数据安全

第6章 数据处理与使用安全

计算机与大数据学院 刘延华

福州大学
FUZHOU UNIVERSITY

第6章 数据处理与使用安全

福州大学
FUZHOU UNIVERSITY

6.1 数据处理安全

数据安全是数字经济发展中最关键的安全挑战之一，随着人工智能在产业和技术两个方面都在加快渡过“探索期”，逐步进入“成长期”， 人工智能发展与数据安全将更加深度地交织在一起，数据安全问题已然成为人工智能突破关键转轨期所必须解决的重要制约瓶颈。

人工智能时代的数据安全

6.1 数据处理安全

- **人工智能发展加剧了传统数据安全风险.**

在以“数字新基建、数据新要素、在线新经济”为重要特征的数字经济发展大背景下，人工智能的新发展必然伴随着数据总量的井喷式爆发，各类智能化数据采集终端的加快增长，数据在多种渠道和方式下的流动更加复杂，数据利用场景更加多样，整体数字空间对于人类现实社会各个领域的融合渗透更趋于深层，这将使得传统数据安全风险持续地扩大泛化。

6.1 数据处理安全

- **人工智能催生了各种新型的数据安全风险.**

人工智能通过训练数据集构造和优化的算法模型，因其对于数据资源特有的处理方式，将会带来**数据污染、数据投毒、算法歧视**等一系列的新型数据安全问题。同时，人工智能在自动化网络攻击、数据黑产的应用，使得传统网络安全和数据安全威胁更加复杂，对国家和企业现有的数据安全治理能力形成巨大冲击。

6.1 数据处理安全

（1）**数据污染**可能会导致人工智能算法模型失效.

- 数据污染是数据质量的技术性治理问题，是指数据与人工智能算法不适配，导致算法训练成本激增甚至完全失效。
- 数据污染产生的原因包括训练数据集规模过小、多样性或代表性不足、异构化严重、数据集标注质量过低、缺乏标准化的数据治理程序、数据投毒攻击等。
- 在数据与模型算法适配度极低的情况下，进行模型训练时将会明显带来反复优化、测试结果不稳定等问题，使得人工智能运行的成本大大提高，严重的数据污染甚至直接导致人工智能算法模型完全不可用。

6.1 数据处理安全

（2）恶意数据投毒攻击导致人工智能决策错误.

- **数据投毒**是指恶意攻击者人为地在训练数据集中定向添加异常数据或是篡改数据，通过破坏原有训练数据的概率分布而导致模型产生分类或聚类错误，从而引发人工智能的决策偏差或错误，最终产生恶意攻击者所期待的结果。
- 在自动驾驶、智能工厂等对实时性要求极高的人工智能场景中，数据投毒对人工智能核心模块产生的定向干扰将会直接扩散到智能设备终端（如智能驾驶汽车的刹车装置、智能工厂的温度分析装置等），从而产生灾难性事故后果。

6.1 数据处理安全

（3）数据偏差可能会导致人工智能决策带有歧视性

- **数据偏差**是指人工智能算法决策中所使用的训练数据，因地域数字化发展不平衡或社会价值的倾向偏见，使得数据所承载的信息带有难以用技术手段消除的偏差，从而导致人工智能的决策结果带有歧视性。
- 当下的人工智能主要是通过对训练样本数据的结构和概率进行特征统计，构建输入数据与输出结果的相关度，而并非通过抽象化的逻辑推演获取真正的因果关系，同时机器学习算法带有“黑箱”的不可解释性，因此这种因数据偏差导致的决策歧视难以使用技术性完全解决。

肤色 与 人脸识别

不同区域 与 饮食偏好

6.1 数据处理安全

- 在大

6.2 数据流通安全

（1）人工智能多主体之间的数据交互存在泄露和滥用隐患。

由于人工智能产业生态体系中各主体之间进行数据交互而导致的数据泄露或滥用主要包括三种类型：

其一，由于大量人工智能企业会委托第三方公司或采用众包的方式实现海量数据的采集、标注、分析和算法优化，因而数据将会在供应链的各个主体之间形成复杂、实时的交互流通链路，可能会因为各主体数据安全能力的参差不齐，产生数据泄露或滥用的风险。

6.2 数据流通安全

其二，当下多数人工智能初创企业普遍使用开源学习框架，即通过谷歌、微软、亚马逊等互联网巨头公开的模块化基础算法进行应用开发，因此初创企业对于开源框架、第三方软件包、数据库和其他相关组件等均存在较大的依赖性，且由于缺乏严格的测试管理和安全认证，将会面临不可预期的系统漏洞、数据泄露和供应链断供的安全风险。

6.2 数据流通安全

其三，通过边缘计算的方式进行人工智能系统开发及数据训练是目前企业较为流行的做法趋势，人工智能云服务平台和开发者、应用者的数据交互，将会使部署在云侧和端侧的数据面临比传统信息系统更加复杂的安全挑战。

6.2 数据流通安全

（2）数据孤岛和数据壁垒问题可能导致滋生数据黑产。

由于人工智能发展处于“探索期”向“成长期”过渡的阶段性特点，对于底层数据资源的竞争仍是人工智能企业最关键的市场竞争力体现。然而成熟的数据要素市场尚未形成，数据合法、便捷、安全、低成本的交易流通机制仍是空白，远远无法满足人工智能企业发展对于数据资源的需求。同时，在政府与企业之间、大企业与小企业之间、行业与行业之间，因数据确权、数据安全等问题存在着诸多法律和技术上的数据壁垒，形成了“数据孤岛”，不仅极大制约着人工智能的发展，也成为滋生数据黑产的主要经济动因。

6.2 数据流通安全

（3）人工智能产生的跨境数据流动引发数据安全问题。

在全球数字经济发展不均衡的大背景下， 大型科技巨头在人工智能的数据资源供给、数据分析能力、算法研发优化、产品设计应用等环节分散在不同的国家，而小型初创企业也需要诸多第三方平台和数据分析公司的支撑。因此，无论是企业内部还是与第三方合作，在人工智能技术研发和场景应用中均需要常态化、持续性、高速率、低延时的跨境数据流动。

6.2 数据流通安全

比如，在智能网联汽车领域，智能汽车产生的路况、地图、车主信息等大量数据可能回传境外的汽车制造商，进行产品优化升级和售后服务支撑，将会带来个人敏感数据和重要数据出境后的安全不可控风险。这种人工智能发展引发的跨境数据流动，不仅因各国日益趋严的数据安全规制和本地化要求而面临极大的政策障碍，更将对主权国家的国家安全、数据主权等带来复杂的挑战。

6.2 数据流通安全

（1）

6.3 数据使用安全

（1）智能化的深度挖掘将会威胁公民隐私和国家安全。

深度挖掘是指人工智能技术能够对用户分散、公开甚至匿名化的数据项进行关联分析，从而获得用户无意公开的信息特征和隐私。当前，随着大数据分析和用户画像技术的快速发展，个性化服务变得越来越普遍，各类平台和企业对于用户“数字轨迹”数据的采集成为其提供精准化产品服务的核心基础，这种对于用户习惯行为的长期跟踪和深度分析将使得公民隐私面临安全风险。同时，随着人工智能关联分析技术的发展，通过对公民分散的、单个无意义的数据项进行组合关联分析，能够形成对特定个体识别和特征画像的数据集（比如活动场所、行动习惯、政治态度、宗教信仰等）。这种技术不但本身会直接威

6.3 数据使用安全

（2）对人工智能的逆向还原攻击将会侵犯商业秘密.

