# 从数据泄露看大数据的利与弊

盛熙然 王乐 冯海桐 丁春浩

2025年2月8日

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

# 目 录

# 01

# 大数据简介

# 大数据的好处

随着数字时代的不断推进，大数据的使用已经成为了我们生活中不可或缺的科技工具，应用在方方面面。

大数据可以用来作为决策参考依据，公司和政府机构可以通过分析需求，总结经验来提升产品或服务，帮助做出正确的决策。

大数据可以用于效率提升，在工厂和公司中，大数据可以用来判断生产的各个环节的重要性并进行比较，由此可以优化生产流程，提升生产效率。

大数据也为用户带来了使用体验的提升，软件可以通过大数据分析用户的喜好和关注点，为用户提供更优质，更加个性化的服务。

# 大数据的问题

然而，随之也涌现出了诸多问题，其中非常严重的一项就是数据泄露。

无论是公司，企业还是政府，在利用大数据进行分析时，其数据来源和保密性都存在着问题。在收集大量数据的同时，用户的隐私很可能被泄露。

# 02

# 数据泄露

# 数据泄露的概念

数据泄露指的是私人信息或者机密数据在不知情或者未经许可的情况下从系统中被获取的数据安全事件。对个人主要是指社会保障号码、银行账户、财务数据、医疗保健信息、知识产权和客户记录等。

数据泄露发生的原因有外部行为和内部行为两类：
外部行为主要包括恶意软件，网络钓鱼和社会工程等。
内部行为则大部分由于访问权限的滥用。

对于企业来说，没能妥善保存用户数据会损害其信誉，破坏用户的信任，严重的则需要承担赔偿责任等。
对于个人来说，个人信息被泄露可能被骚扰和攻击。

# 举例说明

1. Web 服务供应商：
从 2013 年到 2016 年，一家大型美国 Web 服务提供商成为有记录以来几乎最大规模的数据泄露攻击目标。黑客通过一系列包含链接的电子邮件获得了对全部 30 亿用户姓名、出生日期、电话号码、密码、安全问题和答案以及电子邮件地址的访问权限。直到该公司被收购之后，此次数据泄露的程度才得以公开，这导致收购要约减少了 3.5 亿美元。

2. 征信机构：
2017 年，黑客侵入一家美国征信机构，窃取了超过 1.47 亿美国人的个人数据。如今，它被认为是与身份盗窃相关的最大网络犯罪之一。网络攻击者首先获得了网络访问权限，然后进入其他服务器访问个人信息，包括社会保障号码、驾照号码和信用卡号。最终，该公司支付了 14 亿美元的罚金和损失修复费用。

# 负面影响

然而，在我国，对个人信息的保护远不够，未能有效地推进收集到个人信息的企业和公司对于这些数据的保护。这主要体现在相关法律制定不够完善，以及处罚力度不够等原因，相较于国外对于个人信息的保护以及违法的处罚力度，我国在这方面还远远不够。

