CAICT 中国信通院

北京国际大数据交易所
Beijing International Data Exchange

集智 蓝皮报告

数据清洗、去标识化、匿名化业务规程（试行）

中国信息通信研究院产业与规划研究所

北京国际大数据交易所

2023年11月

版权声明

前　言

为规范数据处理行为，指导组织正确开展数据清洗、去标识化、匿名化处理等业务活动及相应的技术测试评估，支撑数据共享、交易、开放等流通活动合规、有序进行，激活数据要素市场，依据《个人信息保护法》《数据安全法》，结合《北京市数字经济促进条例》《北京市数字经济全产业链开放发展行动方案》等法规政策要求，在北京市经济和信息化局指导下，中国信息通信研究院产业与规划研究所、北京国际大数据交易所联合编制本报告。

组织依据法律法规要求及相关业务场景需要，对其控制的数据资源进行清洗、去标识化、匿名化处理，是为满足数据处理目的对原始数据逐步深入加工改造的过程，是提升数据可用性和安全性的关键数据处理活动。

本报告以业务操作规程形式为组织提供数据清洗、去标识化、匿名化处理的流程和方法指引，可以作为组织提升自身数据质量和可用性的指引方法，作为数据交易中介机构审核交易数据合规性、安全性和可流通性的参考规则，以及作为相关认证、检测机构结合应用场景针对相关技术进行安全测试评估的评价工具，支持、鼓励数据加工、咨询、安全、检测、认证等第三方数据服务机构发展。

本报告所描述的技术方法适用于广义的数据范畴，包括但不限于个人数据、企业数据、物联网数据等，但特殊数据类型需要遵守相应的特别管理要求。本报告所描述的数据清洗、去标识化、匿名化处理，是基于数据资源的加工处理过程。有“数”才能对“数”

进行处理，通过采集、标识、编码形成数据资源，是对数据进行清洗、去标识化、匿名化处理的前提。前者是形成数据的基础，后者是维护数据质量和安全的关键。

本报告主要描述各数据处理活动的基本原理和通用技术方法，需要结合实际场景具体适用。本报告所引用的部分技术方法参考了《GB/T 37964-2019 信息安全技术 个人信息去标识化指南》等相关标准指南，在此基础上，结合《个人信息保护法》等法律法规的界定，根据技术特性和处理效果，对去标识化技术和匿名化技术进行了区分。本报告所描述的相关技术方法仍在不断丰富、演进和迭代，相关应用场景也在不断发展变化，本报告将持续跟踪观察，适时更新、补充、调整和校正。欢迎各组织积极反馈技术适用情况和建议，提供技术适用场景和实践案例。

目 录

表 目 录

习近平总书记在2023年中国国际服务贸易交易会全球服务贸易峰会上发表视频致辞指出，要“推动数据基础制度先行先试改革”。《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》要求“创新技术手段，推动个人信息匿名化处理”。**规范数据清洗、去标识化、匿名化处理，有助于提升数据的可用、可信、可流通、可追溯水平，推动数据要素强化优质供给，是建立合规高效、场内外结合的数据要素流通和交易制度的重要内容。**具体来说，为满足数据可用性和安全性进行的数据清洗、去标识化、匿名化处理，是数据产品进场上市的条件，也是数据资产登记、交易的前提，更是数据应用、建模释放二次衍生价值的底线。本报告通过明晰数据清洗、去标识化、匿名化处理三者之间的关系，总结各项处理活动的处理目的、流程、技术方法及环境要求，以期为相关组织开展相应数据处理活动和测试评估提供参考。

一、处理目标及相互关系

（一）数据清洗是数据可用的保障

数据清洗是运用一定方法修正识别到的数据问题，实现数据的规范性、完整性、一致性、准确性和可溯源性，提高数据质量的过程。数据清洗旨在满足数据的可用性要求，是数据资源预处理的第一步，也是保证后续处理结果准确、科学、有效的重要一环。**数据清洗作为数据后续开发利用的基础，是数据去标识化和匿名化处理的前置步骤。**

（二）去标识化是数据脱敏的关键

数据去标识化是指数据经过处理，使其在不借助额外信息的情况

下无法识别特定自然人或相关标识符的过程。**数据去标识化处理强调标识符的“不可识别性”，即对数据内含的相关敏感信息内容进行脱敏处理，通过去除、替换、模糊等方法，达到不借助额外信息的情况下无法识别特定自然人或相关标识符的效果。**

数据去标识化与在先的标识形成过程分属数据处理的不同阶段及场景。标识形成是产生数据的过程，使得被标识对象据此可以被组织进行有效管理和开发利用。数据去标识化是标识数据产生后的加工处理过程，旨在提升标识信息的安全防护水平，确保敏感的标识内容不被未经授权的主体获取和利用。去标识化处理是强化标识数据安全性的重要保障。例如，制造业企业通过对产品、零部件、设备进行标识，形成了可精准定位产品和设备的数据资源，在委托外部第三方技术开发商进行相关应用系统开发时，需要对含有敏感内容或涉及商业秘密的数据进行去标识化处理。

数据去标识化处理暗含了相关标识符具有“复原”的可能，去标识化无法单独实现匿名化的法律效力。例如，对个人信息进行去标识化处理后的数据，仍属于个人信息范畴。

（三）匿名化是去标识化的强化

数据匿名化是指数据经过处理，无法识别特定自然人或相关标识符且不能复原的过程。**数据匿名化处理在强调标识符的“不可识别性”基础上，要求标识符同时满足“难以复原性”标准，是数据去标识化的进一步处理，即数据去标识化后应用相关技术使相关标识符难以复原的过程。**经匿名化处理后数据的初始效用将受到较大程度的改变。

与数据去标识化相比，经匿名化处理后的数据即便借助了额外信息也难以识别特定自然人和被处理的标识符。例如，对个人信息进行匿名化处理后的数据，不再属于个人信息范畴。但匿名化处理仅是描述应用匿名化技术的过程，并非描述数据达到绝对匿名化的状态，完满、绝对的不可复原状态无法100%确定。

表1 数据清洗、去标识化、匿名化处理的技术特点和差异

加工后数据	**改造程度** （相对原始数据）	**数据有用性** （针对个体记录）	**数据安全性** （脱敏程度）
清洗后数据	低	高	低 （单独可识别）
去标识化数据	中	中	中 （不借助额外信息不可识别）
匿名化数据	高	低	高 （借助额外信息也难以复原的不可识别）

来源：中国信息通信研究院

去标识化技术和匿名化技术没有严格界分，二者核心都是通过技术手段对标识信息进行脱敏处理，实现对敏感数据内容的保护，实践中两类技术通常可以组合使用实现预期处理效果。**本报告根据抗重新识别的风险能力大小和对敏感内容安全防护程度的差异，将相关技术划分为去标识化技术和匿名化技术。仍保留原始数据个体颗粒度的，纳入去标识化技术方法范畴；不再保留原始数据个体颗粒度，或原始数据记录的真实性已受到显著减损，或原始数据记录不对外披露的，纳入匿名化技术方法范畴。**

二、数据处理原则

（一）合法合规

组织开展数据清洗、去标识化和匿名化处理，应满足我国法律、法规、规章和标准规范对数据安全和个人信息保护的有关规定，不得不当损害国家、社会和第三方组织及个人的合法正当权益。

（二）安全优先

组织应采取相应的管理和技术措施，保证数据加工处理过程的安全性。数据的安全性考虑是组织开展数据去标识化、匿名化处理活动的首要目的，以降低数据在后续流通、应用环节的安全风险，降低数据安全事故发生概率。

