# 数据经济学

# 第三章：数据的供给

陈希路

暨南大学经济学院

2026 年春

# 章节目录

## 背景：新型生产要素

数据是数字化、网络化、智能化的基础，深刻改变生产、生活和社会治理方式。

**与传统四大要素的本质区别：**

1. **产生过程**：
   - 通过数字化和信息化技术生成
   - 来源包括互联网、物联网、传感器等
   - 特点：规模化、高速化、多样化
2. **权属划分**：
   - 涉及生成者、持有者、使用者
   - 需明确产权、使用权、收益权
3. **参与生产**：
   - 方式：融合、分析、应用

# 本章结构

我们将从三个方面探讨数据的供给：

- **第一节：数据的来源**
  - 数据如何产生？
- **第二节：数据的权属划分**
  - 产生之后的权益归属
- **第三节：数据如何参与生产**
  - 成为新的生产要素及其作用

# 第一节

# 数据的来源

# 数据的来源

## 解构数据由何而来

我们将从三个维度进行探讨:

1. 数据的生产过程
2. 数据产生的方式
3. 数据的分类

**意义**: 充分认识数据作为生产要素的重要地位，以及其带来的新机遇和新挑战

# 案例解析：滴滴出行

## 从原始轨迹到智能调度，数据的汇聚与质变

滴滴出行每天处理海量订单与轨迹数据，展现了典型的数据生产过程：

- **数据采集**：车辆 GPS、车内传感器、用户手机定位持续产生原始位置信息
- **数据清洗与存储**：自动剔除信号漂移与错误定位，依靠云计算基础设施进行海量存储
- **处理与解读**：结合路况动态运算，将原始数据转化为供需预测与最优路线规划知识
- **经济学视角的质变**：单个行车轨迹仅仅是孤立信息，大规模聚合后则打通了城市交通的信息大动脉，正式转化为核心生产要素

# 数据的生产过程

## 视角一：信息科学（循环迭代的过程）

从技术角度看，数据的产生包含以下步骤：

1. **数据采集**：
   - 来源：传感器、应用程序、网站等
   - 方式：手动输入、自动提取、传感器收集
2. **数据清洗**：
   - 去除噪声、错误、重复和不一致性
3. **数据存储**：
   - 介质：数据库、数据仓库、云存储
4. **处理与解读**：
   - 转换、集成、分析、挖掘 $\rightarrow$ 凝练出知识和信息

# 数据的生产过程

## 视角二：经济学（量变引起质变）

数据参与经济活动并成为第五大生产要素，是通过技术的积累实现的。

- **技术基础**：
  - 基础软件升级
  - 硬、软、云、网等设施出现
- **质变过程**：
  - 样本规模随着储存能力提升而增大
  - 作用增强：从分散 $\rightarrow$ 汇聚 $\rightarrow$ 聚集 $\rightarrow$ 聚合
- **“新基建”**：
  - 云计算、大数据、物联网、AI 打通了经济社会发展的信息“大动脉”

# 案例解析：三一重工与字节跳动

## 1. 生产过程中的数据生成：三一重工树根互联“根云平台”

- **应用场景**：挖掘机全面接入工业互联网
- **创新要素体现**：机器设备持续回传运行工况数据，企业据此优化零部件设计，进而向客户提供预测性设备维护服务

## 2. 消费过程中的数据生成：字节跳动内容推荐系统

- **应用场景**：短视频平台基于海量用户行为推送内容
- **数据生成逻辑**：用户的停留时长、点赞、划过等隐性行为生成消费记录，持续喂养与迭代推荐模型
- **潜在挑战**：极度精准的个性化分发容易引发信息茧房效应

