数据脱敏**3** | 脱敏技术与法律效果评价可以机械对应吗？

合规科技系列文章 **Law-Tech Series**

高速发展的时代背景下，一方面行业分工在层层细化，一方面跨学科的交叉研究又越来越不可或缺。科技与法律表面上是两个相去甚远的专业领域，但就数据治理与隐私保护而言，只有跨界互通才可能找到最佳的解决方案。

“合规科技专题文章”旨在兼顾科技与法律的双重视角，深度解读数据技术的逻辑原理与数据合规的法律要求，从而促进技术人与法律人的双向理解，探讨数据利用与个人权益协调发展的可行方案。

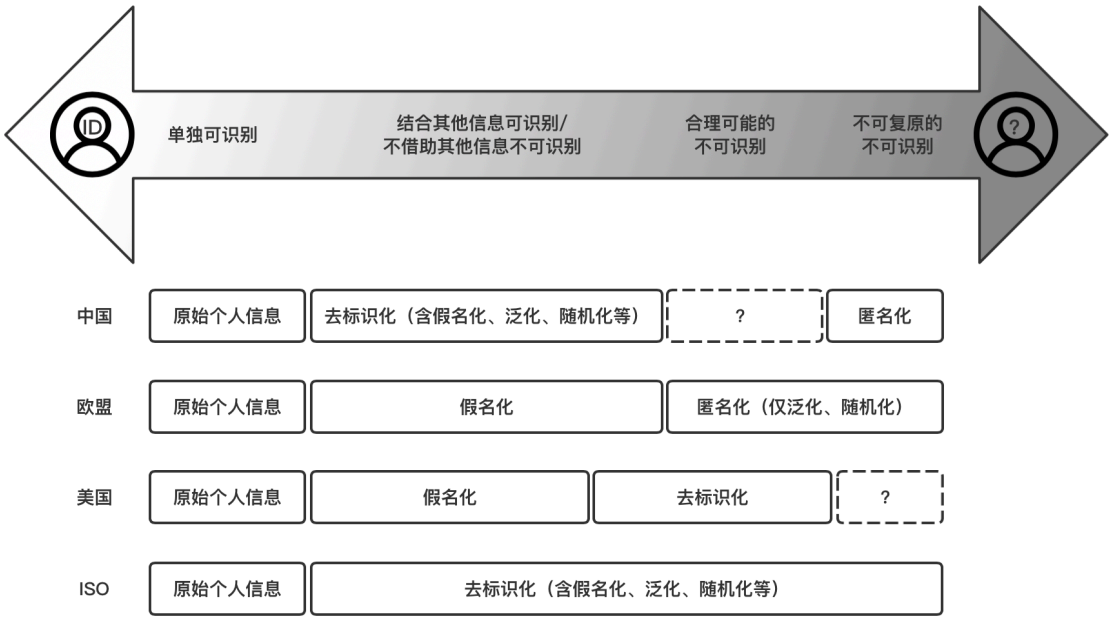
“大数据”已然从热词变成日常，而数据在释放无限潜力的同时，也引发了隐私泄露的巨大隐患。从若干年前科技公司野蛮生长，到近年来数据立法接踵而至，信息社会正在两极之间寻求平衡。数据脱敏提供了这样一种可能性——通过降低数据与主体之间的关联，可以同时保留较高的隐私保护程度和较大的数据利用价值。

“数据脱敏”专题文章将梳理匿名化、去标识化、假名化等一系列相关概念，分析中国、欧盟、美国等法域对不同概念的法律评价，介绍数据脱敏的技术方案与隐私模型，探讨各个业务场景下的行业实践案例与法律落地方案，以推动数据利用和隐私保护的平衡发展。