（三）平衡效用

组织应根据业务目标和安全保护要求，面向场景化应用需求，选择恰当的清洗、去标识化和匿名化处理路径和技术，在确保安全的前提下，强调数据质量要求，尽可能满足预期效用，促进数据安全性和可用性的有效平衡。

（四）技管结合

组织应综合利用技术和管理两方面措施实现数据处理的最佳效果，根据工作目标和数据安全要求制定适当的策略，选择合适的模型和技术，建立完善的管理架构、操作权限和责任机制，将技术和管理措施嵌入数据清洗、去标识化、匿名化处理全流程，并定期跟踪评估和持续改进。

（五）有效溯源

组织应明确各环节的数据处理权限和流程，对数据清洗、去标识化、匿名化设置访问控制程序，采取措施清晰记录数据处理过程的细节、使用的参数和控制措施，及时发现已经出现或可能出现的偏差或不当操作，支撑后续对数据处理过程进行维护、审计和追溯。

三、数据清洗规程

（一）处理目的

组织实施数据清洗活动，应保证清洗加工过程和输出结果符合以下要求:

1.规范性

数据来源合法，数据的格式、质量及存储标准应统一，应使用相同度量单位描述同一场景下的同类数据，满足数据互联互通要求，不存在空值、无效值，响应依据规范标准的各种查询和各种计算。

2.准确性

应对数据所指向的内容客观、真实、准确描述，可对清洗前后的数据进行内外部比对校验，并对具有时效要求的数据根据时间特性及时更新，确保清洗加工不造成数据失真、错漏。

3.完整性

清洗后的数据应保证数据的连续性、完整性，源数据应在源头或备份表中能找到，数据在字段、记录内容或数据集内不应有重复值。

4.一致性

各字段内的数据应与字段描述一致，同一个数据在同一时刻在不同数据库、应用和系统中应保持一致。

5.可溯源性

应在数据清洗转换前对原始数据进行备份，对清洗过程所使用的方法、参数和路径进行记录，保证原始数据可溯源，便于后续查证或重新使用。

（二）处理流程

数据清洗的流程通常包括抽取清洗对象、明确清洗规则、标识错误数据、数据修正处理、数据转换检验、评估清洗结果六个步骤。

1.抽取清洗对象

（1）明确清洗对象

选取需要进行清洗处理的数据，明确清洗的数据范围、类型、性质、体量、内容、关系、质量等信息，全面分析清洗标的的情况，对清洗数据进行分类分级。

（2）对清洗对象进行抽取

清洗对象的抽取应当允许对结构、半结构和非结构等不同类型数据进行抽取，包括对数据的全量抽取和增量抽取，数据抽取后的表结构应与抽取来源的表结构保持一致。

2.定义清洗规则

（1）确定清洗效果和目标

根据清洗的必要性，分析对应数据资源的特点和清洗复杂程度，

结合业务要求或用户和其他相关方的需求，明确清洗的程度和需要达到的质量效果。

（2）确定清洗逻辑规则

结合所抽取的清洗对象的数据特点，以需求为导向，以应用为目标，以数据的可用性为评价标准，明确各数据错误类型的判断标准及相应的修正处理方式。

3.标识错误数据

（1）筛选错误数据

分析筛选出数据资源中存在的数据问题和对应的数据。按照常见错误数据的类型，对数据问题进行分类，针对性进行错误标识，并支持对已标识的错误数据进行查询定位。可采用统计学、关联规则、业务区分等方法来对目标数据进行错误检测，识别出数据的错误类型并进行标识。例如，通过使用统计学方法（例如均值、标准差、范围或分位数）对数据进行分析和可视化，发现异常值或离群值，从而标识错误数据。

（2）常见错误类型

残缺数据：数据中缺失一些记录，或一条记录中缺失一些值，或两者都缺失。

偏差数据：数据没有严格按照要求记录，包括格式内容错误、逻辑错误、不合规数据等。

重复数据：数据中出现多条相同记录，或多条记录反映同一内容，

通常发生在数据来自不同来源、数据多次采集、瑕疵数据更正备份等情形。

其他错误：数据未能准确反映所描述的对象的其他情形，如非结构化或半结构化数据、无意义数据、不相关数据等。

4.数据修正处理

对已标识的残缺数据、偏差数据、重复数据和其他错误数据分别采用针对性的方法和工具进行处理。常见的数据清洗工具包括软件工具、脚本等类型。选择清洗方法和策略时，应根据清洗目标和业务需要，结合数据错误类型，采取删除、填充、更换等不同的方式处理，具体可参考本节“（三）常见技术方法”。

5.数据转换检验

（1）错误数据转换

对错误数据的格式、信息代码、值的冲突进行转换。数据转换前应检查需要转换的数据规则和字段是否一致。

（2）转换结果检验

一是内容检验，即对转换后数据内容的完整性、全面性进行检验，包括非空检验和数据量检验。

二是格式检验，即对照数据格式样例或相关标准对转换后数据格式的规范性、一致性进行检验。

三是逻辑检验，即结合相关联数据对转换后数据逻辑是否符合预先设定的范围、区间、大小、数值关系等规则的约束性要求进行检验。

四是合规检验，即结合业务场景的合规要求对转换后数据内容是否符合法律法规和强制性标准的要求进行检验。

四是合规检验，即结合业务场景的合规要求对转换后数据内容是否符合法律法规和强制性标准的要求进行检验。

6.评估清洗结果

数据清洗后及时评价输出结果是否符合事先设定清洗规则和规范性、准确性、完整性、一致性、可溯源性等目标要求，并从业务角度评估清洗后数据的有用性，判断是否可以支撑后续加工处理活动。

（三）常见技术方法

1.残缺数据处理

组织应当按照所需处理数据的字段缺失比例和重要性，采取差异化的策略进行处理。重要性高，缺失率低的字段，可以通过计算结果填充并进行核验；重要性高，缺失率高的字段，重新采集获取或通过其他渠道取数补全；重要性低，缺失率低的字段，不做处理或简单填充；重要性低，缺失率高的字段，可以选择删除该字段。

（1）删除缺失值

当样本数量充足，且出现缺失值的样本占比相对较小时，可以备份当前数据后，直接删除后期加工处理不需要的字段和缺失值。

（2）填充缺失内容

存在缺失率较低但相对重要的数据项时，可以通过计算填充并进行核验的方式进行补全，包括不同指标的计算结果填充和同一指标的

计算结果填充。

不同指标的计算结果填充: 即通过数据项与数据项之间的逻辑联系，采取相应的计算方法得到缺失内容。包括热卡填补法、最近距离决定填补法、回归填补法、多重填补方法、K-最近邻法、有序最近邻法等。例如，数据中年龄字段缺失，可以从公民身份证号中提取年龄字段。

同一指标的计算结果填充: 即通过对同一指标列的数据采取均值、中位数、众数等方式进行计算，将相应结果进行填充，多用于数值型数据。例如，某一记录的身高数据缺失，可以使用该字段的均值进行填充。

（3）重新采集数据补全

存在缺失率较高且相对重要的数据项时，可以通过线下补充收集、业务知识或经验推测、新增抽取其他数据源数据等方式，进行关联对比后填补。

2.偏差（异常）数据处理

组织应当对未符合规范要求，存在格式、逻辑及内容不匹配等方面偏差的数据进行处理。

（1）格式不规范数据

对存在格式不规范等问题的数据进行处理，包括全、半角处理和无效字符处理。按照事先定义的规则进行全、半角符号统一，以半自动校验结合半人工方式发现错误字符，进行自动化修正或人工修正。

