是通过算法对数据的自动处理实现的，主观上它没有泄露任何人的数据和隐私，但它在某种程度上侵犯了消费者的经济利益。

无论是大数据“杀熟”，还是非法动态定价，企业都是想尽办法制造与消费者、竞争对手甚至合作方的信息不对称，凭借掌握的大数据形成资源优势，操纵价格改变区域或个体供需关系，进而非法牟利。在这个过程中，一旦数据被用“歪”了，也就助长了价格操纵带来的非

法利益，成为价格操纵的“帮凶”。

3.4 数据产权


微课视频

数据不同于石油等传统资源，其价值在于流动。无论政府还是部分企业，都拥有非常丰富的大数据资源，但是大部分都被束之高阁，有数据需求的企业无法获取。其中横亘的第一道“天堑”就是数据产权问题。

【案例 3-8】 2017 年，华为与腾讯两大互联网公司被曝出在获取用户数据方面存在分歧，发生了一场火药味颇浓的争论。2016 年底华为推出一款手机，它可以根据微信聊天内容自动加载地址、天气等信息，对于这项功能，华为并不认为自己侵犯了用户的隐私权。华为认为用户只有通过设置后，公司才能搜集到相关的数据，公司对数据的收集是经过用户同意的。但腾讯指出，华为不仅获取腾讯的数据，还侵犯了用户的隐私权。

该事件集中体现了数据产业链和生态链中围绕数据的白热化竞争，也从侧面反映了当前数据产权不清晰所带来的问题。要弄清数据产权，仅了解数据是不够的，还需要着眼于数据的整个产业链条。

3.4.1 数据资产化

数据作为一种原材料，通过数据分析建模的加工挖掘，能产生新的价值，因此数据已成为新的生产力来源。从数据处理的角度来讲，这里的“数据”可分为原始数据和衍生数据。原始数据是指来自上游系统的，没有做过任何加工的数据。衍生数据是指通过对原始数据进行加工处理后产生的数据。衍生数据经过算法筛选、分析、处理和挖掘之后，形成具有一定使用价值的数据。

原始数据是不可再生的数据，而衍生数据是可再生数据。当数据量较小时，原始数据体现数据价值，因此从数据内容中可以直接获取数据价值。当数据量较大时，原生数据的直观价值锐减，重点应为数据之间相关性的价值挖掘，这就是所谓的衍生数据价值。大数据经济环境下，企业追逐的数据价值基本都体现在衍生数据上，而衍生数据价值的高低取决于对原始数据加工和计算的准确程度。

随着大数据时代的到来，数据分析处理技术的提升使得一个个数据抽象的描述逐步成为可视的计量，也成为大数据进入国民经济体系的一个良好途径。那么，数据能否和其他财产一样成为资产呢？数据资产化，并没有一个标准的定义，通俗地讲就是将数据看作企业资产的一系列探索和尝试。从会计学角度来看，资产是指由企业过去的交易事项形成的、由企业拥有或者控制的、预期会给企业带来经济利益的资源。当然，作为资产其成本和价值还需要可靠计量，如此一来，数据虽然看似具备这些特点，但细究起来却总不如传统资产那样完美符合这些条件。其实，并非所有的数据都能成为数据资产，除非同时满足可被计量、可被控制、可被变现的属性。

近年来，我国数据产业发展迅速，数据产业链中的一大亮点就是数据交易产业。2015 年 4 月 15 日，贵阳大数据交易所正式挂牌。随后，中关村数海数据资产评估中心有限公司也获批成立，这是我国首家数据资产登记确权赋值的服务机构。许多数据创新型企业通过数据资产登记评估等资产化获得挂牌上市的机会，还有很多企业成功将数据资产作为一种

新型资产进行抵押，实现了融资。由此可见，当前数据资产化是大势所趋，数据的经济价值必然成为人们追逐的热点。

3.4.2　数据产权

数据产权，即数据主体对产生的数据所享有的权益，包括数据的所有权、占有权、使用权、收益权、处置权等。不同于其他形式的资产，产权具有唯一性和排他性，数据产权因数据的可复制性而不具有排他性，数据产权的认定比以往任何权利的认定都更为复杂，数据的收集、挖掘、开发、利用、共享与交易等环节都与数据产权的认定息息相关，数据产权界定是数据要素有效配置的基础。