人工智能应用的公开访问接口，利用一系列技术手段逆向还原出人工智能的算法模型和训练数据。由于算法模型在部署应用中通常需要将公共访问接口发布给用户使用，攻击者可以利用神经网络等人工智能算法对训练数据集的记忆，通过公共访问接口对算法模型进行黑盒访问，从而分析系统的输入输出和其他外部信息，并推测系统模型的参数及训练数据中的隐私信息。甚至部分攻击者能够通过构造出与目标模型相似度非常高的模型，进行不断地优化逼近，从而实现对算法模型的窃取，进而还原出模型训练和运行过程。逆向还原攻击对算法模型、参数特征的窃取将直接威胁企业的知识产权和网络资产安全，而其对训练数据隐私信息的窃取将对个人隐私构成安全威胁。

6.3 数据使用安全

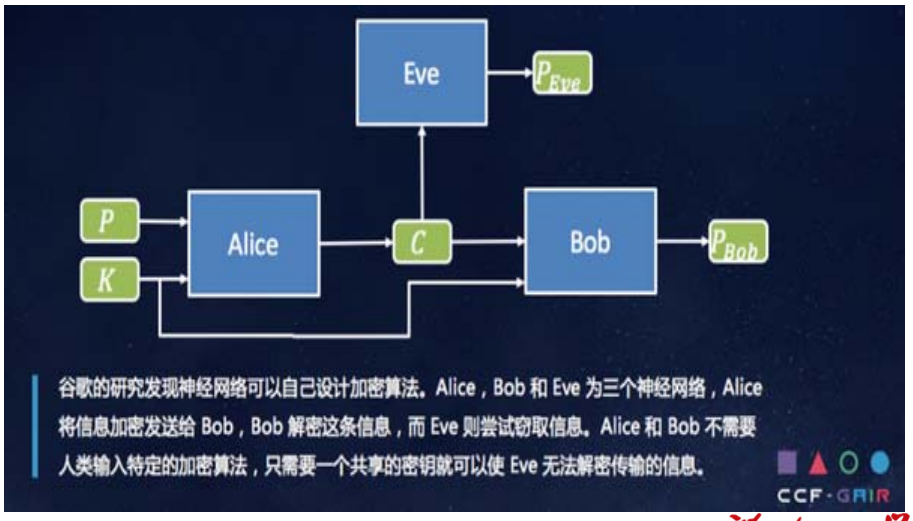
（3）对抗样本攻击将会导致人工智能决策。

错误对抗样本攻击是指在样本数据输入中添加细微、无法识别的干扰信息，导致模型在正常运转中输出一个错误的结果。此类对抗样本攻击既可以是网络空间的虚拟信号错误，也可以是物理世界的实体识别错误。比如在智能网联汽车的无人驾驶中，通过对实体停车或限速标志的精确更改，使得算法模型将其误识别为其他标识，从而引发交通事故。

6.3 数据使用安全

- 人工智能还能生成新的加密算法，Alice和Bob和Eve是三个神经网络，A和B之间共享一个Key，但是没有一个算法，这种情况下经过多轮的训练之后，A和B之间可以用它们自己产生的算法来对数据进行加密，E却不能解密。就是我只是让A和B之间有了一个密钥而已，并没有事先加载任何具体加密算法。这就是在人工智能年代，可以利用AI算法来做很多很好的事情。

6.3 数据使用安全



6.3 数据使用安全

- 北京大学的学者提出基于生成式对抗网络GAN的恶意软件生成器MalGan，经过训练之后可以生成杀毒软件难以检测的对抗样本。这都是人工智能给你带来坏的方面。
- AI自身数据安全的问题，前面是人工智能可以帮助我们，人工智能可以做恶，这里是人工智能自身的安全问题。主要包括三个方面的安全：一个是训练数据安全，一个是模型参数安全，一个是AI应用安全。

6.3 数据使用安全

- 在训练数据安全方面，训练数据有时候是我们花大价钱得来的，比如说人脸数据以及一些医学应用数据，如人体的MRI，是以非常大的代价得来的。如果这个训练数据被偷走了，成本浪费是一个方面，另一方面涉及到用户隐私泄露。比如人脸数据等。

6.3 数据使用安全

- 还有模型参数安全，花了很大代价训练出来的模型，模型的价值在于参数，如果被你轻易学走拿去了，对企业来说是巨大的损失。
- 还有AI应用安全，在这个图里，对人眼来看这还是一个STOP的标签，但是一些算法认不出，那在智能驾驶来说就是很致命的问题。

6.3 数据使用安全

- 左边的人脸数据通过人脸识别系统，给它一个随机输入，然后进行不停的更新，经过多轮训练之后可以拿到右边的数据，最终可以把右边的数据给拿出来，你看到对于人脸来说，我看到右边这张脸，已经可以看出来他是谁了。

6.3 数据使用安全

- 训练数据安全还有一种，当做合作学习的时候，大家都是上传本地参数，可以保护本地隐私，最新研究发现，在很多情况下仍然有可能造成本地的训练数据泄露。
- 模型参数安全，攻击者可以与AI系统进行交互，这个模型如果是线性模型的话，就把它的参数求到，如果是非线性的话，利用梯度学习重新获得近似的学习，可以得到99%的准确率，跟原先训练好的模型，可以把参数给拿到。
- 如果我已经知道模型参数了，甚至可以把模型的超参数也可以拿到，参数是你训练出来的，但是超参数不是不能从数据里得到，而是人工指定的，人工指定的值则是大量的摸索实践经验得来的。这就表明，参数和超参数都有可能被窃取。

6.3 数据使用安全

- 攻击者在已知模型参数的情况下，可以窃取模型算法的超参数

目标函数 $L(w) = L(X, y, w) + \lambda R(w)$

模型参数　超参数

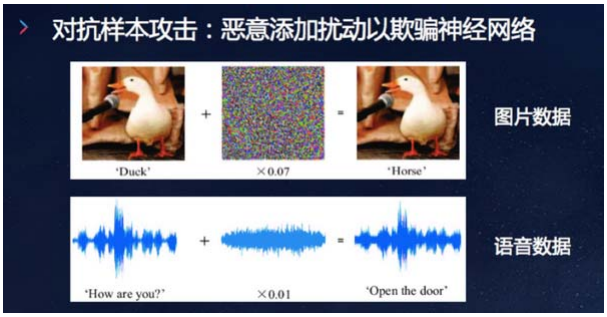
- 学习到的模型参数往往是目标函数的极小值点，梯度为0

$\partial_w L(w) = \partial_w L(X, y, w) + \lambda \partial_w R(w)$　　λ的线性方程

- 利用最小二乘法估计线性方程中的超参数 λ

6.3 数据使用安全

- 下面讲一个AI应用安全中的对抗样本攻击。举例来说，下面的图片数据一开始是一个鸭，加了一点噪音之后就被识别成马。但是对肉眼来说还是一个鸭子，无非就是模糊了一点。对于声音识别来说，左边本来是讲how are you，加了一点噪音就被识别成open the door。



6.3 数据使用安全

- 主要的攻击模式，一个是黑盒模式，攻击者不知道AI系统的算法和参数，但可以与之交互，他能够尽他最大的可能，把训练数据、模型参数、超参数欺骗模型。对于白盒攻击，攻击者能够知道AI系统算法及参数，主要的攻击算法包括快速梯队算法、投射梯队算法等等。

A Taxonomy and Terminology of Adversarial Machine Learning

https://doi.org/10.6028/NIST.IR.8269-draft

6.4 数据去标识化

数据去标识化，关注的待去标识化的数据集是微数据（以记录集合表示的数据集，逻辑上可通过表格形式表示）。**《信息安全技术-个人信息去标识化指南》**

去标识化不仅仅是对数据集中的直接标识符、准标识符进行删除或变换，而且应当结合后期应用场景考虑数据集被重标识的风险，进而选择恰当的去标识化模型和技术措施，并实施合适的效果评估。

6.4 数据去标识化

去标识化目标

1）对直接标识符和准标识符进行删除或变换，避免攻击者根据这些属性直接识别或者结合其它信息识别出原始个人信息主体；

2）控制重标识的风险，根据可获得的数据情况和应用场景选择合适的模型和技术，将重标识的风险控制在可接受范围内，确保重标识风险不会随着新数据发布而增加，确保数据接收方之间的潜在串通不会增加重标识风险；