（2）逻辑冲突数据

对存在不符合逻辑约束要求、相互间存在冲突的数据进行处理，可通过直接推理、关联修正和逻辑重构等方式进行，并再次进行校验。

直接推理：了解数据潜在的逻辑规则，采取逻辑推理法，直接处理简单逻辑错误的数据。

关联修正：借助分箱、聚类、回归等方法识别逻辑错误数据，通过相互验证的方法修正矛盾内容。

逻辑重构：对于重要性较高的不合理数据进行人工干预，或重新采集数据，引入更多数据源进行逻辑的重新梳理并再次进行校验。

（3）内容不匹配数据

对存在噪声数据、超出明确取值范围，以及数据中存在敏感信息或内容不符合要求等数据进行处理。通过设定判定规则，借助自动化手段判断数据是否在规则范围内，不在规则范围内的，进行警告及人工处理。

噪声数据：对噪声值进行平滑处理，或在不影响数据结构和后续使用情况下，将噪声数据进行删除处理。

离群值数据：判断超出明确取值范围数据的来源是否可靠，数据的存在是否合理，合理的数据予以保留，不合理数据予以调整。

内容不对应数据：识别内容与字段要求不匹配的问题类型，如人工填写错误、导入数据时没有对齐、数据源端业务系统缺陷等，通过关联、修正或重新采集等方式匹配相应字段进行填补。

3.重复数据处理

将具有相同含义的数据判定为重复数据，包括相同数据和相似数据。

相同数据：形式、含义和内容均相同的数据，根据来源权威性和应用场合，选择最恰当渠道来源的数据，或在不影响数据保真度和完整性的情况下进行合并处理。

相似数据：识别相似数据的各自含义，判断数据的实质含义上是否存在差异，实质含义相同的数据按照相同数据进行处理，实质含义有差异的数据，不能界定为重复数据，应分别保留。

4.其他错误数据处理

针对数据未能准确反映所描述的对象的其他情形，可以采取以下通用方式进行处理：

将非结构化和半结构化数据转化为结构化数据；将无意义数据、不相关数据在进行必要性和相关性评估后进行删除，提升后续数据处理效率；对仍存在问题未处理的错误数据存入问题数据库，便于后续查证或重新使用

四、数据去标识化规程

（一）处理目的

组织实施数据去标识化，应当确保经过处理的数据达到以下效果：

1.标识不可识别

对数据中的直接标识符和准标识符进行处理，避免未经授权的主体无需借助其他额外信息，直接根据这些标识内容便可以识别出原始

信息主体或相关标识符。

2.控制被识别风险

将去标识化后的数据可能被未经授权的主体再次识别的风险控制在可接受的范围内，确保标识符暴露的风险不会因数据接收方之间的潜在串通或新数据的增加而增加。

3.兼顾数据效用目标

有效平衡数据的安全性和可用性，选择合适的去标识化模型和技术，确保去标识化后的数据尽量满足数据开发利用的预期目的和效用，在数据安全前提下最大发挥去标识化数据应用价值。

（二）处理流程

数据去标识化的流程通常包括确定去标识化对象、制定去标识化目标和计划、识别相关标识符、对标识符进行处理、验证审核处理结果、评估重新标识风险六个步骤。

1.确定去标识化对象

组织对于自身合法取得、合法持有，并实际控制的数据，应当基于外部和内部的多方面因素的考量确定需要进行去标识处理的数据范围。

（1）法规标准要求

根据国家、地区或行业的相关政策、法律、法规等的强制性规定，判断待收集、存储、使用、加工或向第三方提供的数据是否涉及去标识化的相关要求。例如，《个人信息保护法》第 51 条要求，个人信息

处理者应当采取加密、去标识化等安全技术措施，防止未经授权的访问以及个人信息泄露、篡改、丢失。

（2）组织策略要求

根据自身数据管理要求，或者按照与相关合作方约定，判断数据进行内外部应用时是否需要进行去标识化处理。例如，将个人信息对外展示时，参考《GB/T 35273—2020 信息安全技术 个人信息安全规范》，涉及通过界面展示个人信息的（如显示屏幕、纸面），个人信息控制者宜对需展示的个人信息采取去标识化处理等措施，降低个人信息在展示环节的泄露风险。

（3）数据来源方要求

根据数据采集时是否存在对数据来源方等作出了去标识化的相关承诺或约定，判断对数据进行加工或向第三方提供时是否需要进行去标识化处理。例如，组织已在产品隐私政策中声明，将用户个人信息用于对外提供学术研究或描述的结果时，承诺对结果中所包含的个人信息进行去标识化处理。

2.制定去标识化目标

均衡数据安全性和可用性两方面需求，确定数据去标识化处理需要达到的效果。

（1）明确标识被识别风险的控制要求

分析数据的来源、性质、类型，梳理待处理数据是否涉及法律法规要求和相关承诺，结合去标识化后数据的主要用途和使用范围，考虑可能采用的去标识化模型和技术的应用方向及能力，综合评价组织

对相关标识符和准标识符被重新识别的风险的不可接受程度。

（2）明确满足数据可用性的最低要求

结合数据去标识化后的用途，评估相关技术方法的应用对初始数据的改造程度，分析数据去标识化后对业务活动的可能影响，提出数据有用性的最低要求。

3.识别相关标识符

根据去标识化的目标，针对需要去标识化的数据，识别出需要进行处理的直接标识符和准标识符。组织可以通过以下方法识别：

（1）查表识别

组织通过预先建立标识符元数据索引表，待具体识别时，将待识别数据的各个属性名称或字段名称，逐个与元数据表中的标识符进行比对。标识符元数据索引表应当包括标识符名称、含义、格式要求、常用数据类型、常用字段名称等信息。查表识别法适用于数据集格式和属性相对明确的去标识化场景。

（2）规则判定

组织通过总结可能涉及直接标识符和准标识符的数据格式和规律，确立相关标识符识别规则，然后通过运行软件程序，自动化地从数据集中识别出标识数据。

结构化数据和非结构化数据的标识识别均可适用规则判定法。如通过建立身份证号识别规则，识别非结构化存储的司法判决书中的身份证号。

（3）人工分析

在必要场景下，组织通过人工发现和确定数据集中的直接标识符和准标识符。人工分析法适用性较强，当数据集中有特别含义的数据，或数据具有特殊值、容易引起注意的值，或者数据集中的多个不同数据子集之间存在关联、引用关系时，人工分析可以针对性地识别和分析。

4.对标识符进行处理

对数据集进行去标识化前，应当先通过数据清洗，形成规范化或满足特定格式要求的数据。在此基础上，针对不同特征和处理要求的数据类型，考虑去标识化的影响，在可接受的被重新识别风险范围内尽量满足数据可用性的最低要求，选取有效的去标识化技术方法和模型进行处理。具体可参考本节“（三）常见技术方法”。

技术选择需要考量相关因素包括：数据是否可以删除，是否需要保留至少若干个类别的数据项；去标识后的数据是否需要保持唯一性、可逆性，是否需要保持原有的数据格式、表达顺序、统计特征等；是否可以对属性值实施随机噪声添加；以及运用该去标识化技术的成本考量、可承受的重新标识风险范围和业务影响等。