2017 年，欧盟委员会发布《打造欧洲数据经济》报告，明确了欧洲数字单一市场战略的三大目标：一是最大限度发挥数据效益，便于对机器生成的数据的获取和共享；二是保护投资、资产和机密数据，建立完善的投资和创新激励机制；三是确保数据持有人、处理者和服务提供商在价值链内公平分享利益。在这一背景下，欧洲就非个人数据和数据生产者权利展开研究，提出了新型数据产权，以规范市场和交易。

2020 年 3 月，中央国务院《关于构建更加完善的要素市场化配置体制机制的意见》（以下简称《意见》）提出“研究根据数据性质完善产权性质”。在数据要素分配过程中，“数据产权”的概念被反复提及，主要是对如何拥有数据要素及如何分配数据要素的财产权益尚无明确的规则。2022 年 12 月，中共中央、国务院印发《关于构建数据基础制度更好发挥数据要素作用的意见》，全方位构建了数据要素市场的顶层设计，其中指出“探索数据产权结构性分置制度”，并明确提出“建立数据资源持有权、数据加工使用权、数据产品经营权等分置的产权运行机制”。这一“三权分置”思路，有助于破解当前数据确权面临的诸多难题。

一直以来，数据产权的争议归结为三大核心问题：数据归谁所有？谁可以使用数据？数据收益如何分配？即所谓的数据资产的“所有权”“使用权”和“收益权”。“三权分置”将数据资源的这三种权益分别归属于不同的主体，即数据所有者、数据使用者、数据经济收益主体三个主体。数据所有者是指数据产生者或者拥有者，如政府、企业、个人等；数据使用者是指对数据进行使用的主体，如企业、研究机构、政府部门等；数据经济收益主体是指从数据产生的经济收益中受益的主体，如政府、企业等。

1. 数据所有权

典型案例是新浪微博起诉脉脉抓取和使用微博用户信息案，该案被业界称为“大数据引发不正当竞争第一案”。

【案例 3-9】 脉脉作为一款社交软件，通过与新浪微博合作，能够利用用户的新浪微博发现新朋友，并帮助他们建立联系。根据二者之间签订的《开发者协议》，脉脉只能获得新浪微博用户的姓名、性别、头像、电子邮箱这些信息。然而在合作期间，未经微博平台许可，脉脉调用了大量微博用户的教育信息，职业信息和手机号码。此外，在合作终止后，脉脉仍将其用户手机通信录里的联系人与新浪微博用户对应，并展示在脉脉用户“一度人脉”中。

新浪微博认为，脉脉非法获取了教育信息、职业信息以及手机号码等在高级权限下才能调取的信息，违反了与新浪微博签订的《开发者协议》，获得了本属于新浪微博的用户信息。本案暴露的一个问题就是数据权属问题，即新浪微博是否合法取得用户在该平台上的所有数据。

近年来，政府作为公共事务管理机关，无时无刻不在收集社会各界的数据信息，比如身份证信息、指纹信息、信用信息及出行信息等。公民的社保缴费记录，患者的就诊记录，企业的工商登记信息……这些数据产权到底属于谁？政府部门、企业还是个人？如何作出清晰界定，将直接决定谁享有数据的权益。

该问题被提出时，法学界研究者的主要思考一般是从数据权属开始的。采用与传统财产权相似的设计方法，使得智力成果的权属设计最终成为了知识产权制度。然而这一思路，在数据语境下有着很多的障碍。虽然人们习惯于把数据比作石油或黄金，但与此二者不同，数据价值的体系是以大规模汇聚为基本前提的，并且有着递增的边际收益，而单个的或片段的数据，财富创造能力相当有限。所以，《意见》并没有使用“所有权”这一概念，而是代之为“持有权”。

数据资源的持有与不同主体有关。从大类上分，数据可以分为公共数据、企业数据和个人数据，相应地，数据持有主体也就应当包括政府、企业和个人。对于公共数据，一般来说应由管理部门持有和控制；对于企业在生产经营活动中采集加工的不涉及个人信息和公共利益的企业数据，由这些市场主体享有数据持有、支配和收益的权利；而对于承载个人信息的数据，则个人持有，或者由特定的数据处理者按个人授权范围采集、持有和使用。

