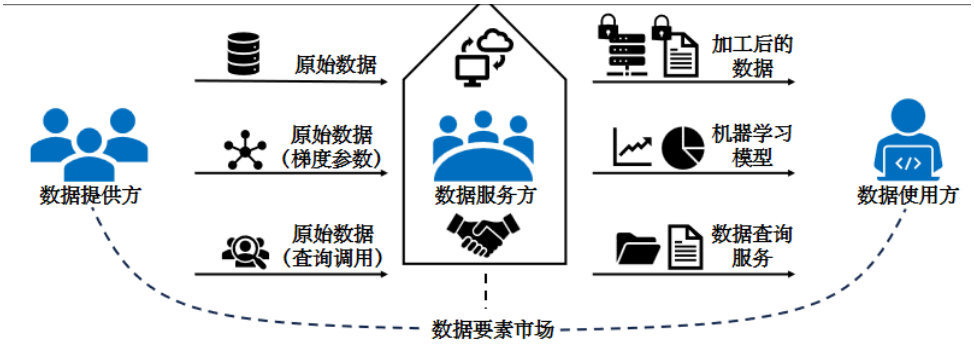
Lecture 9：数据定价基础

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

考虑一个数据卖家需要出售一个关系实例（即一个表格）$D \in I$（其中I是卖家拥有的全体关系实例），数据买家可以通过查询来购买数据

- 买家可以通过对D进行一组查询$Q = (Q_1, ..., Q_p)$来购买数据，其中Q称为查询向量（query vector）
- 每次查询（一个 SQL 语句Q可以视为一个确定性函数，将表格D映射到答案 $Q(D)$，从而整个查询向量得到的答案可以记为 $Q(D)$

一个定价函数 $p(Q, D)$ 根据查询输入的向量以及表格决定出一个价格

直观：如果一个查询向量Q_1得到的结果是查询向量Q_2的子集，那么Q_1的价格一定低于Q_2的

定义：假设$D \in I$是一个表格，如果对于任意满足$D_2(D) = Q_2(D')$ 的$D' \in I$，都有$Q_1(D) = Q_1(D')$，则称在表格 D 下查询向量Q_2可以决定Q_1.

直观来看，查询向量Q_2可以决定Q_1表明Q_2比Q_1更强，或者说Q_2包含了Q_1，这与我们的直觉相符。

定义：如果对于所有满足在表格D下查询向量Q_2可以决定Q_1的表格 D，都有

$$p(Q_2, D) \geq p(Q_1, D)$$

则称定价函数p是信息无套利的.

组合无套利研究买家进行两次分别查询与一次性查询两次的结果之间的关联

- 可以和买香蕉的故事进行类比，简单而言就是一次查询如果可以拆成两次查询完成，那么一次查询的价格不能大于两次查询的价格之和，否则买家可以通过分别两次查询绕开一次查询的定价，实现套利
- 我们记$Q = Q_1 \parallel Q_2$为查询Q_1和Q_2的连接，即查询 Q 可以拆成Q_1和Q_2两次查询分别执行

定义：如果对于所有表格D，都有

$$p(Q_1 \parallel Q_2, D) \leq p(Q_1, D) + p(Q_2, D)$$

则称定价函数p是组合无套利的.

事实上，判断在表格D下查询向量Q_2是否可以决定Q_1是很难计算的，为了解决这一困难，我们给出如下定义便于进一步讨论：

定义：令$S \subset I$是任意一个子集，称之为支撑集（support），并定义Q关于S的冲突集（conflict set）为

$$C_S(Q, D) = \{D' \in S \mid Q(D) \neq Q(D')\}$$

我们可以将任意一个查询 Q 映射到对应的 S 上的捆（bundle），事实上捆的计算在 S 比较小的时候是很容易的，因为只需要对 S 中的每个表格逐个验证是否满足上述定义即可。

定义：对于一个集合函数$f : 2^S \to \mathbb{R}^+$，即将 S 的一个子集映射到正实数的函数，我们称函数 f 是满足

- 单调性（monotone）的，如果对于任意的$A \subset B$，都有$f(A) \leq f(B)$
- 次可加性（subadditive）的，如果对于任意的集合 A 和 B，都有$f(A \cup B) \leq f(A) + f(B)$

定理：令$S \subset I$，并令$f : 2^S \to \mathbb{R}^+$是一个集合函数，则定价函数$p(Q, D) = f(C_S(Q, D))$是无套利的，当且仅当函数f是单调、次可加的.