5.验证数据处理结果

对数据去标识化结果进行验证，确保处理后的数据在安全性和可用性方面符合预设要求。

（1）安全性验证

验证经去标识化处理后数据的安全性，确保所生成数据被重新识别的风险在组织预设的可接受风险范围内。组织可以通过检查生成的数据结果、检查去标识化过程及记录、开展入侵者测试等方式验证去标识化数据的安全性。

（2）有用性验证

分析去标识化后的数据对于预期应用和业务的影响，判断处理后数据的质量是否还能满足预期业务用途。组织可以对原始数据和去标识化后数据分别执行统计计算，并对计算结果进行比较，判断去标识化后的计算结果是否仍可接受。

6.评估被识别风险

对去标识化后的数据进行标识符被识别的风险进行评估，与预期可接受的风险阈值进行比较。若风险超出阈值，需继续进行调整直到满足要求。标识符被识别风险评估常见的流程包括评估准备、定性评估、定量评估、形成评估结论等环节，组织可借鉴《GB/T 42460-2023 信息安全技术 个人信息去标识化效果评估指南》进行流程设计。

按照标识符被识别的风险从高到低，可以将相应的风险阈值划分为高风险、较高风险、可控风险、低风险 4 个等级。

高风险（4 级）：能直接识别主体或敏感属性的数据，即包含直接标识符的数据；较高风险（3 级）：仅消除直接标识符的数据，即删除了直接标识符，但仍包含准标识符的数据；可控风险（2 级）：消除直接标识符和准标识符的数据，即对直接标识符和准标识符均进行了处理，在不借助额外信息的情况下，无法识别或关联识别个人信

息主体或特定标识内容；低风险（1 级），不再保留个体颗粒度的聚合数据，如总计数、最大值、最小值、平均值等。

（三）常见技术方法

本报告将仍保留原始数据个体颗粒度的技术类型，纳入去标识化技术方法范畴。部分技术方法参考了《GB/T 37964-2019 信息安全技术 个人信息去标识化指南》。组织根据需要选择相应的去标识化技术，常见的去标识化技术包括数据抽样技术、加解密技术、假名化技术、抑制遮盖技术等，不同技术之间可以结合使用。

1.数据抽样技术

数据抽样是通过选取数据集中有代表性的子集来对原始数据集进行分析和评估。对数据集进行随机抽样能够增加识别出特定标识符的不确定性，可以作为后续应用其他技术强化去标识化效果的初步处理。

数据抽样的方式较多，需要根据数据集的特点和预期的使用场景进行选择，包括随机抽样、等距抽样、分层抽样、整群抽样等。

2.加解密技术

加解密技术是指利用算法对数据进行加密和解密操作，以密码学为基础构建加密函数，输入敏感数据和相关标识符，输出处理后的加密隐藏数据。同时在有需要的时候，可以对数据进行解密操作，即在拥有密钥的条件下，可以对标识符进行复原。常见的数据加密方法包括确定性加密、保序加密、保留格式加密、同态加密等。

确定性加密：指通过确定性加密结果替代数据中的标识符值。确

定性加密是一种非随机加密方法，可以保证数据真实可用，一定程度上保证数据在统计处理、隐私防挖掘方面的有用性，也可以生成用于精准匹配搜索、数据关联及分析的微数据。对确定性加密结果的分析多用于检查数据值是否相等。

保序加密：指通过保序加密值替代微数据中的标识符值。保序加密同样是一种非随机加密方法，密文的排序与明文的排序相同。对保序加密结果的分析多用于检查数据是否相等和排序关系比较。

保留格式加密：指加密过程要求密文与明文具有相同的格式，可用保留格式加密值替代微数据中的标识符值。保留格式加密可以保证加密后的数据具有与原始数据相同的格式和长度，有助于在不需要修改应用系统匹配格式的情况下实现去标识化。

同态加密：指将原始数据加密后，对得到的密文进行特定的运算，得到的计算结果等价于基于原始明文数据直接进行相同计算所得到的数据结果。同态加密是一种随机加密，对经过同态加密的数据进行处理得到相同的输出结果，处理过程不会泄露任何原始内容。

3.假名化技术

假名化技术是指使用虚构的名称或数值，替换原始数据的直接标识符或准标识符的过程。假名化技术保留了原始数据的唯一性特点，也被称为编码。不同数据在假名化处理后依然可以进行关联，并且不会泄露原始标识符。当需要唯一区分数据值并且没有保留关于原始属性的直接标识符的字符或任何其他隐含信息时，可以使用假名化技术。假名可以独立生成或借助密钥编码生成。

独立生成假名：即不依赖于被替代的原始值，生成独立于标识符的假名创建技术，如使用随机值代替标识符原始值。组织需要创建假名与原始标识的分配表，并采取适当的技术与管理措施限制和控制对该分配表的访问。

基于密钥的假名编码：即基于密码技术的标识符派生假名创建技术，通过对属性值采用加密或散列等密码技术生成假名，也被称为对标识符进行“密钥编码”。其中加密技术生成的假名可以用合适的密钥及对应的算法解密。

4.抑制遮盖技术

抑制遮盖技术即对需要进行处理的标识符或数据项进行删除或屏蔽。抑制技术主要适用于分类数据，可用于数值与非数值数据属性，执行相对容易，通过直接删除或屏蔽降低关联识别的风险，且可以保持数据的真实性，但会造成一定程度的信息缺失。但过多的抑制会影响数据的效用，为保证数据的可用性，组织需要对抑制的数据项数量和范围设定上限。抑制遮盖需要是永久性的，而不仅仅是“隐藏”功能，如果底层数据仍然可访问或编辑，则未达到抑制遮盖效果。根据抑制方式的差异，抑制遮盖技术可以分为直接删除或字符掩码屏蔽。

直接删除：即从数据集中直接删除相关标识符，或删除标识符中的部分属性或内容，或者删除涉及特定属性标识符的数据记录。

字符掩码：通过使用一致的符号（例如“*”或“x”）来替换原数据标识符或标识符中的部分数值。区别于仍具有唯一性的假名，进行同一属性的数值所替换的字符掩码均为相同，具有一致性。

五、数据匿名化规程

（一）处理目的

1.促使标识难以复原

数据匿名化处理是数据去标识化后应用相关技术使相关标识符难以复原的过程，是数据去标识化的进一步处理。与数据去标识化相比，经匿名化处理后的数据即便借助了额外信息也难以识别特定自然人和已被处理的标识符。

2.符合风险可接受水平

任何数据均有被复原的可能。数据匿名化处理并非追求完美、绝对的匿名化状态，强调的是运用匿名化技术将原始数据相关标识符的可识别性降低到监管和组织可接受的风险水平。如果信息主体和相关标识符的识别需要不合理的时间、努力或资源，则不视为是可复原的。

3.支持统计、训练用途

经匿名化处理的数据，数据颗粒度、精确度受到影响，不再保留个体数据记录。例如，经匿名化处理的个人信息，不再属于个人信息范畴。与基于个体特征识别的用户画像、设备定位等用途不同，对数据匿名化处理主要为了支撑统计分析、算法训练、科学研究等场景。

（二）处理流程

数据匿名化的流程通常包括明确匿名化处理对象、设定匿名化处理目标、先行去标识化处理、实施数据匿名化处理、评估匿名化效果、定期追踪复原风险六个步骤。

1.确定匿名化对象

根据法律要求和业务用途，确定需要进行匿名化处理的数据类型和范围。

（1）按照监管要求确定处理对象

例如，组织遵照《汽车数据安全管理若干规定(试行)》要求，因保证行车安全需要，在无法征得个人同意采集到车外个人信息且需要向车外提供时，对相关数据进行匿名化处理，包括删除含有能够识别自然人的画面，或者对画面中的人脸信息等进行局部轮廓化处理等。