# 数据产生的方式

数据的应用和产生是密不可分的。

1. 生产过程中的生成
   - **作为投入要素**：传感器监测设备运行、产品质量
   - **作为创新要素**：监测产品性能 $\rightarrow$ 改进优化 $\rightarrow$ 提高竞争力；数据驱动的创新
2. 消费过程中的生成
   - **来源**：
     - 互动记录（搜索、购买、评论）
     - 物理世界转移（气象、交通）
   - **应用**：个性化推荐、定价策略
   - **挑战**：信息茧房、算法牢笼

# 数据产生的方式

## 和传统要素的区别：自我迭代效应

传统要素（如土地）总量一定。然而，数据拥有自我迭代效应，通过反馈循环逐步提升质量、丰富性和价值。

**体现形式**：

- **生成阶段**：设备升级支持更高精度；反馈循环减少噪声。
- **众包模式与人机协同**：
  - *DataTang*：工人在完成任务（如图像识别）的同时产生新数据
  - 同时，模型也在学习人类偏好（如基于人类反馈的强化学习机制）
- **AI 模型迭代**：前沿推理大模型利用强化学习与自我对弈生成海量数据
- **开源共享与基础研究**：大学、行业研究机构和开源项目通过共享和利用数据，直接推动了人工智能、自动化等前沿领域的创新与发展

# 数据的分类：基于产生主体

**1. 个人数据（微观层面）**

- **定义**：属于个人、与个人生产生活相关的数据（身份、健康等）
- **保护原则**：
  - 需遵循法律法规（如欧盟 GDPR、美国 CCPA）
  - 原则：合法合规、透明公正、目的明确、数据最小化、安全保护、权利保障
- **特殊界定**：
  - 在公共利益需要时（如疫情期间），部分个人数据（行程）可能被公开并成为公共数据

# 数据的分类：基于产生主体

## 2. 企业数据（中观层面）

- **定义**：与企业经营状况、运营收支相关的数据（财务、销售、客户信息等）
- **来源**：
  - 内部：业务系统
  - 外部：市场调研、行业数据
- **面临的问题**：
  - **质量问题**：准确性、完整性不足
  - **安全挑战**：泄露、滥用风险
  - **流通障碍**：数据孤岛、集成难题
  - **合规压力**：法律法规要求

# 案例解析：招商银行

## 痛点：内部数据孤岛

传统大型金融机构往往面临核心业务部门之间数据互不相通的治理难题。

- **面临的挑战**：零售、对公、信用卡等部门各自掌握独立数据源，导致客户画像极度破碎，形成明显的内部数据孤岛
- **解决方案**：构建全行统一的金融大数据平台，将各条业务线数据进行集中存储与标准化清洗
- **最终成效**：实现企业内部数据的自由调用与低成本集成，进而能够为客户提供更精准的财富管理与信贷定价服务

# 数据的分类：基于产生主体

**3. 政府/公共数据（宏观层面）**

- **定义**：与国家和政府相关，从宏观层面上生成、统计和获取的数据
- **来源**：国家和地区政府的统计机构
- **用途**：
  - 支持政策制定、社会管理、公共服务
  - 作为公共资源对外开放
- **特有挑战**：
  - **数据治理**：解读过程中，委托–代理人结构容易引发数据的扭曲和操纵问题

# 第二节

# 数据的权属

# 案例解析：新浪微博诉脉脉案（中国数据确权第一案）

## 案例背景

脉脉（职场社交 App）通过 API 接口获取了新浪微博用户的关系链数据，甚至违规抓取了未授权脉脉的微博用户数据，引发新浪微博起诉。

**法院的裁判逻辑（印证“模糊数据产权”与合同协调）：**

- **个人数据权益**：用户对个人信息享有控制权
- **企业数据权益**：微博经过大量投入积累了用户数据，享有竞争性财产权益，但不是绝对所有权
- **确权规则的诞生**：确立了三重授权原则
  - 第三方使用数据必须经过：**用户授权** $\rightarrow$ **平台授权** $\rightarrow$ **用户再次授权**