还有一个导致信息泄露的原因就是“信息交易”，用户的私人信息被当作商品在企业公司之间进行交易，而从中获得利润。

然而类似的行为最终结果少有刑事处罚，罚款也在几万到百万左右。

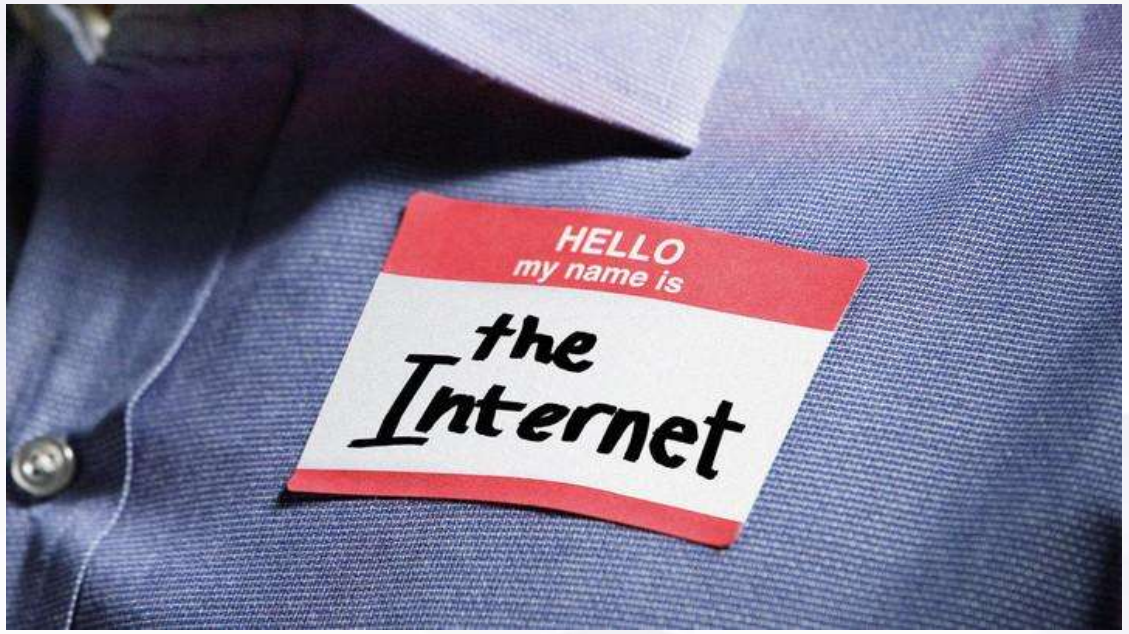
# 03

# 引申问题

# 引申问题

同时，在使用这些信息的过程中也存在着一些问题。
这体现在对于用户的隐私权利的保护中。
即在使用这些数据的过程能否保证数据的公平性、透明性和责任性？

# 公平性 – 信息标签




给用户打标签在互联网时代已经司空见惯

可能信息标签对于你来说比较陌生，换个说法，用户画像是不是十分熟悉。

品牌给用户打标签，其实是一件司空见惯的事情。而且不管你接受与否，大概率是避免不了的了。

# 常见的数据标签

**互联网数据“杀熟”：**
互联网厂商利用自己所拥有的用户数据，在出售同一件商品或同一种服务时，老用户的价格要高于新用户。

**各个品牌应用：**
用户消费画像、用户行为画像、用户兴趣画像、金融领域：风险画像，电商领域：商品的类目偏好、品类偏好、品牌偏好。

**个性化推荐的依据：**
视频平台的个性化推送游戏的老用户回归奖励老乡鸡的“老友券”。高品质用户、羊毛党等等三六九等的划分。

# 案例一

## “大数据杀熟”第一案

2020年7月，胡女士通过某旅游平台App订购了舟山一家酒店，支付款2889元，但在离开酒店后，发现酒店的实际挂牌价仅为1377元，差价达一倍多。最后，法院按“退一赔三”标准予以支持胡女士的赔偿诉求。此案被行业称为“大数据杀熟”第一案。