（2）遵循最小必要原则确定处理对象

例如，征信机构按照《征信业务管理办法》规定，在个人不良信息保存期限届满时，将个人不良信息在对外服务和应用中删除；作为样本数据继续使用的，进行匿名化处理。

（3）履行约定或承诺义务确定处理对象

例如，组织按照《GB/T 35273-2020 信息安全技术 个人信息安全规范》规定，在相关数据超出个人信息约定的存储期限或达成处理目的后，以及组织停止运营其产品或服务时或用户注销账户时，对个人信息进行删除或匿名化处理。

（4）基于业务开展需要确定处理对象

例如，国家卫生健康委等四部门发布的《涉及人的生命科学和医学研究伦理审查办法》中，将“使用匿名化的信息数据开展研究”作为“免除伦理审查”的情形之一，组织为减少科研业务不必要的合规负担，使用匿名化数据开展涉及人的生命科学和医学研究。

2.设定匿名化目标

满足安全性要求是数据匿名化处理的首要目标。组织应结合业务场景和安全防护管理要求，根据数据的性质、使用环境和使用的匿名化技术等，结合匿名化数据的主要用途和使用场景，对标识符被复原的可能性进行分析，评估相应的风险，设定可被组织和监管部门接受和认可的风险阈值。

3.先行去标识化处理

组织应将去标识化作为匿名化处理的一部分执行，结合前述数据去标识化业务规程，识别相关直接标识符和准标识符，针对性进行去标识化处理，先行满足数据的“不可识别性”要求，达到数据在不借助额外信息的情况下无法直接识别特定自然人或相关标识符的效果，为后续的匿名化操作奠定基础。

4.实施匿名化处理

组织针对已去标识化的数据应用匿名化技术，使未获得授权主体不能轻易地将该数据与可能包含额外信息的其他数据相结合，从而难以复原特定自然人信息或相关标识符。不同匿名化技术的技术特点不同，选择处理技术时，应当结合数据类型和性质、业务场景、处理目的等进行综合考量，相关技术具体可参考本节“（三）常见技术方法”。选择匿名化技术过程中需要考虑以下因素：

一是考虑所采用的匿名化技术进行处理后数据是否仍满足预期效用。匿名化处理可能对原始数据格式、数值和表达方式进行较大变动，将对原始数据的保真性、颗粒度形成较大影响。

二是考虑将相关匿名化技术和去标识化技术组合使用，形成系统性匿名化处理方案。例如，如果某个属性类别的数值直接删除不会影响数据效用，可以选择抑制遮盖技术对相关数据项予以删除处理。

三是考虑不同匿名化技术的适用场景。结合技术特点和目标要求选择相应技术。如针对连续值属性的数据可以采用噪声添加、数据扰动等随机化技术，针对无需体现个体数据记录的情形可以采用聚合统计等技术。同时，针对同一场景或同一数据类型的匿名化处理，也可多种匿名化技术结合使用。

5.评估匿名化效果

组织应用适当的匿名化技术后，应当对匿名化处理的效果进行分析评估。计算标识符被复原或重新标识风险的方法需要综合考虑数据因素和环境因素。《GB/T 42460-2023 信息安全技术 个人信息去标识化效果评估指南》提供了“基于 K 匿名模型的重标识风险计算方案及评估事例”，可供组织借鉴参考。

k-匿名值是一种计算数据集重新识别风险水平的方法，指数据集中可以分组在一起的相同记录的最小数量。在评估数据集的总体重新识别风险时，通常采用最小值来表示最坏情况。k-匿名值较高意味着重新识别的风险较低，k 匿名性值较低意味着风险较高。K-匿名值为 1 表示记录是唯一的。k-匿名值需要结合实际场景、处理目标和安全等级要求进行具体设定。在可能的情况下，应设置更高的 k-匿名阈值，以最小化任何重新识别风险。需注意，k-匿名可能不适用于所有类型的数据集或其他复杂情形。

6.定期追踪复原风险

组织应当定期追踪内外部相关主体对匿名化处理数据的使用情况，评估新技术、新数据、新主体的引入可能带来的标识符被复原的新隐患，考虑数据的流通范围、可能的技术演变等，以及未知的跨库数据可能导致与匿名数据集匹配的情形，进而采取适当措施保护相关标识符免受复原识别和披露的风险。

（三）常见技术方法

本报告将不再保留原始数据个体颗粒度，或原始数据记录真实性已受到显著减损，或原始数据记录不对外披露的技术类型，纳入匿名化技术方法范畴。部分技术方法参考了《GB/T 37964-2019 信息安全技术 个人信息去标识化指南》。组织可结合具体场景单独或组合选用聚合统计、泛化、随机化、数据合成、隐私计算等技术进行处理。

1.聚合统计技术

聚合统计技术指将数据集从记录列表转换为汇总值或相关统计值的方法，可以视为求和、计数、平均、最大值与最小值等一系列统计技术的集合。由于聚合统计技术的输出是“统计值”，该值有利于对数据进行整体报告或分析，产生的结果能够代表原始数据集中的所有记录，且不会披露任何个体记录，很大程度上降低了个体的标识符被重新识别的风险。当组织不需要单独的数据记录且聚合数据足以满足预期效用时可以采用聚合统计技术。

例如，2022 年我国 18-80 岁女性平均体重 59.8kg，如果以平均体重来标识数据集中每个人的体重值，则未获得授权主体无法根据体重

属性将某一条数据记录（女，北京，1.63m，59.8kg，1990 年 9 月 1 日）关联到特定个人。

使用聚合统计技术应注意两方面的应用要求：一是数据聚合统计可能会显著改变数据的初始用途，因为输出的结果为统计值，无法反映每一单独数据记录的特征；二是应用聚合统计技术对原始数据的样本量具有一定要求，若原始数据记录的数量很少，则结合其他数据容易推断出其中具体的单独数据记录的特征。

2.泛化技术

泛化技术也是一种概括方法，又被称为离散化处理，是通过降低数据所选属性的颗粒度、精度，对数据进行更概括、抽象描述的匿名化技术。使用泛化技术的目标是减少属性唯一值的数量，使得被泛化后的值被数据集中多个记录所共享，从而增加某个特定数据记录被推测出的难度。例如，将一个人的年龄转换为年龄范围，或将精确位置转换为不太精确的位置。

数据泛化的程度需要均衡预期目的和风险控制两方面要求。数据范围过大可能意味着数据效用的显著损失，数据范围过小可能意味着几乎不修改数据，特定数据记录仍然很容易重新识别。常见的泛化方法包括取整、顶层与底层编码等。

取整：即为数值型标识符选定一个取整基数，然后将每个具体值向上或向下取整至最接近取整基数的倍数。向上还是向下取整按概率确定，该概率值取决于观察值与最接近取整基数倍数的接近程度。例如，如果取整基数为 10，观察值为 7，应将 7 向上取整至 10，概率

为 0.7，若向下取整至 0，概率为 0.3。同时还可以按要求进行受控取整，如确保取整值的求和结果与原始数据的求和取整值相同。

顶层与底层编码：即为数值型标识符设定一个可能的取值范围，用高于或低于所设定的临界值的描述替换某一特定数据记录在该属性上的具体数值，主要适用于连续或分类有序的数据类型。例如，将某一员工的薪水值设置为“高于 10000 元”，其中“10000”为高收入值的界限，而不记录准确的金额。