# 数据权属的定义

## 什么是数据权属?

数据权属是指数据的**合法所有权和控制权**。

- **政策指引**:
  - 中共中央、国务院《关于构建数据基础制度更好发挥数据要素作用的意见》(即**"数据二十条"**), 为数据确权与流通奠定了基础
- **重要意义**:
  - 保护权益: 保护产生、采集、处理过程中的权益
  - 确保合法使用: 促进共享、交易、授权
  - 隐私与安全: 避免滥用、泄露或未经授权的访问
- **确立方式**:
  - 商业场景: 合同和协议(如使用协议)
  - 法律法规: 数据保护法
  - 知识产权: 专利、商标、著作权

# 数据权属面临的挑战

随着数据复杂性和跨境流动性的增加，界定和保护面临挑战

## 复杂场景

- 多方参与的数据合作。
- 跨国数据流动。
- 人工智能生成数据。

**对策**：需要从法律、技术、政策等多方面进行综合考虑

# 数据的确权：核心争议

学界普遍承认数据具有**财产权属性**，但在归属问题上存在分歧

**用户数据的三种确权方案：**

1. **企业拥有数据产权**：
   - 理由：数据是企业在提供服务过程中产生的资产
2. **个人拥有数据产权**：
   - 理由：数据是用户使用服务时产生的个人信息，个人应决定是否分享及获利
3. **模糊数据产权**：
   - 理由：不完全属于一方，应在二者间达成平衡，通过合同协调

# 案例解析：AIGC 时代的数据确权新挑战

## 从违规抓取到训练数据侵权

大模型时代，数据确权的核心矛盾已经转移到了互联网公开数据与 AI 训练的边界界定上。

**典型争议与司法探讨：**

- **《纽约时报》诉 OpenAI 案**：
  - 未经授权使用受版权保护的新闻文章训练大模型，是否构成侵权？
  - AI 企业主张这属于合理使用，类似于人类的阅读与学习
- **AI 绘画平台侵权争议**：
  - 画师的原创作品被爬取，用于训练生成式 AI，生成了风格高度相似的画作
- **思考**：如何在**保护原创者数据权益（提供激励）**与**促进行业技术创新（降低 AI 研发成本）**之间寻找最优均衡？

# 确权的方法论：从博弈到合约

- **数据的本质与痛点**：
  - 数据是一种生成品而非天然存在的禀赋，涉及信息提供者和采集者
  - 核心定性：信息的提供往往是无意识的，且**信息提供方通常是数据要素负外部性的主要承受者**
- **协商机制**：
  - 理论上：合作博弈过程，与贡献、估值、谈判地位相关
  - 现实困难：高昂的协商成本和监管成本
- **解决方案**：
  - 引入分级授权机制，以控制负外部性
  - 合同标准化

# 场景复杂性与理论支撑

不同生成场景下，利益主体的诉求不同：

- **个人数据**：涉及隐私保护和人格权
- **企业数据**：涉及商业秘密
- **公共数据**：涉及公共信息安全

## 海伦·尼森鲍姆（Helen Nissenbaum）的理论

**场景一致性**：隐私的内涵在不同情境中有不同的表现，应在信息传播的具体情境中讨论使用的适当性。

# 案例解析：场景一致性与隐私保护

## 为什么我们要打击 App 越界收集个人信息？

- **符合场景一致性（合理使用）**：
  - 导航 App（如高德地图）要求获取**地理位置**
  - 社交 App（如微信）要求获取**麦克风和摄像头**权限以进行视频通话
- **违背场景一致性（侵犯权益）**：
  - 手电筒 App 要求读取**通讯录**和**地理位置**
  - 输入法 App 在非使用状态下**监听麦克风**以推送精准广告

**政策呼应**

工信部出台的《App 违法违规收集使用个人信息行为认定方法》中强调的原则，正是场景一致性理论在我国数字经济治理中的直接体现。

# 解决方案：数据产权初始合约

- **定义**：在数据生成之前，由提供者和采集者协商明确采集范围、产权划分的合约
- **作用**：
  - 是数据进入市场流通的前提
  - 实现产权划分和授权的统一
- **分配方式**：
  - 按比例分割整体产权
  - 细分具体权利（如使用权、转让权）进行分配