“数据脱敏”专题往期文章链接

- 数据脱敏1 | “数据脱敏”是一个法律概念或技术概念吗？
- 数据脱敏2 | 不同法域下匿名化、去标识化、假名化的含义一致吗？

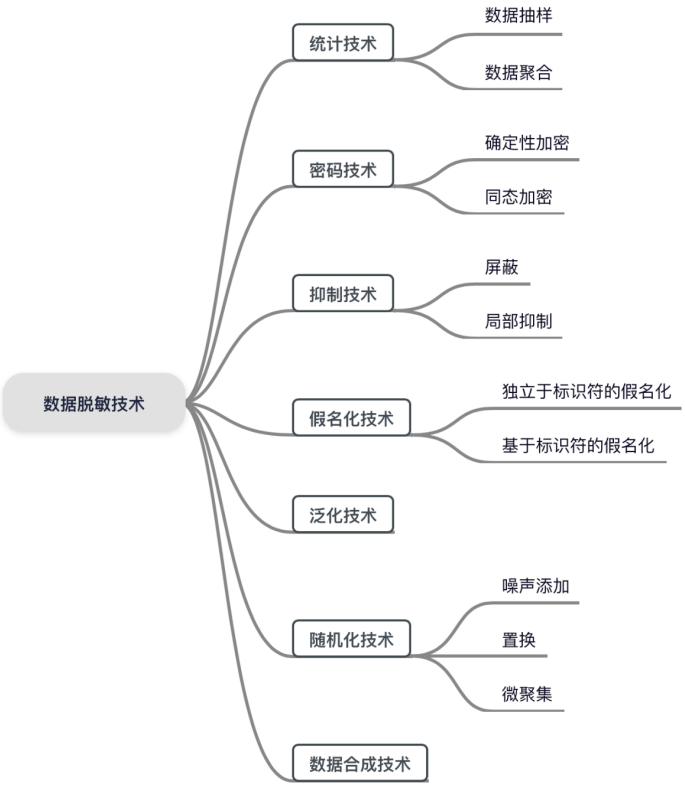
上期回顾：中国、欧盟、美国等法域都有匿名化（**anonymization**）、去标识化（**de-identification**）、假名化（**pseudonymization**）等概念，但各个法域对同一概念的定义存在差异，即对不可识别的程度要求不同。



注：括号内的概念为技术手段，括号外的概念为效果评价。

中国、欧盟对常用的脱敏技术制定了介绍性指南。实践中，一个常见的误区是将特定技术机械地对应特定的数据脱敏效果评价，例如，泛化技术、随机化技术就是匿名化。但实际上，各国立法并没有对一类技术进行概括性评价，而是对技术处理所实现的具体效果进行法律评价，因为同一技术在特定的实施强度和应用场景下，可以实现不同程度的脱敏效果。

本文将介绍统计、密码、抑制、假名化、泛化、随机化、数据合成等数据脱敏技术的基本原理，并举例说明同一技术的效果跨度。每种技术的特点和阈值各不相同，实践中基于特定的场景和目标，可以选择适合的技术及实施强度，从而平衡数据的可用性和安全性。



一. 数据脱敏的技术与原理

(一)统计技术

1. 数据抽样（**sampling**）：从原始数据集抽取若干数量、若干属性的样本，从而使攻击者无法确定某个目标个体是否在抽样数据集之中。

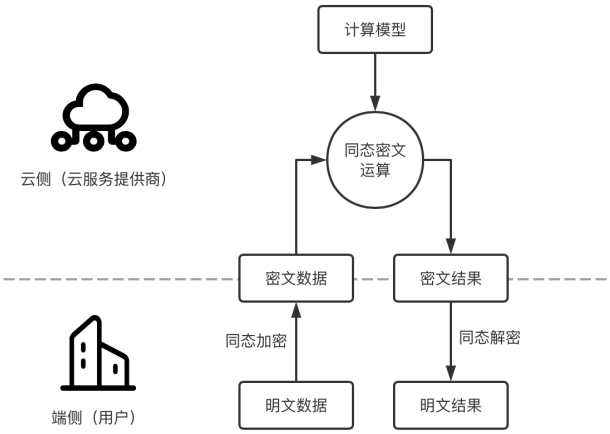
2. 数据聚合（**aggregation**）：对原始数据集的属性值进行统计，仅输出统计特性（例如求和、计数、平均值、最大值、最小值、方差、标准差等），从而降低披露个体信息的风险。

(二)密码技术

1. 确定性加密（**deterministic encryption**）（属于非随机加密）：通过密钥对数据进行加密。

2. 同态加密（**homomorphic encryption**）（属于随机加密）：允许人们对加密数据进行运算，运算结果解密后等同于对不加密的数据进行相同运算的结果。

以云计算场景为例。在传统模式下，用户需要信任云服务提供商不会窃取甚至泄露用户数据。而同态加密可从根本上解决数据处理过程的保密与安全，同时保护用户的数据和云服务提供商的计算模型。



(三)抑制技术

1. 屏蔽（**masking**）：删除直接标识符或其中一部分，例如将手机号或身份证号的其中四位变成****。

2. 局部抑制（**local suppression**）：删除特定的属性值，以免它和其他属性相结合而识别个体。这种技术通常适用于比较稀有的属性值，例如罕见的Rh阴性血型。

(四)假名化技术

假名化（**pseudonymization**）是指用生成的假名代替标识符的原始值。

1. 独立于标识符的假名化：用假名代替标识符的原始值，并创建假名与原始值的分配表。此处的假名是指与标识符无关的随机值，还可以采取“多对一”（多个原始值对应一个假名）的方式，但这会降低数据的可用性。