3.随机化技术

随机化技术指通过随机修改数据属性的值，使得随机化处理后的值区别于原来的真实值。随机化技术降低了未经授权主体从同一数据记录中根据其他属性值推导出某一属性值的能力，会对原始数据记录的真实性造成一定影响。常见的随机化技术有数据扰动、数据置换等。

数据扰动：又称噪声添加，即通过添加随机值来修改数据中的值，同时尽可能保持该属性在数据集中的原始统计特性，包括属性的分布、平均值、方差、标准偏差、协方差以及相关性。数据扰动的程度应当控制在一定范围内容，如果扰动程度太小，匿名化效果较弱；如果扰动程度太大，最终值将与原始值相差太大，数据集的效用可能会降低。数据扰动通常用于数值型标识符，例如对日期前后随机+/-3 个自然日。

数据置换：相当于一种洗牌，即重新排列数据属性中的标识符，使之无法与原始记录对应，但各个属性的值仍在数据集中表示，保持了原有数据集中所选属性整体的准确统计分布。数值型标识符和非数值型标识符均可使用数据置换技术。在保持所选属性之间原有相关性

的情况下，置换算法可用于单个或多个属性。例如，对姓名进行假名化处理后，对职位、性别、年龄等进行乱序重排。

4.数据合成技术

数据合成技术是显著修改原有数据的所有属性，重新合成产生新的微数据的方法。合成数据集与原始数据的特征相符，可根据所选的统计特性随机生成，但不会体现原始数据的任何特定记录。但若是合成后数据与原始数据的拟合度过高可能会存在被关联识别风险。

通常合成数据的生成会在假名化的基础上，采用随机化技术与抽样技术对真实数据集进行多次或连续转换。合成数据通常适用于应用程序开发、测试和应用，将其作为真实数据的替代项，帮助数据开发主体获得与基于真实数据的处理同样的效果。

5.隐私计算技术

隐私计算技术是指在保护数据本身不对外泄露的前提下实现数据分析计算的技术集合，通过对所涉及的隐私信息进行描述、度量、评价和融合等操作，形成一套符号化、公式化且具有量化评价标准的隐私计算方法，达到对数据“可用不可见”的目的。目前主流的隐私计算技术主要分为三大方向：一是以多方安全计算为代表的基于密码学的隐私计算技术；二是以联邦学习为代表的人工智能与隐私保护技术融合衍生的技术；三是以可信执行环境为代表的基于可信硬件的隐私计算技术。

多方安全计算：是指在无可信第三方的情况下，多个参与方共同计算一个目标函数，在不泄露己方数据的同时完成数据计算，并且保

证每一方仅获取自己的计算结果，无法通过计算过程中的交互数据推测出其他任意一方的输入数据。多方安全计算通常应用于联合数据分析、数据可信交换、分布式投票、隐私竞标和拍卖、黑名单安全查询、数据库检索等场景。

联邦学习：是指实现在本地原始数据不出库的情况下，各方通过对中间加密数据的流通、参数交换和处理，共同建立虚拟的共有模型，完成多方联合的机器学习训练。联邦学习可以从技术上有效解决数据孤岛问题，让参与方在不泄露各自拥有的用户数据的基础上，实现联合建模和 AI 协作，加速隐私计算在不同场景的应用与落地。根据参与方的数据分布和特征重叠情况的不同，可以分为横向联邦学习、纵向联邦学习和联邦迁移学习。

可信执行环境：是指将需要保护的数据和代码存储在可信执行环境中，即通过软硬件方法在中央处理器中构建一个安全的区域，对这些数据和代码的任何访问都必须通过基于硬件的访问控制，防止它们在使用中未经授权被访问或修改，从而保证其内部加载的程序和数据在机密性和完整性上得到保护。可信执行环境是一种硬件解决方案，安全性较高，但运维成本相应上升，多用于本地和远程验证场景。

六、数据处理环境要求

（一）管理制度要求

组织应当遵守法律法规及强制性标准的相关要求，衔接自身数据管理制度，制定数据清洗、去标识化、匿名化处理各环节的审批流程，推进数据分类分级管理，梳理特殊数据类型的内、外部特别管理要求，

在此基础上细化数据清洗、去标识化、匿名化处理的权限要求和操作规范，并嵌入组织内部管理机制。

（二）技术能力要求

组织应当强化数据处理的基础技术保障，具备数据收集、存储、加工、分析、挖掘和安全防护的各类技术工具，具有安全、便捷、高效的技术应用系统和可信环境，熟知数据清洗、去标识化、匿名化处理的常见技术方法和应用特点，结合业务场景和内外部要求，统筹组合形成平衡数据安全要求和业务应用目的的有效数据处理技术方案。

（三）人员能力要求

组织应当提升内部人员的数据处理能力和安全防护水平，明确各岗位数据合规职责和数据处理权限要求，定期组织数据处理技能培训和安全合规教育，要求参与数据清洗、去标识化、匿名化处理的人员应当具备相应的数据处理能力，严格按照数据安全管理制度和流程进行操作。必要情况下，组织可以寻求第三方技术服务机构、法律服务机构、审计咨询机构、数据安全防护机构、检测认证机构等协助提供技术能力和业务合规支持。

（四）过程控制要求

组织应当推进数据处理过程的实时可控和动态审计，采取措施清晰记录数据清洗、去标识化、匿名化处理过程的细节、使用的参数和执行情况，监控审查去标识化各步骤实施过程，及时发现已经出现或可能出现的错误或偏差，有效采取措施进行纠正和防护，并对监控审查过程进行记录，便于日后审查、维护、回溯和审计。同时加强对第

三方接收者的数据授权和授权跟踪管理，采取技术保障措施和商业流程防范去标识、匿名化数据的再识别和意外泄露。

（五）事故管理要求

组织应当完善数据处理风险和安全事件管理机制，做好数据风险识别、风险评估、风险处置等工作，制定并实施数据安全事件应急预案，针对不同等级的风险采取针对性的风险处置措施，关注涉及数据标识符、数据映射表、匿名化处理记录表等信息的泄露风险，防范恶意重新标识行为。发生数据泄露、篡改、丢失等安全事件的，应当立即采取补救措施，及时通知管理机构并按规定告知相关数据主体。

附件一：常见直接标识符和准标识符示例

附件二：常见标识符的去标识化或匿名化参考

附件三：部分数据处理技术方法应用建议

附件一：常见直接标识符和准标识符示例

（一）直接标识符示例

直接标识符通常表现为在特定环境下可以单独识别特定自然人或数据所描述特定对象的识别号码、特征或代码。需注意，标识符的识别难度并不与数据的敏感程度直接挂钩。本报告分别列举了个人数据、企业数据、物联网数据的部分直接标识符示例，常见的直接标识符包括但不限于:

类型	序号	常见直接标识符
个人数据	1	姓名
	2	公民身份号码
	3	护照号
	4	工作学习编号，包括工号、学号等
	5	电话号码
	6	传真号码
	7	银行账户
	8	驾照号
	9	车牌号
	10	社会保障号码
	11	健康卡号码
	12	病历号码
	13	网络账号、昵称等
	14	网络身份标识号（ID）
	15	个人移动终端设备标识符
	16	详细住址
	17	电子邮件地址
	18	个人行踪轨迹