1）个人数据所有权

对个人数据持有者来说，对数据的控制放在知情同意权的设置上。我国司法裁判中也明确承认企业以用户知情授权为前提。同时，许多学者也建议在个人数据所有权下设置知情同意权、数据修改权、数据被遗忘权。这种做法使企业在数据清洗前对收集的个人数据的控制权依赖于用户，只有用户才能根据自己的意志对数据财产加以管控。

2）企业数据所有权

企业作为经济活动的组织形式，强调企业数据产权，目的是对经济活动中产生的数据确定权属。企业数据可被分为三种类型：第一类数据是企业作为一个主体，自身独有的数据；第二类数据是企业作为平台或者介质，掌握的用户数据；第三类数据是企业经营活动所获取的数据。

显然，企业对第一类数据享有当然产权。第二类数据中，涉及用户个人信息的，如用户的手机号码、身份证信息、邮箱、微信、QQ、支付宝等信息专属用户所有，企业对此不应享有任何权益，而是还负有保护义务。第二类数据中涉及个人隐私的，如家庭住址、工作单位、行程信息等，企业有义务维护用户的隐私权，未经用户同意，不得公开和利用其隐私。第三类数据是企业在经营活动中，对其掌握的经营信息予以加工、生产的数据，在不侵害个人数据产权的前提下，享有其数据产权。例如，用户在自媒体平台发布的视频、图片和文字，属于用户原创内容，就内容本身而言，无论是著作权还是其他数据权利当然归属于生产该内容的用户。但是，平台也应根据其贡献享有相应的数据权利。平台需要投入大量人力、物力、存储、计算资源，才能让相应的作品以特定的方式存储、播放，并且能够被平台用户观看、评论、转发。经过这一转化之后，这一作品同时构成“平台数据”的组成部分，平台享有与作者权利并行不悖的数据权利。

新浪微博起诉脉脉案件中，新浪微博向脉脉提供了 Open API 接口，用于获取新浪微博用户的头像、名称、标签等信息，但脉脉不仅抓取了上述信息，还非法抓取新浪微博用户的教育信息、职业信息，并非法使用这些信息。这些信息属于用户的隐私数据，在未获得用户知

情同意的情况下，脉脉是无权获取用户的教育信息和职业信息的。新浪微博作为数据提供方，不仅应将用户数据信息作为竞争优势来加以保护，还应将保护用户数据信息作为企业的社会职责，采取相应的技术措施提升 Open API 合作模式中相应权限的控制。第三方应用基于 Open API 合作模式利用用户信息时，除应取得数据提供方同意外，还应再次取得用户的同意，尊重用户的自由选择权。数据提供方与第三方应遵守《开发者协议》，在读取和运用用户数据时，以取得“用户授权＋平台授权＋用户授权”的三重授权为原则，以保护用户的隐私权、知情权和选择权为底线，以公平、诚信为行为准则，维护互联网络的公共竞争秩序，实现数据经济的合作共赢。

3）公共数据所有权

公共数据是政府部门在履行职责过程中获得的各类数据资源。由于公共数据所负载信息的特殊性，所以将其所有权专门提出。公共数据的作用一般体现在两个方面：一方面是服务于政府行政管理，另一方面需开发其市场价值以促进数字经济。当前，我国公共数据所有权分为公共数据归属权与管理权。政府是国家的行政机关，其在行使国家权力时搜集的数据理论上应该归国家所有，属于全体公民，而政府作为公共数据的收集者、控制者，赋予其管理权最为便捷。所以，公共数据归属权属于国家，政府享有数据的管理权，同时通过政府履行数据公开义务，以满足公众对公共数据的需求，达到个体要素与社会要素的平衡。

那公共数据中涉及个人隐私的数据是否归个人所有呢？首先，我国政府收集数据的目的并不是盈利，而是公共管理，这就不存在利益交换的问题。例如，我们到公安局登记更改身份信息，公安局搜集这部分数据，当然不是为了出卖个人信息来挣钱，而是为了维护社会安全。其次，公共数据以国家强制力作保障，可以有效地保障个人隐私不受侵犯。第三，个人数据所有权中涉及的知情同意权、删除权、被遗忘权等，因涉及公共利益，也不应配置给个人。所以，将公共数据的原生权利配置给个人并无实际意义和价值。

