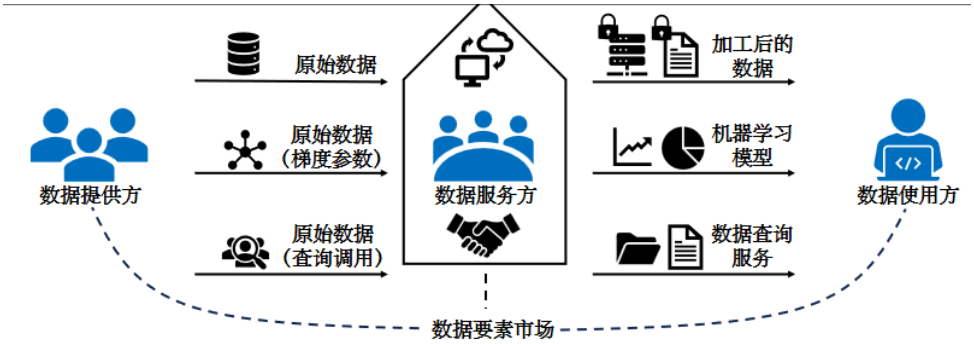
Lecture 9：数据定价基础

一、数据交易的基本框架与数据定价的要求



数据服务方（Data sever）/ 数据交易平台（Data marketplace）/ 数据中介（Data broker）

- 承担数据收集、管理、合规与安全等任务，保障数据交易正常运行
- 将数据提供者的数据转化为产品（查询、机器学习模型），出售给数据使用方，获得收益，将收益按公平分配规则（之后会介绍）分配给数据提供方（可以提取部分收入作为处理费用与利润）
- 现实中的数据交易通常有中介，但也存在无中介的买卖双方直接交易的简单场景

数据提供方（Data owner，或 Data seller，即数据卖家）：为数据交易提供原始数据

- 原始数据可能经过加工得到查询结果或机器学习模型
- 加工后出售的一份数据可能来源于很多数据提供者的数据，因此收益需要根据公平分配的原则分配给这些数据提供者

数据使用方（Data user，或 Data buyer，即数据买家）

- 向数据中介提出购买数据产品的需求
- 为了得到需要的数据产品，需要付出一定的价格

针对数据的特性，数据定价的挑战，在数据交易的框架下，数据定价需要满足哪些要求？保持传统市场中定价的基本要求

- 效用最大化 / 帕累托最优 / 预算约束 / 社会福利最大化：之前已经介绍
- 平衡预算（budget balance）：构建市场的成本应当小于等于市场的收益，否则数据市场的建立会带来“财务赤字”
- 个人理性（individual rationality, IR）：市场中每个参与人参与市场的效用比不参与市场的效用高
  - 近似而言，不参与市场的效用可以视为 0，因此个人理性的要求就是每个参与人的参与收益大于成本
  - 但是数据的外部性比较明显，因此即使不参与市场也可能产生效用
  - 例如与你竞争的公司在数据市场中购买了机器学习模型，可以预测下一季度客户对不同商品的偏好，如果你没有买，会因此丧失客户，因此不参与市场的效用为负数
- 无嫉妒（envy-free）：一个数据买家不会嫉妒另一个数据买家的情况，即他不会觉得用另一个人支付的钱买那个人的产品，比用自己支付的钱买自己的产品更值得：来源于经济学中经典的分蛋糕问题，参见 Selfridge-Conway 算法