6.4 数据去标识化

去标识化目标

3）在控制重标识风险的前提下，结合业务目标和数据特性，选择合适的去标识化模型和技术，确保去标识化后的数据集尽量满足其预期目的（有用）。

6.4 数据去标识化

去标识化原则

a) 合规

应满足我国法律法规和标准规范对个人信息安全保护的有关规定，并持续跟进有关法律法规和标准规范。

b) 个人信息安全保护优先

应根据业务目标和安全保护要求，对个人信息进行恰当的去标识化处理，在保护个人信息安全的前提下确保去标识化后的数据具有应用价值。

6.4 数据去标识化

去标识化原则

c) 技术和管理相结合

根据工作目标制定适当的策略，选择适当的模型和技术，综合利用技术和管理两方面措施实现最佳效果。包括设定具体的岗位，明确相应职责；对去标识化过程中形成的辅助信息（比如密钥，映射表等）采取有效的安全防护措施等。

6.4 数据去标识化

去标识化原则

d) 充分应用软件工具

针对大规模数据集的去标识化工作，应考虑使用软件工具提高去标识化效率、保证有效性。

e) 持续改进

在完成去标识化工作后须进行评估和定期重评估，对照工作目标，评估工作效果（包括重标识风险和有用性）与效率，持续改进方法、技术和工具。并就相关工作进行文档记录。

6.4 数据去标识化

去标识化影响

对数据集进行去标识化，**会改变原始数据集**，可能影响数据**有用性**。

业务应用使用去标识化后的数据集时应充分认识到这一点，并考虑数据集变化可能带来的影响。

6.4 数据去标识化

重标识风险

1）重标识方法

- 隔离：基于是否能唯一确定一个信息主体，将属于一个信息主体的记录隔离出来；
- 关联：将不同数据集中关于相同信息主体的信息关联；
- 推断：通过其它属性的值以一定概率推断出一个属性的值。

6.4 数据去标识化

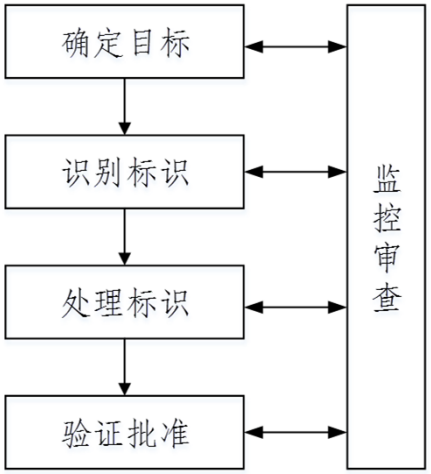
重标识风险

2）重标识攻击

- **重标识**一条记录属于一个特定信息主体；
- **重标识**一条特定记录的信息主体；
- 尽可能多的将记录和其对应的信息主体**关联**；
- **判定**一个特定的信息主体在数据集中是否存在；
- **推断**和一组其它属性关联的敏感属性。

6.4 数据去标识化

去标识化流程



6.4 数据去标识化

识别标识

识别标识符的方法包括查表识别法、规则判定法和人工分析法。

- **查表识别法**指预先建立元数据表格，存储需去标识化的直接标识符和准标识符名称，在识别标识数据时，将待识别数据的各个属性名称或字段名称，逐个与元数据表中记录进行对比，以此识别出标识数据。

6.4 数据去标识化

识别标识

- **查表识别法**指预先建立元数据表格，存储需去标识化的直接标识符和准标识符名称，在识别标识数据时，将待识别数据的各个属性名称或字段名称，逐个与元数据表中记录进行对比，以此识别出标识数据。

查表识别法适用于**数据集格式和属性已经明确**的去标识化场景，如采用关系型数据库，在表结构中已经明确姓名、身份证号等标识符字段。

6.4 数据去标识化

识别标识

但是，查表识别法可能存在无法识别出的标识符情况：

✓ 业务系统存储数据时未采用常用的字段名称，如使用“备注”字段存储身份证号；

✓ 数据中存在混乱或错误情况，如该“备注”字段前100条记录的值为空，而后10000条记录的值为用户身份证号码。

6.4 数据去标识化

识别标识

- **规则判定法**是指通过建立自动化程序，分析数据集规律，从中自动发现需去标识化的直接标识符和准标识符。

规则判定法不仅仅适用于结构化数据应用场景，也适用于某些半结构化和非结构化数据应用场景，如对于非结构化存储的司法判决书，可以通过建立身份证号识别规则和开发程序，从司法判决书中自动识别出所有的身份证号。

6.4 数据去标识化

识别标识

- **人工分析法**是通过人工发现和确定数据集中的直接标识符和准标识符。

在对业务处理、数据集结构、相互依赖关系和对数据集之外可用数据等要素分析的基础上，综合判断数据集重标识风险后，直接指定数据集中需要去标识化的直接标识符和准标识符。

人工分析法在结构化、半结构化和非结构化数据应用场景下都可使用。但考虑到工作量和复杂程度等因素，人工分析法更适用于数据结构简单的数据集。

6.4 数据去标识化

识别标识

人工分析法的优势：

a) 数据集中的多个不同数据子集之间存在关联、引用关系时，如通过数据挖掘算法，可联合分析数据集中多个非常见标识符属性后识别出唯一的用户身份；

b) 数据集中有特别含义的数据，或者数据具有特殊值、容易引起注意的值，从而可能被用来重标识时，如超出常人的身高、独特的地理坐标、罕见的病因等。

相比较于查表识别法和规则判定法，人工分析法能够更加准确地识别出标识符。

6.4 数据去标识化

处理标识

处理标识步骤分为**预处理**、**选择模型技术**、**实施去标识化**三个阶段工作。

- **预处理**

是对数据集正式实施去标识化前的准备过程。一般地，预处理是对数据集施加某种变化，使其有利于后期进行处理。预处理阶段工作可参考如下方法进行：

a) 形成规范化，或满足特定格式要求的数据；

b) 对数据抽样，减小数据集的规模；

c) 增加或扰乱数据，改变数据集的真实性。

6.4 数据去标识化

处理标识

- **选择模型技术**

不同类型的数据需要采用不同的去标识化技术，所以在去标识化的早期阶段，重要的一步是确定数据的类型和业务特性，选择合适的去标识化模型和技术。

- ✓ 是否需要对重标识风险进行量化；
- ✓ 聚合数据是否够用；
- ✓ 数据是否可删除；
- ✓ 是否需要保持唯一性；
- ✓ 是否需要满足可逆性；
- ✓ 是否需要保持原有数据值顺序；
- ✓ 是否需要保持原有数据格式，如数据类型、长度等保持不变；

6.4 数据去标识化

处理标识

- **选择模型技术**

- ✓ 是否需要保持统计特征，如平均值、总和值、最大值、最小值等；
- ✓ 是否需要保持关系型数据库中的实体完整性、参照完整性或用户自定义完整性；
- ✓ 是否可以更改数据类型，比如在针对字符串类型的“性别”（男/女）进行去标识化时，是否可以变成数字类型表示（1/0）；
- ✓ 是否需要满足至少若干个属性值相同，以加强数据的不可区分性；
- ✓ 是否可以对属性值实施随机噪声添加，对属性值做小变化；
- ✓ 去标识化的成本约束。