2. 数据使用权

关于数据主体采集到的数据谁可以用，各界亦无共识。从现实情形来看，当前数据的大规模使用主体有两个，一是政府，二是企业。那么，这些主体是否可以随意使用采集到的数据？比如，政府如果是出于公益或公共管理的目的使用数据，可以不受安全性之外的条件限制，但如果是出于商业目的将采集到的个人数据用于数据交易，是否就要有限制或有条件地使用？

数据使用权属于一种类似用益物权的概念，包括控制、开发、许可、转让等权利内容。然而这些权利的行使，将会受到数据所有权等其他权利的限制。比如，数据处理者应当采取加密、去标识化、匿名化等技术措施和其他必要措施来保障数据安全，在发生数据安全事件时，应当立即采取处置措施，及时告知用户并向有关主管部门报告。

个人数据使用侧重于人格权的行使与保护。数据知情同意权是个人数据权利之起点，范围包括个人数据的收集方式、收集内容、存储及处理方式等，同时也应包括收集的目的、可能对个人产生的后果，明确的同意方式及同意效力的覆盖阶段。

企业使用数据强调用权与限权的结合。将数据产权的构建放入真实的数据产业链中，数据利益在不同阶段呈现出不同的倾向。比如，“数据清洗”前，权利主体有用户与数据收集企业，利益诉求集中于用户的数据人格利益与企业的数据财产利益；在“数据清洗”后，法律视角中的权利主体只有数据收集企业，而利益诉求也只局限于数据财产利益。

企业数据使用应区分完全数据产权与定限数据产权，并且注意构建数据使用过程的权利限制体系，企业数据使用同样应当进行合理限制，从企业数据产权取得的源头就应当遵循合法采集标准。当平台企业自行收集时，采集前取得用户同意和授权；当平台企业许可他人收集时，遵循“用户授权＋平台授权＋用户授权”的三重授权原则；当收集行为未经许可发生时，是违法行为。

3. 数据收益权

通过使用数据产生巨额的经济收益，那么这份巨额收益是如何进行分配的呢？是分配给数据的产生者个人，还是赋予数据的收集加工者政府或企业呢？这个问题的回答，牵动着众多主体的利益。新浪微博起诉脉脉案件中，从其司法判决可以看出，作为投入努力和资源进行数据收集的企业，可以享有竞争法意义上的保护，即可以将该数据作为资产进行利用、许可，并从中获益，他人未经许可和授权不得随意进行信息抓取和利用。

从参与数据产业链的主体来看，共有三类主体拥有数据收益权：个人、企业和政府。

1）个人的数据收益权

个人作为原始数据的持有者，是目前被严重忽视的主体。理论上，用户作为数据的源头，天然应该具有处置其数据的决定权以及获取收益的权利。但由于个人数据价值太低，光靠货币来体现这种价值，在实践中的可操作性不强。例如，淘宝上的浏览记录具有很高的经济价值，阿里巴巴可以通过这些浏览记录有针对性地推送商品，增加用户购买的概率。如果单看一个人某天的淘宝记录可能无法衡量其所具有的价值，因此个人依靠卖单个数据来赚取利益显然是行不通的。个人如果明确同意企业收集其个人信息，那么当企业通过这些数据获得经济或其他利益时，个人也应该部分享有数据收益权，这种收益权不一定非要以现金或货币的方式体现，企业也可以为用户提供一些除货币之外的免费增值服务，例如，免费使用网站、获得免费使用券、免费参加企业活动等，方式可灵活多样。

2）企业的数据收益权

企业的数据收益权主要体现在经济利益的获得，其理论是建立在数据所有权和使用权基础之上的。企业对于其合法处理的数据享有的收益权，该权益是在不违反法律、行政法规的强制性规定以及公序良俗的前提下获得的。

3）政府的数据收益权

政府是庞大的数据收集主体，如果出于公益目的将基础数据用于开放共享，应当是免费提供。公共数据的权属归于国家，实质上归全民所有，为公共利益而存在。但如果在数据的使用上有成本投入，比如开展深度挖掘、可视化等配套服务，则可以适当收费。如果公共数据是用于商业交易目的，就完全可以收费，其收益归政府所有，并以适当形式造福于民，让民众也能享受数据红利。