2. 基于标识符的假名化：通过加密、散列/哈希等密码技术，在标识符原始值的基础上生成假名。加密技术通常是一一对应的，可以通过密钥和算法进行解密，还原标识符的原始值。散列函数是一种单向运算，保密性更好，而难以逆向还原原始值。

(五)泛化技术

泛化（**generalization**）是指降低属性值的粒度，对属性进行更抽象、更概括的描述。例如，将姓名泛化成姓，将市泛化成省，对数字进行取整、设置区间、最大值或最小值。例如，将年收入的确切数值泛化成10万及其以下、10万到100万、100万及其以上三个区间，从而使更多的个体共享同一属性值、降低重标识的概率。

(六)随机化技术

随机化（**randomization**）是指随机修改属性值，这将破坏数据集的真实性。

1. 噪声添加：添加随机值/噪声到某一属性中，同时尽可能维持该属性的原始统计特征。

2. 置换：对数据集中某一属性的值进行重新排序，即，将某一个体的属性值置换给另一个体。

3. 微聚集：对某一属性进行排序和分组，接近的属性值分为一组，并用每组的平均值来代替该组的所有原始值。

数据主体	身高（原始值）	噪声添加	置换	微聚集
张三	1.63	1.66（+0.03）	1.71	1.67
李四	1.71	1.69（–0.02）	1.75	1.67
王五	1.75	1.73（–0.02）	1.81	1.78
赵六	1.81	1.82（+0.01）	1.63	1.78

(七)数据合成技术

数据合成技术（**synthetic data**）是指通过人工方式生成数据集，该合成数据集与原始数据集的特性相符。

二. 法律对脱敏技术的评价

我国的《个人信息去标识化指南》和国际标准化组织的《隐私增强数据去标识化术语和技术分类》（ISO/IEC 20889）将上述七种技术作为去标识化技术进行了列举说明，而欧盟的《关于匿名化技术的意见》将其中的泛化技术、随机化技术作为主要的匿名化技术。但是，这并不意味着使用特定技术必将实现特定效果。泛化技术、随机化技术既可能实现去标识化、也可能实现匿名化，主要取决于特定的技术方案实现了哪种程度的不可识别。

关于脱敏技术的法律评价，应当注意以下几点：

1. 不同技术的脱敏能力存在差异

不同的脱敏技术有其特点和阈值。例如，统计、泛化、随机化等技术有可能实现“不可复原的不可识别”，但假名化技术最多实现“结合其他信息可识别/不借助其他信息不可识别”的效果。

欧盟《关于匿名化技术的意见》中特别强调，假名化无法作为匿名化的方法之一，因为假名化虽然降低了数据集和数据主体身份之间的联系，但数据主体仍有可能被间接识别。例如，用户在社交网络上使用的昵称即是一种假名，但结合用户发布的其他信息，仍有可能识别该用户的身份。

2. 同一技术的实施强度存在差异

就同一脱敏技术而言，其具体的实施强度也存在差异。例如，身份证号中屏蔽的数字如果是生日而不是最后四位，则安全效果较差，因为个人经常在好友庆生、入职信息表、注册会员等场景下暴露自己的生日，容易还原原始的身份证号。

泛化技术也可以设置不同的颗粒度，例如对地址数据的泛化，从精确的门牌号到小区、街道、区县、地市、省、国家，不可识别的程度不断加深，但数据的价值也随之折损。因此，欧盟的《关于匿名化技术的意见》一方面认可泛化技术可以实现匿名化，一方面强调它并不是在一切情形下都能有效实现匿名化。

3. 具体场景也会影响技术处理的效果

对脱敏技术的法律评价不是抽象的，而是基于具体的应用场景，因为特定的情形会影响技术处理的效果。例如，统计技术往往可以隐匿个人，但一旦结合背景知识仍可能暴露个人。假设在某个社区中，患有高血压的人数为40人，从40这个统计值中，一般无法识别出患者的身份；但是，当搬来一个新住户后，如果患病人数变为41人，则可以判断出该新住户患有高血压。因此，统计数据在特定场景下并不必然是匿名的。

本期小结与下期预告：数据脱敏可以采用统计、密码、抑制、假名化、泛化、随机化、数据合成等技术及其组合。不同技术的脱敏能力存在差异，同一技术的实施强度存在差异，具体场景也会影响技术处理的效果，因此，法律对脱敏技术的评价并不是一刀切的，而是具体考量技术所实现的效果。那么，下一个需要回答的问题是，法律上如何衡量脱敏的效果？下期文章将为您介绍定性、定量这两类衡量标准。