	19	生物识别码，包括指纹和声纹等识别码
	20	全脸图片图像及其他任何可比对的图像
企业数据	1	组织机构名称
	2	营业执照编号
	3	统一社会信用代码
	4	法定代表人姓名
	5	税务登记证号
	6	社会保险登记证号码
	7	统计登记证号码
	8	银行账户信息
	9	组织许可证号
	10	企业注册地址
	11	网络和系统账号信息
	12	网站标识码，互联网协议（IP）地址号
	13	网络通用资源定位符（URL）
	14	合同编号
	15	商业发票编号
物联网数据	1	设备标识符和序列号
	2	设备位置信息
	3	设备使用记录
	4	设备故障或警报记录
	5	商品条码
	6	货运设备识别码
	7	集装箱识别代码
	8	医疗器械唯一标识（UDI）
	9	数字版权唯一标识符（DCI）
	10	气象数字对象标识符（MOID）

（二）准标识符示例

准标识符通常指在相应环境下无法单独识别特定自然人或数据所描述的特定对象，但结合其它信息可以进行识别的属性、号码、特征或代码。本报告分别列举了个人数据、企业数据、物联网数据的部分准标识符示例，准标识符范围较广，常见的准标识符包括但不限于：

类型	序号	常见准标识符
个人数据	1	性别
	2	出生日期或年龄
	3	事件日期（例如入院、手术、出院、访问相关日期）
	4	地理范围（例如邮政编码、建筑名称、地区）
	5	血型、身高、体重等体征
	6	疫苗接种状态、病史等健康状况
	7	国籍、籍贯
	8	族裔血统、民族
	9	宗教信仰
	10	语言
	11	职务、工作单位、部门等职业信息
	12	婚姻状况
	13	受教育水平
	14	学习、工作年限
	15	收入状况
企业数据	1	组织设立时间
	2	组织信用评级
	3	资产设备情况
	4	员工情况
	5	客户分布
	6	产品类型

	7	供应链渠道
	8	营收情况
	9	系统日志
	10	工艺参数
物联网数据	1	传感节点标识信息
	2	环境参数信息（温度、湿度、气压、风速、光线等）
	3	设备规格信息
	4	设备健康状态
	5	生产日期
	6	检验日期

附件二：常见标识符的去标识化或匿名化参考

去标识化和匿名化的相关技术和方法没有严格界分，匿名化技术抗重新识别的风险能力相对更高。组织可以根据相关技术特点统筹组合使用，形成平衡数据安全要求和业务应用目的的有效数据处理方案。本报告借鉴《GB/T 37964-2019 信息安全技术 个人信息去标识化指南》列举了部分标识符的去标识化或匿名化参考，更多标识符的处理方法组织还可以参考该标准附录 C“去标识化模型和技术的选择”。

标识符	去标识化或匿名化方法参考
姓名	**假名化。**构建常用人名字典表，并从中选择一个来表示，如先构建常用的人名字典表，包括龚小虹、黄益洪、龙家锐等，假名化时根据按照顺序或随机选择一个人名代替原名。如使用“龚小虹”取代“张三丰”。
	加解密技术。采用密码或其他变换技术，将姓名转变成另外的字符，并保持可逆特性。如使用密码和字符编码技术，使用“SGIHLIKHJ”代替“张三丰”，或使用“Fzf”代替“Bob”。
	抑制遮盖。直接删除姓名或使用统一的“×”来表示。如所有的姓名都使用“***”代替。
	泛化编码。使用概括、抽象的符号来表示，如使用“张先生”来代替“张三”，或使用“张某某”来代替“张三”。
	随机化替代。使用随机生成的汉字来表示，如使用随机生成的“辰筹猎”来取代“张三丰”。
身份证号	**加解密技术。**采用密码或其他变换技术，将身份证号转变成另外的字符，并保持可逆特性。如使用密码和字符编码技术，使用“SF39F83”代替“440524188001010014”。
	部分抑制遮盖。屏蔽身份证号中的一部分，以保护个人信息。如“440524188001010014”可以使用“440524*********0014”

	“440524188*****0014”代替。上述方法可分别用在需保密出生日期、保密出生日期但允许对数据按时代统计分析等场景。
	全部抑制遮盖。直接删除身份证号或使用统一的“*”来表示。如所有的身份证号都使用“**********”代替。
	数据合成。采用重新产生的数据替代原身份证号，如使用数据集中的记录顺序号替代原身份证号，或随机产生符合身份证号编码规则的新身份证号代替原始值。
电话号码	**加解密技术。**采用密码或其他变换技术，将电话号码转变成另外的字符，并保持可逆特性。如使用密码和字符编码技术，使用“15458982684”代替“1988888888”。
	部分抑制遮盖。屏蔽电话号码中的一部分，以保护号码信息。如“19888888888”可以使用“198*********”“198****8888”代替。
	全部抑制遮盖。直接删除电话号码或使用统一的“*”来表示。如所有的电话号码都使用“******”代替。
	随机化替代。使用随机生成的一串数字来表示，如使用随机生成的“2346544580”来取代“19888888888”。
企业合同编号	**部分抑制遮盖。**为了便于合同管理，企业合同编号通常由合同类型、发起部门、区域字母、签订日期、随机数字等部分组成。向不具备特定权限的员工披露合同信息时，可以将企业合同编号中的一部分或多部分屏蔽。如将劳动合同“LD-RL-BJ-20200701-A3”中的区域字母、日期中的月份、随机数字屏蔽，使用“LD-RL-**-2020****-**”替代。
	聚合统计。采用相关统计值代表原始数据记录，反映合同签订情况。如对企业 2020 年签订的设备采购类合同进行统计，转化为 2020 年某业务部门签订北京区域的采购合同 18 项。
地址	**部分抑制遮盖。**屏蔽地址中的一部分，以保护地址信息。如使用“江西省 XX 市 XX 县”来代替"江西省吉安市安福县”
	全部抑制遮盖。直接删除姓名或使用统一的“*”来表示。如

	所有的地址都使用“**********”代替。
	泛化编码。使用概括、抽象的符号来表示，如“江西省吉安市安福县”使用“南方某地”或“J省”来代替。
	数据合成。采用重新产生的数据替代原地址数据。如使用“黑龙江省鸡西市特铁县北京路23号”代替“江西省吉安市安福县安平路1 号”。
环境参数信息	**部分抑制遮盖。**对室外传感器设备采集的温度、风速信息予以屏蔽，只显示湿度、气压、光线等信息，使其无法辨识是哪一特定位置传感器采集的环境参数。
	泛化编码。降低室外传感器设备采集的温度、湿度等信息精度，为其设置一定阈值，如温度＞30℃，空气相对湿度低于70%。
	聚合统计。采用相关统计值代表原始数据记录。如将某地过去30日每天的温度、风速记录，转化为某地上月平均气温30℃，风力≥8级的日数为5天。
日期	**部分抑制遮盖。**对日期中的一部分做屏蔽，如1880年某月1日代替1880年1月1日。
	全部抑制遮盖。直接删除日期数据或使用统一的“×”来表示。如所有的数值都使用“某年某月某日”代替。
	泛化编码。使用概括、抽象的日期来表示，如使用1880年代替1880年1月1日。
	随机化-数据置换。使用数据集中其他记录的相应数值代替本记录的数值。如设定规则，将记录集中的所有的日期数据取出并全部打乱位置后（其他属性数据位置不变）放回到原数据集中。这种方法有利于保持数据集的统计特性。
	随机化-数据扰动（噪声添加）。将相对微小的随机数，加到原始数值上并代替原始数值。如对于出生日期1880年1月1日，产生随数值32天，加到原始数值后将其变为1880年2月2日。
	数据合成。采用重新产生的数据替代原日期数据，如使用“1972年8月12日”代替“1880年1月1日”。