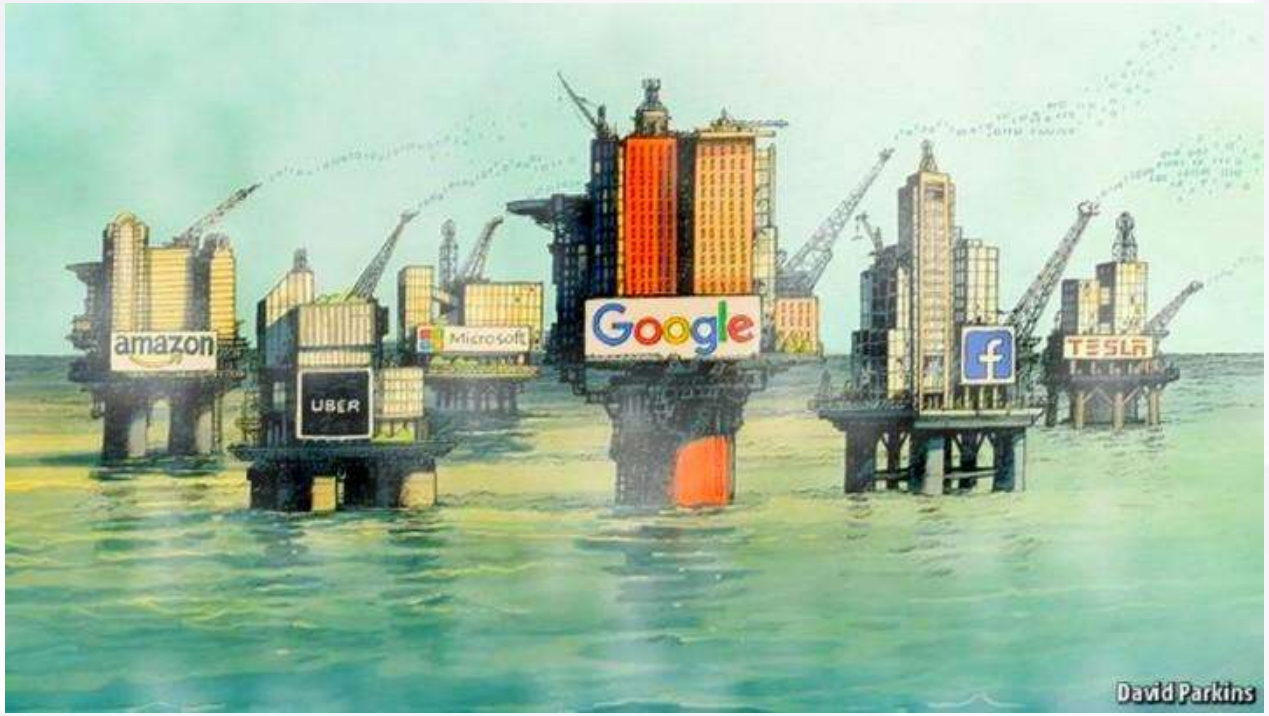
# 案例二

2021年11月，《中华人民共和国个人信息保护法》开始实施，作为中国首部指向个人数据保护的立法，规定要求完善互联网注册制度，对诱导沉迷大数据杀熟等领域做了全面要求和规范。

京东、美团、饿了么以及滴滴等数十家企业签署了《平台企业维护公平竞争市场秩序承诺书》，并作出承诺：不非法收集、使用消费者个人信息，不利用数据优势“杀熟”。

现在个人信息得到了良好的保护？企业做到了它的承诺？

# 数据标签的价值




英国数学家 Clive Humby 曾说过——数据是新时代的石油。

不同的是，石油是不可再生的能源。而标签数据，更像可再生能源。

# 实际案例

## 用户数据创造的利益用户能否分一杯羹

加利福尼亚州州长 Gavin Newsom 就曾提出了一项雄心勃勃的「数据红利」计划，希望 Facebook 和 Google 这样的大型互联网公司将向用户支付一部分来自用户数据的收入。

Facebook 的联合创始人 Chris Hughes 也曾经用因石油而诞生的阿拉斯加永久基金类比互联网。（阿拉斯加每年会把至少 25% 的矿产租金、专营税、联邦矿产等收入放进永久基金，再作为分红分给其居民。）

# 小结

希望在不远的未来，用户能够享有选择权，有权利去关闭采集标签信息和建立用户画像；用户能够享有知情权，能够清楚知道各个厂商给自己的标签和画像；个人信息能够更完善的被保护，更合理的被使用。

# 透明性 - 信息透明度

信息透明度，通常指的是在信息流通和管理过程中，能够有效公开、共享和传递信息的程度。在当今社会，得益于数字技术和互联网的广泛应用，信息透明度得到提升，公众、企业和政府等不同主体之间的信息交流变得更加快捷与方便。

# 信息透明度的好处

## 增强信任与公正

在政府和企业中，公开透明的决策过程有助于防止腐败、贪污等不正当行为的发生。公众对于决策者和管理者的信任会随着信息的公开程度而增强，从而增强社会的整体公正性。

## 促进效率与决策的优化

在信息高度透明的环境中，决策者能够获得更加准确、全面的信息，从而作出更加科学、合理的决策。

## 保障个体权益

在电子商务、金融等领域，信息透明能够帮助消费者做出明智的选择，避免遭遇欺诈和不公平待遇。在健康医疗、社会福利等方面，透明的政策和信息能够保障公民知情权，避免因信息不对称导致的不公平现象。

## 促进创新和竞争

信息透明化降低了行业壁垒，使得市场参与者能够更加了解竞争对手的动态，促进技术创新和市场竞争。

# 信息透明度的弊端分析

尽管信息透明度有诸多好处，但从计算机伦理学的角度来看，其过度追求透明化可能带来一系列伦理和社会问题。以下将重点分析信息透明度带来的潜在弊端。

- 个人隐私的侵犯
- 信息的误用与滥用
- 社会的不平等加剧

# 透明度过高的弊端

1.侵犯个人隐私：数字化时代，涉及到个人身份、行为等的信息经常被采集和存储在各种系统中。而信息透明度高可能导致过度曝光，特别是在社交媒体等领域。而这些信息的泄露可能进一步产生上述的不公平的信息标签化的问题。