微课视频

3.5 数据开放共享

3.5.1 数据开放共享现状

曾有研究表明，政府掌握了80%的社会信息资源。一般提到数据开放共享，广义上包括政府与企业之间的数据开放共享，以及企业与企业之间的数据开放共享，而狭义上指的是

政府数据开放共享。随着大数据思维与大数据应用技术不断普及，政府数据开放共享成为社会民众的权利诉求。因此，如何尽可能地将政府数据向社会公众开放、利用开放数据提高政府的社会参与度和透明度、服务经济社会生活、增加人民的福祉成为各个国家的发展战略问题。

英国、美国、加拿大等国家积极推动政府数据开放。英国在 data.gov.uk 发布了教育、健康、交通、环境、经济等方面的数据。截至 2019 年 4 月，美国开放数据平台 data.gov 已发布了 23.8 万数据集，涉及农业、气候、教育、能源、金融、健康、海洋、公共安全以及科学研究等方面。一些国际组织如联合国、欧盟、经合组织、世界银行以及开放知识基金会等也在政府数据开放方面贡献力量。欧洲议会、欧盟理事会和委员会的谈判代表就《公共部门信息重用指令》达成协议，新命名的《开放数据和公共部门信息指令》确定了一系列高价值的数据集。经合组织定期对成员国的政府数据开放情况进行调查并组织政府数据开放专家组会议，2018 年发布了《开放政府数据报告》。世界银行也发布了开放数据准备度评估工具，推动世界各国的开放数据进程。开放知识基金会作为非营利性组织，为各国政府数据开放提供了许多建设性意见。

2015 年 8 月国务院发布《促进大数据发展行动纲要》(以下简称《纲要》)，《纲要》明确提出“推动政府数据开放共享”，并将形成公共数据资源合理开放共享的法规制度和政策体系作为 5～10 年的目标，我国政府数据开放的顶层设计被正式提上日程。2016 年，国家进一步将“加快政府数据开放共享”列入《中华人民共和国国民经济和社会发展第十三个五年规划纲要》。2016 年 12 月 27 日，《“十三五”国家信息化规划》将“数据资源共享开放行动”列入优先行动。2017 年 12 月，习近平主席在中共中央政治局第二次集体学习时强调：“推动实施国家大数据战略……推进数据资源整合和开放共享”“构建以数据为关键要素的数字经济”。2019 年两会期间，“政府数据开放”成为很多代表委员关注的重要议题。

和西方国家相比，我国的开放数据政策建设较为欠缺。虽然也通过国家大数据战略明确了政府数据开放的方向，但国家政策涉及的内容仍不具体。除国家大数据战略外，我国并没有专门的政府数据开放政策，相关的政策内容较为零散、缺乏系统性。各级政府的开放数据缺乏统一的标准规范，导致诸多地方政府开放数据步伐很慢，严重阻碍了我国开放数据的进程。因此亟待在国家大数据战略的指导下形成目标明确、层次分明并且切实可行的政策体系。

3.5.2　数据开放共享的主要方式

1. 数据开放

数据开放主要是指政府数据面向公众开放。该方式主要适合于非敏感、不涉及个人隐私的数据，并且需要保证数据经过二次加工或聚合分析后仍不会产生敏感数据。

政府数据不仅仅涉及政府在日常政务开展过程中形成的一些数据，还应包括公共部门(如图书馆、医院、学校、第三方服务机构)在经营过程中形成的数据。广义层面的政府数据，涉及到公共部门内部数据以及在运营过程中生产、创造、收集、处理和存储的数据。

2. 数据交换

数据交换主要是政府部门之间、政府与企业之间通过签署协议或合作等方式开展的非营利性数据开放共享。一般有两种情况，一是为信用较好或有关联的实体之间提供数据交

换机制，由第三方机构为双方提供交换区域技术及服务，这种交换适用于非涉密或保密程度比较低的数据，另一种是针对敏感数据封装在业务场景中的闭环交换。通过安全标记、多级授权、基于标准的访问控制、多租户隔离、数据族谱、血缘追踪及安全审计等安全机制构建安全的交换平台空间，确保数据可用不可见。