6.4 数据去标识化

处理标识

- **实施去标识化**

根据选择的去标识化模型和技术，对数据集实施操作。

a) 若存在多个需要去标识化的标识符，则根据数据特点和业务特性设定去标识化的顺序；

b) 依次选择相应的工具或程序；

c) 设置工具或程序的属性和参数，如设置数据源、用户名/口令、算法参数等；

d) 依次执行去标识化工具或程序，获得结果数据集。

6.4 数据去标识化

验证批准

对数据集去标识化后进行验证，以**确保生成的数据集在重标识风险和数据有用性方面都符合预设的目标**。

在验证满足目标过程中，需对去标识化后重标识风险进行评估，计算出实际风险，与预期可接受风险阈值进行比较，若风险超出阈值，需继续进行调整直到满足要求。

由于重标识技术和重标识攻击的能力在迅速演变，需要由内部专业人员或权威的外部组织**定期展开验证评估**。

6.4 数据去标识化

◘ 验证批准

验证去标识化有效性的方法包括：

✓ 检查生成的数据文件，以确保文件数据或元数据中不包含直接标识符和准标识符；

✓ 检查生成的数据文件，以确保所得数据符合既定目标要求；

✓ 评估去标识化软件所使用的默认假设；

✓ 进行有动机的入侵者测试，看看是否有具备合格能力的外部人员可以使用公开的数据集执行重标识；

✓ 让团队利用内部数据进行有针对性的入侵者测试，模拟违规者或敌对内幕人士可能发生的情况。

6.4 数据去标识化

◘ 验证批准

验证数据有用性：

◘ 去标识化降低了数据质量和生成数据集的有用性。因此，需要考虑去标识化后的数据集对于预期的应用仍然有用。

◘ 存在一些方法用于验证数据有用性。例如，内部人员可对原始数据集和去标识化的数据集执行统计计算，并对结果进行比较，以查看去标识化后是否导致不可接受的更改。组织可让可信的外部人员检查去标识化数据集，以确定数据能被用于预期目的。

6.4 数据去标识化

◘ 验证批准

评审批准去标识化工作：

- ◘ 在完成处理标识和验证结果后，组织管理层应依据数据发布共享用途、重标识风险、数据有用性最低要求等因素，以及验证结果、去标识化各步骤实施过程中的监控审查记录等因素，做出是否认可数据去标识化结果的决定。
- ◘ 批准由组织最高管理层或更高层的主管部门来执行。

6.4 数据去标识化 -常用去标识化技术

◘ 统计技术

统计技术是一种对数据集进行去标识化或提升去标识化技术有效性的常用方法，主要包含**数据抽样**和**数据聚合**两种技术。

- **数据抽样（Sampling）**

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的，它是提升去标识化技术有效性的重要方法。

应注意以下几个方面：

a) 从数据集中抽取样本的方法很多，各方法差异很大，需根据数据集的特点和预期的使用场景来选择。

6.4 数据去标识化 -常用去标识化技术

统计技术

- **数据抽样（Sampling）**

b) 数据抽样经常用于去标识化的预处理，对数据集进行**随机抽样能够增加识别出特定个人信息主体的不确定性**，从而可以提高后续应用的其它去标识化技术的有效性。

c) 数据抽样可以简化对数据集的计算量，因此，在对大样本的数据集进行去标识化时，**首先进行抽样，然后再采用某项特定的技术进行去标识化**。

6.4 数据去标识化 -常用去标识化技术

统计技术

- **数据抽样（Sampling）**

例如，攻击者想通过将样本某一记录的属性与外部信息相匹配而识别出特定主体，在采用抽样的情况下，数据主体是否存在于样本数据集还不能确定，所以无法确定该记录是否与特定数据主体相对应。

6.4 数据去标识化 -常用去标识化技术

统计技术

- **数据抽样（Sampling）**

例子：某市从1000万市民中随机抽取1万人的信息进行去标识化工作，采用屏蔽技术，只保留4项信息（性别、学历、籍贯、身高）。如果攻击者通过外部其它信息关联而确定样本中某人的出生日期，得到如下记录：

记录甲：（男，本科，北京，1.75米，1980年9月1日）

如果市民A的情况完全符合记录甲，攻击者并不能确定记录甲就是指市民A，因为A并不一定在此抽样数据集中。

6.4 数据去标识化 -常用去标识化技术

统计技术

- **数据聚合（Aggregation）**

作为一系列统计技术（如求和、计数、平均、最大值与最小值）的集合，应用于微数据中的属性时，产生的结果能够代表原始数据集中的所有记录。

应注意以下几个方面：

a) 数据聚合**可能会降低数据的有用性**；因为得到的是统计值，无法反映独立数据记录的特征。

6.4 数据去标识化 -常用去标识化技术

统计技术

- **数据聚合（Aggregation）**

b) 数据聚合对重标识攻击非常有效；数据聚合的输出是“统计值”，该值有利于对数据进行整体报告或分析，而不会披露任何个体记录。

例如：2012年我国18岁及以上成年男性平均身高1.67米。如果数据集以平均身高来标识数据集中每个人的身高值，则记录（男，本科，北京，1.67米，1980年9月1日）中，身高属性值对攻击者识别身份主体没有什么作用。。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **确定性加密（Deterministic encryption）**

确定性加密是一种非随机对称加密。在去标识化过程中应用时，确定性加密用加密结果替代微数据中的标识符值。

应注意以下几个方面：

a) 确定性加密可以保证数据真实可用，即相同的两个数据用同一密钥进行加密将产生两个一样的密文。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **确定性加密（Deterministic encryption）**

b) 确定性加密可以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，确定性加密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局限于检查数据值是否相等。

c) 对确定性加密的重标识攻击主要在于不具备密钥使用权时的攻击；关联性攻击则可能适用于采用同一密钥进行确定性加密的密文，攻击的成功与否很大程度上取决于对加密算法参数的选。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **保序加密（Order-preserving encryption）**

保序加密是一种非随机对称加密。用作去标识化技术时，保序加密可用加密值替代微数据中的标识符值。

应注意以下几个方面：

a) 密文的排序与明文的排序相同。

b) 保序加密可以在有限的范围内保证加密结果在统计处理、隐私防挖掘、数据外包存储与处理等场景中的有用性。保序加密可以产生用于范围/区间匹配搜索、分析的微数据。对保序加密结果的分析局限于检查数据相等和排序比较关系。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **保序加密（Order-preserving encryption）**

c) 保序加密数据的完全重标识仅可能适用于拥有密钥的一方。关联性攻击能否成功很大程度上取决于保序加密方案的参数选择。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **保留格式加密（Format-preserving encryption）**

保留格式加密是一种适宜于去标识化技术的加密方法，加密要求密文与明文具有相同的格式，当作为去标识化技术的一部分加以采用时，保留格式加密可用加密值替代标识符值。

应注意以下几个方面：

a) 某些保留格式加密具有确定性加密技术一样的特点，如相同数据在同一密钥下加密生成同样的密文，且可以通过生成微数据进行精准匹配搜索、数据关联分析等。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **保留格式加密（Format-preserving encryption）**

b) 保留格式加密适用于多种格式的数据，包括字符型、数字型、二进制等，加密结果也是同类型数据。

c) 和其它加密技术不一样，在给定有限符号集的情况下，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，这有助于在不需要应用修改的情况下，实现去标识化。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **同态加密（Homomorphic encryption）**

同态加密是一种随机加密。当作为去标识化技术的一部分加以采用时，对加密数据进行处理，但是处理过程不会泄露任何原始内容。同时，拥有密钥的用户对处理过的数据进行解密后，得到的正好是处理后的结果。同态加密用加密值替代微数据中的标识符值。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **同态加密（Homomorphic encryption）**

a) 对经过同态加密的数据进行处理得到一个输出，将这一输出进行解密，其结果与用同一方法处理未加密的原始数据得到的输出结果是一样的。

b) 与传统的确定性加密方案相比，同态加密的性能一般较低，存储成本较高。

c) 同态加密方案具有语义上的安全性，使得在不具备访问私钥权限时无法实现重标识攻击。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **同态秘密共享（Homomorphic secret sharing）**

同态秘密共享可将一个秘密拆分为“若干份额”，可利用拆分后秘密信息的特定子集来重构原始的秘密，如果对用于重构秘密的所有份额执行相同的数学运算，则其结果等价于在原始秘密上执行相应数学运算的结果。