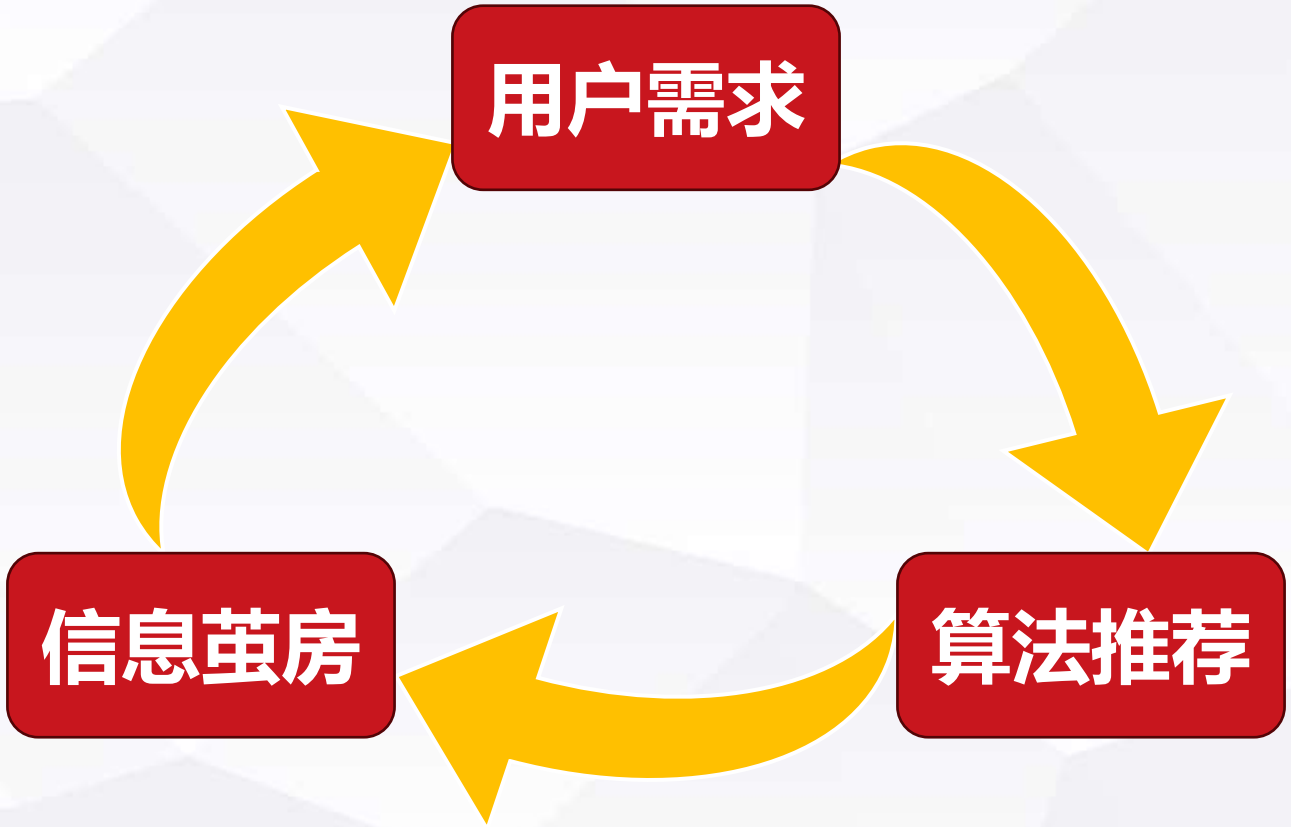
2.信息的误用和滥用：在信息高度透明的环境中，数据的准确性和可验证性成为至关重要的因素。如果信息被错误地传播或者被恶意篡改，它可能会被用于误导、操控或煽动公众。

3.社会的不平等加剧：由于社会中的不同群体对于信息的掌握和解读存在差异，透明度的提升可能会加剧社会的不平等。技术、知识和教育水平较低的群体，可能无法有效理解和利用透明化的公共信息，从而在面对决策时处于不利地位。

# 责任性 - 信息茧房

信息茧房是指人们关注的信息领域会习惯性地被自己的兴趣所引导，从而将自己的生活桎梏于像蚕茧一般的“茧房”中的现象。

# 信息茧房的成因



# 信息茧房的弊端

## 信息窄化：

算法推荐技术将用户所接触到的信息赋以三种标签：用户明确感兴趣的准推荐类内容，用户可能感兴趣的待推荐类内容，用户明确不感兴趣的不推荐类内容，基于用户留存的目的，平台需要对用户获取内容进行比重控制，并定期对用户采取激励手段。