# 数据的授权：制度背景与核心机制

- **制度背景**：中央深改委提出建立数据产权制度，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运行机制

- **核心解决方案**：数据分级授权机制
  - 让用户和平台企业基于市场原则，达成不同级别的数据授权协议

# 案例解析：数据产权三权分置的商业实践

## 国家电网与银行的电费贷合作（可用不可见）

**如何不泄露原始数据，又能让数据产生价值？**

中小微企业缺乏财务数据，但拥有真实的用电数据。银行需要数据审批贷款。

**“三权分置”在实际中的对应：**

1. **数据资源持有权**：国家电网持有海量企业的原始用电量数据，但不直接出售原始明细，以保护企业隐私
2. **数据加工使用权**：国家电网引入隐私计算（如联邦学习、多方安全计算）技术，在原始数据不出域的前提下，联合训练风控模型进行加工
3. **数据产品经营权**：将加工后的企业经营健康度指数上架数据交易所，银行付费调用该结果进行信贷风控

# 数据分级授权机制的优势

- **降低交易成本**:
  - 平台无须处理复杂的权属纠纷，直接采用市场化协议
- **灵活设计**:
  - 将数据分为**基础数据、衍生数据、派生数据**等等级
  - 用户自主选择授权级别
- **市场效应**:
  - **平台**：获得的数据要素总量增加（因部分授权用户上升）
  - **合规**：遵循了最小必要原则
  - **福利**：提升用户福利和社会福利

# 实施路径：协议条款的标准化

- **借鉴经验**：
  - 类似电商交易合同、软件授权协议中的标准化条款
  - 降低协商成本和司法裁决的不确定性
- **分级思路**：
  - 依据：后续数据流动的广度和深度（负外部性大小）
  - 级别设置：从“拒绝授权”到“完全授权”

# 一种有效的协商方案

分级授权机制综合了“分散化协商”与“标准化协议”的优点。

**保留分散化特点**

- 允许信息提供方根据补偿条件决定是否授权
- 避免了单一协议对多元性激励的不足

**具备标准化效率**

- 将协议划分为多个级别
- 降低了一对一协商的高昂成本

**结论**

分级授权机制能够在保障各方权益的基础上，促进数据的合理生成和使用。

# 第三节

# 数据供给侧的成本函数

# 数据供给侧的核心问题

## 核心问题

数据如何参与生产？参与生产的绩效如何？

在探讨这些问题之前，我们必须首先明确：

- **数据的成本构成**：从产生过程、经济学、交易成本等不同视角
- **数据的成本函数**：基于已有逻辑构建的数学表达

# 案例解析：蔚来汽车基于云服务的数据管理方案

## 海量数据带来的存储挑战

智能电动汽车每天产生海量车端行驶数据，传统的自建机房方案将面临难以承受的存储成本。

**低成本的数据处理策略：**

- 借助公有云可扩展的基础设施，免去了高昂的本地服务器采购成本
- 将海量车联网明细数据存储在公有云对象存储上，实现存储与计算分离，从而显著降低数据存储成本

# 数据的成本构成：产生过程视角

从数据的产生过程来看，成本主要由以下环节构成：

- 核心环节成本
  - **获取成本**：取决于收集方法（传感器、购买、人工测量）
  - **处理成本**：提取有效信息的技术成本
  - **存储成本**：取决于数据量、安全级别、介质类型
  - **传输成本**：取决于距离、带宽、协议
- 影响因素与策略
  - **影响因素**：数据质量、准确性水平、基础设施可用性、分析技能
  - **节约策略**：
    - 数据压缩、重复数据删除、数据去标识化、数据虚拟化
    - 使用基于云的存储和处理服务（可扩展且价格合理）