当作为去标识化技术的一部分加以采用时，同态秘密共享可用信息共享算法得出的两个或以上若干份额替代数据记录中的任何标识符或敏感属性。这样，便可将这些若干份额分配给两个或以上的份额持有者。这些份额持有者的数量通过秘密共享方案加以确定。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **同态秘密共享（Homomorphic secret sharing）**

有效的同态秘密共享的特性是，相同份额持有者共享机密的两个值可与加密方案的同态运算相结合，产生代表原始属性运算结果的新份额。此外，同态密钥共享可与安全的多方计算相结合，以便对去标识化数据进行任何安全运算。同态密钥共享并不会降低数据的真实性。

虽然同态密钥共享有着相对低的计算性能开销，但存在与份额持有者之间交换份额的额外开销。

6.4 数据去标识化 -常用去标识化技术

密码技术

- **同态秘密共享（Homomorphic secret sharing）**

共享秘密数据的存储开销是有限的。基于安全多方计算执行的数据去标识化的处理技术是灵活的，但根据所采用的不同方案，可能会导致高昂的成本。

同态密钥共享会产生微数据的分布式实例，该类实例可被同态运算或安全多方计算技术处理。同态加密方案是随机的，攻击者只有控制所有份额持有者才能实现重标识攻击。

6.4 数据去标识化 -常用去标识化技术

◘ 抑制技术

✓ 抑制技术即对不满足隐私保护的数据项删除，不进行发布。包括从所有记录中选定的属性（如屏蔽）、对所选定的属性值（例如，局部抑制）、或是从数据集中选定的记录（例如，记录抑制）进行的删除操作。抑制技术主要适用于分类数据。

✓ 抑制技术可用于防止基于关联规则推导的攻击，因为不发布能最大化降低关联规则支持度和置信度的属性值，从而破坏关联规则推导攻击。

✓ 抑制技术适用于数值与非数值数据属性，执行相对比较容易，且可以保持数据的真实性。

6.4 数据去标识化 -常用去标识化技术

◘ 抑制技术

✓ 抑制技术会导致信息丢失，抑制技术处理后的数据有被重标识的风险，因此需要与其它去标识化技术相结合以降低数据的重标识风险。

✓ 过多的抑制会影响数据的效用，所以在具体应用时，为保证数据的可用性，要对抑制的数据项数量设定一个上限值。

6.4 数据去标识化 -常用去标识化技术

抑制技术

- **屏蔽（Masking）**

屏蔽技术是一种基线去标识化技术，该技术包括从数据集中删除所有直接标识符，并尽可能剥离数据集中所有记录的部分或全部剩余标识符。删除直接标识符的一部分，使其既不是直接标识符也不是唯一标识符，也是一种屏蔽技术。

使用屏蔽技术后，通常还会对数据集使用其它去标识化技术。

在将屏蔽技术作为唯一的去标识化技术的系统中，应采取安全措施和组织其它的管理措施去保护未被识别的数据。

6.4 数据去标识化 -常用去标识化技术

抑制技术

- **屏蔽（Masking）**

屏蔽技术也有其它一些叫法，如：

a)**部分数据移除**：指在屏蔽过程中不会删除所有标识符。

b)**数据隔离**：指屏蔽需要有严格的安全措施，以确保对数据集的授权访问，如访问控制和相应的合约条款

c)**数据限制**：指在有特定目的的环境中收集数据时进行数据抑制的情况。

6.4 数据去标识化 -常用去标识化技术

抑制技术

- **局部抑制（Local suppression）**

局部抑制技术是一种去标识化技术，即从所选记录中删除特定属性值，该特定属性值与其它标识符结合使用可能识别出相关个人信息主体。通常应用局部抑制技术来移除准标识符在泛化后仍然出现的稀有值（或这些值的稀有组合）。

局部抑制技术应用于分类值，而泛化通常应用于数值，其共同目标是增加共享其标识符值的记录数。

6.4 数据去标识化 -常用去标识化技术

抑制技术

- **记录抑制（Record suppression）**

“记录抑制”是一种从数据集中删除整个记录或一些记录的去标识化技术。典型应用场景为删除包含稀有属性（如异常值）组合的记录。

6.4 数据去标识化 -常用去标识化技术

◘ 抑制技术

抑制技术--示例::

◘ 在某个具体应用中，需要对**不同工作年限的薪资水平**进行分析，原始数据集包括{姓名，性别，薪水，工作年限，职务}，采用如下步骤进行去标识化：

- ✓ **姓名**是直接标识符，需要应用抑制技术删除；通过{职务、工作年限}或者{职务、性别}也可以推导出该组织中的一部分员工，因此应用抑制技术删除职务属性；
- ✓ 剩下的{性别，薪水，工作年限}，有被重标识的风险，需要结合泛化技术，对“薪水”、“工作年限”属性值进行泛化处理，如薪水泛化为5k-10k、10k-15k、15k-20k等，工作年限泛化为0-3年、4-6年等；
- ✓ 如果数据记录中只有1人工作年限为0-3年，薪水为15k-20k，则能够定位到某个员工，应用抑制技术删除该条记录。

6.4 数据去标识化 -常用去标识化技术

◘ 假名化技术

- ● 假名化技术是一种使用假名替换直接标识（或其它敏感标识符）的去标识化技术。假名化技术为每一个人信息主体创建唯一的标识符，以取代原来的直接标识或敏感标识符。不同数据集中的相关记录在进行假名化处理后依然可以进行关联，并且不会泄露个人信息主体的身份。
- ● 在使用假名化技术的过程中，通常会使用一些辅助信息。这些辅助信息包括从原始数据集中删除的标识符、假名分配表或密钥等，采取必要的措施来保护这些辅助信息有利于降低重标识风险。

6.4 数据去标识化 -常用去标识化技术

假名化技术

- 假名创建技术主要包括独立于标识符的假名创建技术和基于密码技术的标识符派生假名创建技术。假名创建技术的选择需要考虑以下因素：创建假名的成本、散列函数的抗碰撞能力、以及重标识过程中假名被还原的手段。
- 独立于标识符的假名创建技术不依赖于被替代的属性原始值，而是独立生成，典型方法为用随机值代替属性原始值。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，这一过程也称为对数据集中的属性进行“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，而散列技术是一种单向的数学运算。

6.4 数据去标识化 -常用去标识化技术

假名化技术

- **独立于标识符的假名创建**

独立于标识符的假名创建技术不依赖于被替代的属性原始值，而是独立生成，典型方法为用随机值代替属性原始值。

使用该类技术时需创建假名与原始标识的分配表。根据去标识化的目标，应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。

6.4 数据去标识化 -常用去标识化技术

假名化技术

- **基于密码技术的标识符派生假名创建**

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，这一过程也称为对数据集中的属性进行“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，而散列技术是一种单向的数学运算。

6.4 数据去标识化 -常用去标识化技术

假名化技术

- **基于密码技术的标识符派生假名创建**

采用多种密码技术的组合可更好地保护属性原始值。

采用加密方法来创建假名的计算成本很高，但非常有效。应采取特殊措施来保护密钥，防止密钥被未授权访问，包括密钥与数据分离，不与第三方共享密钥，安全地删除密钥以防重标识等。

6.4 数据去标识化 -常用去标识化技术

假名化技术

- **基于密码技术的标识符派生假名创建**

散列函数的单向运算及抗碰撞能力等特性，使其适用于假名化过程。但是，当散列算法和所用密钥是已知的，且有可能遍历散列函数生成数值空间时，散列函数是可逆的。因此使用密钥散列函数时可增加另一随机输入，增强其对抗暴力搜索攻击的能力，防止未经授权的重标识。即使采用了安全的散列技术，如果在使用或执行散列算法中发生了疏忽，或未经授权共享密钥，均可能导致数据的重标识。