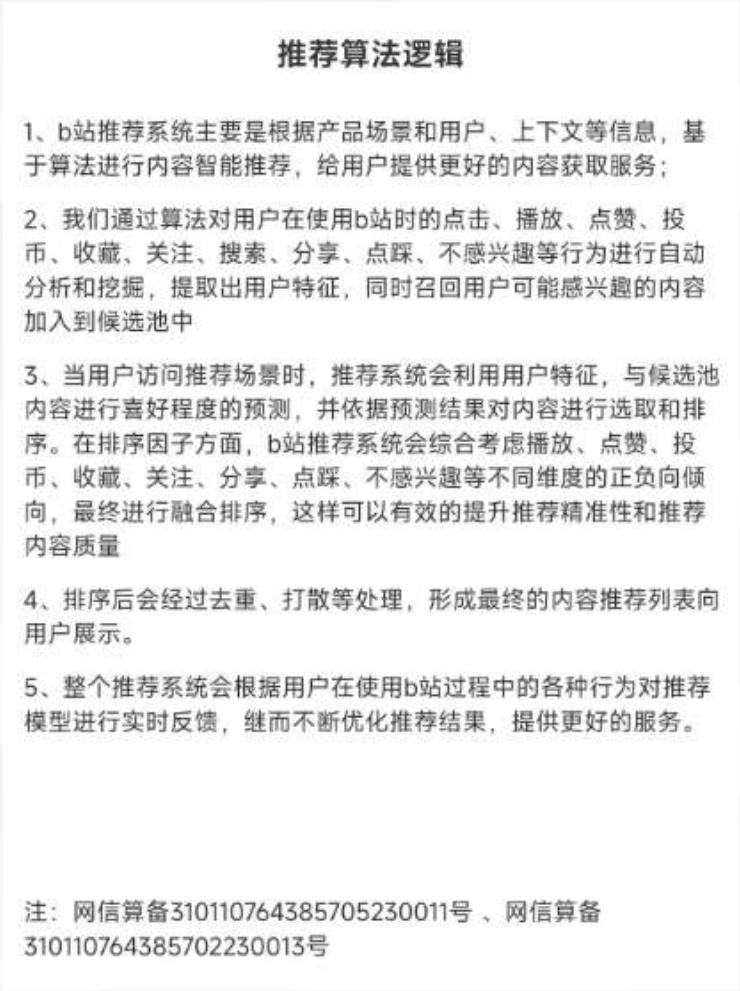
## 群体极化：

不论群体是用户主观选择加入，或是平台标记用户而被动形成，均以群体的形式，将用户限制在定制化和封闭化的信息环境中，在单一声音构成的内容环境中，长期处于与自身持有相似的观点的群体内，受沉默的螺旋和回声室效应作用，用户将受自身所处群体一贯持有态度影响，从而不断强化自身态度。

# 数据收集与推荐算法

深入整治“信息茧房”、诱导沉迷问题。构建“信息茧房”防范机制，提升推送内容多样性丰富性。严禁推送高度同质化内容诱导用户沉迷。不得强制要求用户选择兴趣标签，不得将违法和不良信息记入用户标签并据以推送信息，不得超范围收集用户个人信息用于内容推送。规范设置“不感兴趣”等负反馈功能。

——**《四部门：开展“清朗·网络平台算法典型问题治理”专项行动》**



推荐算法逻辑

1、b站推荐系统主要是根据产品场景和用户、上下文等信息，基于算法进行内容智能推荐，给用户提供更好的内容获取服务；

2、我们通过算法对用户在使用b站时的点击、播放、点赞、投币、收藏、关注、搜索、分享、点踩、不感兴趣等行为进行自动分析和挖掘，提取出用户特征，同时召回用户可能感兴趣的内容加入到候选池中

3、当用户访问推荐场景时，推荐系统会利用用户特征，与候选池内容进行喜好程度的预测，并依据预测结果对内容进行选取和排序。在排序因子方面，b站推荐系统会综合考虑播放、点赞、投币、收藏、关注、分享、点踩、不感兴趣等不同维度的正负向倾向，最终进行融合排序，这样可以有效的提升推荐精准性和推荐内容质量

4、排序后会经过去重、打散等处理，形成最终的内容推荐列表向用户展示。

5、整个推荐系统会根据用户在使用b站过程中的各种行为对推荐模型进行实时反馈，继而不断优化推荐结果，提供更好的服务。

注：网信算备310110764385705230011号 、网信算备310110764385702230013号

# 04

# 总结

# 总结

即使法律上对于数据泄露及其引申问题的管控不够，我们也可以从伦理道德角度对其进行批判。

义务论：在公司成功地获取到用户的私人数据之后，其便有义务保护和利用好这些数据，这不仅体现在网站或者应用中包含的用户协议中，更体现在一种信任关系上，公司获得这些数据是为了更好地盈利，用户提供这些数据是为了得到更好的服务。

结果论：公司的数据泄露不只是可能使得用户被信息骚扰，更有可能招致安全隐患。对公司来讲，会破坏与用户间的信任关系，以及利润的下降，从结果上看，同样是错误行为。


上海交通大学

SHANGHAI JIAO TONG UNIVERSITY

# 感谢聆听

饮水思源 爱国荣校