3. 数据交易

数据交易主要是对数据明码标价进行买卖，目前市场上比较多的第三方数据交易平台提供的主要是这种模式。基于大数据交易所的交易模式是目前我国大数据交易的主要模式，比较典型的代表有贵阳大数据交易所、长江大数据交易所及东湖大数据交易平台等。

3.5.3 数据开放共享困难重重

2018 年 7 月，国务院印发了《关于加快推进全国一体化在线政务服务平台建设的指导意见》，对政务信息共享提出了科学的目标和要求，数据应用开放的关键是打破数据孤岛，让数据互联互通，达到数据和信息共享。

但是，当前我国政府信息化建设依然存在标准不一致和重复建设的问题，部门条块分割比较严重，各部门之间沟通困难。出于权限和利益问题的考虑，很多单位将政府数据资源部门化、专属化、利益化，导致“数据割据”问题严重。例如，我们每个公民的个体信息分别掌握在工商部门、银行、保险、公安、医院、社保、运营商等不同机构手里，真要打通和融合各个部门掌握的数据是很困难的事情。

近年来，数据安全问题频频发生，大到给国家安全和经济社会发展造成严重的潜在危害，小到给公民个人造成巨大的精神伤害和经济损失。正是由于数据在收集、存储、使用、交换和销毁等各个环节都存在极大的安全隐患，很多政府部门和大型互联网企业在数据开放共享中都心存忧虑，担心因数据泄露或遭黑客攻击而带来严重后果，不敢推动数据开放进程。

除了对数据泄露等安全事件的恐惧，还有些出于对数据伦理的考虑。

【案例 3-10】 2018 年 10 月，科技部官网公布了对复旦大学附属华山医院、华大基因、药明康德、昆皓睿诚、厦门艾德生物、阿斯利康等 6 家单位的行政处罚。涉事单位阿斯利康未经许可将已获批项目的剩余样本转运至厦门艾德生物医药科技股份有限公司和昆皓睿诚医药研发有限公司，开展超出审批范围的科研活动。厦门艾德生物未经许可接收阿斯利康投资 30 管样本，拟用于试剂盒研发相关活动，而昆皓睿诚则未经许可接收阿斯利康 567 管样本并保存。这是科技部首次公开涉及人类遗传资源的行政处罚。

这种带有人类生物特征的数据，是一个国家、每一个公民所不可触碰的资源底线，这类数据的开放共享将涉及更多既有挑战性又复杂的问题，加重数据主体在开放共享数据时的顾虑。

此外，数据质量问题也是阻碍数据开放共享的又一“拦路虎”。改革开放以来，我国政府进行了一系列改革，政府统计数据正在朝着越来越全面、客观反映国家经济社会发展情况的方向发展。但统计工作是一个比较复杂的系统工程，需要多个部门加强配合协调，按计划进行统计信息的收集、汇总和分析，才能形成统计数据分析结论。只要其中一个环节出现问题，就会直接导致统计数据的准确性下降同时。缺乏明确的解释和统一的统计口径，也会导致统计数据混乱。很多数据即使收集起来，也无法进行对比分析和统一转化，因而直接影响

了政府统计数据的全面性、真实性和准确性，损害了政府公信力和权威形象。因此，有些部门和单位为了不承担数据开放共享后数据质量存在问题所带来的麻烦，宁可不开放共享。

从政策层面上来讲，目前关于数据开放共享的法律法规也十分匮乏。关于数据开放共享，国家和地方出台了一些管理制度，但主要针对政府行为。对于数据开放共享的原则、数据分类和开放边界、数据格式、质量标准、互操作性等尚没有明确的规范，而且数据在采集、传输、存储、处理、交换甚至销毁等各个阶段，其所有者和使用者往往都不同，存在数据所有权和使用权分离的情况，很容易导致数据滥用、数据权属不明确以及无法进行数据定价等问题。这些问题都导致数据开放共享难以操作，出现问题也找不到相应的法律依据加以解决。

3.6　大数据科技人员的伦理责任


微课视频

3.6.1　大数据伦理责任特点

大数据伦理责任是具有普遍意义的伦理责任在大数据时代的具体化表现。因此它具有责任伦理的一般特征。同时，由于数据管理和网络社会自身的自由性、开放性和虚拟性的特点，数据伦理责任又有自己的特殊性，表现为自律性、广泛性和实践性。