# 案例解析：自动驾驶的数据闭环与成本控制

## 海量无效数据带来的处理成本陷阱

高阶智能驾驶汽车（如萝卜快跑、特斯拉）每天产生海量视频数据，若全部回传并人工标注，数据的获取、传输与处理成本将呈指数级爆炸。

**低成本数据处理策略：**

- **影子模式与触发式采集**：
  - 车端 AI 实时运行但不控制车辆，仅在 AI 预测与人类驾驶员实际操作发生分歧（如紧急接管）时，才触发异常数据回传
- **自动化标注与端到端模型**：
  - 减少对人工标注的依赖，利用云端大模型自动处理海量 Corner Case（长尾恶劣场景）
- 通过降低无价值数据的传输与筛选成本，实现数据的提纯与迭代

# 数据的成本构成：经济学视角

从经济学视角看，数据的成本构成由两方面决定：

1. **数据包含的信息价值**（数据质量）
2. **数据相关的基础设施技术水平**

## 数据的本质与功能

- **技术本质**：字符和字节的堆叠排列，构成多样的数据空间
- **经济功能**：
  - 第五大生产要素
  - 具有规范性：能将不同类型信息转化为同一形式资源
  - *例子*：将“微博评论的情绪”（文字）与“宏观经济指标”（数字）统一纳入

# 数据的成本构成：交易成本视角

## 交易成本的定义（Coase 的观点）

为达成交易，在全部成本中除去传统生产成本以外的**间接的时间和货币成本**。

**数据成本的具体分类（Williamson 的观点）：**

- **搜索成本**：寻找交易对象。
- **信息成本**：查询对方需求和资质。
- **议价成本**：签订合同的细节开销。
- **决策成本**与**监督成本**：预防违约行为。

**注意**：最低粒度的数据可能没有价值，需要被组合、聚合才能体现价值，这增加了数据交易成本

# 学术界对数据成本的四部分解构

1. **生产成本**:
   - 涉及采集、加工、存储、移交
   - 维持数据作为产品的基础性运转
2. **搜索查询成本**:
   - 寻找符合要求数据集付出的费用、时间、精力及风险
   - 平台建设可降低此成本
3. **议价成本**:
   - 从未达成共识到签订合同阶段的各项细节开销
4. **监督成本**:
   - 履行义务、应对数据泄露等违约行为的约束成本

*补充：还有接近于零的复制成本，以及追踪验证成本*

# 案例解析：上海数据交易所破解流通难题

## 传统数据交易的困境

过去企业购买数据，往往面临找不到卖家、不敢轻易相信数据质量等现实难题。

**交易所如何降低交易成本：**

- **降低搜索查询成本**：提供统一的数据产品交易大厅
- **降低监督成本**：数据交易双方应签署数据交易协议，确保双方正确履行义务

# 数据成本界定的争议与挑战

- **非竞争性的双刃剑**：
  - 分享和复制通常不影响价值
  - 但少数情况下（商业竞争）可能给所有者带来竞争劣势
  - 防止二次转售是一项挑战
- **信息不对称（委托代理问题）**：
  - 卖方希望隐瞒事实，可能通过对数据的操纵来诱导买方
  - 甚至可能出现与解读者合谋隐藏信息的道德风险行为
- **特殊的成本结构**：
  - **初始创作成本高**：需大量人工干预、翻译、融合
  - **再生产成本低**：摩尔定律下，整合存储成本降低，再生产边际成本接近于零
- **价格外部性**：公开价格可能泄露数据价值

# 企业案例：万得信息金融数据库的投入与产出

## 金融信息服务商的商业逻辑

金融行业对数据准确性和及时性要求极高，打造一套专业数据库的成本结构非常典型。

- **初始创作成本高**：整合全球海量金融市场数据，需要大量人工进行干预、翻译和清洗，前期创作成本非常高
- **再生产成本低**：一旦数据库搭建完成，向新增金融机构客户开放账号的物理复制成本趋近 0
- **非竞争性**：多位基金经理同时登录系统查看同一只股票的财报数据，彼此获取信息的价值完全不受影响