6.4 数据去标识化 -常用去标识化技术

假名化技术

示例：：

在某个具体的应用中，需要从外部某数据库中抽取包含人名的有效数据以供分析，采用如下步骤进行去标识化：

a)构建常用人名字典表

常用人名字典表有200个常用人名构成：龚小虹、黄益洪、龙家锐、龚尧莞、齐新燕、车少飞、龙家铸、赖鸿华、龙宣霖、连丽英……

6.4 数据去标识化 -常用去标识化技术

假名化技术

b)制定人名与假名的分配技术。

分配技术采用纯随机方式，对于每一个标识符（人名），随机生成一个不小于1并且不大于200的随机数，从字典表中的对应位置获取假名，进行替换。

c)使用字典表和分配技术，完成对人名的去标识化。

在去标识过程中，在遇到人名“辛培军”时，随机生成了数5，则使用字典中的排列第5的名字“齐新燕”替换“辛培军”。

该示例使用随机方式构建分配规则，采用了多对一的方式，在保留适当可用性的同时，降低了数据的重标识风险。

6.4 数据去标识化 -常用去标识化技术

泛化技术

泛化技术是指一种降低数据集中所选属性粒度的去标识化技术，对数据进行更概括、抽象的描述。泛化技术实现简单，能保护记录级数据的真实性。

使用泛化技术的目标是减少属性唯一值（更概括地说，是指多个属性值的组合集的唯一值）的数量，使得被泛化后的值（或多个值的集合）被数据集中多个记录所共享，从而增加某特定个人信息主体被推测出的难度。因此，通常选择对标识符属性进行泛化，但是根据具体情况也可考虑对任何属性（特别是敏感属性）进行泛化。

6.4 数据去标识化 -常用去标识化技术

泛化技术

- **取整（Rounding）**

取整涉及到为所选的属性选定一个取整基数，然后将每个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。

受控取整也是可行的，例如确保取整值的求和结果与原始数据的求和取整值相同。

6.4 数据去标识化 -常用去标识化技术

泛化技术

- **顶层与底层编码（Top and bottom coding）**

泛化技术为某一属性设定一个可能的最大（或最小）阈值。顶层与底层编码技术使用表示顶层（或底层）的阈值替换高于（或低于）该阈值的值。

该技术适用于连续或分类有序的属性。例如，如果一个人的薪水非常高，则可将该用户的薪水值设置为“高于X元”，其中“X”为高收入值的界限，而不记录或报告准确的金额。

6.4 数据去标识化 -常用去标识化技术

□ 随机化技术（Randomization techniques）

随机化技术作为一种去标识化技术类别，指通过随机化修改属性的值，使得随机化处理后的值区别于原来的真实值。该过程降低了攻击者从同一数据记录中根据其它属性值推导出某一属性值的能力。

随机化技术并不能保证数据在记录集的真实性。为达到特定的目标，有效随机化过程需要逐项定制，定制过程中需要详细了解数据特性，并选取合适的参数。

6.4 数据去标识化 -常用去标识化技术

□ 随机化技术（Randomization techniques）

- **噪声添加（Noise addition）**

噪声添加是一种随机化技术，通过添加随机值、“随机噪声”到所选的连续属性值中来修改数据集，同时尽可能保持该属性在数据集中的原始统计特性。该类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。

6.4 数据去标识化 -常用去标识化技术

□ 随机化技术（Randomization techniques）

- **置换（Permutation）**

置换是在不修改属性值的情况下对数据集记录中所选属性的值进行重新排序的一种技术。因此，置换保持了整个数据集中所选属性的准确统计分布。

置换技术适用于数字与非数字值。因为观察到的不一致性可能有助于对置换算法实施逆向工程，需要考虑如何来确保生成的数据集是一致与真实的。

6.4 数据去标识化 -常用去标识化技术

□ 随机化技术（Randomization techniques）

- **置换（Permutation）**

不同置换技术的区别在于方法与复杂性的差别。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。

通常情况下，采用逆向工程可以将数据恢复到原始状态，从而加大受控重标识的可能性，因此把随机化算法引入到置换中会增强对抗重标识攻击的能力。

6.4 数据去标识化 -常用去标识化技术

□ 随机化技术（Randomization techniques）

- **微聚集（Microaggregation）**

“微聚集”是指用某种算法方式计算出来的平均值代替连续属性所有值的去标识化技术。对于每种连续属性，或对于所选的一组连续属性，数据集中的所有记录都进行了分组，具有最近属性值的记录属于同一组，而且每一组中至少有k个记录。每一种属性的新值替换为该属性所在组中的平均值。每组中的各个值越接近，数据的有效性就保持得越好。

微聚集的输出是微数据，该技术不能保证数据的真实性。

微聚集技术的不同之处在于：选择的属性、属性值之间的相似性计算方式以及其它考虑因素。

6.4 数据去标识化 -常用去标识化技术

□ 数据合成技术（Synthetic data）

数据合成是一种以人工方式产生微数据的方法，用以表示预定义的统计数据模型。

应注意以下几个方面：

a) 合成数据集与原始数据特性相符，但不包含现有个人信息主体有关的任何数据，但是，若合成后的数据与原始数据的拟合度过高可能会导致敏感信息泄露。

6.4 数据去标识化 -常用去标识化技术

□ 数据合成技术（Synthetic data）

b)创建合成数据的方法很多。理论上，数据可根据所选的统计特性随机生成。该类模型的关键特征主要体现在每种属性（总体与子总体）的分布以及属性之间的内部关系。实际上，合成数据的生成会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。合成数据通常用于测试工具与应用。

c)合成数据可用于开发查询。合成数据可用作真实数据的替代项：数据管理者能在实际数据中重现在合成数据中执行的查询，以确保基于合成数据的处理能够同样正确应用于真实数据。

利用差分隐私机制可以保证合成数据的隐私。

聚集”是指用

6.4 数据去标识化 -常用去标识化技术

□ 数据合成技术（Synthetic data）

b)创建合成数据的方法很多。理论上，数据可根据所选的统计特性随机生成。该类模型的关键特征主要体现在每种属性（总体与子总体）的分布以及属性之间的内部关系。实际上，合成数据的生成会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。合成数据通常用于测试工具与应用。

c)合成数据可用于开发查询。合成数据可用作真实数据的替代项：数据管理者能在实际数据中重现在合成数据中执行的查询，以确保基于合成数据的处理能够同样正确应用于真实数据。

利用差分隐私机制可以保证合成数据的隐私。

聚集”是指用

数据安全

福州大学 FUZHOU UNIVERSITY

6.4 数据去标识化 -常用去标识化技术

K-匿名模型（K-anonymity model）

K-匿名模型是在发布数据时保护个人信息安全的一种模型。K-匿名模型要求发布的数据中，指定标识符（直接标识符或准标识符）属性值相同的每一等价类至少包含K个记录，使攻击者不能判别出个人信息所属的具体个体，从而保护了个人信息安全。在使用K-匿名模型整合得到的数据集中，各记录之间的关联性是有限的（1/K）。

可独立或综合使用附录A中的各种去标识化技术，以符合K-匿名模型的要求。抑制技术、泛化技术及微聚集均适用于数据集中的各种属性，以实现期望的结果。

6.4 数据去标识化 -常用去标识化技术

K-匿名模型（K-anonymity model）

K-匿名模型还包括一些增强概念，如L-多样性和T-接近性。

✓ **L-多样性（L-diversity）**

L-多样性是针对属性值差异性不大的数据集提出的一种增强概念。为防止确定性推导，L-多样性要求在K-匿名的基础上，实现每一等价类在每一敏感属性上存在至少L个不同值。

6.4 数据去标识化 -常用去标识化技术

K-匿名模型（K-anonymity model）

✓ **T-接近性（T-closeness）**

T-接近性是L-多样性的增强概念，适用于发布数据集的敏感属性分布要尽可能贴近整个数据集的敏感属性分布。针对属性值分布不规则、属性值范围很小或者已被分类的数据集，为防止概率性推导，要求任何等价类中敏感属性的分布与整个数据集中相应属性的分布之间的距离小于阈值T。