数值型标识符	**部分抑制遮盖。**使用数值的高位部分代替原有数值，如百分制考试成绩全部使用去掉个位数、保留十位数的数值代替。
	全部抑制遮盖。直接删除数值或使用统一的“*”来表示。如所有的数值都使用“*****”代替。
	聚合统计。使用概括、抽象的符号或统计数值来表示，如“有4个人，分别是蓝色、绿色和浅褐色的眼睛"来代替"有1个人是蓝色眼睛，2个人是绿色的眼睛，1个人是浅褐色的眼睛"。
	泛化-顶层和底层编码。大于或小于一个特定值的处理成某个固定值。例如，年龄超过70岁的一律用“大于70岁”描述，以保障满足此条件的人数多于20 000人。
	随机化-数据置换。使用数据集中其他记录的相应数值代替本记录的数值。如设定规则，将记录集中的所有的身高数据取出并全部打乱位置后(其他属性数据位置不变)放回原数据集中。这种方法可以保持数据集的统计特性不变。
	随机化-数据扰动（噪声添加）。相对原始数据，产生微小的随机数，将其加到原始数值上并代替原始数值。如对于身高1.72m，产生随机数值-0.11m，加到原始数值后将其变为1.61m。
	数据合成。采用重新产生的数据替代原始数据，数据产生方法可以采用确定性方法或随机性方法。如使用“19”岁年龄代替“45”岁年龄。
	联邦学习。如在第三方服务商协调下，某地银行和商超在本地数据不出库情况下通过中央服务器进行联合建模，银行基于掌握的交叉用户性别、年龄数据，匹配该用户在商超的购物频次、消费金额等数据，预测该用户的信用情况。商超基于掌握的交叉用户购物频次、消费金额数据，匹配该用户在银行记录的性别、年龄数据，形成不同性别、不同年龄段用户的消费偏好，辅助产品销售。

附件三：部分数据处理技术方法应用建议

为向相关组织提供具体技术的选取参考，本报告就常见的和正在积极推动落地的部分数据处理技术进行介绍。部分技术实现简单易操作，但难以单独解决复杂问题。部分技术理论可行，但落地场景需要持续拓展，商用化程度还需提高。相应技术的具体适用方案组织可以向相关技术提供方咨询。

技术方法	技术特点	适用场景	应用限制	应用现状
同态加密	对原始数据进行加密，使得加密数据和原始数据进行相同处理时，结果相同。可在不解密情况下对密文进行计算和分析	医疗数据模型构建、电子商务验证、隐私数据求交集和检索等	计算消耗和性能要求较高，存储成本较高	技术应用理论上可行，商用化程度还需提高
抑制遮盖	直接删除或采用字符掩码屏蔽隐私数据或部分字段，并保证底层数据无法访问或编辑，可以保持数据的真实性	主要适用于分类数据，可用于数值与非数值数据属性	容易导致信息丢失，过多抑制会影响数据效用	执行相对容易，日常应用较多，但需要与其他技术结合提升抗风险能力
聚合统计	将个体的数据转化为求和、平均、最大值与最小值等统计值，产生的结果能够代表原始数据集的所有记录，但不会披露任何个体记录	适用于连续数据的整体报告或分析，且不需要反映每一单独数据记录的特征	对个体特征的分析受限，可能会降低数据的有用性；对原始数据样本量有一定要求	适用场景有限，多见于整体数据披露，需结合应用目选择具体统计方法

泛化		降低数据所选属性的颗粒度、精度，对数据进行更概括、抽象的描述，可以保护数据的真实性	多用于数值处理，用于可被概括处理且仍对预期目的有用的数据	需注意泛化范围的选取，过大可能过度破坏数据精确度，过小面临较大重新识别风险	泛化技术实现简单，简单数据可直接用电子表格软件处理，日常数据处理中较常用
随机化		随机修改数据属性值，使得处理后的值区别于原来的真实值，包括添加噪声数据进行扰动和置换相互间的属性值等	适用于需要保留个体数据记录，并保留所选属性的统计分布特征的情形	会对原始数据记录真实性造成影响，对个体数据准确性有要求时避免使用	随机化技术实现简单，已得到成熟应用，在进行整体分析时较常用
数据合成		显著修改所有数据属性，根据原始数据的统计特性重新拟合产生新的微数据，不再保留原始数据集的数值	可作为原始数据的替代项适用于应用程序开发、测试和应用	如果合成后数据与原始数据拟合度过高可能会导致信息泄露	在实际应用中逐渐增多，对计算能力和计算效率要求较高
隐私计算	**多方安全计算**	在无可信第三方的情况下，多个参与方共同计算一个目标函数，每个参与方除计算结果外不能获得其他参与方输入的任何数据	通常应用于联合数据分析、数据可信交换、隐私竞标和拍卖、安全查询检索等场景	具有去中心化特征，对计算能力和各参与方联合协同要求较高	商用程度待提升，参与方越多，算法越复杂，计算成本越高，多数只支持两方计算，应用场景有限

	联邦学习	在中央服务器或服务提供商协调下，多个实体协作实现联合建模和AI协作，而本地原始数据不出库	应用于联合数据统计、联合机器学习建模等场景	需第三方参与，不同参与方的设备、数据存在异构性，需要大量协调和标准化	目标受众和落地场景仍较少。算力需求高，通信网络成本高。需与区块链等技术结合
	可信执行环境	通过提供一个可信的、隔离的安全区域（飞地）进行数据处理，防止数据未经授权的访问和修改	多用于本地和远程验证，如移动金融支付、指纹认证、面部和声纹识别等	硬件运维成本相应上升，同时需要防止硬件漏洞和加强人员及密钥管理	应用场景受限，硬件依赖导致某些场景无法适用，某些设备有国产化要求

参考资料

[1]GB/T 35273-2020 信息安全技术 个人信息安全规范

[2]GB/T 37964-2019 信息安全技术 个人信息去标识化指南

[3]GB/T 42460-2023 信息安全技术 个人信息去标识化效果评估指南

[4]GB/T 36344-2018 信息技术 数据质量评价指标

[5]GB/T 38606-2020 物联网标识体系 数据内容标识符

[6]DB52/T 1540.3-2020 政务数据 第 3 部分：数据清洗加工规范

[7]DB31/T 1311-2021 数据去标识化共享指南

[8]中华人民共和国全国人民代表大会常务委员会.中华人民共和国数据安全法.2021 年 6 月 10 日.

[9]中华人民共和国全国人民代表大会常务委员会.中华人民共和国个人信息保护法.2021 年 8 月 20 日.

[10]北京市人民代表大会常务委员会.北京市数字经济促进条例.2022 年 11 月 25 日.

[11]北京市经济和信息化局.北京市数字经济全产业链开放发展行动方案.2022 年 5 月 30 日.

[12]ISO/IEC 2st CD 20889,Information technology-Security techniques-Privacy enhancing data de-identification techniques,June 2017.

[13]ISO/IEC 38505, Information technology-Governance of IT-Governance of data-Part 1: Application of ISO/IEC 38500 to the governance of data,March 2017.

[14]Personal Data Protection Commission Singapore, Guide to Basic

Anonymisation.31 March 2022, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Advisory-Guidelines/Guide-to-Basic-Anonymisation-31-March-2022.ashx.

[15]Article 29 Data Protection Working Party (European Commission). Opinion 05/2014 on Anonymisation Techniques. 10 April 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

中国信息通信研究院 产业与规划研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-68033816

传真：010-68033234

网址：www.caict.ac.cn