3.6.2　大数据科技人员的伦理责任

大数据科技人员的伦理责任主要表现在以下五个方面。

1）尊重个人自由

大数据时代，尊重个人自由，很大程度上表现为自觉地、发自内心地尊重个人隐私，遵守伦理道德。

2）强化技术保护

通过不断完善信息系统安全性能，部署防火墙、入侵检测系统、防病毒系统、认证系统，采取访问过滤、动态密码保护、登录限制、网络攻击追踪方法等技术手段，强化应用数据的脱敏处理、存取管理和业务审计，确保系统中的用户个人信息得到更加稳妥的安全技术防护。

3）严格操作规程

制定严密的数据管理和追责制度，包括数据获取、清洗存储、传输、分享、交易、关联分析等环节的权限管理和访问日志，规范所有能接触到数据及算法的人员的操作行为。同时，对于重要和关键数据，要建立多重访问控制规则，提高信息外泄成本，降低风险。

4）加强行业自律

努力培育和强化行业自律机制，发挥行业自律的灵活性和专业化优势，弥补法律法规滞后的缺陷，重点行业应制定自律规范和自律公约，规范大数据的使用方法和标准流程。

5）承担社会责任

共同承担建设安全可信、平等、以及惠民的大数据社会的责任，避免发明伤害他人。涉嫌歧视、损害名誉、降低道德水平的大数据产品和服务，在企业私利和社会公德之间履行好大数据科技创新人员的社会责任。

3.6.3 大数据科技人员的行为规范

针对网络和大数据应用对社会生活产生的巨大作用，2014 年 6 月，IEEE 发布《国际电气电子工程师学会行为标准》中提出 5 项规范：

（1）尊重他人；

（2）公平待人；

（3）避免伤害他人、财务、名誉或聘用关系；

（4）克制而不报复；

（5）遵守与 IEEE 有业务往来的各国适用法律及 IEEE 的政策和流程。

文中特别提到要尊重他人隐私，保护他们的个人信息和数据，不在现实生活和网络空间中做危害人类的事情，不用错误或恶意的方式侵害他人身体、财产、数据、名誉和聘用关系，不在网上和其他场所传播关于他人的恶意谣言、诽谤、污言秽语和物理伤害。与此前 1990 年发布的《国际电气电子工程师学会伦理条例》的 10 条规定相比，更突出网络和大数据的特点，更落实到行为层面。因此，具有更好的指导性和可操作性。表 3-1 罗列了针对大数据生命周期的不同阶段，合乎大数据伦理的行为规范。

表 3-1 大数据生命周期不同阶段所遵守的行为规范

	原则 1	原则 2	原则 3	原则 4
数据采集阶段	知情同意原则：用户有权作出 Y/N 选择	自由选择原则：用户可以决定数据被采集的范围	随时可删原则：用户随时可以彻底删除数据	——
数据交易阶段	明白交易原则：明确所交易的数据范围、使用规范、定价策略等	保证用户数据安全：脱敏个人信息并安全存放、传输	再交易告知原则	使用可审计、权利可撤销原则
数据应用阶段	隐私保护原则	价值维护原则	——	——

思考讨论

1. 大数据有哪些特点？

2. 随着大数据的兴起，可能带来哪些道德伦理问题？

3. 如何应对大数据技术带来的伦理问题？

4. 在日常生活中如何做好个人信息保护？

5. 个人信息泄露的根源和途径是什么？

6. 浅谈大数据背景下个人信息的法律保护。

7. 分析一下数据安全的现状以及如何保障数据安全？

8. 大数据时代，企业收集数据的方式多种多样，网络爬虫采集是其中之一。网络爬虫（Web Crawler）是依照一定规则主动抓取网页的程序，是搜索引擎获得信息的渠道之一。浅谈网络爬虫的利与弊？

9. 数据产权是什么？怎样做好数据产权的界定？

10. 数据开放共享如何实现？它有什么意义？

11. 现阶段对数据开放共享有哪些困难与挑战？

12. 大数据科技人员的伦理责任主要表现在哪些方面？

13. 据不完全统计，包括滴滴出行、携程、飞猪、美团、京东在内的多家互联网平台均被曝出存在“杀熟”的情况。“千人千价”也深刻改变了商家和消费者之间的关系。请问大数据是如何杀熟的？我们该如何避免“被杀”？