6.4 数据去标识化 -常用去标识化技术

K-匿名模型（K-anonymity model）

数据集的重标识风险度量包括如下两个关键因素：

➢ **每行记录重标识概率的计算方法**

数据集中的每一行都包含有关个体的信息，存在重标识的概率。对于给定的行，重标识的概率取决于数据集中其它行对于准标识符的属性是否具有相同的值。

数据集中的“等价类”是指具有与准标识符属性相同值的数据记录行。例如，在具有性别，年龄和最高教育水平的属性列的数据集中，所有满足“35岁以上且具有大专学位的老年男子”的数据记录，形成一个等价类。等价类的大小等于准标识符具有相同值的行数。

6.4 数据去标识化 -常用去标识化技术

K-匿名模型（K-anonymity model）

对于每一行，重标识的概率等于1除以其等价类的大小，即，给定记录行重标识概率=1/等价类大小。例如，大小为5的等价类中的每一行都有重标识的概率为0.2。因此，具有较大等价类的行，具有较低的重标识概率。

- **根据所使用的发布模型采用适当的风险衡量方法**

虽然每行记录重标识的概率等于1除以其等价类的大小，但是具体的计算数据集中重标识风险的方法，取决于具体使用的发布模型。

6.4 数据去标识化 -常用去标识化技术

K-匿名模型（K-anonymity model）

公开共享数据发布应使用最大风险。因为对于公开数据发布，应假设有攻击者会进行炫耀式攻击。该类攻击将针对数据集中最脆弱的行，即具有最小等价类和最高重标识概率的数据行。因此，应使用所有行中重标识的最大概率来衡量重标识风险。

受控共享数据发布应使用严格的平均风险。受控共享数据发布数据集的访问仅限于选定数量的已鉴别信息接收方，每行数据的重标识概率是均等的，应使用所有行中重标识的平均概率来衡量数据集中重标识风险。为了保护具有高度重标识风险的独特行或等价类，平均值通常建议为0.33，即数据集中等价类的最小尺寸应为3。实际使用时重标识的最大概率也可以定为0.5。

6.4 数据去标识化 -常用去标识化技术

隐私和安全控制水平	动机和能力	重标识攻击概率
高	低	0.05
	中	0.1
	高	0.2
中	低	0.2
	中	0.3
	高	0.4
低	低	0.4
	中	0.5
	高	0.6

6.4 数据去标识化 -常用去标识化技术

- 差分隐私

差分隐私是对数据发布时数据集中的隐私损失进行量化的数学模型。差分隐私确保数据集中任何特定的个人信息主体的存在与否无法从去标识化数据集或系统响应中推导出。即使攻击者能够访问其它相关的数据集，只要隐私损失限定在某一水平，这些保证就会得到保持。

6.4 数据去标识化 -常用去标识化技术

◘ 差分隐私

差分隐私提供：

✓ 隐私数学定义，该定义假定，无论任何个人信息主体添加到数据集或者从数据集中删除，分析结果有近似相同的统计分布。

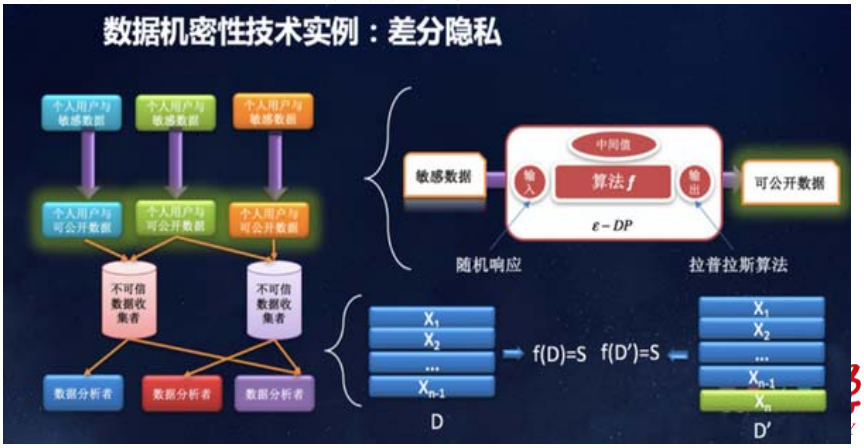
✓ 隐私度量方法，可以监控累积的隐私损失并设置损失限制的“预算”。

差分隐私算法在数据集中增加了一定数量的“随机噪声”，该噪声通过精心选择的概率分布产生。随机噪声既可在采集点（本地模式）添加至每一个人信息主体信息的输入中，也可以添加至差分隐私系统向分析者（服务器模式）提供的输出中。

6.4 数据去标识化 -常用去标识化技术

◘ 差分隐私保护

差分隐私，目前谷歌和iOS都开始使用这一技术。收集整体数据的时候会暴露用户的所有隐私，那是否可以先将个人数据做一些扰动，对敏感数据做随机响应。



6.4 数据去标识化 -常用去标识化技术

- 差分隐私保护

为什么差分隐私保护是很好的工具?它给予了一个定量的方法去衡量你把隐私能够保护得比较好，对任意两个只相差一条数据记录的数据库，这两个数据库在面对相同查询返回结果的时候，差距不会太大。这是非常好的一件事情，可以有一个定量的方法保护隐私强度。

6.4 数据去标识化 -常用去标识化技术

- 差分隐私保护
- 服务器模式（Server model）

差分隐私“服务器模式”通常会将数据以原始值保存在安全的数据库中。为了保护隐私，对查询的响应仅能从软件组件获得。

软件组件会接受系统用户或报表软件的查询，并从数据库获得正确的无噪声回答。但是，在对用户或报表软件做出响应前，软件组件会添加一定量的随机噪声，且该噪声与查询所对应的隐私损失成比例。

6.4 数据去标识化 -常用去标识化技术

- 差分隐私保护
- **服务器模式（Server model）**

软件组件负责持续记录累积的隐私损失并确保该损失不超出隐私预算。一旦隐私预算耗尽，软件组件应针对系统建立逐项定义的策略来确定是停止响应查询，还是采取其它措施。

6.4 数据去标识化 -常用去标识化技术

- 差分隐私保护
- ✓ **本地模式（Local model）**

本地模式适用于执行数据采集的实体不受个人信息主体信任，或采集数据的实体正寻求降低风险并执行数据最小化的情形。在该模型中，首先对属于单个个人信息主体的数据或数据的计算结果进行随机化，以便对数据进行去标识化，然后才将其转移至并存储在服务器中。

特定概率分布生成一个随机量，并添加到每一单独的数据或从属于个人信息主体的数据测量的结果中，以便在采集点对数据进行随机化。

6.4 数据去标识化 -常用去标识化技术

- 差分隐私保护

✓ **本地模式（Local model）**

当源自大量设备的随机化数据聚合并用于采集点的统计分析时，分析结果会紧密与总体的集体行为相关。由于噪声在传输前被添加，因此在很多实例中，源自主体的数据报告会存储在服务器中，无需采取其它隐私保护措施，而且产生的数据库可直接共享并进行查询，无需管理者参与。

6.4 数据去标识化 -常用去标识化技术

- 差分隐私

差分隐私系统的关键考虑因素:

- **概率分布**

在差分隐私的环境下，随机噪声采取随机数的形式，随机数根据所选的概率分布生成。可选的概率分布包括零均值的高斯分布、拉普拉斯分布或指数概率分布。

决定噪声生成器产生噪声高低的参数是标准差，与S/ε成正比，其中S表示给定查询的敏感性，而ε则表示相关的隐私预算。

6.4 数据去标识化 -常用去标识化技术

- 差分隐私
  - **敏感度**

给定查询或函数的敏感度S描述了将一个个人信息主体从数据库中删除时该查询或函数的返回结果最多会改变多少的情况。

为了“隐藏”带来最大变化的个人信息主体，需要将一定比例的噪声添加至该特殊查询或函数的所有返回结果中。

6.4 数据去标识化 -常用去标识化技术

- 差分隐私
  - **隐私预算**

隐私预算ε是差分隐私系统设计的一个选项。

由于噪声的标准差与S/ε成正比，则ε越大，标准差越小，隐私预算开销越小，但通常也会带来较大的隐私风险。

较小的ε会增加标准差，从而增加了较大噪声值添加至实际结果中的概率，因此提供了更大程度的隐私保护。

6.4 数据去标识化 -常用去标识化技术

- 差分隐私
  - 累积隐私损失

差分隐私算法对其应答的每次查询会产生隐私成本或隐私损失。在精心设计的差分隐私算法中，这些损失可以足够小，不使隐私受到侵犯，但这些损失的累积效应最终会导致对隐私的侵犯。

为了计算隐私预算中发生的变化，需对从多次查询中累积损失的概念进行规定。比如在差分隐私算法中出现了含有相似隐私成本C的n次查询，则总体隐私预算开销将不高于nC。

隐私预算耗尽并不意味着对隐私一定有侵犯，而只是表明数学保证的失效。一旦保证失效，攻击者就会利用算法输出并运用推导、关联及其它类型的重标识技术实施攻击，可能会导致重标识攻击的成功实施。

6.4 数据去标识化 -常用去标识化技术

- 姓名的去标识化

姓名是一种常用的标识符，适用的去标识化方法举例如下：

a)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”。这种方法是用在需要保留“姓”这一基本特征的应用场景；

b)抑制屏蔽。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；

c)随机替代。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”；

6.4 数据去标识化 -常用去标识化技术

- 姓名的去标识化

d)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”。这种方法有可能用在需要保持姓名数据可逆变换的场景；

e)可逆编码。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”。

6.4 数据去标识化 -常用去标识化技术

- 身份证号的去标识化

身份证号也是一种常用的标识符，国内身份证号按照GB 11643—1999《公民身份号码》制定的规则进行编码，其结构分为地址码、出生日期码、顺序码和校验码，常见的去标识化方法举例如下：

a)抑制屏蔽。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；

b)部分屏蔽。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”或“******188*********”代替，上述数据可分别用在需要保密出生日期、保密出生日期但允许对数据按时代作统计分析、保密所有信息但允许对出生日期按时代作统计分析等场景

6.4 数据去标识化 -常用去标识化技术

- 身份证号的去标识化

c)可逆编码。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”；

d)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。

6.4 数据去标识化 -常用去标识化技术

- 银行卡号的去标识化

银行卡号在很多应用中和个人身份密切关联，是一种常用的标识符。银行卡号是按照规则进行编码的，其结构分为发卡行标识代码、自定义位和校验码。常见的去标识化方法举例如下：

a) 抑制屏蔽。直接删除银行卡号或使用统一的“*”来表示。如所有的银行卡号都使用“*****”代替；

b)部分屏蔽。屏蔽银行卡号中的一部分，以保护卡号信息。如分别可以屏蔽银行卡号中的发卡行标识代码和自定义位；

6.4 数据去标识化 -常用去标识化技术

- 银行卡号的去标识化

c)可逆编码。采用密码或其他变换技术，将银行卡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。这种方法适用于使用银行卡号做数据库主键的应用场景；

d)数据合成。采用重新产生的数据替代原银行卡号，如使用随机产生符合身份证号编码规则的新银行卡号代替原始值，这种场景适应于对银行卡号做合法性校验的应用场景。

6.4 数据去标识化 -常用去标识化技术

- 地址的去标识化

a)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；

b)抑制屏蔽。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；

c)部分屏蔽。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”；

d)数据合成。采用重新产生的数据替代原地址数据，数据产生方法可以采用确定性方法或随机性方法。如使用“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”。

6.4 数据去标识化 -常用去标识化技术

- 电话号码的去标识化

a)抑制屏蔽。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“000000”代替；

b)部分屏蔽。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”或“*******8888”代替；

c)随机替代。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”；

d)可逆编码。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“19888888888”。

6.4 数据去标识化 -常用去标识化技术

- 数值型标识符的去标识化

a) 泛化编码。使用概括、抽象的符号来表示，如“有四个人，他们分别是蓝色、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”；

b) 抑制屏蔽。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；

c) 顶层和底层编码。大于或者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；

6.4 数据去标识化 -常用去标识化技术

- 数值型标识符的去标识化

d) 部分屏蔽。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；

e) 记录交换。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；

f) 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1.72米，产生随数值-0.11米，加到原始数值后将其变为1.61米；

g) 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。

6.4 数据去标识化 -常用去标识化技术

- 日期的去标识化

在数据集中，日期有多种存在形式，包括出生日期、开始日期、纪念日等。常见的对日期的去标识化方法包括：

a)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；

b)抑制屏蔽。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“某年某日”代替；

c)部分屏蔽。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；

6.4 数据去标识化 -常用去标识化技术

- 日期的去标识化

d)记录交换。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；

e)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15日；

f)数据合成。采用重新产生的数据替代元日期数据，如使用“1972年8月12日”代替“1880年1月1日”岁年龄。

6.4 数据去标识化 -常用去标识化技术

- 地理位置的标识化

地理数据在数据集中的表现形式多种多样。地理位置可以通过地图坐标推断出来（例如，39.1351966，-77.2164013），可以通过街道地址（例如清华园10号）或者邮编（20899）。地理位置也可能隐藏在文本数据中。

一些地理位置是不可标识的（例如，一个拥挤的火车站），而另一些是高度可标识的（例如，一个单身汉居住的房子）。单独的地址可能并不可标识，但是如果将它们表示的位置与个人相关联则会成为可标识的信息。

6.4 数据去标识化 -常用去标识化技术

- 地理位置的标识化

对地理位置信息进行去标识化，采用的噪声值很大程度上取决于外界因素。例如在中心区范围内通过加减100m的范围，而偏远地区通过加减5km来得到充足的模糊化结果。

添加噪声时也要考虑噪声对数据真实性的影响。例如，将一个居民的沿海住所搬迁到内陆甚至跨政治领域范畴的另一个国家，这种方式有时是不可取的。

在一个个体的位置信息被持续记录的情况下，对于地理数据信息的去标识化将会变得尤其有挑战性。这是因为事件地点的特征记录就像是人的指纹一样，有利于重标识，即使是很少量的数据记录也能达到这样的效果。

6.4 数据去标识化 -常用去标识化技术

- 去标识化模型的应用

如果针对重标识风险的量化保证纳入了组织的目标中，则可执行合适的去标识化模型。

对于微数据，K-匿名是提供针对重标识风险的量化保证的一种方法。可利用不同的去标识化技术执行K-匿名。因此，去标识化数据的有效性将由模型中所含的特定去标识化技术决定。例如，如果去标识化数据需要在记录级保持真实性，则随机化技术无法用来实现K-匿名。

6.4 数据去标识化 -常用去标识化技术

- 去标识化模型的应用

差分隐私是一种广泛适用于以下情况的方法：需要可证明的隐私水平，而且针对数据访问及噪声添加是可行的。除了采用不同隐私模型来实现标准的统计分析工具（如平均值、标准偏差及直方图）之外，还可定制适用于特定应用的时常不同的隐私系统，这些应用采用统计工具作为其逻辑的一部分。比如数据挖掘工具（如聚类算法）及机器学习算法（如决策树、支持向量机及回归）。

去标识化模型具有需要在实施时加以确定的一些参数（如K-匿名的k，差分隐私的ε）。选择这些参数值取决于重标识的总体风险和特定用例中的应用要求。

谢谢大家，一起交流学习！

QQ: 10068 0 2383