由此我们得到了一个定价函数无套利的等价条件，事实上单调、次可加性对于函数而言并非很严苛的要求，因此这个等价条件在设计定价函数时有很大的帮助。需要注意的是，这里的函数f取值于冲突集。

为了实现最大化利润，我们要对买家的行动给出一定的假设

1. 我们假设每个买家只购买一组查询Q的查询结果，现实中一个买家查询多次可以视为多个不同买家
2. 买家只有当$p(Q_1, D) \leq v_Q$时才会购买，其中v_Q是买家对Q的估值，即价格要小于等于估值才购买
3. 假设市场是完全信息的，即卖家知道买家的心理价位，这可以通过市场调研得到

现在的问题转化为：在已知所有可能购买查询产品的买家及其估值后，如何利用这一信息，最大化利润，同时满足前面的无套利条件

1. 这里我们假设市场调研的买家就是所有可能购买的买家，这也是一种经典的简化方式
2. 当然数据量很大时可以尝试拟合曲线解决
3. 为了接下来讨论方便，我们假设有m个买家，每个买家i希望购买查询Q_i

选定支撑集$S \subset I$，S的大小为$n = |S|$，我们构建超图（hypergraph）$H = (V, E)$

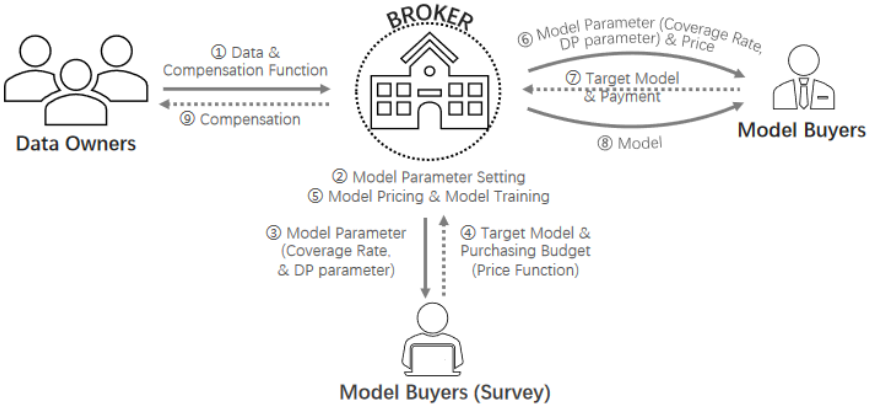
- 超图是一种广义上的图，它的一条边可以连接任意数量的顶点（在普通图中，一条边只能连接两个顶点）
- $V = S$
- $e = \{e_i \mid i = 1, ..., m\}$，其中$e_i = C_S(Q_i, D)$

目标问题：找到一个满足单调性、次可加性的集合函数f可以最大化卖家利润

$$\mathrm{OPT} = \max_{\text{单调、次可加的} f} R(f)$$

其中$R(f) = \sum_{i:v_i \geq f(e_i)} f(e_i)$，即在集合函数$f$下价格小于等于买家估值的价格之和

四、机器学习模型的版本化与无套利定价



数据拥有者: 关心隐私保护程度和收益分配的公平性

- 用ε-differential privacy 来量化 Data owners 的隐私泄露风险
- 用一个单调增的 compensation function c_i 来建模 data owner D_i 对参与训练的、满足ε-differential privacy 的机器学习模型（数据产品）要求的 compensation。

$$c_i(\epsilon) = b_i \cdot (e^{\epsilon})^{\rho_i}$$

ρ_i：卖家即Data owner的Privacy sensitivity

ϵ：Privacy budget

b_i：Base price (来自Shapley value)

$\rho_i \uparrow$ Privacy sensitivity高，增加privacy budget时要求的compensation增长快

模型购买者: 关心最终模型（数据产品）的效用

如何估计模型的最终效用？

- 用参与训练数据的 Shapley value 覆盖率 CR

$$CR(M) = CR\left(D_{i_1}, ..., D_{i_k}\right) = \frac{SV\left(\left\{D_{i_1}, ..., D_{i_k}\right\}\right)}{SV(\{D_1, ..., D_n\})} = \frac{SV\left(\left\{D_{i_1}, ..., D_{i_k}\right\}\right)}{U(\{D_1, ..., D_n\})}$$

- 为了满足 differential privacy 所添加的噪声，用 Privacy Budget 来量化ε

$$P(B_j, M) = V_j \cdot \frac{1}{1 + e^{-\delta_j(CR(M) - \theta_j)}} \cdot \frac{1}{1 + e^{-\gamma_j(\varepsilon - \eta_j)}}$$

给定一组 data owners，一组 model buyers，和模型版本数量l，中间商需要合理划分l个模型版本使得收入最大化的同时兼顾以下几点：

1. 做到公平分配 compensation 激励 data owners 参与模型市场
2. 不同模型版本定价公平，实现无套利
3. 模型（数据产品）生产开销最小化，i.e.，充分利用给定 compensation，模型效用最大化
4. 考虑一个非盈利导向的 broker，以促进数据市场发展与社会福利为目的，在最大化收入的同时最大化模型效用，模型收入直接分配为 data owner 的 compensation