二、数据的版本化与无套利原则

- 免费数据

  - 一些公开的统计数据，不具有出售的价值
  - 但是可以吸引用户进入市场，进而吸引数据提供者进入市场
- 根据使用次数决定价格
  - 每次调用 API，都收取一定的价格
  - 类似于咨询费，咨询 1h 给多少钱，咨询 3h 给多少钱
  - 从供给成本角度看，数据可以零成本复制，例如机器学习模型训练好之后，尽管训练成本可能很高，但每次调用 API 需要的成本几乎为零，因此当 API 调用越多时，边际成本也会下降
  - 但从需求效用角度看，每次买家调用 API 是为了新的数据，那么存在数据需求，因此也有合理性
- 打包定价：
  - 以固定的价格出售一定次数的 API 调用权
  - 上一种方法的升级版
- 订阅制
  - 按订阅时间收费，包含两种收费机制
    - 一类是订阅后权限全部开放
    - 另一类是支付固定订阅费用后，还要加收每次服务的费用（结合订阅制和前面的使用次数定价），类似于移动公司，软件许可等。固定费用用于覆盖固定成本，服务费用提供利润
  - 可以比较有效地解决动态定价问题
  - 产生订阅循环，锁定用户
- 免费增值（Freemium）
  - 用免费服务吸引用户，将用户锁定在平台，然后通过增值服务将部分免费用户转化为收费用户实现变现
  - overleaf，淘宝旺铺服务
  - 不需要再花费大笔的广告费用向用户介绍你服务的各种特性，而是通过用户的免费使用，让用户进行对你提供服务的自我学习以及熟悉
  - 让一部分高级用户支付普通用户使用产品所消耗的费用（二八定律），因此要思考如何提高付费转化率
- 特点总结
  - 有很多经典营销策略，因为参考了类似的情况。服务定价（例如软件使用、移动公司）或数字商品（例如视频平台）也具有高固定成本，低边际成本的特性，而视频平台的内容也具有动态更新的性质
  - 从成本导向转向需求导向。零成本复制使得考虑数据定价时，固定成本只是一个预算平衡的限制，并非核心要点如上方法基本都从买家需求出发，目标是吸引买家，给予买家更好的体验
- 根据使用次数定价以及订阅制的原理比较明显
  - 根据使用次数定价与传统定价方式类似，是一种非常 naive 的方案
  - 订阅制有效解决了动态定价的问题，同时可以留住客户，并且以买家为中心，因为买家感到不满则可以未来退出订阅，影响市场盈利，不像买断那样服务可能越来越差

数据的版本化

- 原始数据版本化：不用一次出售全部数据，可以分成用户感兴趣的几块出售，或根据数据的关键性对数据进行分级分类出售，或添加噪声构造新版本
- 查询数据版本化：可以为任意的 SQL 查询定价
- 机器学习模型版本化：向训练的模型中添加噪声影响模型准确性，从而生成不同版本的模型

版本化的好处：

- 买家侧

- 面向买家的定价策略：买家的选择更自由，可以只购买自己感兴趣的部分
- 有预算约束的买家可能买不起最好的数据，但可以购买低级别的数据

- 卖家侧
  - 有效利用数据零成本复制的特点，构建不同类型的数据产品，吸引不同类型的买家
  - 网络外部性（network externalities），或“需求面的规模经济”(demand-side economies of scale)：使用某一产品的消费者越多，个体消费者在使用该产品时获得的效用就越大
    - 如果其他人没有传真机，那么你购买传真机毫无意义
    - 规模经济：通过扩大生产规模而引起经济效益增加的现象
    - 微软版本化出售操作系统，低价版本吸引客户，建立 Windows 用户生态，越来越多的人愿意用 Windows 系统。增加用户或服务的成本可以忽略，但愿意购买付费服务的用户增加，故实现规模经济
    - 但是在数据市场中似乎看不到网络外部性的好处？
  - 增大卖家利润：价格歧视理论
    - 经济学中的定义：按不同价格出售不同单位的产量
    - 第二级价格歧视：非线性定价，即每单位产品价格不一致，例如批量购买可以打折扣
    - 数据市场中：为较高支付意愿买家出售高价优质数据，为较低支付意愿买家出售低价次级数据
    - 个性化定价：从成本导向转向需求导向，根据用户需求出售产品、制定价格

版本化带来的问题：套利的可能

- 套利的严格定义
  - 套利这一名词来源于金融学，是指利用一个或多个市场存在的各种价格差异，在不冒任何损失风险且无需投资者自有资金的情况下有可能赚取利润的交易策略（或行为）
  - 前面买香蕉的例子可以匹配上面的定义：我们现从卖家手中分三次借到三根香蕉，然后将这三根香蕉捆绑以 10 元出售，最后分三次还 9 元给卖家，这样可以无风险地“空手套白狼”
  - 在金融学中，这叫先“做空”三根香蕉，然后卖出，获得的收益比要还的钱多，从而盈利
  - 数据市场的研究中没有使用这么严谨的定义，事实上你先分三次买入三根香蕉然后捆绑卖出就是一次套利行为，也即市场中存在套利机会
  - 套利行为的存在可能导致非预期的行为，金融学理论研究的重要基石就是无套利原则
- 无套利原则（arbitrage freeness）
  - 即市场中不允许出现如上的套利机会
  - 无套利原则在不同的数据类型上有相同的本质，但内涵略有差异的定义
  - 接下来我们分别介绍查询数据和机器学习模型的无套利定价

三、查询数据的版本化与无套利定价

四、机器学习模型的版本化与无套利定价