# 数据的成本函数：构建逻辑

- **研究现状**：
  - 集中于数据交易过程（数据量、使用权、所有权）
  - 本质在于对数据包含的信息内容进行定价
- **主流形式**：
  - 将**柯布–道格拉斯（Cobb–Douglas）生产函数**与数据要素结合
  - 侧重于数据促进内生增长的讨论

# 数据成本函数的数学表达

假定社会总数据量不变，成本函数可表达为：

$$Y = LD^{\eta} \tag{1}$$

数据的供给价格（边际成本）为：

$$P(D) = \frac{\partial Y}{\partial D} = \eta \cdot \frac{L}{D^{1-\eta}} \tag{2}$$

**变量含义**：

- $Y$：社会总产出
- $L$：劳动力
- $D$：技术进步与数据驱动的变量（数据要素）
- $\eta$：在 $(0,1)$ 区间的变量，反映数据转化为知识信息的效率

注：在完全竞争市场中，数据的边际成本与边际收益（即价格）相等。

# 数据成本函数的前沿学术探讨

在柯布–道格拉斯函数的设定之上，学界对数据如何参与定价有更深入的探讨：

- **Jones & Tonetti（2020）**：
  - 强调数据的非竞争性
  - 认为数据未被定价的原因，是数据参与内生增长往往是企业应用中**自动内生生成**的过程
- **Cong et al.（2022）**：
  - 构建了动态均衡模型，指出数据**既参与生产，又参与创新过程**
  - 边际生产率与数据要素的边际成本共同构成了数据的成本
- **成本函数的拓展：常替代弹性（CES）函数**
  - 作为柯布–道格拉斯函数的推广版本，**CES** 函数也被用于研究数据成本，以更灵活地衡量数据与其他传统要素间的替代关系

# 数据成本评估的现实困境

尽管构建了函数，但实际评估仍面临困难：

1. **质量评估难**：
   - 数据来源众多、格式不一致，质量是开放性问题
2. **传统方法失效**：
   - 现有的无形资产评估方法（重置成本法、收益现值法）很难准确量化数据价值
3. **缺乏统一标准**：
   - 使得提供者和购买者无法对价值达成共识
   - 现有指标（质量、效用、历史成交价）尚未被广泛认可

**展望**：未来随着 AI 技术发展，我们需要更准确地衡量数据资产价值

# 本章总结：数据的来源

数据生成是一个动态的过程，彻底区别于传统要素。

- **产生方式**：在生产过程与消费过程中不断生成，并具备自我迭代效应
- **分类体系**：基于产生主体，分为个人数据、企业数据和公共数据，不同主体面临着不同的隐私、安全与治理挑战

# 本章总结：数据的权属

确权是数据要素市场流通的基础与前提。

- **产权界定**：单一所有权归属存在争议，数据是多方协同创造的生成品
- **核心机制**：基于场景一致性理论，数据分级授权机制与协议条款标准化能够有效降低协商成本，平衡各方利益诉求

# 本章总结：数据供给侧的成本函数

数据具备非竞争性及极为特殊的成本结构。

- **成本组成**：包含生产成本、搜索查询成本、议价成本以及监督成本
- **成本特征**：数据要素呈现初始创作投入极高、再生产边际投入趋近于零的特征
- **定价函数**：学术界主流采用柯布–道格拉斯生产函数进行建模
  - 将数据要素与技术进步纳入统一考量
  - 通过反映数据转化为知识信息的效率来探究供给边际成本

# 本章总结：前瞻与展望

- 供给侧的完善，必须理清数据来源、清晰划定数据权属、合理核算数据成本
- 随着数字基础设施不断升级与确权定价机制日益成熟，数据作为新型生产要素的巨大潜能将被进一步释放，进而推动实体经济向智能化深度转型

# 谢谢大家！